Tạp chí Phát triển Khoa học và Công nghệ – Khoa học Tự nhiên, 5(2):1032-1039

Open Access Full Text Article

Bài nghiên cứu

Mô hình học sâu cho bài toán gán nhãn ngữ nghĩa trên văn bản y sinh

Tuấn Nguyên Hoài Đức[1,*], Lê Đình Việt Huy[2], Trần Tiền Lợi Long Tứ[3]

Use your smartphone to scan this QR code and download this article

TÓM TẮT

Chúng tôi xây dựng một mô hình gán nhãn Cấu trúc Đối số Vị ngữ cho văn bản Y Sinh. Cấu trúc Đối số Vị ngữ là thông tin ngữ nghĩa quan trọng của văn bản, do nó chuyển tải sự kiện chính được nói đến trong mỗi câu. Rút trích được Cấu trúc Đối số Vị ngữ trong câu là tiền đề quan trọng để máy tính có thể giải quyết được hàng loạt bài toán khác liên quan đến ngữ nghĩa của văn bản như rút trích sự kiện, rút trích thực thể, hệ hỏi đáp… Cấu trúc Đối số Vị ngữ phụ thuộc vào lĩnh vực của văn bản. Do đó, trong lĩnh vực Y Sinh, văn bản cần xác định khung Đối số Vị ngữ hoàn toàn mới so với lĩnh vực tổng quát. Với đặc thù phải xử lý trên một khung đối số mới, việc xác định bộ đặc trưng cho học máy là khó và đòi hỏi nhiều công sức chuyên gia. Để giải quyết thách thức này, chúng tôi chọn huấn luyện mô hình của mình bằng phương pháp Học sâu (Deep learning) với Mạng nơ ron bộ nhớ ngắn dài hai chiều (Bi-directional Long Short Term Memory). Học sâu là phương pháp học máy không đòi hỏi con người phải xác định bộ đặc trưng một cách thủ công. Ngoài ra, chúng tôi cũng tích hợp kết nối cao tốc (Highway Connection) giữa những tầng nơ ron ẩn không liên tiếp để hạn chế mất mát đạo hàm. Bên cạnh đó, để khắc phục vấn đề ngữ liệu huấn luyện ít, chúng tôi tích hợp Học sâu với kỹ thuật Học đa tác vụ. Học Đa tác vụ giúp cho tác vụ chính (bài toán gán nhãn Cấu trúc Đối số Vị ngữ) được bổ trợ tri thức từ một tác vụ phụ có liên quan mật thiết là bài toán rút trích Thực thể. Mô hình của chúng tôi đạt F1 = 72% mà không cần chuyên gia thiết kế bất kỳ đặc trưng nào, qua đó cho thấy triển vọng của Học sâu trong bài toán này. Ngoài ra, kết quả thực nghiệm cũng cho thấy Học đa tác vụ là kỹ thuật phù hợp để khắc phục vấn đề ngữ liệu huấn luyện ít trong lĩnh vực Y Sinh vì nó cải thiện được độ đo F1.

Từ khoá: cấu trúc đối số vị ngữ, gán nhãn ngữ nghĩa văn bản, học sâu

[1]*Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM, Việt Nam*

[2]*Công ty TNHH Công nghệ ZAMO LLC, Việt Nam*

[3]*Công ty Gameloft Vietnam, Việt Nam*

Liên hệ

Tuấn Nguyên Hoài Đức, Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM, Việt Nam

Email: tnhduc@fit.hcmus.edu.vn

Lịch sử

- Ngày nhận: 18-7-2020
- Ngày chấp nhận: 01-04-2021
- Ngày đăng: 16-04-2021

DOI : 10.32508/stdjns.v5i2.928

Check for updates

VNU-HCM Press

GIỚI THIỆU

Y Sinh (Biomedicine) là ngành khoa học ứng dụng các kỹ thuật Công nghệ Sinh học vào chăm sóc sức khỏe con người. Ngành khoa học này ngày càng khẳng định tiềm năng to lớn của nó trong chẩn đoán và điều trị bệnh [1]. Kho tri thức của lĩnh vực Y Sinh đang được tích lũy và phát triển không ngừng, và phần lớn ở dạng văn bản. Việc khai thác hiệu quả kho tri thức này sẽ giúp ích rất nhiều cho các hoạt động chăm sóc sức khỏe. Tuy nhiên, với khối lượng văn bản đồ sộ vượt trên khả năng khai thác thủ công của con người, việc khai khoáng kho tri thức Y Sinh một cách tự động bằng máy tính là cần thiết.

Để máy tính có thể đọc hiểu văn bản nhằm rút trích tri thức, trước tiên máy tính cần hiểu được nội dung của từng câu trong văn bản. Nội dung của mỗi câu được truyền tải thông qua một động từ chính, gọi là vị ngữ (predicate) và những đối số (argument) có liên quan ngữ nghĩa đến động từ chính. Vì vậy, một trong những bài toán quan trọng nhằm giúp máy tính đọc hiểu văn bản một cách hiệu quả là bài toán rút trích Cấu trúc Đối số Vị ngữ (Predicate Argument Structure – PAS), hay còn được gọi là bài toán Gán nhãn Ngữ nghĩa (Semantic Role Labeling – SRL).

SRL là một bài toán phụ thuộc lĩnh vực (domain dependence). Khi áp dụng vào một lĩnh vực mới như lĩnh vực Y Sinh, việc xác định bộ đặc trưng nào phù hợp để huấn luyện máy tính hiệu quả là một thách thức. Một hướng tiếp cận cho thách thức này là ứng dụng học sâu (deep learning – DL) vì DL có thế mạnh tự đúc kết được bộ đặc trưng phù hợp, giúp tránh việc chuyên gia phải xây dựng thủ công bộ đặc trưng cho một lĩnh vực rất mới [2]. Công trình của chúng tôi nghiên cứu và thử nghiệm một mô hình DL cho bài toán SRL trên văn bản Y Sinh và phân tích, đánh giá kết quả đạt được của mô hình.

CƠ SỞ LÝ THUYẾT VỀ CẤU TRÚC ĐỐI SỐ VỊ NGỮ

Cấu trúc Đối số Vị ngữ (Predicate Argument Structure – PAS) là kết quả của phương pháp phân tích văn bản ở mức ngữ nghĩa sâu. Trong cấu trúc này thì trung tâm là động từ chính, gọi là vị ngữ, xoay quanh vị ngữ là các đối số (bao gồm cả chủ ngữ của câu). Mỗi đối số đều có một vai trò ngữ nghĩa cụ thể (semantic role).

Trích dẫn bài báo này: Đức T N H, Huy L D V, Tứ T T L L. **Mô hình học sâu cho bài toán gán nhãn ngữ nghĩa trên văn bản y sinh**. *Sci. Tech. Dev. J. - Nat. Sci.;* 5(2):1032-1039.

Thí dụ: Xét câu "Tôi thuê căn phòng của bạn một tháng", câu này có PAS gồm vị ngữ là "thuê" và bốn đối số xoay quanh vị ngữ là:
Đối số 0: "Tôi" (vai trò ngữ nghĩa: Người thuê).
Đối số 1: "Phòng" (vai trò ngữ nghĩa: Vật được thuê).
Đối số 2: "Bạn" (vai trò ngữ nghĩa: Người cho thuê).
Đối số 3: "Một tháng" (vai trò ngữ nghĩa: Thời hạn thuê).
Có nhiều bộ ngữ liệu PAS được xây dựng cho lĩnh vực tổng quát như FrameNet, VerbNet và PropBank[3–5]. Trong đó, PropBank định nghĩa bộ đối số chi tiết nhất cho từng vị ngữ.
PAS trong lĩnh vực Y Sinh có nhiều khác biệt so với PAS trong lĩnh vực tổng quát, bao gồm khác biệt về ý nghĩa của vị ngữ, cũng như là khác biệt về thành phần đối số. Thí dụ: Xét vị ngữ "mutate". Trong Y Sinh, "mutate" có nghĩa là "đột biến" với 4 đối số là: (1) Vị trí exon hoặc nitron xảy ra đột biến, (2) Gene bị đột biến, (3) Hậu quả về kiểu gene, (4) Hậu quả về kiểu hình. Trong khi đó, ở lĩnh vực tổng quát thì "mutate" có nghĩa là "thay đổi" với chỉ 2 đối số là: (1) Tác nhân gây thay đổi, (2) đối tượng bị thay đổi.
Nhận thấy những khác biệt đó, nhiều công trình đã xây dựng những bộ ngữ liệu PAS riêng cho lĩnh vực Y Sinh. Mỗi công trình đều chọn ra những vị ngữ có ý nghĩa quan trọng trong văn bản Y Sinh, là những động từ thường truyền tải các sự kiện Y Sinh quan trọng (như đột biến, mã hóa, giải mã, biểu hiện…), để định nghĩa khung đối số cụ thể cho từng vị ngữ ấy. Các công trình xây dựng những bộ ngữ liệu PAS Y Sinh được biết đến nhiều nhất bao gồm BioProp, PasBIO và GREC.

- BioProp là bộ ngữ liệu bao gồm 1635 câu trích dẫn từ phần tóm tắt (abstract) của 500 bài báo Y Sinh[6]. Hạn chế của Bioprop là vay mượn hoàn toàn bộ đối số từ PropBank, một bộ ngữ liệu của lĩnh vực tổng quát. Do đó, các bộ đối số của BioProp chưa thực sự được chuyên biệt hóa vào lĩnh vực Y Sinh.
- PasBio khắc phục hạn chế của BioProp bằng cách định nghĩa lại toàn bộ các khung đối số cho phù hợp với lĩnh vực Y Sinh[7]. Nhưng hạn chế của công trình này là chưa đầu tư gắn nhãn lại các đối số ấy vào bộ ngữ liệu. Kết quả là bộ ngữ liệu thực sự được gán nhãn các đối số theo định nghĩa của PasBIO chỉ vỏn vẹn hơn 200 câu. Kích thước này là quá nhỏ để dùng trong học máy.
- GREC là bộ ngữ liệu bao gồm 1489 câu trích dẫn từ phần tóm tắt của 677 bài báo Y Sinh[8]. Trong GREC, vị ngữ không chỉ bao gồm động từ chính mà còn bao gồm cả các danh động, với bộ đối số được định nghĩa chuyên biệt cho lĩnh vực Y Sinh. Vì vậy, GREC khắc phục được hạn chế của BioProp, và cũng khắc phục được hạn chế về kích thước ngữ liệu của PasBIO.

NHỮNG NGHIÊN CỨU VỀ SRL

Gán nhãn ngữ nghĩa (Semantic Role Labeling - SRL) là bài toán tự động nhận diện vị ngữ cùng các đối số của nó trong văn bản và gắn nhãn vai trò ngữ nghĩa (gọi tắt là nhãn ngữ nghĩa) cho từng đối số. Vì vậy, SRL còn được gọi là bài toán rút trích PAS. Các nghiên cứu về SRL chia ra 3 hướng tiếp cận: Hướng dựa trên luật, hướng khớp mẫu và hướng học máy trong đó có học sâu.

Hướng dựa trên luật

Hướng dựa trên luật là hướng tiếp cận sớm nhất, sử dụng bộ luật viết thủ công bởi chuyên gia để nhận biết vị ngữ, đối số trong văn bản thô và quyết định nhãn ngữ nghĩa cho đối số. Những công trình tiêu biểu trong hướng này có thể kể đến như: Thuyết Ngữ pháp Cấu trúc Dẫn xuất Đầu tố ngữ (Head-Driven Phrase Structure Grammar - HPSG)[9]; Mô hình khai thác Penn Treebank trong việc dựng luật và khắc phục các trường hợp đối số rỗng, như câu khuyết túc từ hoặc chủ ngữ ngầm định[10,11]; Hệ thống cơ sở của CoNLL 2004 và CoNLL 2005 với một tập luật heuristic để xử lý SRL[12,13]. Ngoài ra, một số công trình tuy tiếp cận theo hướng học máy nhưng vẫn dùng luật heuristic như một giải pháp tinh chỉnh kết quả xử lý[14,15].
Trong lĩnh vực Y Sinh, nhiều công cụ SRL được xây dựng cũng vận dụng bộ luật heuristic dựa trên cây cú pháp để rút trích PAS từ văn bản Y Sinh[16–18]. Mặt hạn chế của hướng dựa trên luật là cần có chuyên gia xây dựng thủ công bộ luật. Chỉ cần chuyển sang một lĩnh vực khác hoặc một ngôn ngữ khác thì lại phải xây dựng lại từ đầu một bộ luật mới. Hơn nữa, bộ luật mà chuyên gia xây dựng cũng không thể nào phủ hết mọi cấu trúc ngữ pháp có thể xuất hiện trong văn bản. Do đó, hướng này tuy cho độ chính xác cao nhưng độ bao phủ lại không cao. Tuy nhiên, ưu điểm của hướng dựa trên luật là nó phù hợp với những ngôn ngữ hoặc lĩnh vực có ít tài nguyên ngôn ngữ, nơi mà kích thước ngữ liệu không đủ để huấn luyện máy tính theo hướng học máy (như các công trình SRL cho tiếng Hà Lan và tiếng Nhật[19,20]).

Hướng khớp mẫu

Hướng khớp mẫu sử dụng các mẫu được định nghĩa sẵn để so khớp vào văn bản, từ đó rút trích được vị ngữ và các đối số kèm theo vai trò ngữ nghĩa của chúng. Trong lĩnh vực tổng quát, ở hầu hết các công trình, bộ mẫu có được là do khai khoáng từ ngữ liệu[21–23]. Hạn

chế của việc khai khoáng bộ mẫu từ ngữ liệu là khó kiểm soát các mẫu thu được do độ nhiễu cao. Vì thế, hướng này vẫn không tránh khỏi phải có sự can thiệp thủ công để rà soát lại bộ mẫu. Trong lĩnh vực Y Sinh, do hạn chế vì kích thước ngữ liệu nên các công trình cần có chuyên gia xây dựng thủ công bộ mẫu [16,24].

Cũng như hướng dựa trên luật, hướng khớp mẫu chỉ phù hợp với những lĩnh vực hoặc những ngôn ngữ hạn chế về kích thước ngữ liệu. Khi kích thước ngữ liệu đủ lớn, hướng học máy vẫn là giải pháp được lựa chọn hàng đầu.

Hướng học máy

Hướng học máy là hướng tiếp cận mới hơn hai hướng kể trên (trong đó, học sâu là kỹ thuật mới nhất). Hướng học máy huấn luyện máy tính thông qua một quá trình học, có thể là học có giám sát, bán giám sát hoặc không giám sát, để sau đó máy tính có thể tự nó thực hiện SRL.

Học máy có giám sát sử dụng bộ ngữ liệu có kích thước đủ lớn đã gán nhãn ngữ nghĩa sẵn để huấn luyện máy tính (như Penn TreeBank; PropBank; FrameNet) [25–29]. Trong lĩnh vực Y Sinh, BIOSMILE là công trình SRL hoàn chỉnh đầu tiên, được huấn luyện bằng MaxEnt trên bộ ngữ liệu BioProp [30].

Thách thức của học máy có giám sát là việc xây dựng bộ ngữ liệu gán nhãn sẵn rất công phu, đòi hỏi thời gian và chi phí. Từ đó, nhiều công trình đã đề xuất các mô hình học máy bán giám sát cho bài toán SRL [31,32]. Trong đó, các cấu trúc PAS được rút trích bằng việc lặp đi lặp lại quá trình tuyển chọn ứng viên trên dữ liệu thô, bắt đầu từ một ít PAS làm ứng viên hạt giống. Các mô hình này không đòi hỏi nhiều ngữ liệu gán nhãn sẵn nên thuận lợi khi chuyển sang ngôn ngữ hoặc lĩnh vực mới, nhưng do tính phân kỳ của các cấu trúc ứng viên nên độ chính xác thấp hơn học máy có giám sát. Đối với học máy, bộ đặc trưng đóng vai trò quan trọng. Hầu hết công trình đều tập trung vào việc tinh chỉnh, bổ sung đặc trưng để cải thiện kết quả của công trình trước đó. Việc chọn đặc trưng gì cho từng lĩnh vực hoặc từng ngôn ngữ cụ thể là do chuyên gia quyết định. Đây là một thách thức đối với lĩnh vực Y Sinh vì các lý do sau:

- Bài toán SRL là phụ thuộc lĩnh vực nên các bộ đặc trưng đã được nghiên cứu trong lĩnh vực tổng quát không thể áp dụng rập khuôn cho lĩnh vực Y Sinh.
- Khó mà quyết định đặc trưng gì là hiệu quả do: (i) Một đối số Y Sinh có nhiều biến thể, (ii) PAS trong Y Sinh xuất hiện trong nhiều cấu trúc ngữ pháp phong phú phức tạp; (iii) Vai trò ngữ nghĩa trong Y Sinh có độ nhập nhằng cao (cùng một danh từ có thể giữ nhiều vai trò ngữ nghĩa).

Vì vậy, công trình của chúng tôi chọn thử nghiệm mô hình học sâu (deep learning) vào bài toán SRL cho văn bản Y Sinh vì thế mạnh của học sâu là không cần xác định thủ công bộ đặc trưng.

PHƯƠNG PHÁP THỰC HIỆN

Mô hình mạng nơ ron mà chúng tôi lựa chọn là Mạng nơ ron bộ nhớ ngắn dài hai chiều (Bi-directional Long Short Term Memory – gọi tắt là Bi-LSTM).

Mạng nơ ron hồi quy truyền thống không giải quyết được vấn đề phụ thuộc xa, một vấn đề quan trọng trong xử lý ngôn ngữ tự nhiên [33]. Do đó, mạng nơ ron bộ nhớ ngắn dài (LSTM) là lựa chọn hợp lý vì nó khắc phục được hạn chế này của mạng hồi quy truyền thống [34]. Mạng LSTM mô phỏng tế bào bộ nhớ con người với các cổng thông tin vào ra. Thông qua các cổng này, tế bào sẽ quyết định thông tin nào được ghi nhớ để phục vụ xử lý.

Tuy nhiên, các tế bào của LSTM chỉ liên kết theo một chiều, một thông tin chỉ được xử lý dựa trên dữ kiện từ các thông tin trước nó. Trong khi đó, mỗi một từ trong văn bản có liên hệ ngữ nghĩa mật thiết với không chỉ những từ trước nó mà cả những từ sau nó. Một cải tiến của mạng LSTM là mạng LSTM hai chiều (Bi-LSTM) đã khắc phục vấn đề này, cho phép xử lý thông tin dựa trên những dữ kiện đi trước và đi sau nó [35]. Bi-LSTM đã được chọn sử dụng trong các nghiên cứu gần đây về SRL [36–38].

Bên cạnh đó, chúng tôi vận dụng Kết nối Cao tốc (Highway Connection – HC), một cải tiến cho mạng BiLSTM được đề xuất cho bài toán SRL [38]. HC là những kết nối thông tầng giữa hai tầng tế bào không liên tiếp, tạo nên sự liên kết không những là 2 chiều mà còn là xuyên tầng giữa các tế bào trong mạng nơ ron, giúp hoạt động học của mạng nơ ron linh hoạt và thông minh hơn. Hệ thống SRL cho văn bản trong lĩnh vực tổng quát được huấn luyện bằng mạng Bi-LSTM có HC đã cho kết quả cao nhất (state-of-the-art) với F1 = 83,2% [38]. Vì vậy, mô hình Bi-LSTM có HC cũng hứa hẹn triển vọng cho SRL trên văn bản Y Sinh.

Kết hợp tất cả những đề xuất nêu trên, mô hình của chúng tôi vẫn còn một thách thức phải quan tâm: kích thước ngữ liệu huấn luyện trong Y sinh rất hạn chế so với lĩnh vực tổng quát (Bộ ngữ liệu GREC gồm 1489 câu). Vì vậy, chúng tôi tích hợp kỹ thuật học đa tác vụ vào mô hình của mình. Học đa tác vụ (Multi-Task Learning) là thuật toán học máy, cho phép huấn luyện các tác vụ có liên quan với nhau trên cùng một mô hình và dữ liệu để bổ trợ nhau. Việc tận dụng kiến thức của những tác vụ liên quan sẽ giúp cải thiện đáng kể hiệu quả của tác vụ chính. Học đa tác vụ được đề xuất cho bài toán SRL khi xử lý trên văn bản tiếng

Indonesia trong lĩnh vực tổng quát với dữ liệu huấn luyện ít và cho thấy F1 được cải thiện 8%[36]. Khi áp dụng vào văn bản Y Sinh, chúng tôi nhận thấy bài toán SRL có liên quan mật thiết với bài toán Rút trích Thực thể (Named Entity Recognition – NER), vì loại thực thể của đối số quyết định vai trò ngữ nghĩa của đối số (Ví dụ loại thực thể DNA chỉ có thể giữ vai trò "tác nhân" của vị ngữ "encode" chứ không thể giữ vai trò "sản phẩm"). Vì vậy, chúng tôi chọn bài toán NER là tác vụ phụ trong mô hình học đa tác vụ của mình để bổ trợ cho tác vụ chính là SRL.

Ngoài ra, DL kết hợp với học máy truyền thống sẽ cho kết quả tốt hơn từng kỹ thuật riêng lẻ[39]. Do đó, trong mô hình của chúng tôi, tầng đầu ra của mạng nơ ron được phân loại một lần nữa bởi mô hình học máy truyền thống là CRF và Softmax (Hình 1). Kết quả thực nghiệm của mô hình sẽ được phân tích trong mục Kết quả thực nghiệm.

KẾT QUẢ THỬ NGHIỆM VÀ THẢO LUẬN

Bộ ngữ liệu được sử dụng để huấn luyện và đánh giá là GREC, được xây dựng bởi trung tâm Text Mining (NaCTeM), Khoa Khoa học Máy tính, Trường Đại học Manchester, Anh quốc[8]. Ưu điểm của GREC là các vị ngữ của câu không chỉ có động từ mà còn bao gồm cả danh động nên độ phủ cao, với 4770 vị ngữ. Đồng thời GREC còn gán nhãn thực thể nên rất thuận lợi cho học đa tác vụ. Về phương pháp đánh giá, chúng tôi dùng phương pháp đánh giá chéo 10 pha (10-fold cross validation)

Chúng tôi thử nghiệm và so sánh kết quả của mô hình với ba mức biểu diễn là chỉ có mức từ (word embedding), chỉ có mức ký tự (character embedding) và mức từ kết hợp với mức ký tự với những số chiều vector khác nhau. Chúng tôi cũng so sánh hiệu quả của Softmax và CRF ở tầng đầu ra, cũng như so sánh hiệu quả của mô hình khi không có học đa tác vụ (Bảng 1) và khi có học đa tác vụ (Bảng 2).

Từ kết quả thử nghiệm cho thấy:

- F1 cao nhất của mô hình có học đa tác vụ cao hơn F1 cao nhất của mô hình học đơn tác vụ 5.14%, đây là một khoảng cách đáng kể. Điều này củng cố giả thiết của chúng tôi về hiệu quả tích cực của học đa tác vụ cũng như việc lựa chọn tác vụ phụ là NER đối với SRL cho văn bản Y Sinh.
- Ở cả mô hình học đơn tác vụ và đa tác vụ đều cho thấy việc tăng số chiều vector không nâng cao F1 đáng kể bằng việc chia mịn mức biểu diễn, từ mức từ thành mức ký tự.
- Ở cả mô hình học đơn tác vụ và đa tác vụ đều cho thấy ở tầng đầu ra của tác vụ SRL, CRF là phù hợp hơn so với Softmax.

KẾT LUẬN

Chúng tôi đã xây dựng được mô hình học sâu cho bài toán SRL trên văn bản Y Sinh với một dữ liệu huấn luyện có kích thước hạn chế. Mô hình của chúng tôi trong lĩnh vực Y Sinh đạt F1 = 72% với chỉ 1389 câu trong dữ liệu huấn luyện. Kết quả này không cách quá xa so với kết quả cao nhất trong lĩnh vực tổng quát đạt F1 = 77% trong công trình tiên phong đề xuất học đa tác vụ cho bài toán SRL với dữ liệu huấn luyện hơn 6000 câu[36]. Kết quả thử nghiệm của chúng tôi cho thấy mô hình học đa tác vụ cũng phù hợp với SRL trong lĩnh vực Y Sinh, một lĩnh vực còn hạn chế về kích thước ngữ liệu gán nhãn sẵn.

Hướng phát triển của chúng tôi là thử nghiệm kết hợp tri thức ngữ pháp với mức biểu diễn từ và ký tự để làm giàu đặc trưng cho mô hình. Đồng thời, chúng tôi sẽ kết hợp kỹ thuật học chủ động (Active learning) với học đa tác vụ để nâng cao hiệu quả của mô hình. Ngoài ra, chúng tôi sẽ nghiên cứu ứng dụng transfer learning từ một mô hình pre-trained để hỗ trợ mô hình học sâu khi tập dữ liệu huấn luyện có kích thước hạn chế.

LỜI CẢM ƠN

Nghiên cứu được tài trợ bởi Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM trong khuôn khổ Đề tài mã số CNTT 2020-13

DANH MỤC TỪ VIẾT TẮT

PAS: Cấu trúc Đối số Vị ngữ (Predicate Argument Structure)

SRL: Gán nhãn Ngữ nghĩa (Semantic Role Labelling)

NER: Gán nhãn thực thể (Named Entity Recognition)

DL : Ứng dụng học sâu (Deep Learning)

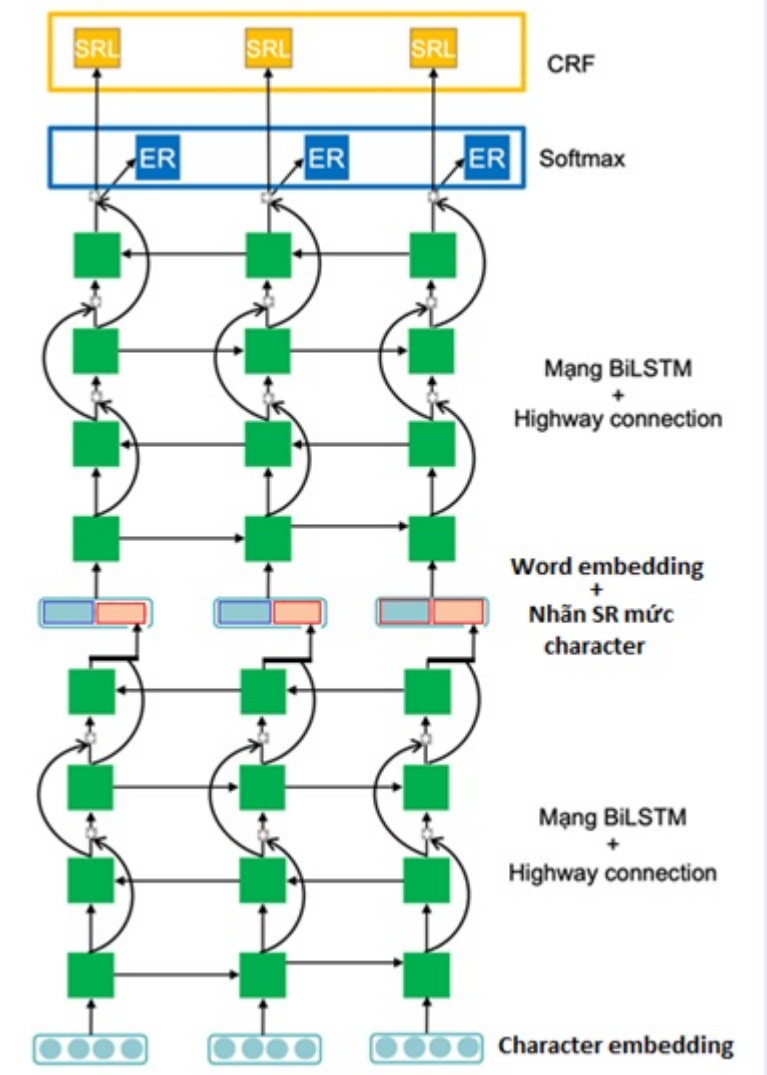
XUNG ĐỘT LỢI ÍCH TÁC GIẢ

Các tác giả tuyên bố rằng họ không có xung đột lợi ích.

ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Tuấn Nguyên Hoài Đức chủ trì đề tài, tiến hành khảo sát hiện trạng, thu thập dữ liệu, phân tích đánh giá giải pháp và viết bài.

Lê Đình Việt Huy và Trần Tiến Lợi Long Tứ tham gia khảo sát hiện trạng, đề xuất giải pháp và lập trình thử nghiệm.



Hình 1: Kiến trúc của mô hình được đề xuất

Bảng 1: Kết quả thực nghiệm với mô hình đơn tác vụ

STT	Số chiều vector	Mức biểu diễn	Lớp đầu ra	P	R	F1
1	100	Word	Softmax	64,12	58,01	60,91
2	100	Word	CRF	67,95	56,13	61,48
3	100	Char	CRF	67,81	63,3	65,48
4	300	Char	CRF	68,62	63,55	65,98
5	100	Word+Char	CRF	72,21	66,34	69,15
6	300	Word+Char	CRF	73,36	66,93	69,99

Bảng 2: Kết quả thực nghiệm với mô hình đa tác vụ.

STT	Số chiều vector	Mức biểu diễn	Lớp đầu ra cho SRL	Lớp đầu ra cho NER	Kết quả SRL		
						R	F1
1	100	Word	Softmax	CRF	68,93	64,31	66,54
2	100	Word	CRF	CRF	69,27	64,97	67,05
3	100	Word	CRF	Softmax	70,04	67,74	68,87
4	100	Char	CRF	Softmax	73,29	67,97	70,53
5	300	Char	CRF	Softmax	74,57	67,90	72,08
6	100	Word+Char	CRF	Softmax	78,03	70,97	74,33
7	300	Word+Char	CRF	Softmax	78,86	71,74	75,13

TÀI LIỆU THAM KHẢO

1. Enderle JD, et al. Introduction to Biomedical Engineering, Academic Press. 2012;p. 16–21.
2. Schmidhuber J. Deep Learning in Neural Networks: An Overview, Neural Networks. 2015;61:85–117. PMID: 25462637. Available from: https://doi.org/10.1016/j.neunet.2014.09.003.
3. Johnson CR, et al. The FrameNet project: Tools for lexicon building, International Conference on Computational Linguistics. 1998;17:86–90.
4. Kipper K, et al. Class-based construction of a verb lexicon, AAAI-2000. 2000;(2000):691–696.
5. Kingsbury P, Palmer M. From Treebank to PropBank, International Conference on Language Resources and Evaluation. 2002;12:38–43.
6. Chou WC, et al. A semi-automatic method for annotating a biomedical proposition bank, The workshop on frontiers in linguistically annotated corpora. 2006;p. 5–12.
7. Wattarujeekrit T, et al. PASBio: predicate-argument structures for event extraction in molecular biology, BMC Bioinformatics. 2004;5:155–163. PMID: 15494078. Available from: https://doi.org/10.1186/1471-2105-5-155.
8. Thompson P, Cotter P, McNaught J, et al. Building a Bio-Event Annotated Corpus for the Acquisition of Semantic Frames from Biomedical Corpora. LREC. 2008;.
9. Pollard C, Sag IA. Head-Driven Phrase Structure Grammar. IL: Univ. of Chicago Press. 1994;.
10. Liakata M, et al. From Trees To Predicate-Argument Structures, International Conference on Computational Linguistics. 2002;20:563–569. Available from: https://doi.org/10.3115/1072228.1072333.
11. Marcus M, et al. The Penn Treebank: Annotating Predicate Argument Structure, The Human Language Technology Workshop. Plainsboro, NJ, 114119. 1994;Available from: https://doi.org/10.3115/1075812.1075835.
12. Carreras X, Màrquez L. Introduction To the CoNLL-2005 shared task: Semantic role labeling, CoNLL. 2005;p. 152–164. Available from: https://doi.org/10.3115/1706543.1706571.
13. Carreras X, Màrquez L. Introduction to the CoNLL-2004 shared task: Semantic role labeling, HLT-NAACL 2004 Workshop 8th Conf. Comput. Natural Lang. Learn. 2004;p. 89–97.
14. Park KM, et al. Two-phase semantic role labeling based on support vector machines, CoNLL. 2004;.
15. Surdeanu M, et al. Semantic role labeling using complete syntactic analysis, CoNLL. 2005;p. 67–72. Available from: https://doi.org/10.3115/1706543.1706586.
16. Chi-San (Althon) Lin, Tony C. Smith, Semantic role labeling via consensus in pattern-matching, CONLL. 2005;5:185–188.
17. Grenager T, et al. Manning, Unsupervised Discovery of a Statistical Verb Lexicon. EMNLP. 2007;06:1–8. Available from: https://doi.org/10.3115/1610075.1610077.
18. Wattarujeekrit T. Exploring Predicate-Argument Relations for Named Entity Recognition in the Molecular Biology Domain, International Conference on Discovery Science. 2005;8:267–280. Available from: https://doi.org/10.1007/11563983_23.
19. Stevens G. XARA: An XML- and rule-based semantic role labeler, The Linguistic Annotation Workshop, Annual Meeting of the Association for Computational Linguistics. 2007;45. PMID: https://doi.org/10.3115/1642059.1642077. Available from: 26110305.
20. Iida R, et al. Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations, the Linguistic Annotation Workshop. 2007;p. 132–139. Available from: https://doi.org/10.3115/1642059.1642081.
21. Riloff E. Automatically Generating Extraction Patterns from Untagged Text, National Conference on Artificial Intelligence. 1996;19:1044–1049.
22. Riloff E. An empirical approach to conceptual case frame acquisition, The Workshop on Very Large Corpora. 1998;6:49–56.
23. Riloff E. Automatically constructing a dictionary for information extraction tasks, National Conference on Artificial Intelligence (AAAI). 1993;1:811–816.
24. Huang M. Discovering patterns to extract protein-protein interactions from full texts. Bioinformatics. 2004;p. 3604–3612. PMID: 15284092. Available from: https://doi.org/10.1093/bioinformatics/bth451.
25. Blaheta D, Charniak E. Assigning function tags to parsed text, the Annual Meeting of the North American Chapter of the ACL. 2000;1:234–240.
26. Gildea D, Jurafsky D. Automatic labeling of semantic roles, Computational Linguistics. 2002;p. 245–288. Available from: https://doi.org/10.1162/089120102760275983.
27. Gildea D, Palmer M. The necessity of parsing for predicate argument recognition, Meeting of the Association for Computational Linguistics. 2002;40:239–246. Available from: https://doi.org/10.3115/1073083.1073124.
28. Surdeanu M, Harabagiu S, et al. Using Predicate-Argument Structure for Information Extraction, Annual Conference on the Association for Computational Linguistics. 2013;41:46–51.
29. Kingsbury P, Palmer M, Marcus M. Adding Semantic Annotation to the Penn TreeBank, The Human Language Technology Conference. 2002;p. 252–256.
30. Tsai RTH, et al. BIOSMILE: Adapting Semantic Role Labeling for Biomedical Verbs: An Exponential Model Coupled with Automatically Generated Template Features, BioNLP Conference. 2006;.
31. Swier RS, Stevenson S. Exploiting a Verb Lexicon in Automatic Semantic Role Labelling, EMNLP. 2005;05:883–890.

32. Swier RS, Stevenson S. Unsupervised Semantic Role Labeling, EMNLP. 2004;04:95–102.
33. Bengio Y, Simard P. Problem of learning long-term Dependencies in Recurrent Network, IEEE Transactions on Neural Networks archive. 1994;5:157–166. PMID: 18267787. Available from: https://doi.org/10.1109/72.279181.
34. Hochreiter S. Long-Short Term Memory, Neural Computation Archive. 1997;9:1735–1780. PMID: 9377276. Available from: https://doi.org/10.1162/neco.1997.9.8.1735.
35. Graves A, rahman Mohamed A, Hinton G. Speech Recognition with Deep Recurrent Neural networks, 1988. ICASSP. 2013;88:90–95. Available from: https://doi.org/10.1109/ICASSP.2013.6638947.
36. Ikhwantri F, et al. Multi-Task Active Learning for Neural Semantic Role Labeling on Low Resource Conversational Corpus, Workshop on Deep Learning Approaches for Low-Resource NLP. 2018;p. 43–50. Available from: https://doi.org/10.18653/v1/W18-3406.
37. Zhou J, Xu W. End-to-end Learning of Semantic Role Labeling Using Recurrent Neural Networks, Annual Meeting of the Association for Computational Linguistics 53 - International Joint Conference on Natural Language Processing. 2015;7:1127–1137. Available from: https://doi.org/10.3115/v1/P15-1109.
38. He L, et al. Deep Semantic Role Labeling: What Works and What's Next, Annual Meeting of the Association for Computational Linguistics. 2017;55:473–483. Available from: https://doi.org/10.18653/v1/P17-1044.
39. Bethard YV. A survey on recent advances in named entity recognition from deep learning models, International Conference on Computational Linguistics. 2018;27:2145 –2158.

Tạp chí Phát triển Khoa học và Công nghệ – Natural Sciences, 5(2):1032-1039

Open Access Full Text Article

Research article

A deep-learning model for semantic role labelling in medical documents

Tuan Nguyen Hoai Duc[1,*], Le Dinh Viet Huy[2], Tran Tien Loi Long Tu[3]

Use your smartphone to scan this QR code and download this article

TÓM TẮT

We built a model labelling the Predicate Argument Structure (PAS) for biomedical documents. PAS is an important semantic information of any document, because it reveals the main event mentioned in each sentence. Extracting PAS in a sentence is an important premise for the computer to solve a series of other problems related to the semantics in text such as event extraction, named entity extraction, question answering system… The predicate argument structure is domain dependent. Therefore, in Biomedical field, it is required to define a completely new Predicate Argument frame compared to the general field. For a machine learning model to work well with a new argument frame, identifying a new feature set is required. This is difficult, manual and requires a lot of expert labor. To address this challenge, we chose to train our model with Deep Learning method utilizing Bi-directional Long Short Term Memory. Deep learning is a machine learning method that does not require defining the feature sets manually. In addition, we also integrate Highway Connection between hidden neuron layers to minimize derivative loss. Besides, to overcome the problem of small training corpus, we integrate Deep Learning with Multi-task Learning technique. Multi-task Learning helps the main task (PAS tagging) to be complemented with knowledge learnt from a closely related task, the NER. Our model achieved F1 = 75.13% without any manually designed feature, thereby showing the prospect of Deep Learning in this domain. In addition, the experiment results also show that Multi-task Learning is an appropriate technique to overcome the problem of little training data in biomedical fields, by improving the F1 score.

Từ khoá: predicate argument structure, semantic role labelling, deep learning

[1]*Faculty of Information Technology, University of Sciences, VNU-HCM, Vietnam.*

[2]*ZAMO LLC Technology Ltd. Company, Vietnam.*

[3]*Gameloft Vietnam Company, Vietnam.*

Liên hệ

Tuan Nguyen Hoai Duc, Faculty of Information Technology, University of Sciences, VNU-HCM, Vietnam.

Email: tnhduc@fit.hcmus.edu.vn

Lịch sử

- Ngày nhận: 18-7-2020
- Ngày chấp nhận: 01-4-2021
- Ngày đăng: 16-4-2021

DOI : 10.32508/stdjns.v5i2.928

Check for updates

Trích dẫn bài báo này: Duc T N H, Huy L D V, Tu T T L L. **A deep-learning model for semantic role labelling in medical documents**. *Sci. Tech. Dev. J. - Nat. Sci.;* 5(2):1032-1039.