Tạp chí Phát triển Khoa học và Công nghệ – Khoa học Tự nhiên 2022, 6(2):2083-2094

Open Access Full Text Article

Bài nghiên cứu

Xây dựng ngữ liệu gán nhãn ngữ nghĩa y sinh bằng hướng tiếp cận bán tự động

Tuấn Nguyên Hoài Đức[1,*], Phạm Hữu Sang[2], Hoàng Văn Thức[3]

TÓM TẮT

Bài báo trình bày một giải pháp bán tự động để xây dựng bộ ngữ liệu gán nhãn ngữ nghĩa Y sinh mang tên PASBio+. Bộ ngữ liệu PASBio+ chứa nhãn Cấu trúc Đối số Vị ngữ, một dữ kiện quan trọng bao quát toàn bộ nội dung chính của câu. Do hơn 86% đối số trong Y sinh khác biệt đáng kể so với đối số trong lĩnh vực tổng quát nên ngữ liệu được gán nhãn theo PASBio, một bộ khung đối số được soạn chuyên biệt dành riêng cho Y sinh. Tiền đề của PASBio+ là 317 câu đã gán nhãn của PASBio. Từ đó, với giải pháp bán tự động này, các chuyên gia chỉ cần gán nhãn thủ công 87 câu để cuối cùng có ngữ liệu gồm 2.500 câu đã gán nhãn đầy đủ. Điều này đạt được nhờ Phương Pháp Ví Dụ Ảo, một kỹ thuật tăng cường dữ liệu mạnh mẽ đầy linh hoạt được áp dụng thành công trong hàng loạt tác vụ khác nhau. Ngữ liệu sinh ra bởi Phương Pháp Ví Dụ Ảo được qui định bằng hai quy tắc tuần tự để đảm bảo tri thức Y sinh luôn được giữ đúng đắn (quy tắc Trao đổi và quy tắc Thay thế). PASBio+ cũng được tăng cường độ phong phú mẫu câu bằng biến thể ngữ pháp của các câu gốc, giúp ngữ liệu có độ phủ rộng trên các cách hành văn tự nhiên đa dạng. Ngoài ra, ngay từ đầu, bộ câu gốc của PASBio cũng được làm giàu bằng nguồn văn bản ngoài, là bộ câu bổ sung được chọn lọc từ ngữ liệu Y sinh GREC. Bên cạnh đó, PASBio+ đạt độ phân bố tần suất rất đồng đều giữa các vị ngữ, nhờ đó loại bỏ vấn đề dữ liệu thưa (data sparsity), giúp hạn chế lỗi quá khớp (overfitting) trong học máy. Kết quả đánh giá thực nghiệm cho thấy bộ ngữ liệu đề nghị này, với vai trò là ngữ liệu huấn luyện, đã giúp mô hình học sâu tăng điểm F thêm 52,2% và 22,5% khi so sánh lần lượt với mô hình huấn luyện bằng ngữ liệu gốc chưa tăng cường và ngữ liệu của lĩnh vực tổng quát.

Từ khoá: cấu trúc đối số vị ngữ, gán nhãn ngữ nghĩa, xây dựng ngữ liệu, tăng cường dữ liệu

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

[2]*Công ty TNHH Giải pháp Việt Bản Đồ, Việt Nam*

[3]*Tổng Công ty Giải pháp Doanh nghiệp Viettel, Việt Nam*

Liên hệ

Tuấn Nguyên Hoài Đức, Bộ môn Hệ thống Thông tin, 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: 30-11-2021
- Ngày chấp nhận: 05-6-2022
- Ngày đăng: 30-6-2022

DOI : 10.32508/stdjns.v6i2.1151

GIỚI THIỆU

Y sinh (Biomedical) là ngành khoa học ứng dụng tri thức sinh học phân tử để chẩn đoán và điều trị trong y khoa. Cùng với sự lớn mạnh của các thành tựu sinh học phân tử, ngành Y sinh dần xác định vững chắc vai trò của mình trong việc chăm sóc sức khỏe con người, và thu hút ngày càng nhiều nỗ lực nghiên cứu liên ngành Sinh-Tin [1–5]. Kho tàng văn bản khoa học của ngành Y sinh ngày càng đồ sộ, với lượng tri thức tuy hữu ích nhưng khổng lồ vượt trên khả năng khai thác thủ công của con người [6,7]. Do đó, con người cần đến sức mạnh máy tính để có thể khai khoáng hiệu quả tri thức quý báu từ những kho tài liệu đồ sộ này [8–11].

Cấu trúc Đối số Vị ngữ (**P**redicate **A**rgument **S**tructure – PAS) là bộ khung ngữ nghĩa của câu, chuyển tải toàn bộ sự kiện quan trọng được nói đến trong câu [1]. Rút trích được PAS, máy tính xem như đã hiểu được nội dung chính trong câu. Do đó, bài toán rút trích PAS, còn gọi là bài toán SRL (**S**emantic **R**ole **L**abelling) là tiền đề quan trọng cho hàng loạt bài toán xử lý ngôn ngữ tự nhiên khác như rút trích sự kiện, nhận diện thực thể, hệ hỗ trợ ra quyết định, tóm tắt văn bản… [12–15].

Để huấn luyện máy tính xử lý bất kỳ tác vụ nào trên văn bản Y sinh cũng cần có một bộ ngữ liệu gán nhãn tác vụ ấy trên chính văn bản Y sinh [3,4,16,17]. Tuy nhiên, với bài toán SRL cho văn bản Y sinh, ngữ liệu gán nhãn sẵn vẫn còn rất hạn chế cả về kích thước và độ phong phú ngữ nghĩa [3,4]. Bên cạnh đó, việc xây dựng một bộ ngữ liệu bằng phương pháp thủ công lại đòi hỏi nhiều nhân lực và thời gian. Vì vậy, bài báo này trình bày việc đề xuất một giải pháp bán tự động nhằm xây dựng bộ ngữ liệu gán nhãn PAS trên văn bản Y sinh tiếng Anh. Bộ ngữ liệu được đặt tên là PASBio+, được thử nghiệm trên văn bản Y sinh để so sánh với các bộ ngữ liệu huấn luyện SRL hiện có và cho thấy hiệu quả cao (xin xem phần "Kết quả thực nghiệm và đánh giá").

CƠ SỞ LÝ THUYẾT VỀ CẤU TRÚC ĐỐI SỐ VỊ NGỮ

Cấu trúc Đối số Vị ngữ (PAS) là một cấu trúc mà trong đó đối tượng trung tâm là động từ chính của câu, gọi

Trích dẫn bài báo này: Đức T N H, Sang P H, Thức H V. **Xây dựng ngữ liệu gán nhãn ngữ nghĩa y sinh bằng hướng tiếp cận bán tự động**. *Sci. Tech. Dev. J. - Nat. Sci.;* 6(2):2083-2094.

là vị ngữ (predicate). Xung quanh vị ngữ là những đối tượng được nói đến trong câu mà có liên quan đến hành động do vị ngữ truyền tải. Tất cả những đối tượng ấy được gọi là các đối số của vị ngữ (argument). Mỗi đối số đều có một vai trò ngữ nghĩa cụ thể (semantic role). Danh sách đối số của mỗi động từ gọi là khung đối số (frameset) của động từ đó. Có nhiều công trình định nghĩa khung đối số cho tất cả động từ trong từ điển như VerbNet[18], FrameNet[19], và PropBank[20,21]. Tất cả đều dành cho lĩnh vực tổng quát, không chuyên biệt vào Y sinh.

Ví dụ: Theo PropBank, câu "I've heard that **John sold this car for 500 USD** last month." có PAS như sau:

- "sold" là vị ngữ (P) chuyển tải sự kiện chính trong câu là sự kiện "bán".
- "John" là đối số thứ nhất (A0) với vai trò ngữ nghĩa là "người bán"
- "car" là đối số thứ hai (A1) với vai trò ngữ nghĩa là "món hàng"
- "500 USD" là đối số thứ tư (A3) với vai trò ngữ nghĩa là "giá bán"

Không phải lúc nào mọi đối số cũng hiện diện đủ. Như trong ví dụ trên còn khuyết hai đối số mà PropBank có định nghĩa: Đối số A2 với vai trò ngữ nghĩa "người mua" và đối số A4 với vai trò ngữ nghĩa "lợi nhuận thu về".

Với cùng một động từ trong từ điển, các công trình khác nhau định nghĩa khung đối số rất khác nhau. Ví dụ như cùng động từ "sell" nêu trên, trong khi PropBank định nghĩa đến 5 đối số thì VerbNet chỉ định nghĩa 3 đối số (người bán, người mua, món hàng) và FrameNet thì chỉ định nghĩa 2 đối số (người bán, món hàng). Vì vậy, hiện tại PropBank vẫn là bộ khung đối số giàu ngữ nghĩa nhất trong lĩnh vực tổng quát.

PAS trong Y sinh chỉ tập trung vào các động từ liên quan đến những sự kiện Y sinh quan trọng chứ không còn dàn trải trên tất cả các động từ của từ điển. Khung đối số trong Y sinh có nhiều khác biệt so với lĩnh vực tổng quát. Nhiều công trình đã đề xuất khung đối số mới cho phù hợp với lĩnh vực Y sinh như GREC[3] và PASBio[1]. Cụ thể, bộ khung đối số giữa lĩnh vực Y sinh và lĩnh vực tổng quát có 4 dạng khác biệt: (i) Dạng 1: Động từ không đổi nghĩa nhưng khung đối số Y sinh có nhiều đối số hơn, ví dụ động từ "Alter" (Bảng 1); (ii) Dạng 2: Động từ không đổi nghĩa nhưng khung đối số Y sinh có ít đối số hơn, ví dụ động từ "Generate" (Bảng 2); (iii) Dạng 3: Động từ đổi nghĩa khi đi vào lĩnh vực Y sinh, ví dụ động từ "Express" (Bảng 3); (iv) Dạng 4: Động từ không đổi nghĩa, số lượng đối số không đổi, nhưng ý nghĩa đối số thay đổi khi đi vào lĩnh vực Y sinh, ví dụ động từ "Modify" (Bảng 4).

Bảng 5 trình bày sự phân bố bộ động từ của PASBio vào 4 dạng nêu trên, cho thấy tổng cộng có đến 86,2% động từ có sự khác biệt khi so sánh trực tiếp khung đối số của nó trong lĩnh vực tổng quát và trong lĩnh vực Y sinh.

Do đó, việc xây dựng bộ ngữ liệu gán nhãn PAS chuyên biệt cho lĩnh vực Y sinh là rất cần thiết. Mặc dù đã có nhiều ngữ liệu gán nhãn PAS cho lĩnh vực tổng quát[20,21] nhưng hiện nay chỉ có 3 tài nguyên ngữ liệu gán nhãn PAS cho văn bản Y sinh:

- BioProp là bộ ngữ liệu bao gồm 1635 câu trích từ phần tóm tắt (abstract) của 500 bài báo Y sinh[4]. Hạn chế của Bioprop là vay mượn 80% khung đối số của VerbNet. Vì BioProp không tập trung chuyên biệt cho đối số Y sinh, nó không thực sự đáp ứng nhu cầu huấn luyện tác vụ SRL trong Y sinh.
- PASBio khắc phục hạn chế của BioProp bằng cách định nghĩa lại toàn bộ khung đối số chuyên biệt vào Y sinh[1]. Nhưng PASBio chỉ gán nhãn 317 câu để minh họa các khung đối số của mình. Kích thước này là quá nhỏ, không thể dùng làm ngữ liệu huấn luyện trong học máy.
- GREC là bộ ngữ liệu bao gồm 1489 câu trích từ phần tóm tắt của 677 bài báo Y sinh[3]. Tuy nhiên, GREC không tập trung định nghĩa bất kỳ mối liên hệ nào giữa đối số và vị ngữ. Điều này làm cho đối số và vị ngữ trong GREC chỉ là 2 tập hợp độc lập phân bố ngẫu nhiên trong ngữ liệu.

Hình 1 minh họa sự khác biệt giữa PASBio và GREC, qua đó cho thấy PASBio có cấu trúc mạch lạc và giàu ngữ nghĩa hơn nhờ mối liên hệ giữa vị ngữ và các đối số. Điều PASBio còn thiếu so với GREC là ngữ liệu gán nhãn còn quá ít để có thể dùng huấn luyện máy tính. Vì thế, một giải pháp bán tự động nhằm làm tăng kích thước của bộ ngữ liệu này được đề xuất.

NHỮNG NGHIÊN CỨU VỀ XÂY DỰNG NGỮ LIỆU

Các công trình xây dựng ngữ liệu được phân thành ba hướng tiếp cận: Hướng tiếp cận thủ công, hướng tiếp cận tự động và hướng tiếp cận bán tự động.

Hướng tiếp cận thủ công

Ở hướng tiếp cận thủ công, ngữ liệu được xây dựng hoàn toàn thủ công bởi chuyên gia. Công cụ hỗ trợ nếu có chỉ là những phần mềm biên tập giúp rà soát ngữ liệu thuận tiện hơn[22]. Nhiều bộ ngữ liệu cho lĩnh vực Y sinh đã được xây dựng bằng hướng tiếp cận này, như ngữ liệu gán nhãn thực thể và quan hệ BioInfer[17], ngữ liệu gán nhãn sự kiện GENIA Event[23], và

Bảng 1: Các khung đối số của động từ "Alter"

Động từ : Alter – Ý nghĩa: Làm thay đổi	
Khung đối số Y sinh (theo PASBio)	Khung đối số tổng quát (theo PropBank)
+ Arg0: Tác nhân làm thay đổi (đột biến, protein) + Arg1: Thứ bị thay đổi (codon, exon, mRNA) + Arg2: Trạng thái kết thúc + Arg3: Trạng thái bắt đầu + Arg4: Vị trí xảy ra thay đổi (mô, tạng, gene)	+ Arg0: Người thực hiện việc thay đổi + Arg1: Thứ bị thay đổi

Bảng 2: Các khung đối số của động từ "Generate"

Động từ : Generate – Ý nghĩa: Sinh ra	
Khung đối số Y sinh (theo PASBio)	Khung đối số tổng quát (theo PropBank)
+ Arg0: Nguồn sinh (gene, quá trình sinh học) + Arg1: Sản phẩm (bản phiên mã, mRNA)	+ Arg0: Người tạo tác + Arg1: Thứ được sinh ra + Arg2: Nguyên liệu + Arg3: Người thụ hưởng + Arg4: Mục đích

Bảng 3: Các khung đối số của động từ "Express"

Khung đối số Y sinh (theo PASBio)	Khung đối số tổng quát (theo PropBank)	
Express = Biểu hiện ra kiểu hình + Arg0: Gene được biểu hiện + Arg1: Đặc điểm kiểu hình + Arg2: Cơ quan mang kiểu hình	Express = Chuyển phát nhanh + Arg0: Người gửi + Arg1: Hàng gửi + Arg2: Người nhận	Express = Diễn đạt + Arg2: Người nói + Arg3: Điều nói ra + Arg4: Người nghe

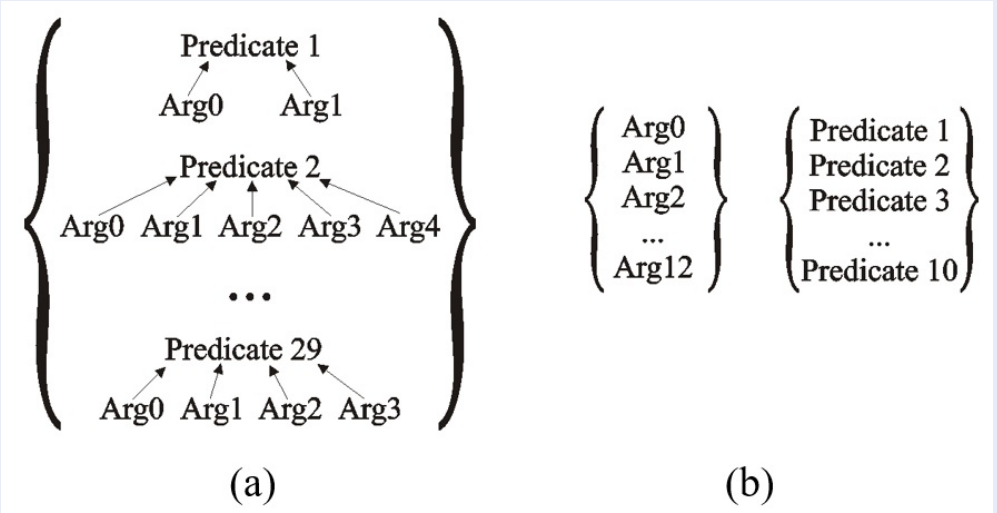
Bảng 4: Các khung đối số của động từ "Modify" cho thấy Arg2 và Arg3 trong Y sinh có ý nghĩa khác

Động từ : Modify – Ý nghĩa: Biến đổi	
Khung đối số Y sinh (theo PASBio)	Khung đối số tổng quát (theo PropBank)
+ Arg0: Tác nhân + Arg1: Thứ bị thay đổi + Arg2: Phương pháp thực hiện + Arg3: Hậu quả	+ Arg0: Tác nhân + Arg1: Thứ bị thay đổi + Arg2: Trạng thái kết thúc + Arg3: Trạng thái bắt đầu

Bảng 5: Sự phân bố các động từ PASBio vào những dạng PAS khác biệt với PropBank

Dạng 1 (34,5%)	Dạng 2 (34,5%)	Dạng 3 (3,4%)	Dạng 4 (13,8%)
alter, confer, develop, disrupt, inhibit, initiate, proliferate, skip, splice, transcribe	begin, block, decrease, generate, lead, lose, recognize, transform, translate, truncate	express	delete, encode, modify, mutate

ngữ liệu gán nhãn ngữ nghĩa GREC nêu trên. Trước tiên, ngữ liệu thô được chia ra cho từng nhóm chuyên gia. Mỗi nhóm có ít nhất 2 chuyên gia gán nhãn độc lập. Sau đó, kết quả của các chuyên gia trong cùng nhóm được đối chiếu để đánh giá mức đồng thuận. Ngữ liệu đạt yêu cầu là ngữ liệu có tỷ lệ đồng thuận cao hơn một ngưỡng định trước. Ngoài nhóm chuyên gia gán nhãn, một số bộ ngữ liệu còn có sự tham gia của giám sát viên để kiểm tra chất lượng gán nhãn và giải quyết trường hợp các chuyên gia không đạt được sự đồng thuận ở một nhãn nào đó. Ưu điểm của hướng tiếp cận này là tính chính xác cao, bám sát mục tiêu đặt ra. Tuy nhiên, thách thức là tốn nhiều thời gian và nhân lực. Ngoài ra, số lượng chuyên gia tham gia gán nhãn thường đông nên sự ảnh hưởng từ suy nghĩ chủ quan của cá nhân mỗi chuyên gia dễ dẫn đến tình trạng không nhất quán trong cách gán nhãn ngữ liệu.



Hình 1: So sánh Cấu trúc Đối số Vị ngữ của PASBio (a) và GREC (b)

Hướng tiếp cận tự động

Ở hướng tiếp cận tự động, ngữ liệu được gán nhãn hoàn toàn tự động. Độ chính xác vì vậy thấp hơn gán nhãn thủ công. Do đó, việc gán nhãn ngữ liệu tự động chủ yếu chỉ áp dụng cho những tác vụ bổ trợ bên cạnh tác vụ chính (vì nhãn của tác vụ chính cần độ chính xác cao). Tiêu biểu là các tác vụ gán nhãn từ loại, ngữ pháp[24,25]. Bài toán gán nhãn từ loại, ngữ pháp đã có bề dày nghiên cứu, đến nay gần như đạt độ chính xác của con người[26–28]. Nhờ đó, mảng này của ngữ liệu có thể gán nhãn hoàn toàn tự động mà không ảnh hưởng đến chất lượng. Nhãn ngữ pháp cũng là dữ kiện quan trọng cho hàng loạt bài toán xử lý ngôn ngữ tự nhiên, nên thường được ngầm hiểu là hiển nhiên phải có trong các bộ ngữ liệu huấn luyện của các bài toán khác. Ưu điểm của hướng tiếp cận tự động là rất tiết kiệm nhân lực và thời gian. Các mô hình học máy cũng giúp cho việc gán nhãn không bị ảnh hưởng bởi suy nghĩ chủ quan của chuyên gia. Tuy nhiên, hướng này không phù hợp với các bài toán đương đại, nơi mà độ chính xác vẫn cần được tiếp tục cải thiện.

Hướng tiếp cận bán tự động

Hướng tiếp cận bán tự động sử dụng công cụ máy tính để xây dựng những bộ ngữ liệu lớn nhằm giảm bớt gánh nặng sức người, đồng thời vẫn cần có chuyên gia để rà soát, tinh chỉnh lại. Penn Treebank là công trình đầu tiên xây dựng ngữ liệu gán nhãn ngữ pháp bằng hướng tiếp cận này[29]. Đối với bài toán SRL cho lĩnh vực Y sinh, hướng tiếp cận bán tự động cũng đã được áp dụng để xây dựng nhiều bộ ngữ liệu như BioProp[30]. Bộ ngữ liệu này được gán nhãn bằng mô hình BIOSMILE[31] và sau đó được các chuyên gia rà soát tinh chỉnh. BIOSMILE được huấn luyện trên bộ ngữ liệu BioProp[4] để xử lý bài toán SRL cho văn bản Y sinh, dựa trên sự kế thừa những thành tựu SRL trước đó ở lĩnh vực tổng quát[32–34]. Bản thân bộ ngữ liệu BioProp cũng được xây dựng bằng phương pháp bán tự động. Nhãn ngữ nghĩa của BioProp được gán tự động bằng một mô hình MaxEnt rồi được chuyên gia tinh chỉnh bằng WordFreak[35]. Đáng tiếc, bộ khung đối số mà BioProp sử dụng chủ yếu dựa trên VerbNet. Do đó, bộ ngữ liệu BioProp cũng như mô hình BIOSMILE chỉ có văn bản là Y sinh, nhưng đối số là phổ thông. Hướng tiếp cận bán tự động cũng là giải pháp phổ biến cho tác vụ tăng cường dữ liệu (data augmentation). Một trong những kỹ thuật hiệu quả để tăng cường dữ liệu là phương pháp Ví dụ ảo[2]. Nhiều công trình đã ứng dụng thành công phương pháp Ví dụ ảo để tăng cường dữ liệu của mình[36–38].

Phương pháp Ví dụ ảo (**V**irtual **E**xample - VE) có thể sinh ra hàng loạt dữ liệu mới từ dữ liệu vốn có theo bất kỳ quy tắc nào mà con người định nghĩa. Mỗi công trình có thể định nghĩa những quy tắc VE phù hợp với tác vụ của mình. Nguyên thủy, VE được sử dụng cho dữ liệu ảnh, nhằm sinh ra ảnh mới bằng việc dịch chuyển điểm ảnh của ảnh gốc[38]. Áp dụng vào dữ liệu văn bản, VE được sử dụng lần đầu ở bài toán phân loại văn bản. Quy tắc sinh VE sơ khai nhất là thêm hoặc bớt một vài từ trong văn bản gốc với nhận định rằng điều này không làm thay đổi phân loại của văn bản[2]. Trong tác vụ nhận diện thực thể Y sinh, VE có thể được sinh ra bằng cách thay mới cụm danh từ trong câu gốc, và huấn luyện ra mô hình đạt độ đo F tăng từ

2% đến 6% so với mô hình huấn luyện bằng ngữ liệu gốc, tùy theo loại thực thể[36,37]. Lợi ích từ VE không chỉ được chứng minh bằng thực nghiệm mà còn được chứng minh bằng toán học, cho thấy VE thực sự có khả năng cải thiện hiệu quả huấn luyện cho mô hình học máy[39].

Tóm lại, hướng tiếp cận bán tự động khắc phục được mặt hạn chế của hai hướng tiếp cận trước, đồng thời cũng dung hòa được ưu điểm của chúng, nhất là với hiệu quả của VE. Do đó, chúng tôi chọn hướng tiếp cận bán tự động với phương pháp Ví dụ ảo để xây dựng bộ ngữ liệu của mình.

PHƯƠNG PHÁP THỰC HIỆN

Gọi O_1 là ngữ liệu gốc gồm 317 câu minh họa của PASBio, được chọn lọc từ toàn văn của các bài báo MEDLINE sao cho bao phủ tối đa cách dùng của 29 động từ thể hiện các sự kiện sinh học phân tử quan trọng[1]. Từ ngữ liệu gốc này, quá trình xây dựng bộ ngữ liệu PASBio+ của chúng tôi gồm các bước sau: (i) Tăng cường ngữ liệu gốc bằng nguồn văn bản ngoài; (ii) Tăng cường số mẫu câu bằng biến thể ngữ pháp; (iii) Tăng cường số thể hiện (instance) của từng mẫu câu bằng ví dụ ảo (VE).

Tăng cường ngữ liệu gốc bằng nguồn văn bản ngoài

Nguồn văn bản ngoài được lựa chọn là GREC[3]. GREC có độ tương đồng nhất định với PASBio vì cùng trích văn bản từ MEDLINE. Trước hết, lọc ra tất cả những câu trong GREC có chứa 29 động từ của PASBio ở bất kỳ dạng nào, gọi bộ câu này là O_2. Kế tiếp, loại bỏ khỏi O_2 những câu chứa 2 động từ PASBio trở lên. Câu có nhiều hơn 1 động từ PASBio gây trở ngại cho phương pháp VE (xem phần "Vấn đề mất nhất quán đối số trong ví dụ ảo"). Sau cùng, với mỗi động từ PASBio, chỉ giữ lại trong O_2 3 câu dài nhất, dựa trên nhận định rằng máy tính học được nhiều hơn từ những câu dài. Vì phương pháp VE có thể phát sinh dữ liệu một cách mạnh mẽ nên bộ hạt giống không cần quá nhiều. Tiết chế số câu hạt giống giúp tiết kiệm sức lao động chuyên gia. Kết quả là bộ O_2 giữ lại 87 câu, một khối lượng để chuyên gia gán nhãn và dễ dàng hoàn thành trong thời hạn 2 tuần.

Hai chuyên gia sinh học phân tử đảm trách gán nhãn PASBio cho bộ O_2. Mỗi chuyên gia gán nhãn cho toàn bộ 87 câu trong O_2 một cách độc lập. Kết quả gán nhãn của hai chuyên gia được xem là đồng thuận nếu giống nhau cả về thành phần đối số và ranh giới mỗi đối số. Những câu chưa đồng thuận được hai chuyên gia thảo luận cho đến khi đạt được sự nhất trí.

Tăng cường số mẫu câu bằng biến thể ngữ pháp

Khác với O_2, O_1 có thể chứa những câu có từ hai động từ PASBio trở lên. Tách O_1 thành O_{1a} gồm 88 câu chỉ có đúng 1 động từ PASBio và O_{1b} chứa 229 câu còn lại. Chỉ có O_{1a} cùng O_2 tham gia vào quá trình phát sinh VE. Nhưng trước khi phát sinh VE, chúng tôi tăng cường bộ hạt giống bằng biến thể ngữ pháp. Tất cả câu trong O_{1a}, O_{1b} và O_2 đều được phát sinh biến thể ngữ pháp. Có hai phép biến đổi ngữ pháp thường gặp nhất trong văn bản khoa học: (i) Thêm hoặc khử mệnh đề tính từ (xem ví dụ ở Bảng 6); (ii) Chuyển đổi thụ động cách/chủ động cách (xem ví dụ ở Bảng 7). Để tối đa hóa số biến thể, phép biến đổi (ii) được áp dụng lên cả những biến thể ngữ pháp sinh ra bởi phép biến đổi (i) nếu có thể. Biến thể ngữ pháp được viết bởi một chuyên gia Anh ngữ, sau đó được kiểm tra lại lần nữa bởi một chuyên gia Anh ngữ khác để rà soát lỗi ngữ pháp và chính tả.

Kết quả cho ra bộ G gồm G_{1a}, G_{1b} và G_2 lần lượt là các bộ biến thể ngữ pháp của O_{1a}, O_{1b} và O_2, đạt tổng cộng 1010 câu. Gọi Seed = $\{O_{1a}, G_{1a}, O_2, G_2\}$ là bộ hạt giống đầu vào cho bước tiếp theo để phát sinh VE.

Tăng cường số thể hiện của từng mẫu câu bằng VE

Cùng một mẫu câu, việc thay một bộ đối số phù hợp vào vị trí tương ứng của các đối số cũ sẽ cho một thể hiện (instance) mới của mẫu câu ấy. Một VE phát sinh tự động như vậy được gọi là đúng đắn khi nhãn của bài toán quan tâm trong VE không bị thay đổi nếu đem cho chuyên gia gán thủ công[2]. Để thực hiện điều này, hai quy tắc phát sinh VE được định nghĩa: (i) Quy tắc trao đổi đối số; (ii) Quy tắc thay thế đối số.

Quy tắc trao đổi đối số

Từ cặp câu gốc cùng vị ngữ, cùng thành phần đối số hiện diện, tạo ra cặp câu mới bằng cách trao đổi bộ đối số giữa cặp câu gốc. Chúng tôi gọi S là tập hợp các VE được sinh ra. Ví dụ:

Cặp câu gốc:

(1) [Transcriptional stimulation]$_{A1}$ has been [abolished]$_P$ by [further deletion of the C-terminal transactivation domain in the Pax5 mutants B8 and B9]$_{A0}$.

(2) [Its BCFA biosynthesis]$_{A1}$ is believed to be [abolished]$_P$ by [this complete removal of FabD from the crude FAS]$_{A0}$.

Cặp ví dụ ảo được sinh ra:

(1') [Its BCFA biosynthesis]$_{A1}$ has been [abolished]$_P$ by [this complete removal of FabD from the crude FAS]$_{A0}$.

Bảng 6: Ví dụ về biến thể ngữ pháp tạo ra bằng khử Mệnh đề tính từ

Câu gốc	Biến thể ngữ pháp
Patient 1 has [a G-to-A transition at the first nucleotide of intron 2]$_{A0}$, which [abolishes]P [normal splicing]$_{A1}$.	[A G-to-A transition at the first nucleotide of intron 2]$_{A0}$ of patient 1 [abolishes]P [normal splicing]$_{A1}$.

Bảng 7: Ví dụ về biến thể ngữ pháp tạo ra bằng chuyển Thụ động cách sang Chủ động cách

Câu gốc	Biến thể ngữ pháp
[The DNA-binding mutant (Stat5aEE-AA) of Stat5a]$_{A1}$ was [generated]P by [mutating amino acids EE (437 and 438) to AA]$_{A0}$.	[The mutation of amino acids EE (437 and 438) to AA]$_{A0}$ [generated]P [the DNA-binding mutant (Stat5aEE-AA) of Stat5a]$_{A1}$.

(2')[Transcriptional stimulation]$_{A1}$ is believed to be [abolished]$_P$ by [further deletion of the C-terminal transactivation domain in the Pax5 mutants B8 and B9]$_{A0}$.

Quy tắc thay thế đối số

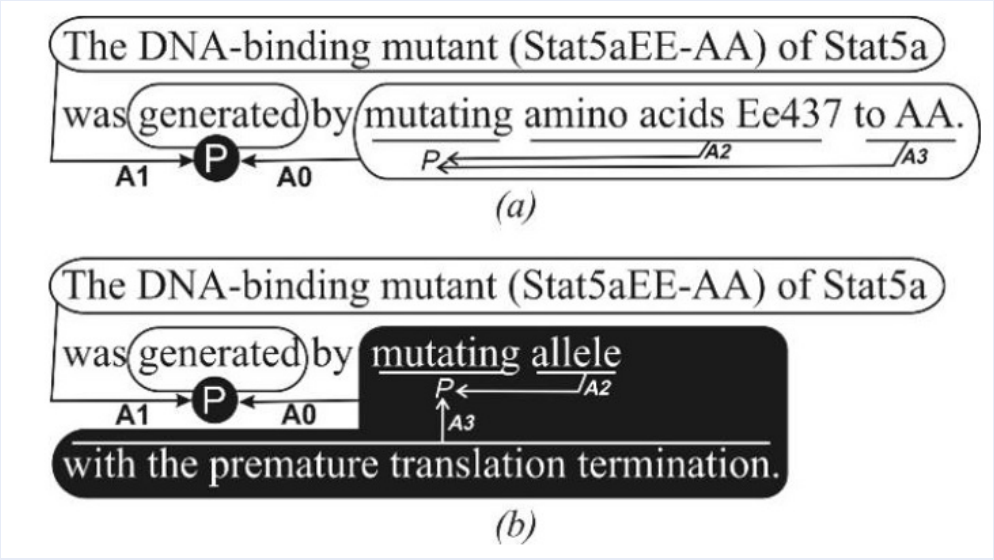
Tạo ra câu mới bằng cách thay thế bộ đối số của câu gốc bằng bộ đối số của một câu khác (câu nguồn) có cùng vị ngữ với câu gốc và có đủ đối số để thay thế. Chúng tôi gọi R là tập hợp các VE được sinh ra. Ví dụ:
Câu nguồn: Besides these side-chain interactions with the 06-alkyl group, structure-based analysis of mutational data suggests that [substitutions at Gly156 and Lys165]$_{A0}$ [confer]$_P$ [resistance]$_{A1}$ to [O6-BG]$_{A2}$ through backbone distortions.
Câu gốc: [The portion of the STATs]$_{A0}$ [conferring]$_P$ [specificity]$_{A1}$ for [either a MAPK or a MAPK substrate kinase (MAPKAP)]$_{A2}$ has not been determined.
V í dụ ảo được sinh ra: [Substitutions at Gly156 and Lys165]$_{A0}$ [conferring]$_P$ [resistance]$_{A1}$ for [O6-BG]$_{A2}$ has not been determined.
Hai câu được gọi là có cùng vị ngữ khi vị ngữ của chúng có cùng nguyên mẫu (Ví dụ: Mutate, mutating, mutation, mutated được xem là cùng một vị ngữ). Kết quả phát sinh ví dụ ảo tạo ra 675 câu ở bộ R và 528 câu ở bộ S, tổng cộng 1203 câu ví dụ ảo. Mã giả của hai quy tắc này có thể được tham khảo ở Phụ lục A.

Vấn đề mất nhất quán đối số trong ví dụ ảo

Nguyên thủy, VE không quan tâm tính logic và tính thực tế mà chỉ quan tâm đến các nhãn của tác vụ đang theo đuổi[2]. Nói cách khác, khi ẩn đi các nhãn trong một VE rồi cho chuyên gia gán nhãn lại, nếu kết quả không đổi thì VE ấy là đúng (Tham khảo chi tiết trong Phụ lục B). Tuy nhiên, một yêu cầu khác cao hơn được đặt ra trong bài báo này khi áp dụng VE vào bài toán SRL: Tất cả các đối số trong VE không chỉ là đứng đúng chỗ của nó mà còn phải thực sự phù hợp với nhau theo tri thức Y sinh (còn những phần ở ngoài khung đối số sẽ không thuộc phạm vi quan tâm). Từ đó đặt ra vấn đề cần tránh: Vấn đề mất nhất quán đối số.
Mất nhất quán đối số trong ví dụ ảo xảy ra khi việc thay mới một bộ đối số lại vô tình làm tổn thương một bộ đối số khác đồng thời tồn tại trong câu. Hình 2 minh họa một ví dụ là câu có hai động từ PASBio ("generate" và "mutate"). Việc phát sinh ví dụ ảo bằng quy tắc thay thế cho A2 và A3 của "mutate" đã vô tình làm cho "generate" phải nhận một A0 mới không phù hợp với A1 cũ của nó (Hình 2b). Bộ đối số mới của "generate" bị sai về tri thức Y sinh: Đột biến liên kết DNA Stat5sEE-AA không sinh ra từ đột biến alene do gián đoạn dịch mã.
Mất nhất quán đối số chỉ có nguy cơ với những câu chứa từ 2 động từ PASBio trở lên, chúng đã được lọc vào bộ O_{1b}. Bộ Seed dùng để phát sinh VE không bao gồm O_{1b}. Tuy O_{1b} có tham gia vào Quy tắc thay thế, nhưng chỉ đóng vai trò câu cho chứ không phải câu nhận. Vì vậy, các VE của chúng tôi hoàn toàn tránh được vấn đề này.

KẾT QUẢ THỬ NGHIỆM VÀ THẢO LUẬN

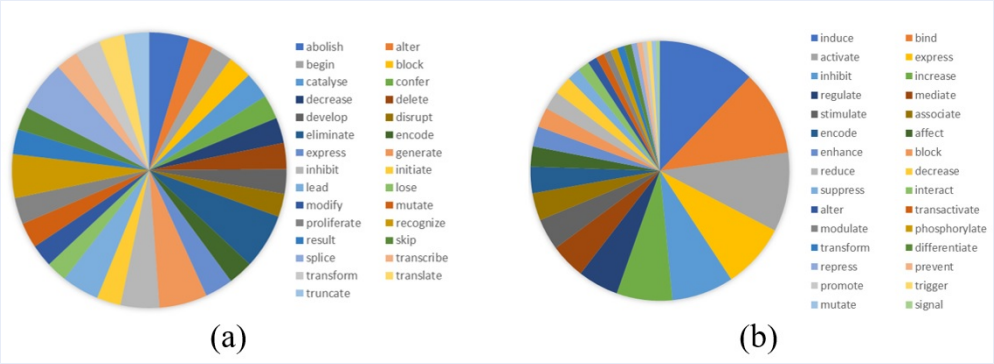
Số câu của từng thành phần ngữ liệu và độ phân bố của các vị ngữ Y sinh trong PASBio+ được thống kê tương ứng trong Bảng 8 và Hình 3. Như vậy, từ 317 câu gốc (O1), bộ ngữ liệu kết quả đạt 2617 câu gán nhãn PASBio. Sự phân bố các vị ngữ trong PASBio+ rất đồng đều (Hình 3a), không tồn tại những tần suất thấp dưới 0,1% như ở BioProp (Hình 3b). Điều này giúp tránh vấn đề dữ liệu thưa (data sparsity), từ đó hạn chế lỗi quá khớp (overfitting).
Ở bước Tăng cường ngữ liệu gốc bằng nguồn văn bản ngoài, sản phẩm là bộ O_2 được gán nhãn hoàn toàn thủ công bởi hai chuyên gia inh học phân tử làm việc độc lập. Những nhãn trong O_2 đã đạt được sự đồng thuận từ kết quả làm việc của hai chuyên gia nên tính đúng đắn được đảm bảo.
Ở bước Tăng cường số mẫu câu, bộ biến thể ngữ pháp G được viết bởi chuyên gia Anh ngữ và kiểm tra bởi



Hình 2: Câu gốc (a) và câu ví dụ ảo sinh ra bị mất nhất quán đối số của vị ngữ "generate" (b)

Bảng 8: Số câu trong các thành phần

Thành phần	O1	O2	G	R		Tổng
Số câu	317	87	1010	675	528	2617



Hình 3: Phân bố vị ngữ Y sinh trong PASBio+ (a) và BioProp (b)

chuyên gia Anh ngữ thứ hai nên đảm bảo tính đúng đắn về Anh ngữ. Quá trình này chỉ thay đổi cấu trúc tổng quát của câu nhưng giữ nguyên bộ đối số nên bảo toàn tính đúng đắn về chuyên môn Y sinh vốn có của câu gốc.

Ở bước phát sinh ví dụ ảo, các đối số cũ và mới đều thuộc cùng một vị ngữ. Đối số được thay thế hoặc trao đổi theo từng bộ trọn vẹn, không tách lẻ. Việc các đối số luôn đi cùng nhau một bộ đảm bảo tri thức Y sinh mà bộ đối số chuyển tải được bảo toàn như câu gốc, không bị thất thoát hay thay đổi.

Để đánh giá hiệu quả của PASBio+ trong cải thiện mô hình học máy, bốn mô hình độc lập M_1–M_4 được huấn luyện với ngữ liệu huấn luyện lần lượt là O_1 + O_2 gồm 404 câu, G gồm 1010 câu, R + S gồm 1203 câu, G + R + S gồm 2213 câu. Ngoài ra, việc so sánh hiệu quả giữa ngữ liệu chuyên biệt vào Y sinh và ngữ liệu tổng quát đòi hỏi huấn luyện thêm mô hình M_5 trên PropBank. Năm mô hình được đánh giá trên cùng ngữ liệu đánh giá là O_1 + O_2. Riêng M_1, vì ngữ liệu huấn luyện và đánh giá trùng nhau, nên được áp dụng đánh giá chéo 10 pha: Số câu của mỗi động từ

PASBio được chia đều ra 10 phần, luân phiên 9 phần để huấn luyện và 1 phần để đánh giá.

Mô hình huấn luyện là mô hình học sâu cho SRL trên văn bản Y sinh[40]. Đây là một mô hình học sâu dựa trên mạng Bi-LSTM có tăng cường thêm Kết nối Cao tốc (Highway Connection). Điều này giúp quá trình lan truyền trong mạng Nơ-ron có thể thực hiện trực tiếp giữa những tầng không liên tiếp, nhờ đó giảm thiểu sự thất thoát đạo hàm. Mô hình này có tích hợp học đa tác vụ với tác vụ bổ trợ là bài toán NER. Tuy nhiên, do PASBio+ không gán nhãn NER nên ở đây chỉ áp dụng phiên bản đơn tác vụ của mô hình này.

Bảng 9 trình bày kết quả thử nghiệm của mô hình M_{2-4} so với M_1. M_2 giúp tăng điểm F thêm 44,4% cho thấy tác dụng tích cực khi máy tính được huấn luyện trên một tập mẫu câu phong phú hơn. M_3 với tập huấn luyện hoàn toàn là các VE phát sinh tự động nhưng giúp tăng điểm F thêm đến 49%, cao hơn cả M_2 dù cho tập huấn luyện của M_2 được viết hoàn toàn thủ công. Điều này cho thấy hiệu quả và triển vọng của VE rất đáng được chú ý. M_4 kết hợp tập huấn luyện của cả M_2 và M_3 giúp tăng điểm F thêm đến 52,2%. Ngoài ra, việc M_4 đạt điểm F cao hơn M_5 22,5% cho thấy ý nghĩa tích cực của một bộ ngữ liệu được gán nhãn chuyên biệt cho lĩnh vực Y sinh để giải bài toán Y sinh, thay vì dùng ngữ liệu của lĩnh vực tổng quát.

KẾT LUẬN

Một bộ ngữ liệu PASBio+ gồm 2617 câu được gán nhãn ngữ nghĩa với khung đối số chuyên biệt cho lĩnh vực Y sinh đã được xây dựng. Bộ ngữ liệu theo định dạng xml, tuân thủ cấu trúc thẻ trong PASBio gốc, và có thể được tải về tại đây. So với 317 câu gốc của PASBio, bộ ngữ liệu này được làm giàu số mẫu câu bằng cách thu thập và gán nhãn thêm nguồn câu phù hợp từ GREC, cũng như biên soạn thêm biến thể ngữ pháp của các câu gốc. Ngữ liệu cũng được tăng cường số thể hiện của từng mẫu câu bằng phương pháp Ví dụ ảo. Kết quả thử nghiệm cho thấy từng thành phần tăng cường đều đóng góp vào sự cải thiện hiệu quả học máy.

Hướng phát triển kế tiếp là nghiên cứu thêm những kỹ thuật tăng cường dữ liệu khác nhằm bổ trợ cho phương pháp Ví dụ ảo, giúp tạo ra dữ liệu lớn hơn. Ngoài ra, bộ ngữ liệu PASBio+ cũng hướng đến tích hợp thêm nhãn thực thể (NER) để ứng dụng học đa tác vụ, một kỹ thuật nhiều tiềm năng trong học sâu.

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 2021-13. Chúng tôi xin chân thành cảm ơn Ban Giám hiệu Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM cùng quý đồng nghiệp đã tạo điều kiện hỗ trợ hết mình giúp chúng tôi hoàn thành mục tiêu đề tài.

DANH MỤC CÁC TỪ VIẾT TẮT

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

SRL: Semantic Role Labelling (Gán nhãn Vai trò Ngữ nghĩa)

VE: Virtual Example (Ví dụ ảo)

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

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 báo.

Tác giả Phạm Hữu Sang và Hoàng Văn Thức tham gia khảo sát hiện trạng, đề xuất giải pháp, lập trình công cụ và triển khai thử nghiệm.

PHỤ LỤC A: ĐẶC TẢ HÌNH THỨC QUY TẮC PHÁT SINH VÍ DỤ ẢO

Ký hiệu: Gọi x.P, x.L và x.A_k lần lượt là vị ngữ (động từ), tập nhãn đối số có hiện diện và đối số thứ k của câu x. Gọi X[i] là câu thứ i trong bộ câu X (Ví dụ: Seed[1] là câu thứ nhất trong bộ Seed, O_{1b}[2] là câu thứ hai trong O_{1b})

Với cặp câu x và y, chúng tôi định nghĩa hai hàm cơ sở được đặc tả bằng mã giả như sau:

Replace (x, y) //Hàm tạo ra câu z bằng cách
//thay đối số trong câu x bằng đối số tương ứng trong câu y
{ z = x.Clone(); //z là bản sao của x
$\forall$ z.A_k : z.$A_k \leftarrow$ y.A_k;
Return z;
}

Swap (x, y) //Hàm tạo ra cặp câu mới z, t từ cặp câu gốc x, y
//bằng cách hoán đổi đối tương ứng giữa x và y
{ z = Replace (x, y);
t = Replace (y, x);
Return {z, t}; //trả về 1 cặp câu
}

Từ hai hàm cơ sở, hai quy tắc phát sinh VE được đặc tả như sau:

(i) Quy tắc trao đổi: Từ cặp câu gốc cùng vị ngữ, cùng thành phần đối số hiện diện, tạo ra cặp câu mới bằng cách trao đổi bộ đối số giữa cặp câu gốc.

Bảng 9: Kết quả thử nghiệm của năm mô hình với ngữ liệu huấn luyện tương ứng

M1: $O_1 + O_2$			M5: PropBank			M2: G			M3: R + S			M4: G + R + S		
	R	F		R	F		R	F		R	F		R	F
16,5	13,3	14,7	56,4	32,7	41,4	61,5	56,9	59,1	67,8	60,1	63,7	70,2	63,9	66,9
Cải thiện so với ngữ liệu gốc (M1)						45	43,6	44,4	51,3	46,8	49	53,7	50,6	52,2
Cải thiện so với ngữ liệu tổng quát (M5)						5,1	24,2	17,7	11,4	27,4	22,3	13,8	31,2	25,5

```
S = ∅; // Ban đầu chưa có VE nào.
∀ Seed[i]
  ∀ Seed[j]
    If (i ≠ j) ∧ (Seed[i].P = Seed[j].P)
      If (Seed[i].L = Seed[j].L)
        S = S + Swap(Seed[i], Seed[j]);
```

(ii) Quy tắc thay thế: Tạo ra câu mới bằng cách thay thế bộ đối số của câu gốc bằng bộ đối số của câu khác (câu nguồn) có cùng vị ngữ với câu gốc và có đủ đối số để thay thế.

```
R = ∅; // Ban đầu chưa có VE nào.
∀ Seed[i]
  ∀O_1b[j]
    If (Seed[i].P = O_1b[j].P)
      If (Seed[i].L ∩ O_1b[j].L = Seed[i].L)
        { new z = Replace(Seed[i], O_1b[j]);
        R = R + {z}; }
```

PHỤ LỤC B: PHẠM VI LOGIC CỦA PHƯƠNG PHÁP VÍ DỤ ẢO

Nguyên thủy, phương pháp Ví dụ ảo không quan tâm tính logic và tính thực tế của các ví dụ ảo (VE) được tạo ra. Ban đầu, VE được tạo ra với dữ liệu ảnh bằng việc dịch chuyển một số điểm ảnh trong ảnh gốc, dù cho ảnh tạo ra theo mắt người nhìn sẽ kỳ quặc nhưng phân loại của nó sẽ không đổi[38]. Áp dụng vào dữ liệu văn bản, VE được sinh ra bằng cách thêm bớt vài từ[2] hoặc thay mới cả cụm danh từ[36,37] dù cho điều đó tạo ra một câu có nội dung không thực tế. Điều mà VE quan tâm là tạo ra nhiều mẫu mới để huấn luyện trong học máy, nên một VE gọi là đúng đắn nếu khi ta ẩn đi các nhãn của tác vụ ta quan tâm và cho chuyên gia gán nhãn thủ công thì nhãn được gán sẽ không khác gì so với các nhãn đã ẩn đi, dù cho nội dung câu ấy sai thực tế. Từ trong bản chất, hai quy tắc phát sinh ví dụ ảo mà chúng tôi đề xuất cho bài toán SRL đã đảm bảo được tiêu chí này vì các đối số được thay thế cho nhau không chỉ là của cùng vị ngữ mà còn cùng loại.

Tuy nhiên, chúng tôi còn tự đặt ra cho mình thêm một yêu cầu cao hơn khi áp dụng VE vào bài toán SRL: Tất cả các đối số trong VE không chỉ là đứng đúng chỗ của nó mà còn phải thực sự phù hợp với nhau theo tri thức Y sinh. Từ đó đặt ra vấn đề mất nhất quán đối số mà chúng tôi đã đề xuất giải pháp khắc phục (xem phần *Vấn đề mất nhất quán đối số trong VE*).

Ngoài ra, những nội dung không thuộc về khung đối số của bài toán SRL sẽ không thuộc phạm vi quan tâm của công trình. Do đó, các VE có thể chứa thông tin phi thực tế nếu thông tin ấy nằm ngoài nội dung Cấu trúc Đối số Vị ngữ. Để dễ hình dung, trước khi nêu ví dụ Y sinh, xin bắt đầu bằng một ví dụ trong lĩnh vực tổng quát, xét câu gốc sau với vị ngữ P là "sáng lập":

"[Bill Gates]$_{A0}$ [sáng lập]$_P$ [tập đoàn Microsoft]$_{A1}$ khi chỉ mới 20 tuổi"

Ta thay nguyên vẹn một bộ đối số khác vào câu gốc này để tạo ra VE sau:

"[Larry Page cùng Sergey Brin]$_{A0}$ [sáng lập]$_P$ [cỗ máy tìm kiếm mang tên Google]$_{A1}$ khi chỉ mới 20 tuổi"

Rõ ràng VE này phi thực tế vì khi sáng lập Google thì Larry Page 25 tuổi chứ không phải 20 tuổi. Tuy nhiên, đó là chi tiết ở ngoài cấu trúc đối số vị ngữ nên phương pháp Ví dụ ảo không quan tâm. Đây vẫn là một VE đúng đắn vì nếu cho con người gán nhãn lại thì kết quả vẫn vậy. Hơn nữa, trong khuôn khổ bộ đối số, Larry Page và Sergey Brin quả thật đã sáng lập Google nên thông tin trong phạm vi cấu trúc đối số vị ngữ là chính xác. Nó còn hữu ích vì đóng góp một A0 có cấu trúc phức tạp (2 tên người thay vì câu gốc chỉ có 1 tên người). Những ví dụ có cấu trúc phong phú hơn luôn có ý nghĩa tích cực cho huấn luyện trong học máy.

Đến đây ta phân tích ví dụ Y sinh đã nêu trong phần *Quy tắc thay thế đối số*, xét câu gốc:

"[The portion of the STATs]$_{A0}$ [conferring]$_P$ [specificity]$_{A1}$ for [either a MAPK or a MAPK substrate kinase (MAPKAP)]$_{A2}$ has not been determined."

Và một ví dụ ảo tạo ra từ nó:

"[Substitutions at Gly156 and Lys165]$_{A0}$ [conferring]$_P$ [resistance]$_{A1}$ for [O6-BG]$_{A2}$ has not been determined."

Ví dụ ảo này phi thực tế, vì việc thay thế ở gốc Gly156 và protein Lys165 gây ra tính kháng O^6-Benzylguanin là đã xác định chứ không phải chưa xác định. Nhưng "xác định rồi" hay "chưa xác định" là thông tin không thuộc về cấu trúc đối số vị ngữ, đây vẫn là một VE

đúng đắn cho học máy vì bộ đối số của nó không những đúng về vị trí đối số, ranh giới đối số mà còn đúng ở mối quan hệ của các đối số với nhau. Nghĩa là trong tri thức Y sinh, các đối số A0, A1 và A2 này thực sự đi chung với nhau trong sự kiện "confer", thể hiện ở câu nguồn sau:

"Besides these side-chain interactions with the 06-alkyl group, structure-based analysis of mutational data suggests that [substitutions at Gly156 and Lys165]$_{A0}$ [confer]$_{P}$ [resistance]$_{A1}$ to [O6-BG]$_{A2}$ through backbone distortions."

Vì điều này, các ví dụ ảo tuy không hoàn toàn trùng khớp với những gì xảy ra trong thế giới thực nhưng vẫn hoàn thành tốt vai trò của ngữ liệu huấn luyện trong học máy.

TÀI LIỆU THAM KHẢO

1. Wattarujeekrit T, et al. PASBio: predicate-argument structures for event extraction in molecular biology, BMC Bioinformatics 5. 2004; 155-174;Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-155.
2. Sasano M. Virtual examples for text classification with support vector machines, Proceedings of 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003). 2003; 208-215;Available from: https://aclanthology.org/W03-1027/.
3. Paul T. Building a bio-event annotated corpus for the acquisition of semantic frames from biomedical corpora, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). 2008; 2159-2166;Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.499.8282.
4. Chou W-C, Tsai RT-H, Su Y-S, Ku W, Sung T-Y, Hsu W-L. A semi-automatic method for annotating a biomedical proposition bank, Proceedings of ACL Workshop on Frontiers in Linguistically Annotated Corpora. 2006;5-12;Available from: https://doi.org/10.3115/1641991.1641993.
5. Marc W. Kirschner. The role of biomedical research in health care reform, New Series - Vol. 266 (5182) 1994; 49-51;Available from: https://doi.org/10.1126/science.7939643.
6. Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: An Overview, Comput Biol. 2003; 10(6):821-55;PMID: 14980013. Available from: https://doi.org/10.1089/106652703322756104.
7. Hajic J. Final report: Natural Language Generation in the Context of Machine Translation, The Center for Language and Speech Processing, The Johns Hopkins University. 2004;Available from: https://www.cs.jhu.edu/~jason/papers/hajic+al.ws02.pdf.
8. Andreas H, Andreas N, Gerhard P. A Brief Survey of Text Mining, GLDV Journal for Computational Linguistics and Language Technology. 2005;19-62;.
9. Jin-Dong K, Tomoko O, Jun'ichi T. Corpus annotation for mining biomedical events from literature, BMC Bioinformatics 9(1):10. 2008; 1-25;PMID: 18182099. Available from: https://doi.org/10.1186/1471-2105-9-10.
10. Han C. Handling Structural Divergences and Recovering Deropped Arguments in a Korean/English Machine Translation System, Association for MT in the Americas. 2000; 40-53;Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.6713.
11. Rindflesch TC. Extracting molecular binding relationships from biomedical text, 6th Conference on Applied Natural Language Processing. 2000; 188-195;Available from: https://doi.org/10.3115/974147.974173.
12. Jae-Hong Eom, B young-Tak Zhang. PubMiner: Machine Learning-based Text Mining for Biomedical Information Analysis, International Conference on Artificial Intelligence: Methodology, Systems, and Applications. 2004; 216-225;Available from: https://doi.org/10.1007/978-3-540-30106-6_22.
13. Tanabe L. MedMiner: An internet text-mining tool for biomedical information, with application to gene expression profiling, BioTechniques Vol. 27 (6) 2018; 1210-1217;PMID: 10631500. Available from: https://doi.org/10.2144/99276bc03.
14. Novichkova S. MedScan, a NLP engine for Medline abstracts, Bioinformatics. 2003; 1699-1706;PMID: 12967967. Available from: https://doi.org/10.1093/bioinformatics/btg207.
15. Sekimizu T. Identifying the interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts, Genome Inform. 1998; 62-71;.
16. Andrew D. BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies, KR-MED 2006 "Biomedical Ontology in Action". 2006; 87-94;Available from: http://ceur-ws.org/Vol-222/krmed2006-p10.pdf.
17. Sampo P. BioInfer: a corpus for information extraction in the biomedical domain, BMC Bioinformatics. 2007; 1-24;PMID: 17291334. Available from: https://doi.org/10.1186/1471-2105-8-50.
18. Kipper K, Dang HT, Palmer M. Class based construction of a verb lexicon, 17th National Conference on Artificial Intelligence (AAAI-2000). Austin, TX. 2000; 691-696;.
19. Baker CF. The Berkeley FrameNet project, 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics. 1998; 86-90;Available from: https://doi.org/10.3115/980845.980860.
20. Kingsbury P, Palmer M. From Treebank to PropBank, 3rd International Conference on Language Resources and Evaluation. 2002. 1989-1993;Available from: http://www.lrec-conf.org/proceedings/lrec2002/pdf/283.pdf.
21. Kingsbury P, Palmer M, Marcus M. Adding Semantic Annotation to the Penn TreeBank, Human Language Technology Conference. 2002; 1-5;.
22. Neves M, Leser U. A survey on annotation tools for the biomedical literature, Brief Bioinform. 2014; 327-340;PMID: 23255168. Available from: https://doi.org/10.1093/bib/bbs084.
23. Tateisi. Syntax Annotation for the GENIA Corpus, Proceedings of the Second International Joint Conference on Natural Language Processing: Companion Volume. 2005; 220-225;Available from: https://aclanthology.org/I05-2038/.
24. Francis W. Nelson, Henry Kucera. Computational Analysis of Present-Day American English. International Journal of American Linguistics Vol.5 Num.1. 1969; 71-75;Available from: https://doi.org/10.1086/465045.
25. Garside R, and Smith N. A hybrid grammatical tagger: CLAWS4, Linguistic Information from Computer Text Corpora. Longman, London. 1997; 102-121;.
26. DeRose, Steven J. 1988. Grammatical category disambiguation by statistical optimization, Computational Linguistics 14(1). 1988;31-39;.
27. Kenneth Ward Church. A stochastic parts program and noun phrase parser for unrestricted text, Proceedings of the Second Conference on Applied Natural Language Processing. Association for Computational Linguistics Stroudsburg. 1988; 136-143;.
28. van Halteren H. Improving Accuracy in NLP Through Combination of Machine Learning Systems. Computational Linguistics. 27(2). 2001; 199-229;Available from: https://doi.org/10.1162/089120101750300508.
29. Taylor A. The Penn Treebank: An Overview. Abeillé A. (eds) Treebanks. Text, Speech and Language Technology, vol 20. Springer, Dordrecht. 2003; 5-22;Available from: https://doi.org/10.1007/978-94-010-0201-1_1.
30. Tsai RT, et al. Semi-automatic conversion of BioProp semantic annotation to PASBio annotation, BMC Bioinformatics 9, S18. 2008; 1-12;Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-S12-S18.

31. Tsai RT-H. BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features, BMC Bioinformatics 2007,8(1). 2007; 325-339;.
32. Gildea D. Automatic Labeling of Semantic Roles, Proceedings of the 38th Annual Conference of the Association for Computational Linguistics. 2002; 512-520;.
33. Pradhan. Semantic Role Labeling Using Different Syntactic Views, Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). 2005. 581-588 ;Available from: https://doi.org/10.3115/1219840.1219912.
34. Surdeanu. Using Predicate-Argument Structures for Information Extraction, Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 2003; 8-15;Available from: https://dl.acm.org/doi/10.3115/1075096.1075098.
35. Morton, Thomas, LaCivita, Jeremy. WordFreak: An Open Tool for Linguistic Annotation, Proceedings of HLT-NAACL. 2003; 17-18;Available from: https://doi.org/10.3115/1073427.1073436.
36. Eunji Y, et al. SVM-Based Biological Named Entity Recognition Using Minimum Edit-Distance Feature Boosted by Virtual Examples, International Joint Conference on Natural Language Processing. 2004; 807-814;Available from: https://doi.org/10.1007/978-3-540-30211-7_86.
37. Song Y, Kim E, Lee GG, Yi BK. POSBIOTM-NER: a trainable biomedical named-entity recognition system, Bioinformatics. 2005; 2794-2796;PMID: 15814561. Available from: https://doi.org/10.1093/bioinformatics/bti414.
38. Schölkopf B, Burges C, Vapnik V. Incorporating invariances in support vector learning machines, Lecture Notes in Computer Science. 1996; 47-52;Available from: https://doi.org/10.1007/3-540-61510-5_12.
39. Niyogi P, et al. Incorporating prior information in machine learning by creating virtual examples, Proceedings of IEEE, vol. 86. 1998; 2196-2207;Available from: https://doi.org/10.1109/5.726787.
40. Đức H. A deep-learning model for semantic role labelling in medical documents, Science and Technology Development Journal - Natural Sciences, 5(2), 2021; 1032-1039;Available from: http://stdjns.scienceandtechnology.com.vn/index.php/stdjns/article/view/928.

Science & Technology Development Journal – Natural Sciences 2022, 6(2):2083-2094

Open Access Full Text Article

Research article

A semi-automatic approach to biomedical semantic role corpus construction

Tuan Nguyen Hoai Duc[1,*], Hoang Van Thuc[2], Pham Huu Sang[3]

ABSTRACT

A semi-automatic solution to build a iomedical semantic role corpus named PASBio+ was proposed. The corpus was annotated with a predicate argument tructure, the important information that revealed the main content of a sentence. Because more than 86% of the arguments in the biomedical domain significantly differed from those in the general domain, this proposed corpus was labeled on top of 317 labeled sentences from PASBio, the argument frameset specifically designed for the Biomedical domain. From these sentences, the proposed semi-automatic solution additionally generated 87 sentences which were manually annotated by our experts. More instances were further generated by using the virtual example method, a powerful and flexible data augmentation technique that ha been successfully applied in a wide range of tasks. Specifically, two sequential rules (the swap rule and the replace rule) were proposed to ensure that the biomedical knowledge was always kept correct. PASBio+ was also augmented by adding grammatical variants of the original sentences which kept the corpus having a wide coverage of diverse natural writing styles. In addition, from the very beginning, the PASBio's original sentence set was also enriched by an external text source which was an additional set of sentences selected from GREC iomedical corpus. As a result, a corpus with 2,500 fully labeled sentences with a uniform frequency distribution among predicates was obtained, thereby eliminating the problem of data sparsity and helping to restrict the overfitting in machine learning. The experimental results showed that when using the augmented corpus to train a semantic role labeling model, an increase in the F score by 52.2% or 22.5% were obtained compared to those trained by using the original PASBio corpus or a general domain one, respectively.

Key words: predicate-argument structure, semantic role labelling, corpus construction, data augmentation

[1]*Faculty of Information Technology, University of Sciences, VNUHCM, Vietnam*

[2]*Viettel Business Solutions Corporation, Vietnam*

[3]*VietBanDo Solution CO.,LTD, Vietnam*

Correspondence

Tuan Nguyen Hoai Duc, Faculty of Information Technology, University of Sciences, VNUHCM, Vietnam

Email: tnhduc@fit.hcmus.edu.vn

History

- Received: 30-11-2021
- Accepted: 05-6-2022
- Published: 30-6-2022

DOI : 10.32508/stdjns.v6i2.1151

Cite this article : Duc T N H, Thuc H V, Sang P H. **A semi-automatic approach to biomedical semantic role corpus construction**. *Sci. Tech. Dev. J. - Nat. Sci.;* 2022, 6(2):2083-2094.