MỞ ĐẦU

1. Tính cấp thiết của luận án

Số lượng bài báo khoa học được công bố ngày nay đang gia tăng với tốc độ chưa từng có, dẫn đến thách thức đáng kể cho các nhà nghiên cứu, đặc biệt là những người trẻ và thiếu kinh nghiệm, trong việc xác định các tài liệu liên quan và có chất lượng cao để trích dẫn. Trước tình trạng quá tải thông tin từ hàng loạt ấn phẩm khoa học được công bố mỗi năm, các hệ thống khuyến nghị trích dẫn tự động có tiềm năng giảm bớt gánh nặng này. Những hệ thống này có thể cung cấp các đề xuất phù hợp, hỗ trợ các nhà nghiên cứu định hướng hiệu quả trong khối lượng thông tin khổng lồ.

Các phương pháp tiếp cận hiện nay đối với bài toán khuyến nghị trích dẫn vẫn tồn tại một số hạn chế. Hạn chế đầu tiên nằm ở việc các mô hình khuyến nghị chưa tận dụng đầy đủ thông tin từ các bài báo khoa học. Một trong những nghiên cứu tiên phong trong lĩnh vực này được thực hiện bởi Ebesu [10] và Färber [11], trong đó họ đề xuất một kiến trúc linh hoạt dựa trên cơ chế mã hóa-giải mã (*encoder-decoder*) có tên là mạng nơ-ron trích dẫn (Neural Citation Network - NCN). Mặc dù mô hình này đã đạt hiệu quả vượt trội so với các phương pháp cùng thời trên các bộ dữ liệu RefSeer và arXiv CS, nó vẫn còn những hạn chế đáng kể, đặc biệt là việc chưa tích hợp toàn diện các thông tin quan trọng từ bài báo vào quá trình huấn luyện mô hình, chẳng hạn như tiêu đề, tác giả, năm xuất bản và nơi công bố.

Hạn chế thứ hai liên quan đến việc các mô hình khuyến nghị hiện tại chưa tận dụng những tiến bộ mới nhất trong lĩnh vực học sâu. Chẳng hạn, các mô hình khuyến nghị kép như DualLCR [12] và DualLCR-design [13], được nhóm Medić và Šnajder giới thiệu lần lượt vào năm 2020 và 2022, vẫn dựa trên cơ chế Bộ nhớ dài-ngắn hai chiều (Bidirectional Long-Short Term Memory, BiLSTM) [14]. Tương tự, mô hình BERT-GCN do nhóm nghiên cứu Jeong [15] phát triển cũng chưa tích hợp các tiến bộ mới nhất về xử lý ngôn ngữ tự nhiên và đồ thị liên kết trích dẫn trong các bài báo khoa học.

Hạn chế thứ ba liên quan đến việc các mô hình khuyến nghị trích dẫn hiện nay chủ yếu tập trung vào ngữ cảnh trích dẫn và nội dung của bài báo ứng viên [16] [17], trong khi chưa khai thác hiệu quả siêu dữ liệu của bài báo, bao gồm tên tác giả, năm xuất bản và nơi công bố. Những yếu tố này có vai trò quan trọng trong việc định hình xu hướng trích dẫn của các nhà khoa học, bởi lẽ họ thường ưu tiên trích dẫn các tác giả có uy tín, các công bố mới hoặc các bài báo đăng tải tại các tạp chí hoặc hội nghị hàng đầu trong lĩnh vực nghiên cứu của mình.

2. Mục tiêu của luận án

Áp dụng các tiến bộ mới nhất từ các mô hình học sâu để phát triển một mô hình hoàn toàn mới hoặc đề xuất các giải pháp cải thiện hiệu năng cho các mô hình khuyến nghị trích dẫn tiên tiến.

3. Đối tượng và phạm vi nghiên cứu của luận án

Luận án tập trung nghiên cứu và phân tích một số khía cạnh liên quan đến bài toán khuyến nghị trích dẫn, bao gồm:

- Các mô hình học sâu tiên tiến hiện có dành cho bài toán khuyến nghị trích dẫn.
- Các cải tiến trong mô hình học sâu, những tiến bộ nổi bật trong xử lý ngôn ngữ tự nhiên, cùng các phương pháp biểu diễn dữ liệu khác nhau từ bài báo khoa học.
- Các chỉ số đánh giá hiệu suất và các bộ dữ liệu thường được sử dụng trong các mô hình khuyến nghị trích dẫn tiên tiến hiện nay.

4. Phương pháp nghiên cứu

Nghiên cứu lý thuyết: Tập trung nghiên cứu và phân tích các kết quả hiện có của các hệ thống khuyến nghị trích dẫn tiên tiến hiện nay, đánh giá ưu nhược điểm của các hệ thống này và đề xuất các phương án cải tiến nhằm nâng cao hiệu suất và độ chính xác của kết quả khuyến

nghị thông qua việc ứng dụng các kỹ thuật và mô hình học sâu. Đồng thời, xem xét các chỉ số đánh giá hiệu suất và các bộ dữ liệu phổ biến được sử dụng trong các mô hình khuyến nghị trích dẫn.

Nghiên cứu thực nghiệm: Thực hiện cài đặt và triển khai các mã nguồn trên các bộ dữ liệu phổ biến trên môi trường thực nghiệm, nhằm đo lường và đánh giá các kết quả đạt được từ các phương án đề xuất.

5. Các đóng góp của luận án

Với mục tiêu cải thiện hiệu suất của các mô hình khuyến nghị trích dẫn hiện đại, luận án đã có những đóng góp đáng kể như sau:

- Theo hướng tiếp cập lọc nội dung, đưa ra các giải pháp nâng cao hiệu suất cho mô hình mạng nơ-ron trích dẫn NCN [10] [11] (công bố trong công trình CT1).
- Theo hướng tiếp cận lọc nội dung kết hợp lọc đồ thị, phát triển một mô hình mới có tên RHN-DualLCR, bao gồm các giải pháp cải thiện hiệu suất cho mô hình khuyến nghị trích dẫn kép DualLCR đã được Medić và Šnajder công bố trước đó [12] [13] (công bố trong công trình CT2 và CT4).
- Theo hướng tiếp cận lọc nội dung và lọc đồ thi, giới thiệu mô hình khuyến nghị trích dẫn mới có tên SciBERT-GraphSAGE, bằng cách kết hợp hai tiến bộ gần đây trong xử lý ngôn ngữ tự nhiên cho bài báo khoa học SciBERT [18] và cấu trúc đồ thị GraphSAGE [19] (công bố trong công trình CT3 và CT5).

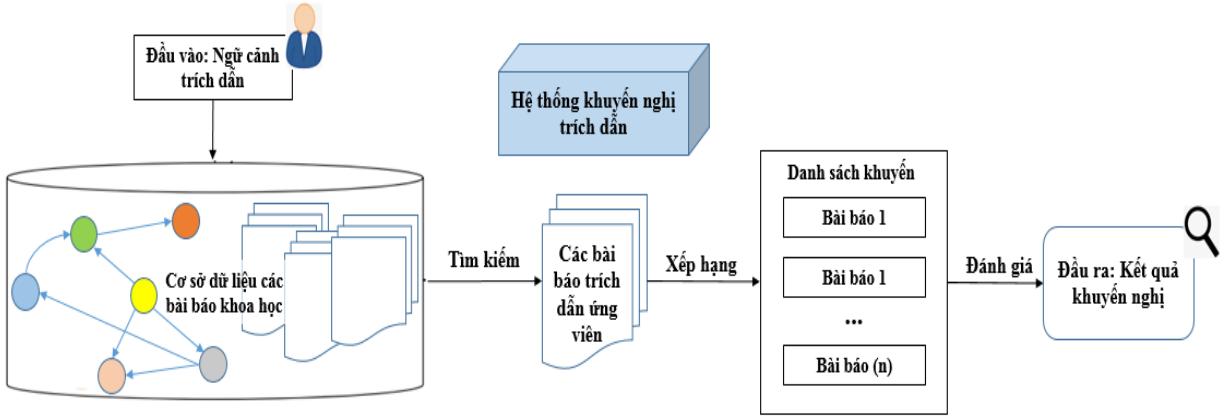
6. Bố cục của luận án

Luận án bao gồm phần mở đầu và các chương nội dung chính được sắp xếp như sau: Chương 1 trình bày tổng quan các nghiên cứu liên quan, phân tích những hạn chế của các kết quả nghiên cứu trước đây. Các chương 2, 3 và 4 tập trung vào các đóng góp chính của luận án, mỗi chương trình bày các phương pháp được đề xuất nhằm cải thiện hiệu quả của các mô hình khuyến nghị hiện đại. Phần kết luận tổng hợp những đóng góp chính của luận án, đề xuất các hướng nghiên cứu phát triển trong tương lai và nêu những vấn đề quan tâm của NCS. Cuối cùng, luận án liệt kê danh mục các công trình đã công bố của NCS và tài liệu tham khảo.

Chương 1. TỔNG QUAN NGHIÊN CỨU

1.1. Giới thiệu bài toán khuyến nghị trích dẫn

Bài toán khuyến nghị trích dẫn (*citation recommendation*) được nhóm nghiên cứu của McNee đưa ra lần đầu vào năm 2002 [1]. Theo nghiên cứu này, hoạt động của mô hình khuyến nghị trích dẫn điển hình như được mô tả trong Hình 1.1 như sau:.



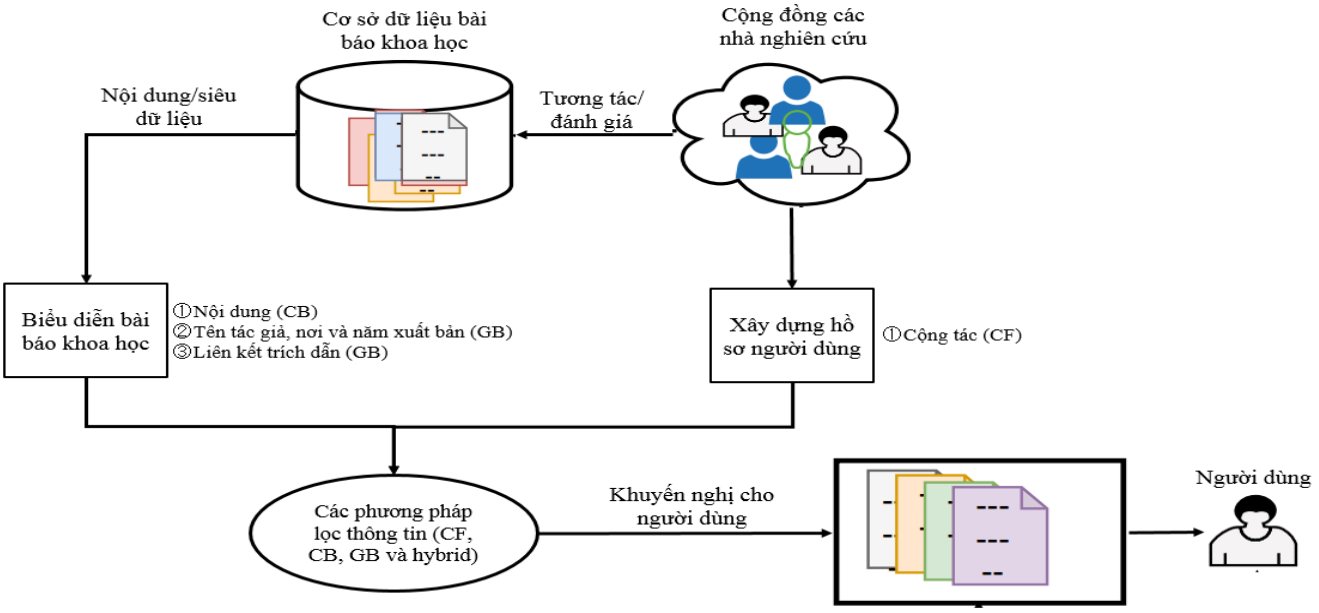
Hình 1.1 Sơ đồ luồng xử lý của mô hình khuyến nghị trích dẫn

Nhìn chung, mục tiêu mô hình khuyến nghị trích dẫn là đề xuất các bài báo/trích dẫn cho người dùng bằng cách khai thác sở thích và mối quan tâm nghiên cứu của họ. Về mặt hình thức, mô hình khuyến nghị trích dẫn có thể được định nghĩa: (P) là một tập hợp các bài báo có thể được đề xuất cho các nhà nghiên cứu (U) và (Γ) là một hàm tiện ích đo lường mức độ hữu

ích của một bài báo $(p_i) \in (P)$ đối với một người dùng cụ thể $(u_i) \in (U)$. Về mặt toán học, nó có thể được biểu diễn dưới dạng $(\Gamma) = (U) \times (P) \rightarrow (K)$, trong đó (K) là tập hợp khuyến nghị. Đối với người dùng $(u) \in (U)$, mô hình đề xuất một số bài báo $(p_i) \in (P)$ mà tối đa hóa (Γ) cho người dùng, thường được biểu diễn thông qua xếp hạng do người dùng đưa ra.

1.2. Tổng quan các nghiên cứu liên quan hiện nay

Nhóm của Beel [6] đã phân loại các mô hình khuyến nghị trích dẫn dựa trên các phương pháp mà mô hình áp dụng: lọc cộng tác (*collaborative filtering, CF*), lọc nội dung (*content-based filtering, CB*), lọc dựa trên đồ thị (*graph-based filtering, GB*) và mô hình kết hợp (*hybrid*).



Hình 1.2. Mô hình khuyến nghị trích dẫn trong đó nội dung bài báo và hồ sơ người dùng được khai thác bằng các phương pháp lọc thông tin khác nhau

1.2.1. Mô hình lọc cộng tác

Mô hình lọc cộng tác đưa ra các khuyến nghị bằng cách tận dụng xếp hạng trước đây của người dùng và xếp hạng từ những người dùng khác. Sự tương đồng giữa người dùng và hạng mục được xác định thông qua ma trận xếp hạng người dùng-hạng mục (user-item matrix), được duy trì và cập nhật thường xuyên để đảm bảo tính chính xác của các khuyến nghị. Tuy nhiên, các mô hình này thường gặp khó khăn trong trường hợp dữ liệu thưa thớt, khi có quá ít thông tin đánh giá về các tài liệu nghiên cứu [7][8][9].

1.2.2. Mô hình lọc nội dung

Mô hình CB phân tích nội dung của tài liệu truy vấn và tìm các tài liệu tương tự. Mô hình này thực hiện theo các bước: ①Nhúng tài liệu (embedding): chuyển đổi văn bản thành vectơ số đại diện cho nội dung của bài báo (Doc2vec) ⇒②Tìm hàng xóm gần nhất: xác định hàng xóm gần nhất (trích dẫn tiềm năng) của nó trong không gian vectơ ⇒③Xếp hạng lại trích dẫn tiềm năng (Okapi BM25) ⇒ ④Khuyến nghị: theo danh sách đã xếp hạng

Mô hình CB hoàn toàn tập trung vào nội dung của bài báo và không yêu cầu các siêu dữ liệu như địa điểm, thời điểm công bố hay số lần trích dẫn. Điều này làm cho mô hình đặc biệt hữu ích trong trường hợp siêu dữ liệu không đầy đủ hoặc bị thiếu [10][11][12][13][14]. Tuy nhiên, mô hình này cũng tồn tại một số hạn chế như: không tận dụng siêu dữ liệu; chưa ứng dụng đầy đủ các thành tựu mới trong xử lý ngôn ngữ tự nhiên; và chưa khai thác toàn bộ các thông tin không phải siêu dữ liệu, chẳng hạn như tiêu đề bài báo.

1.2.3. Mô hình lọc dựa trên đồ thị

Mô hình lọc dựa trên đồ thị tận dụng liên kết trích dẫn để khuyến nghị các bài báo có liên quan [15][16][17][18][19][20]. Mô hình này thực hiện các bước ①Xây dựng đồ thị: xây dựng

các nút trong đó biểu thị các bài báo và các cạnh biểu thị liên kết trích dẫn giữa chúng ⇒② Nhúng nút (node embedding): Các bài báo được nhúng vào không gian vectơ bằng các kỹ thuật như GCN, HIN, GAT, GraphSAGE…⇒③Tính toán độ tương tự giữa các vectơ nhúng để xác định các trích dẫn tiềm năng⇒④Xếp hạng dựa trên điểm tương đồng và các xếp hạng cao được đề xuất làm trích dẫn.

Phương pháp này khai thác hiệu quả các mối quan hệ trích dẫn giữa các bài báo, cung cấp thông tin sâu sắc về mức độ liên quan và tác động của bài báo trong lĩnh vực nghiên cứu.

1.2.4. Mô hình kết hợp

Mỗi loại mô hình đều có những ưu và nhược điểm riêng, do đó, việc kết hợp các kỹ thuật từ mô hình lọc cộng tác (CF), lọc nội dung (CB) và lọc dựa trên đồ thị (GB) là xu hướng tất yếu nhằm khai thác tối đa thông tin từ các bài báo. Các nghiên cứu tiêu biểu theo hướng tiếp cận này bao gồm các mô hình như DualLCR (CB+CF) [21][22], BERT-GCN (CB+GB) [23], MP-BERT4CR (CB+GB) [24], và RecCite (CB+CF) [25]. Tuy nhiên, các mô hình kết hợp này vẫn tồn tại một số hạn chế, chẳng hạn như chưa tận dụng triệt để các thông tin bổ sung của bài báo hoặc chưa khai thác đầy đủ các thành tựu mới nhất trong học sâu, đặc biệt là trong xử lý ngôn ngữ tự nhiên và mạng tích chập đồ thị.

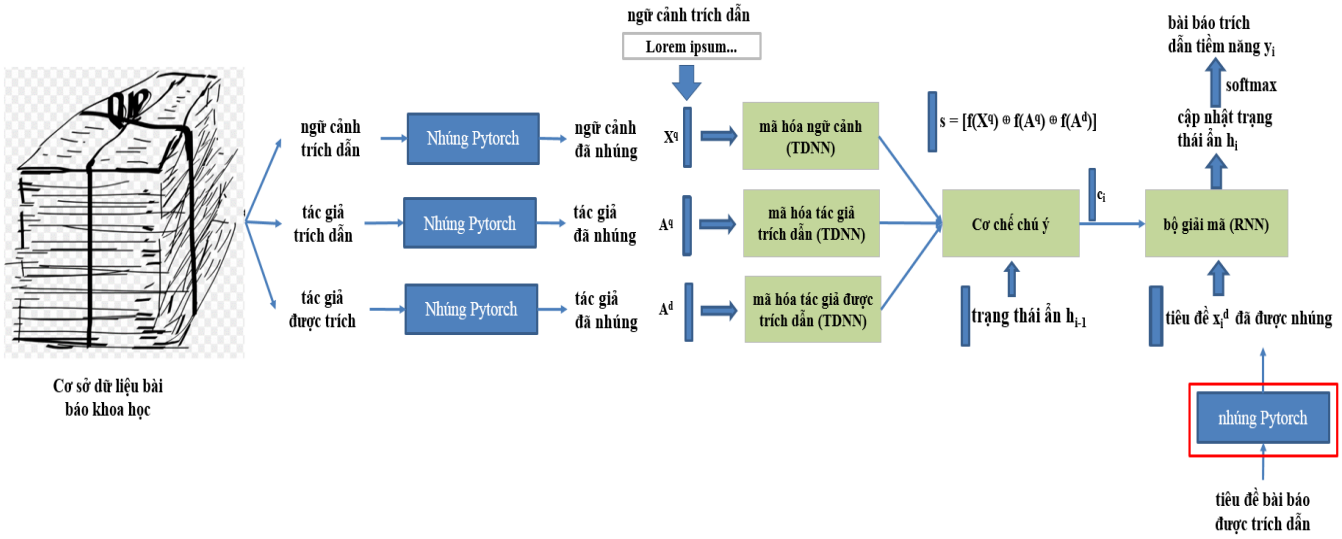
Chương 2. MÔ HÌNH ENHANCED-NCN BỔ SUNG THÊM THÔNG TIN TIÊU ĐỀ VÀ SỬ DỤNG PHÉP NHÚNG BERT

2.1. Mở đầu

Chương 2 trình bày chi tiết về đề xuất cải tiến mô hình NCN của hai nhóm nghiên cứu Ebesu [10] và Färber [11] bằng cách bổ sung thêm thông tin của bài báo và sử dụng phép nhúng BERT. Các kết quả trong chương này được công bố trong công trình CT1.

2.2. Phân tích vấn đề tồn tại của mô hình NCN

Mô hình mạng nơ-ron trích dẫn (*Neural Citation Network - NCN*) là một trong những mô hình đầu tiên được công bố để giải quyết bài toán khuyến nghị trích dẫn. NCN lần đầu tiên được giới thiệu vào năm 2017 bởi nhóm nghiên cứu của Ebesu và Yi Fang [10], và sau đó được cải tiến vào năm 2020 bởi nhóm nghiên cứu của Färber [11]. Như mô tả trong Hình 2.1, mô hình NCN bao gồm ba thành phần chính: bộ mã hóa, bộ giải mã, và cơ chế chú ý.



Hình 2.1. Kiến trúc tổng thể của mô hình NCN

2.2.1. Bộ mã hóa

Bộ mã hóa trong mô hình NCN được thiết kế nhằm chuyển đổi ngữ cảnh trích dẫn và tên tác giả được trích dẫn hoặc đang được trích dẫn thành các đặc trưng đại diện chứa thông tin quan trọng về ngữ cảnh và tác giả tương ứng. Bộ mã hóa này bao gồm hai thành phần chính: mã hóa ngữ cảnh trích dẫn (*citation context encoding*) và mã hóa tác giả (*author encoding*).

Mã hóa ngữ cảnh trích dẫn chịu trách nhiệm mã hóa bối cảnh trích dẫn trong các bài báo khoa học. Thành phần này sử dụng mạng nơ-ron có độ trễ thời gian (*Time-Delay Neural Network - TDNN*) do nhóm nghiên cứu Collobert [64] giới thiệu. TDNN cho phép lan truyền song song qua mạng, giúp tính toán đồng thời tất cả các ánh xạ đặc trưng (*feature maps*). Trong mô hình NCN, TDNN bao gồm một lớp chập (*convolutional layer*), tiếp theo là lớp gộp (*pooling layer*) và lớp kết nối đầy đủ (*fully connected layer*).

Để tạo ra các đề xuất trích dẫn bao gồm thông tin tác giả, NCN cũng tích hợp một bộ mã hóa tác giả, có kiến trúc tương tự như bộ mã hóa ngữ cảnh. Bộ mã hóa tác giả được áp dụng cho (1) phần nhúng tên tác giả ($\boldsymbol{A}^q$) của tài liệu từ ngữ cảnh truy vấn và (2) phần nhúng tên tác giả ($\boldsymbol{A}^d$) của tất cả các bài báo trong cơ sở dữ liệu. Quá trình mã hóa tác giả được thực hiện nhiều lần bằng cách sử dụng TDNN với các kích thước bộ lọc vùng khác nhau trong lớp chập. Biểu diễn cuối cùng của văn bản được ký hiệu là kết quả của việc tích hợp mã hóa ngữ cảnh và mã hóa tác giả.

$$\mathbf{s}_j = [f(\boldsymbol{X}^q) \oplus f(\boldsymbol{A}^q) \oplus f(\boldsymbol{A}^d)]_j \quad (2.1)$$

trong đó ($\boldsymbol{X}^q$) biểu diễn cho một ngữ cảnh trích dẫn.

2.2.2. Bộ giải mã

Bộ giải mã trong mô hình NCN là một mạng nơ-ron hồi quy (*Recurrent Neural Network - RNN*) sử dụng đơn vị hồi quy có kiểm soát (*Gated Recurrent Units - GRU)* [65] làm cơ chế kiểm soát (*gating mechanism*) và tích hợp cơ chế chú ý [66]. Bộ giải mã này được áp dụng cho tiêu đề của tất cả các tài liệu tiềm năng có thể được sử dụng làm trích dẫn cho ngữ cảnh truy vấn. Chức năng chính của bộ giải mã là tạo ra điểm số cho mỗi tài liệu trong cơ sở dữ liệu nhằm xác định mức độ phù hợp của tài liệu đó như một trích dẫn cho một ngữ cảnh truy vấn cụ thể. Các điểm số này sau đó có thể được sử dụng để đề xuất trích dẫn phù hợp với ngữ cảnh truy vấn.

2.2.3. Cơ chế chú ý

NCN sử dụng cơ chế chú ý được giới thiệu ban đầu bởi nhóm của Bahdanau [66]. Với cơ chế chú ý này, các mã hóa (s_j) bắt nguồn từ bộ mã hóa ngữ cảnh và tác giả được gán cho các trọng số phụ thuộc vào đầu ra (h_{i-1}) của bộ giải mã cho từ đứng trước (i). Kết quả là một vectơ ngữ cảnh (c_i) được tạo thành từ tổng có trọng số của đầu ra bộ mã hóa (s_j) theo mức độ liên quan của chúng. Cơ chế chú ý được sử dụng để nhấn mạnh vào các mã hóa đặc biệt quan trọng đối với bước thời gian hiện tại. Cơ chế chú ý được xây dựng dưới dạng mạng nơ-ron truyền thẳng FNN kết thúc bằng lớp softmax để chuyển đổi vectơ chú ý (a_{ij}) thành điểm chú ý (α_{ij}). Những điều này cho thấy tầm quan trọng của đầu ra bộ mã hóa (s_j) đối với từ thứ (i) trong tiêu đề của bài báo hiện đang được giải mã.

Mô hình NCN sử dụng cơ chế chú ý được giới thiệu lần đầu bởi nhóm nghiên cứu của Bahdanau [66]. Cơ chế này gán trọng số cho các mã hóa (s_j) được tạo ra bởi bộ mã hóa ngữ cảnh và tác giả, dựa trên đầu ra (h_{i-1}) của bộ giải mã từ thời điểm trước ($i-1$). Kết quả của cơ chế chú ý là một vectơ ngữ cảnh (c_i), được tính toán dưới dạng tổng có trọng số của các đầu ra từ bộ mã hóa (sj) dựa trên mức độ liên quan của chúng. Cơ chế chú ý này được thiết kế để tập trung vào các mã hóa quan trọng nhất đối với thời điểm hiện tại trong chuỗi thời gian. Nó được triển khai dưới dạng một mạng nơ-ron truyền thẳng (*Feedforward Neural Network - FNN*) và kết thúc bằng một lớp *softmax*, nhằm chuyển đổi vectơ chú ý (a_{ij}) thành các điểm chú ý (α_{ij}). Những điểm này thể hiện mức độ quan trọng của mỗi đầu ra từ bộ mã hóa (sj) đối với từ thứ (i) trong tiêu đề của bài báo đang được giải mã.

2.2.4. Hạn chế của mô hình NCN

Mặc dù NCN là một trong những mô hình khuyến nghị trích dẫn nổi tiếng và đã được trích dẫn trong hơn 170 công trình nghiên cứu, nhưng mô hình này vẫn tồn tại một số hạn chế đáng kể như sau:

(1) Biến đổi nhúng dữ liệu văn bản:

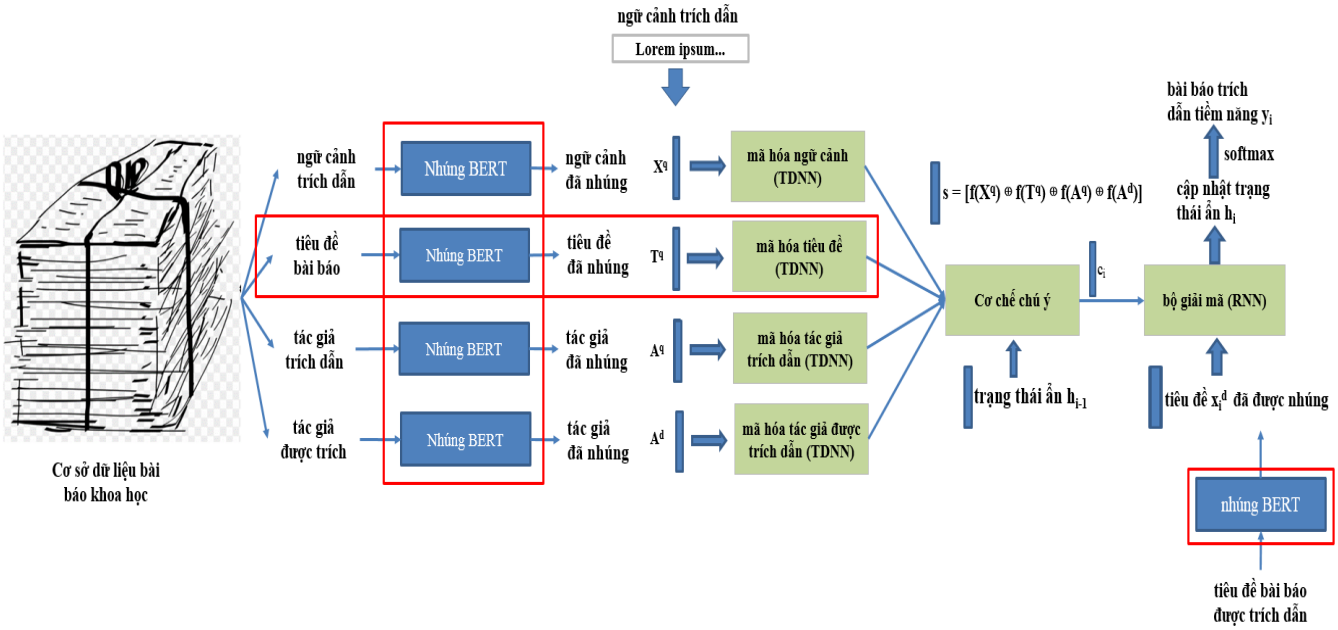
Dữ liệu bài báo ở dạng văn bản cần được biến đổi thành các biểu diễn nhúng (embedding) trước khi đưa vào bộ mã hóa. Tuy nhiên, mô hình NCN hiện nay đang sử dụng hàm torch.nn.Embedding của thư viện PyTorch để thực hiện phép biến đổi này. Hàm torch.nn.Embedding là một công cụ tạo biểu diễn vectơ dày đặc cho các đối tượng rời rạc, thường được áp dụng trong các tác vụ xử lý ngôn ngữ tự nhiên để ánh xạ các biến phân loại (như từ hoặc chỉ mục) tới không gian vectơ liên tục. Tuy nhiên, việc sử dụng hàm này vẫn còn đơn giản và chưa tận dụng được các kỹ thuật nhúng tiên tiến hơn.

(2) Chưa tích hợp tiêu đề bài báo:

Tiêu đề của bài báo là thông tin rất quan trọng vì nó chứa đựng ý nghĩa cô đọng nhất về nội dung của toàn bộ bài báo. Tuy nhiên, như được trình bày trong Hình 2.1, kiến trúc của mô hình NCN hiện tại chưa tích hợp thông tin tiêu đề bài báo vào quá trình mã hóa. Hạn chế này ảnh hưởng đáng kể đến hiệu suất của mô hình, làm giảm khả năng khai thác tối đa thông tin quan trọng từ bài báo.

2.3. Cải tiến mô hình NCN

Dựa trên những phân tích về hạn chế của mô hình NCN hiện tại, nghiên cứu sinh (NCS) đã thực hiện hai cải tiến nhằm nâng cao hiệu suất của mô hình NCN: (1) Thay thế phép nhúng torch.nn.Embedding bằng phép nhúng của BERT [54], một thành tựu mới trong lĩnh vực xử lý ngôn ngữ tự nhiên, và (2) Bổ sung thông tin tiêu đề của bài báo trích dẫn vào mô hình để thực hiện mã hóa. Các phần cải tiến của mô hình NCN được minh họa trong Hình 2.2, với các thay đổi được đánh dấu trong vùng khoanh đỏ. Mô hình NCN sau khi cải tiến được gọi là Enhanced-NCN.



Hình 2.2. Kiến trúc tổng thể của mô hình Enhanced-NCN

2.3.1. Phép nhúng BERT

BERT (*Bidirectional Encoder Representations from Transformers* - Biểu diễn bộ mã hóa hai chiều từ bộ biến đổi) là một kỹ thuật học máy dựa trên bộ biến đổi (*Transformer*) được sử dụng trong huấn luyện trước và xử lý ngôn ngữ tự nhiên. BERT được công bố vào năm 2019 bởi Jacob và các cộng sự của Google [54]. Đây là một mô hình ngôn ngữ mạnh mẽ có khả năng tạo ra các phần nhúng theo ngữ cảnh cho các từ và câu từ dữ liệu văn bản. Phần nhúng này được biểu diễn dưới dạng các vectơ chiều thấp (*low-dimensional vector*), giúp nắm bắt ý nghĩa và mối quan hệ giữa các từ và câu, từ đó hỗ trợ tốt hơn cho các tác vụ hoặc mô hình liên quan khác.

2.3.2. Thêm tiêu đề bài báo vào mô hình

Mặc dù tiêu đề bài báo trích dẫn là một yếu tố quan trọng, cung cấp thông tin liên quan đến truy vấn của người dùng và hỗ trợ hệ thống khuyến nghị xác định kết quả trích dẫn phù

hợp, nhưng trong các phiên bản NCN trước đây, cả nhóm nghiên cứu của Ebesu và Yi Fang [10], cũng như nhóm của Färber [11], đều chỉ sử dụng thông tin về bối cảnh trích dẫn, tác giả trích dẫn và tác giả được trích dẫn. Tiêu đề bài báo trích dẫn đã bị bỏ qua trong quá trình nhúng dữ liệu trước khi đưa vào bộ mã hóa. Để cải thiện hiệu suất của NCN, NCS đã tích hợp khả năng mã hóa tiêu đề của bộ mã hóa vào mô hình Enhanced-NCN. Kết quả cuối cùng của văn bản sau khi xử lý trong Enhanced-NCN là sự kết hợp giữa mã hóa ngữ cảnh, mã hóa tiêu đề và mã hóa tác giả, được tính toán như sau:

$$\mathbf{s}_j = [f(\boldsymbol{X}^q) \oplus f(\boldsymbol{T}^q) \oplus f(\boldsymbol{A}^q) \oplus f(\boldsymbol{A}^d)]_j \quad (2.2)$$

trong đó ($\boldsymbol{T}^q$) biểu diễn cho tiêu đề của bài báo.

2.4. Tiến hành thực nghiệm với mô hình Enhanced- NCN

2.4.1. Xây dựng mô hình Enhanced-NCN

Mô hình Enhanced-NCN được phát triển dựa trên mã nguồn của mô hình NCN[1] từ nghiên cứu của nhóm Färber [11] thông qua việc tích hợp mô hình BERT, một trong những công cụ xử lý ngôn ngữ tự nhiên tiên tiến nhất hiện nay, đồng thời bổ sung bộ mã hóa dành cho tiêu đề bài báo và tích hợp chúng vào mô hình cải tiến Enhanced-NCN. Nghiên cứu sinh (NCS) sử dụng Python phiên bản 3.8.5 và PyTorch phiên bản 1.7.1 để xây dựng mô hình này. Đối với BERT, NCS đã sử dụng BertTokenizer và BertModel từ thư viện transformers của Python. Mô hình Enhanced-NCN cũng tận dụng thư viện torchtext để chuyển đổi dữ liệu thành định dạng phù hợp với PyTorch, hỗ trợ các bước tiền xử lý. Ngoài ra, NCS sử dụng thư viện SpaCy kết hợp với torchtext[2] để mã hóa dữ liệu.

Dữ liệu sau khi được bổ nghĩa hóa và loại bỏ từ dừng (*stopword*) bằng cách kết hợp SpaCy[3] và tập từ dừng nltk[4] được số hóa thông qua từ vựng BERT với kích thước 30,522 mã thông báo, áp dụng cho ngữ cảnh trích dẫn, tiêu đề bài báo và tác giả trích dẫn/được trích dẫn. Để tối ưu hóa việc xử lý dữ liệu theo lô (batch), NCS áp dụng kỹ thuật phân lô (*bucketing technique*) từng được sử dụng bởi Ebesu và Fang [10]. Tương tự, trong phần giải mã, NCS vẫn giữ nguyên chức năng xếp hạng Okapi BM25[5] để chọn lọc trước tiêu đề cho một ngữ cảnh trích dẫn nhất định, như trong nghiên cứu ban đầu của nhóm Ebesu và Fang [10].

2.4.2. Bộ dữ liệu thực nghiệm

Cả hai nhóm nghiên cứu của Ebesu [10] và Färber [11] đều sử dụng hai bộ dữ liệu RefSeer và arXiv CS trong các nghiên cứu của mình. Tuy nhiên, nhóm Färber [11] chỉ ra rằng bộ dữ liệu RefSeer không thể tái tạo từ nghiên cứu của nhóm Ebesu [10]. Do đó, tương tự như nhóm Färber, trong luận án này cũng chỉ đánh giá mô hình Enhanced-NCN trên bộ dữ liệu arXiv CS.

Bộ dữ liệu gốc arXiv CS[6] bao gồm 1,7 triệu bài báo khoa học thuộc lĩnh vực khoa học máy tính, với đa dạng các trường con và chủ đề nghiên cứu. Dữ liệu này chứa siêu dữ liệu quan trọng của từng bài báo, chẳng hạn như tiêu đề, tác giả, tóm tắt, danh mục và tài liệu tham khảo. Bộ dữ liệu này có thể phục vụ nhiều ứng dụng như phân tích xu hướng, khuyến nghị trích dẫn, dự đoán danh mục, xây dựng đồ thị tri thức và tìm kiếm ngữ nghĩa. Các bài báo trong tập dữ liệu trải dài từ tháng 1 năm 1993 đến tháng 4 năm 2021 và được cập nhật hàng tháng. Bộ dữ liệu này có định dạng JSON và có thể tải xuống từ Hugging Face[7]. Tập dữ liệu arXiv CS thu gọn được sử dụng để kiểm tra hiệu suất mô hình Enhanced-NCN bao gồm 502,355 bản ghi, với các thông tin: ngữ cảnh trích dẫn, tác giả trích dẫn, tên bài báo và tác giả

[1] https://github.com/timoklein/neural_citation
[2] https://pytorch.org/text/stable/index.html
[3] https://spacy.io/
[4] https://www.nltk.org/
[5] https://pypi.org/project/rank-bm25/
[6] https://arxiv.org/
[7] https://huggingface.co/datasets/arxiv_dataset

được trích dẫn. NCS đã giới hạn ngữ cảnh trích dẫn và tiêu đề bài báo ở độ dài lần lượt là 100 và 30 từ, nhằm cân bằng giữa hiệu suất mô hình và thời gian huấn luyện. Không giống như nghiên cứu của nhóm Färber [11], NCS đã tích hợp tiêu đề bài báo vào mô hình để thực hiện mã hóa tiêu đề, từ đó cải thiện đáng kể hiệu suất hệ thống. Để đào tạo và đánh giá mô hình, dữ liệu arXiv CS được chia thành ba phần: 80% để huấn luyện (*training*), 10% để xác nhận (*validation*), và 10% để kiểm tra (*test*).

2.4.3. Chỉ số đánh giá

Hầu hết các nghiên cứu về bài toán khuyến nghị trích dẫn sử dụng các chỉ số đánh giá phổ biến như mức tăng tích lũy chiết khấu chuẩn hóa (*Normalized Discounted Cumulative Gain - NDCG*), xếp hạng đối ứng trung bình (*Mean Reciprocal Rank - MRR*), độ chính xác trung bình (*Mean Average Precision - MAP*), Recall Top@K và Hits@K để đánh giá hiệu suất của mô hình. Trong nghiên cứu của nhóm Färber [11], chỉ số Recall@10 được sử dụng để đánh giá hiệu suất của NCN. Do đó, trong luận án này cũng chỉ sử dụng tiêu chí Top@10 để đánh giá hiệu suất của mô hình Enhanced-NCN sau khi cải tiến.

2.5. Đánh giá kết quả thực nghiệm với mô hình Enhanced- NCN

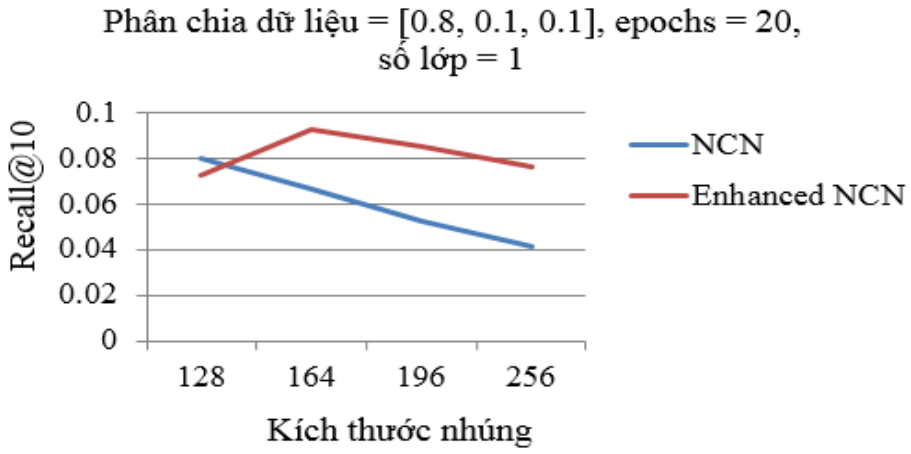
Để tìm kiếm cấu hình tham số tối ưu cho mô hình Enhanced-NCN, NCS đã thực hiện điều chỉnh bốn siêu tham số chính: tỉ lệ phân chia dữ liệu (*split data*) khi huấn luyện, số lượng lớp (*number of layers*), số lần lặp huấn luyện (*epoch*), và kích thước nhúng (*embedding size*). Quá trình này nhằm so sánh hiệu suất của mô hình Enhanced-NCN với kết quả từ nhóm Färber [11]. Như thể hiện trong Bảng 2.1, NCS đã thử nghiệm nhiều giá trị khác nhau của các siêu tham số này để tìm ra cấu hình tốt nhất cho mô hình đề xuất. Bối cảnh trích dẫn đóng vai trò quan trọng trong việc cung cấp thông tin để đưa ra các đề xuất trích dẫn. Tiêu đề bài báo được trích dẫn thường chứa đựng các thông tin liên quan trực tiếp đến nội dung cần trích dẫn và thường là yếu tố đầu tiên các nhà nghiên cứu chú ý đến khi tìm kiếm tài liệu phù hợp.

Bảng 2.1. So sánh kết quả mô hình NCN của nhóm Färber [11] và mô hình Enhanced-NCN

Tên mô hình	Điều chỉnh	Split data	Number of layers	Epochs	Embedding size	Recall@10
NCN của nhóm Färber [11]	Embedding size	[0.8, 0.1, 0.1]	1	20	128	**0.0801**
		[0.8, 0.1, 0.1]	1	20	164	0.0663
		[0.8, 0.1, 0.1]	1	20	196	0.0527
		[0.8, 0.1, 0.1]	1	20	256	0.0413
	Number of layers	[0.8, 0.1, 0.1]	2	20	128	**0.1074**
		[0.8, 0.1, 0.1]	3	20	128	0.0867
Enhanced-NCN	Embedding size	[0.8, 0.1, 0.1]	1	20	128	0.0723
		[0.8, 0.1, 0.1]	1	20	164	**0.0921**
		[0.8, 0.1, 0.1]	1	20	196	0.0853
		[0.8, 0.1, 0.1]	1	20	256	0.0763
	Number of layers	[0.8, 0.1, 0.1]	2	20	164	**0.1285**
		[0.8, 0.1, 0.1]	3	20	164	0.1115

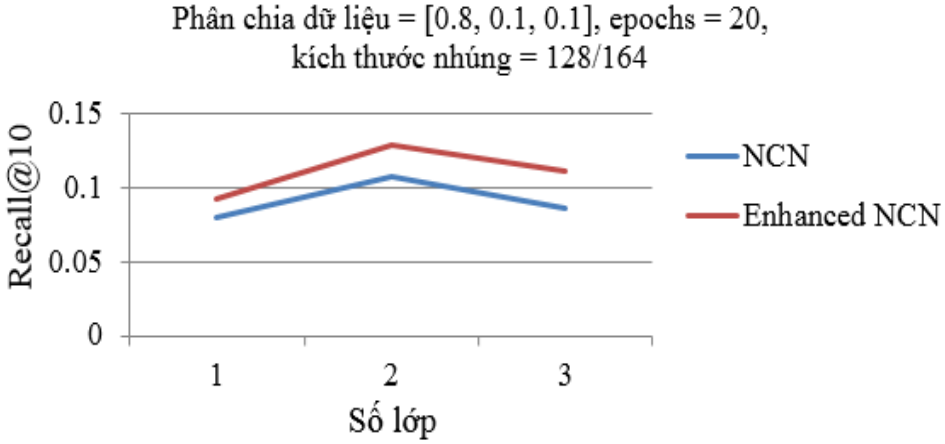
Kết quả trong Bảng 2.1 cho thấy rằng khi bối cảnh trích dẫn được xử lý tốt và tiêu đề bài báo được tích hợp vào mô hình Enhanced-NCN, hiệu suất của mô hình được cải thiện đáng kể so với mô hình NCN của nhóm Färber [11]. Sự gia tăng thông tin đầu vào trong mô hình Enhanced-NCN đòi hỏi việc tăng kích thước nhúng để đạt được kết quả tối ưu. Khi đặt số lớp là 1, NCS đã thử nghiệm với các kích thước nhúng khác nhau, lần lượt là 128, 164, 196 và 256. Kết quả tốt nhất đạt được khi kích thước nhúng = 164, với Recall@10 = 0.0921, cao hơn đáng

kể so với Recall@10 = 0.0801 (với kích thước nhúng = 128) của mô hình NCN gốc từ nhóm Färber [11]. Kết quả này được minh họa chi tiết trong Hình 2.3.



Hình 2.3. So sánh kết quả Recall@10 của mô hình Enhanced-NCN và mô hình NCN của nhóm Färber [11], được điều chỉnh về kích thước nhúng

Để tối ưu hóa Recall@10 cho mô hình Enhanced-NCN, NCS tiếp tục điều chỉnh số lượng lớp trong mô hình. Kết quả thực nghiệm cho thấy hiệu suất tốt nhất đạt được khi số lớp = 2, áp dụng cho cả mô hình NCN và Enhanced-NCN. Với cấu hình này, Recall@10 của mô hình Enhanced-NCN đạt 0.1285, cao hơn đáng kể so với Recall@10 = 0.1074 của mô hình NCN từ nhóm Färber [11]. Tuy nhiên, việc tăng số lớp lên 3 hoặc nhiều hơn không mang lại sự cải thiện nào thêm. Kết quả này được minh họa rõ ràng trong Hình 2.4.



Hình 2.4. So sánh kết quả Recall@10 của mô hình Enhanced-NCN và mô hình NCN của nhóm Färber [11], được điều chỉnh về số lượng lớp

2.6. Kết luận chương 2

Trong chương 2, NCS đã tiến hành cải tiến mô hình mạng nơ-ron trích dẫn (NCN), một mô hình tiên tiến được nhóm nghiên cứu của Ebesu và Fang [10] giới thiệu vào năm 2017 và được nhóm của Färber [11] cải tiến vào năm 2020. Cải tiến này bao gồm việc tích hợp mô hình BERT để tiền xử lý ngữ cảnh trích dẫn và bổ sung bộ mã hóa dành cho tiêu đề bài báo, cho phép sử dụng tiêu đề như một nguồn dữ liệu đầu vào quan trọng trong mô hình khuyến nghị trích dẫn. Hiệu suất của mô hình cải tiến Enhanced-NCN đã được đánh giá trên bộ dữ liệu arXiv CS. Kết quả thực nghiệm cho thấy mô hình Enhanced-NCN đạt hiệu suất cao hơn đáng kể so với mô hình của nhóm Färber [11] khi sử dụng cùng chỉ số đánh giá Recall Top@10. Ngoài ra, NCS đã cung cấp phân tích chi tiết về ảnh hưởng của các tham số khác nhau đối với hiệu suất của mô hình Enhanced-NCN. Thông tin này đóng vai trò quan trọng trong việc tối ưu

hóa mô hình, đồng thời định hướng cho các nghiên cứu trong tương lai nhằm cải thiện hiệu quả hoạt động của hệ thống khuyến nghị trích dẫn.

Chương 3. MÔ HÌNH RHN-DUALLCR SỬ DỤNG MẠNG HỒI QUY RHN VÀ PHÉP NHÚNG SCIBERT

3.1. Mở đầu

Chương 3 trình bày chi tiết về mô hình khuyến nghị trích dẫn RHN-DualLCR, được phát triển bằng cách cải tiến mô hình của nhóm nghiên cứu Medić và Šnajder [12]. Cải tiến này sử dụng mạng cao tốc hồi quy (Recurrent Highway Networks - RHN) và phép nhúng SciBERT để nâng cao hiệu suất của mô hình. Các kết quả nghiên cứu trong chương này đã được công bố trong công trình số CT2 và CT4.

3.2. Phân tích vấn đề tồn tại của mô hình DualLCR

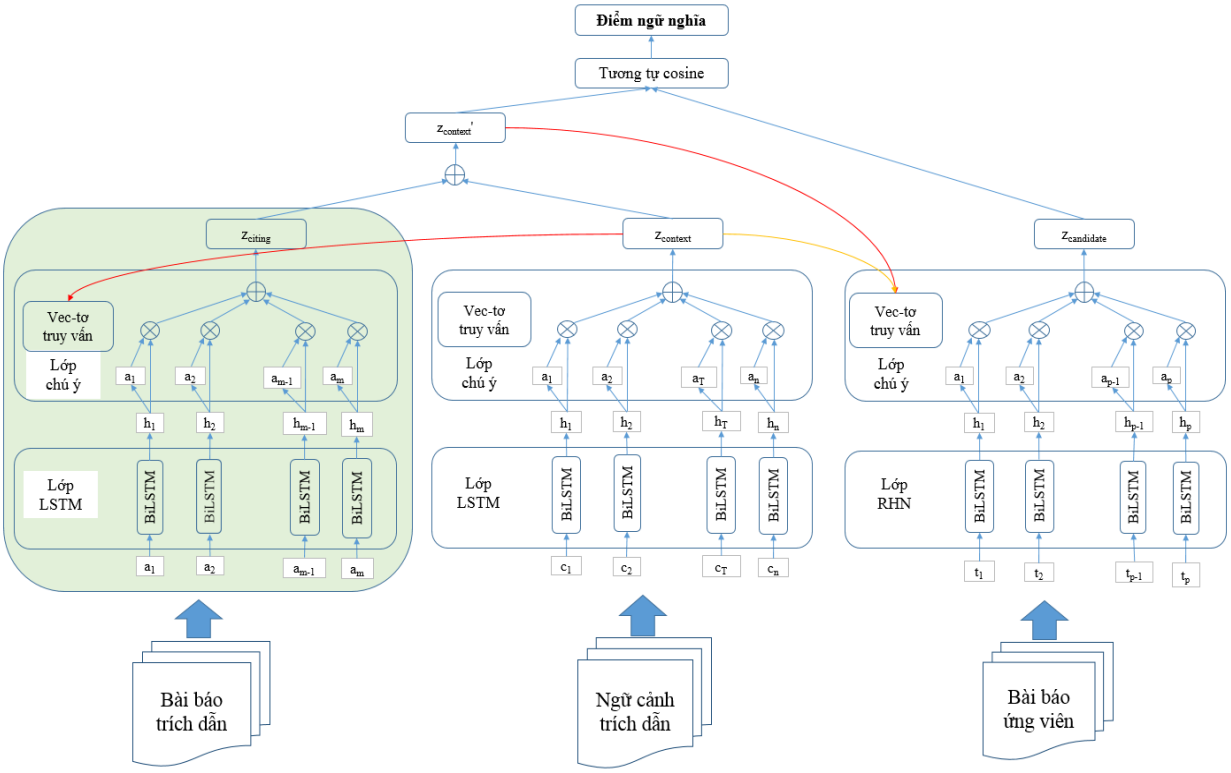
Mô hình DualLCR, được nhóm nghiên cứu Medić và Šnajder [12] công bố vào năm 2020, tập trung giải quyết bài toán khuyến nghị trích dẫn. Trong khi hầu hết các phương pháp trước đây chỉ sử dụng văn bản xung quanh vị trí trích dẫn để biểu diễn ngữ cảnh [15][56][49][44][16], Medić và Šnajder đã đề xuất cách biểu diễn ngữ cảnh tích hợp thêm thông tin toàn cục, chẳng hạn như tiêu đề và tóm tắt của bài báo được trích dẫn.

Để đưa ra đề xuất trích dẫn cho một ngữ cảnh cụ thể, đầu vào của mô hình DualLCR bao gồm năm loại thông tin: (1) ngữ cảnh trích dẫn dạng văn bản, (2) tiêu đề và tóm tắt của bài báo chứa trích dẫn (gọi là bài báo trích dẫn), (3) tiêu đề và tóm tắt của bài báo ứng cử viên, (4) danh sách tác giả của bài báo được trích dẫn, và (5) tần suất trích dẫn mà bài báo ứng cử viên nhận được trong vòng (y) năm gần nhất cùng tổng số lần trích dẫn của nó. Đầu ra của mô hình là tổng điểm khuyến nghị, cho biết mức độ phù hợp của bài báo ứng cử viên để được trích dẫn trong ngữ cảnh đầu vào.

Mô hình DualLCR được cấu trúc thành hai mô-đun chính: mô-đun ngữ nghĩa và mô-đun thông tin học thuật. Điểm khuyến nghị cuối cùng là tổng có trọng số của các điểm được tạo ra bởi hai mô-đun này. Trực giác đằng sau tổng điểm có trọng số là, tùy thuộc vào ngữ cảnh, các tác giả có thể ưu tiên trích dẫn những bài báo có ảnh hưởng lớn trong cộng đồng nghiên cứu (bài báo có điểm thông tin học thuật cao) hoặc những bài báo liên quan trực tiếp đến chi tiết cụ thể trong nghiên cứu của họ (chẳng hạn như các bài báo cung cấp cơ sở lý thuyết hoặc phương pháp mà họ sử dụng). Do đó, trong trường hợp đầu tiên, mô hình sẽ đặt trọng số cao hơn cho điểm thông tin học thuật, trong khi ở trường hợp thứ hai, trọng số sẽ nghiêng nhiều hơn về điểm ngữ nghĩa.

3.2.1. Mô-đun ngữ nghĩa

Tương tự như nhóm nghiên cứu của Dai [17], mô hình DualLCR cũng sử dụng bộ nhớ dài-ngắn hai chiều BiLSTM [14] để biểu diễn bối cảnh trích dẫn, nhưng cũng để thể hiện cả bài viết được trích dẫn và thông tin toàn cầu từ bài viết trích dẫn. Văn bản trước khi đưa vào mô-đun ngữ nghĩa được phân đoạn và mã hóa bằng thư viện SpaCy. Trích dẫn mục tiêu và các trích dẫn khác được che dấu bằng các phần giữ chỗ TARGETCIT và OTHERCIT tương ứng. Tất cả ba đầu vào dạng văn bản đều được truyền qua hai lớp giống nhau: BiLSTM và lớp chú ý (*attention layer*).



Hình 3.1. Cấu trúc của mô-đun ngữ nghĩa trong mô hình DualLCR [12]

Gọi (n) là tổng số mã thông báo của chuỗi đầu vào, ký hiệu là: $s = (t_1,...,t_n)$. Mỗi mã thông báo (t_i) được ánh xạ tới vectơ nhúng d_e chiều $x_i \in \mathbb{R}^{d_e}$ để tạo ra một chuỗi có dạng như $x = (x_1, \ldots, x_n)$ bằng cách sử dụng các phần nhúng được huấn luyện trước từ của phép nhúng Bhagavatula [60]. Sau đó, chuỗi đã cho (x) được chuyển qua lớp BiLSTM với kích thước trạng thái ẩn là (d_h), trong đó đầu ra ($\boldsymbol{h}_i$) ở mỗi bước (i) được hình thành bằng cách nối các trạng thái ẩn tiến và lùi: $\boldsymbol{h}_i = [\rightarrow \boldsymbol{h}_i; \leftarrow \boldsymbol{h}_i], \boldsymbol{h}_i \in \mathbb{R}^{2d_h}$. Các trạng thái ẩn của chuỗi đầu vào ($\boldsymbol{s}$) được chuyển qua lớp chú ý phụ [66] để tạo ra chuỗi nhúng cuối cùng ($\boldsymbol{z}_s$). Cho vectơ truy vấn đầu vào ($\boldsymbol{q}$) và vectơ trạng thái ẩn ($\boldsymbol{h}_i$), điểm chú ý (*attention score*) cho mỗi bước (i) là:

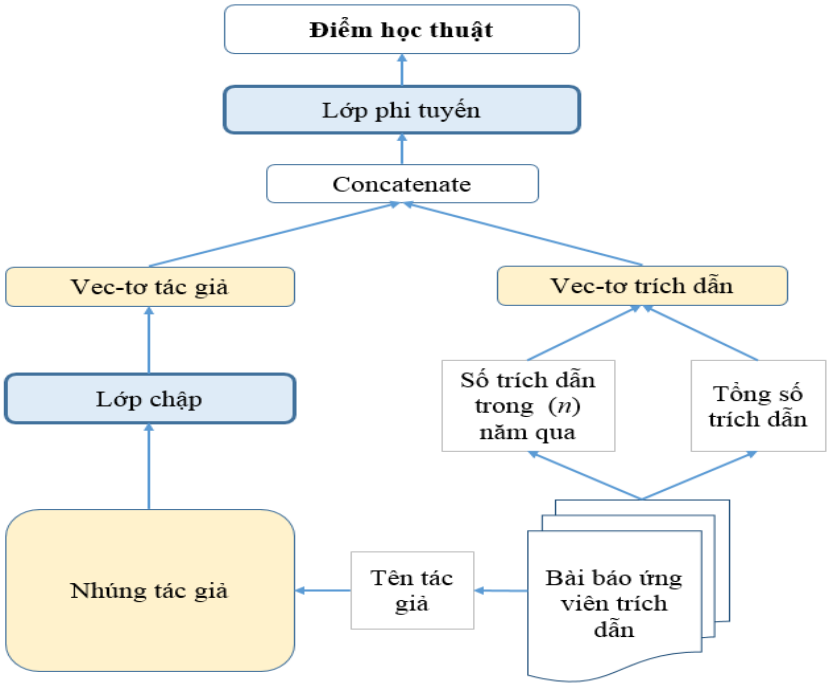
$$a_i = \boldsymbol{v} \cdot tanh(W \cdot [\boldsymbol{q}; \boldsymbol{h}_i]) \quad (3.1)$$

Trong đó $\boldsymbol{v}$ và W là tham số của mô hình DualLCR. Điểm chú ý được chuẩn hóa, áp dụng cho các trạng thái ẩn tương ứng và được tính tổng để tạo ra chuỗi nhúng (z_s) cuối cùng. Một vectơ truy vấn (q) khác nhau được sử dụng tùy thuộc vào loại đầu vào. Để làm cho việc biểu diễn ngữ cảnh cụ thể hơn cho trích dẫn đang được dự đoán, mô hình DualLCR sử dụng trạng thái ẩn tương ứng với vị trí của trích dẫn (*citation's placeholder*) ($\boldsymbol{h}_T$) của ngữ cảnh trích dẫn. Đối với văn bản của bài báo trích dẫn, mô hình DualLCR sử dụng phép nhúng chuỗi (*sequence embedding*) cuối cùng của ngữ cảnh ($\mathbf{z}_{\text{context}}$), trong khi đối với văn bản của bài báo ứng viên để trích dẫn, tổng số bài viết trích dẫn và nội dung nhúng ngữ cảnh ($\mathbf{z}_{\text{citing}} + \mathbf{z}_{\text{context}}$) được sử dụng. Điều này cho phép mô hình DualLCR tập trung vào thông tin theo ngữ cảnh cụ thể trong cả bài viết trích dẫn và bài viết được trích dẫn. Cụ thể hơn, bằng cách sử dụng trạng thái ẩn của vị trí trích dẫn làm vectơ truy vấn để tính điểm chú ý đối với mã thông báo của ngữ cảnh trích dẫn, mô hình DualLCR sẽ tập trung vào mã thông báo trong ngữ cảnh có liên quan để có thể thu được biến đổi nhúng của vị trí trích dẫn. Tương tự, bằng cách sử dụng phép nhúng của ngữ cảnh trích dẫn để làm truy vấn các văn bản của bài báo trích dẫn hay được trích dẫn, mô hình DualLCR sẽ tập trung vào các mã thông báo trong văn bản có liên quan đến ngữ cảnh trích dẫn nhất định, vì văn bản của bài báo sẽ mô tả các khía cạnh khác nhau của một vấn đề nghiên cứu và không phải tất cả chúng đều cần phải phù hợp như nhau đối với bối cảnh trích dẫn hiện tại.

Với ngữ cảnh trích dẫn (c) và bài báo ứng cử viên (p), hàm chấm điểm ngữ nghĩa $s_{sem}(c, p)$ được định nghĩa là độ tương tự cosine giữa biến đổi nhúng của ngữ cảnh nâng cao và biến đổi nhúng của bài báo ứng cử viên.

3.2.2. Mô-đun thông tin học thuật

Khi bối cảnh ngữ nghĩa chấp nhận một số trích dẫn, Medić và Šnajder [12] cho rằng các tác giả nói chung sẽ muốn trích dẫn các bài báo nổi tiếng trong cộng đồng nghiên cứu. Đây chính là nhiệm vụ của mô-đun thông tin học thuật. Mô-đun này lấy tên tác giả và số lượng trích dẫn của bài báo trích dẫn ứng viên (p) làm đầu vào và tạo ra một điểm thông tin học thuật duy nhất.



Hình 3.2. Cấu trúc của mô-đun thông tin học thuật trong mô hình DualLCR [12]

Cấu trúc của mô-đun học thuật được thể hiện ở Hình 3.2. Tương tự như cách làm của Ebesu và Fang [10], mô hình DualLCR thể hiện tên tác giả bài báo dưới dạng biến đổi nhúng. Danh sách tên tác giả của bài báo (a) = (a_1, . . ., a_m) trước tiên được chuyển thành chuỗi các biến đổi nhúng của tên tác giả ($\boldsymbol{a_e}$) = (a_{e1}, ..., a_{em}). Tiếp theo, chuỗi ($\boldsymbol{a_e}$) được chuyển qua một lớp tích chập, sau đó là phép biến đổi tổng hợp tối đa và phi tuyến (*max-pooling and non-linear transformation*) để cuối cùng tạo ra biến đổi nhúng cho danh sách tên tác giả, rồi biến đổi nhúng này được kết hợp với tổng số trích dẫn và số lượng trích dẫn của bài viết trong (y) năm qua. Cuối cùng, toàn bộ vectơ được truyền qua lớp phi tuyến tính để tạo ra điểm thông tin học thuật $s_{bib}(p)$.

3.2.3. Điểm khuyến nghị cuối cùng

Điểm khuyến nghị tổng hợp cuối cùng $s_{fin}(c, p)$ được tính bằng tổng trọng số của điểm $s_{sem}(c, p)$ và $s_{bib}(p)$. Trọng số điểm có được bằng cách chuyển biến đổi nhúng của ngữ cảnh trích dẫn ($\mathbf{z}_{context}$) qua một lớp phi tuyến (*non-linear layer*) với hai giá trị ở đầu ra.

3.2.4. Các hạn chế của mô hình DualLCR

Trong khi hầu hết các mô hình khuyến nghị trích dẫn trước đây khi biểu diễn ngữ cảnh thì chỉ sử dụng văn bản xung quanh vị trí trích dẫn, thì mô hình DualLCR [12] đã tăng cường biểu diễn ngữ cảnh bằng thông tin tổng thể của bài báo. Cụ thể, mô hình DualLCR đã đưa tiêu đề và phần tóm tắt của bài báo trích dẫn vào phần biểu diễn ngữ cảnh, do đó, so với các mô hình hiện tại thì mô hình DualLCR đã được nâng lên đáng kể về mặt hiệu suất. Tuy nhiên, mô hình này vẫn còn những nhược điểm như sau:

(1) Bộ nhớ dài-ngắn hai chiều BiLSTM:

Mô hình DualLCR vẫn đang sử dụng bộ nhớ dài-ngắn hạn hai chiều BiLSTM [14] để biểu diễn ngữ cảnh tăng cường. BiLSTM thường được sử dụng cho các tác vụ xử lý ngôn ngữ tự nhiên như phân loại văn bản (*text classification*) hay dịch máy, tuy nhiên BiLSTM cũng có một số hạn chế như:

· BiLSTM có thể gặp phải vấn đề biến mất hoặc bùng nổ gradient, xảy ra khi gradient trở nên quá nhỏ hoặc quá lớn trong quá trình truyền ngược (*backpropagation*), làm cho mạng khó huấn luyện

· BiLSTM có thể tốn kém về mặt tính toán và tốn nhiều bộ nhớ vì nó yêu cầu hai lớp LSTM cho mỗi hướng và một lớp móc nối (*concatenation layer*) để hợp nhất các đầu ra.

· BiLSTM có thể nhạy cảm với nhiễu và các giá trị ngoại lệ trong dữ liệu vì nó dựa trên giả định rằng chuỗi đầu vào trơn tru và nhất quán, tuy nhiên văn bản của bài báo khoa học thì không như vậy.

· BiLSTM có thể gặp khó khăn trong việc lập mô hình phụ thuộc lâu dài vì thông tin từ các phần tử ở xa có thể bị loãng hoặc bị lãng quên theo thời gian.

(2) Phép nhúng ngữ cảnh AI2:

Mô hình DualLCR đang sử dụng phép nhúng được sử dụng trong nghiên cứu của nhóm Bhagavatula [60] để nhúng ngữ cảnh trích dẫn cũng như thông tin (bao gồm tiêu đề và phần tóm tắt) của bài báo trích dẫn và được trích dẫn. Đây là phép nhúng AI2 (*AI2 embedding*) được công bố từ năm 2017 bởi nhóm nghiên cứu đến từ Viện trí tuệ nhân tạo Allen[8] (*Allen Institute for Artificial Intelligence, AI2*). Phép nhúng AI2 dựa trên ký tự và có thể nắm bắt các đặc điểm phức tạp của việc sử dụng từ (như cú pháp và ngữ nghĩa) cũng như cách sử dụng này khác nhau tùy theo ngữ cảnh ngôn ngữ. Không giống như cách nhúng từ truyền thống, phép nhúng AI2 tạo ra các cách nhúng có chức năng của toàn bộ câu đầu vào, cung cấp sự hiểu biết phong phú hơn về nghĩa của từ. Tuy nhiên phép nhúng AI2 này vẫn bộc lộ những hạn chế như sau so với phép nhúng SciBERT:

· Hạn chế về ngữ cảnh: Các phần nhúng AI2 được đào tạo trên các ngữ liệu cụ thể, điều này có thể hạn chế tính hiệu quả của chúng trong các lĩnh vực hoặc bối cảnh không được thể hiện rõ trong dữ liệu đào tạo. Ví dụ: các phần nhúng được đào tạo trên văn bản web chung có thể không hoạt động tốt trên các văn bản khoa học chuyên ngành.

· Biểu diễn tĩnh: Nếu phần nhúng AI2 thuộc loại truyền thống, không theo ngữ cảnh (như BERT hoặc SciBERT), thì mỗi từ có một cách biểu diễn cố định bất kể ngữ cảnh của nó. Điều này có thể gây khó khăn cho các từ đa nghĩa (từ có nhiều nghĩa) vì sắc thái của ý nghĩa trong các bối cảnh khác nhau bị mất đi.

· Xu hướng: Các phần nhúng được đào tạo trước có thể phản ánh và duy trì các thành kiến có trong dữ liệu đào tạo. Đây là một vấn đề thiết yếu khi các phần nhúng được sử dụng trong quá trình ra quyết định hoặc trong bối cảnh mà sự công bằng và vô tư là rất quan trọng.

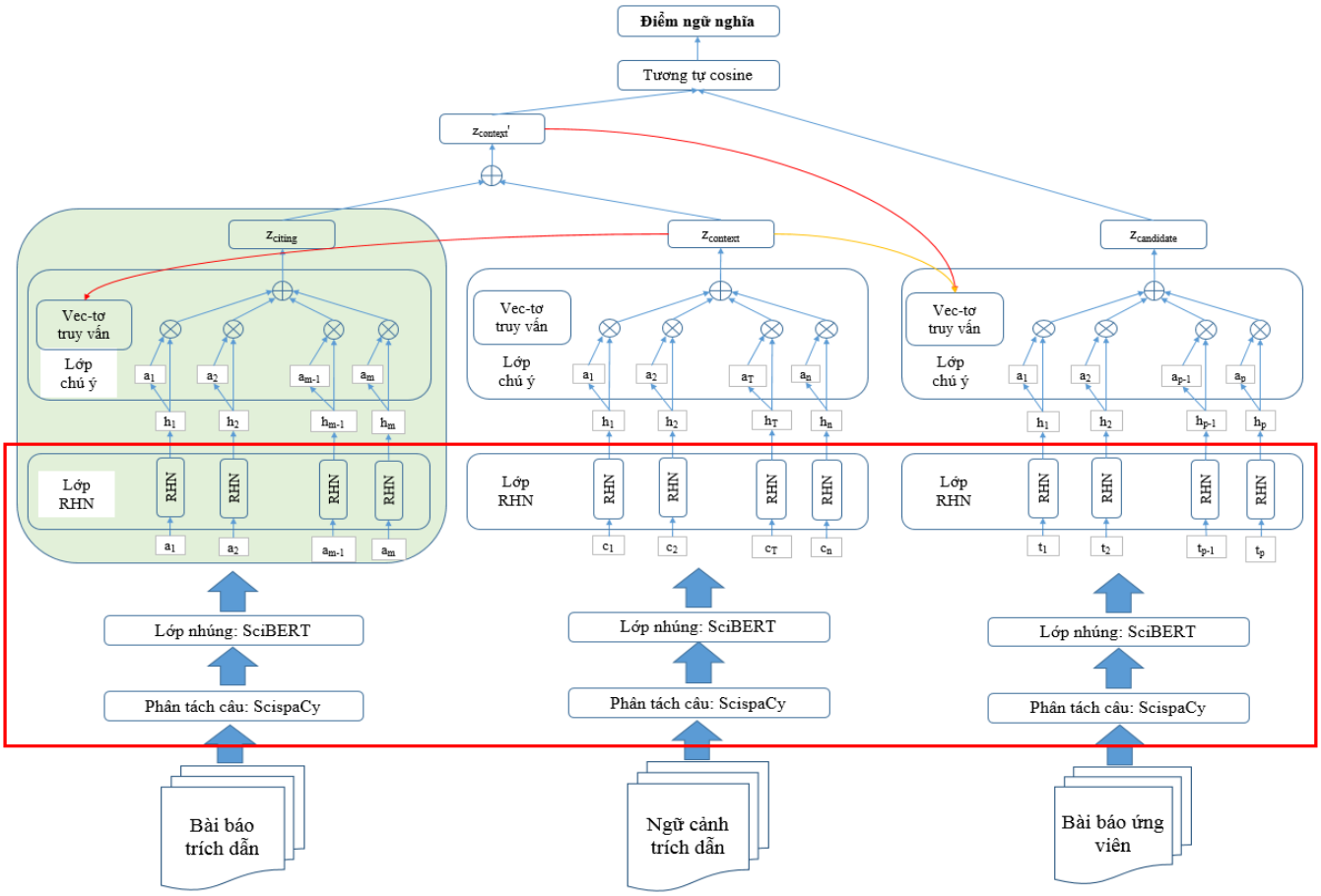
· Từ ngoài từ vựng: Các từ nhúng được huấn luyện trước gặp khó khăn với những từ không được nhìn thấy trong quá trình đào tạo. Điều này đặc biệt có vấn đề trong các lĩnh vực như khoa học và công nghệ, nơi các thuật ngữ mới liên tục được đặt ra.

3.3. Cải tiến mô hình DualLCR

Dựa trên những phân tích về hạn chế của mô hình DualLCR [12] hiện tại, NCS đã áp dụng hai phương pháp sau đây để nâng cao hiệu suất của mô hình DualLCR: (1) Thay thế bộ nhớ dài-ngắn hạn hai chiều BiLSTM bằng mạng cao tốc hồi quy RHN và (2) Thay thế phép

[8] https://allenai.org/

nhúng AI2 từ công bố của nhóm nghiên cứu Bhagavatula [60] bằng phép nhúng SciBERT [18]. Phần thay đổi của mô hình DualLCR này được khoanh đỏ trong Hình 3.3 bên dưới. Mô hình DualLCR đã cải tiến được đặt tên là RHN-DualLCR.



Hình 3.3. Cấu trúc của mô-đun ngữ cảnh trong mô hình RHN-DualLCR

3.3.1. Mạng cao tốc hồi quy RHN

Với các mô hình huấn luyện, quá trình chuyển đổi phi tuyến từ bước này sang bước khác trong các nhiệm vụ xử lý tuần tự rất phức tạp, điều này khiến cho việc huấn luyện các mạng nơ-ron hồi quy với các hàm chuyển tiếp "sâu" trở thành thách thức ngay cả với các bộ nhớ dài-ngắn hạn hai chiều BiLSTM. Nhằm giải quyết vấn đề này, nhóm nghiên cứu Zilly [20] đã giới thiệu một nghiên cứu lý thuyết mới về các mạng hồi quy sử dụng lý thuyết vòng tròn Geršgorin [68]. Nghiên cứu này đã giúp hiểu thêm về mô hình hóa và tối ưu hóa, cũng như cải thiện hiệu suất của mô hình BiLSTM. Dựa trên phân tích này, họ đã công bố mô hình mạng cao tốc hồi quy RHN, là một mô hình mở rộng của BiLSTM để cho phép chuyển đổi từng bước sâu hơn. RHN được coi là một mô hình mạnh mẽ được tạo ra để tận dụng độ sâu ngày càng tăng trong quá trình chuyển đổi lặp lại trong khi vẫn duy trì các đặc điểm đào tạo dễ dàng của BiLSTM. Kiến trúc đề xuất đã được đánh giá trong các thực nghiệm mô hình hóa ngôn ngữ khác nhau để chứng minh hiệu suất và tính hiệu quả. Ví dụ, trong một nghiên cứu sử dụng kho ngữ liệu Penn Treebank [69], chỉ cần tăng độ sâu chuyển tiếp từ 1 lên 10, độ phức tạp ở cấp độ từ của mô hình đã giảm từ 90.6 lên 65.4 trong khi vẫn giữ nguyên số tham số. Hơn nữa, khi đánh giá trên các bộ dữ liệu Wikipedia lớn hơn [70] về dự đoán ký tự (text8 và enwik8), RHN đã đánh bại tất cả các mô hình trước đó, đạt được entropy 1.27 bit cho mỗi ký tự. Theo cách tiếp cận của NCS, RHN được áp dụng trong mô hình khuyến nghị trích dẫn sẽ đạt được sự biểu diễn theo ngữ cảnh tuần tự chuyên sâu hơn về mối quan hệ giữa các bài báo mục tiêu

và các trích dẫn ứng cử viên. Do đó, áp dụng RHN được coi là một hướng đi đầy hứa hẹn để nâng cao hiệu quả của các hệ thống khuyến nghị trích dẫn hiện có.

3.3.2. Phép nhúng SciBERT

Phép nhúng SciBERT được công bố bởi nhóm nghiên cứu Beltagy [18] là một biến thể chuyên biệt của BERT với mục tiêu đào tạo đặc biệt cho văn bản khoa học. SciBERT được huấn luyện trên một lượng lớn các bài báo khoa học, tài liệu nghiên cứu và nội dung học thuật khác. Mục tiêu của SciBERT là nắm bắt ngôn ngữ và cấu trúc độc đáo có trong tài liệu khoa học, giúp nó trở nên hiệu quả hơn đối với các nhiệm vụ liên quan đến phân tích văn bản khoa học. SciBERT đã được chứng minh là có hiệu quả trong các nhiệm vụ xử lý ngôn ngữ tự nhiên trong văn bản khoa học khác nhau và đã góp phần tạo nên những tiến bộ trong việc trích xuất thông tin có giá trị từ tài liệu khoa học. SciBERT được đào tạo trên bộ 1.14 triệu tài liệu nghiên cứu được chọn ngẫu nhiên từ Semantic Scholar [26]. Các bài báo này bao gồm 18% bài báo về khoa học máy tính và 82% bài báo từ lĩnh vực y sinh rộng hơn và toàn bộ nội dung của các bài báo, thay vì chỉ là phần tóm tắt, được sử dụng để đào tạo. Kho văn bản này có trung bình 154 câu trên mỗi bài báo, tương đương với 2,769 mã thông báo, nên kích thước của toàn bộ kho văn bản là 3.17 tỷ mã thông báo. Con số này có thể so sánh với kích thước của kho dữ liệu được sử dụng để huấn luyện BERT là 3.3 tỷ mã thông báo. Thử nghiệm của nhóm nghiên cứu cho thấy với bộ dữ liệu của các bài báo khoa học như ACL-ARC hay RefSeer, kết quả của SciBERT tốt hơn nhiều so với BERT [54]. Trong nghiên cứu được trình bày ở chương 3 này, NCS đã áp dụng mô hình SciBERT để nhúng bối cảnh trích dẫn, tiêu đề và tóm tắt của cả các bài báo được trích dẫn và trích dẫn trước khi đưa nó vào mạng cao tốc hồi quy RHN của mô-đun ngữ nghĩa. Hiệu suất ấn tượng của mô hình SciBERT trong xử lý ngôn ngữ tự nhiên trong các bài báo khoa học được kỳ vọng sẽ nâng cao hiệu quả của mô hình DualCLR hiện tại đối với vấn đề khuyến nghị trích dẫn.

3.4. Tiến hành thực nghiệm với mô hình RHN-DualLCR

3.4.1. Xây dựng mô hình RHN-DualCLR

NCS đã xây dựng lại mã nguồn[9] của hệ thống khuyến nghị trích dẫn từ bài báo của nhóm nghiên cứu Medić và Šnajder [12] bằng cách thêm SciBERT [18], mô hình tốt nhất hiện nay để xử lý ngôn ngữ tự nhiên của các bài báo khoa học, cũng như sử dụng RHN [20] để thay thế lớp BiLSTM hiện tại. NCS đã sử dụng Python phiên bản 3.8.5 và PyTorch phiên bản 1.7.1 để xây dựng mô hình nâng cao. Đối với mô hình SciBERT được huấn luyện trước, NCS sử dụng AutoTokenizer và AutoModel từ thư viện transformers của Python. NCS phân chia các câu trong bối cảnh trích dẫn và tóm tắt của bài viết bằng cách sử dụng ScispaCy[10] [71] đã được tối ưu hóa cho văn bản khoa học. Sau khi được phân tách, các câu văn trong các bài báo thực hiện nhúng theo mô hình SciBERT.

3.4.2. Bộ dữ liệu thực nghiệm

Để so sánh NCS cũng đã đánh giá mô hình RHN-DualLCR nâng cao của mình trên hai tập dữ liệu RefSeer và ACL-ARC như được sử dụng trong bài báo gốc của Medić và Šnajder [12]. Cả hai bộ dữ liệu này cũng thường được sử dụng để tính toán hiệu suất của các hệ thống khuyến nghị trích dẫn được công bố gần đây [10] [15] [16].

Bảng 3.1. Thống kế tập dữ liệu theo số lượng bối cảnh trích dẫn và bài báo [12]

Dataset	Huấn luyện	Xác thực	Kiểm tra	Bài báo
ACL-ARC	30,390	9,381	9,585	19,711
RefSeer	3,521,582	124,551	126,021	624,957

[9] https://github.com/zoranmedic/DualLCR
[10] https://github.com/allenai/SciSpaCy

3.4.3. Chỉ số đánh giá

Trong nghiên cứu của Medić và Šnajder [12] đang sử dụng 2 chỉ số đánh giá hiệu suất của mô hình khuyến nghị trích dẫn là Recall Top@K và xếp hạng đối ứng trung bình MRR, do đó để so sánh hiệu suất của mô hình đã cải tiến RHN-DualLCR thì NCS cũng sẽ đánh giá dựa trên 2 tiêu chí này.

3.5. Đánh giá kết quả thử nghiệm

Nội dung nghiên cứu trong chương 3 này tập trung vào việc nâng cao mô hình đề xuất trích dẫn tiên tiến nhất DualLCR của nhóm Medić và Šnajder [12], và do đó NCS đánh giá hiệu quả thực nghiệm của mô hình đã cải tiến RHN-DualLCR với mô hình DualLCR. Tương tự như Medić và Šnajder [12], NCS đã đánh giá hiệu suất đề xuất trích dẫn bằng cách sử dụng các chỉ số đánh giá tiêu chuẩn xếp hạng đối ứng trung bình MRR và Recall@K (R@K).

Theo kết quả được báo cáo bởi Medić và Šnajder [12], DualCon-ws đạt kết quả tốt nhất cho tiêu chí MRR với bộ dữ liệu ACL-600. Hơn nữa, với tập dữ liệu ACL-600 này, DualEnh-ws thu được kết quả tốt nhất với tiêu chí R@10. Với bộ dữ liệu ACL-200, DualEnh-s đạt kết quả tốt nhất với tiêu chí R@10 trong khi DualEnh-ws nhận được điểm MRR tốt nhất. Đối với tập dữ liệu RefSeer, DualEnh-ws có kết quả tốt nhất với cả chỉ số đánh giá R@10 và MRR. Kết quả thực nghiệm của bài báo này cho chứng minh rằng việc làm phong phú khi biểu diễn ngữ cảnh trích dẫn bằng cách thêm thông tin toàn cục là có lợi khi ngữ cảnh trích dẫn ngắn hơn, nhưng quá trình làm phong phú này không cần thiết khi chúng dài hơn, vì các ngữ cảnh dài hơn đã cung cấp đủ thông tin cho khuyến nghị trích dẫn.

Bảng 3.2. So sánh kết quả từ Medić và Šnajder [12] và mô hình nâng cao RHN-DualLCR

Mô hình		**ACL-600**		**ACL-200**		**RefSeer**	
		R@10	**MRR**	**R@10**	**MRR**	**R@10**	**MRR**
DualLCR [12]	**DualCon-ws**	0.689	**0.368**	0.647	0.335	0.406	0.206
	DualEnh-s	0.662	0.315	**0.716**	0.341	0.437	0.230
	DualEnh-ws	**0.699**	0.357	0.703	**0.366**	**0.534**	**0.280**
RHN-DualLCR	**DualCon-ws**	0.701	**0.391**	0.661	0.354	0.428	0.223
	DualEnh-s	0.683	0.342	**0.748**	0.363	0.461	0.256
	DualEnh-ws	**0.756**	0.379	0.718	**0.403**	**0.582**	**0.307**

Medić và Šnajder đã tiếp tục thực hiện một nghiên cứu thực nghiệm [13] về tác động của ba lựa chọn thiết kế trong mô hình do chính họ công bố trước đó [12], vì vậy NCS cũng so sánh thêm kết quả thu được từ mô hình RHN-DualLCR với kết quả từ nghiên cứu thực nghiệm của họ để làm nổi bật thêm những đóng góp của mình. Medić và Šnajder đã tiến hành thực nghiệm về ba lựa chọn thiết kế: tham số của mô hình lọc trước, chế độ huấn luyện, chiến lược lấy mẫu phủ định trên hai bộ dữ liệu thường được sử dụng ACL-200 và RefSeer, cho nên NCS so sánh lần lượt mô hình RHN-DualLCR với các lựa chọn thiết kế trên bộ dữ liệu này. Với 2 bộ dữ liệu ACL-200 và RefSeer, RHN-DualLCR đạt thành tích tốt nhất với 2 biến thể DualEnh-s và DualEnh-ws nên NCS chỉ đem 2 biến thể này ra để so sánh. Kết quả so sánh được thể hiện lần lượt ở Bảng 3.3, 3.4 và 3.5.

Bảng 3.3. So sánh hiệu suất của model lọc trước BM25 và SPECTER [13] với RHN-DualLCR

Mô hình		**ACL-200**		**RefSeer**	
		R@10	**MRR**	**R@10**	**MRR**
DualLCR-design [13]	**BM25**	0.254	0.077	0.173	0.055
	SPECTER	0.170	0.080	0.119	0.055
RHN-DualLCR	**DualEnh-s**	**0.748**	0.363	0.461	0.256
	DualEnh-ws	0.718	**0.403**	**0.582**	**0.307**

Bảng 3.4. So sánh hiệu suất của mô hình sắp xếp lại Text+Bib cho các chế độ huấn luyện nghiêm ngặt [13] với RHN-DualLCR

Mô hình		ACL-200		RefSeer	
		R@10	MRR	R@10	MRR
DualLCR-design [13]	**BM25(Text)**	0.531	0.268	0.300	0.116
	SPECTER(Text)	0.551	0.287	0.297	0.139
	BM25(Text+Bib)	0.725	0.401	0.324	0.147
	SPECTER(Text+Bib)	0.729	0.339	0.301	0.137
RHN-DualLCR	**DualEnh-s**	**0.748**	0.363	0.461	0.256
	DualEnh-ws	0.718	**0.403**	**0.582**	**0.307**

Bảng 3.5. So sánh hiệu suất của các chiến lược lấy mẫu phủ định của thiết kế DualLCR [13] với RHN-DualLCR

Mô hình		ACL-200		RefSeer	
		R@10	MRR	R@10	MRR
DualLCR-design [13]	**Cited(Text+Bib)**	0.746	**0.413**	0.265	0.123
	Graph neighbors	0.676	0.363	0.418	0.210
RHN-DualLCR	**DualEnh-s**	**0.748**	0.363	0.461	0.256
	DualEnh-ws	0.718	0.403	**0.582**	**0.307**

Để chứng minh rõ hơn hiệu suất của mô hình RHN-DualLCR, ngoài việc so sánh với kết quả của hai mô hình DualLCR [12] và DualLCR-design [13] do nhóm Medić và Šnajder công bố, NCS tiếp tục so sánh với 3 mô hình tiên tiến nhất hiện nay cho bài toán khuyến nghị trích dẫn. Đó là các mô hình: (1) HAtten [16] bao gồm hai giai đoạn: tìm nạp trước và giai đoạn sắp xếp lại; (2) Mô hình mạng nơ-ron trích dẫn NCN được đề xuất bởi Ebesu và Fang [10] và cải tiến mới của nhóm Färber [11]; (3) BERT-GCN [15] kết hợp BERT cho xử lý văn bản và mạng tích chập đồ thị GCN [53] cho bộ mã hóa siêu dữ liệu của các bài báo. Kết quả của việc so sánh hiệu suất các mô hình được trình bày ở Bảng 3.6 như sau:

Bảng 3.6. So sánh kết quả từ 3 mô hình khuyến nghị trích dẫn tiên tiến với mô hình RHN-DualLCR

Mô hình	ACL-200		RefSeer	
	R@10	MRR	R@10	MRR
HAtten [16]	0.281	0.148	0.214	0.115
NCN [10] [11]	0.438	0.282	0.291	0.267
BERT-GCN [15]	0.685	0.378	0.423	0.281
RHN-DualLCR	**0.748**	**0.403**	**0.582**	**0.307**

3.6. Kết luận chương 3

Trong nội dung nghiên cứu của chương 3 này, NCS đã cải tiến mô hình khuyến nghị trích dẫn DualLCR hiện có do Medić và Šnajder [12] [13] công bố bằng cách áp dụng các kết quả nghiên cứu gần đây của mạng nơ-ron hồi quy của học sâu cũng như các thành tựu nghiên cứu trong xử lý ngôn ngữ tự nhiên, đặc biệt là ngôn ngữ trong các bài báo khoa học. NCS đã tích hợp ScispaCy [71] để phân tách các câu trong bối cảnh trích dẫn và tóm tắt của bài báo, sử dụng mô hình SciBERT [18] để thực hiện phép nhúng cho văn bản khoa học của các bài báo ở đầu vào và thay thế BiLSTM [14] bằng RHN [20], một mô hình đã mở rộng kiến trúc BiLSTM để cho phép chuyển tiếp giữa các bước có độ sâu lớn hơn một. NCS đã đánh giá hiệu suất của mô hình được đề xuất bằng ba bộ dữ liệu ACL-200, ACL-600 và RefSeer và đạt được kết quả

cải thiện đáng kể khi so sánh với mô hình của Medić và Šnajder trong 2 bài báo gốc [12] [13] khi sử dụng cùng hai tiêu chuẩn đánh giá R@10 và MRR. Mô hình RHN-DualLCR cũng đã chứng minh được hiệu quả hơn so với 3 mô hình khuyến nghị trích dẫn tiên tiến hiện nay, đó là các mô hình: HAtten [16], NCN [10] [11] và BERT-GCN [15]. Hơn nữa, nội dung nghiên cứu của chương 3 này cũng bao gồm phần tiến hành thực nghiệm điều chỉnh để xem xét các siêu tham số khác nhau có thể ảnh hưởng như thế nào đến hiệu suất của RHN [20] và NCS đã đề xuất các cách sử dụng các kết quả thực nghiệm này để nâng cao hiệu quả của mô hình trong tương lai.

Chương 4. MÔ HÌNH KHUYẾN NGHỊ TRÍCH DẪN MỚI SỬ DỤNG SCIBERT VÀ GRAPHSAGE

4.1. Mở đầu

Trong phạm vi nghiên cứu của chương 4 này, NCS xây dựng một mô hình khuyến nghị trích dẫn nhận biết ngữ cảnh mới bằng cách kết hợp hai thành tựu nghiên cứu tiên tiến nhất hiện nay cho các kỹ thuật học biểu diễn dữ liệu văn bản/ngữ cảnh và đồ thị biểu thị liên kết trích dẫn: SciBERT [18] và GraphSAGE [19].

Các kết quả nghiên cứu ở chương này được công bố ở công trình số 3 và 5.

4.2. Vấn đề tồn tại của mô hình khuyến nghị hiện nay

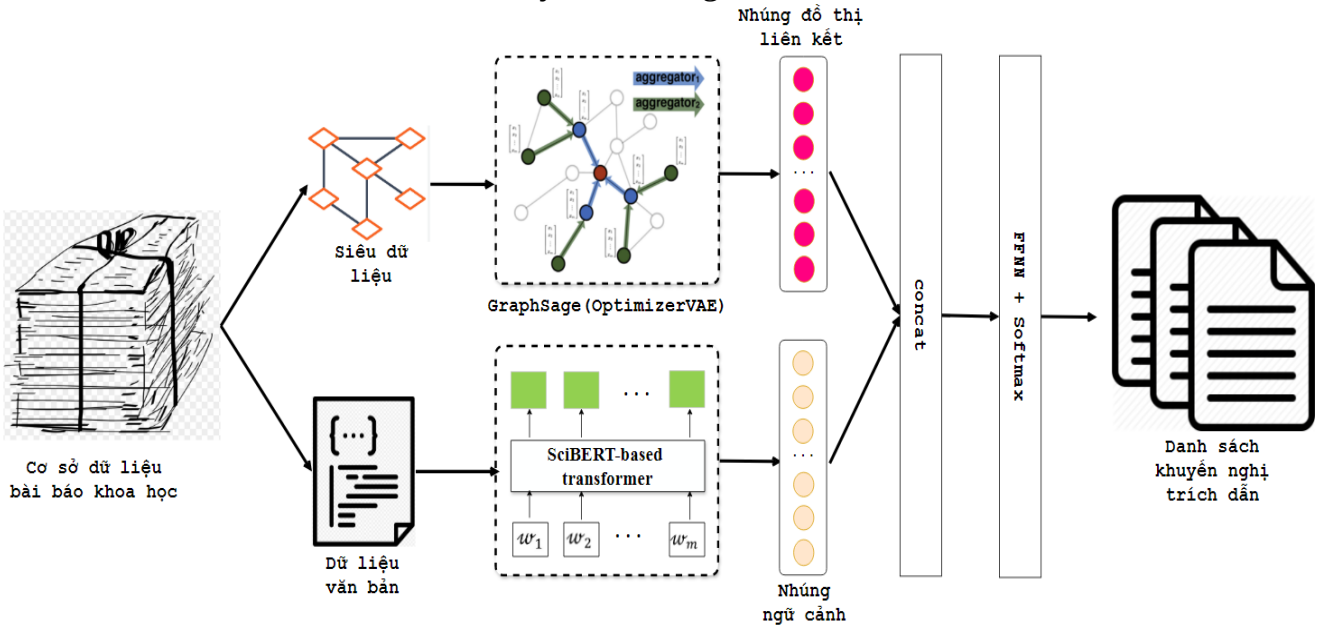
Phần lớn các phương pháp tiếp cận hiện tại đối với bài toán khuyến nghị trích dẫn nhận biết ngữ cảnh chỉ tập trung vào nội dung của cả ngữ cảnh trích dẫn và các bài báo khoa học [13] [16] [72]. Cách tiếp cận này nhằm mục đích kết nối khoảng cách ngữ nghĩa giữa các yếu tố này mà không xem xét thông tin vượt ra ngoài nội dung ngữ nghĩa của các bài báo khoa học. Thực tế trong các công bố khoa học có những yếu tố bổ sung như tác giả, thông tin hội nghị/tạp chí và năm xuất bản, có mức độ quan trọng khác nhau trong việc hỗ trợ các nhà nghiên cứu hiểu được sự tương đồng về ngữ nghĩa giữa các bài báo khoa học và bối cảnh trích dẫn. Ví dụ: một tác giả được liên kết với một bài báo khoa học có thể là đồng tác giả với các bài báo liên quan khác. Tương tự như vậy, một hội nghị/tạp chí công bố một bài báo khoa học cụ thể cũng có thể xuất bản các bài báo khác có chủ đề tương tự. Vì vậy, việc bổ sung thông tin về tác giả và nơi xuất bản (tên hội nghị hoặc tạp chí) được kỳ vọng sẽ làm tăng hiệu quả của hệ thống khuyến nghị trích dẫn. Ngoài ra, thông tin về năm xuất bản của bài báo cũng cần được quan tâm đúng mức đối với các mô hình khuyến nghị trích dẫn. Cụ thể, khi tìm tài liệu trích dẫn thì các nhà nghiên cứu luôn có xu hướng trích dẫn những bài báo mới nhất và cập nhật nhất. Dựa vào những giả định này, NCS đã phát hiện ra rằng hiệu suất khuyến nghị trích dẫn theo ngữ cảnh không chỉ bị ảnh hưởng hoàn toàn bởi sự giống nhau về mặt ngữ nghĩa giữa ngữ cảnh trích dẫn và các bài báo khoa học mà nó cũng phụ thuộc vào các yếu tố khác, chẳng hạn như tác giả, địa điểm và năm xuất bản của các bài báo này.

Có một số các công bố gần đây, tiêu biểu như là nghiên cứu của nhóm Jeong [15] đã đề xuất mô hình BERT-GCN, trong đó kết hợp BERT để mã hóa ngữ cảnh trích dẫn và thông tin của bài báo, còn mạng đồ thị tích chập GCN để mã hóa các thông tin siêu dữ liệu (tác giả, địa điểm và năm công bố) của bài báo. Các công bố của các nhóm nghiên cứu khác [11] [12] cũng đã bước đầu đưa thêm các thông tin siêu dữ liệu vào mô hình của họ. Tuy nhiên các mô hình này đều có thể cải thiện được nếu áp dụng các thành tựu nghiên cứu về mạng tích chập đồ thị trong thời gian gần đây.

4.3. Xây dựng mô hình khuyến nghị trích dẫn mới với SciBERT và GraphSAGE

Nội dung phần này mô tả chi tiết xây dựng một mô hình khuyến nghị trích dẫn nhận biết ngữ cảnh hoàn toàn mới bằng cách kết hợp hai thành tựu nghiên cứu cập nhật nhất hiện nay cho các kỹ thuật học biểu diễn dữ liệu dựa trên văn bản/ngữ cảnh SciBERT [18] và đồ thị biểu thị liên kết trích dẫn GraphSAGE [19]. Như đã đề cập ở chương 3, SciBERT là một biến thể

của mô hình xử lý ngôn ngữ tự nhiên BERT [54] nhưng đã được điều chỉnh đặc biệt cho các nhiệm vụ trong lĩnh vực phân tích văn bản khoa học và y sinh. NCS hy vọng rằng việc sử dụng SciBERT được huấn luyện trước để biểu diễn câu theo ngữ cảnh sẽ mang lại hiệu quả cao. Dữ liệu khoa học chẳng hạn như các bài báo, thì ngoài dữ liệu văn bản vẫn còn chứa nhiều siêu dữ liệu khác nhau như liên kết trích dẫn giữa các bài báo, tác giả, thông tin địa điểm và năm xuất bản. Do đó NCS ứng dụng mô hình GraphSAGE để mô tả mối liên kết trích dẫn giữa các bài báo và rút ra biểu diễn đã được huấn luyện từ chúng.



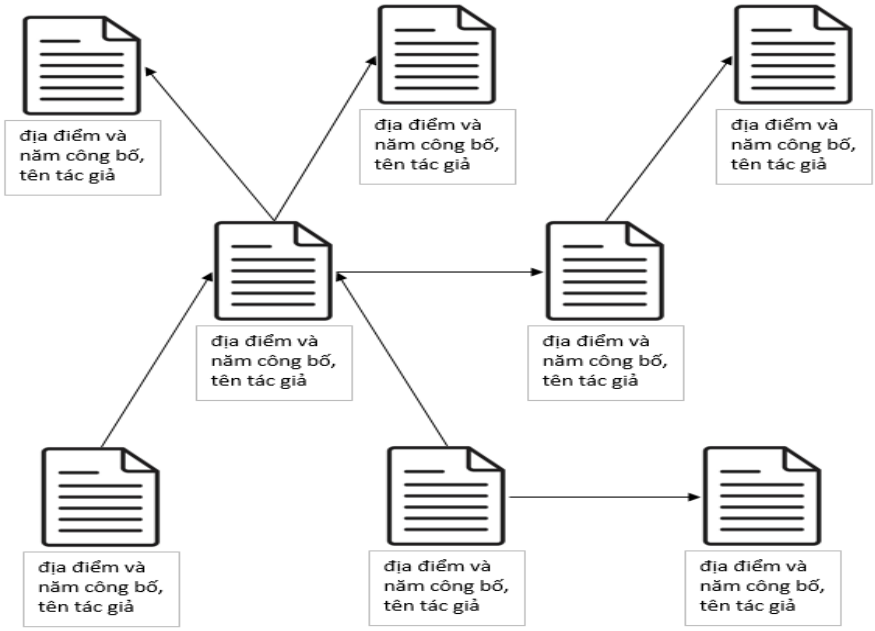
Hình 4.1. Sơ đồ kiến trúc tổng thể của mô hình SciBERT-GraphSAGE

Hình 4.1 minh họa kiến trúc của mô hình khuyến nghị trích dẫn SciBERT-GraphSAGE, trong đó bao gồm một bộ mã hóa ngữ cảnh để lấy các phần nhúng văn bản thông qua SciBERT và một bộ mã hóa liên kết trích dẫn để trích xuất các thành phần nhúng đồ thị bằng GraphSAGE. Cả hai bộ mã hóa đều được huấn luyện trước bằng cách sử dụng dữ liệu đồ thị ngữ cảnh và trích dẫn từ bài báo. Tiếp theo, dữ liệu được đưa vào các mô hình được huấn luyện trước này và các phần nhúng do mỗi bộ mã hóa tạo ra sẽ được kết hợp lại với nhau. Những phần nhúng kết hợp này sau đó được chuyển qua mạng nơ-ron chuyển tiếp. Đầu ra được xử lý thông qua lớp softmax và hàm entropy chéo (*cross-entropy*) được sử dụng làm hàm mất mát cho quá trình huấn luyện của mô hình.

4.4. Bộ mã hóa đồ thị liên kết trích dẫn sử dụng GraphSAGE

Nhóm nghiên cứu của Hamilton [19] đã giới thiệu GraphSAGE (**Graph SA**mpling and **AG**gregation) như một phiên bản mở rộng và cải tiến của mạng tích chập đồ thị GCN [53]. Khái niệm cốt lõi đằng sau mô hình GraphSAGE là rút ra các biểu diễn cấu trúc cục bộ có thứ tự cao và quy nạp của các nút từ các đồ thị biểu diễn trích dẫn đã cho. Khác với mạng tích chập đồ thị GCN trước đó, việc tổng hợp tính năng cho nút mục tiêu dựa vào tập hợp con được lấy mẫu của các thuộc tính nút lân cận của nó, thay vì dựa vào tập hợp lân cận đã hoàn chỉnh như mạng tích chập đồ thị GCN [53]. Với thuộc tính này, việc áp dụng các mô hình GraphSAGE vào khuyến nghị trích dẫn đã trở nên nổi bật. Các nghiên cứu gần đây đã chứng minh rằng GraphSAGE được coi là một nền tảng học máy dựa trên đồ thị mạnh mẽ được thiết kế để học cách biểu diễn từ các đồ thị tỷ lệ lớn một cách hiệu quả. Bằng cách coi các bài báo học thuật là các nút trong đồ thị, siêu dữ liệu (tác giả, thông tin về địa điểm và năm xuất bản) là thuộc tính của nút và mối liên kết trích dẫn giữa các bài báo là các cạnh của đồ thị, mạng nơ-ron trích dẫn tạo thành cấu trúc đồ thị tự nhiên, trong đó các bài báo được kết nối với nhau thông qua mối quan hệ trích dẫn của chúng. Chi tiết về quá trình này có thể được tham khảo trong Hình 4.2 bên dưới. Tận dụng khả năng kết nối vốn có này, GraphSAGE cung cấp một

cách tiếp cận đầy hứa hẹn để nâng cao khuyến nghị trích dẫn. Nó vượt trội trong việc học cách biểu diễn nút bằng cách lấy mẫu và tổng hợp thông tin từ các nút lân cận trong đồ thị. Bằng cách này, RHN không chỉ có thể nắm bắt được ý nghĩa ngữ nghĩa của từng bài báo mà còn mã hóa ngữ cảnh được cung cấp bởi các mối quan hệ trích dẫn của chúng. Do đó, các biểu diễn nút đạt được phản ánh vị trí của các bài báo trong bối cảnh học thuật rộng hơn, nhằm nắm bắt những điểm tương đồng, ảnh hưởng và mối liên hệ theo chủ đề giữa các bài báo.



Hình 4.2. Tạo các nút và cạnh cho GraphSAGE từ siêu dữ liệu của các bài báo

4.5. Tiến hành thực nghiệm với mô hình SciBERT-GraphSAGE

4.5.1. Xây dựng mô hình SciBERT-GraphSAGE

NCS đã xây dựng mô hình SciBERT-GraphSAGE cho hệ thống khuyến nghị trích dẫn bằng cách kết hợp bộ mã hóa ngữ cảnh trích dẫn SciBERT và mã hóa liên kết trích dẫn GraphSAGE. Để có thể mã hóa ngữ cảnh trích dẫn, NCS đã khởi tạo SciBERT bằng mô hình được huấn luyện trước được cung cấp bởi nhóm của Beltagy[11] [18]. Tương tự, NCS đã chỉnh sửa mã nguồn của mô hình GraphSAGE[12] [19] để có thể mã hóa đồ thị liên kết trích dẫn. Tất cả các mô hình được xây dựng với Python phiên bản 3.8.5 và TensorFlow phiên bản 2.7.0. NCS đã trích xuất các vectơ ngữ cảnh nhúng và vectơ đồ thị trích dẫn bằng cách sử dụng các lớp SciBERT và GraphSAGE, được xây dựng thông qua các quy trình huấn luyện riêng biệt. Trong SciBERT, số lượng lớp chú ý (*number of attention heads*) là 12, ngăn xếp (*stack*) bộ mã hóa là 12 và trình tối ưu hóa Adam optimizer [74] được sử dụng. Tốc độ học (η) là 2×10^{-5}, epsilon (ϵ) là 1×10^{-6}, với tham số beta 1 (β_1) được đặt là: 0.9, beta 2 (β_2) được đặt là 0.999 và tốc độ giảm trọng số là 0.01. Mô hình cũng thiết lập độ dài chuỗi tối đa là 128, bộ đệm (*padding*) là 0 nếu độ dài ngắn hơn 128 và kích thước ẩn là 768. Đối với GraphSAGE, số vòng lặp huấn luyện (*epoch*) là 200, kích thước ẩn đầu tiên tương ứng với với số lượng bài báo trong bộ dữ liệu và thứ nguyên ẩn thứ hai là 768, kích thước tập (*batch*) giống với tổng kích thước tài liệu (giảm độ dốc toàn lô), trình tối ưu hóa là OptimizerVAE [73] và tốc độ học là 0.01.

[11] https://github.com/allenai/scibert
[12] https://github.com/williamleif/GraphSAGE

4.5.2. Bộ dữ liệu thực nghiệm

Chương 4 này đã đánh giá tính hiệu quả của mô hình SciBERT-GraphSAGE trên ba bộ dữ liệu chuẩn thường được sử dụng cho hệ thống khuyến nghị trích dẫn cục bộ, bao gồm: ACL-200 [12], RefSeer [12] và FullTextPeerRead [15]. NCS đã sử dụng các bộ dữ liệu này để đánh giá hiệu suất vượt trội của mô hình SciBERT-GraphSAGE với 5 mô hình khuyến nghị trích dẫn tiên tiến hiện nay là: CACR [72], BERT-GCN [15], HAtten [16], DualLCR [12], DualLCR theo thiết kế [13]. Số liệu thống kê của ba bộ dữ liệu này được hiển thị ở trong Bảng 4.1.

Bảng 4.1. Thống kê 3 bộ dữ liệu (số lượng ngữ cảnh trích dẫn và bài báo)

Tên bộ dữ liệu	Số lượng ngữ cảnh			Số bài báo	Năm công bố
	Huấn luyện	Xác thực	Kiểm tra		
ACL-200	30,390	9,381	9,585	19,711	2009 - 2015
FullTextPeerRead	9,363	1,043	6,841	4,898	2007 - 2017
RefSeer	3,521,582	124,551	126,021	624,957	- 2014

4.5.3. Chỉ số đánh giá

Để đánh giá hiệu quả của mô hình SciBERT-GraphSAGE, NCS sử dụng 3 chỉ số đánh giá thông dụng cho các hệ thống khuyến nghị trích dẫn là độ chính xác trung bình MAP, xếp hạng đối ứng trung bình MRR và Recall Top@K.

4.6. Đánh giá kết quả thực nghiệm

Trong phần này, NCS trình bày phân tích so sánh toàn diện về hiệu suất của mô hình SciBERT-GraphSAGE với một số mô hình khuyến nghị trích dẫn cục bộ tiên tiến có kết quả tốt nhất gần đây. Mục tiêu của phân tích này là đánh giá chặt chẽ điểm mạnh và khả năng của mô hình SciBERT-GraphSAGE so với các tiêu chuẩn tiên tiến hiện có. Bằng cách tiến hành so sánh chuyên sâu, NCS mong muốn cung cấp những hiểu biết sâu sắc về những cải thiện hiệu suất đã đạt được bằng phương pháp tiếp cận mới của mình và nêu bật những đóng góp tiềm năng của phương pháp đó cho lĩnh vực này. Để xác thực những cải tiến của mô hình mới, NCS đã đưa mô hình của mình vào một quy trình đo điểm chuẩn nghiêm ngặt dựa trên 5 mô hình tiên tiến nhất được công nhận rộng rãi trong cộng đồng nghiên cứu về khuyến nghị trích dẫn cục bộ. Việc đánh giá được thực hiện trên ba bộ dữ liệu được sử dụng phổ biến và thường được sử dụng: ACL-200 [12], RefSeer [12] và FullTextPeerRead [15]. Những bộ dữ liệu này bao gồm nhiều vấn đề phức tạp, phản ánh các tình huống và thách thức mà trong thế giới thực hay gặp phải với các hệ thống khuyến nghị trích dẫn cục bộ. Bằng cách sử dụng các bộ dữ liệu đa dạng như vậy, NCS mong muốn nắm bắt được sự hiểu biết toàn diện về hiệu suất của mô hình SciBERT-GraphSAGE trong nhiều bối cảnh khác nhau.

Với bộ dữ liệu FullTextPeerRead [15], mô hình SciBERT-GraphSAGE được so sánh với các mô hình tiên tiến như sau: (1) mô hình CACR [72] có cả bộ mã hóa bài báo và bộ mã hóa ngữ cảnh trích dẫn dựa trên mô hình LSTM; (2) mô hình BERT-GCN [15] kết hợp BERT cho bộ mã hóa văn bản và GCN cho bộ mã hóa siêu dữ liệu; (3) mô hình HAtten [16] bao gồm giai đoạn tìm nạp trước và giai đoạn xếp hạng lại; (4) mô hình SciBERT-base [76] được huấn luyện để dự đoán bài báo mà sẽ được trích dẫn từ một ngữ cảnh. Kết quả so sánh được thể hiện ở Bảng 4.2 như sau:

Bảng 4.2. Kết quả so sánh hiệu suất của mô hình SciBERT-GraphSAGE với 4 mô hình tiên tiến nhất trên bộ dữ liệu FullTextPeerRead

Mô hình	MAP	MRR	Recall@5	Recall@10	Recall@10	Recall@10	Recall@10
CACR [72]	0.1551	0.1549	0.2154	0.2761	0.4128	0.4794	0.5516
BERT-GCN [15]	0.4181	0.4179	0.4864	0.5291	0.6093	0.6495	0.6994
HAtten [16]	0.1672 *	0.1670	0.2780*	0.3060	0.4850*	0.5270*	0.5560*
SciBERT-base [76]	0.454	0.466	-	-	-	-	-
SciBERT-GraphSAGE	**0.5162**	**0.5163**	**0.6217**	**0.6744**	**0.7636**	**0.8099**	**0.8504**

Trong kết quả công bố của họ, nhóm của Gu [33] không trình bày kết quả đánh giá thử nghiệm với các tiêu chí MAP và Recall@K, K = 5, 30, 50, 80 cho mô hình HAtten của họ. Để so sánh được, NCS đã chạy thực nghiệm lại mô hình HAtten [13] với bộ dữ liệu FullTextPeerRead và đánh dấu (*) trong các thành tích này. Các tác giả của mô hình SciBERT-base [76] cũng không công bố thực nghiệm của họ với tiêu chí Recall@K. Kết quả từ Bảng 4.3 cho thấy, trong số các mô hình được công bố gần đây thì kết quả thử nghiệm của 2 mô hình BERT-GCN [15] và SciBERT-base [76] với bộ dữ liệu FullTextPeerRead cho thành tích khả quan nhất. Tuy nhiên, bằng cách kết hợp SciBERT và GraphSAGE là những mô hình mới hơn và cải tiến so với BERT [54] và GCN [53], mô hình SciBERT-GraphSAGE thậm chí còn vượt trội hơn so với 2 mô hình này từ 22% đến 28% ở tất cả các chỉ số so sánh.

Với hai bộ dữ liệu ACL-200 [12] và RefSeer [12], mô hình SciBERT-GraphSAGE tiếp tục được so sánh với các thành tựu nghiên cứu được công bố gần đây cho bài toán khuyến nghị trích dẫn cục bộ: (1) mô hình HAtten [16] bao gồm giai đoạn tìm nạp trước và giai đoạn sắp xếp lại; (2) mô hình DualEnh và DualCon [12] sử dụng cả nội dung ngữ nghĩa và dữ liệu thông tin học thuật để đánh giá chất lượng của từng bài báo ứng viên trích dẫn tiềm năng; (3) mô hình DualLCR-design [13] đã tối ưu hóa các tham số để mang lại hiệu suất tốt nhất. Kết quả so sánh được thể hiện ở Bảng 4.3 như sau.

Bảng 4.3. Kết quả so sánh hiệu suất của mô hình SciBERT-GraphSAGE với 4 mô hình tiên tiến nhất trên 2 bộ dữ liệu ACL-200 và RefSeer

Bộ dữ liệu	ACL-200		RefSeer	
Mô hình	**MRR**	**Recall@10**	**MRR**	**Recall@10**
HAtten [16]	0.148	0.281	0.115	0.214
DualCon [12]	0.340	0.693	0.206	0.406
DualEnh [12]	0.366	0.716	0.280	0.534
DualLCR-design [13]	0.413	0.746	0.210	0.418
SciBERT-GraphSAGE	**0.472**	**0.765**	**0.308**	**0.565**

Kết quả thực nghiệm từ Bảng 4.3 cũng cho thấy, trong các công bố gần đây, Medić và Šnajder [12] [13] đã đề xuất các mô hình có kết quả tốt nhất. Với bộ dữ liệu ACL-200, mô hình DualEnh [12] của họ thu được kết quả vượt trội so với các mô hình khác, trong khi với bộ dữ liệu RefSeer, mô hình DualLCR-design [13] cũng thu được kết quả tốt hơn. Tuy nhiên,

[13] https://github.com/nianlonggu/Local-Citation-Recommendation

mô hình SciBERT-GraphSAGE thậm chí còn mang lại thành tích tốt hơn. Với bộ dữ liệu ACL-200, sử dụng số liệu MRR và Recall@10, kết quả của NCS lần lượt tốt hơn 14% và 3%, trong khi với bộ dữ liệu RefSeer, mô hình SciBERT-GraphSAGE cũng cho kết quả tốt hơn lần lượt là 10% và 6% cho các chỉ số tương ứng.

4.7. Kết luận chương 4

Trong chương 4 này NCS đã đề xuất một mô hình mới cho bài toán khuyến nghị trích dẫn, đó là mô hình lai ghép SciBERT-GraphSAGE, là sự kết hợp của SciBERT và GraphSAGE. Mô hình SciBERT [18] là phiên bản chuyên biệt của BERT [54] đã được huấn luyện đặc biệt cho các nhiệm vụ trong lĩnh vực nghiên cứu khoa học và kỹ thuật. GraphSAGE [19] là một nền tảng học máy mạnh mẽ dựa trên đồ thị được thiết kế để tìm hiểu các biểu diễn từ đồ thị thị có số lượng nút lớn một cách hiệu quả. Bằng cách coi các bài báo học thuật là các nút, siêu dữ liệu là các thuộc tính của nút và các trích dẫn là các cạnh trong đồ thị, mạng trích dẫn tạo thành một cấu trúc đồ thị tự nhiên, trong đó các bài báo được kết nối với nhau thông qua các mối quan hệ trích dẫn của chúng. Kết quả thực nghiệm khi so sánh trên 3 bộ dữ liệu tiêu chuẩn thường dùng (ACL-200 [12], RefSeer [12] và FullTextPeerRead [15]) với 6 mô hình tiên tiến nhất (CACR [72], BERT-GCN [15], HAtten [16], DualEnh và DualCon [12], DualLCR-design [13] và SciBERT-base [76]) đều cho kết quả vượt trội ở 3 chỉ số (MAP, MRR và Recall@K), chứng tỏ phương pháp tiếp cận của NCS là đúng đắn và đã đạt được những kết quả thực sự nổi bật.

KẾT LUẬN VÀ KIẾN NGHỊ

1) Kết luận:

Luận án nghiên cứu hướng tiếp cận áp dụng các thành tựu mới nhất của các mô hình học sâu cho bài toán khuyến nghị trích dẫn. Luận án đã có những đónh góp chính sau đây:

1) Đề xuất các giải pháp đề nâng cao hiệu suất cho mô hình mạng nơ-ron trích dẫn NCN.
2) Đề xuất một mô hình mới tên là RHN-DualLCR, trong đó bao gồm các giải pháp để nâng cao hiệu suất cho mô hình mạng song song 2 bước DualLCR cho bài toán khuyến nghị trích dẫn đã được công bố bởi Medić và Šnajder [12] [13].
3) Đề xuất mô hình khuyến nghị trích dẫn mới SciBERT-GraphSAGE bằng cách kết hợp 2 thành tựu mới nhất trong xử lý ngôn ngữ tự nhiên SciBERT [18] và tạo dữ liệu nhúng của liên kết trích dẫn bằng đồ thị GraphSAGE [19].

Các kết quả nghiên cứu này đều đã được đăng tải trên các tạp chí chuyên ngành uy tín.

2) Kiến nghị:

Trong các hướng nghiên cứu tiếp theo, NCS tập trung vào 2 hướng sau:

(1) Áp dụng các kết quả của mạng không đồng nhất (heterogeneous network) và graph convolutional networks mà nhóm nghiên cứu đã đạt được cho bài toán khuyến nghị trích dẫn.

(2) Thực nghiệm các kết quả nghiên cứu trên các dataset lớn hơn như là Microsoft Academic Graph[14], PubMed[15] hay DBLP[16].

[14] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph
[15] https://www.nlm.nih.gov/databases/download/pubmed_medline.html
[16] https://dblp.uni-trier.de/xml/

DANH MỤC CÁC BÀI BÁO ĐÃ XUẤT BẢN LIÊN QUAN ĐẾN LUẬN ÁN

1. **Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Bay Vo, "Enhanced context-aware citation recommendation with auxiliary textual information based on an auto-encoding mechanism," *Applied Intelligence*, 53(14), 2023, pp. 17381–17390, ISSN/eISSN:0924-669X/1573-7497, https://doi.org/10.1007/s10489-022-04423-1, Scopus indexed (Q2), SCIE IF: 5.086 (Q2).
2. **Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Bay Vo, "Enhancing local citation recommendation with recurrent highway networks and SciBERT-based embedding", *Expert Systems with Applications* (ESWA), Volume 243, 2024, 122911, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.122911 Scopus indexed (Q1), SCIE IF: 8.5 (Q1).
3. **Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Ngoc Thanh Nguyen and Bay Vo, "An effective context-aware citation recommendation model with SciBERT and GraphSAGE", *IEEE Transactions on Systems, Man and Cybernetics: Systems*, Volume 55, Issue 2, pp. 852-863, Feb. 2025, ISSN/eISSN: 2168-2216/2168-2232. https://doi.org/10.1109/TSMC.2024.3490774 Scopus indexed (Q1), SCIE IF: 8.7 (Q1).
4. **Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Bay Vo, "Enrich textual information for Hierarchical-Attention Text Encoder in Local Citation Recommendation", *Kỷ yếu Hội thảo quốc gia lần thứ XXV: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông* – Hà Nội, 8-9/12/2022 (National Symposium of Selected ICT Problems – VNICT(@) 2022), ISBN:978-604-67-2508-4.
5. **Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Bay Vo, "A context-aware citation recommendation model with SciBERT and GraphSAGE", *Kỷ yếu Hội thảo quốc gia lần thứ XXVI: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông* – Bắc Ninh, 5-6/10/2023 (National Symposium of Selected ICT Problems – VNICT(@) 2023) ISBN: 978-604-67-2746-0.