Phraseformer: Trích xuất cụm từ khóa đa phương thức bằng cách sử dụng Transformer và Nhúng đồ thị

N. Nikzad–Khasmakhi[Một], Mohammad–Reza Feizi–Derakhshi[Một,*], Meysam Asgari-Chenaghlu[Một], MA Balafar[b], Ali-Reza Feizi-Derakhshi[Một], Taymaz Rahkar-Farshi[AC], Majid Ramezani[Một], Zoleikha Jahanbakhsh-Nagadeh[một, d], Elnaz Zafarani-Moattar[một, e], Mehrdad Ranjbar-Khadivi[một, f]

[Một]*Phòng thí nghiệm Hệ thống Trí tuệ Máy tính, Khoa Kỹ thuật Máy tính, Đại học Tabriz, Tabriz, Iran*
[b]*Khoa Kỹ thuật Máy tính, Đại học Tabriz, Tabriz, Iran*
[c]*Khoa Kỹ thuật phần mềm, Đại học Ayvansaray, Istanbul, Thổ Nhĩ Kỳ*
[d]*Khoa Kỹ thuật Máy tính, Chi nhánh Naghadeh, Đại học Hồi giáo Azad, Naghadeh, Iran*
[e]*Khoa Kỹ thuật Máy tính, Chi nhánh Tabriz, Đại học Hồi giáo Azad, Tabriz, Iran*
[f]*Khoa Kỹ thuật Máy tính, Chi nhánh Shabestar, Đại học Hồi giáo Azad, Shabestar, Iran.*

trừu tượng

Lý lịchTrích xuất từ khóa là một chủ đề nghiên cứu phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên. Từ khóa là thuật ngữ mô tả thông tin phù hợp nhất trong tài liệu. Vấn đề chính mà các nhà nghiên cứu đang phải đối mặt là làm thế nào để trích xuất hiệu quả và chính xác các từ khóa cốt lõi từ một tài liệu. Tuy nhiên, các phương pháp trích xuất từ khóa trước đây đã sử dụng các tính năng văn bản và biểu đồ, nhưng vẫn còn thiếu mô hình có thể học và kết hợp các tính năng này một cách tốt nhất.

phương phápTrong bài báo này, chúng tôi phát triển một phương pháp trích xuất cụm từ khóa đa phương thức, cụ thể là*Người tạo cụm từ*, sử dụng kỹ thuật nhúng biến áp và đồ thị. Trong Phraseformer, mỗi từ khóa được đề xuất sẽ được trình bày bởi một vectơ là sự kết hợp của các biểu diễn học tập văn bản và cấu trúc. Người tạo cụm từ mất những lợi thế của các nghiên cứu gần đây như BERT và ExEm để duy trì cả hai cách biểu diễn. Ngoài ra, Phraseformer xử lý tác vụ trích xuất cụm từ khóa như một vấn đề gắn nhãn trình tự được giải quyết bằng cách sử dụng phân loại nhiệm vụ.

Kết quảChúng tôi phân tích hiệu suất của Phraseformer trên ba bộ dữ liệu bao gồm Inspec, SemEval2010 và SemEval 2017 theo điểm F1. Ngoài ra, chúng tôi điều tra hiệu suất của các bộ phân loại khác nhau trên phương pháp Phraseformer trên tập dữ liệu Inspec. Kết quả thực nghiệm chứng minh tính hiệu quả của phương pháp Phraseformer trên ba bộ dữ liệu được sử dụng. Ngoài ra, trình phân loại Rừng ngẫu nhiên đạt được điểm F1 cao nhất trong số tất cả các bộ phân loại.

Kết luậnDo sự kết hợp giữa BERT và ExEm có ý nghĩa hơn và có thể thể hiện ngữ nghĩa của từ tốt hơn. Do đó, Phraseformer hoạt động tốt hơn đáng kể so với các phương pháp đơn phương thức.

*Từ khóa:*Học biểu diễn đa phương thức, Trích xuất từ khóa, Biến áp, Nhúng đồ thị

*Đồng tác giả
*Địa chỉ email:*n.nikzad@tabrizu.ac.ir (N. Nikzad–Khasmakhi),mfeizi@tabrizu.ac.ir (Mohammad–Reza

1. Giới thiệu

Có sự tăng trưởng phi thường về số lượng tài liệu văn bản có sẵn cho người dùng trên các mạng xã hội khác nhau. media. Do đó, điều quan trọng là phải có những cách hiệu quả và hiệu quả để truy xuất và tóm tắt tất cả các tài liệu này [1]. Trích xuất từ khóa hoặc cụm từ khóa là giải pháp xác định tập hợp các thuật ngữ kết luận ý chính của một tài liệu [2]. Quá trình này giúp người đọc có thể nhanh chóng lĩnh hội được nội dung của tài liệu. Cụm từ khoá Các phương pháp trích xuất có thể được sử dụng hiệu quả bởi nhiều ứng dụng khai thác văn bản như lập chỉ mục, trực quan hóa, tóm tắt, phát hiện và theo dõi chủ đề, phân cụm và phân loại [3, 4].

Một số nghiên cứu đã được thực hiện để giải quyết vấn đề trích xuất cụm từ khóa. Một trong sô đo cách tiếp cận là các mô hình văn bản chỉ tập trung vào nội dung của tài liệu để lấy từ khóa. Một nhóm của một người đã sử dụng các phân tích từ vựng và cú pháp. Các cách tiếp cận khác của lớp này tận dụng ưu điểm của thống kê số như tần số thuật ngữ (TF) hoặc tần số tài liệu nghịch đảo tần số thuật ngữ (TF-IDF) [5]. Mặt khác, các phương pháp tiếp cận dựa trên biểu đồ tạo ra một nhóm phương pháp khác nhau bằng cách xây dựng một biểu đồ Của từ. Trong lớp này, các nút trung tâm nhất minh họa các từ khóa [1]. Ngoài ra, chúng ta có thể tìm thấy các mô hình lai sử dụng kết hợp các phương pháp dựa trên văn bản và biểu đồ để chọn từ khóa. Trong mọi trường hợp, chủ đề của Phương pháp kết hợp nào phù hợp và mạnh mẽ nhất để bắt các cụm từ khóa đang được thảo luận.

Trong bài báo này, chúng tôi mong muốn đề xuất một phương pháp kết hợp giữa mô hình dựa trên đồ thị và mô hình văn bản. Cũng, chúng tôi xem nhiệm vụ trích xuất cụm từ khóa là các vấn đề về phân loại và gắn thẻ theo trình tự. Sau khi tiền xử lý Loại bỏ các từ dừng và dấu chấm câu, chúng tôi xây dựng một biểu đồ sự xuất hiện không có trọng số và vô hướng cho tất cả các tài liệu trong kho văn bản. Sau đó, chúng tôi sử dụng ba kỹ thuật nhúng biểu đồ bao gồm ExEm, Node2vec và DeepWalk để tìm hiểu cách biểu diễn cấu trúc của các từ. Mặt khác, mỗi tài liệu riêng lẻ được đưa vào BERT Transformer để nhúng từ. Trong bước tiếp theo, chúng tôi ghép hai phần nhúng thu được từ biểu đồ và nội dung của tài liệu để tạo một cách biểu thị duy nhất cho mỗi từ. Sau đó, chúng tôi xây dựng việc trích xuất cụm từ khóa dưới dạng nhiệm vụ ghi nhãn theo trình tự chỉ định gắn thẻ BIO cho từng từ trong tài liệu. Ngoài ra, phương pháp đề xuất của chúng tôi xử lý trình tự ghi nhãn như một nhiệm vụ phân loại. Đầu ra của phân loại là thẻ BIO trong đó một từ ở đầu (B) của cụm từ khóa, bên trong (I) của cụm từ khóa và bên ngoài cụm từ khóa (O). Những đóng góp chính của bài viết này có thể được tóm tắt như sau:

- Theo hiểu biết tốt nhất của chúng tôi, công việc này là công việc đầu tiên áp dụng cách tiếp cận đa phương thức để trích xuất các cụm từ khóa bằng cách sử dụng kỹ thuật nhúng Transformer và đồ thị.

Feizi–Derakhshi),m.asgari@tabrizu.ac.ir (Meysam Asgari-Chenaghlu),balafarila@tabrizu.ac.ir (MA Balafar), derakhshi96@ms.tabrizu.ac.ir (Ali-Reza Feizi-Derakhshi),taymazfarshi@ayvansaray.edu.tr (Taymaz Rahkar-Farshi), m_ramezani@tabrizu.ac.ir (Majid Ramezani),zoleikha.jahanbakhsh@srbiau.ac.ir (Zoleikha Jahanbakhsh-Nagadeh), e.zafarani@iaut.ac.ir (Elnaz Zafarani-Moattar),mehrdad.khadivi@iaushab.ac.ir (Mehrdad Ranjbar-Khadivi)

- Ngoài ra, theo hiểu biết của chúng tôi, đây là nghiên cứu triển vọng chính sử dụng trực tiếp các kỹ thuật nhúng biểu đồ để chuyển đổi các từ trong biểu đồ xuất hiện đồng thời thành các vectơ có chiều thấp và sử dụng các vectơ này để tìm các cụm từ khóa.
- Chúng tôi coi vấn đề trích xuất cụm từ khóa từ tài liệu như một nhiệm vụ gán nhãn theo trình tự. Hơn nữa, chúng tôi quan sát việc ghi nhãn trình tự dưới dạng nhiệm vụ phân loại bằng cách sử dụng mã hóa BIO cơ chế.

Phần còn lại của bài báo được cấu trúc như sau: Phần 2 tổng quan các công trình liên quan. Phần 3 trình bày của chúng tôi phương pháp được đề xuất và giải thích nó một cách chi tiết. Mô tả các thí nghiệm của chúng tôi được trình bày trong Phần 4. Phần 5 cung cấp các kết quả thử nghiệm. Cuối cùng, Phần 6 kết luận bài viết.

2. Công việc liên quan

Các kỹ thuật thoát được sử dụng để trích xuất từ khóa có thể thuộc ba nhóm: văn bản, dựa trên biểu đồ và các mô hình lai. Các phương pháp tiếp cận văn bản tạo ra các từ khóa trực tiếp từ văn bản gốc bằng cách áp dụng các các kỹ thuật xử lý ngôn ngữ. Trong khi đó, các phương pháp dựa trên biểu đồ sẽ chuyển đổi tài liệu thành biểu đồ xuất hiện đồng thời trong đó các nút biểu thị các từ và các cạnh hiển thị mối quan hệ giữa hai từ trong cửa sổ ngữ cảnh. Mặt khác, các mô hình lai tận dụng cả cách trình bày văn bản và biểu đồ của tài liệu để phát hiện từ khóa. Trong các đoạn văn sau, chúng ta sẽ điều tra ba loại này một cách chi tiết hơn. TRONG Nhìn chung, những đóng góp chính trong nghiên cứu của chúng tôi được tóm tắt như dưới đây.

Trong mô hình văn bản, mục đích là tạo từ khóa trực tiếp từ văn bản gốc [5]. Mô hình đơn giản nhất trong danh mục này sử dụng kỹ thuật TF-IDF để trích xuất từ khóa. Sau đó, các nghiên cứu tập trung vào máy phương pháp học tập để đào tạo một bộ phân loại để nắm bắt các từ khóa. Với sự ra đời của phương pháp học sâu các kiến trúc Mạng thần kinh chuyển đổi (CNN), mạng thần kinh tái phát (RNN) như Bộ nhớ ngắn hạn dài (LSTM) và Máy biến áp ngày nay là những giải pháp phổ biến cho nhiệm vụ này. Có một số cách tiếp cận văn bản bao gồm KEA [6], KP-Miner [7], WINGNUS [8], RAKE [9], YAKE [10], TNT-KID [11], [12, 13, 14, 15, 16].

Ý tưởng chính của các phương pháp dựa trên biểu đồ là xây dựng các tài liệu dạng biểu đồ đồng xuất hiện. Mạng đồng xuất hiện hiển thị sự tương tác của các từ trong kho ngữ liệu. Trong biểu đồ này, các từ đại diện cho các nút và có một ranh giới giữa hai từ nếu những từ này cùng xuất hiện trong một cửa sổ. Sau khi xây dựng các đồ thị đồng xuất hiện, một số thước đo trung tâm như mức độ, độ gần, độ giữa và vectơ riêng là áp dụng trên đó để tìm cụm từ khóa. Trong các phương pháp này, các từ khóa được xác định bởi các nút trung tâm nhất. Một số phương pháp bao gồm TextRank [17], CollabRank [18], DegExt[19], NE-Rank [20], TopicRank [21], Xếp hạng vị trí [22], M-GCKE [5], [23, 24, 25, 26, 27, 1] sử dụng lý thuyết đồ thị để chọn từ khóa.

Các mô hình lai cố gắng tham gia hai loại đã đề cập trước đó. Những mô hình này tính điểm cho các từ từ cả biểu đồ xuất hiện đồng thời và nội dung tài liệu. Những cách tiếp cận khác nhau sử dụng những cách khác nhau để kết hợp các điểm số này. Các tác giả trong nghiên cứu [28, 29, 30, 31] đã đề xuất các phương pháp lai.

Các nghiên cứu [32, 33, 34, 35, 36] đã tiến hành đánh giá các kỹ thuật trích xuất từ khóa và cụm từ khóa. TRONG nghiên cứu của chúng tôi, bằng cách kết hợp các mô hình văn bản và dựa trên biểu đồ, đồng thời sử dụng gắn thẻ và phân loại theo trình tự, chúng tôi cố gắng phát triển một phương pháp trích xuất từ khóa hiệu quả có thể loại bỏ những hạn chế của các phương pháp trước đó. học.

3. Phương pháp đề xuất

Trong phần này, chúng tôi đề xuất một khung, được gọi là Phraseformer, giúp trích xuất các cụm từ khóa thông qua việc học biểu diễn ngữ cảnh từ thông tin văn bản và biểu diễn nút trong mạng đồng thời. Các bước xử lý tổng thể của Phraseformer được giải thích trong Hình 1. Luồng khung bao gồm bốn bước chính các bước như sau:

Học văn bản:Tìm cách trình bày văn bản cho tất cả các từ bằng BERT Transformer. Phần này của Phraseformer cung cấp sự hiểu biết sâu sắc hơn để đánh giá sự giống nhau về ngữ nghĩa giữa các từ.

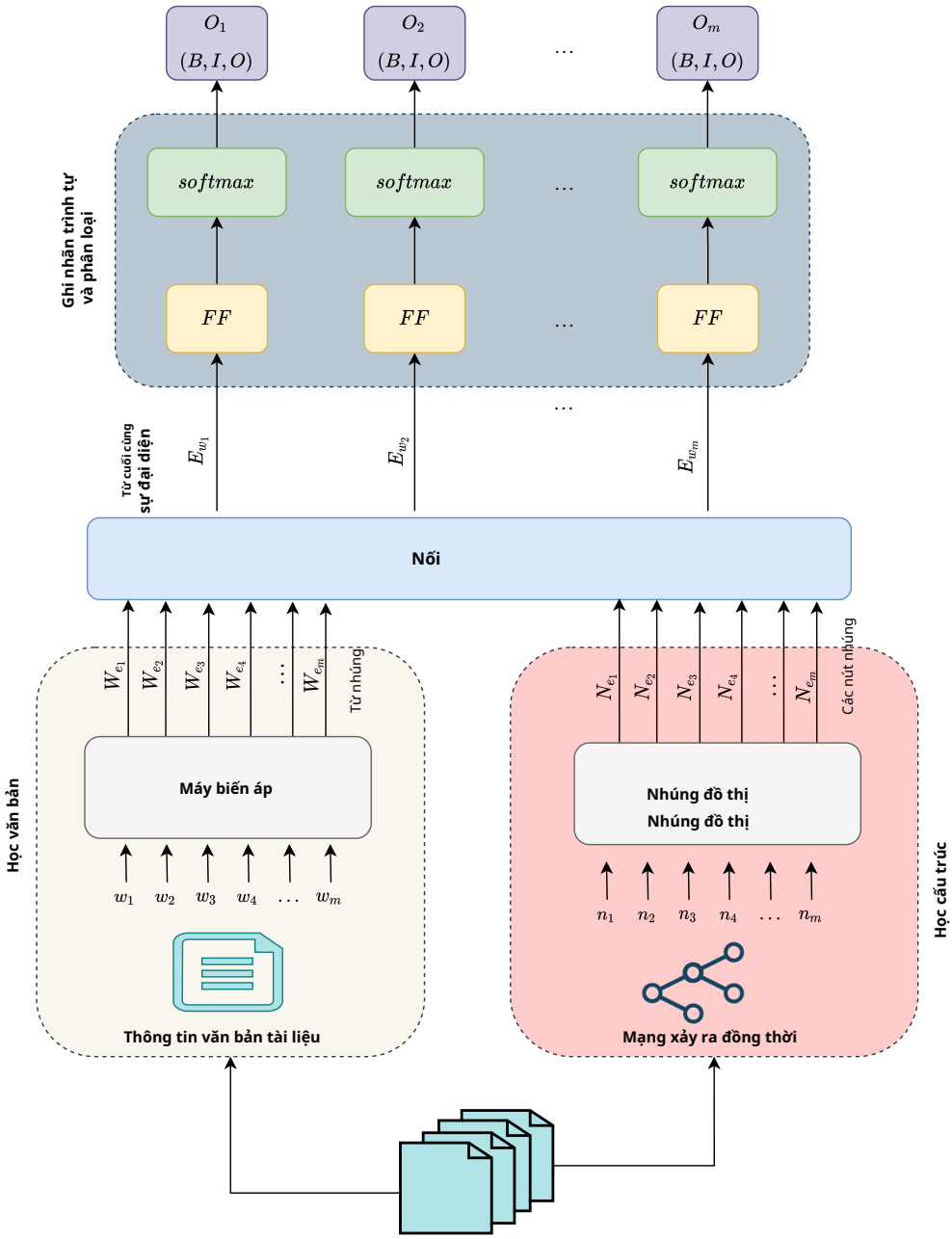
Học cấu trúc:Tạo ngữ cảnh của mỗi từ từ biểu đồ xuất hiện bằng cách nhúng biểu đồ các kỹ thuật để học biểu diễn cấu trúc.

Đại diện từ cuối cùng:Nối thông tin văn bản và thông tin cấu trúc và tạo ra một biểu diễn duy nhất cho mỗi từ.

Ghi nhãn và phân loại trình tự:Xây dựng việc trích xuất cụm từ khóa như một tác vụ ghi nhãn theo trình tự và gắn nhãn cho từng từ dựa trên sơ đồ gắn thẻ BIO thông qua một lớp được kết nối đầy đủ để phân loại từ đó.

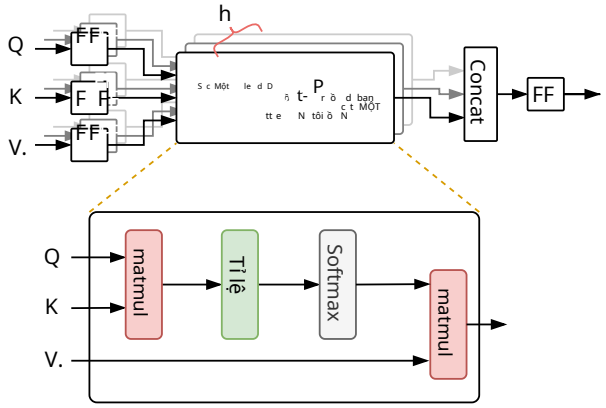
3.1. Học văn bản

Bước đầu tiên của Phraseformer là tạo vectơ văn bản cho mỗi từ. Với sự ra đời của Transformers, cách làm việc với dữ liệu văn bản đã thực sự thay đổi. Máy biến áp khắc phục nhược điểm của Kiến trúc RNN và CNN. Áp dụng khả năng tự chú ý cho phép Transformers có tính song song hơn nhiều hơn các kiến trúc khác [37]. Một Transformer bao gồm các thành phần mã hóa và giải mã như trong Hình 2. Thành phần mã hóa bao gồm một số khối mã hóa có hai lớp: lớp Chú ý nhiều đầu và lớp Mạng thần kinh chuyển tiếp nguồn cấp dữ liệu [38]. Ở phía bên kia, các khối có Mặt nạ Lớp Chú ý nhiều đầu trước lớp chuyển tiếp nguồn cấp dữ liệu tạo thành phần giải mã [39]. Ngoài ra, cả hai các thành phần có cùng số khối. Có nhiều mô hình khác nhau dựa trên cấu trúc máy biến áp chẳng hạn như BERT [40], OpenGPT [41, 42], XLNet [43] và ELMo [44]. Trong nghiên cứu này, kỹ thuật học biểu diễn văn bản là BERT Transformer có cấu trúc được trình bày trong Hình 3. Một trong những kỹ thuật quan trọng Ưu điểm của BERT so với các mô hình như Word2Vec là BERT tạo từ nhúng cho mỗi từ

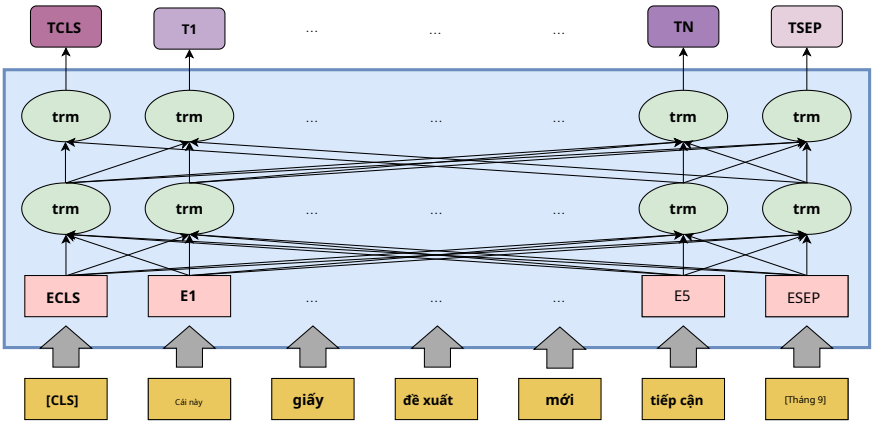


Hình 1: Cấu trúc tổng thể của Phraseformer.

dựa trên từng câu hoặc từng tài liệu mà từ đó có trong đó. Điều đó có nghĩa là BERT có khả năng nắm bắt ngữ cảnh của một từ trong tài liệu. Thông tin văn bản của một tài liệu bao gồm các tài liệu ' tựa đề, tóm tắt. Khối bên trái trong Hình 1 hiển thị quá trình học các vectơ từ trong văn bản.



Hình 2: Kiến trúc mô hình Transformer [45].



Hình 3: Kiến trúc máy biến áp BERT [46].

3.2. Học cấu trúc

Bước thứ hai là tìm hiểu vectơ cấu trúc cho mỗi từ dựa trên ba thuật toán nhúng đồ thị bao gồm ExEm [47] , Node2vec [48] và DeepWalk [49] trong mạng đồng xảy ra. Sự xuất hiện đồng thời mạng được định nghĩa là một biểu đồ $G = (Đ, Ã)$ trong đó các nút đại diện cho các từ và mỗi cạnh $e \in E$ thể hiện mối quan hệ cùng xuất hiện giữa từ $v_{Tôi}$ và từ v_j xuất hiện trong cửa sổ ngữ cảnh. Vì số lượng thông tin có thể thu được từ một tài liệu đơn độc là hữu hạn, chúng tôi xây dựng biểu đồ sự xuất hiện trên tất cả các tài liệu thay vì một tài liệu duy nhất, như khối bên phải trong Hình 1 được hiển thị. Cần lưu ý rằng trước khi xây dựng biểu đồ này, chúng tôi sẽ loại bỏ các từ dừng và dấu chấm câu. Vậy nhiệm vụ của chúng ta là học

các biểu diễn tiềm ẩn theo chiều của các từ trong biểu đồ này bằng cách sử dụng kỹ thuật nhúng biểu đồ. Có các thuật toán học biểu diễn nút khác nhau. Theo kết quả được báo cáo bởi các nghiên cứu khác nhau, chúng tôi chọn ba phương pháp nhúng biểu đồ dựa trên bước đi ngẫu nhiên để mang lại hiệu suất tốt hơn. Trong DeepWalk, một tập hợp các bước đi ngẫu nhiên được tạo bằng cách bắt đầu từ mọi nút trong biểu đồ. Trong khi đó, Node2vec đề xuất một phương pháp bước đi ngẫu nhiên thiên vị là phiên bản sửa đổi của DeepWalk. Phương pháp này sử dụng hai tham số để kiểm soát không gian tìm kiếm. Hơn nữa, ExEm là một kỹ thuật khác sử dụng lý thuyết tập hợp ưu thế để tạo ra các bước đi ngẫu nhiên. ExEm mô tả đặc điểm của các vùng lân cận cục bộ bằng cách bắt đầu mỗi đường dẫn được lấy mẫu bằng một nút thống trị. Ngoài ra, thông tin cấu trúc toàn cầu được nắm bắt bằng cách chọn một nút thống trị khác trong bước đi ngẫu nhiên Cả ba cách tiếp cận đều đưa các bước đi ngẫu nhiên này vào mô hình mạng thần kinh Skip-gram để tìm hiểu cách biểu diễn nút.

3.3. Đại diện từ cuối cùng

Sau khi chúng ta có được biểu diễn vectơ của mỗi từ về thông tin văn bản và cấu trúc mạng đồng xuất hiện, bước tiếp theo là kết hợp các thông tin này thành một biểu diễn duy nhất. Chúng tôi tin rằng Việc ghép thông tin dựa trên văn bản và dựa trên cấu trúc có thể khám phá tốt hơn tiềm năng của các từ trong là từ khóa. Vì vậy, chúng tôi trình bày mỗi từ bằng một vectơ đơn độc là sự kết hợp giữa văn bản và vectơ cấu trúc. Ví dụ đối với từ $w_{Tôi}$, chúng tôi biểu diễn $E_{wTôi} = W_{eTôi} + N_{eTôi}$ Ở đâu $W_{eTôi}$ Và $N_{eTôi}$ biểu thị các biểu diễn học tập văn bản và cấu trúc cho từ $w_{Tôi}$, tương ứng.

3.4. Ghi nhãn và phân loại trình tự

Ghi nhãn trình tự là một loại nhiệm vụ nhận dạng mẫu trong miền NLP nhằm phân loại các từ trong văn bản và gán một lớp hoặc nhãn cho mỗi từ trong một chuỗi đầu vào nhất định [50, 51]. Có rất nhiều kỹ thuật đối với nhiệm vụ ghi nhãn trình tự bao gồm Mô hình Markov ẩn [52], Trường ngẫu nhiên có điều kiện (CRF) [53] và các phương pháp học sâu.

Trong bài báo này, chúng tôi coi vấn đề trích xuất cụm từ khóa từ tài liệu như một nhiệm vụ gán nhãn theo trình tự. Ngoài ra, chúng tôi quan sát việc ghi nhãn trình tự dưới dạng một nhiệm vụ phân loại bằng cách sử dụng sơ đồ mã hóa BIO như nhãn đầu ra. Vì vậy, mô hình của chúng tôi mất $E_{w_1}, E_{w_2}, \ldots, E_{w_{tôi}}$ làm đầu vào và gán cho mỗi từ một nhãn $ờ_{Tôi} \in \{B, tôi, O\}$ Ở đâu B chỉ ra rằng $w_{Tôi}$ là sự khởi đầu của một cụm từ khóa, $TÔI$ biểu thị rằng $w_{Tôi}$ nằm trong một cụm từ khóa và cuối cùng $ờ$ minh họa rằng $w_{Tôi}$ nằm ngoài cụm từ khóa. Có thể thấy từ Hình 1 rằng cấu trúc được kết nối đầy đủ và lớp softmax được sử dụng để đưa ra quyết định phân loại.

4. Đánh giá thực nghiệm

Trong phần nghiên cứu này, chúng tôi sẽ làm rõ những bộ dữ liệu nào đã được sử dụng để đánh giá Phraseformer. Ngoài ra, chúng tôi sẽ mô tả các thuật toán cơ bản để so sánh phương pháp được đề xuất của chúng tôi với chúng. bên trong Phần tiếp theo, chúng tôi sẽ trình bày các phiên bản khác nhau của Phraseformer. Cuối cùng, các số liệu đánh giá sẽ được chỉ định.

4.1. Tập dữ liệu

Để đánh giá hiệu quả của Phraseformer, chúng tôi sử dụng ba bộ dữ liệu bao gồm Inspec [54], SE-2010 [55] và SE-2017 [56]. Chúng tôi sẽ mô tả các bộ dữ liệu này trong các đoạn tiếp theo.

Kiểm trabao gồm các bản tóm tắt các bài báo về Khoa học Máy tính được thu thập từ năm 1998 đến 2002.

SE-2010chứa đầy đủ các bài báo khoa học được lấy từ Thư viện kỹ thuật số ACM. Trong của chúng tôi thử nghiệm, chúng tôi đã sử dụng phần tóm tắt của các bài báo.

SE-2017bao gồm các đoạn được chọn từ 500 bài báo tạp chí ScienceDirect từ Khoa học Máy tính, Lĩnh vực Khoa học Vật liệu và Vật lý.

Cần lưu ý rằng do việc xây dựng việc trích xuất cụm từ khóa như một nhiệm vụ ghi nhãn trình tự, chúng ta xem xét các cụm từ khóa xuất hiện trong phần tóm tắt của các bài viết trong ba bộ dữ liệu. Bảng 1 cho thấy số liệu thống kê[1]. của ba tập dữ liệu trên.

Bảng 1: Thống kê tập dữ liệu.

Tập dữ liệu	Loại tài liệu	# bác sĩ	# Chìa khóa vàng (mỗi tài liệu)	# Token cho mỗi tài liệu	Chìa khóa vàng vắng mặt
Kiểm tra	trừu tượng	2000	29230 (14.62)	128,20	37,7%
SemEval2010	Giấy	243	4002 (16,47)	8332.34	11,3%
SemEval2017	Đoạn văn	493	8969 (18.19)	178,22	0,0%

4.2. Mô hình cơ sở

Trong nghiên cứu này, chúng tôi so sánh phương pháp Phraseformer với bốn phương pháp dựa trên biểu đồ và ba phương pháp văn bản. Trong các đoạn sau, chúng tôi sẽ giải thích các kỹ thuật này chi tiết hơn.

Xếp hạng văn bản [17] là phương pháp dựa trên biểu đồ đơn giản nhất dựa trên thuật toán PageRank.

DeepWalk [49] là một kỹ thuật nhúng đồ thị dựa trên các bước đi ngẫu nhiên. Phương pháp này đại diện cho mỗi nút dưới dạng một vectơ có chiều thấp.

Nút2vec [48] là phiên bản sửa đổi của DeepWalk sử dụng bước đi ngẫu nhiên thiên vị để chuyển đổi các nút thành vectơ.

ExEm [47] là một cách tiếp cận dựa trên bước đi ngẫu nhiên sử dụng lý thuyết tập hợp thống trị để tạo ra các bước đi ngẫu nhiên. ExEm$_{w2v}$và ExEm$_{fl}$là hai phiên bản khác nhau của ExEm.

Word2Vec [57] học cách biểu diễn vector của các từ. Phương pháp này chuyển tài liệu qua hai lớp chuyển tiếp nguồn cấp dữ liệu sang các vectơ được tạo [37].

BiLSTM-CRF [58] là một phương pháp văn bản coi việc trích xuất cụm từ khóa là ghi nhãn trình tự nhiệm vụ sử dụng kiến trúc BiLSTM-CRF.

[1]Chúng tôi có được thông tin này từhttps://github.com/LIAAD/KeywordExtractor-Datasets

BERT [40] là một cách tiếp cận văn bản sử dụng cấu trúc biến đổi để thu được cách trình bày tài liệu.

4.3. Các biến thể của phương pháp

Chúng tôi trình bày bốn biến thể của Phraseformer sử dụng các kỹ thuật nhúng biểu đồ khác nhau cho cấu trúc học hỏi. Trình tạo cụm từ(BERT, DeepWalk), Trình tạo cụm từ(BERT, Node2vec), Trình tạo cụm từ(BERT, ExEm$_{w2v}$) và Trình tạo cụm từ(BERT, ExEm$_{ft}$) là các mô hình Phraseformer khác nhau sử dụng DeepWalk, Node2vec, ExEm$_{w2v}$và ExEm$_{ft}$các cách tiếp cận tương ứng để có được biểu diễn từ từ biểu đồ cùng xuất hiện.

4.4. Số liệu

Để đo lường hiệu quả thử nghiệm, chúng tôi đã sử dụng điểm F1 được tính toán như sau:

$$F1-điểm = 2 \times \frac{\frac{Y \cap Y'}{Y'} \times \frac{Y \cap Y'}{Y}}{\frac{Y \cap Y'}{Y'} + \frac{Y \cap Y'}{Y}} \quad (1)$$

đây Y'và Ylần lượt là từ khóa được dự đoán và từ khóa thực.

5. Kết quả thí nghiệm

Phần này trình bày lại các kết quả thử nghiệm. Đầu tiên, chúng tôi so sánh Phraseformer với đường cơ sở. Sau đó, chúng tôi kiểm tra các bộ phân loại khác nhau cho phần phân loại và báo cáo kết quả.

5.1. So sánh cơ bản

chúng tôi so sánh mô hình đa phương thức của mình với các phương pháp dựa trên văn bản và biểu đồ. Bảng 2 trình bày kết quả. Đúng như dự đoán, mô hình của chúng tôi vượt trội hơn đáng kể so với tất cả các phương pháp. Từ kết quả ta có những quan sát sau:Tôi)Có thể kết luận rằng các phương pháp Phraseformer đạt được điểm F1 cao nhất. Trình tạo cụm từ(BERT, ExEm$_{ft}$) tăng cường hiệu suất bằng cách6.4%, 19.94%Và13.70%so sánh với BERT trên các bộ dữ liệu Inspec, SE-2010 và SE-2017 tương ứng. Hơn nữa, Phraseformer(BERT, ExEm$_{ft}$) vượt trội hơn ExEm bởi112.6%, 290.4%Và130.4%trên các tập dữ liệu đã đề cập. Những điểm số cao này chứng tỏ rằng giả thuyết của chúng tôi về việc sử dụng phương pháp học tập đa phương thức, gắn nhãn và phân loại theo trình tự là đúng.ii)Rõ ràng là hai phương pháp dựa trên biểu đồ và văn bản cho thấy hiệu suất kém.iii)Rõ ràng là văn bản phương pháp có kết quả khả quan so với các mô hình dựa trên đồ thị.iv)Hơn nữa, BERT cho thấy tốt hơn quả hơn các phương pháp văn bản khác. Bởi vì BERT nhúng các từ vào vectơ bằng cách xem xét ý nghĩa của các từ trong câu.v)Ngoài ra, kết quả của các phương pháp dựa trên biểu đồ chứng minh rằng kỹ thuật nhúng biểu đồ đạt được điểm F1 cao nhất so với TextRank. Ngoài ra, các phiên bản khác nhau của ExEm còn nhiều hơn thế. thành công hơn các phương pháp nhúng đồ thị khác. Lý do là ExEm tuân theo nguyên tắc đồng âm và sự tương đương về vai trò cấu trúc trong việc học các biểu diễn nút với sự trợ giúp của các nút thống trị.

Bảng 2: So sánh với các phương pháp cơ bản (điểm F1).

Loại		Người mẫu	Tập dữ liệu		
			Kiểm tra	SE-2010	SE-2017
Dựa trên đồ thị	phương pháp	Xếp hạng văn bản [17]	0,1780	0,1990	0,2090
		Đi sâu [49]	0,3190	0,1102	0,2887
		Nút2vec [48]	0,3138	0,1098	0,2863
		ExEm$_{w2v}$[47]	0,3273	0,1233	0,2911
		ExEm$_{ft}$[47]	0,3286	0,1246	0,2913
Dựa trên văn bản	phương pháp	Word2Vec [57]	0,4730	0,2080	0,2920
		BiLSTM-CRF [58]	0,5930	0,3570	0,5210
		BERT [40]	0,6564	0,4056	0,5904
Hỗn hợp	phương pháp	Trình tạo cụm từ(BERT, Deep-Đi bộ)	0,6844	0,4722	0,6570
		Người tạo cụm từ(BERT, Node2vec)	0,6868	0,4746	0,6594
		Người tạo cụm từ(BERT, ExEm$_{w2v}$)	0,6970	0,4848	0,6696
		Người tạo cụm từ(BERT, ExEm$_{ft}$)	0,6987	0,4865	0,6713

5.2. Trình phân loại

Trong phần thử nghiệm này, chúng tôi mong muốn điều tra xem bộ phân loại nào phù hợp nhất để ghi nhãn trình tự và nhiệm vụ phân loại để tìm các cụm từ khóa. Bảng 3 minh họa hiệu suất của các trình phân loại khác nhau trên phương pháp Phraseformer trên tập dữ liệu Inspec. Từ kết quả có thể thấy rằng bộ phân loại Rừng ngẫu nhiên đạt được hiệu suất tốt nhất trong số tất cả các phân loại, như được hiển thị bằng kết quả được gạch chân. Mặt khác tay, các kết quả đậm nét trong bảng này trình bày những mô hình nào của Phraseformer có thể đạt được giá trị cao nhất trên mỗi bộ phân loại. Rõ ràng là sự kết hợp giữa BERT và ExEm có ý nghĩa hơn và có thể thể hiện ngữ nghĩa của từ tốt hơn. Do đó, Phraseformer có thể trích xuất chính xác các từ khóa.

Bảng 3: So sánh các phân loại trên tập dữ liệu Inspec (điểm F1)

Người mẫu	Trình phân loại			
	Rừng ngẫu nhiên	SVM	Hồi quy logistic	Đã kết nối đầy đủ
Người tạo cụm từ(BERT, Sâu-Đi bộ)	0,7024	0,6564	0,6714	0,6844
Người tạo cụm từ(BERT, Node2vec)	0,7048	0,6588	0,6738	0,6868
Người tạo cụm từ(BERT, ExEm$_{w2v}$)	0,7150	0,6707	0,6857	0,6970
Người tạo cụm từ(BERT, ExEm$_{ft}$)	0,7167	0,6690	0,6840	0,6987

6. Kết luận

Trong bài viết này, một phương pháp tiếp cận đa phương thức, Phraseformer, để trích xuất các cụm từ khóa từ tài liệu được đề xuất. Trong cách tiếp cận này, hai phương thức bắt nguồn từ biểu đồ xảy ra đồng thời và nội dung của tài liệu. Hơn nữa, chúng tôi xây dựng việc trích xuất cụm từ khóa như một nhiệm vụ ghi nhãn trình tự được giải quyết bằng mô hình phân loại. Để biểu diễn thông tin của các phương thức dưới dạng vectơ, chúng tôi sử dụng BERT Transformer và nhúng biểu đồ kỹ thuật. Việc sử dụng thông tin dựa trên biểu đồ với sự trợ giúp của ngữ nghĩa văn bản mang lại cái nhìn sâu sắc hơn trình bày các từ khóa mang lại phương pháp trích xuất cụm từ khóa mạnh mẽ hơn. Cuối cùng, để xác thực tính hiệu quả của phương pháp Phraseformer được đề xuất, chúng tôi tiến hành thử nghiệm trên ba bộ dữ liệu và kết quả cho thấy Phraseformer hoạt động tốt hơn đáng kể như thế nào so với các phương pháp đơn phương thức.

Nhìn nhận

Dự án này được hỗ trợ bởi khoản tài trợ nghiên cứu của Đại học Tabriz (Số S/806).

Tuyên bố

Kinh phí

Dự án này được hỗ trợ bởi khoản tài trợ nghiên cứu của Đại học Tabriz (Số S/806).

Xung đột lợi ích

Các tác giả không có xung đột lợi ích khi tuyên bố rằng có liên quan đến nội dung của bài viết này.

Phê duyệt đạo đức

Bài viết này không chứa bất kỳ nghiên cứu nào với người tham gia hoặc động vật được thực hiện bởi bất kỳ cơ quan nào.

các tác giả.

Người giới thiệu

[1] DA Vega-Oliveros, PS Gomes, EE Milios, L. Berton, Chỉ mục đa trung tâm để trích xuất từ khóa dựa trên biểu đồ, Xử lý & Quản lý thông tin 56 (6) (2019) 102063.doi:10.1016/j.ipm.2019.102063.

[2] MW Berry, J. Kogan, Khai thác văn bản: ứng dụng và lý thuyết, John Wiley & Sons, 2010.

[3] S. Lahiri, Từ khóa tại nơi làm việc: Điều tra việc trích xuất từ khóa trong các ứng dụng truyền thông xã hội, Ph.D. luận án (2018).

[4] C. Zhang, Trích xuất từ khóa tự động từ tài liệu bằng các trường ngẫu nhiên có điều kiện, Tạp chí Hệ thống thông tin tính toán 4 (3) (2008) 1169–1180.

[5] B. Wang, B. Yang, S. Shan, H. Chen, Phát hiện các chủ đề nóng từ dữ liệu lớn học thuật, IEEE Access 7 (2019) 185916–185927. doi:10.1109/ACCESS.2019.2960285.

[6] IH Witten, GW Paynter, E. Frank, C. Gutwin, CG Nevill-Manning, Kea: Trích xuất cụm từ khóa tự động thực tế, trong: Thiết kế và Khả năng sử dụng của Thư viện Kỹ thuật số: Nghiên cứu điển hình ở Châu Á Thái Bình Dương, IGI toàn cầu, 2005, trang 129–152.doi: 10.4018/978-1-59140-441-5.ch008.

[7] SR El-Beltagy, A. Rafea, Kp-miner: Hệ thống trích xuất cụm từ khóa cho tài liệu tiếng Anh và tiếng Ả Rập, Hệ thống thông tin 34 (1) (2009) 132–144.doi:10.1016/j.is.2008.05.002.

[8] TD Nguyễn, M.-T. Lương, Wingnus: Trích xuất cụm từ khóa sử dụng cấu trúc logic tài liệu, trong: Kỷ yếu hội thảo quốc tế lần thứ 5 về đánh giá ngữ nghĩa, 2010, trang 166–169. URLhttps://www.aclweb.org/anthology/S10-1035

[9] S. Rose, D. Engel, N. Cramer, W. Cowley, Trích xuất từ khóa tự động từ các tài liệu riêng lẻ, Khai thác văn bản: ứng dụng và lý thuyết 1 (2010) 1–20.doi:10.1002/9780470689646.ch1.

[10] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, Yake! trích xuất từ khóa từ các tài liệu đơn lẻ bằng nhiều tính năng cục bộ, Khoa học thông tin 509 (2020) 257–289.doi:10.1016/j.ins.2019.09.013.

[11] M. Martinc, B. Škrlj, S. Pollak, Tnt-kid: Trình gắn thẻ thần kinh dựa trên máy biến áp để nhận dạng từ khóa, bản in trước arXiv arXiv:2003.09166 (2020).

[12] M. Basaldella, E. Antolli, G. Serra, C. Tasso, Mạng thần kinh tái phát lstm hai chiều để trích xuất cụm từ khóa, trong: Hội nghị nghiên cứu Ý về Thư viện kỹ thuật số, Springer, 2018, trang 180–187.doi:10.1007/978-3-319-73165-0_18.

[13] R. Alzaidy, C. Caragea, CL Giles, ghi nhãn chuỗi Bi-lstm-crf để trích xuất cụm từ khóa từ các tài liệu học thuật, trong: Hội nghị web toàn cầu, 2019, trang 2551–2557.doi:10.1145/3308558.3313642.

[14] M. Tang, P. Gandhi, MA Kabir, C. Zou, J. Blakey, X. Luo, Phân loại ghi chú tiến độ và trích xuất từ khóa bằng mô hình học sâu dựa trên sự chú ý với bert, arXiv bản in trước arXiv:1910.05786 (2019) .

[15] J. Wang, F. Song, K. Walia, J. Farber, R. Dara, Sử dụng mạng thần kinh tích chập để trích xuất các từ khóa và cụm từ khóa: Một nghiên cứu điển hình về các bệnh do thực phẩm, trong: Hội nghị quốc tế IEEE lần thứ 18 năm 2019 về học máy Và Ứng dụng (ICMLA), IEEE, 2019, trang 1398–1403.doi:10.1109/ICMLA.2019.00228.

[16] Y. Kim, JH Lee, S. Choi, JM Lee, J.-H. Kim, J. Seok, HJ Joo, Xác thực thuật toán xử lý ngôn ngữ tự nhiên học sâu để trích xuất từ khóa từ báo cáo bệnh lý trong hồ sơ sức khỏe điện tử, Báo cáo khoa học 10 (1) (2020) 1–9. doi:10.1038/s41598-020-77258-w.

[17] R. Mihalcea, P. Tarau, Textrank: Đưa trật tự vào văn bản, trong: Kỷ yếu của hội nghị năm 2004 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên, 2004, trang 404–411. URLhttps://www.aclweb.org/anthology/W04-3252

[18] X. Wan, J. Xiao, Collabrank: hướng tới cách tiếp cận hợp tác để trích xuất cụm từ khóa một tài liệu, trong: Kỷ yếu của Hội nghị quốc tế lần thứ 22 về Ngôn ngữ học tính toán (Coling 2008), 2008, trang 969–976. URLhttps://www.aclweb.org/anthology/C08-1122

[19] M. Litvak, M. Last, H. Aizenman, I. Gobits, A. Kandel, Degext—một công cụ trích xuất cụm từ khóa dựa trên biểu đồ độc lập với ngôn ngữ, trong: Những tiến bộ trong việc làm chủ web thông minh–3, Springer, 2011, trang . 121–130.doi:10.1007/978-3-642-18029-3_13.

[20] A. Bellaachia, M. Al-Dhelaan, Ne-rank: Một trích xuất cụm từ khóa dựa trên biểu đồ mới trên twitter, trong: Hội nghị quốc tế IEEE/WIC/ACM 2012 về Trí tuệ web và Công nghệ tác nhân thông minh, Tập. 1, IEEE, 2012, trang 372–379. doi:10.1109/WI-IAT.2012.82.

[21] A. Bougouin, F. Boudin, B. Daille, TopicRank: Xếp hạng chủ đề dựa trên biểu đồ để trích xuất cụm từ khóa, trong: Kỷ yếu của Hội nghị chung quốc tế lần thứ sáu về xử lý ngôn ngữ tự nhiên, Liên đoàn xử lý ngôn ngữ tự nhiên châu Á, Nagoya, Nhật Bản , 2013, trang 543–551. URLhttps://www.aclweb.org/anthology/I13-1062

[22] C. Florescu, C. Caragea, Xếp hạng vị trí: Một cách tiếp cận không giám sát để trích xuất cụm từ khóa từ các tài liệu học thuật, trong: Kỷ yếu của Hội nghị thường niên lần thứ 55 của Hiệp hội Ngôn ngữ học tính toán (Tập 1: Tài liệu dài), 2017, trang 1105 – 1115.doi:10.18653/v1/P17-1102.

[23] F. Boudin, So sánh các biện pháp trọng tâm để trích xuất cụm từ khóa dựa trên biểu đồ, trong: Kỷ yếu của hội nghị chung quốc tế lần thứ sáu về xử lý ngôn ngữ tự nhiên, 2013, trang 834–838. URLhttps://www.aclweb.org/anthology/I13-1102

[24] WD Abilhoa, LN De Castro, Phương pháp trích xuất từ khóa từ các tin nhắn Twitter được biểu diễn dưới dạng biểu đồ, Toán học ứng dụng và tính toán 240 (2014) 308–325.doi:10.1016/j.amc.2014.04.090.

[25] A. Tixier, F. Malliaros, M. Vazirgiannis, Cách tiếp cận dựa trên suy biến đồ thị để trích xuất từ khóa, trong: Kỷ yếu của hội nghị năm 2016 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên, 2016, trang 1860–1870.doi:10.18653/v1/D16-1191.

[26] MS El BazzI, D. Mammass, T. Zaki, A. Ennaji, Mô hình xếp hạng dựa trên biểu đồ để trích xuất cụm từ khóa tự động từ tài liệu tiếng Ả Rập, trong: Hội nghị công nghiệp về khai thác dữ liệu, Springer, 2017, trang 313–322 .doi:10.1007/978-3-319-62701-4_25.

[27] F. Boudin, Trích xuất cụm từ khóa không được giám sát bằng biểu đồ nhiều phần, trong: Kỷ yếu Hội nghị năm 2018 của Chi hội Bắc Mỹ của Hiệp hội Ngôn ngữ học tính toán: Công nghệ ngôn ngữ con người, 2018, tr. 667–672.doi:10.18653/v1/N18-2105.

[28] S. Danesh, T. Sumner, JH Martin, Sgrank: Kết hợp các phương pháp thống kê và đồ họa để cải thiện tính năng tiên tiến trong trích xuất cụm từ khóa không giám sát, trong: Kỷ yếu của hội nghị chung lần thứ tư về ngữ nghĩa từ vựng và tính toán, 2015, trang . 117–126.doi:10.18653/v1/S15-1013.

[29] Y. Zhang, Y. Chang, X. Liu, SD Gollapalli, X. Li, C. Xiao, Mike: trích xuất cụm từ khóa bằng cách tích hợp thông tin đa chiều, trong: Kỷ yếu của Hội nghị ACM 2017 về Quản lý Thông tin và Tri thức, 2017, trang 1349–1358.doi:10.1145/3132847.3132956.

[30] D. Mahata, J. Kuriakose, R. Shah, R. Zimmermann, Key2vec: Trích xuất cụm từ khóa được xếp hạng tự động từ các bài báo khoa học bằng cách sử dụng tính năng nhúng cụm từ, trong: Kỷ yếu của Hội nghị Hiệp hội Ngôn ngữ học tính toán Bắc Mỹ năm 2018 : Công nghệ ngôn ngữ con người, Tập 2 (Bài viết ngắn), 2018, trang 634–639.doi: 10.18653/v1/N18-2100.

[31] D. Mahata, RR Shah, J. Kuriakose, R. Zimmermann, JR Talburt, Xếp hạng theo trọng số chủ đề của từ khóa từ tài liệu văn bản bằng cách sử dụng tính năng nhúng cụm từ, trong: Hội nghị IEEE 2018 về xử lý và truy xuất thông tin đa phương tiện (MIPR), IEEE , 2018, trang 184–189.doi:10.1109/MIPR.2018.00041.

[32] S. Siddiqi, A. Sharan, Kỹ thuật trích xuất từ khóa và cụm từ khóa: tổng quan tài liệu, Tạp chí quốc tế về ứng dụng máy tính 109 (2) (2015).

[33] ZA Merrouni, B. Frikh, B. Ouhbi, Trích xuất cụm từ khóa tự động: khảo sát và xu hướng, Tạp chí Hệ thống thông tin thông minh (2019) 1–34doi:10.1007/s10844-019-00558-9.

[34] Ö. Ünlü, A. Çetin, Khảo sát về trích xuất từ khóa và cụm từ khóa bằng học sâu, trong: Hội nghị chuyên đề quốc tế lần thứ 3 năm 2019 về nghiên cứu đa ngành và công nghệ đổi mới (ISMSIT), IEEE, 2019, trang 1–6.doi:10.1109/ISMSIT.2019. 8932811.

[35] E. Papagiannopoulou, G. Tsoumakas, Đánh giá về trích xuất cụm từ khóa, Đánh giá liên ngành của Wiley: Khai thác dữ liệu và khám phá kiến thức 10 (2) (2020) e1339.doi:10.1002/widm.1339.

[36] N. Firoozeh, A. Nazarenko, F. Alizon, B. Daille, Trích xuất từ khóa: Các vấn đề và phương pháp, Kỹ thuật ngôn ngữ tự nhiên 26 (3) (2020) 259–291.doi:10.1017/S1351324919000457.

[37] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Phân loại văn bản dựa trên học sâu: Đánh giá toàn diện, bản in trước arXiv arXiv:2004.03705 (2020).

[38] M. Asgari-Chenaghlu, M.-R. Feizi-Derakhshi, M.-A. Balafar, C. Motamed, và những người khác, Topicbert: Phương pháp tiếp cận biểu đồ bộ nhớ dựa trên phương pháp học chuyển đổi máy biến áp để phát hiện chủ đề truyền thông xã hội đa phương thức, bản in trước arXiv arXiv:2008.06877 (2020).

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, AN Gomez, L. Kaiser, I. Polosukhin, Chú ý là tất cả những gì bạn cần, bản in trước arXiv arXiv:1706.03762 (2017).

[40] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Đào tạo trước các máy biến áp hai chiều sâu để hiểu ngôn ngữ, bản in trước arXiv arXiv:1810.04805 (2018).

[41] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Cải thiện khả năng hiểu ngôn ngữ bằng cách đào tạo trước mang tính tổng quát (2018).
URLhttps://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

[42] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Mô hình ngôn ngữ là những người học đa nhiệm không được giám sát, blog OpenAI 1 (8) (2019) 9.

[43] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, QV Le, Xlnet: Huấn luyện trước tự hồi quy tổng quát để hiểu ngôn ngữ, arXiv preprint arXiv:1906.08237 (2019).

[44] ME Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Cách trình bày từ theo ngữ cảnh sâu sắc, bản in trước arXiv arXiv:1802.05365 (2018).

[45] M. Asgari-Chenaghlu, MR Feizi-Derakhshi, L. Farzinvash, C. Motamed, Một phương pháp học sâu đa phương thức để nhận dạng thực thể được đặt tên từ phương tiện truyền thông xã hội, bản in trước arXiv arXiv:2001.06888 (2020).

[46] N. Nikzad-Khasmakhi, M. Balafar, MR Feizi-Derakhshi, C. Motamed, Berters: Học biểu diễn đa phương thức cho hệ thống khuyến nghị của chuyên gia với máy biến áp, bản in trước arXiv arXiv:2007.07229 (2020).

[47] N. Nikzad-Khasmakhi, M. Balafar, MR Feizi-Derakhshi, C. Motamed, Exem: Chuyên gia nhúng sử dụng lý thuyết tập hợp thống trị với các phương pháp học sâu, bản in trước arXiv arXiv:2001.08503 (2020).

[48] A. Grover, J. Leskovec, Node2vec: Học tính năng có thể mở rộng cho mạng, trong: Kỷ yếu của Hội nghị quốc tế ACM SIGKDD về Khám phá kiến thức và khai thác dữ liệu, 2016.arXiv:1607.00653, doi:10.1145/2939672.2939754.

[49] B. Perozzi, R. Al-Rfou, S. Skiena, DeepWalk: Học trực tuyến về các biểu hiện xã hội, trong: Kỷ yếu của Hội nghị quốc tế ACM SIGKDD về Khám phá tri thức và khai thác dữ liệu, 2014.arXiv:1403.6652, doi:10.1145/ 2623330.2623732.

[50] Z. He, Z. Wang, W. Wei, S. Feng, X. Mao, S. Jiang, Một cuộc khảo sát về những tiến bộ gần đây trong việc ghi nhãn trình tự từ các mô hình học sâu, bản in trước arXiv arXiv:2011.06727 (2020).

[51] A. Akhundov, D. Trautmann, G. Groh, Ghi nhãn trình tự: Một cách tiếp cận thực tế, bản in trước arXiv arXiv:1808.03926 (2018).

[52] J. Kupiec, Gắn thẻ phần lời nói mạnh mẽ bằng cách sử dụng mô hình markov ẩn, Lời nói & ngôn ngữ máy tính 6 (3) (1992) 225–242. doi:10.1016/0885-2308(92)90019-Z.

[53] J. Lafferty, A. McCallum, FC Pereira, Các trường ngẫu nhiên có điều kiện: Mô hình xác suất để phân đoạn và gắn nhãn dữ liệu chuỗi, trong: Hội nghị quốc tế lần thứ 18 về Học máy 2001 (ICML 2001), 2001, tr. 282–289.doi: 10.5555/645530.655813.

[54] A. Hulth, Trích xuất từ khóa tự động được cải tiến mang lại nhiều kiến thức ngôn ngữ hơn, trong: Kỷ yếu của hội nghị năm 2003 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên, 2003, trang 216–223.doi:10.3115/1119355.1119383.

[55] SN Kim, O. Medelyan, M.-Y. Kan, T. Baldwin, Semeval-2010, nhiệm vụ 5: Trích xuất cụm từ khóa tự động từ các bài báo khoa học, trong: Kỷ yếu của Hội thảo quốc tế lần thứ 5 về đánh giá ngữ nghĩa, 2010, trang 21–26.

[56] I. Augenstein, M. Das, S. Riedel, L. Vikraman, A. McCallum, Semeval 2017 nhiệm vụ 10: Trích xuất các cụm từ khóa và quan hệ khoa học từ các ấn phẩm khoa học, trong: Hội thảo quốc tế lần thứ 11 về đánh giá ngữ nghĩa (SemEval-2017)), 2017, tr. 546–555.doi:10.18653/v1/S17-2091.

[57] T. Mikolov, K. Chen, G. Corrado, J. Dean, Ước tính hiệu quả các cách biểu diễn từ trong không gian vectơ, arXiv preprint arXiv:1301.3781 (2013).

[58] D. Sahrawat, D. Mahata, M. Kulkarni, H. Zhang, R. Gosangi, A. Stent, A. Sharma, Y. Kumar, RR Shah, R. Zimmermann, Trích xuất cụm từ khóa từ các bài báo học thuật dưới dạng ghi nhãn theo trình tự bằng cách sử dụng các phần nhúng được ngữ cảnh hóa, bản in trước arXiv arXiv:1910.08840 (2019).