Tạp chí Phát triển Khoa học và Công nghệ – Khoa học Tự nhiên 2023, 7(4):2750-2762

Open Access Full Text Article

Bài nghiên cứu

Mô hình học sâu nhúng cú pháp cho tác vụ gán nhãn ngữ nghĩa văn bản y sinh

Tuấn Nguyên Hoài Đức[1,2,*], Lưu Trường Dương[3], Huỳnh Quốc Duy[3]

TÓM TẮT

Bài báo trình bày việc xây dựng một mô hình học sâu cho tác vụ gán nhãn ngữ nghĩa trên văn bản Y Sinh. Gán nhãn ngữ nghĩa là tác vụ thiết thực giúp máy tính hiểu rõ sự kiện chính được nói đến trong mỗi câu, và là tiền đề quan trọng để giải quyết được hàng loạt tác vụ khác liên quan đến ngữ nghĩa như rút trích sự kiện, rút trích thực thể, hệ hỏi đáp... Gán nhãn ngữ nghĩa là tác vụ phụ thuộc lĩnh vực. Trong lĩnh vực Y Sinh, ngữ nghĩa không chỉ dựa trên khung đối số vị ngữ khác biệt rất nhiều so với lĩnh vực tổng quát mà còn được chuyển tải bằng những cấu trúc ngữ pháp và quan hệ phụ thuộc phức tạp hơn. Ba loại tri thức được tích hợp vào mô hình học sâu để thích ứng với đặc thù này, gồm có tri thức ngữ cảnh có được từ mô hình ngôn ngữ tiền huấn luyện trên ngữ liệu lớn của chuyên ngành Y Sinh, tri thức về phụ thuộc giữa các từ có được từ cây phân tích quan hệ phụ thuộc và tri thức về cấu trúc ngữ pháp câu có được từ cây phân tích ngữ pháp thành phần. Để xử lý hiệu quả các loại tri thức ngữ pháp vốn dĩ được biểu diễn như đồ thị, Mạng Đồ thị Chú ý (Graph Attention Network) được chọn để phát huy thế mạnh học trên đồ thị của nó. Ngoài ra, nhúng chỉ định vị ngữ (Predicate Indicator Embedding) cũng được tích hợp để góp phần nâng cao hiệu quả mô hình. Kết quả thực nghiệm cho thấy hai loại tri thức cú pháp nêu trên kết hợp với nhúng chỉ định vị ngữ có thể giúp F1 tăng đến 20%.

Từ khoá: cấu trúc đối số vị ngữ, gán nhãn ngữ nghĩa văn bản, học sâu, ngữ pháp thành phần, quan hệ phụ thuộc, văn bản Y Sinh

[1]*Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên TP. Hồ Chí Minh, Việt Nam*

[2]*Đại học Quốc gia TP. Hồ Chí Minh, Việt Nam*

[3]*Công ty TNHH Antsomi Việt Nam, Việt Nam*

Liên hệ

Tuấn Nguyên Hoài Đức, Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên TP. Hồ Chí Minh, Việt Nam

Đại học Quốc gia TP. Hồ Chí Minh, Việt Nam

Email: tnhduc@fit.hcmus.edu.vn

Lịch sử

- Ngày nhận: 10-4-2023
- Ngày chấp nhận: 9-10-2023
- Ngày đăng: 31-12-2023

DOI :

https://doi.org/10.32508/stdjns.v7i4.1279

GIỚI THIỆU

Ngành Y Sinh (Biomedicine) đã xác định được thế mạnh là chăm sóc sức khỏe con người [1,2]. Vai trò của Y Sinh càng thể hiện rõ hơn trong đại dịch Covid-19, khi mà những nghiên cứu về Sinh học Phân tử, nhất là về vật chất di truyền, đóng vai trò quan trọng. Vì vậy, ngành khoa học này đã thu hút nhiều nghiên cứu, và vì thế kho tri thức Y Sinh càng được tích lũy nhiều đến mức đã vượt quá khả năng khai thác thủ công của con người [3]. Việc khai thác kho văn bản to lớn này, thí dụ như kho văn bản của cơ sở dữ liệu MEDLINE, bằng sức mạnh điện toán mở ra nhiều triển vọng khai phá hiệu quả tri thức trong ấy để giúp ích trong chẩn đoán và điều trị bệnh [4,5].

Để làm được điều này, trước tiên máy tính phải hiểu được từng sự kiện được nói đến trong mỗi câu. Các sự kiện này được chuyển tải thông qua một cấu trúc bao gồm động từ chính trong câu, gọi là vị ngữ (predicate) và tất cả các đối tượng xoay quanh động từ này trong câu, gọi là các đối số (argument). Toàn bộ cấu trúc này gọi là *cấu trúc đối số vị ngữ* (Predicate Argument Structure – PAS). Do đó, tác vụ gán nhãn PAS cho văn bản là một tác vụ thiết thực. Tác vụ này còn được gọi là tác vụ gán nhãn ngữ nghĩa (Semantic Role Labelling – SRL) vì mỗi đối số trong PAS đều có một vai trò ngữ nghĩa kèm theo.

Tác vụ SRL không phải một bài toán mới mẻ trên văn bản tổng quát. Tuy nhiên, đây vẫn là một tác vụ nhiều thách thức đối với văn bản Y Sinh vì ngữ liệu gán nhãn sẵn ít ỏi và PAS trong Y Sinh có nhiều đặc thù khác xa trong văn bản tổng quát khiến cho việc thiết kế đặc trưng khó khăn hơn. Vì thế, một mô hình học sâu dựa trên mô hình ngôn ngữ tiền huấn luyện cần phải có để khắc phục tất cả những khó khăn này [6]. Quan trọng hơn, tri thức về cú pháp đóng vai trò rất tích cực trong tác vụ SRL [7–9]. Vì vậy, một giải pháp nhúng toàn diện tri thức cú pháp được đề nghị để nâng cao hiệu quả của mô hình. Kết quả thử nghiệm của chúng tôi trên bộ ngữ liệu PASBio+ cho thấy hiệu ứng khác nhau mà các loại tri thức ngữ pháp khác nhau tác động lên dự đoán của mô hình SRL trên văn bản Y Sinh. Các giải pháp đã đóng góp gồm có: (i) Đề xuất một mô hình học sâu cho tác vụ SRL trên văn bản Y Sinh dựa trên việc tinh chỉnh (fine tuning) một kiến trúc transformer mạnh mẽ được tiền huấn luyện trên ngữ liệu lớn của ngành Y Sinh; (ii) Thông tin cú pháp đóng vai trò rất quan trọng trong SRL [10], vì vậy, đề xuất giải pháp nhúng hai loại cây cú pháp là cây quan hệ phụ thuộc (dependency parse tree) và cây ngữ

Trích dẫn bài báo này: Đức T N H, Dương L T, Duy H Q. **Mô hình học sâu nhúng cú pháp cho tác vụ gán nhãn ngữ nghĩa văn bản y sinh**. *Sci. Tech. Dev. J. - Nat. Sci.* 2023; 7(4):2750-2762.

pháp thành phần (constituency-based parse tree) vào mô hình học sâu để nâng cao hiệu quả gán nhãn ngữ nghĩa của mô hình; (iii) Thông qua thực nghiệm, đã phân tích tầm ảnh hưởng của hai loại cây cú pháp này lên tác vụ SRL khi sử dụng riêng lẻ và khi sử dụng kết hợp.

Bài báo này được trình bày như sau: Phần *Cơ sở lý thuyết về cấu trúc đối số vị ngữ* cung cấp những khái niệm nền tảng về *cấu trúc đối số vị ngữ* và sự khác biệt giữa lĩnh vực tổng quát với lĩnh vực Y Sinh, đồng thời phân tích một số bộ ngữ liệu hiện có. Phần *Những nghiên cứu về SRL* khái quát hiện trạng nghiên cứu của tác vụ gán nhãn ngữ nghĩa và nhấn mạnh vai trò của tri thức ngữ pháp trong tác vụ này. Phần *Phương pháp thực hiện* mô tả chi tiết mô hình đãi đề xuất. Kết quả thử nghiệm và những thảo luận liên quan được trình bày trong phần *Kết quả thực nghiệm và thảo luận*. Sau cùng, phần *Kết luận* khái quát lại những kết quả đạt được và đề xuất hướng phát triển.

Cơ sở lý thuyết về cấu trúc đối số vị ngữ

Trong Xử lý ngôn ngữ tự nhiên, *cấu trúc đối số vị ngữ* (PAS) là cách biểu diễn mối quan hệ giữa vị ngữ (thành phần cốt lõi của câu diễn tả một hành động hoặc trạng thái) và một khung đối số (frameset) chứa các đối số đi kèm (các thành phần trong câu biểu thị các thực thể liên quan đến vị ngữ). Mỗi đối số được gán nhãn với vai trò ngữ nghĩa (semantic role) của nó để mô tả vai trò của đối số ấy trong hành động hoặc trạng thái mà vị ngữ chuyển tải.

Thí dụ: Xét câu "Dennis borrowed this book from Janes", câu này có PAS gồm vị ngữ là động từ "borrow" và ba đối số xoay quanh vị ngữ là:

Đối số 0: "Dennis" (Vai trò ngữ nghĩa: Người mượn).

Đối số 1: "this book" (Vai trò ngữ nghĩa: Vật được mượn).

Đối số 2: "Janes" (Vai trò ngữ nghĩa: Người cho mượn).

Thí dụ trên cho thấy chỉ cần nhận biết được PAS là máy tính đã nắm hết nội dung chính của câu. Trong lĩnh vực tổng quát, có một số bộ ngữ liệu gán nhãn PAS đã được xây dựng như FrameNet, VerbNet và PropBank[11–13]. Trong đó, PropBank định nghĩa bộ đối số chi tiết nhất cho từng vị ngữ.

Trong lĩnh vực Y Sinh, PAS có nhiều khác biệt so với PAS trong lĩnh vực tổng quát, và thường thấy nhất là khác biệt về vai trò ngữ nghĩa của các đối số. Xét động từ "mutate" làm thí dụ, đối số của động từ này trong lĩnh vực tổng quát và lĩnh vực Y Sinh có vai trò ngữ nghĩa hoàn toàn khác nhau. Trong lĩnh vực tổng quát, các đối số liên quan động từ "mutate" có ngữ nghĩa là: Tác nhân gây thay đổi, vật bị thay đổi, trạng thái trước thay đổi, trạng thái sau thay đổi[13]. Trong khi đó, với lĩnh vực Y sinh, các đối số liên quan động từ "mutate" lại có ngữ nghĩa là: Vị trí xảy ra đột biến (mô hoặc tạng), gene bị đột biến, hậu quả về kiểu gene, hậu quả về kiểu hình[5].

Nhận thấy những khác biệt đó, một số bộ ngữ liệu PAS chuyên biệt cho lĩnh vực Y Sinh đã được xây dựng. Nếu như PAS tổng quát định nghĩa khung đối số cho tất cả động từ trong từ điển thì PAS trong Y Sinh chỉ chú ý những vị ngữ quan trọng trong văn bản Y Sinh, là những động từ truyền tải các sự kiện Y Sinh quan trọng (như đột biến, mã hóa, giải mã, biểu hiện…). Tuy nhiên, chưa có một sự thống nhất chung giữa các công trình về danh sách các động từ này. Các công trình xây dựng những bộ ngữ liệu PAS cho Y Sinh được biết đến nhiều nhất là BioProp, GREC, BioVerbNet và PASBio+.

BioProp là bộ ngữ liệu gồm 1.635 câu, được trích từ đoạn tóm tắt của 500 công bố khoa học về Y Sinh[1]. Điểm hạn chế của Bioprop là chỉ có bối cảnh ngữ liệu là Y Sinh, còn bộ khung đối số cho từng động từ thì vay mượn hoàn toàn từ PropBank. Do đó, các bộ khung đối số của BioProp thực chất là khung đối số tổng quát chứ không phải khung đối số Y Sinh.

GREC là bộ ngữ liệu bao gồm 1.489 câu, được trích từ đoạn tóm tắt của 677 công bố khoa học về Y Sinh[14]. So với BioProp, GREC vượt trội ở chỗ vị ngữ không chỉ bao gồm động từ mà còn có cả các danh động, với bộ đối số được định nghĩa chuyên biệt cho lĩnh vực Y Sinh. Ngoài ra, bộ ngữ liệu GREC còn được gán nhãn thực thể Y Sinh (Bio Named Entity). Tuy nhiên, hạn chế của GREC là không định nghĩa mối quan hệ giữa đối số và vị ngữ. Nói cách khác, GREC định nghĩa tập hợp vị ngữ và tập hợp đối số Y Sinh là hai tập hợp độc lập được phân bố ngẫu nhiên trong văn bản mà không có bất kỳ quan hệ nào với nhau.

BioVerNet là thành phần bổ sung của VerbNet, trong đó thêm vào những động từ Y Sinh mà VerbNet còn thiếu[15]. Tuy nhiên, BioVerbNet có hai hạn chế lớn là số câu gán nhãn sẵn quá ít (chỉ gồm 521 câu) và giữ nguyên khung đối số tổng quát kế thừa từ VerbNet.

PASBio+ là bộ ngữ liệu gán nhãn PAS được xây dựng bằng phương pháp bán tự động gồm 2.617 câu[16]. PASBio+ khắc phục được hạn chế của BioProp và BioVerbNet vì nó dựa trên PASBio, một công trình đặc tả chi tiết từng khung đối số chuyên biệt cho lĩnh vực Y Sinh, không vay mượn đối số của lĩnh vực tổng quát[5]. PASBio+ cũng khắc phục được hạn chế của GREC do đã đặc tả được rõ ràng mối liên hệ của từng đối số với vị ngữ của nó. Cuối cùng, PASBio+ vượt trội về mặt kích thước ngữ liệu so với BioProp và GREC với số lượng câu gần như gấp đôi.

Đối với tác vụ SRL, việc lựa chọn khung đối số có thể quyết định toàn bộ quá trình xử lý của mô hình. Sau khi phân tích các ưu và khuyết điểm của các bộ ngữ liệu cùng với khung đối số như trên, PASBio+ với bộ khung đối số PASBio được chọn cho đề ghị này.

Những nghiên cứu về SRL

Thời gian đầu, bài toán SRL được giải bằng hai hướng tiếp cận phổ biến là hướng dựa luật (rule-based) và hướng khớp mẫu (pattern matching). Hướng dựa luật đòi hỏi một bộ luật được viết thủ công bởi chuyên gia nên tốn kém và khó bao phủ hết mọi lối hành văn phong phú của ngôn ngữ tự nhiên. Tuy nhiên, nó phù hợp với các ngôn ngữ và lĩnh vực ít ngữ liệu như tiếng Hà Lan, tiếng Nhật, lĩnh vực Y Sinh… [17–21]. Hướng khớp mẫu so khớp các mẫu có sẵn vào văn bản để gán nhãn PAS. Hầu hết các mẫu này có được từ khai khoáng dữ liệu [22]. Dù vậy, với lĩnh vực Y Sinh vốn dĩ rất ít ngữ liệu để khai khoáng, các bộ mẫu thường có được từ sự biên soạn thủ công bởi chuyên gia, dẫn đến sự tốn kém và tính bao phủ thấp [23,24].

Từ khi học máy (nhất là kỹ thuật học sâu với độ chính xác ấn tượng) ra đời thì các hướng tiếp cận cũ chỉ còn giữ vai trò bổ trợ. Bên cạnh các công trình học sâu nỗ lực tích hợp tri thức cú pháp vào mô hình thì cũng có nhiều công trình học sâu cố gắng tìm hướng đi khác để tránh những xử lý phức tạp liên quan đến cây cú pháp, từ đó hình thành hai hướng đi chính:

Học sâu không nhận biết cú pháp

Do độ khó của tác vụ phân tích cú pháp không thua gì tác vụ SRL, nhiều công trình học sâu đã bỏ qua tri thức về cú pháp để chứng minh phân tích cú pháp không hẳn là điều kiện tiên quyết cho SRL. Khi này, tác vụ SRL được giải như một bài toán sequence-to-sequence thông thường với các mô hình học sâu quen thuộc như RNN, BiLSTM… và các mô hình xác suất cổ điển hỗ trợ ở lớp đầu ra như Softmax, CRF… [25,26]. Tuy nhiên, vấn đề mất mát đạo hàm (vanishing gradient) có thể làm giảm hiệu quả của các mô hình. Để giải quyết điều này, nhiều kỹ thuật bổ trợ đã được ứng dụng nhằm khắc phục mất mát đạo hàm, như Kết nối Cao tốc (highway connections) hay thuật toán phân tích ngữ pháp phân loại kết hợp A* (A* CCG Parsing) [9,27].

Những mô hình học sâu cho tác vụ SRL không nhận biết cú pháp chủ yếu khác nhau ở cách định nghĩa bài toán. Theo đó, có hai hướng định nghĩa phổ biến là hướng cụm từ và hướng phân tích quan hệ phụ thuộc. Hướng cụm từ (span) xem một chuỗi bao gồm một hoặc nhiều từ làm đơn vị xử lý và tập trung vào việc dự đoán xem cụm từ ấy có phải một đối số hay không [9,25]. Hướng còn lại xem tác vụ SRL là một bài toán phân tích quan hệ phụ thuộc. Các công trình theo hướng này cho thấy sự khả thi về việc xây dựng hệ thống gán nhãn ngữ nghĩa theo hướng quan hệ phụ thuộc mà không cần dùng hoặc dùng rất ít thông tin cú pháp [28,29].

Những công trình nêu trên đều chia tác vụ SRL thành hai tác vụ con, đó là xác định vị ngữ và xác định đối số, với hai mô hình con riêng biệt. Sau đó, công trình đầu tiên dùng chung một mô hình để xác định cả vị ngữ lẫn đối số cũng chỉ dựa trên mạng BiLSTM thông thường, nhưng có tăng cường thêm bộ ghi điểm biaffine giúp hợp nhất việc dự đoán vai trò ngữ nghĩa của đối số và nghĩa của vị từ [30,31].

Dù cho những công trình không nhận biết cú pháp cố gắng tìm những phương án bất khả tri cú pháp (syntax agnostic), vai trò của tri thức cú pháp trong tác vụ SRL là không thể phủ nhận và những mô hình theo hướng nhận biết cú pháp vẫn cho hiệu quả trội hơn.

Học sâu có nhận biết cú pháp

Các nghiên cứu theo hướng này đều đồng ý rằng việc phân tích cây phụ thuộc cung cấp hình thức biểu diễn tốt hơn để gán nhãn vai trò cho các đối số [10]. Phương pháp thường thấy là sử dụng một mô hình bổ trợ để tính toán ra dạng biểu diễn cú pháp phù hợp và nhúng vec-tơ này vô dữ liệu đầu vào của mô hình SRL [32,33]. Từ đó, nhiệm vụ nhúng tri thức cú pháp được chia thành hai nhiệm vụ con là mã hóa cây cú pháp và nhúng vec-tơ biểu diễn cú pháp vào mô hình SRL.

Để mã hóa cây cú pháp, nhiều mô hình đặc biệt đã được đề xuất. Trong số đó, TreeLSTM là mô hình tiên phong, nó là một biến thể của mạng LSTM với khả năng mã hóa cây phụ thuộc với các yếu tố phân nhánh tùy ý [34]. Các mô hình mã hóa thông tin cú pháp trên đây chủ yếu được thiết kế để tạo ra các biểu diễn cú pháp ở cấp độ câu hoặc ngữ. Đối với những mô hình SRL định nghĩa tác vụ như một bài toán sequence-to-sequence thì việc biểu diễn thông tin cú pháp ở cấp độ từ sẽ phù hợp hơn. Mạng Tích chập Đồ thị (Graph Convolutional Network, GCN) hứa hẹn đáp ứng được điều này. GCN là một loại mạng nơ-ron hoạt động dựa trên đồ thị, phù hợp cho việc mô hình hóa đồ thị cú pháp. Với mỗi nốt trên đồ thị (trong trường hợp tác vụ SRL là mỗi từ trong câu), GCN mã hóa thông tin liên quan về các nốt gần đó như là 1 vector. GCN có hai biến thể chuyên biệt vào hai loại cú pháp quan trọng. Mạng Tích chập Đồ thị Cú pháp (Syntactic Graph Convolutional Network, SGCN) dùng để mã hóa quan hệ phụ thuộc [35]. Mạng Tích chập Đồ thị Chuỗi từ (Span Graph Convolutional Network, SpanGCN) sử dụng để mã hóa các

cấu trúc thành phần câu (constituent)[36]. Kết quả là mỗi từ trong câu đều có được một vector mã hóa tất cả những mối quan hệ cú pháp liên quan đến nó. Các vector này có thể được mô hình SRL sử dụng làm dữ liệu đầu vào như các vector nhúng từ (word embedding) thông thường.

Để nhúng vector mã hóa cú pháp vào mô hình SRL, công trình tiên phong sử dụng Mô hình Bộ nhớ ngắn/dài (LSTM) để mã hóa mối quan hệ phụ thuộc giữa vị ngữ và các đối số của nó rồi nhúng vector mã hóa vào mô hình SRL như một vector nhúng câu (sentence embedding)[32]. Mô hình Bộ nhớ ngắn/dài nhận biết cú pháp (Syntax-aware Long Short Term Memory, SA-LSTM) đóng góp một cải tiến, mô hình này học cho mỗi loại quan hệ phụ thuộc một trọng số riêng, từ đó có thể chỉ ra tầm quan trọng của từng loại quan hệ phụ thuộc khác nhau trong tác vụ SRL[33].

Việc phân tích cú pháp không chỉ phức tạp về tính toán mà còn kéo dài thời gian xử lý. Vì vậy, một số nghiên cứu đã tránh phân tích cú pháp bằng cách chọn tập trung học các nhãn phụ thuộc, với khả năng cung cấp thông tin quan trọng cho vai trò ngữ nghĩa, mà không cần sử dụng biểu diễn cú pháp đầy đủ của câu[37,38]. Bản chất của các công trình này là mô hình học đa tác vụ (multi-task learning). Trong đó, tác vụ chính là SRL và tác vụ phụ là phân tích cú pháp. Tri thức học được tự tác vụ phụ sẽ bổ trợ tích cực cho tác vụ chính.

Tóm lại, các công trình nêu trên đều cho thấy tri thức cú pháp đóng góp tích cực cho tác vụ SRL. Trong đó, các thuộc tính cú pháp phụ thuộc và ngữ pháp thành phần nhận được phần lớn sự chú ý do sự liên quan chặt chẽ của chúng với PAS.

PHƯƠNG PHÁP THỰC HIỆN

Tri thức cú pháp được sử dụng

Hai loại cây cú pháp phổ biến nhất trong xử lý ngôn ngữ tự nhiên để nhúng vào mô hình SRL đã được lựa chọn trong đề xuất này: Cây quan hệ phụ thuộc (dependency parse tree) và cây ngữ pháp thành phần (constituency-based parse tree).

Cây ngữ pháp thành phần trong xử lý ngôn ngữ tự nhiên phản ánh cấu trúc ngữ pháp của một câu dựa trên các bộ phận cấu thành câu ấy. Trong cây ngữ pháp thành phần, câu được chia lại được quy thành các đơn vị hoặc thành phần nhỏ hơn. Các thành phần này bao gồm các cụm từ, mệnh đề và các đơn vị ngữ pháp khác được tạo nên bởi các từ trong câu. Xét câu sau trong bộ ngữ liệu PASBio+ làm thí dụ: "Patient 1 has a G-to-A transition at the first nucleotide of intron 2". Câu này có cây ngữ pháp thành phần như minh họa trong Hình 1 dưới đây.

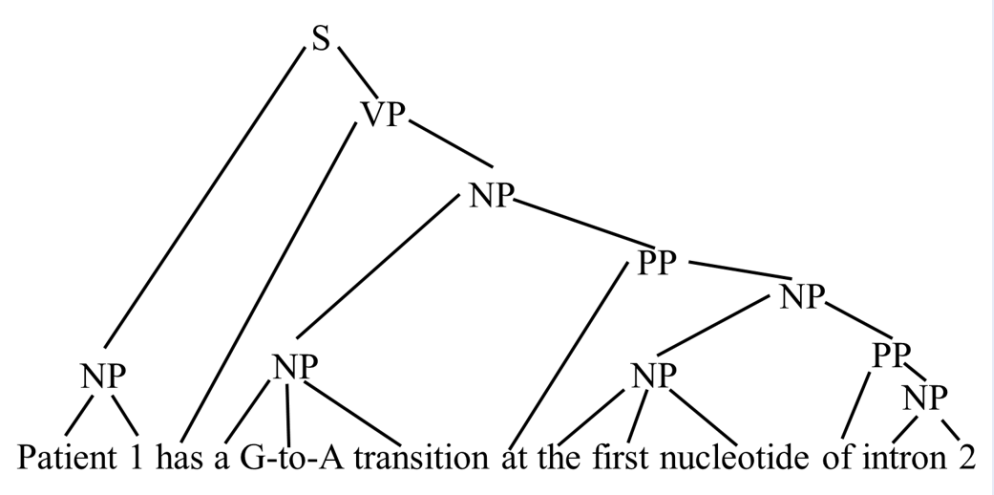
Cây quan hệ phụ thuộc là một dạng đồ thị được sử dụng trong xử lý ngôn ngữ tự nhiên để phân tích mối quan hệ giữa các từ trong câu. Trong đồ thị này, mỗi từ trong câu được biểu diễn dưới dạng một nút và mối quan hệ giữa chúng được biểu diễn thành các cạnh. Các cạnh này biểu thị các mối quan hệ ngữ pháp khác nhau giữa các từ, chẳng hạn như chủ ngữ/tân ngữ, từ bổ nghĩa/đầu tố. Cũng với câu thí dụ trên, cây quan hệ phụ thuộc được minh họa trong Hình 2.

Mô hình được đề xuất

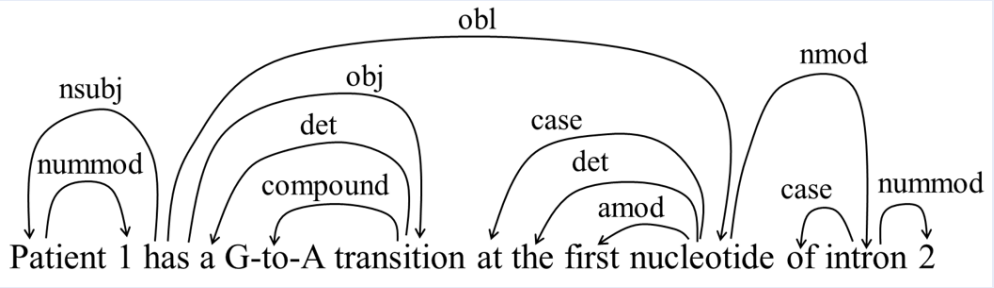
Dữ liệu đầu vào của mô hình SRL gồm hai loại tri thức: Tri thức ngữ cảnh của từ trong câu và tri thức cú pháp (bao gồm hai loại cây cú pháp nêu trên). Hai loại tri thức này được mã hóa thành bộ nhúng ngữ cảnh và bộ nhúng cú pháp (tất cả đều ở mức từ). Hai loại vector này được nối (concatenating) lại với nhau thành một vector duy nhất đưa vào mô hình xử lý trung tâm trước khi đi qua lớp softmax để tính toán phân bố xác suất cuối cùng cho các nhãn. Ngoài ra, ma trận liền kề cũng được sử dụng giúp cho mô hình xử lý trung tâm nắm bắt được vị trí các từ có liên quan với nhau trong câu. Kiến trúc tổng thể của mô hình được minh họa trong Hình 3.

Mã hóa ngữ cảnh

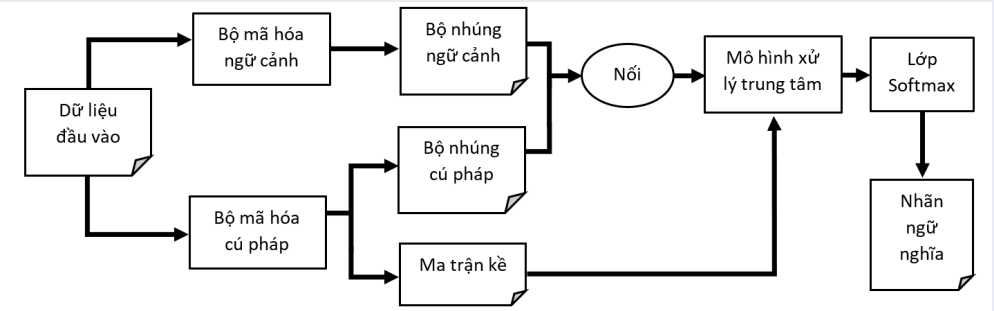
Mã hóa ngữ cảnh là tạo ra một vector biểu diễn cho một từ *w*, trong đó thâu tóm toàn bộ thông tin về tầm ảnh hưởng của các từ khác trong câu lên việc phân loại từ *w* đang xét. Lĩnh vực Y Sinh là một lĩnh vực hạn chế về tài nguyên ngôn ngữ. Một bộ ngữ liệu lớn gán nhãn SRL là không sẵn có. Bộ ngữ liệu huấn luyện PASBio+ cũng chỉ có hơn 2.600 câu, một kích thước chấp nhận được nhưng không phải tối ưu. Vì vậy, để nâng cao hiệu quả gán nhãn ngữ nghĩa, cần tận dụng sức mạnh của các mô hình ngôn ngữ được tiền huấn luyện trên ngữ liệu chuyên ngành với kích thước khổng lồ để tạo ra vector mã hóa ngữ cảnh chất lượng. Trong đề xuất này, BioBERT được lựa chọn làm bộ mã hóa ngữ cảnh cho mô hình[39]. Lựa chọn này đến từ ba lý do: (i) BioBERT được huấn luyện trên ngữ liệu Y Sinh nên hỗ trợ tốt cho tác vụ SRL trên văn bản Y Sinh; (ii) Bản chất BioBERT dựa trên cơ chế self-attention và chính là mô hình transformer, một mô hình mạnh mẽ với độ chính xác ấn tượng trong hàng loạt tác vụ xử lý ngôn ngữ tự nhiên gần đây[40]; (iii) BioBERT sử dụng thuật toán Wordpiece để giảm thiểu vấn đề không tìm thấy từ vựng, một vấn đề thường gặp đối với văn bản Y Sinh vốn chứa nhiều chuỗi ký hiệu đặc biệt của chuyên ngành.



Hình 1: Cây ngữ pháp thành phần của một thí dụ từ PASBio+



Hình 2: Cây quan hệ phụ thuộc của một thí dụ từ PASBio+



Hình 3: Kiến trúc tổng thể của mô hình

Mã hóa các cây quan hệ phụ thuộc

Trước tiên, sử dụng thư viện SpaCy để phân tích mỗi câu thành một cây quan hệ phụ thuộc[41]. Dựa trên cây quan hệ phụ thuộc này, với mỗi từ w_i trong câu, hai vector được sử dụng để mã hóa các quan hệ phụ thuộc liên quan w_i:

i) Bộ nhúng quan hệ phụ thuộc là một vector dạng one-hot có 46 chiều, bao gồm 10 chiều đầu tiên mã hóa từ loại của w_i, 18 chiều tiếp theo mã hóa quan hệ phụ thuộc đi ra từ w_i, và 18 chiều cuối cùng mã hóa quan hệ phụ thuộc đi vào w_i. Sở dĩ có số chiều như vậy là vì cây quan hệ phụ thuộc của SpaCy có 10 nhãn từ loại phân biệt và 18 nhãn quan hệ phụ thuộc phân biệt (Phụ lục A).

ii) Ma trận kề là một vector 2 chiều dạng one-hot, trong đó giá trị 1 tại vị trí dòng i cột j thể hiện rằng từ w_i có quan hệ phụ thuộc đến từ w_j. Còn muốn biết cụ thể w_i phụ thuộc vào w_j bằng mối quan hệ gì thì chỉ cần tra qua bộ nhúng quan hệ phụ thuộc của w_i và w_j sẽ rõ. Hai vector này, cùng với bộ nhúng ngữ cảnh bởi BioBERT, đều được sử dụng làm dữ liệu đầu vào cho mô hình xử lý trung tâm để giải bài toán SRL.

Mã hóa cây ngữ pháp thành phần

Sử dụng thư viện NLTK để phân tích mỗi câu thành một cây ngữ pháp thành phần[42]. Sau đó, tương tự như việc mã hóa cây quan hệ phụ thuộc nêu trên, cây ngữ pháp thành phần cũng được mã hóa thành hai vector là Bộ nhúng ngữ pháp thành phần và Ma trận kề để đưa vào mô hình xử lý trung tâm nhằm giải bài toán SRL. Ma trận kề ở bước này có nguyên lý hoàn toàn giống với ma trận kề của bộ mã hóa cây quan hệ phụ thuộc. Còn Bộ nhúng ngữ pháp thành phần của một từ w_i trong câu là một vector one-hot với giá trị 1 tại vị trí i và tại tất cả vị trí lân cận của i mà ứng với những từ khác nằm trong cùng một ngữ (phrase) với w_i, và giá trị 0 ở các vị trí còn lại (Phụ lục B).

Mã hóa vị trí vị ngữ

Để nhấn mạnh tầm quan trọng của vị ngữ đối với việc xác định đối số, sử dụng Nhúng chỉ định vị ngữ (Predicate Indicator Embedding) để tăng cường sự đóng góp của vị ngữ[9]. Trước tiên, tạo một vector one-hot với giá trị 1 đánh dấu vị trí của vị ngữ trong câu và giá trị 0 ở những vị trí còn lại. Sau đó, vector one-hot này được nối với bộ nhúng thông tin cú pháp rồi đưa tới các bước tính toán tiếp theo. Thông tin này có thể giúp quá trình mã hóa có thể tạo ra các bộ biểu diễn hàm chứa vị trí cụ thể của vị ngữ, từ đó giúp tăng hiệu quả cho mô hình[43].

Mô hình xử lý trung tâm

Vì các bộ nhúng cú pháp mang thông tin mã hóa từ cấu trúc đồ thị, nên các loại mạng neuron truyền thống gặp phải hạn chế lớn là là chúng bỏ qua tất cả các kết nối giữa các nút trong đồ thị mà chỉ đơn thuần đưa ra quyết định dựa vào các đặc trưng được lưu cục bộ ở bản thân từng nút. Do đó, Mạng Neuron Đồ thị (Graph Neural Network, GNN) có thể phù hợp hơn khi xử lý trên các bộ nhúng cú pháp[35]. Từ cơ chế truyền thông điệp của GNN nguyên thủy, nhiều biến thể cải tiến đã ra đời, trong đó biến thể có tích hợp cơ chế attention chính là Mạng Đồ thị Chú ý (Graph ATtention network, GAT)[44]. Cơ chế attention đã chứng tỏ được sức mạnh vượt trội trong xử lý ngôn ngữ tự nhiên. Vì lý do nêu trên, GAT được lựa chọn làm mô hình xử lý trung tâm của đề xuất này. Bộ đặc trưng cuối cùng sau khi được GAT học từ các bộ nhúng ngữ cảnh và cú pháp sẽ được đưa vào hàm softmax để tính phân bố xác suất của các nhãn ngữ nghĩa đầu ra cuối cùng.

KẾT QUẢ VÀ THẢO LUẬN

Dữ liệu thử nghiệm của đề xuất này là PASBio+, một bộ ngữ liệu gán nhãn PAS được xây dựng bằng phương pháp bán tự động gồm 2617 câu[16]. PASBio+ được gán nhãn dựa trên bộ khung đối số PASBio, công trình đặc tả chi tiết cho mỗi động từ một khung đối số chuyên biệt cho lĩnh vực Y Sinh[5]. Bộ ngữ liệu PASBio+ được chia ra thành 3 tập dữ liệu để huấn luyện, đánh giá và kiểm thử tương ứng theo tỉ lệ 60/20/20. Tất cả mô hình của đề xuất được kiểm thử theo phương pháp kiểm chéo 5 pha (5-fold cross validation) và lấy điểm số trung bình.

Các mô hình thử nghiệm: lần lượt huấn luyện và so sánh năm mô hình: (i) Mô hình 1: Chỉ có mã hóa ngữ cảnh và mã hóa vị trí vị ngữ; (ii) Mô hình 2: Mô hình 1 + thêm mã hóa cây quan hệ phụ thuộc; (iii) Mô hình 3: Mô hình 1 + mã hóa cây ngữ pháp thành phần; (iv) Mô hình 4: Mô hình 1 + mã hóa cây quan hệ phụ thuộc và cây ngữ pháp thành phần; (v) Mô hình 5: Mô hình 1 nhưng không có mã hóa vị trí vị ngữ. Các mô hình trên được chạy trên máy có CPU 16 nhân. Tổng thời gian huấn luyện và kiểm thử được chia đều cho 16 nhân CPU nên thời gian thực chạy nhờ đó mà giảm đi 16 lần.

Kết quả thử nghiệm được trình bày trong Bảng 1 và Bảng 2.

Ngoài việc lấy kết quả cho toàn bộ ngữ liệu, điểm F1 cho từng vị ngữ trong danh sách 29 vị ngữ của PASBio cũng được thống kê. Bộ ngữ liệu PASBio+ được tạo ra bằng phương pháp bán tự động, trong đó có những câu tuy khác nhau về nội dung nhưng có cấu trúc ngữ

Bảng 1: Kết quả thực nghiệm thống kê trên toàn bộ ngữ liệu

Mô hình	Precision	Recall	F1	Tổng thời gian huấn luyện (giờ)	Tổng thời gian kiểm thử (giờ)
Mô hình 1	83,34	83,34	83,34	570	35
Mô hình 2	88,41	88,41	88,41	633	42
Mô hình 3	90,37	90,37	90,37	642	43
Mô hình 4	87,29	87,29	87,29	662	45
Mô hình 5	70,10	70,10	70,10	565	34

pháp tương tự nhau. Vì vậy, cũng thống kê tỷ lệ cấu trúc ngữ pháp phân biệt trên tổng số câu của từng vị ngữ trong ngữ liệu để làm cơ sở cho phần thảo luận.

Kết quả thực nghiệm tổng quát (Bảng 1) cho thấy việc nhúng tri thức cú pháp giúp tăng F1 đáng kể. Trong đó, cây ngữ pháp thành phần cải thiện mô hình tốt nhất, giúp F1 tăng 6%, trội hơn cây quan hệ phụ thuộc 1,96%. Điều này cho thấy tuy cả hai loại cây cú pháp đều hữu ích nhưng tri thức từ cây ngữ pháp thành phần thực sự hữu ích hơn tri thức từ cây quan hệ phụ thuộc. Tuy nhiên, ngữ pháp thành phần đòi hỏi thời gian xử lý cao hơn quan hệ phụ thuộc, đây là sự đánh đổi cần cân nhắc khi triển khai vào ứng dụng thực tế. Ngoài ra, sự kết hợp 2 loại cú pháp (Mô hình 4) tuy đòi hỏi thời gian xử lý dài nhất nhưng hiệu quả thấp hơn các mô hình chỉ nhúng một loại cú pháp. Điều này có thể là do việc nhúng cả hai loại ngữ pháp gây ra phần nhiễu trội hơn phần tri thức hữu ích, dẫn tới hiệu quả mô hình giảm xuống. Đáng chú ý nhất là đóng góp vượt trội của Nhúng chỉ định vị ngữ (Predicate Indicator Embedding), thể hiện ở chỗ Mô hình 5 bị giảm F1 đến hơn 13% so với Mô hình 1 chỉ vì bị vắng mặt Nhúng chỉ định vị ngữ. Công trình của này có thể là công trình đầu tiên đưa Nhúng chỉ định vị ngữ vào SRL. Điều này có được nhờ thuận lợi của lĩnh vực Y Sinh chỉ tập trung vào một bộ vị ngữ hữu hạn do PASBio chỉ định.

Khi xét hiệu quả mô hình trên từng vị ngữ (Bảng 2), nhận thấy có sự phân hóa khá rõ rệt. Một số vị ngữ cải thiện F1 vượt bậc sau khi được nhúng cú pháp, như các vị ngữ *express* (tăng từ 72,8 lên 98,5), *transform* (tăng từ 76,9 lên 92,2) hay *eliminate* (tăng từ 78,5 lên 87,8). Khi xem xét ngữ liệu huấn luyện của các vị từ này, nhận thấy chúng hầu như luôn là động từ chính trong câu, nghĩa là làm trung tâm của các cây cú pháp. Vì vậy, các cây cú pháp này dễ dàng hỗ trợ mô hình xác định các đối số xung quanh tốt hơn.

Vị ngữ *mutate* có F1 trung bình thấp nhất, nguyên nhân chủ yếu do *mutate* trong văn bản Y sinh có 3 dạng khác nhau: *mutate*, *mutation*, và *mutated*. Do đó, trong câu nó vừa có thể là danh từ, hoặc động từ, hoặc tính từ, gây khó khăn cho mô hình xử lý trung tâm. Ngược lại, vị ngữ *proliferate* có F1 trung bình cao nhất bởi vì trong dữ liệu huấn luyện, đối số của nó vừa ít nhất, vừa ngắn nhất. Tác vụ SRL cho *proliferate* nhờ vậy mà đơn giản hơn các vị ngữ khác.

Ngoài ra, hai vị ngữ là splice và catalyst có F1 ở mô hình 2 thấp hơn mô hình 1, cho thấy quan hệ phụ thuộc bị phản tác dụng. Khi phân tích ngữ liệu, chúng tôi nhận thấy quan hệ phụ thuộc của hai vị ngữ này khá phức tạp. Hình 4 và Hình 5 dưới đây đơn cử thí dụ về cây quan hệ phụ thuộc của splice và catalyst, qua đó cho thấy chúng có đồng thời khoảng một chục quan hệ phụ thuộc, phần lớn trong số đó không liên quan đến đối số, dẫn tới gây nhiễu đáng kể cho mô hình.

Một điều đáng chú ý là ảnh hưởng của tỷ lệ biến thể ngữ pháp phân biệt trong ngữ liệu. Trái ngược với phần lớn các vị ngữ, một ít vị ngữ có tỷ lệ biến thể ngữ pháp trong ngữ liệu thấp nhất, dưới 15% (vùng tô xám ở đầu ảng 2) lại cho F1 ở Mô hình 2 cao hơn Mô hình 3. Điều này cho thấy số biến thể ngữ pháp nghèo nàn đã dẫn đến việc trích xuất các đặc trưng về ngữ pháp thành phần không đem lại nhiều ý nghĩa bằng quan hệ phụ thuộc, và cũng cho thấy chất lượng của quan hệ phụ thuộc không nhạy cảm với tỷ lệ biến thể ngữ pháp trong ngữ liệu.

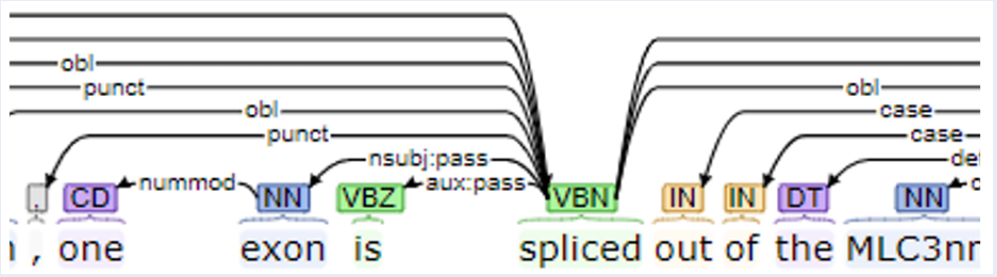
Cuối cùng, đối với các vị ngữ có tỷ lệ biến thể ngữ pháp cao nhất, trên 75% (vùng tô xám ở cuối ảng 2), kết quả của Mô hình 4 lại là kết quả cao nhất. Điều này cho thấy rằng một lượng biến thể ngữ pháp phong phú sẽ hỗ trợ mô hình trung tâm học được cách phối hợp hiệu quả giữa quan hệ phụ thuộc với ngữ pháp thành phần, từ đó giúp cho Mô hình 4 đạt F1 cao nhất.

KẾT LUẬN

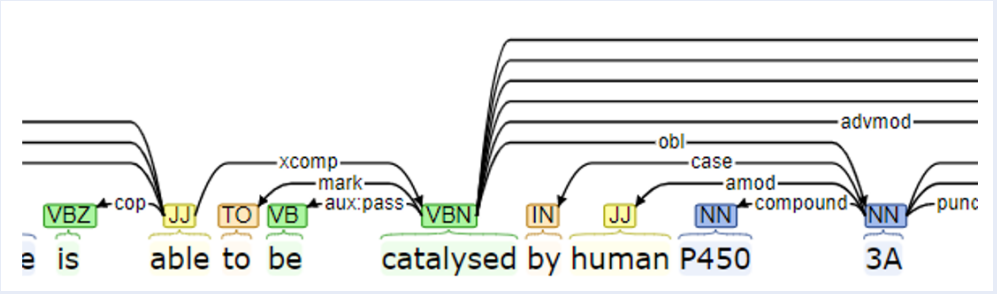
Một mô hình học sâu cho tác vụ SRL trên văn bản Y Sinh đã được xây đựng, trong đó kết hợp ba loại tri thức là tri thức ngữ cảnh từ mô hình ngôn ngữ tiền huấn luyện, tri thức quan hệ phụ thuộc, và tri thức ngữ pháp thành phần. Kết quả thực nghiệm cho thấy trên tổng quát thì tri thức ngữ pháp thành phần cải

Bảng 2: Điểm F1 trung bình thống kê cho từng vị ngữ

Vị ngữ	Tỷ lệ biến thể ngữ pháp (%)	Mô hình 1	Mô hình 2	Mô hình 3	Mô hình 4	Mô hình 5
lead	12	82,97	95,34	91,72	89,32	68,34
block	10	82,41	90,27	89,80	87,39	70,60
eliminate	13	78,55	87,87	87,41	79,84	65,96
inhibit	16	76,72	80,57	84,90	80,38	68,16
abolish	17	80,02	86,62	86,93	87,86	71,28
alter	17,5	88,56	92,39	95,95	85,89	76,15
recognize	18	82,18	82,20	83,04	75,19	64,33
result	18	86,40	88,78	90,52	84,27	76,72
delete	21,5	86,73	86,84	86,84	81,26	63,88
lose	23,5	85,61	90,81	92,03	86,92	68,51
begin	24	83,18	94,59	97,67	94,27	72,89
catalyst	26	92,17	90,98	96,40	89,47	75,72
confer	26	82,57	91,38	94,71	86,57	72,24
transform	29,5	76,95	90,37	92,26	80,89	71,27
initiate	33	83,79	92,29	96,84	82,65	68,65
proliferate	34	95,90	97,97	97,79	96,41	68,74
disrupt	38	89,28	91,12	96,22	85,48	69,80
transcribe	41	77,73	78,36	79,10	78,93	72,21
truncate	57	86,67	93,09	96,66	88,63	68,34
mutate	57,5	72,25	78,14	78,34	73,08	65,08
splice	70,5	89,14	88,67	96,63	86,36	73,13
skip	74,5	86,54	95,16	94,16	94,19	65,63
translate	75,5	84,67	90,86	92,45	94,72	72,10
decrease	80,5	83,52	91,43	92,83	94,91	72,13
develop	81	92,54	93,84	93,98	95,43	70,57
express	85	72,81	81,58	88,40	98,50	65,96
generate	86	76,22	77,27	80,47	92,58	69,95
modify	91,5	83,29	88,91	89,05	92,75	66,64
encode	97,5	77,47	72,18	77,83	87,22	67,92



Hình 4: Một cây quan hệ phụ thuộc của vị ngữ splice



Hình 5: Một cây quan hệ phụ thuộc của vị ngữ catalyst

thiện mô hình tốt nhất (F1 tăng 6%). Tuy nhiên, ở một số vị ngữ có tỷ lệ biến thể ngữ pháp phong phú trong ngữ liệu thì sự kết hợp cả hai loại tri thức cú pháp phát huy hiệu quả cao nhất và có thể tăng F1 lên tối đa gần 20% (trường hợp vị ngữ translate). Ngoài ra, Nhúng chỉ định vị ngữ cho SRL Y Sinh cũng được thử nghiệm và cho thấy vector này giúp F1 tăng đáng kể (hơn 13%).

Tuy nhiên, mô hình này cần đánh đổi bằng thời gian xử lý khi phải phân tích từng câu ra hai cây cú pháp trước khi có thể đưa vào mô hình. Vì vậy, hướng phát triển là sử dụng học chuyển giao (transfer learning) để tái sử dụng hai loại tri thức cú pháp từ mô hình khác. Điều này sẽ giúp tránh được việc phân tích cú pháp ở pha kiểm thử, từ đó hướng đến triển vọng triển khai vào ứng dụng trong nghiệp vụ thực tế được thuận lợi hơn.

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 2023-01.

DANH MỤC TỪ VIẾT TẮT

PAS: Predicate Argument Structure
SRL: Semantic Role Labelling
GAT: Graph ATtention network

XUNG ĐỘT LỢI ÍCH TÁC GIẢ

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

ĐÓNG GÓP CỦA 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.

Lưu Trường Dương và Huỳnh Quốc Duy tham gia khảo sát hiện trạng, đề xuất giải pháp và lập trình thử nghiệm.

PHỤ LỤC A

Về các nhãn cú pháp liên quan đến cây quan hệ phụ thuộc được phân tích bởi thư viện SpaCy, sau khi sử dụng ngưỡng tần số 1% để lọc, chúng tôi chọn được 10 từ loại thường gặp nhất và 18 quan hệ phụ thuộc phổ biến nhất trong ngữ liệu PASBio+, được trình bày trong Bảng 3 và Bảng 4.

PHỤ LỤC B

Bảng liệt kê 10 nhãn ngữ, hay còn gọi là nhãn ngữ pháp thành phần, đại diện cho 10 loại ngữ khác nhau có thể hiện diện trên cây ngữ pháp thành phần.

Bảng 3: 10 nhãn từ loại phổ biến trong PASBio+

Từ loại	Ý nghĩa	Từ loại	Ý nghĩa
NOUN	Chỉ một vật hay một người.	VERB	Động từ
PROPN	Chỉ tên của một vật, một người hay một nơi nào đó	SYM	Các biểu tượng, ký hiệu như dấu +, dấu %,…
PUNCT	Dấu chấm câu	CCONDJ	Liên từ (như or, and, but)
NUM	Biểu thị số lượng, tần suất, phân số	ADP	Giới từ
ADJ	Tính từ	PART	Các tiểu từ, phải liên kết với từ khác mới có nghĩa như 's hay not

Bảng 4: 18 nhãn quan hệ phụ thuộc phổ biến trong PASBio+

Quan hệ phụ thuộc	Ý nghĩa	Quan hệ phụ thuộc	Ý nghĩa
compound	Định hình đầu tố (head) của một chuỗi từ ghép	npadvmod (noun phrase as adverbial modifier)	Các trường hợp mà cụm danh từ làm bổ ngữ trong câu
punct (punctuation)	Liên kết giữa một từ và dấu chấm câu	cc (cooordination)	Mối quan hệ giữa một yếu tố của liên từ và từ liên kết phối hợp của liên từ
nmod (nominal modifier)	Phụ thuộc danh nghĩa giữa danh từ với một đặc trưng trong câu	nsubj (nominal subject)	Một cụm danh từ là chủ đề cú pháp và là tác nhân khởi đầu (proto-agent) trong mệnh đề
amod (adjectival modifier)	Bất kỳ cụm tính từ nào dùng để sửa đổi ý nghĩa của cụm danh từ	dobj (direct object)	Một cụm danh từ là chủ thể chính của động từ
pobj (object of a proposition)	Phần đầu tố (head) của cụm danh từ theo sau giới từ hoặc các trạng từ "here" và "there".	prep (prepositional modifier)	Bất kỳ cụm giới từ nào dùng để sửa đổi ý nghĩa của động từ, tính từ, danh từ hoặc thậm chí một giới từ khác
conj (conjunction)	Mối quan hệ giữa hai yếu tố được nối với nhau bằng liên từ	nsubjpass (passive nominal subject)	Một cụm danh từ là chủ thể cú pháp chính của một mệnh đề gián tiếp
appos (appositional modifier	Một cụm danh từ nằm ngay bên phải của cụm danh từ đầu tiên, dùng để bổ nghĩa cho cụm đó	acl (clausal modifier of noun)	Mối quan hệ giữa các mệnh đề hữu hạn hoặc không hữu hạn sửa đổi ý nghĩa một danh từ
det (determiner)	Mối quan hệ giữa phần đầu tố (head) của cụm danh từ và từ bổ nghĩa cho nó (the, which)	root	Gốc của câu
nummod (numeric modifier)	Bất kỳ cụm số nào dùng để sửa đổi ý nghĩa của danh từ với một số lượng	case	Phụ thuộc giữa bất kỳ giới từ nào và danh từ

Bảng 5: Các nhãn ngữ (phrase) trong PASBio+

Nhãn ngữ pháp thành phần (Constituent tag)	Ý nghĩa	Nhãn ngữ pháp thành phần (Constituent tag)	Ý nghĩa
NP (Noun Phrase)	Cụm danh từ	ADJP (Adjective Phrase)	Cụm tính từ
VP (Predicate)	Cụm động từ	PRT (particle)	Một từ đơn giản, không thể tách ra thêm
PP (Preposition Phrase)	Cụm giới từ	INTJ (interjection)	Từ cảm thán
ADVP (Adverb Phrase)	Cụm trạng từ	CONJP (conjunction phrase)	Cụm liên từ
SBAR (subordinate clause)	Mệnh đề bắt đầu bằng "that" hoặc một liên từ	LST (list markers)	Các dấu câu

Bảng 6: Các thông số cấu hình khi chạy mô hình

Tên thông số	Giá trị thông số được cấu hình
Số epoch khi train	50
Chiều dài tối đa của 1 câu sau khi token hóa bằng WordPiece	256
Kích thước của batch	32
Hệ số học	0,00003
Số GAT head (trong cơ chế multi-head)	12
Số đơn vị GAT trong mỗi head	64
Số lớp GAT	1
Tỉ lệ dropout trong GAT	0 (không dùng dropout)

PHỤ LỤC C

Các thông số cấu hình Mô hình xử lý trung tâm khi huấn luyện trên tập ngữ liệu PASBio+:

Các mô hình trên được chạy trên máy có cấu hình: Intel(R) Xeon(R) CPU E5-2609 v4 @ 1.70GHz, CPU 16 core, Ram 32 GB.

TÀI LIỆU THAM KHẢO

1. Wen-Chi Chou, "A Semi-Automatic Method for Annotating a Biomedical Proposition Bank," in Workshop on Frontiers in Linguistically Annotated Corpora; 2006;Available from: https://aclanthology.org/W06-0602/.
2. Kirschner MW, Marincola E, Teisberg EO. The role of biomedical research in health care reform. Science;266(5182). doi: 10.1126/science.7939643, PMID 7939643 in . Science. 1994;266(5182):49-51;PMID: 7939643. Available from: https://doi.org/10.1126/science.7939643.
3. Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: an overview. J Comput Biol. 2003;10(6):821-55;PMID: 14980013. Available from: https://doi.org/10.1089/106652703322756104.
4. Kim J-D, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature. BMC Bioinformatics. 2008;9:10;PMID: 18182099. Available from: https://doi.org/10.1186/1471-2105-9-10.
5. Wattarujeekrit T, Shah PK, Collier N. PASBio: predicate-argument structures for event extraction in molecular biology. BMC Bioinformatics. 2004;5:155;PMID: 15494078. Available from: https://doi.org/10.1186/1471-2105-5-155.
6. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85-117;PMID: 25462637. Available from: https://doi.org/10.1016/j.neunet.2014.09.003.
7. Roth M, Lapata M. Neural semantic role labeling with dependency path embeddings. Available from: https://aclanthology.org/P16-1113/. In: the 54th Annual Meeting of the Association for Computational Linguistics; 2016;Available from: https://doi.org/10.18653/v1/P16-1113.
8. Strubell E, Verga P, Andor D, Weiss D, McCallum A. Linguistically informed self-attention for semantic role labeling. Available from: https://arxiv.org/abs/1804.08199. In: the 2018 Conference on Empirical Methods in Natural Language Processing; 2018;Available from: https://doi.org/10.18653/v1/D18-1548.
9. He Luheng, Lee K, Lewis M, Zettlemoyer L. Deep semantic role labeling: what works and What's Next. Available from: https://aclanthology.org/P17-1044/. In: the 55th Annual Meeting of the Association for Computational Linguistics; 2017;Available from: https://doi.org/10.18653/v1/P17-1044.
10. Punyakanok Vasin, Roth D, Yih W. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. Comp Linguist. 2008;34(2):257-87;Available from: https://doi.org/10.1162/coli.2008.34.2.257.
11. Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet project. Available from: https://aclanthology.org/P98-1013. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics; 1998;Available from: https://doi.org/10.3115/980845.980860.
12. Kipper K. Hoa Trang Dang và Martha Palmer, [Class-based construction of a verb lexicon]. In: the Seventeenth National Conference on Artificial Intelligence; 2000;Available from: https://cdn.aaai.org/AAAI/2000/AAAI00-106.pdf.

13. Kingsbury P, Palmer M. From TreeBank to PropBank. Available from: https://aclanthology.org/L02-1283/. In: the Third International Conference on Language Resources and Evaluation; 2002;.
14. Thompson P, Cotter P, McNaught J, Ananiadou S, Montemagni S, Trabucco A et al. Building a bio-event annotated corpus for the acquisition of semantic frames from biomedical corpora. Available from: https://aclanthology.org/L08-1231/. In: the Sixth International Conference on Language Resources and Evaluation; 2008;.
15. Majewska O, Collins C, Baker S, Björne Jari, Brown SW, Korhonen A et al. BioVerbNet: a large semantic-syntactic classification of verbs in biomedicine. J Biomed Semantics. 2021;12(1):12;PMID: 34266499. Available from: https://doi.org/10.1186/s13326-021-00247-z.
16. Đức TNH, Hoàng VT, Phạm HS. A semiautomatic approach to biomedical semantic role corpus construction. VNUHCM J Nat Sci. 2022;6(2):2083-94;.
17. Stevens G. XARA: an XML- and rule-based semantic role labeler; 2007. In: The Linguistic Annotation workshop;PMID: 26110305. Available from: https://doi.org/10.3115/1642059.1642077.
18. Iida R, Komachi M, Inui K, Matsumoto Y. Annotating a Japanese text corpus with predicate-argument and coreference relations; 2007. In: The Linguistic Annotation workshop;Available from: https://doi.org/10.3115/1642059.1642081.
19. Pollard C, Sag IA. Head-driven phrase structure grammar. In: the 57th Annual Meeting of the Association for Computational Linguistics; 1994;.
20. Liakata M, Stephen G. Pulman. In: From trees to predicate-argument the 19th international conference on Computational linguistics; 2002;Available from: https://doi.org/10.3115/1072228.1072333.
21. Wattarujeekrit T, Collier N. Exploring predicate-argument relations for named entity recognition in the molecular biology domain. In: IFIP Working Conference on Database Semantics; 2005;Available from: https://doi.org/10.1007/11563983_23.
22. Riloff E. Automatically generating extraction patterns from untagged text. In: the thirteenth national conference Artificial intelligence; 1996;.
23. Lin C-S (A), Smith TC. Semantic role labeling via consensus in pattern-matching. In: Conference on Computational Natural Language Learning; 2005;Available from: https://doi.org/10.3115/1706543.1706577.
24. Huang M, Zhu X, Hao Y, Payan DG, Qu K, Li M. Discovering patterns to extract protein-protein interactions from full texts. Bioinformatics. 2004;20(18):3604-12;PMID: 15284092. Available from: https://doi.org/10.1093/bioinformatics/bth451.
25. Zhou J, Xu W. End-to-end learning of semantic role labeling using recurrent neural. In: the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; 2015;Available from: https://doi.org/10.3115/v1/P15-1109.
26. FitzGerald N, Täckström O, Ganchev K, Das D. Semantic role labeling with neural network factors. Available from: https://aclanthology.org/D15-1112.pdf. In: the 2015 Conference on Empirical Methods in Natural Language Processing; 2015;Available from: https://doi.org/10.18653/v1/D15-1112.
27. Lewis M, Steedman M. A* CCG parsing with a supertag-factored model. Available from: https://aclanthology.org/D14-1107. In: the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014;Available from: https://doi.org/10.3115/v1/D14-1107.
28. Marcheggiani D, Frolov A, Titov I. A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling. Available from: https://aclanthology.org/K17-1041. In: the 21st Conference on Computational Natural Language Learning (CoNLL 2017); 2017;Available from: https://doi.org/10.18653/v1/K17-1041.
29. Cai J, He Shexia, Li Zuchao, Zhao H. A full end-to-end semantic role labeler, syntactic-agnostic over syntactic-aware. In: the 27th International Conference on Computational Linguistics; 2018;.
30. Li Zuchao, He Shexia, Zhao H, Zhang Y, Zhang Zhuosheng, Zhou X et al. Dependency or span, end-to-end uniform semantic role labeling. AAAI. 2018;33(1):6730-7;Available from: https://doi.org/10.1609/aaai.v33i01.33016730.
31. Dozat T, Manning CD. Deep biaffine attention for neural dependency parsing. ICLR. 2017;2017;.
32. Roth M, Lapata M. Neural semantic role labeling with dependency path embeddings. Available from: https://aclanthology.org/P16-1113. In: the 54th Annual Meeting of the Association for Computational Linguistics; 2016;Available from: https://doi.org/10.18653/v1/P16-1113.
33. Qian F, Sha L, Chang B, Liu L-C, Zhang M. Syntax Aware LSTM model for Semantic Role Labeling. In: the 2nd Workshop on Structured Prediction for Natural Language Processing; 2017;Available from: https://doi.org/10.18653/v1/W17-4305.
34. Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. In: the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; 2015;Available from: https://doi.org/10.3115/v1/P15-1150.
35. Marcheggiani D, Titov I. Encoding sentences with graph convolutional networks. Available from: https://aclanthology.org/D17-1159. In: the 2017 Conference on Empirical Methods in Natural Language Processing; 2017;Available from: https://doi.org/10.18653/v1/D17-1159.
36. Marcheggiani D, Titov I. Graph convolutions over constituent trees for syntax-aware semantic role labeling. Available from: https://aclanthology.org/2020.emnlp-main.322. In: the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020;Available from: https://doi.org/10.18653/v1/2020.emnlp-main.322.
37. Cai R, Lapata M. Syntax-aware semantic role labeling without parsing. Trans Assoc Comp Linguist. 2019;7:343-56;Available from: https://doi.org/10.1162/tacl_a_00272.
38. Swayamdipta S, Thomson S, Lee K, Zettlemoyer L, Dyer C, Smith NA. Syntactic scaffolds for semantic structures. In: the 2018 Conference on Empirical Methods in Natural Language Processing; 2018;Available from: https://doi.org/10.18653/v1/D18-1412.
39. Lee Jinhyuk, Yoon Wonjin, Kim Sungdong, Kim Donghyeon, Kim Sunkyu, So CH et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40;PMID: 31501885. Available from: https://doi.org/10.1093/bioinformatics/btz682.
40. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al. Attention is all you need.. In: Neural information processing systems (NIPS 2017). Long Beach, CA; 2017;Available from: https://arxiv.org/abs/1706.03762.
41. spaCy: industrial-strength NLP [online];Available from: https://github.com/explosion/spaCy.
42. "NLTK: Natural Language Toolkit," [online];Available from: https://www.nltk.org/.
43. Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations. In: Vol. 2018; 2018. NAACL;Available from: https://doi.org/10.18653/v1/N18-2074.
44. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. ICLR. 2018;2018;.

Science & Technology Development Journal – Natural Sciences 2024, 7(4):2750-2762

Open Access Full Text Article

Research article

5A syntax–aware deep–learning model for biomedical semantic role labelling

Tuan Nguyen Hoai Duc[1,2,*], Luu Truong Duong[3], Huynh Quoc Duy[3]

ABSTRACT

A deep learning model for biomedical semantic role labeling was build. Semantic role labeling is a useful task that enables the computer to comprehend the key facts expressed in each sentence, and is a necessary first step in the resolution of several other semantic-related tasks, such as event extraction, entity extraction, and Q-A systems... Semantic role labeling is a domain-dependent task. In the biomedical field, semantics are transmitted via more intricate grammatical structures and dependencies in addition to being built on a predicate argument frameset that differs greatly from that of the general domain. To effectively account for these unique characteristics, three types of information were integrated into this deep learning model: Context knowledge obtained from a pre-trained language model trained on a substantial corpus of biomedical texts, dependencies derived from the dependency parse trees and sentence structure obtained from constituency parse trees. To handle grammatical information that is naturally represented as graphs, the Graph Attention Network which is well-known for its remarkable graph learning capabilities, was used. To further boost the model effectiveness, predicate indicator embedding was additionally included in the proposed model. According to experimental findings, the two above-indicated forms of syntactic information, along with the predicate indicator embedding, could boost F1 by up to 20%.

Key words: predicate-argument structure, semantic role labelling, deep learning, constituency parsing, dependency parsing, biomedical text

[1]*Faculty of Information Technology, University of Sciences, Ho Chi Minh City, Vietnam*

[2]*Vietnam National University Ho Chi Minh City, Vietnam*

[3]*Antsomi Vietnam Company Limited, Vietnam*

Correspondence

Tuan Nguyen Hoai Duc, Faculty of Information Technology, University of Sciences, Ho Chi Minh City, Vietnam

Vietnam National University Ho Chi Minh City, Vietnam

Email: tnhduc@fit.hcmus.edu.vn

History

- Received: 10-4-2023
- Accepted: 9-10-2023
- Published Online: 31-12-2023

DOI :

https://doi.org/10.32508/stdjns.v7i4.1279

VNUHCM PRESS

Cite this article : Duc T N H, Duong L T, Duy H Q. **5A syntax–aware deep–learning model for biomed-ical semantic role labelling**. *Sci. Tech. Dev. J. - Nat. Sci.* 2023; 7(4):2750-2762.