TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Ứng dụng mô hình học sâu cho bài toán trích xuất đồng thời thực thể và quan hệ trong văn bản

ĐẶNG DUY ANH
anh.dd183471@sis.hust.edu.vn

Ngành: Khoa học máy tính

Giảng viên hướng dẫn: TS. Nguyễn Thị Kim Anh _____

Chữ kí GVHD

Khoa: Khoa học máy tính

Trường: Công nghệ thông tin và Truyền thông

HÀ NỘI, 08/2023

LỜI CAM KẾT

Họ và tên sinh viên: Đặng Duy Anh
Điện thoại liên lạc: 0986609276
Email: anh.dd183471@sis.hust.edu.vn
Lớp: Khoa học máy tính 03 - K63
Hệ đào tạo: Kỹ sư chính quy

Tôi – *Đặng Duy Anh* – cam kết Đồ án Tốt nghiệp (ĐATN) là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của *PGS.TS. Nguyễn Thị Kim Anh*. Các kết quả nêu trong ĐATN là trung thực, là thành quả của riêng tôi, không sao chép theo bất kỳ công trình nào khác. Tất cả những tham khảo trong ĐATN – bao gồm hình ảnh, bảng biểu, số liệu, và các câu từ trích dẫn – đều được ghi rõ ràng và đầy đủ nguồn gốc trong danh mục tài liệu tham khảo. Tôi xin hoàn toàn chịu trách nhiệm với dù chỉ một sao chép vi phạm quy chế của nhà trường.

Hà Nội, ngày 2 tháng 8 năm 2023

Tác giả ĐATN

Đặng Duy Anh

LỜI CẢM ƠN

Tôi xin chân thành cảm ơn PGS.TS. Nguyễn Thị Kim Anh đã tận tâm hướng dẫn, định hướng và giúp đỡ để tôi hoàn thành đồ án một cách tốt nhất.

Tôi cũng xin gửi lời cảm ơn đến các thầy cô, gia đình, bạn bè, những người đã đồng hành, chia sẻ và giúp đỡ tôi trong quá trình học tập và cuộc sống.

TÓM TẮT NỘI DUNG ĐỒ ÁN

Đồ án tập trung xây dựng một mô hình trích xuất đồng thời thực thể và quan hệ ở trên mức tài liệu, trong đó mô hình cần thực hiện trích xuất các từ đề cập tham chiếu đến thực thể, phân giải đồng tham chiếu, phân loại các thực thể và các quan hệ giữa chúng. Khác với các bài toán trích xuất thông tin trên mức câu dựa vào chú thích các từ tham chiếu ở trong câu, mô hình trong đề tài trích xuất và tìm ra các mối quan hệ ở trên mức thực thể. Để làm được điều này, mô hình sử dụng kiến trúc các tầng mạng Neural Network và mạng đồ thị tích chập Graph Convolution Network để phát hiện ra các thực thể và mối quan hệ của chúng ở trên cả mức nội câu và liên câu. Mô hình được đánh giá ở trên tập dữ liệu công khai DocRED, cho kết quả tốt và được báo cáo để tham chiếu cho tương lai.

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

DANH MỤC THUẬT NGỮ VÀ TỪ VIẾT TẮT

Thuật ngữ	**Ý nghĩa**
FFNN	Feed Forward Neural Network
GCN	Graph Convolution Network
GNN	Graph Neural Network
LSTM	Long Short Term Memory
MLM	Masked Language Model
NLP	Natural Language Processing
NSP	Next Sentence Prediction

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

1.1 Đặt vấn đề

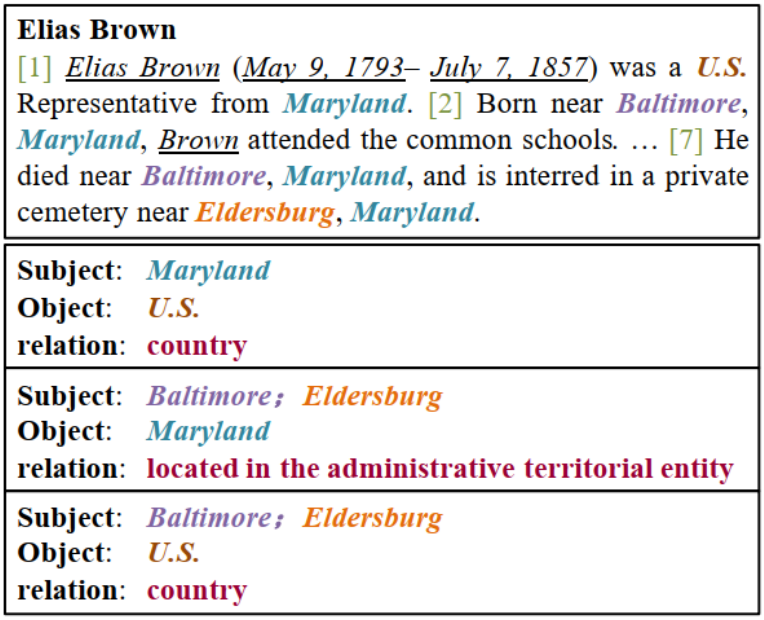
Trích xuất thông tin là quá trình tự động tách lấy thông tin quan trọng và hữu ích từ các nguồn dữ liệu không cấu trúc như văn bản, tài liệu, trang web, email và các nguồn thông tin khác để chuyển đổi thành dữ liệu dạng có cấu trúc để phân tích và sử dụng. Trích xuất thông tin thường bao gồm các bước: nhận dạng thực thể, trích xuất mối quan hệ, trích xuất sự kiện, tóm tắt thông tin, phân loại và gán nhãn. Đề tài sẽ tập trung vào hai bước là nhận dạng thực thể và trích xuất mối quan hệ.

Trích xuất thực thể là quá trình xác định và trích xuất các thực thể có ý nghĩa từ văn bản hoặc dữ liệu không cấu trúc. Các thực thể là các đối tượng hoặc đối tượng có ý nghĩa trong ngữ cảnh như tên riêng, địa điểm, thời gian, ngày tháng, số lượng, sản phẩm, tên công ty, tên tổ chức,.. Trích xuất mối quan hệ là quá trình nhận diện và trích xuất các mối quan hệ giữa các thực thể từ văn bản. Các mối quan hệ này sẽ thể hiện sự liên kết, và tương tác giữa các thực thể. Ứng dụng của trích xuất mối quan hệ là rất đa dạng và quan trọng trong xử lý ngôn ngữ tự nhiên và khai thác dữ liệu. Chúng có thể được sử dụng để xây dựng và khai thác cơ sở tri thức, hệ thống hỏi đáp tự động, trợ lý ảo, phân tích tin tức, ý kiến, thông tin y tế, dự đoán xu hướng,..

Sau khi bộ dữ liệu quy mô lớn DocRED [1] được giới thiệu, bài toán trích xuất mối quan hệ ở mức tài liệu đã nhận được nhiều sự quan tâm hơn. Việc trích xuất này yêu cầu suy diễn giữa các câu văn để xác định các thực thể toàn cục và phân loại các trường hợp mối quan hệ ở mức thực thể, trong đó mỗi thực thể là một nhóm các đề cập đồng tham chiếu trong cả tài liệu. Trong chuỗi nghiên cứu tập trung vào trích xuất mối quan hệ trên thực thể này, các nghiên cứu gần đây đã có đạt được kết quả tốt hơn trên việc suy diễn toàn cục khi dữ liệu đầu vào đã xác định các thực thể [2] [3] [4] [5]. Tuy nhiên, việc trích xuất toàn diện các thực thể toàn cục và mối quan hệ cùng một lúc chưa nhận được nhiều sự chú ý, điều này tạo thêm gánh nặng cho mô hình cần giải quyết các đề cập, đồng tham chiếu và mối quan hệ cùng một lúc. Vậy nên, đồ án sẽ tập trung giải quyết bài toán này, để khi mô hình nhận được một tài liệu, nó có thể xác định được toàn bộ các bộ ba (e_h, e_t, r) , trong đó e_h là chủ thể, e_t là đối tượng, r là quan hệ giữa hai thực thể. Một ví dụ được đánh giá là đúng chỉ khi cặp thực thể đầu cuối e_h, e_t và quan hệ của nó r được xác định đúng.

1.2 Các giải pháp hiện tại và hạn chế

Trích xuất đồng thời thực thể và mối quan hệ vẫn đang là một lĩnh vực nghiên cứu tích cực trong xử lý ngôn ngữ tự nhiên. Các nghiên cứu về lĩnh vực này có thể



Hình 1.1: Một ví dụ trong tập dữ liệu DocRED

được chia thành các nhóm phương pháp sau.

Với nhóm học đa tác vụ, các mô hình học đa tác vụ này được đào tạo một mô hình duy nhất để thực hiện cùng lúc cả nhiệm vụ trích xuất thực thể và mối quan hệ. Phương pháp này nhằm tận dụng các biểu diễn chung và có thể cải thiện hiệu suất khi dữ liệu huấn luyện bị hạn chế. Tuy nhiên, cân bằng đóng góp của mỗi nhiệm vụ và thiết kế hàm mất mát hiệu quả có thể khó khăn. Ngoài ra, việc cải thiện trong một nhiệm vụ không phải lúc nào cũng dẫn đến cùng một mức độ cải thiện trong nhiệm vụ khác, làm cho quá trình tối ưu hóa phức tạp.

Với các mô hình dựa trên các xâu, những mô hình này xác định vị trí các xâu là thực thể đồng thời phân loại các quan hệ của chúng. Nhược điểm của nhóm các mô hình này là chúng có thể gặp khó khăn với các thực thể lồng nhau hoặc các cấu trúc chồng chéo phức tạp.

Các mô hình mạng neural đồ thị (GNN) đã cho thấy tiềm năng trong việc thu thập thông tin về thực thể và mối quan hệ trong một biểu diễn đồ thị thống nhất. Những mô hình này có thể truyền thông tin giữa các thực thể và mối quan hệ, tận dụng thông tin ngữ cảnh hiệu quả. Nhưng huấn luyện các mô hình GNN có thể tốn kém tính toán, đặc biệt là đối với các đồ thị lớn. Ngoài ra, thiết kế các cấu trúc đồ thị phù hợp và định nghĩa cơ chế truyền thông điệp đòi hỏi kiến thức chuyên môn và thử nghiệm.

Các mô hình dựa trên kiến trúc Transformer như BERT, RoBERTa và ELECTRA đã được áp dụng rộng rãi cho các nhiệm vụ trích xuất đồng thời thực thể và mối

quan hệ. Các mô hình này có thể được điều chỉnh lại cho nhiệm vụ cụ thể này bằng cách cung cấp đầu vào dưới dạng thích hợp và điều chỉnh hàm mất mát tương ứng. Hạn chế của các mô hình này thường yêu cầu lượng dữ liệu được gán nhãn lớn để điều chỉnh lại đạt hiệu suất tối ưu. Việc tạo ra các bộ dữ liệu nhãn như vậy có thể tốn kém và tốn thời gian, đặc biệt là đối với các ngôn ngữ có tài nguyên hạn chế hoặc thuộc các lĩnh vực cụ thể.

Nhiều nghiên cứu đã kết hợp tích hợp các thông tin về cấu trúc, hoặc các kĩ thuật đặc thù khác vào các mô hình tiền huấn luyến để cải thiện hiệu quả của mô hình. Tuy nhiện việc tối ưu và điều chỉnh lại các mô hình tiền huấn luyện cho các nhiệm vụ dự đoán có cấu trúc có thể khó khăn.

1.3 Mục tiêu và định hướng giải pháp

Mục tiêu của đồ án là xử lý các tài liệu chứa nhiều câu và trích xuất các đề cập tham chiếu đến thực thể, phân cụm chúng thành các thực thể, và dự đoán ra loại của các thực thể và mối quan hệ ở cấp độ thực thể. Mô hình bao gồm bốn thành phần cụ thể cho từng nhiệm vụ, dựa trên cùng một bộ mã hóa và biểu diễn đề cập, và được huấn luyện theo cách đồng thời. Việc huấn luyện đồng thời không chỉ cải thiện tính đơn giản và hiệu quả, mà còn được thúc đẩy bởi việc nhiều nhiệm vụ có thể có lợi từ nhau: Ví dụ, việc biết loại của hai thực thể (ví dụ: Người, Tổ chức) có thể tăng cường mối quan hệ giữa chúng (ví dụ: CEO của).

Một tài liệu ban đầu sẽ được mã hóa để thu được chuỗi nhúng được bối cảnh hóa của tài liệu. Vì mục tiêu của đồ án là thực hiện trích xuất mối quan hệ từ đầu đến cuối, nên các thực thể cũng như các đề cập tương ứng của chúng trong tài liệu không được biết trong dữ liệu đầu vào.

Đồ án đề xuất mô hình đa mức. Đầu tiên là tầng xác định các đề cập trên toàn bộ tài liệu. Tiếp đó, các đề cập đã được phát hiện sẽ được phân cụm vào các thực thể bằng tầng phân giải đồng tham chiếu. Tầng tiếp theo sẽ tổng hợp các biểu diễn của đề cập về cùng một thực thể và tiến hành phân loại. Cuối cùng, các quan hệ giữa các thực thể sẽ được suy luận dựa trên tầng mạng neural hoặc lan truyền đồ thị.

1.4 Đóng góp của đồ án

Đồ án có đóng góp chính như sau: Đồ án xây dựng mô hình đầu cuối (end to end) trích xuất đồng thời thực thể và quan hệ trong tài liệu. Trong đó với nhiệm vụ trích xuất quan hệ, đồ án đề xuất hai phương pháp là phân loại đa ví dụ và phân loại dựa trên đồ thị.

1.5 Bố cục đồ án

Phần còn lại của báo cáo đồ án tốt nghiệp được sắp xếp như sau.

Chương 2 trình bày về nền tảng lý thuyết của bài toán, bao gồm ngữ cảnh của bài toán cùng các kết quả nghiên cứu tương tự, các nhiệm vụ cần giải quyết trong bài toán trích xuất đồng thời thực thể và quan hệ ở trên mức tài liệu. Đồng thời, chương cũng khái quát các lý thuyết mà đồ án sử dụng trong bài toán như việc mã hóa văn bản với mô hình BERT, phân cụm phân cấp sử dụng liên kết hoàn chỉnh, mạng đồ thị tích chập.

Trong chương 3, đồ án trình bày cụ thể về các giải pháp cho bài toán trích xuất thực thể và quan hệ trong văn bản. Công việc bao gồm: mã hóa văn bản, phát hiện các đề cập, phân giải đồng tham chiếu, phân loại thực thể và cuối cùng là xác định mối quan hệ giữa các thực thể sử dụng hai cách là học đa ví dụ và lan truyền đồ thị.

Chương 4 bao gồm các kết quả về lựa chọn tham số, tối ưu mô hình, đồng thời báo cáo kết quả kiểm thử của mô hình trên tập dữ liệu DocRED[1]. Chương sẽ thống kê, so sánh và nhận xét các kết quả thử nghiệm được.

Cuối cùng, chương 5 tổng kết các kết quả, đóng góp của đồ án và đề ra các hướng phát triển trong tương lai.

CHƯƠNG 2. NỀN TẢNG LÝ THUYẾT

2.1 Ngữ cảnh của bài toán

Trích rút đồng thời thực thể và quan hệ là quá trình tự động nhận diện và trích xuất thông tin từ văn bản liên quan đến các thực thể và quan hệ giữa chúng. Trong ngữ cảnh xử lý ngôn ngữ tự nhiên, bài toán này yêu cầu hệ thống có khả năng tự động xác định các thực thể khác nhau (như người, địa điểm, tổ chức, sản phẩm, v.v.) trong văn bản và nhận biết các quan hệ hoặc mối liên hệ giữa các thực thể đó. Ví dụ, với văn bản đầu vào: "Barack Obama sinh năm 1961 tại Hawaii. Ông là cựu Tổng thống Hoa Kỳ. Michelle Obama là vợ của ông. Cả hai kết hôn năm 1992.", mô hình cần trích rút tập các thực thể: "Barack Obama", "Hawaii", "Tổng thống Hoa Kỳ", "Michelle Obama", và tập các quan hệ: ("Barack Obama", "sinh tại", "Hawaii"), ("Barack Obama", "là", "Tổng thống Hoa Kỳ"), ("Barack Obama", "có vợ là", "Michelle Obama"), ("Barack Obama", "kết hôn năm", "1992"). Trong bài toán này, hệ thống phải hiểu và xử lý ngôn ngữ tự nhiên, nhận biết các từ và cụm từ có ý nghĩa là các thực thể, sau đó xác định mối quan hệ giữa các thực thể đó. Đây là một bài toán phức tạp trong lĩnh vực NLP, và để giải quyết nó, có thể sử dụng các phương pháp học máy và học sâu, cũng như sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên tiên tiến.

Có một số thách thức lớn trong việc trích xuất mỗi quan hệ hiệu quả ở mức tài liệu. Trước hết, trong bài toán trích rút đồng thời thực thể và quan hệ, mô hình cần trích rút các quan hệ có chiều. Vì hai kết quả ("Barack Obama", "có vợ là", "Michelle Obama") và ("Michelle Obama", "có chồng là", "Barack Obama") là khác nhau.

Tiếp đó, mô hình cần tính đến sự tương tác giữa các quan hệ, đặc biệt là quan trọng cho những quan hệ chồng chéo, tức là các quan hệ chia sẻ các đề cập đến thực thể chung. Ví dụ, ("Barack Obama", "là tổng thống", "Hoa Kỳ") có thể suy ra từ ("Barack Obama", "lãnh đạo", "Hoa kỳ"); hai bộ ba này được gọi là trùng lặp cặp thực thể. Trường hợp khác là bộ ba trước cũng có thể suy ra từ ("Barack Obama", "sống ở", "Nhà Trắng") và ("Nhà Trắng", "là dinh thự tổng thống", "Hoa Kỳ"), trong đó hai bộ ba sau được gọi là trùng lặp một thực thể. Những sự tương tác như vậy, bất kể thông qua suy diễn trực tiếp hay gián tiếp, đối với các mô hình nhận dạng cùng thực thể và trích xuất quan hệ là đặc biệt khó khăn, vì các thực thể không được đề cập trong dữ liệu đầu vào.

Hơn nữa, các thực thể chủ thể và đối tượng liên quan đến một quan hệ có thể xuất hiện trong các câu khác nhau. Do đó, một quan hệ không thể được xác định chỉ

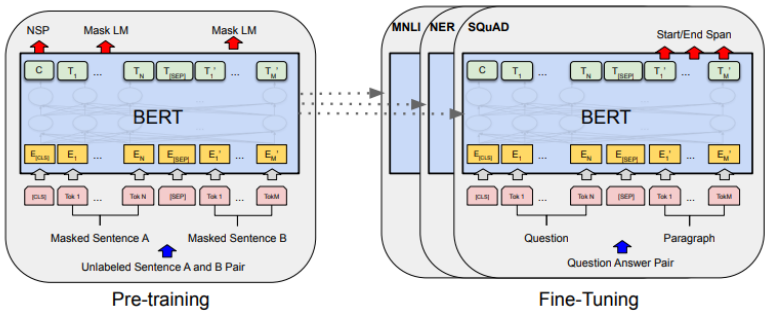
dựa vào một câu duy nhất. Thứ hai, cùng một thực thể có thể được đề cập nhiều lần trong các câu khác nhau. Thông tin ngữ cảnh giữa các câu phải được tổng hợp để biểu diễn thực thể một cách tốt hơn. Thứ ba, việc nhận diện nhiều quan hệ đòi hỏi các kỹ thuật suy luận logic. Điều này có nghĩa là các quan hệ này chỉ có thể được trích xuất thành công khi các thực thể và quan hệ khác, thường xuất hiện ở các câu khác nhau, được nhận dạng một cách ngụ ý hoặc rõ ràng. Như hình 1.1 thể hiện, dễ dàng nhận ra các quan hệ trong cùng một câu ("Maryland", "quốc gia", "Hoa Kỳ"), ("Baltimore", "nằm trong thực thể hành chính", "Maryland"), và ("Eldersburg", "nằm trong thực thể hành chính", "Maryland") vì chủ thể và đối tượng xuất hiện trong cùng một câu. Tuy nhiên, dự đoán các quan hệ giữa Baltimore và Hoa Kỳ, cũng như giữa Eldersburg và Hoa Kỳ là khó khăn, vì các đề cập của chúng không xuất hiện trong cùng một câu và có sự phụ thuộc xa. Ngoài ra, việc nhận dạng hai ví dụ quan hệ này cũng đòi hỏi suy luận logic. Ví dụ, Eldersburg thuộc Hoa Kỳ vì Eldersburg nằm trong Maryland, thuộc về Hoa Kỳ.

2.2 Các kết quả nghiên cứu tương tự

Trích xuất mối quan hệ là một trong những vấn đề xử lý ngôn ngữ tự nhiên (NLP) được nghiên cứu nhiều nhất đến thời điểm hiện tại. Có nhiều phương pháp trích xuất quan hệ của các cặp thực thể trên mức câu như sử dụng mẫu[6], mạng neural[7] hay đồ thị GCN kết hợp với cơ chế chú ý[8]. Các nghiên cứu này đã cho kết quả ấn tượng khi trích xuất quan hệ chỉ trên mức câu.

Tuy nhiên, vì các mối quan hệ phức tạp hơn chỉ có thể được diễn tả bằng nhiều câu, nên cần phải trích xuất quan hệ ở trên mức tài liệu. Tùy thuộc vào cách tiếp cận trong việc xử lý ngữ cảnh, có hai xu hướng chung trong lĩnh vực này. Phương pháp dựa trên đồ thị thường tích hợp ngữ cảnh vào các đồ thị tài liệu dựa trên heuristic và thực hiện lý luận đa bước qua các kỹ thuật neural tiên tiến[2][3]. Phương pháp thứ hai là dựa trên Transformer, với việc tận dụng sức mạnh của các mô hình ngôn ngữ tiền huấn luyện (pre-trained language models) để mã hóa các phụ thuộc ngữ cảnh phạm vi xa[4][5][9]. Tất cả các mô hình trên có điểm chung là giả định rằng các thực thể và đề cập của chúng đã được xác định sẵn. Ngược lại, phương pháp của đồ án trích xuất các đề cập, gom nhóm chúng thành các thực thể và phân loại các mối quan hệ cùng nhau.

Các nghiên cứu trước đó cũng đã đề xuất các mô hình đầu cuối (end to end) trích xuất thực thể và quan hệ trên mức câu. Với phương pháp mã hóa mỗi câu bằng một mạng Bi-LSTM, tác giả sử dụng trạng thái ẩn mã hóa ở tầng cuối để khởi tạo một hoặc nhiều bộ giải mã LSTM để giải mã các bộ ba mối quan hệ[10]. Hay một nghiên cứu khác đã đề xuất phương pháp lan truyền thông tin về thực thể và



Hình 2.1: Mô hình BERT[13]

quan hệ trên một đồ thị từ vựng được học tự động các liên kết bằng cách áp dụng hai pha GCN lên trên bộ mã hóa LSTM-GCN[11].

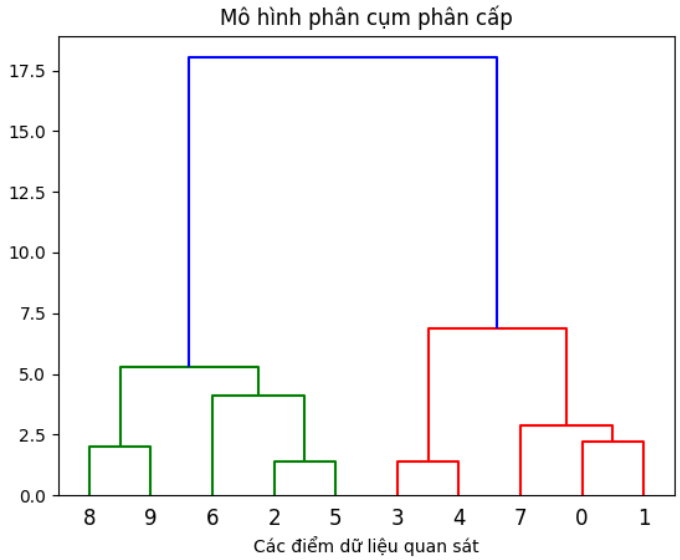
2.3 Mô hình ngôn ngữ BERT

BERT (Bidirectional Encoder Representation from Transformer) là một mô hình biểu diễn từ theo 2 chiều ứng dụng kỹ thuật Transformer. BERT được ứng dụng trong các bài toán NLP để huấn luyện trước các biểu diễn từ (pre-train word embedding). Điểm khác biệt của BERT đó là nó có thể điều hòa cân bằng bối cảnh theo cả 2 chiều trước và sau để thu được một mô hình ngôn ngữ với ngữ nghĩa phong phú hơn. Mô hình Transformer có khả năng đồng thời xử lý toàn bộ các từ trong câu bằng cơ chế chú ý, không phụ thuộc vào thứ tự của từ. Cũng chính đặc điểm này giúp cho mô hình có khả năng học bối cảnh của từ dựa trên cả ngữ cảnh trước và sau nó, bao gồm cả các từ bên trái và bên phải.

Mô hình BERT được cấu thành bởi một kiến trúc đa tầng, bao gồm nhiều lớp Bidirectional Transformer encoder, dựa trên mô tả ban đầu trong bài báo [12]. Văn bản đầu vào được phân chia thành các mã (token) giống như trong mô hình transformer, và mỗi mã sẽ được biến đổi thành một vector tại đầu ra của BERT.

Một mô hình BERT được đào tạo bằng cách sử dụng mô hình ngôn ngữ ẩn (MLM) và dự đoán câu tiếp theo (NSP) đồng thời. Mỗi mẫu huấn luyện cho BERT là một cặp câu từ một tài liệu. Hai câu có thể liên tiếp trong tài liệu hoặc không. Một mã [CLS] sẽ được thêm vào đầu câu đầu tiên (đại diện cho lớp) và một mã [SEP] sẽ được thêm vào cuối mỗi câu (như là dấu phân tách). Sau đó, hai câu sẽ được nối lại với nhau thành một chuỗi các mã để trở thành một mẫu huấn luyện. Một tỷ lệ nhỏ các mã trong mẫu huấn luyện sẽ được che giấu bằng một mã đặc biệt [MASK] hoặc được thay thế bằng một mã ngẫu nhiên.

Trước khi đưa vào mô hình BERT, các mã trong mẫu huấn luyện sẽ được biến đổi thành các vector nhúng, với việc thêm mã hóa vị trí và đặc biệt cho BERT, thêm các vector nhúng phân đoạn để đánh dấu liệu mã đó đến từ câu đầu tiên hay



Hình 2.2: Trực quan phân cụm phân cấp

câu thứ hai.

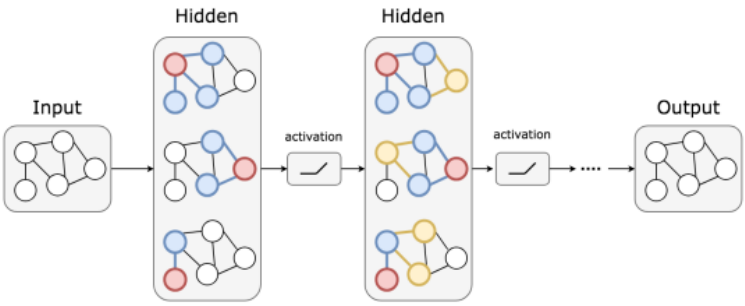
Mỗi mã đầu vào cho mô hình BERT sẽ tạo ra một vector đầu ra. Vector đầu ra tương ứng với mã ẩn có thể tiết lộ mã gốc ban đầu là gì. Vector đầu ra tương ứng với mã [CLS] ở đầu có thể tiết lộ xem hai câu có liên tiếp nhau trong tài liệu hay không. Sau đó, các trọng số được đào tạo trong mô hình BERT có thể hiểu biết rất tốt về ngữ cảnh ngôn ngữ.

Hiện nay, nhiều phiên bản khác nhau của mô hình BERT đã được giới thiệu. Khi thay đổi các tham số của kiến trúc Transformer, bao gồm: L - số lượng các khối trong transformer, H - kích thước của vector nhúng (hay còn gọi là hidden size), A - số lượng đầu ra (head) trong lớp multi-head (mỗi một đầu ra sẽ áp dụng cơ chế chú ý), ta sẽ thu được các phiên bản mô hình BERT khác nhau. Hai kiến trúc BERT phổ biến hiện nay được sử dụng nhiều là $BERT_{BASE}(L = 12, H = 768, A = 12)$ với tổng tham số là 110 triệu và $BERT_{LARGE}(L = 24, H = 1024, A = 16)$ với tổng tham số là 340 triệu.

2.4 Phân cụm phân cấp

Phân cụm là thuật toán làm việc trên dữ liệu không có nhãn, có tác dụng nhóm các điểm dữ liệu có tính chất tương tự vào các cụm khác nhau. Mục tiêu của phân cụm là tìm cách tổ chức dữ liệu sao cho các điểm dữ liệu trong cùng một cụm có tính chất giống nhau hoặc gần giống nhau, trong khi các điểm thuộc các cụm khác nhau có tính chất khác biệt.

Trong mô hình của đề tài sử dụng phân cụm phân cấp. Phân cụm phân cấp là thuật toán phân cụm xây dựng các cụm lồng nhau bằng cách hợp nhất hoặc tách



Hình 2.3: Mạng đồ thị tích chập (GCN)

chúng liên tiếp. Hệ thống phân cấp các cụm này được biểu diễn dưới dạng cây. Gốc của cây là cụm duy nhất tập hợp tất cả các mẫu, lá là cụm chỉ có một mẫu. Thuật toán được thực hiện bằng cách tiếp cận từ dưới lên, mỗi quan sát bắt đầu trong cụm riêng của nó và các cụm được hợp nhất liên tục với nhau. Các tiêu chí liên kết xác định số liệu được sử dụng cho chiến lược hợp nhất gồm có: liên kết tổng - giảm thiểu tổng bình phương sự khác biệt trong tất cả các cụm, liên kết tối đa - giảm thiểu khoảng cách tối đa giữa các quan sát của các cặp cụm, liên kết trung bình - giảm thiểu mức trung bình của khoảng cách giữa tất cả các quan sát của các cặp cụm, và liên kết đơn giảm thiểu khoảng cách giữa các lần quan sát gần nhất của các cặp cụm. Trong đề tài mô hình sử dụng liên kết hoàn chỉnh.

Thuật toán phân cụm có thể được thực hiện với các độ đo khác nhau, như khoảng cách Manhattan (L1), khoảng cách Euclid (L2), khoảng cách cosin hay bất kì hàm khoảng cách nào.

2.5 Mạng đồ thị tích chập (GCN)

Các mạng GCN là các mạng neural hoạt động trực tiếp trên kiến trúc đồ thị[14]. Giống như mạng nơ-ron tích chập (CNN), Mạng đồ thị tích chập (GCN) tiến hành việc tích chập các đặc trưng của các nút lân cận và cũng truyền thông tin của một nút đến các nút lân cận gần nhất. Như được thể hiện trong hình 2.3, bằng cách xếp các lớp GCN lên nhau, mạng GCN có thể trích xuất các đặc trưng khu vực cho mỗi nút.

Một lớp GCN lấy các đặc trưng mới của nút bằng cách xem xét đặc trưng của các nút lân cận bằng phương trình sau đây:

$$h_u^{l+1} = \sigma\left(\sum_{v \in N(u)} \left(W^l h_v^l + b^l\right)\right)$$

trong đó u là nút mục tiêu và $N(u)$ đại diện cho các lân cận của u, bao gồm cả

chính nút u, h_v^l biểu diễn đặc trưng ẩn của nút v tại tầng l; W và b là các trọng số có thể học được, để ánh xạ đặc trưng của một nút sang các nút lân cận trong đồ thị, và $h \in \mathbb{R}^f$,$W \in \mathbb{R}^{f \times f}$, và $b \in \mathbb{R}^f$, trong đó f là kích thước đặc trưng, σ là một hàm kích hoạt. Bằng cách lan truyền thông tin, mỗi tầng GCN sẽ học được các đặc trưng ẩn của dữ liệu đồ thị ban đầu.

Kết chương: Như vậy, chương đã trình bày các nhiệm vụ cần phải thực hiện và kết quả đầu ra cần đạt được trong bài toán trích xuất đồng thời thực thể và quan hệ. Bài toán này có những thách thức riêng như phải nhận biết các quan hệ hai chiều, các quan hệ chồng chéo, và các quan hệ phải suy luận trên nhiều câu. Chương cũng đề cập đến các nghiên cứu tương tự và cách các nghiên cứu này giải quyết các khía cạnh khác nhau của bài toán trích rút thực thể và quan hệ. Và cuối cùng, chương trình bày một số các lý thuyết được sử dụng để xây dựng mô hình như mô hình ngôn ngữ BERT, thuật toán phân cụm phân cấp, mạng đồ thị tích chập. Chi tiết cách ứng dụng các lý thuyết này và phương pháp cụ thể xây dựng mô hình sẽ được trình bày ở chương 3.

CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT

3.1 Kiến trúc mô hình

Mô hình xử lý các tài liệu chứa nhiều câu và trích xuất các đề cập đến thực thể, phân cụm chúng thành các thực thể, từ đó dự đoán ra loại các thực thể và mối quan hệ giữa chúng. Mô hình bao gồm năm thành phần cho từng nhiệm vụ, thực hiện các nhiệm vụ dựa trên cùng một bộ mã hóa và biểu diễn trên mức đề cập, và được huấn luyện cùng một lúc. Tầng đầu tiên là tầng mã hóa, tài liệu ban đầu được mã hóa bằng mô hình ngôn ngữ BERT[13], để thu được một chuỗi nhúng theo ngữ cảnh của văn bản $(e_1, e_2, ...e_n)$. Tầng thứ hai là tầng phát hiện các đề cập, là vị trí các từ tham chiếu đến các thực thể. Tầng thứ ba là tầng phân giải đồng tham chiếu để phân cụm các thực thể. Tầng tiếp thứ tư là tầng nhận diện và phân loại các thực thể. Và cuối cùng tầng năm là tầng phân loại các quan hệ được suy diễn dựa trên các đề cập đã được phát hiện. Kiến trúc mô hình được minh họa như hình 3.1.

3.2 Xác định các đề cập

Mô hình thực hiện tìm kiếm trên tất cả các cụm từ của tài liệu với độ dài các cụm từ không vượt quá một siêu tham số L. Cách tiếp cận này cho phép phát hiện các đề cập chồng chéo. Gọi $s := (e_i, e_{i+1}, ..., e_{i+k})$ là biểu diễn của các đề cập ứng cử viên. Ban đầu, mô hình thu được biểu diễn của các đề cập bằng cách sử dụng hàm max-pooling để tổng hợp các véc-tơ nhúng của các từ đề cập tương ứng.

$$\mathbf{e}(s) = \text{max-pooling}\left(\mathbf{e}_i, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_{i+k}\right) \tag{3.1}$$

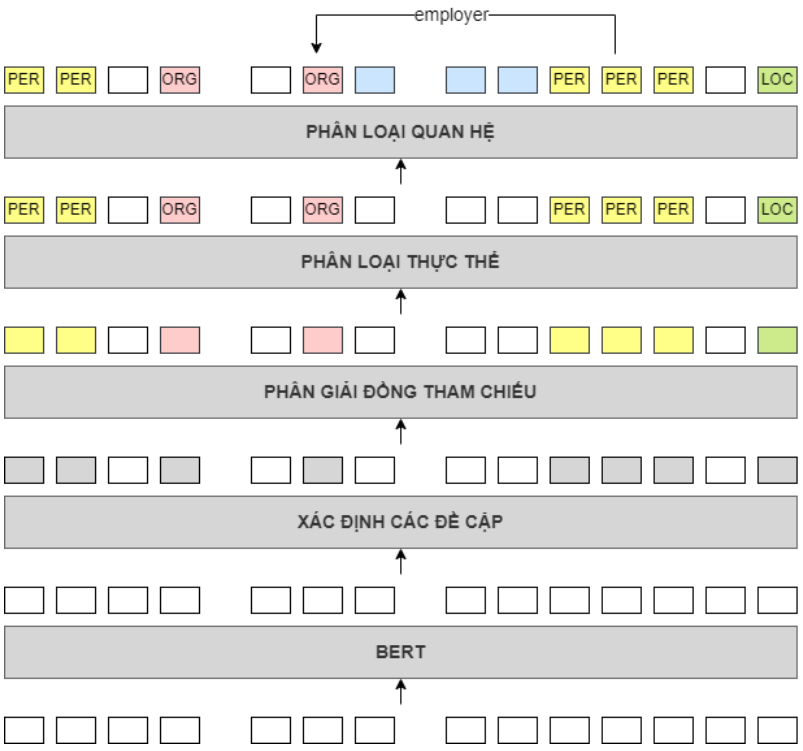
Bộ phân loại đề cập sẽ lấy các biểu diễn đề cập $e(s)$ để thực hiện phân loại nhị phân và sử dụng hàm kích hoạt sigmoid để thu được xác suất biểu diễn s là một đề cập thực thể:

$$\hat{y}^s = \sigma\left(\text{FFNN}^s(\mathbf{e}(s)\right) \tag{3.2}$$

$FFNN^s$ là một mạng lan truyền tiến hai tầng và được kích hoạt bởi hàm ReLu. Mô hình sử dụng ngưỡng lọc α^s trên điểm tin cậy thu được, và sẽ giữ lại tất cả các đề cập có giá trị $\hat{y}^s \geq \alpha^s$ và sinh ra tập hợp S chứa các đề cập mà được dự đoán chính là các đề cập tham chiếu đến các thực thể.

3.3 Phân giải đồng tham chiếu

Các đề cập cùng tham chiếu đến một thực thể (ví dụ các đề cập “Elizabeth II.” và "the Queen") có thể cùng xuất hiện rải rác ở trên văn bản đầu vào. Để trích xuất



Hình 3.1: Kiến trúc mô hình

các quan hệ trên mức thực thể, các đề cập như này cần được nhóm vào các cụm thực thể cấp tài liệu bằng tầng phân giải đồng tham chiếu. Mô hình phân loại các cặp đề cập $(s_1, s_2) \in S$ đã được phát hiện trước đó xem chúng có phải là đồng tham chiếu hay không, bằng cách nối hai biểu diễn đề cập $e(s_1)$ và $e(s_2)$. Từ đó một biểu diễn cặp đề cập x^c được hình thành bằng phép nối:

$$\mathbf{x}^c = \mathbf{e}\left(s_1\right) \circ \mathbf{e}\left(s_2\right) \tag{3.3}$$

Tương tự như với phân loại đề cập, biểu diễn cặp đề cập được đưa qua một bộ phân lớp nhị phân sử dụng hàm kích hoạt sigmoid, để thu được điểm tương đồng giữa hai đề cập:

$$\hat{y}^c = \sigma\left(\mathrm{FFNN}^c\left(\mathbf{x}^c\right)\right) \tag{3.4}$$

với $FFNN^c$ có kiến trúc tương tự như $FFNN^s$. Từ các điểm đồng tham chiếu này, mô hình tạo thành ma trận độ tương đồng $C \in \mathbb{R}^{m \times m}$ (với m là tổng tất cả các đề cập xuất hiện trong tài liệu), ma trận sẽ chứa tất cả các điểm tương đồng của từng cặp đề cập. Sau khi áp dụng một ngưỡng lọc α^c, mô hình phân cụm các đề cập sử dụng phân cụm phân cấp đã trình bày ở trên với kiểu liên kết hoàn chỉnh, để sinh ra tập $\mathcal{E}$ chứa các cụm là các cụm thực thể.

3.4 Phân loại các thực thể

Tầng này phân loại thực thể thành một loại, ví dụ như thực thể đó là loại địa điểm hay người. Ban đầu tầng thu được biểu diễn của thực thể x^e bằng cách tổng hợp biểu diễn của các đề cập thuộc về cụm thực thể đó $\{s_1, s_2, \ldots, s_t\} \in \mathcal{E}$ bằng hàm max-pooling. Bằng cách biểu diễn như vậy, mô hình có thể tổng hợp được thông tin từ các đề cập xuất hiện ở những chỗ khác nhau trên tài liệu.

$$\mathbf{x}^e = \text{max-pooling}(\mathbf{e}(s_1), \mathbf{e}(s_2), .., \mathbf{e}(s_t)) \tag{3.5}$$

Sau đó, phân loại thực thể được thực hiện dựa trên biểu diễn thực thể x^e vừa thu được. x^e được truyền vào bộ phân loại softmax, để thu được phân bố xác suất trên các loại thực thể:

$$\hat{y}^e = \text{softmax}(\text{FFNN}^e(\mathbf{x}^e)) \tag{3.6}$$

Từ đó, loại thực thể có xác suất cao nhất sẽ được gán cho thực thể.

3.5 Phân loại các quan hệ

Tầng cuối cùng gán các loại mối quan hệ cho các cặp thực thể. Các quan hệ này có hướng, tức là có phân biệt thực thể nào cấu thành đầu, đuôi của mối quan hệ, và tài liệu đầu vào có thể diễn đạt nhiều mối quan hệ giữa các đề cập khác nhau của cùng một cặp thực thể. Gọi R là tập hợp các loại mối quan hệ có trong tập huấn luyện. Tầng phân loại mối quan hệ xử lý từng cặp thực thể $(s_1, s_2) \in \mathcal{E} \times \mathcal{E}$, dự đoán xem có mối quan hệ nào từ R được diễn đạt giữa các thực thể này. Để làm được điều này, mô hình tính điểm cho mỗi bộ ba (e_1, r_i, e_2), trong đó e_1 là điểm đầu của mỗi quan hệ r_i với e_2 là điểm đuôi.

Tầng có hai mô hình phân loại khác nhau, một là mô hình phân loại đa ví dụ, hai là mô hình phân loại dựa trên đồ thị.

3.6 Mô hình phân loại quan hệ đa ví dụ

Mô hình phân loại đa ví dụ hoạt động dựa trên mức độ đề cập. Mô hình coi các cặp đề cập như các biến tương quan và dự đoán các quan hệ bằng cách tổng hợp thông tin trên các cặp đề cập này. Với mỗi cặp cụm thực thể $e_1 = \left\{s_1^1, s_2^1, ..., s_{t_1}^1\right\}$ và $e_2 = \left\{s_1^2, s_2^2, ..., s_{t_2}^2\right\}$, mô hình thực hiện biểu diễn từng cặp đề cập cho mọi $(s_1, s_2) \in e_1 \times e_2$. Biểu diễn này thu được bằng cách nối các véc tơ nhúng của thực thể x^e (biểu thức 3.5) cùng với các vector biểu diễn đề cập $e(s)$ (biểu thức 3.1)

$$u(s_1, s_2) = (e(s_1) \circ x_1^e) \circ (e(s_2) \circ x_2^e) \tag{3.7}$$

Sau đó, véc tơ ngữ cảnh giữa hai đề cập $c(s_1, s_2)$ được nối vào để thu được biểu diễn có ngữ cảnh. Véc tơ ngữ cảnh $c(s_1, s_2)$ được tính bằng hàm max-pooling của các biểu diễn từ nằm giữa hai đề cập s_1 và s_2. Giả sử tập $\{e_1, e_2, \ldots, e_t\}$ là các biểu diễn từ nằm giữa s_1 và s_2, véc tơ $c(s_1, s_2)$ được tính bằng công thức:

$$c(s_1, s_2) = \text{max-pooling}(e_1, e_2, \ldots, e_t) \tag{3.8}$$

Véc tơ ngữ cảnh này cung cấp một cái nhìn tập trung hơn về tài liệu và đặc biệt có lợi cho các đầu vào dài, cũng là những đầu vào nhiễu. Biểu diễn cuối cùng của cặp đề cập sau khi được nối véc tơ ngữ cảnh như sau:

$$u'(s_1, s_2) = u(s_1, s_2) \circ c(s_1, s_2) \tag{3.9}$$

Biểu diễn cặp đề cập này sau đó được đưa vào tầng lan truyền tiến để thu được véc tơ nhúng có số chiều bằng véc tơ nhúng ban đầu (768):

$$u''(s_1, s_2) = \text{FFNN}^p(u'(s_1, s_2)) \tag{3.10}$$

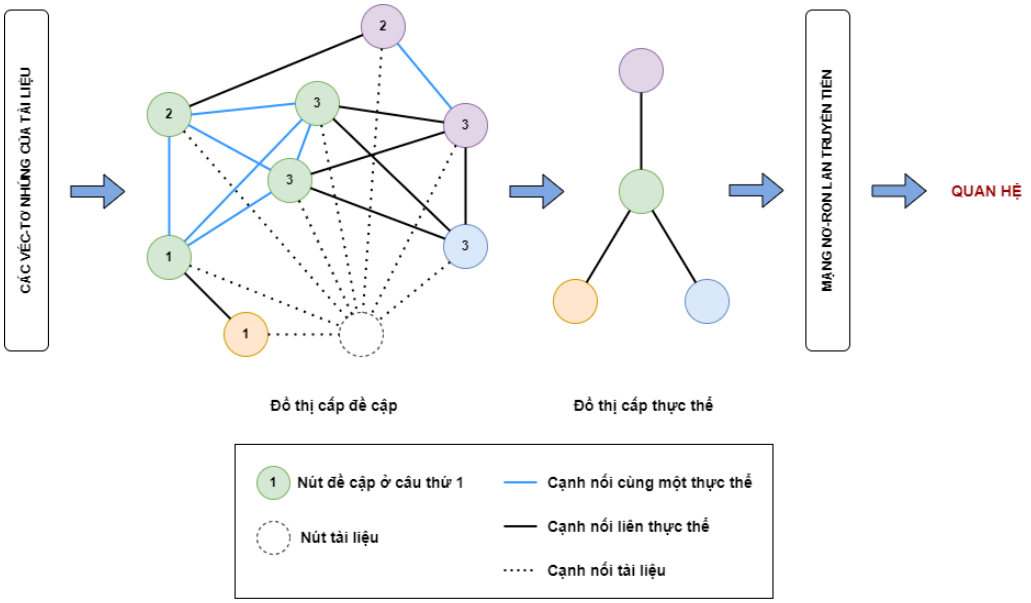
Tiếp theo, từng cặp biểu diễn đề cập của hai cụm thực thể được tổng hợp bằng hàm max-pooling:

$$x^r = \text{max-pooling}(\{u''(s_1, s_2) | s_1 \in e_1, s_2 \in e_2\}) \tag{3.11}$$

Sau đó, x^r được nối với véc tơ nhúng kiểu của thực thể w_1^e, w_2^e và được đưa qua một tầng FFNN 2 tầng. Các véc tơ nhúng này được xây dựng bằng một bảng tra cứu và sẽ trả ra véc tơ nhúng với chỉ số tương ứng. Như trong tập dữ liệu có tất cả 6 loại thực thể, thì một bảng sẽ được tạo ra với 6 phần tử. Khi truyền vào chỉ số của loại thực thể, bảng sẽ trả ra véc tơ nhúng với loại thực thể đó.

$$x^p = x^r \circ w_1^e \circ w_2^e \tag{3.12}$$

Và cuối cùng mô hình sử dụng hàm kích hoạt sigmoid cho phân loại đa nhãn và gán bất kỳ loại mối quan hệ nào vượt qua ngưỡng α^r. Mô hình sẽ dự đoán kết quả quan hệ trên cả hai chiều (s_1, r_i, s_2) và (s_2, r_i, s_1) để suy ra hướng của các quan hệ



Hình 3.2: Kiến trúc đồ thị

bất đối xứng.

$$\hat{y}^r = \sigma\left(\text{FFNN}^p(x^p)\right) \tag{3.13}$$

3.7 Mô hình phân loại quan hệ dựa trên đồ thị

Để trích xuất được thông tin cấp tài liệu và tương tác giữa các đề cập và thực thể, mô hình xây dựng hai đồ thị liên tiếp bao gồm một đồ thị cấp đề cập (Mention-level Graph) và sau đó là một đồ thị cấp thực thể (Entity-level Graph). Chi tiết kiến trúc đồ thị như hình 3.2.

Đồ thị cấp đề cập có hai loại nút khác nhau: nút các đề cập và nút tài liệu. Mỗi nút đề cập đại diện cho một đề cập cụ thể của một thực thể. Đồ thị còn có một nút tài liệu nhằm mô hình thông tin tổng quan về tài liệu. Nút này có thể hoạt động như một điểm trung gian để tương tác với các đề cập khác nhau và do đó giảm khoảng cách xa giữa chúng trong tài liệu.

Có ba loại cạnh trong đồ thị cấp đề cập:

- Cạnh nối cùng một thực thể: Các đề cập tham chiếu đến cùng một thực thể sẽ được nối với nhau bằng cạnh này. Cạnh này giúp biểu diễn thông tin sự tương tác giữa các đề cập của cùng một thực thể.
- Cạnh nối liên thực thể: Hai đề cập của hai thực thể khác nhau nhưng lại cùng xuất hiện trong một câu thì sẽ được nối bởi cạnh này. Bằng cách này, sự tương tác giữa các thực thể có thể được mô hình hóa bởi các đề cập của chúng.
- Cạnh nối tài liệu: Tất cả các đề cập sẽ được nối vào một nút chung là nút tài

liệu bằng cạnh nối tài liệu. Với cách kết nối như vậy, nút tài liệu có thể chú ý đến tất cả các đề cập và tạo ra tương tác giữa tài liệu và các đề cập. Ngoài ra, khoảng cách giữa hai nút đề cập tối đa là hai với nút tài liệu là trung gian. Do đó, mô hình có thể biểu diễn các phụ thuộc có khoảng cách xa tốt hơn.

Tiếp đó, mô hình áp dụng mạng đồ thị tích chập [14] ở trên đồ thị cấp đề cập vừa xây dựng để tổng hợp thông tin giữa các nút lân cận. Dựa vào nút u ở tầng thứ l, lan truyền đồ thị có thể được thực hiện như sau:

$$h_{m,u}^{(l+1)} = \sigma(\sum_{k \in K} \sum_{v \in N_k(u)} W_{m,k}^{(l)} h_{m,v}^{(l)} + b_{m,k}^{(l)}) \tag{3.14}$$

trong đó K là tập ba loại cạnh đã được đề cập ở trên, $W_{m,k}^{(l)} \in \mathbb{R}^{d \times d}$ và $b_{k,m}^{(l)} \in \mathbb{R}^d$ là các tham số có thể huấn luyện. $N_k(u)$ đại diện cho các hàng xóm của nút u được kết nối bởi loại cạnh thứ k. σ là hàm kích hoạt (ReLU).

Vì mỗi tầng trong GCN biểu diễn các đặc trưng ở các mức trừu tượng khác nhau, nên do đó sẽ có ý nghĩa các cấp độ khác nhau, mô hình đề xuất ghép nối các trạng thái ẩn của các tầng khác nhau để thu được biểu diễn cuối cùng của nút u:

$$m_u = h_{m,u}^{(0)} \circ h_{m,u}^{(1)} \circ ... \circ h_{m,u}^{(N)} \tag{3.15}$$

Trong đó $h_{m,u}^{(0)}$ là biểu diễn ban đầu của nút u và $h_{m,u}^{(N)}$ là biểu diễn thu được của nút u tại các tầng thứ n. Giá trị biểu diễn của nút tại tầng ban đầu là véc tơ biểu diễn đề cập ở công thức 3.1. Đối với nút tài liệu, ban đầu nó được khởi tạo bằng trung bình tất cả các véc tơ nhúng của các từ trong tài liệu.

Sau khi xây dựng đồ thị cấp độ đề cập, mô hình xây dựng đồ thị cấp thực thể. Đầu tiên, các nút đề cập tham chiếu đến cùng một thực thể được hợp nhất thành nút thực thể để thu được các nút trong đồ thị thực thể. Với một nút thực thể e_u^0 được đề cập đến N lần, nó sẽ được biểu diễn bằng trung bình của N các biểu diễn mức đề cập:

$$e_u^0 = \frac{1}{N} \sum_n m_n \tag{3.16}$$

Sau đó, mô hình gộp các cạnh liên thực thể mà kết nối các thực thể trong cùng một câu như đề cập ở trên để thu được các cạnh trong đồ thị cấp thực thể. Mô hình tiếp tục sử dụng mạng đồ thị tích chập (GCN) trên đồ thị thực thể để thu được các đặc trưng của nút thực thể. Biểu diễn của nút u tại tầng $l+1$ tại đồ thị thực thể như

sau:

$$h_{e,u}^{(l+1)} = \sigma(\sum_{v \in N(u)} W_e^{(l)} h_{e,v}^{(l)} + b_e^{(l)}) \tag{3.17}$$

trong đó $W_e^{(l)} \in \mathbb{R}^{d \times d}$ và $b_e^{(l)} \in \mathbb{R}^d$ là các tham số có thể huấn luyện. $N(u)$ đại diện cho các hàng xóm là các thực thể có kết nối đến nút u. σ là hàm kích hoạt (ReLU).

Sau đó biểu diễn cuối cùng của nút e cũng thu được bằng cách nối các biểu diễn ẩn của mỗi tầng trong đồ thị:

$$e_u = h_{e,u}^{(0)} \circ h_{e,u}^{(1)} \circ ... \circ h_{e,u}^{(N)} \tag{3.18}$$

trong đó $h_{e,u}^{(0)}$ là biểu diễn ban đầu của nút u và có giá trị bằng e_u^0.

Để dự đoán cho từng quan hệ giữa cặp các nút thực thể trong đồ thị, tương tự như bộ phân loại ở trên, mô hình xây dựng biểu diễn cho các cặp thực thể (e_h, e_t) theo biểu thức:

$$x^p = (e_h \circ w_1^e) \circ (e_t \circ w_2^e) \circ m_{doc} \tag{3.19}$$

trong đó (e_h, e_t) là các nút biểu diễn thực thể đầu-cuối trong đồ thị, w_1^e, w_2^e là các véc tơ nhúng của kiểu thực thể. Và véc tơ m_{doc} giúp tổng hợp thông tin giữa các câu và cung cấp biểu diễn liên quan đến toàn bộ tài liệu. Và cuối cùng, mô hình có thể xác định quan hệ giữa các thực thể bằng bộ phân loại đa lớp (đa quan hệ) tương tự như kiến trúc mạng FFNN 2 tầng như biểu thức 3.13.

$$\hat{y}^r = \sigma\left(\text{FFNN}^p(x^p)\right) \tag{3.20}$$

3.8 Hàm đánh giá lỗi

Mô hình thực hiện việc huấn luyện đa nhiệm vụ có giám sát, trong đó mỗi tài liệu huấn luyện chứa các kết quả chân lý cho tất cả bốn nhiệm vụ con: xác định vị trí đề cập, phân giải đồng tham chiếu, cũng như phân loại thực thể và mối quan hệ. Mô hình tối ưu lỗi tổng hợp của cả bốn thành phần:

$$\mathcal{L} = \beta_s \cdot \mathcal{L}^s + \beta_c \cdot \mathcal{L}^c + \beta_e \cdot \mathcal{L}^e + \beta_r \cdot \mathcal{L}^r \tag{3.21}$$

trong đó $\mathcal{L}^s, \mathcal{L}^c, \mathcal{L}^r$ thể hiện các hàm lỗi nhị phân cross-entropy của việc phân

loại các đề cập, đồng tham chiếu và quan hệ. Mô hình sử dụng hàm lỗi cross-entropy $\mathcal{L}^e$ cho phân loại thực thể.

Kết chương: Chương 3 đã trình bày kiến trúc mô hình cũng như chi tiết các kĩ thuật được thực hiện trong từng tầng của kiến trúc, các tầng này tương ứng với từng nhiệm vụ: xác định các đề cập, phân giải đồng tham chiếu, phân loại thực thể và phân loại quan hệ. Đặc biệt, với tầng phân loại quan hệ mô hình có sử dụng hai phương pháp khác nhau đó là phân loại bằng bộ phân lớp đa ví dụ và phân loại bằng xây dựng đồ thị. Hai phương pháp này sẽ được đánh giá và so sánh ở chương 4.

CHƯƠNG 4. ĐÁNH GIÁ THỰC NGHIỆM

4.1 Tập dữ liệu DocRED

Tập dữ liệu mô hình sử dụng là tập dữ liệu công khai DocRED[1]. DocRED (Document-Level Relation Extraction Dataset) là một bộ dữ liệu trích xuất mối quan hệ được xây dựng từ Wikipedia và Wikidata. Mỗi tài liệu trong bộ dữ liệu được con người đánh dấu với các đề cập đến thực thể có tên, thông tin về đồng tham chiếu, các mối quan hệ nội câu và ngoại câu, cũng như bằng chứng hỗ trợ cho mỗi mối quan hệ. Để trích xuất các thực thể và suy luận về các mối quan hệ của chúng, DocRED yêu cầu đọc nhiều câu trong một tài liệu và tổng hợp tất cả thông tin của tài liệu đó.

DocRED bao gồm 132,375 thực thể và 56,354 các mối quan hệ được gán nhãn trên 5,053 tài liệu.

DocRED bao gồm 6 loại thực thể khác nhau, bao gồm người (18,5%), địa điểm (30,9%), tổ chức (14,4%), thời gian (15,8%) và số (5,1%). Nó cũng bao gồm một tập hợp đa dạng các tên thực thể không thuộc loại đã nêu trên (15,2%), chẳng hạn như các sự kiện, tác phẩm nghệ thuật và luật pháp. Mỗi thực thể được chú thích trung bình với 1,34 đề cập.

Có 96 loại mối quan hệ thuộc một loạt các danh mục rộng, bao gồm các mối quan hệ liên quan đến khoa học (33,3%), nghệ thuật (11,5%), thời gian (8,3%), cuộc sống cá nhân (4,2%), v.v. Điều này có nghĩa là các mối quan hệ không bị ràng buộc trong bất kỳ lĩnh vực cụ thể nào. Hơn nữa, các loại mối quan hệ được tổ chức theo một hệ thống phân cấp và phân loại rõ ràng, có thể cung cấp thông tin phong phú cho các hệ thống trích xuất mối quan hệ cấp tài liệu. Chi tiết các mối quan hệ xem trong bảng A.1.

4.2 Lấy mẫu dữ liệu

- Xác định vị trí các đề cập: Mô hình tận dụng tất cả các đề cập có giá trị chân lý đúng ở trong tài liệu để làm nhãn dương, và lấy mẫu một số lượng ngẫu nhiên N_s các đề cập độ dài không quá L_s làm các nhãn âm.
- Phân giải đồng tham chiếu: Mô hình cũng lấy tất cả các cụm thực thể có giá trị chân lý đúng ở trong tài liệu để làm nhãn dương, và lấy mẫu một số lượng ngẫu nhiên N_c các cặp đề cập không thuộc về cùng một cụm làm các nhãn âm.
- Phân loại các thực thể: Vì phân loại các thực thể chỉ được thực hiện ở trên các cụm mà chắc chắn sẽ thuộc về một loại thực thể nào đó, nên bộ phân loại này được huấn luyện ở trên tất cả các cụm có giá trị chân lý đúng ở trong tài liệu.

Hình 4.1: Kết quả đánh giá mô hình phân lớp đa ví dụ trên tập tối ưu

- Phân loại các quan hệ: Mô hình sử dụng các quan hệ có giá trị chân lý đúng giữa các cụm thực thể làm nhãn dương và lấy ra N_r mẫu từ các cụm thực thể mà không có quan hệ với nhau làm nhãn âm.

4.3 Chia tập dữ liệu

Từ tập dữ liệu DocRED ban đầu, đồ án thực hiện chia ngẫu nhiên dữ liệu thành 3 tập với tập huấn luyện (2991 tài liệu), tập tối ưu (300 tài liệu), và tập kiểm thử (700 tài liệu).

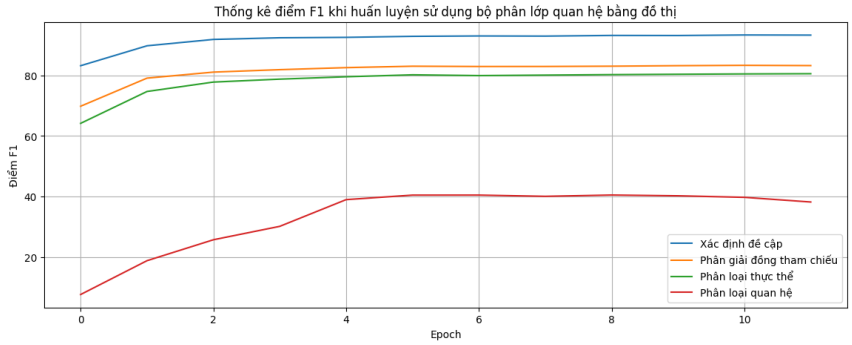
4.4 Cài đặt các tham số

Mô hình sử dụng mô hình ngôn ngữ $\text{BERT}_{\text{BASE}}$ (cased) để mã hóa tài liệu[13]. Mô hình huấn luyện với giải thuật tối ưu Adam và với tốc độ học là 0.001, áp dụng tỉ lệ dropout là 0.2. Đối với đồ thị GCN phân loại quan hệ, mô hình sử dụng một đồ thị cấp đề cập 2 tầng và một đồ thị cấp thực thể 2 tầng, các đồ thị này cũng sử dụng tỉ lệ dropout là 0.2. Kích thước của véc tơ nhúng kiểu thực thể w^e là 25. Đề tài chọn được các ngưỡng để phân loại nhãn: $\alpha^s = 0.85$ (ngưỡng phân loại đề cập), $\alpha^c = 0.85$ (ngưỡng phân cụm), $\alpha^r = 0.6$ (ngưỡng phân loại quan hệ). Số lượng các mẫu ấm được lấy ngẫu nhiên là $N_s = N_c = N_r = 200$ và trọng số các hàm lỗi của từng nhiệm vụ là ($\beta_s = \beta_c = \beta_r = 1, \beta_e = 0.25$). Mô hình sử dụng kích thước nhóm ví dụ huấn luyện (batch size) bằng 1 (tương ứng với mỗi tài liệu).

4.5 Kết quả kiểm thử

Để thu được mô hình tốt nhất, đồ án tiến hành đánh giá mô hình trên tập tối ưu và sẽ dừng huấn luyện nếu sau một vài vòng lặp kết quả của mô hình không tốt lên. Sau đó, mô hình tốt nhất trên tập tối ưu sẽ được đánh giá ở trên tập kiểm thử để có được nhận xét cuối cùng.

Trong quá trình khảo sát, đồ án ghi nhận với mô hình sử dụng bộ phân lớp đa ví dụ, mô hình hội tụ sau khoảng 20 epoch (chi tiết ở hình 4.1), còn với mô hình sử dụng phân lớp quan hệ bằng đồ thị, mô hình đạt được trạng thái tốt nhất sau



Hình 4.2: Kết quả đánh giá mô hình phân lớp quan hệ sử dụng đồ thị trên tập tối ưu

	Phân lớp đa ví dụ			**Phân lớp sử dụng đồ thị**		
Nhiệm vụ	Precision	Recall	F1	Precision	Recall	F1
Xác định các đề cập	93.65	92.66	93.15	92.74	92.49	92.61
Phân giải đồng tham chiếu	82.71	83.32	83.02	81.52	85.13	82.32
Phân loại các thực thể	80.17	80.75	80.46	78.83	80.38	79.60
Phân loại các quan hệ	43.60	38.76	41.04	41.61	37.73	39.58

Bảng 4.1: Kết quả đánh giá hai mô hình trên tập kiểm thử

khoảng 11 epoch (chi tiết ở hình 4.2).

Kết quả đánh giá hai mô hình trên tập kiểm thử thu được như bảng 4.1. Nhận thấy, mô hình phân lớp đa ví dụ cho kết quả tốt hơn mô hình phân lớp sử dụng đồ thị.

Kết chương: Chương đã trình bày thông tin chi tiết về tập dữ liệu được sử dụng để huấn luyện và đánh giá mô hình. Ngoài ra, chương cũng báo cáo cách chia và lấy mẫu từ tập dữ liệu để huấn luyện, kiểm thử, các siêu tham số của mô hình đã được cài đặt. Cuối cùng, chương 4 báo cáo các kết quả đánh giá và đưa ra so sánh hai mô hình trích xuất thực thể và quan hệ, đó là mô hình dựa trên phân lớp đa ví dụ và mô hình phân lớp sử dụng đồ thị.

CHƯƠNG 5. KẾT LUẬN

5.1 Kết luận

Trích rút đồng thời thực thể và quan hệ trong tài liệu là một bài toán thách thức và có đa dạng hướng tiếp cận khác nhau. Đồ án đã giới thiệu mô hình để trích xuất đồng thời các thực thể cùng với quan hệ của chúng bao gồm các thành phần: xác định các đề cập, phân giải đồng tham chiếu, phân loại các thực thể, và đặc biệt là phân loại các quan hệ với hai phương pháp là sử dung bộ phân lớp đa ví dụ và bộ phân lớp dựa trên đồ thị. Đồ án cũng đã so sánh, đánh giá kết quả mô hình xây dựng được ở trên một tập dữ liệu công khai DocRED.

5.2 Hướng phát triển trong tương lai

Trong tương lai, đồ án vẫn có thể phát triển thêm bằng cách thử các phương pháp tiếp cận khác để cải thiện hiệu suất của mô hình. Ví dụ như lớp các phương pháp sử dụng kiến trúc Transformer và các mô hình tiền huấn luyện tiên tiến hiện nay cho các bài toán trích rút đồng thời thực thể và quan hệ nói riêng hay trích rút thông tin nói chung trong văn bản. Ngoài ra, do đặc thù trong tập dữ liệu DocRED không gán nhãn các đại từ quan hệ, nên việc suy diễn các tài liệu xuất hiện nhiều đại từ quan hệ sẽ cho kết quả chưa tốt, đây cũng là hướng nên nghiên cứu giải quyết và cải tiến sau này.

TÀI LIỆU THAM KHẢO

[1] Y. Yao, D. Ye, P. Li **andothers**, "Docred: A large-scale document-level relation extraction dataset," *arXiv preprint arXiv:1906.06127*, 2019.

[2] B. Li, W. Ye, Z. Sheng, R. Xie, X. Xi **and** S. Zhang, "Graph enhanced dual attention network for document-level relation extraction," **in***Proceedings of the 28th international conference on computational linguistics* 2020, **pages** 1551–1560.

[3] S. Zeng, Y. Wu **and** B. Chang, "Sire: Separate intra-and inter-sentential reasoning for document-level relation extraction," 2021.

[4] B. Xu, Q. Wang, Y. Lyu, Y. Zhu **and** Z. Mao, "Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction," **in***Proceedings of the AAAI conference on artificial intelligence* **volume** 35, 2021, **pages** 14 149–14 157.

[5] W. Xu, K. Chen **and** T. Zhao, "Discriminative reasoning for document-level relation extraction," *arXiv preprint arXiv:2106.01562*, 2021.

[6] W. Zhou, H. Lin, B. Y. Lin **andothers**, "Nero: A neural rule grounding framework for label-efficient relation extraction," **in***Proceedings of The Web Conference 2020* 2020, **pages** 2166–2176.

[7] R. Cai, X. Zhang **and** H. Wang, "Bidirectional recurrent convolutional neural network for relation classification," **in***Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2016, **pages** 756–765.

[8] Z. Guo, Y. Zhang **and** W. Lu, "Attention guided graph convolutional networks for relation extraction," *arXiv preprint arXiv:1906.07510*, 2019.

[9] W. Zhou, K. Huang, T. Ma **and** J. Huang, "Document-level relation extraction with adaptive thresholding and localized context pooling," **in***Proceedings of the AAAI conference on artificial intelligence* **volume** 35, 2021, **pages** 14 612–14 620.

[10] X. Zeng, D. Zeng, S. He, K. Liu **and** J. Zhao, "Extracting relational facts by an end-to-end neural model with copy mechanism," **in***Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2018, **pages** 506–514.

[11] T.-J. Fu, P.-H. Li **and** W.-Y. Ma, "Graphrel: Modeling text as relational graphs for joint entity and relation extraction," **in***Proceedings of the 57th annual meeting of the association for computational linguistics* 2019, **pages** 1409–1418.

[12] A. Vaswani, N. Shazeer, N. Parmar **andothers**, "Attention is all you need," *Advances in neural information processing systems*, **jourvol** 30, 2017.

[13] J. Devlin, M.-W. Chang, K. Lee **and** K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] T. N. Kipf **and** M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

PHỤ LỤC

A. TẬP DỮ LIỆU DOCRED

A.1 Các loại quan hệ có trong tập dữ liệu DocRED

Chi tiết các loại quan hệ có trong tập dữ liệu DocRED được liệt kê như dưới bảng sau đây:

STT	Mã	Tên quan hệ	Ý nghĩa
1	P1376	capital of	thủ đô
2	P136	genre	thể loại
3	P137	operator	nhà điều hành
4	P131	located in the administrative territorial entity	nằm trong đơn vị hành chính lãnh thổ
5	P607	conflict	xung đột
6	P527	has part	có một phần
7	P1412	languages spoken, written or signed	ngôn ngữ được nói, viết hoặc ký
8	P206	located in or next to body of water	nằm trong hoặc bên cạnh vùng nước
9	P205	basin country	quốc gia lưu vực
10	P449	original network	mạng ban đầu
11	P127	owned by	được sở hữu bởi
12	P123	publisher	nhà xuất bản
13	P86	composer	nhà soạn nhạc
14	P840	narrative location	vị trí tường thuật
15	P355	subsidiary	công ty con
16	P737	influenced by	bị ảnh hưởng bởi
17	P740	location of formation	vị trí hình thành
18	P190	sister city	thành phố kết nghĩa
19	P576	dissolved, abolished or demolished	giải thể, bãi bỏ hoặc phá bỏ
20	P194	legislative body	cơ quan lập pháp
21	P112	founded by	thành lập bởi
22	P118	league	liên đoàn
23	P17	country	đất nước
24	P19	place of birth	nơi sinh
25	P3373	sibling	anh chị em ruột
26	P6	head of government	người đứng đầu chính phủ

27	P276	location	vị trí
28	P1001	applies to jurisdiction	áp dụng cho thẩm quyền
29	P580	start time	thời gian bắt đầu
30	P582	end time	thời gian kết thúc
31	P585	point in time	thời điểm
32	P463	member of	thành viên của
33	P676	lyrics by	viết lời bởi
34	P674	characters	nhân vật
35	P264	record label	hãng thu âm
36	P108	employer	nhà tuyển dụng
37	P102	member of political party	thành viên của đảng chính trị
38	P25	mother	mẹ
39	P27	country of citizenship	quốc tịch
40	P26	spouse	vợ chồng
41	P20	place of death	nơi mất
42	P22	father	bố
43	P807	separated from	tách khỏi
44	P800	noteable work	công việc đáng chú ý
45	P279	subclass of	phân lớp của
46	P1336	territory claimed by	lãnh thổ được tuyên bố bởi
47	P577	publication date	ngày xuất bản
48	P570	date of death	ngày mất
49	P571	inception	khởi đầu
50	P178	developer	người phát triển
51	P179	series	loạt
52	P272	production company	công ty sản xuất
53	P170	creator	người sáng tạo
54	P170	parent taxon	đơn vị phân loại mẹ
55	P172	ethnic group	nhóm dân tộc
56	P175	performer	người biểu diễn
57	P176	manufacturer	nhà chế tạo
58	P39	position held	vị trí nắm giữ
59	P30	continent	lục địa
60	P31	instance of	trường hợp của
61	P36	capital	thủ đô
62	P37	official language	ngôn ngữ chính thức
63	P35	head of state	nguyên thủ quốc gia

64	P400	platform	nền tảng
65	P403	mouth of the watercourse	cửa sông
66	P361	part of	một phần của
67	P364	original language of work	ngôn ngữ gốc của tác phẩm
68	P569	date of birth	ngày sinh
69	P710	participant	người tham gia
70	P1344	participant of	người tham gia của
71	P488	chairperson	chủ tịch
72	P241	military branch	chi nhánh quân đội
73	P162	producer	nhà sản xuất
74	P161	cast member	diễn viên
75	P166	award received	giải thưởng nhận được
76	P40	child	con cái
77	P1441	present in work	hiện đang làm việc
78	P156	followed by	theo dõi bởi
79	P155	follows	theo dõi
80	P150	contains administrative territorial entity	chứa thực thể lãnh thổ hành chính
81	P551	residence	nơi cư trú
82	P706	located on terrain feature	nằm trên đặc điểm địa hình
83	P159	headquarters location	địa điểm trụ sở chính
84	P495	country of origin	nước xuất xứ
85	P58	screenwriter	biên kịch
86	P749	parent organization	tổ chức mẹ
87	P54	member of sports team	thành viên của đội thể thao
88	P57	director	đạo diễn/giám đốc
89	P50	author	tác giả
90	P1366	replaced by	thay thế bởi
91	P1365	replaces	thay thế
92	P937	work location	địa điểm làm việc
93	P140	religion	tôn giáo
94	P69	educated at	được giáo dục tại
95	P1198	unemployment rate	tỷ lệ thất nghiệp
96	P1056	product or material produced	sản phẩm hoặc vật liệu được sản xuất

Bảng A.1: Danh sách các loại quan hệ trong tập dữ liệu DocRED

B. KẾT QUẢ CHẠY CHƯƠNG TRÌNH

B.1 Kết quả chạy kiểm thử mô hình trên tập kiểm thử

Chi tiết kết quả chạy mô hình phân lớp đa ví dụ được trình bày ở bảng B.1. Kết quả bao gồm các điểm precision, recall, f1-score trên từng lớp của các ví dụ trong tập kiểm thử.

Phân loại các đề cập				
type	precision	recall	f1-score	support
Binary	93.65	92.66	93.15	17926
micro	93.65	92.66	93.15	17926
macro	93.65	92.66	93.15	17926
Phân giải đồng tham chiếu				
type	precision	recall	f1-score	support
Binary	82.71	83.32	83.02	13593
micro	82.71	83.32	83.02	13593
macro	82.71	83.32	83.02	13593
Xác định các thực thể				
type	precision	recall	f1-score	support
MISC	72.85	69.14	70.95	2278
LOC	81.15	82.72	81.93	4033
PER	80.21	83.95	82.04	2081
TIME	89.02	88.30	88.66	2598
NUM	85.90	87.32	86.60	844
ORG	71.28	73.18	72.22	1760
micro	80.17	80.75	80.46	13594
macro	80.07	80.77	80.40	13594
Phân loại các quan hệ				
type	precision	recall	f1-score	support
P580	50.00	20.83	29.41	24
P740	25.00	15.38	19.05	13
P140	9.68	17.65	12.50	34
P170	27.78	14.71	19.23	34
P1001	28.85	26.79	27.78	56
P39	0.00	0.00	0.00	4
P577	57.14	53.24	55.12	293
P1056	100.00	16.67	28.57	6

P178	32.76	33.33	33.04	57
P69	50.75	58.62	54.40	58
P1441	49.06	52.00	50.49	50
P54	56.45	58.82	57.61	119
P131	44.70	39.64	42.02	936
P272	27.78	27.78	27.78	18
P171	12.50	27.27	17.14	11
P607	43.90	39.56	41.62	91
P20	60.53	60.53	60.53	38
P1344	30.00	31.03	30.51	29
P551	0.00	0.00	0.00	4
P676	33.33	11.76	17.39	17
P737	0.00	0.00	0.00	1
P840	50.00	16.67	25.00	12
P150	55.44	47.66	51.26	449
P17	48.55	50.54	49.53	2026
P57	51.85	51.85	51.85	54
P175	53.33	52.41	52.87	290
P495	15.87	6.45	9.17	155
P22	26.67	24.00	25.26	50
P749	22.22	9.09	12.90	22
P19	64.29	56.25	60.00	128
P403	17.24	21.74	19.23	23
P40	27.45	18.42	22.05	76
P50	53.23	51.56	52.38	64
P206	22.22	12.24	15.79	49
P279	9.09	3.57	5.13	28
P264	42.37	37.31	39.68	134
P1366	0.00	0.00	0.00	16
P156	16.67	1.92	3.45	52
P112	15.79	11.11	13.04	27
P1336	0.00	0.00	0.00	4
P31	4.55	3.45	3.92	29
P488	14.29	10.00	11.76	10
P807	0.00	0.00	0.00	1
P37	2.33	2.94	2.60	34
P172	12.50	5.56	7.69	18

P205	18.18	8.33	11.43	24
P276	44.44	12.50	19.51	32
P364	14.29	30.00	19.35	10
P800	35.29	23.08	27.91	26
P937	40.00	16.67	23.53	24
P161	48.24	58.57	52.90	140
P35	9.52	8.33	8.89	24
P571	32.22	29.00	30.53	100
P108	34.78	20.51	25.81	39
P1365	0.00	0.00	0.00	11
P127	0.00	0.00	0.00	72
P576	33.33	5.00	8.70	20
P3373	31.63	25.62	28.31	121
P1412	4.00	2.94	3.39	34
P123	61.90	44.83	52.00	29
P706	44.44	9.52	15.69	42
P1198	0.00	0.00	0.00	1
P241	16.00	16.00	16.00	25
P710	15.79	26.09	19.67	23
P585	47.06	25.81	33.33	31
P194	24.39	30.30	27.03	33
P118	46.67	40.38	43.30	52
P30	32.98	43.06	37.35	72
P674	17.65	10.00	12.77	30
P355	66.67	5.71	10.53	35
P137	25.00	7.69	11.76	26
P179	40.00	36.36	38.10	22
P463	29.27	23.76	26.23	101
P176	25.00	16.67	20.00	30
P400	62.26	58.93	60.55	56
P582	28.57	22.22	25.00	9
P27	27.76	28.21	27.98	553
P190	0.00	0.00	0.00	2
P6	15.38	12.50	13.79	32
P155	20.00	1.64	3.03	61
P569	71.17	74.18	72.64	213
P570	65.32	64.20	64.76	176

P102	36.76	32.47	34.48	77
P26	32.56	22.58	26.67	62
P361	44.55	27.22	33.79	180
P527	32.81	24.71	28.19	170
P58	28.00	35.00	31.11	20
P162	20.00	11.76	14.81	34
P25	25.00	10.00	14.29	20
P159	30.00	22.64	25.81	53
P449	45.00	58.06	50.70	31
P136	27.27	11.11	15.79	27
P166	30.56	28.95	29.73	38
P36	31.58	23.08	26.67	26
P86	60.00	11.54	19.35	52
P1376	50.00	22.73	31.25	22
micro	43.60	38.76	41.04	8787
macro	31.14	23.76	25.35	8787

Bảng B.1: Kết quả chạy mô hình phân lớp đa ví dụ trên tập kiểm thử

Chi tiết kết quả chạy mô hình phân lớp sử dụng đồ thị được trình bày ở bảng B.2.

Phân loại các đề cập				
type	precision	recall	f1-score	support
Binary	92.74	92.49	92.61	17926
micro	92.74	92.49	92.61	17926
macro	92.74	92.49	92.61	17926
Phân giải đồng tham chiếu				
type	precision	recall	f1-score	support
Binary	81.52	83.13	82.32	13593
micro	81.52	83.13	82.32	13593
macro	81.52	83.13	82.32	13593
Xác định các thực thể				
type	precision	recall	f1-score	support
MISC	70.07	69.36	69.71	2278
NUM	83.62	85.31	84.46	844
PER	81.06	84.09	82.55	2081
LOC	80.84	82.02	81.43	4033

TIME	86.83	88.30	87.56	2598
ORG	68.84	72.44	70.60	1760
micro	78.83	80.38	79.60	13594
macro	78.54	80.25	79.38	13594

Phân loại các quan hệ				
type	precision	recall	f1-score	support
P150	55.15	46.55	50.48	449
P582	40.00	22.22	28.57	9
P40	22.22	13.16	16.53	76
P178	36.36	21.05	26.67	57
P57	49.23	59.26	53.78	54
P264	36.88	38.81	37.82	134
P190	0.00	0.00	0.00	2
P127	0.00	0.00	0.00	72
P463	24.62	31.68	27.71	101
P1366	0.00	0.00	0.00	16
P37	5.88	8.82	7.06	34
P58	28.57	30.00	29.27	20
P400	57.41	55.36	56.36	56
P241	17.39	16.00	16.67	25
P176	35.00	23.33	28.00	30
P495	17.14	7.74	10.67	155
P551	0.00	0.00	0.00	4
P570	67.24	66.48	66.86	176
P123	66.67	34.48	45.45	29
P156	0.00	0.00	0.00	52
P585	45.00	29.03	35.29	31
P3373	27.54	31.40	29.34	121
P19	64.46	60.94	62.65	128
P449	32.31	67.74	43.75	31
P279	15.38	7.14	9.76	28
P137	0.00	0.00	0.00	26
P206	25.00	4.08	7.02	49
P569	66.95	75.12	70.80	213
P749	25.00	4.55	7.69	22
P17	52.32	46.84	49.43	2026
P740	33.33	23.08	27.27	13

P800	42.86	34.62	38.30	26
P577	55.16	47.44	51.01	293
P39	0.00	0.00	0.00	4
P54	61.02	60.50	60.76	119
P840	30.00	25.00	27.27	12
P175	50.51	51.38	50.94	290
P706	36.36	9.52	15.09	42
P1376	37.50	13.64	20.00	22
P527	31.43	32.35	31.88	170
P172	8.33	5.56	6.67	18
P50	55.74	53.12	54.40	64
P170	31.58	17.65	22.64	34
P1365	0.00	0.00	0.00	11
P710	16.28	30.43	21.21	23
P1198	0.00	0.00	0.00	1
P580	38.46	20.83	27.03	24
P69	38.14	63.79	47.74	58
P171	7.69	9.09	8.33	11
P674	28.57	20.00	23.53	30
P364	16.67	30.00	21.43	10
P86	50.00	1.92	3.70	52
P140	13.95	17.65	15.58	34
P1344	20.45	31.03	24.66	29
P30	29.89	36.11	32.70	72
P276	28.57	12.50	17.39	32
P571	30.00	27.00	28.42	100
P576	33.33	5.00	8.70	20
P1441	51.02	50.00	50.51	50
P361	36.36	26.67	30.77	180
P26	38.46	24.19	29.70	62
P166	22.22	31.58	26.09	38
P112	7.14	7.41	7.27	27
P937	28.12	37.50	32.14	24
P20	51.28	52.63	51.95	38
P25	0.00	0.00	0.00	20
P179	30.77	36.36	33.33	22
P1056	33.33	16.67	22.22	6

P118	48.84	40.38	44.21	52
P6	10.53	12.50	11.43	32
P102	31.40	35.06	33.13	77
P607	39.77	38.46	39.11	91
P108	25.93	17.95	21.21	39
P807	0.00	0.00	0.00	1
P159	27.27	22.64	24.74	53
P162	17.65	8.82	11.76	34
P22	14.55	16.00	15.24	50
P355	0.00	0.00	0.00	35
P136	27.27	11.11	15.79	27
P737	0.00	0.00	0.00	1
P205	40.00	8.33	13.79	24
P403	14.29	8.70	10.81	23
P194	18.06	39.39	24.76	33
P155	0.00	0.00	0.00	61
P488	7.14	10.00	8.33	10
P131	39.04	38.25	38.64	936
P1336	0.00	0.00	0.00	4
P1001	33.33	17.86	23.26	56
P31	5.26	3.45	4.17	29
P676	33.33	5.88	10.00	17
P36	25.00	19.23	21.74	26
P272	26.09	33.33	29.27	18
P35	8.33	8.33	8.33	24
P27	26.43	30.02	28.11	553
P161	41.18	55.00	47.09	140
P1412	5.00	2.94	3.70	34
micro	41.61	37.73	39.58	8787
macro	27.12	23.41	23.72	8787

Bảng B.2: Kết quả chạy mô hình phân lớp sử dụng đồ thị trên tập kiểm thử

B.2 Chạy thử mô hình

Khi mô hình nhận vào đầu vào là một tài liệu chứa nhiều câu văn, mô hình sẽ trả về kết quả bao gồm: vị trí các từ đề cập đến thực thể, các cụm thực thể cùng với loại thực thể, và mối quan hệ giữa các thực thể. Chi tiết kết quả xem ở bảng B.3.

Câu đầu vào	Barack Obama was born in 1961 in Hawaii. He is the former President of the United States. Michelle Obama is his wife. Both got married in 1992.
Output	[**"tokens"**: ["Barack", "Obama", "was", "born", "in", "1961", "in", "Hawaii", ".", "He", "is", "the", "former", "President", "of", "the", "United", "States", ".", "Michelle", "Obama", "is", "his", "wife", ".", "Both", "got", "married", "in", "1992", "."], **"mentions"**: ["start": 5, "end": 6, "start": 7, "end": 8, "start": 0, "end": 2, "start": 15, "end": 18, "start": 19, "end": 21, "start": 29, "end": 30], **"entities"**: ["mentions": [2, 4], "type": "PER", "mentions": [3], "type": "LOC", "mentions": [5], "type": "TIME", "mentions": [1], "type": "LOC", "mentions": [0], "type": "TIME"], **"relations"**: ["head": 0, "tail": 1, "type": "P27", "head": 0, "tail": 4, "type": "P569", "head": 1, "tail": 3, "type": "P150", "head": 3, "tail": 1, "type": "P17"]]
Ý nghĩa	Có 5 thực thể: Barack Obama, Hoa Kỳ, năm 1992, Hawaii, năm 1961. Có 4 mối quan hệ: Barack Obama có quốc tịch Hoa Kỳ, Barack Obama sinh năm 1961, Hoa Kỳ chứa thực thể lãnh thổ hành chính Hawaii, Hawaii nằm trong Hoa Kỳ.

Bảng B.3: Kết quả chạy thử chương trình