TRÍCH XUẤT THỰC THỂ TRONG AN TOÀN THÔNG TIN SỬ DỤNG HỌC SÂU

Nguyễn Ngọc Điệp, Nguyễn Thị Thanh Thủy
Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Hiện nay, với sự gia tăng nhanh chóng của các nguồn tài liệu liên quan đến lĩnh vực an toàn thông tin, việc trích xuất tự động các thông tin quan trọng từ các nguồn tài liệu này là một nhu cầu cấp thiết. Một trong những loại thông tin phổ biến cần trích xuất đó là các thực thể có tên, như tên chương trình phần mềm, tin tặc, chương trình mã độc, lỗ hổng, công nghệ, các kỹ thuật,... Tuy nhiên, do tính phức tạp, đa dạng, có nhiều đặc trưng về chuyên ngành riêng của các nguồn tài liệu này, việc xác định các thực thể có tên hiện còn gặp rất nhiều khó khăn. Gần đây, có một số phương pháp tiếp cận để giải quyết bài toán này, trong đó nổi trội hơn là các phương pháp dựa trên học sâu, là các kỹ thuật tiên tiến nhất, được sử dụng nhiều trong lĩnh vực trích xuất thông tin. Trong bài báo này, chúng tôi trình bày một phương pháp trích xuất thực thể có tên trong an toàn thông tin sử dụng các kỹ thuật học sâu, là mô hình kết hợp gồm word2vec, BERT, BiLSTM và CRF. Đồng thời, chúng tôi cũng đề xuất một phương pháp để tăng cường, bổ sung dữ liệu cho các thực thể có số lượng ít trong tập dữ liệu. Kết quả cho thấy mô hình đề xuất có độ chính xác khá cao, với độ đo F_1 lên tới 72,86% khi thử nghiệm trích xuất thực thể có tên trên tập dữ liệu văn bản an toàn thông tin. Phương pháp tăng cường dữ liệu đề xuất cũng đạt được hiệu quả khả quan.

Từ khóa: An toàn thông tin, trích xuất thực thể, BiLSTM, CRF, BERT.

I. GIỚI THIỆU

Sự phát triển nhanh chóng của công nghệ Internet kéo theo ngày càng nhiều những mối de dọa cho người dùng và các công ty trên toàn thế giới. Mỗi ngày, có rất nhiều các báo cáo điều tra về các mối đe dọa, sự cố về an toàn thông tin cùng nhiều các văn bản về vấn đề an ninh mạng khác như các hướng dẫn, chính sách, các công cụ, công nghệ được cung cấp trên Internet. Việc xác định và phân loại các thông tin về an toàn thông tin một cách tự động đóng vai trò cấp thiết trong nhiều ứng dụng, và hỗ trợ cho nhiều đối tượng người dùng khác nhau như nhân viên kiểm toán, nhà phân tích bảo mật, các nhà nghiên cứu, v.v. Một trong những loại thông tin phổ biến, quan trọng cần được xác định là các thực thể có tên (sau đây sẽ gọi tắt là thực thể) trong các văn bản an toàn thông tin. Tuy nhiên, do tính phức tạp và đa dạng của văn bản trong lĩnh vực này, việc xác định các thực thể này là một công việc có nhiều thách thức.

Về cơ bản, việc xác định các thực thể trong an toàn thông tin là bài toán nhận dạng thực thể có tên (NER) trong xử lý ngôn ngữ tự nhiên. Các thực thể có thể là chương trình phần mềm, thiết bị, công nghệ, tin tặc hay chương trình độc hại, lỗ hổng (CVE), v.v. Một trong các phương pháp tiếp cận ban đầu nhanh chóng và hiệu quả để nhận dạng các thực thể này là dựa trên luật. Các phương pháp dựa trên luật có thể trích xuất các thực thể theo mẫu như email, địa chỉ IP hay các lỗ hổng phổ biến, hoặc dựa vào tập từ điển để nhận dạng ra các thực thể đã biết. Tuy nhiên phương pháp này không phù hợp đối với các trường hợp phức tạp của văn bản an toàn thông tin, với cấu trúc văn bản không theo quy tắc thông thường, xuất hiện nhiều thực thể có tên mới, đồng thời yêu cầu chi phí cao về cả thời gian, con người và tiền bạc để duy trì, cập nhật kịp thời thông tin mới nhất xuất hiện liên tục trong thời gian tính bằng phút hoặc thậm chí bằng giây.

Tiếp đó, nhiều phương pháp học máy khác nhau được áp dụng để trích xuất thực thể mới trong các văn bản an toàn thông tin như dựa trên Conditional random fields (CRF) [2, 3], Support vector machines (SVM) [4], Expectation regularization [5], Bootstrapping algorithm [6], Maximum entropy model (ME) [7] nhưng tính hiệu quả chưa thực sự cao, dù cho các phương pháp này đã đạt được kết quả tốt khi nhận dạng thực thể mới trong các văn bản thông thường trong xử lý ngôn ngữ tự nhiên. Nguyên nhân là các mô hình này cần phải xác định nhiều đặc trưng thủ công và bỏ qua mối tương quan của các thực thể, kéo theo đó là việc khó có thể đáp ứng được với các ứng dụng quy mô lớn [1].

Một bước tiến nhảy vọt về xử lý ngôn ngữ tự nhiên trong những năm gần đây là học sâu. Đây là mô hình mạng nơ-ron có thể tự học hiệu quả các đặc trưng tổ hợp phi tuyến, trong khi các phương pháp cổ điển như CRF chỉ có thể học các tổ hợp tuyến tính của các đối tượng đã xác định. Việc mở rộng khả năng truy cập thông tin, tăng sức mạnh xử lý phần cứng, đặc biệt là GPU, các chức năng kích hoạt khác nhau, v.v. đã làm cho học sâu trở nên đặc biệt thiết thực và có tính khả thi. Gần đây, nhiều phương pháp hiệu quả hơn cho các ứng dụng khác nhau trong lĩnh vực an toàn thông tin dựa trên học sâu đã được đề xuất. Trong ứng dụng nhận dạng thực thể, một số mô hình hiệu quả như Long short-term memory (LSTM) [8, 9], mô hình kết hợp giữa LSTM và CRF với mô hình word2vec [8] và tốt hơn nữa là các phương pháp dựa trên BERT [17]. Các mô hình dựa trên BERT đã tận dụng được tri thức học được từ các văn bản có sẵn, giúp cho việc nhận dạng thực thể trong văn bản an

Tác giả liên hệ: Nguyễn Ngọc Điệp,
Email: diepnn@ptit.edu.vn
Đến tòa soạn: 31/10/2021, chỉnh sửa: 29/11/2021, chấp nhận đăng: 9/12/2021.

toàn thông tin chính xác hơn so với các phương pháp trước đó. Tuy nhiên, do văn bản an toàn thông tin là một lĩnh vực đặc thù, chứa rất nhiều từ vựng liên quan đến bảo mật như tên tệp, giá trị băm, tên các công cụ tấn công, nên nếu chỉ tận dụng các tri thức từ các văn bản có sẵn phổ biến trong xử lý ngôn ngữ tự nhiên thì sẽ không đạt được độ chính xác cao.

Trong nghiên cứu này, chúng tôi đề xuất sử dụng mô hình kết hợp bao gồm word2vec, BERT [17], BiLSTM [21] và CRF [22] để giải quyết bài toán trích xuất thực thể có tên trong văn bản an toàn thông tin. Đây là sự kết hợp những ưu điểm của các mô hình như [8] và [17]. Mô hình kết hợp này hiệu quả trong lĩnh vực hẹp và có tính chuyên môn cao như an toàn thông tin, với việc kết hợp khả năng tích hợp các đặc trưng cho từ, trích xuất từ mô hình ngôn ngữ ngữ cảnh như BERT, và mô hình ngôn ngữ phi ngữ cảnh nhưng lại đặc biệt hiệu quả cho các từ mới và phức tạp là FastText cùng các đặc trưng từ mức ký tự. Kết quả thực nghiệm trong phần sau cho thấy, mô hình hoạt động có độ chính xác cao trên tập dữ liệu văn bản an toàn thông tin.

Phần còn lại của bài báo được tổ chức như sau. Phần II mô tả các nghiên cứu liên quan. Phần III trình bày đề xuất phương pháp thực hiện trích xuất thực thể có tên trong an toàn thông tin. Kết quả và những phân tích thực nghiệm được trình bày trong phần Phần IV. Cuối cùng, Phần V là kết luận bài báo và định hướng nghiên cứu.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Trong những năm gần đây đã có nhiều nghiên cứu về trích xuất thông tin trong lĩnh vực an toàn thông tin. Trong đó các thông tin được trích xuất có thể là thông tin về các mối đe dọa liên quan đến các cuộc tấn công và lỗ hổng tiềm ẩn trong văn bản web [15], tên các sản phẩm phần mềm, hệ điều hành [2], tên mã độc, địa chỉ IP, tên miền [7],… Tuy nhiên, việc trích xuất các loại thông tin trong lĩnh vực an toàn thông tin còn gặp rất nhiều khó khăn do thiếu nguồn dữ liệu đã được chú thích [7], các nguồn dữ liệu không đồng nhất [16], đồng thời trên thực tế các nguồn thông tin mới liên tục xuất hiện. Việc này ảnh hưởng khá nhiều đến các phương pháp tiếp cận giải quyết bài toán.

Một bài toán trích xuất thông tin quan trọng trong an toàn thông tin là trích xuất thực thể an toàn thông tin. Có hai hướng tiếp cận chính thường được sử dụng cho bài toán này đó là hướng tiếp cận dựa trên luật, và hướng tiếp cận dựa trên học máy thống kê. Hướng tiếp cận dựa trên luật cần có chuyên gia xử lý và sinh luật, thường được thực hiện bằng cách xây dựng các biểu thức chính quy, hoặc các quy tắc heuristic nên rất khó khăn, tốn kém thời gian và khó thực hiện với quy mô lớn [10, 11].

Hướng tiếp cận dựa trên học máy thống kê được nghiên cứu và phát triển nhiều hơn do không phụ thuộc vào tri thức chuyên gia, được thực hiện một cách tự động, đồng thời được đánh giá là có độ chính xác cao. Do vậy thực tế có khá nhiều các nghiên cứu tiếp cận giải quyết bài toán theo hướng này [2, 3, 4, 5, 6, 7]. Bridges và cộng sự [7] đề xuất Mô hình Maximum Entropy được huấn luyện với nhiều kho dữ liệu bảo mật và đạt được hiệu suất cao trong việc xác định và phân loại các thực thể. Jones và cộng sự [6] triển khai một thuật toán bootstrapping để trích xuất các thực thể và mối quan hệ của chúng từ các văn bản bảo mật. Joshi và cộng sự [2] sử dụng trường ngẫu nhiên có điều kiện CRF để xác định các thực thể, khái niệm và quan hệ liên quan đến an ninh mạng từ nguồn dữ liệu về lỗ hổng bảo mật và từ các nguồn văn bản. Lal và cộng sự [3] cũng sử dụng thuật toán trường ngẫu nhiên có điều kiện CRF để trích xuất các khái niệm và thực thể liên quan đến an ninh mạng bằng cách sử dụng một tập hợp các đặc trưng từ các văn bản bảo mật được chú thích thủ công.

Gần đây các phương pháp hiệu quả hơn cho trích xuất thực thể an toàn thông tin dựa trên học sâu được đề xuất. Trong nghiên cứu [8], Gasmi và cộng sự kết hợp ưu điểm của phương pháp BiLSTM (bidirectional long short-term memory) và trường ngẫu nhiên có điều kiện (CRF) để cải thiện độ chính xác của việc trích xuất thực thể, kết quả đạt được tốt hơn so với phương pháp chỉ sử dụng CRF truyền thống. Nghiên cứu của Qin và cộng sự [9] đề xuất một mô hình kết hợp của mạng nơ-ron được gọi là FT-CNN-BiLSTM-CRF. Tác giả sử dụng các đặc trưng ngữ cảnh cho huấn luyện mô hình trích xuất, kết quả đạt 86% tính theo độ đo F_1 trên tập dữ liệu thử nghiệm.

Nghiên cứu này cũng sử dụng mô hình học sâu kết hợp ưu điểm của BiLSTM và CRF tương tự như nghiên cứu [8], [9]. Tuy nhiên chúng tôi còn kết hợp thêm ưu điểm của các các phương pháp biểu diễn từ khác nhau để mô hình hiệu quả hơn, bao gồm BERT với khả năng biểu diễn ngôn ngữ ngữ cảnh, kết hợp với khả năng biểu diễn từ hiệu quả cho các từ mới, ít xuất hiện như FastText và charCNN. Các kết hợp này tạo ra mô hình hiệu quả cho việc trích xuất thực thể có tên trong văn bản an toàn thông tin.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Ý tưởng chính đề xuất cho kiến trúc trích xuất thực thể trong văn bản an toàn thông tin là kết hợp nhiều phương pháp biểu diễn từ hiệu quả vào kiến trúc mạng nơ-ron học sâu, bao gồm đặc trưng ngữ cảnh BERT, đặc trưng biểu diễn từ theo kiểu *n*-gram với FastText, đặc trưng biểu diễn từ mức ký tự. Các đặc trưng biểu diễn từ phi ngữ cảnh biểu diễn mỗi từ bằng một véc-tơ hữu ích trong miền dữ liệu chuyên ngành như an toàn thông tin do các từ ngữ sử dụng không mơ hồ, các thực thể có tên thường có ý nghĩa duy nhất, không phụ thuộc vào ngữ cảnh hay các mối liên hệ trong văn bản. Tuy nhiên, kể cả văn bản chuyên ngành an toàn thông tin cũng không thể tránh khỏi sự mơ hồ, đa nghĩa, phụ thuộc ngữ cảnh, và khi đó đặc trưng ngữ cảnh BERT sẽ rất hữu ích khi có thể biểu diễn chính xác ngữ nghĩa của từ trong câu. Ngoài ra, các đặc trưng dựa trên FastText kết hợp với đặc trưng của từ ở mức ký tự rất hiệu quả trong biểu diễn các từ mới.

Phần này trình bày lý thuyết về một số mô hình học sâu có liên quan, sau đó là mô tả bài toán và đề xuất phương pháp trích xuất thực thể có tên trong văn bản an toàn thông tin dựa trên kết hợp nhiều đặc trưng biểu diễn từ và kiến trúc BiLSTM-CRF. Mô hình đề xuất gồm 2 phần chính: (1) xây dựng véc-tơ từ được biểu diễn theo các cách khác

nhau và (2) kiến trúc mạng nơ-ron BiLSTM-CRF để đưa ra các dự đoán từ các đặc trưng từ kết hợp.

Về biểu diễn từ, chúng tôi sử dụng các phương pháp trích xuất khác nhau: biểu diễn từ mức ký tự dựa trên mạng CNN, FastText để biểu diễn từ theo *n*-gram ký tự và mô hình BERT để biểu diễn từ theo ngữ cảnh. Sau đó, kết hợp các đặc trưng có được của các phương pháp biểu diễn từ thành một véc-tơ tổng trước khi cung cấp cho kiến trúc mạng học sâu BiLSTM-CRF. Mạng có các lớp BiLSTM và lớp CRF để biểu diễn câu và suy luận nhãn tương ứng.

Để hiểu chi tiết hơn về kiến trúc của mô hình đề xuất, trước hết chúng tôi giới thiệu sơ bộ về các mô hình học sâu có liên quan như phần A dưới đây, sau đó mô tả về bài toán và mô hình đề xuất.

A. *Một số mô hình học sâu*

1) Mạng nơ-ron tích chập (CNN)

CNN là mạng rất nổi tiếng do có hiệu năng cao và ít sử dụng các tham số học hơn. Mạng này bao gồm ba loại tầng là tầng tích chập, tầng gộp và tầng được kết nối đầy đủ. Trong tầng tích chập của mạng CNN, phép toán tích chập được thực hiện bằng cách sử dụng một số bộ lọc trượt qua đầu vào và học đặc trưng từ dữ liệu đầu vào. Tầng gộp được sử dụng để kết hợp thông tin qua các vùng không gian kề nhau bằng cách giảm kích thước của tầng trước đó. Có các loại tầng gộp khác nhau bao gồm gộp cực tiểu, gộp cực đại và gộp trung bình. Mạng được kết nối với một tầng dày đặc ở cuối để các đặc trưng có thể được ánh xạ phân loại.

2) Mạng bộ nhớ dài-ngắn (LSTM)

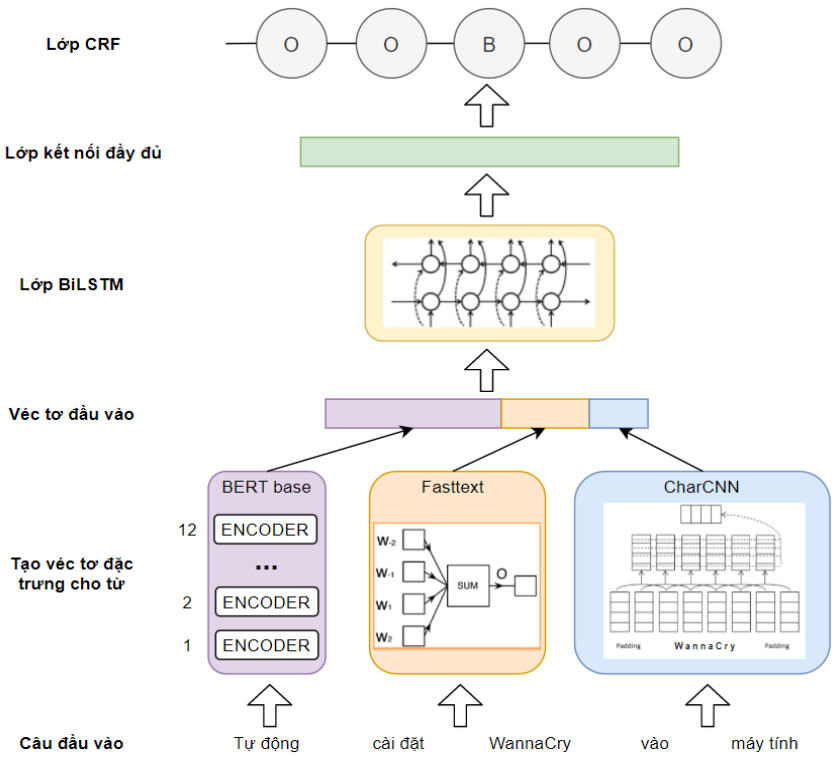
Mạng LSTM là một dạng đặc biệt của mạng nơ-ron hồi quy (RNN), được đưa ra để giải quyết vấn đề triệt tiêu gradient trong RNN. LSTM có khả năng học được các phụ thuộc xa, có thể ghi nhớ có chọn lọc các mẫu trong một thời gian dài mà không cần phải huấn luyện (trong khi RNN chỉ có thể xử lý dữ liệu ngắn hạn). LSTM có kiến trúc là dạng chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron, trong đó mỗi mô-đun có 4 tầng tương tác với nhau (khác với RNN chuẩn chỉ có 1 tầng mạng nơ-ron).

Mấu chốt của mạng LSTM là trạng thái tế bào, chạy xuyên suốt tất cả các nút mạng giúp thông tin có thể dễ dàng di chuyển và không bị thay đổi. Việc thêm hoặc bớt thông tin cần thiết cho trạng thái tế bào được thực hiện (sàng lọc) bởi các cổng.

Một LSTM có 3 cổng để duy trì và điều hành trạng thái của tế bào. Mỗi cổng được kết hợp bởi một tầng mạng sigmoid và một phép nhân. Đầu ra của tầng sigmoid là một số trong khoảng [0, 1], mô tả số lượng thông tin được qua. Nếu đầu ra là 1 thì cho tất cả các thông tin đi qua, nếu đầu ra là 0 thì không cho thông tin nào qua cả.

3) Các dạng kiến trúc hai chiều

Để hiểu ngữ cảnh tốt hơn và giải quyết những điểm mơ hồ trong văn bản, các cấu trúc hồi quy hai chiều (bidirectional) được sử dụng để học thông tin từ các dấu thời gian trong quá khứ và cả tương lai. Mỗi cấu trúc này có hai loại kết nối, trong đó một loại đi về phía trước theo thời gian và loại còn lại đi lùi lại theo thời gian. Các kết nối nhằm trợ giúp trong việc học các biểu diễn trong quá khứ và tương lai. Một số dạng mô-đun có thể có cấu trúc hồi quy hai chiều là RNN, LSTM hoặc GRU.



Hình 1. Kiến trúc mô hình trích xuất thực thể trong văn bản an toàn thông tin

4) Trường ngẫu nhiên có điều kiện (CRF)

CRF là một mô hình đồ thị xác suất được sử dụng rộng rãi trong các tác vụ gán nhãn tuần tự như NER, gán thẻ giọng nói và nhận dạng giọng nói. CRF kết hợp các đặc điểm của mô hình entropy cực đại và mô hình Markov ẩn (HMM). CRF giảm nhẹ giả thiết về độc lập có điều kiện như trong HMM và sử dụng nhiều đặc trưng toàn cục hơn. Hơn nữa, CRF dựa trên cùng một dạng hàm mũ như mô hình entropy cực đại, thực hiện huấn luyện mô hình với ít dữ liệu hơn.

CRF là một loại mô hình đồ họa xác suất phân biệt, thường áp dụng trong việc dự đoán tuần tự và nhận dạng thực thể có tên. CRF có thể nhận biết thông tin ngữ cảnh từ các nhãn trước đó, do đó tạo ra hiệu suất dự đoán tốt hơn.

5) BERT

BERT là viết tắt của cụm từ Bidirectional Encoder Representation from Transformer, có nghĩa là mô hình biểu diễn từ theo hai chiều, ứng dụng kỹ thuật Transformer. BERT được thiết kế để huấn luyện trước các biểu diễn từ. Điểm đặc biệt ở BERT đó là nó có thể điều hòa cân bằng ngữ cảnh theo cả 2 chiều trái và phải.

Cơ chế attention của Transformer sẽ truyền toàn bộ các từ trong câu văn bản đồng thời vào mô hình một lúc mà không cần quan tâm đến chiều của câu. Do đó Transformer được xem như là huấn luyện hai chiều (bidirectional). Đặc điểm này cho phép mô hình học

được bối cảnh của từ dựa trên toàn bộ các từ xung quanh nó bao gồm cả từ bên trái và từ bên phải.

Mô hình BERT của Google được huấn luyện trên một kho dữ liệu lớn của văn bản không gán nhãn, bao gồm toàn bộ Wikipedia (lên tới 2500 triệu từ) và Book Corpus (lên tới 800 triệu từ). Khi huấn luyện trên kho dữ liệu lớn như vậy, mô hình học và có được sự hiểu biết thực sự sâu sắc về cách thức hoạt động của ngôn ngữ.

B. *Mô tả bài toán*

Giả sử cho một tập văn bản an toàn thông tin $\boldsymbol{L}$, với mỗi văn bản T trong tập văn bản $\boldsymbol{L}$, thông tin cần trích xuất là các thực thể có tên trong văn bản T, ký hiệu là E. Trong một văn bản T có thể có nhiều thực thể có tên E. Xét mỗi câu S trong văn bản T. S sẽ được sử dụng làm đầu vào cho bài toán. Mỗi câu đầu vào S trong một văn bản pháp quy được biểu diễn thành một chuỗi các từ (token) $S=w_1w_2...w_n$, với n là số các từ có trong câu. Từ mỗi câu đầu vào S, cần trích xuất các thực thể E.

Về ngữ nghĩa của bài toán trích xuất thực thể trong văn bản an toàn thông tin ở đây, thực thể có tên là một chuỗi con của các từ liên tiếp đề cập đến một đối tượng liên quan đến ngữ cảnh an toàn thông tin, chẳng hạn như thiết bị, sự kiện, tin tặc, địa điểm, chương trình phần mềm, công nghệ, vi rút…

C. *Mô hình đề xuất*

Mô hình đề xuất cho việc trích xuất thực thể trong văn bản an toàn thông tin là mô hình dựa trên kiến trúc BiLSTM-CRF, khai thác sự kết hợp của các đặc trưng ngữ cảnh và phi ngữ cảnh, đặc trưng biểu diễn từ theo mức ký tự. Hình 1 trình bày kiến trúc của mô hình này.

1) *Trích xuất đặc trưng*

Đặc trưng mức từ có được từ phương pháp nhúng từ với kỹ thuật FastText được Facebook giới thiệu [18], có hiệu suất tốt hơn mô hình Word2vec[23] trong nhiều ứng dụng. Nguyên nhân là FastText biểu thị mỗi từ dưới dạng *n*-gram ký tự thay vì học trực tiếp véc-tơ cho các từ, từ đó giúp nắm bắt ý nghĩa của các từ ngắn hơn và cho phép biểu diễn các hậu tố và tiền tố. Sử dụng FastText cũng cho phép biểu diễn ý nghĩa cho các từ không phổ biến trong lĩnh vực an toàn thông tin, không có trong văn bản huấn luyện, chẳng hạn như ký hiệu các lỗ hổng phổ biến CVE, các công nghệ, tên chương trình phần mềm, v.v.

Đặc trưng mức từ có được từ FastText là đặc trưng phi ngữ cảnh, do đó không mã hóa được các từ đa nghĩa, phụ thuộc ngữ cảnh trong câu. Để nắm bắt được mối tương quan giữa các từ trong một câu, cần sử dụng khả năng biểu diễn từ theo ngữ cảnh từ mô hình BERT [19]. Đây là kỹ thuật học máy dựa trên các Transformer được dùng cho việc học chuyển giao trong xử lý ngôn ngữ tự nhiên bởi Google. Mô hình này là một mô hình học sẵn (pre-trained) học ra các véc-tơ đại diện theo ngữ cảnh 2 chiều của từ trong câu. Trong mô hình đề xuất, chúng tôi sử dụng một mô hình RoBERTa cỡ nhỏ huấn luyện cho đa ngôn ngữ, gồm 12 lớp là 12 bộ mã hóa (encoder) của mô hình Transformer, mỗi lớp tạo ra một véc-tơ 768 chiều để mã hóa một từ. Vì mỗi lớp trong RoBERTa nắm bắt các cấp độ ngữ cảnh khác nhau, nên sẽ hợp lý hơn khi sử dụng nhúng từ nhiều lớp hơn là chỉ sử dụng lớp cuối cùng. Do đó, chúng tôi nối 3 lớp cuối cùng để tạo thành biểu diễn 2034 (768*3) cho một từ.

Các đặc trưng mức từ rất phổ biến và đạt được nhiều thành công trong các ứng dụng xử lý ngôn ngữ tự nhiên. Tuy nhiên, các đặc trưng này cũng tồn tại một số điểm yếu trong xử lý văn bản thuộc lĩnh vực chuyên ngành như an toàn thông tin vì những văn bản này chứa rất nhiều mã và ký hiệu đặc thù chuyên ngành, như các ký hiệu lỗ hổng CVE, các phần mềm và phiên bản tương ứng, hay tên mã các công nghệ. Những cụm từ này hầu như không có ý nghĩa trong ngôn ngữ tự nhiên nhưng mang nhiều thông tin về nhãn của chúng. Do đó, chúng tôi sử dụng mô hình CharCNN để trích xuất các véc-tơ nhúng cấp độ ký tự. Mô hình bao gồm các lớp tích chập 1D, maxpooling và lớp kết nối đầy đủ. Mô hình nhận các chuỗi ký tự đầu vào ở dạng one-hot rồi chuyển qua một lớp embedding để ánh xạ các véc-tơ vào một không gian 30 chiều mới. Lớp tích chập 1D tiếp theo bao gồm 30 kernel với kích cỡ là 3 để duyệt trên các véc-tơ này. Sau đó véc-tơ được làm phẳng và chuyển qua một lớp kết nối đầy đủ 128 unit.

Kết hợp các véc-tơ đặc trưng của ba phương pháp biểu diễn từ ở trên bằng cách nối lại với nhau tạo thành một véc-tơ nhiều chiều biểu diễn cho mỗi từ. Véc-tơ này là đầu

vào cho mạng nơ-ron sâu BiLSTM-CRF được mô tả dưới đây.

2) *Kiến trúc mạng nơ-ron BiLSTM-CRF*

Nhiệm vụ trích xuất thực thể trong văn bản an toàn thông tin được xây dựng dưới dạng bài toán phân loại nhiều đầu ra, trong đó mỗi từ trong câu sẽ được gán một nhãn. Một mạng gồm 2 lớp BiLSTM được sử dụng để chuyển các véc-tơ biểu diễn từ (token) thành véc-tơ biểu diễn câu.

Về cơ bản, LSTM truyền thông tin theo một hướng chỉ có thông tin quá khứ trong lớp không cho phép biết thông tin từ hướng các lớp mạng LSTM hai chiều (BiLSTM) học từ cả hai hướng, cho phép tạo ra các đặc trưng véc-tơ phong phú so với các mô hình LSTM một chiều [12]. Việc áp dụng mô hình này cho phép nắm bắt được nhiều ngữ cảnh nhất có thể, đồng thời còn có thể ngăn ngừa mất mát thông tin [15]. Như thể hiện trong kiến trúc ở Hình 1, các véc-tơ đầu vào kết hợp từ ba phương pháp biểu diễn từ được đưa vào theo cả hai hướng của LSTM. Các đầu ra của BiLSTM lại được sử dụng trong lớp mạng kết nối đầy đủ trước khi vào lớp CRF nhằm suy luận nhãn cho các từ ban đầu. CRF ở đây giúp tận dụng các chuỗi nhãn tốt nhất trong một câu đầu vào nhất định thay vì chỉ các vị trí riêng lẻ.

3) *Bổ sung dữ liệu huấn luyện*

Nhiều dữ liệu thực thể quan trọng trong tập văn bản an toàn thông tin có thể có số lượng rất ít trong tập dữ liệu đã được chú thích, ví dụ như tên của phần mềm độc hại, hay tin tặc. Điều này là do trong thực tế thông thường tên của phần mềm độc hại hay tin tặc không được biết đến vào thời điểm xảy ra cuộc tấn công mà chỉ thường xuất hiện sau thời gian đó. Do vậy, số lượng dữ liệu về các thực thể loại này sẽ không cân đối so với dữ liệu về các thực thể khác. Sự mất cân đối này làm cho mô hình khó có thể học được chính xác biểu diễn của các thực thể, khiến cho hiệu năng nhận dạng thấp.

Để khắc phục tình trạng thiếu hoặc mất cân đối dữ liệu thực thể theo loại, có thể áp dụng một số ý tưởng về bổ sung dữ liệu như trong [13] đưa ra. Tuy nhiên trong tập dữ liệu an toàn thông tin, hạn chế nằm ở số lượng thực thể của một hay một số loại quá ít so với các thực thể còn lại, do đó ở đây chúng tôi chọn phương pháp thay thế thực thể đã được đề cập bằng một thực thể khác cùng kiểu. Cụ thể là đối với mỗi cụm từ thể hiện thực thể cần bổ sung thêm được đề cập trong một câu, chúng tôi sử dụng phân phối nhị thức để quyết định ngẫu nhiên xem có nên thay thế nó hay không. Nếu có, chọn ngẫu nhiên một thực thể khác từ tập huấn luyện ban đầu có cùng loại với thực thể cần thay thế. Trình tự nhãn BIO (Beginning, Inside, Outside) tương ứng có thể được thay đổi tương ứng tùy thuộc kích thước của cụm từ. Ví dụ: Với một câu ban đầu là: "*Một khi nạn nhân mở tệp, mã độc sẽ được kích hoạt và tự động cài đặt EMOTET vào máy tính của nạn nhân.*", thì câu bổ sung là "*Một khi nạn nhân mở tệp, mã độc sẽ được kích hoạt và tự động cài đặt WannaCry vào máy tính của nạn nhân.*"

IV. THỰC NGHIỆM VÀ KẾT QUẢ

A. *Tập dữ liệu*

Nghiên cứu này sử dụng tập dữ liệu Sec_col1 [14] để thử nghiệm mô hình đề xuất trong nhiệm vụ trích xuất thực thể an toàn thông tin. Đây là tập dữ liệu được thu thập với tổng số 861 văn bản, gồm nhiều bài báo, bài đăng và các nhận xét trong lĩnh vực an toàn thông tin. Tổng số có hơn 400 nghìn từ (token), hơn 14 nghìn thực thể được gán nhãn được chia thành 9 loại: người, tổ chức, địa điểm, sự kiện, chương trình máy tính, thiết bị, công nghệ, phần mềm độc hại/lỗ hổng và tin tặc. Số lượng thực thể tin tặc và phần mềm độc hại/lỗ hổng có số lần xuất hiện tương đối ít so với các nhãn còn lại, tương ứng là 60 và 400 lần xuất hiện.

B. *Thiết lập thực nghiệm*

Hiệu năng của mô hình trích xuất được đo bằng độ đo F_1, được tính từ độ chính xác (precision), độ bao phủ (recall) theo các công thức như sau:

$$Precision = \frac{|A \cap B|}{|A|}$$

$$Recall = \frac{|A \cap B|}{|B|}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Các tham số A và B ở công thức trên tương ứng là tập các thực thể được nhận ra và tập hợp các thực thể đúng (được gán nhãn bởi người gán nhãn), của một loại thực thể cụ thể (ví dụ như *tin tặc*). Thử nghiệm được thực hiện theo phương pháp kiểm tra chéo với dữ liệu được chia thành 5 phần.

Chúng tôi áp dụng cơ chế *mini-batch* để huấn luyện mô hình đề xuất, trong đó: *batch size* là 16; bộ tối ưu Adam optimizer được sử dụng với *learning rate* là $1e^{-5}$, độ dài tối đa của câu đầu vào là 128. Chúng tôi cũng áp dụng cơ chế dừng sớm để ngăn mô hình bị tình trạng quá khớp. Cụ thể, quá trình huấn luyện sẽ dừng khi hiệu suất trên tập dữ liệu kiểm chứng không được cải thiện nào trong ít nhất 5 *epoch* liên tiếp. Số Transformer block là 12, với kích thước của véc-tơ trạng thái ẩn là 768. Mô hình BERT đã đào tạo trước sử dụng cho đa ngôn ngữ multilingual-bert-base. Kĩ thuật WordPiece được sử dụng để tách từ trong câu đầu vào. Ngưỡng xác định cho vị trí bắt đầu và kết thúc của thực thể là không xác định do thực thể tham chiếu có thể ở bất kỳ vị trí nào. Cấu hình của máy tính thử nghiệm có CPU IntelCore i7 9700K, RAM 32 GB, GPU GeForceGTX 2080Ti và đĩa cứng SSD.

C. *Kết quả thực nghiệm*

Phần dưới đây sẽ mô tả các thực nghiệm để đánh giá các đặc trưng quan trọng cũng như hiệu năng của mô hình trích xuất thực thể an toàn thông tin đã đề xuất khi so sánh với các mô hình cơ sở khác.

1) Đánh giá hiệu năng của các kết hợp đặc trưng khác nhau

Chúng tôi đã sử dụng kết hợp một số đặc trưng đã đề xuất cho mô hình để hiểu rõ hơn về đóng góp của từng đặc trưng đối với hiệu suất trích xuất thực thể. Nhiều cách kết

hợp đặc trưng biểu diễn từ khác nhau đã trình bày ở trên được thực hiện, hoặc bỏ lần lượt từng đặc trưng biểu diễn từ trong tổ hợp các đặc trưng, sau đó kết hợp với mạng nơ-ron với các lớp BiLSTM và CRF để trích xuất thực thể.

Bảng I. Hiệu năng của các kết hợp đặc trưng khác nhau trong mô hình BiLSTM-CRF

STT	Đặc trưng	F_1 (%)
1	BERT-CharCNN-FastText	72,86
2	BERT-CharRNN- FastText	72,34
3	CharCNN- FastText	68,41
4	CharRNN- FastText	68,23
5	BERT- FastText	71,93
6	BERT-Glove	71,62
7	BERT-CharCNN-Glove	72,25

Kết quả thử nghiệm trong Bảng I cho thấy, đặc trưng mức ký tự có vai trò quan trọng trong việc trích xuất chính xác thực thể. Hiệu suất của mô hình BiLSTM-CRF dựa trên CharCNN và CharRNN (đặc trưng số 1 và 2) tốt hơn mô hình không có đặc trưng mức ký tự này (đặc trưng số 5 và 6) từ 0,41% đến 1,24%. Khi so sánh giữa hai phương pháp biểu diễn mức ký tự là CharCNN và CharRNN, mô hình dựa trên charCNN (đặc trưng số 1 và số 3) đạt được độ chính xác tốt hơn từ 0,18% tới 0,52% so với các mô hình dựa trên charRNN (đặc trưng số 2 và số 4 trong Bảng I).

Đối với đặc trưng BERT, có thể thấy rằng việc thêm biểu diễn từ sử dụng BERT đã tăng hiệu suất tổng thể đáng kể, từ hơn 68% (không có BERT) lên tới hơn 72% (có BERT). Mức tăng lên tới 4,11% khi xem xét cặp đặc trưng số 2 và số 4; và thậm chí lên tới 4,45% với cặp đặc trưng số 1 và số 3. Mức tăng này hơn hẳn các mức tăng còn lại của các cách kết hợp đặc trưng khác, cho thấy tầm quan trọng của BERT với khả năng biểu diễn từ theo ngữ cảnh thật sự hiệu quả.

Để đánh giá sự phù hợp của phương pháp biểu diễn dựa trên nhúng từ đối với khả năng trích xuất thực thể trong văn bản an toàn thông tin, chúng tôi thay thế FastText bằng Glove, một phương pháp nhúng từ bằng véc-tơ toàn cục [20] (các đặc trưng số 1, 5 so với các đặc trưng số 6, 7 trong Bảng I). Mặc dù chênh lệch không lớn nhưng FastText vẫn thể hiện hiệu năng tốt hơn so với Glove. Điều này chứng tỏ biểu diễn từ kiểu *n*-gram của FastText có lẽ phù hợp với văn bản chuyên ngành an toàn thông tin hơn như đã phân tích trước đó.

Chú ý rằng, mô hình sử dụng cấu trúc mạng nơ-ron tương tự như cấu trúc sử dụng trong [8]. Tuy nhiên so với [8], cách thức kết hợp các đặc trưng như đề xuất trong bài báo giúp tạo nên hiệu năng vượt trội hơn, với mức chênh khoảng 3-4%. Sự khác biệt này có được từ đặc trưng BERT. Kết quả này giúp nhấn mạnh một lần nữa về khả năng biểu diễn chính xác ngữ nghĩa của từ trong câu của BERT.

2) Đánh giá hiệu năng của kiến trúc mạng nơ-ron đề xuất

Để đánh giá hiệu năng của kiến trúc mạng nơ-ron đề xuất BiLSTM-CRF, chúng tôi sẽ so sánh kết quả với mạng chỉ có lớp BiLSTM và bộ phân lớp softmax với mạng chỉ có CRF để phân lớp mà không có BiLSTM. Đầu vào vẫn là đặc trưng kết hợp từ 3 kiểu biểu diễn từ: biểu diễn từ mức ký tự CharCNN, biểu diễn từ được trích xuất từ FastText và đặc trưng BERT. Ngoài ra, chúng tôi cũng đánh giá kết quả của việc bổ sung dữ liệu gán nhãn với mô hình đề xuất.

Bảng II. Hiệu năng của các mô hình với đặc trưng kết hợp

STT	Mô hình	F_1 (%)
1	BERT-CharCNN- FastText **+BiLSTM-CRF**	**72,86**
2	BERT-CharCNN- FastText **+BiLSTM-CRF (có bổ sung dữ liệu)**	**73,61**
3	BERT-CharCNN- FastText **+BiLSTM**	68,53
4	BERT-CharCNN- FastText **+CRF**	65,72

Như thể hiện trong Bảng II, mô hình dựa trên sự kết hợp BiLSTM-CRF được đề xuất vượt trội hơn so với các phương pháp khác trên tập dữ liệu ban đầu với giá trị F_1 là 72,86 %, tốt hơn hơn 7,14% so với mô hình CRF và 4,33% so với mô hình BiLSTM. Kết quả này là hợp lý tương tự kết quả có được từ các nghiên cứu trích xuất thực thể từ văn bản thông thường: mạng nơ-ron kết hợp lớp BiLSTM và lớp CRF thường cho kết quả vượt trội hơn so với các kiến trúc chỉ có các lớp mạng đơn lẻ về hiệu quả phân loại. Mạng BiLSTM có thể nắm bắt các đặc điểm của cả hai hướng bao gồm phần văn bản trước phần văn bản sau trong câu. Phương pháp dựa trên kết hợp BiLSTM-CRF có thế mạnh của cả hai mô hình, mang tới sự cải thiện về độ chính xác phân loại do CRF giúp tận dụng các chuỗi nhãn tốt nhất trong câu đầu vào thay vì chỉ các vị trí riêng lẻ.

Kết quả thử nghiệm mô hình trên tập dữ liệu có bổ sung dữ liệu gán nhãn mới dựa trên phương pháp tăng cường dữ liệu đề xuất trong phần III cho thấy, việc tăng cường dữ liệu có hiệu quả đáng kể với mức tăng lên tới 1,05%, với F_1 từ 72,86% lên tới 73,91%.

V. KẾT LUẬN

Nghiên cứu này đã đề xuất một mô hình học sâu để trích xuất hiệu quả thực thể có tên trong văn bản thuộc lĩnh vực an toàn thông tin với sự kết hợp của ba phương pháp biểu diễn từ gồm BERT, FastText và biểu diễn từ theo mức ký tự dựa trên CNN, cùng với kiến trúc mạng BiLSTM-CRF. Kết quả của nghiên cứu chỉ ra rằng, kết hợp ưu điểm của các phương pháp biểu diễn từ khác nhau gồm: BERT - biểu diễn từ mang thông tin ngữ cảnh trong câu; FastText - đặc trưng phi ngữ cảnh mang thông tin ngữ nghĩa của từ, hỗ trợ tốt các từ mới trong văn bản; và đặc trưng CharCNN - ký tự mang thông tin hình thái, tiền tố và hậu tố của từ, cùng

với mạng học sâu BiLSTM-CRF, tốt hơn so với các mô hình học sâu khác trong bài toán trích xuất thực thể cho dữ liệu văn bản an toàn thông tin. Ngoài ra, phương pháp tăng cường dữ liệu đề xuất cho các thực thể có số lượng hạn chế có hiệu quả đáng kể, với mức tăng hơn 1% so với thử nghiệm trên tập dữ liệu ban đầu.

Trong những nghiên cứu tới, chúng tôi sẽ xem xét mở rộng khai thác về mối quan hệ giữa các thực thể trong văn bản an toàn thông tin dựa trên ý tưởng là sự ràng buộc lẫn nhau giữa quan hệ và thực thể có thể giúp việc nhận dạng thực thể tốt hơn. Ví dụ: quan hệ giữa các thực thể có thể giúp xác định phần mềm bị tấn công và công cụ thực hiện tấn công bằng cách sử dụng thông tin trích xuất từ mô tả lỗ hổng phần mềm.

LỜI CẢM ƠN

Nghiên cứu sinh Nguyễn Thị Thanh Thủy được tài trợ bởi Tập đoàn Vingroup – Công ty CP và hỗ trợ bởi chương trình học bổng đào tạo thạc sĩ, tiến sĩ trong nước của Quỹ Đổi mới sáng tạo Vingroup (VINIF), Viện Nghiên cứu Dữ liệu lớn (VinBigdata), mã số VINIF.2020.TS.94.

TÀI LIỆU THAM KHẢO

[1] Yi, Feng, Bo Jiang, Lu Wang, and Jianjun Wu. "Cybersecurity named entity recognition using multi-modal ensemble learning." IEEE Access 8 (2020): 63214-63224.

[2] Joshi, Arnav, Ravendar Lal, Tim Finin, and Anupam Joshi. "Extracting cybersecurity related linked data from text." In 2013 IEEE Seventh International Conference on Semantic Computing, pp. 252-259. IEEE, 2013.

[3] Lal, Ravendar. "Information Extraction of Security related entities and concepts from unstructured text." (2013): 54.

[4] Deliu, Isuf, Carl Leichter, and Katrin Franke. "Extracting cyber threat intelligence from tin tặc forums: Support vector machines versus convolutional neural networks." In 2017 IEEE International Conference on Big Data (Big Data), pp. 3648-3656. IEEE, 2017.

[5] Ritter, Alan, Evan Wright, William Casey, and Tom Mitchell. "Weakly supervised extraction of computer security events from twitter." In Proceedings of the 24th international conference on world wide web, pp. 896-905. 2015.

[6] Jones, Corinne L., Robert A. Bridges, Kelly MT Huffer, and John R. Goodall. "Towards a relation extraction framework for cyber-security concepts." In Proceedings of the 10th Annual Cyber and Information Security Research Conference, pp. 1-4. 2015.

[7] Bridges, Robert A., Corinne L. Jones, Michael D. Iannacone, Kelly M. Testa, and John R. Goodall. "Automatic labeling for entity extraction in cyber security." arXiv preprint arXiv:1308.4941 (2013).

[8] Gasmi, Houssem, Abdelaziz Bouras, and Jannik Laval. "LSTM recurrent neural networks for cybersecurity named entity recognition." ICSEA 11 (2018): 2018.

[9] Qin, Ya, Guo-wei Shen, Wen-bo Zhao, Yan-ping Chen, Miao Yu, and Xin Jin. "A network security entity recognition method based on feature template and CNN-BiLSTM-CRF." Frontiers of Information Technology & Electronic Engineering 20, no. 6 (2019): 872-884.

[10] Liao, Xiaojing, Kan Yuan, XiaoFeng Wang, Zhou Li, Luyi Xing, and Raheem Beyah. "Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence." In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 755-766. 2016.

[11] Balduccini, Marcello, Sarah Kushner, and Jacquelin Speck. "Ontology-driven data semantics discovery for cyber-security." In International Symposium on Practical Aspects of Declarative Languages, pp. 1-16. Springer, Cham, 2015.

[12] Kim, Gyeongmin, Chanhee Lee, Jaechoon Jo, and Heuiseok Lim. "Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network." International journal of machine learning and cybernetics 11, no. 10 (2020): 2341-2355.

[13] Dai, Xiang, and Heike Adel. "An analysis of simple data augmentation for named entity recognition." arXiv preprint arXiv:2010.11683 (2020).

[14] Sirotina, Anastasiia, and Natalia Loukachevitch. "Named entity recognition in information security domain for russian." In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 1114-1120. 2019.

[15] Mulwad, Varish, Wenjia Li, Anupam Joshi, Tim Finin, and Krishnamurthy Viswanathan. "Extracting information about security vulnerabilities from web text." In 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, vol. 3, pp. 257-260. IEEE, 2011.

[16] More, Sumit, Mary Matthews, Anupam Joshi, and Tim Finin. "A knowledge-based approach to intrusion detection modeling." In 2012 IEEE Symposium on Security and Privacy Workshops, pp. 75-81. IEEE, 2012.

[17] Tikhomirov, Mikhail, et al. "Using bert and augmentation in named entity recognition for cybersecurity domain." International Conference on Applications of Natural Language to Information Systems. Springer, Cham, 2020.

[18] Bojanowski, Piotr, et al. "Enriching word vectors with subword information." Transactions of the Association for Computational Linguistics 5 (2017): 135-146.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[20] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[21] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." 2013 IEEE international conference on acoustics, speech and signal processing. Ieee, 2013.

[22] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

[23] Mikolov, Tomas, et al. "Efficient estimation of word representations in véc-tơ space." arXiv preprint arXiv:1301.3781 (2013).

ENTITY EXTRACTION IN INFORMATION SECURITY USING DEEP LEARNING

Abstract: With the rapid increase of documents related to information security, automatic extraction of important information from these sources is an urgent need. One of the common types of information that needs to be extracted are named entities, such as software programs, hackers, malicious programs, vulnerabilities, technologies, techniques, etc. Due to the complexity, diversity, and unique domain characteristics of these sources, named entity recognition is still facing many

difficulties. Currently, there are a number of approaches to solve this problem, among which are methods based on deep learning, which are the most advanced techniques, widely used in the field of information extraction. In this paper, we present a method to extract named entities in information security using a deep learning technique, which is a combination model including word2vec, BERT, BiLSTM and CRF. In addition, we propose a new form of data augmentation method for entities with a small number in the data set. The results show that the proposed model obtained quite high accuracy, with the F1 score up to 72.86% on information security data set. The proposed method for data augmentation also achieved positive results.

Keywords: information security, entity extraction, BiLSTM, CRF, BERT.

Nguyễn Ngọc Điệp. Nhận học vị Tiến sĩ năm 2017. Hiện đang công tác tại Khoa Công nghệ Thông tin 1 và Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, an toàn thông tin, xử lý ngôn ngữ tự nhiên.

Nguyễn Thị Thanh Thủy. Nhận học vị Thạc sĩ năm 2009. Hiện đang công tác tại Khoa Công nghệ Thông tin 1 và Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, xử lý ngôn ngữ tự nhiên.