Tạp chí Phát triển Khoa học và Công nghệ – Khoa học Tự nhiên 2025, 9(3):3402-3415

Open Access Full Text Article

Bài nghiên cứu

Giải thích tầm quan trọng của đặc trưng từ vựng trong tác vụ gán nhãn ngữ nghĩa

Tuấn Nguyên Hoài Đức*

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

TÓM TẮT

Mặc dù các mô hình học sâu mang lại hiệu quả dự đoán ấn tượng trong các tác vụ xử lý ngôn ngữ tự nhiên (NLP), chúng cũng dấy lên lo ngại về độ tin cậy và những vấn đề đạo đức bởi tính chất hộp đen của chúng. Vì vậy, một bài toán mới đang thu hút nhiều nỗ lực nghiên cứu là giải thích mô hình NLP thông qua việc làm rõ tầm quan trọng của các đặc trưng đầu vào đối với dự đoán đầu ra của các mô hình học sâu này. Trong các mức dự đoán của NLP, dự đoán mức chuỗi từ liên quan đến nhiều tác vụ quan trọng như nhận diện thực thể, gán nhãn ngữ nghĩa, rút trích sự kiện… Tuy nhiên, dự đoán mức chuỗi từ lại chưa được quan tâm trong các kỹ thuật giải thích hiện có. Bài báo này giới thiệu một cách tiếp cận giải thích tầm quan trọng của đặc trưng từ vựng đối với dự đoán mức chuỗi từ trong tác vụ gán nhãn ngữ nghĩa. Phương pháp của chúng tôi có thể đánh giá đồng thời tầm ảnh hưởng và độ hữu ích của từng từ vựng trong câu đối với kết quả dự đoán. Chúng tôi cũng đề xuất một cách tiếp cận mới về cơ chế làm nhiễu dữ liệu dựa trên việc lựa chọn từ thay thế một cách thông minh nhằm che giấu hiệu quả đặc trưng từ vựng, qua đó khắc phục những hạn chế của các kỹ thuật làm nhiễu dữ liệu phổ biến hiện nay. Thông qua thực nghiệm trên văn bản Y Sinh, chúng tôi chứng minh hiệu quả phương pháp giải thích của mình trong việc cung cấp lời giải thích vừa dễ hiểu đối với con người, vừa trung thực với quá trình xử lý bên trong mô hình NLP.

Từ khoá: Tính khả diễn giải, NLP khả diễn giải, giải thích mô hình, gán nhãn ngữ nghĩa văn bản

GIỚI THIỆU

Trong xử lý ngôn ngữ tự nhiên (NLP), việc chuyển từ các mô hình trong suốt cổ điển như mô hình dựa trên luật và cây quyết định sang các mô hình học sâu như BERT đã gây ra lo ngại về độ tin cậy của các mô hình hộp đen này. Độ phức tạp của các mô hình học sâu khiến con người khó hiểu rõ những yếu tố chi phối dự đoán đầu ra, dẫn đến sự hoài nghi về độ ổn định và vấn đề đạo đức trong các mô hình NLP. Đạo đức trong NLP rất quan trọng vì các mô hình NLP có ảnh hưởng trực tiếp đến ngôn ngữ, nhận thức và văn hóa con người, từ đó dẫn đến nguy cơ duy trì các định kiến xã hội nếu bản thân các mô hình này tiềm ẩn định kiến [1, 2]. Do đó, nghiên cứu về tính khả diễn giải trong các mô hình NLP, gọi tắt là X-NLP (một nhánh của Trí tuệ Nhân tạo Khả diễn giải, gọi tắt là XAI), ngày càng thu hút sự quan tâm của xã hội [3]. Liên minh châu Âu thậm chí đã đưa yêu cầu về tính khả diễn giải vào các đạo luật của họ, như EU's General Data Protection Regulation (GDPR) để bảo vệ quyền lợi của người dùng. Khả năng diễn giải giúp đảm bảo các mô hình NLP ra quyết định không thiên vị, tuân thủ các chuẩn mực về đạo đức và tính công bằng trong việc ứng dụng AI [4]. X-NLP tập trung vào các phương pháp giúp con người giám sát và hiểu rõ hơn về việc đưa ra dự đoán của các mô hình NLP, phổ biến nhất là giải thích bằng cách làm rõ ảnh hưởng của các đặc trưng đầu vào lên dự đoán đầu ra [5].

Trong các tác vụ NLP, gán nhãn ngữ nghĩa (Semantic Role Labeling – SRL) có tầm ảnh hưởng lớn. SRL giúp máy tính hiểu tri thức quan trọng từ dữ liệu văn bản bằng cách rút trích các sự việc và đối tượng được nói đến trong câu. Đối tượng xử lý của SRL là cấu trúc đối số vị ngữ (Predicate Argument Structure – PAS), bao gồm động từ chính của câu và các đối tượng liên quan gọi là các đối số. SRL không chỉ nhận biết mà còn gán cho mỗi đối số một vai trò ngữ nghĩa cụ thể, qua đó trình bày rõ ràng sự việc trong câu. Năng lực này quan trọng đối với các tác vụ đọc hiểu văn bản như rút trích sự kiện và nhận dạng thực thể. Trong lĩnh vực Y Sinh, với kho tri thức chuyên ngành đồ sộ vượt quá khả năng khai thác thủ công, SRL giúp khai thác tri thức Y Sinh hiệu quả hơn, gián tiếp nâng cao chất lượng chăm sóc sức khỏe [6]. Tuy nhiên, phần lớn công trình X-NLP tập trung giải thích những kết quả dự đoán ở mức từ, mức câu, và mức văn bản [7, 8, 9], trong khi việc giải thích kết quả dự đoán ở mức chuỗi từ như tác vụ SRL vẫn là một vấn đề mở cần thêm nỗ lực nghiên cứu.

Vì vậy, chúng tôi nghiên cứu và đề xuất một giải pháp X-NLP cho tác vụ SRL, lấy bối cảnh thử nghiệm là văn

Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên TP. Hồ Chí Minh, 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

Email: tnhduc@fit.hcmus.edu.vn

Lịch sử

- Ngày nhận: 16-07-2024
- Ngày sửa đổi: 14-07-2025
- Ngày chấp nhận: 14-07-2025
- Ngày đăng: 12-09-2025

DOI :
https://doi.org/10.32508/stdjns.v9i3.1397

Check for updates

VNUHCM PRESS

Trích dẫn bài báo này: Hoài Đức T N. **Giải thích tầm quan trọng của đặc trưng từ vựng trong tác vụ gán nhãn ngữ nghĩa**. *Sci. Tech. Dev. J. - Nat. Sci.* 2025; 9(3):3402-3415.

bản trong lĩnh vực Y Sinh. Văn bản Y Sinh thường chứa thông tin quan trọng liên quan đến chăm sóc sức khỏe và trị bệnh, nơi độ chính xác và tính minh bạch là vô cùng cần thiết vì mọi quyết định đều gắn liền với trách nhiệm lớn về sức khỏe và tính mạng con người. Việc hiểu để tin tưởng vào quyết định của mô hình NLP là rất cần thiết. Sai sót hoặc hiểu lầm trong việc đọc hiểu văn bản Y Sinh có thể gây ra hậu quả nghiêm trọng[10]. Do đó, đây là một trong những lĩnh vực mà khả năng giải thích dự đoán của mô hình NLP là đặc biệt có ý nghĩa.

Đặc trưng được dùng để giải thích là đặc trưng từ vựng, một loại đặc trưng hiển thị tường minh trong văn bản mà bất kỳ ai cũng có thể nhìn thấy. Do đó, so với các đặc trưng ẩn hoặc đặc trưng chuyên sâu về ngôn ngữ học, đặc trưng từ vựng có thế mạnh trong việc đem lại lời giải thích dễ hiểu và thân thiện với người dùng.

Cấu trúc của bài báo này như sau: Phần 1 giới thiệu tầm quan trọng của việc giải thích các mô hình NLP học sâu, qua đó nêu bật ý nghĩa nghiên cứu của chúng tôi. Phần 2 khái quát một số kiến thức nền tảng về giải thích tầm quan trọng của đặc trưng trong NLP. Phần 3 sẽ trình bày tổng quan hiện trạng nghiên cứu trong lĩnh vực này, qua đó phân tích các hạn chế của những nghiên cứu hiện có để đề ra mục tiêu cho nghiên cứu của chúng tôi. Các phương pháp đề xuất của chúng tôi sẽ được mô tả chi tiết trong Phần 5. Kết quả thực nghiệm và các thảo luận liên quan sẽ được trình bày chi tiết ở Phần 6. Cuối cùng, Phần 7 sẽ đưa ra kết luận và hướng nghiên cứu trong tương lai.

TỔNG QUAN

Phần này trình bày tổng quan về SRL và X-NLP, gồm mô tả tác vụ SRL (Phần Tác vụ gán nhãn ngữ nghĩa), các loại giải thích (Phần Phân loại lời giải thích), chất lượng giải thích (Phần Chất lượng lời giải thích) và các phương pháp giải thích (Phần Các hướng tiếp cận X-NLP).

Tác vụ gán nhãn ngữ nghĩa

Mục tiêu của tác vụ gán nhãn ngữ nghĩa (Semantic Role Labeling - SRL) là trích xuất Cấu trúc Đối số Vị ngữ (Predicate Argument Structure - PAS) từ văn bản. Tác vụ SRL sẽ xác định các thành phần của PAS trong một câu, bao gồm động từ chính (vị ngữ) và các đối tượng liên quan của nó (đối số), mỗi đối số mang một vai trò ngữ nghĩa khác nhau. PAS cho các động từ tiếng Anh được ghi chép trong nhiều bộ ngữ liệu, như VerbNet[11], FrameNet[12], PropBank[13], chủ yếu dành cho lĩnh vực tổng quát hơn là các lĩnh vực chuyên môn. Ví dụ, trong câu "Dennis sold his apartment to John for 20,000 USD", PAS bao gồm "sold" là vị ngữ (động từ chính) và bốn đối số của nó là:

- "Dennis" – Vai trò ngữ nghĩa: Người bán
- "his apartment" – Vai trò ngữ nghĩa: Người bán
- "John" – Vai trò ngữ nghĩa: Người mua
- "20,000 USD" – Vai trò ngữ nghĩa: Giá trị giao dịch

Trong lĩnh vực Y Sinh, PAS chỉ bao gồm các động từ liên quan đến các sự kiện Y Sinh quan trọng, thay vì bao quát tất cả các động từ tiếng Anh. Các khung đối số trong Y Sinh khác biệt đáng kể so với ngôn ngữ tổng quát. Một số công trình, như GREC[14] và PASBio[15], đã soạn lại khung đối số chuyên biệt cho các động từ Y Sinh.

Phân loại lời giải thích

Giải thích trong NLP được phân loại theo hai khía cạnh chính: Phạm vi và thời điểm giải thích.

Phạm vi giải thích bao gồm:

- Giải thích cục bộ: Giải thích dự đoán cho từng dữ liệu đầu vào.
- Giải thích toàn cục: Cung cấp cái nhìn toàn diện về quá trình dự đoán của mô hình NLP.
- Giải thích theo lớp: Giải thích chung cho tất cả dự đoán của một lớp đầu ra.

Thời điểm giải thích bao gồm:

- Giải thích nội tại: Tạo ra lời giải thích song song với dự đoán của mô hình
- Giải thích hậu nghiệm: Tạo ra lời giải thích sau khi mô hình đã dự đoán, yêu cầu các bước xử lý bổ sung và một mô hình giải thích tách biệt.

Lựa chọn loại giải thích nào là phụ thuộc vào bối cảnh và chi tiết cụ thể của mô hình NLP. Giải thích nội tại quan trọng trong phát triển mô hình mới, và sẽ sớm trở thành một tiêu chí chính thức trong việc đánh giá mô hình. Ngược lại, giải thích hậu nghiệm cho phép hiểu các mô hình phức tạp đang có sẵn mà không cần can thiệp cấu trúc hay hiệu suất mô hình[16].

Chất lượng lời giải thích

Các giải pháp X-NLP hướng đến hai tiêu chí về chất lượng: Tính dễ hiểu và tính trung thực[5].

- Tính dễ hiểu đảm bảo các lời giải thích phải dễ hiểu và hữu ích cho con người, hỗ trợ họ đánh giá độ tin cậy và gỡ lỗi của mô hình. Sự tham gia của con người là rất quan trọng khi đánh giá tính dễ hiểu. Ví dụ, người dùng dựa trên các lời giải thích để chọn ra mô hình NLP phù hợp nhất, hoặc để dự đoán kết quả của mô hình trên dữ liệu đầu vào mới.
- Tính trung thực đảm bảo lời giải thích phản ánh chính xác cách hoạt động bên trong của mô hình. Nó phải tiết lộ được cả ưu điểm và nhược điểm trong logic

của mô hình. Đạt được điều này là một thách thức và đòi hỏi nhiều nỗ lực nghiên cứu [17].

Dung hòa giữa tính dễ hiểu và tính trung thực là không dễ dàng. Các giải thích tập trung vào con người có thể chỉ nhấn mạnh những khía cạnh mà họ hiểu và quan tâm, thay vì cung cấp cái nhìn toàn diện về tất cả các yếu tố ảnh hưởng đến đầu ra của mô hình. Điều này có thể dẫn đến các giải thích không đầy đủ, làm giảm tính trung thực.

Ở các mô hình giải thích hậu nghiệm, người ta cũng mong muốn tính độc lập mô hình (model agnosticism), mặc dù không bắt buộc. Các kỹ thuật giải thích độc lập mô hình có thể giải thích dự đoán của bất kỳ mô hình NLP nào mà không cần biết kiến trúc hoặc tham số của nó. Điều này tạo ra các giải thích chỉ dựa trên dữ liệu, bao gồm dữ liệu đầu vào, đầu ra hoặc dữ liệu ẩn (latent data), mang lại sự linh hoạt trong việc giải thích các mô hình NLP khác nhau. Nghiên cứu của chúng tôi cũng đi theo hướng giải thích hậu nghiệm để có thể giải thích các mô hình hộp đen mà không đòi hỏi quyền tiếp cận vào kiến trúc bên trong mô hình NLP như các phương pháp giải thích dựa trên đạo hàm (gradient) hoặc điểm attention.

Các hướng tiếp cận X-NLP

Tiêu chí phổ biến nhất để phân loại các hướng tiếp cận X-NLP là nhìn từ góc độ người dùng để phân loại dựa trên nội dung giải thích. Điều này là vì nội dung đóng vai trò quyết định đối với chất lượng lời giải thích. Theo thuật ngữ được giới thiệu bởi tác giả Danilevsky và cộng sự, X-NLP được chia thành bốn hướng tiếp cận dựa trên nội dung giải thích [3]:

• Giải thích bằng tầm quan trọng của đặc trưng: Loại giải thích này định lượng tầm quan trọng của các đặc trưng mà mô hình NLP sử dụng khi đưa ra dự đoán.

• Giải thích dựa trên ví dụ: Các giải thích này sử dụng các ví dụ cụ thể để làm rõ lý do đằng sau các dự đoán.

• Giải thích dựa trên nguồn gốc: Cách giải thích này cung cấp thông tin chi tiết về quá trình lập luận của mô hình NLP, thể hiện chuỗi logic liên quan đến các dự đoán của chúng.

• Giải thích dựa trên quy nạp khai báo: Các giải thích này đúc kết các hướng dẫn cơ bản, như bộ luật hoặc mã giả, mà mô hình NLP dựa vào khi đưa ra dự đoán.

Trong các hướng tiếp cận này, giải thích bằng tầm quan trọng của đặc trưng thu hút nhiều nỗ lực nghiên cứu nhất vì nó gắn liền chặt chẽ với cơ chế của học máy và rất thân thiện với người dùng, với các đặc trưng đầu vào dễ nhận biết như từ ngữ và cụm từ làm nguyên liệu giải thích [5]. Do đó, chúng tôi chọn giải thích bằng tầm quan trọng của đặc trưng làm hướng tiếp cận cho nghiên cứu của mình.

Ở một góc nhìn khác, các hướng tiếp cận X-NLP cũng chia ra hai phân nhánh riêng biệt:

• Giải thích nội tại (intrinsic explanation): Lời giải thích được tạo ra đồng thời với các dự đoán của mô hình NLP bằng cách sử dụng dữ liệu trung gian được tạo ra trong quá trình dự đoán. Với giải thích nội tại, mô hình NLP và mô hình giải thích thường cùng là một mô hình. Các mô hình truyền thống có tính trong suốt như cây quyết định hoặc mô hình dựa trên luật là thuộc nhóm này.

• Giải thích hậu học (post-hoc explanation): Lời giải thích được tạo ra sau khi mô hình đã đưa ra dự đoán. Vì vậy, mô hình NLP và mô hình giải thích hậu học thường là hai mô hình khác nhau.

Giải thích hậu học đòi hỏi các bước xử lý bổ sung sau khi mô hình NLP đã đưa ra dự đoán của mình. Mô hình giải thích hậu học và mô hình NLP thường là hai mô hình riêng biệt. Do đó, giải thích hậu học có thế mạnh trong việc giải thích các mô hình hộp đen đang tồn tại mà không đòi hỏi phải tiếp cận hoặc can thiệp vào kiến trúc mô hình. Điều này tạo nên tính độc lập mô hình (model agnostic), một điều không thể đạt được ở giải thích nội tại. Vì vậy, chúng tôi định hướng nghiên cứu của mình theo nhánh giải thích hậu học để để xuất phương pháp giải thích có nhiều khả năng mở rộng trong tương lai.

NHỮNG NGHIÊN CỨU LIÊN QUAN

Phép giải thích bằng tầm quan trọng đặc trưng, ký hiệu là I, sẽ đo lường tầm quan trọng của n đặc trưng trong không gian đặc trưng F của mẫu dữ liệu x, đối với dự đoán y của mô hình NLP M:

$$I(x, y, M): F \rightarrow I^n \quad (1)$$

Trong đó, $I \subset R$, và thường là [0,1] hoặc [-1, 1]. Phương pháp này chủ yếu dùng cho giải thích cục bộ, bao gồm cả giải thích nội tại và giải thích hậu nghiệm. Các nghiên cứu thường đo lường ảnh hưởng của một đặc trưng f đến các dự đoán của mô hình NLP bằng cách loại bỏ f khỏi đầu vào và ghi lại các thay đổi trong dự đoán:

$$_f J(x, y, M) = \triangle_{f \leftarrow null} p(y; \theta_M) \quad (2)$$

Trong khi hầu hết các nghiên cứu tập trung vào việc đánh giá tầm quan trọng của f khi kết hợp với các đặc trưng khác [18], [19], một số ý kiến cho rằng nên đánh giá ảnh hưởng độc lập của từng đặc trưng [20], [21]. Cũng có tranh luận về tầm quan trọng của tần suất đặc trưng: một số tác giả cho rằng các đặc trưng xuất hiện thường xuyên ít ảnh hưởng hơn [22], trong khi những tác giả khác tin rằng chúng có ảnh hưởng lớn đến dự đoán của mô hình [5].

Giải thích nội tại

Ước lượng tầm quan trọng đặc trưng bằng đạo hàm

Phương pháp này đo lường tầm quan trọng của một đặc trưng đầu vào bằng cách sử dụng đạo hàm. Cụ thể, một đặc trưng đầu vào sẽ có độ quan trọng cao khi dự đoán đầu ra rất nhạy cảm với những biến động ở đặc trưng ấy:

$$E_{gradiant}(x,c) = L_p(\nabla_x\, p(c\,;\theta)) \tag{3}$$

Nói cách khác, chỉ một biến động nhỏ ở đặc trưng ấy cũng có thể dẫn tới thay đổi đáng kể trong dự đoán đầu ra của mô hình NLP[23]. Trong biểu thức (2), $\nabla_x\, p(c\,;\theta)$ đại diện cho đạo hàm của xác suất mà mô hình NLP dự đoán cho lớp đầu ra c đối với đầu vào x, và θ là các tham số của mô hình. L_p là những phép chuẩn hóa vectơ dùng để tính toán độ lớn của các vectơ đạo hàm, bao gồm chuẩn L1 (Manhattan), L2 (Euclidean), và L_∞ (chuẩn Maximum hoặc Infinity).

Trong các mô hình tuyến tính, phương pháp sử dụng đạo hàm này hoạt động tốt và đưa ra các giải thích liên quan trực tiếp đến cách mô hình đưa ra quyết định. Tuy nhiên, trong các mô hình phức tạp hơn (phi tuyến tính), phương pháp này chỉ đưa ra giá trị xấp xỉ, và đôi khi không phản ánh chính xác tầm quan trọng của một từ trong câu[24].

Ước lượng tầm quan trọng đặc trưng bằng attention

Giải thích dựa trên attention là một kỹ thuật phổ biến khác dùng cho các giải thích cục bộ và nội tại. Phương pháp này định lượng tầm quan trọng của đặc trưng bằng cách sử dụng điểm attention. Tuy nhiên, các nghiên cứu đưa ra quan điểm mâu thuẫn về chất lượng giải thích của nó. Một số nghiên cứu cho rằng điểm attention không phải là lời giải thích thực sự[25], trong khi những nghiên cứu khác lại đề xuất rằng khi được áp dụng đúng cách, attention có thể mang lại những thông tin ý nghĩa[26].

Cũng có nhận định rằng điểm attention tự thân nó không đem lại các giải thích rõ ràng, nhưng attention saliency có thể cung cấp những thông tin ý nghĩa[27]. Không giống như điểm attention, attention saliency làm nổi bật các vùng đặc trưng giàu ý nghĩa tương ứng với các nhãn đầu ra khác nhau. Để cải thiện độ tin cậy trong giải thích, điểm attention có thể được kết hợp với kết quả giải thích lấy từ các phương pháp giải thích khác[28]. Qua đó, nhóm tác giả không chỉ nâng cao hiệu quả mô hình NLP mà còn đem lại lời giải thích từ điểm attention nhưng tương đồng với các kỹ thuật giải thích khác. Điều này cho thấy rằng các giải thích dựa trên attention có tiềm năng nhưng cần được nghiên cứu sâu hơn.

Điểm attention cũng được dùng để xác định từ nào trong câu là quan trọng đối với quá trình dự đoán[8,29,30]. Ngoài ra, token [CLS] trong các mô hình BERT cũng giúp đo lường tầm quan trọng của từ trong tác vụ phân loại văn bản[8]. Token [CLS] đại diện cho toàn bộ chuỗi đầu vào và rất quan trọng cho các dự đoán NLP. Công trình này tập trung vào vector ngữ cảnh của token [CLS] ở lớp cuối cùng trong BERT, lấy trung bình cộng các ma trận self-attention của tất cả các attention-head trong lớp này để tính toán tầm quan trọng của đặc trưng từ vựng trong cả câu. Đối với tác vụ phát sinh hình ảnh từ văn bản, cơ chế attention được liên kết giữa văn bản và hình ảnh, gọi là attention xuyên ngữ cảnh[30]. Attention xuyên ngữ cảnh được dùng để xác định từ nào trong văn bản đầu vào có ảnh hưởng lớn đến vùng nào trong hình ảnh được tạo ra.

Tuy nhiên, như đã trình bày, nghiên cứu của chúng tôi thuộc nhánh giải thích hậu học, vì vậy các hướng tiếp của nhánh giải thích nội tại không phải là trọng tâm nghiên cứu của chúng tôi. Các nghiên cứu liên quan dưới đây sẽ tập trung vào giải thích hậu học.

Làm nhiễu dữ liệu (Data perturbation)

Tầm quan trọng của đặc trưng thường được đánh giá bằng kỹ thuật làm nhiễu dữ liệu, nghĩa là cố tình thay đổi dữ liệu đầu vào để quan sát ảnh hưởng gây ra cho các dự đoán đầu ra, qua đó đo lường tầm quan trọng của vùng đặc trưng bị thay đổi[31,32]. Ví dụ, LIME (Local Interpretable Model-agnostic Explanations) tạo ra một tập dữ liệu đầu vào bị làm nhiễu $\widetilde{X}$ bằng cách ngẫu nhiên loại bỏ các đặc trưng khỏi đầu vào[18]. Sau đó, LIME gán trọng số $w_x(\widetilde{x})$ cho mỗi mẫu bị làm nhiễu $\widetilde{x} \in \widetilde{X}$, dựa trên thước đo độ tương đồng D (ví dụ, độ tương đồng cosine) với đầu vào gốc x chứa n đặc trưng:

$$w_x(\widetilde{x}) = \sqrt{exp\left(\frac{-D(x,\widetilde{x})}{(0.75*\sqrt{n})^2}\right)} \tag{4}$$

Việc gán trọng số này biến $\widetilde{X}$ i thành một phân phối gần như tuyến tính. Sau đó, LIME huấn luyện một mô hình hồi quy tuyến tính trên tập dữ liệu có trọng số này, và các hệ số β_i trong phương trình hồi quy (5) đại diện cho tầm quan trọng của mỗi đặc trưng f_i bị loại bỏ trong $\widetilde{x}_i$:

$$g(\widetilde{x}) = \beta_0 + \beta_1\widetilde{x}_1 + \beta_2\widetilde{x}_2 + \cdots + \beta_n\widetilde{x}_n \tag{5}$$

Để khớp $g(x)$ với mô hình NLP M đang được giải thích, LIME sử dụng hàm mất mát bình phương có

trọng số như trong phương trình (6). Cách tiếp cận này cho phép LIME giải thích các mô hình NLP phức tạp bằng một mô hình hồi quy tuyến tính dễ hiểu hơn làm mô hình thay thế.

$$L(M, g, w_x) = \sum_{\widetilde{x}, \widetilde{x'} \in \widetilde{X}} w_x(\widetilde{x})(M(\widetilde{x}) - g(\widetilde{x'}))^2 \quad (6)$$

SHAP, một phiên bản khác của LIME, sử dụng Lý thuyết Trò chơi để tính toán các tương tác đặc trưng bằng cách ước lượng giá trị Shapley của mỗi đặc trưng[19]. SHAP mang lại các giải thích phù hợp với cách suy nghĩ của con người hơn so với LIME. Trọng số trong phương trình (4) trở thành:

$$w_x(\widetilde{x}, \overrightarrow{cv}) = \frac{(n-1)}{\binom{n}{|\overrightarrow{cv}|} |\overrightarrow{cv}| (n - |\overrightarrow{cv}|)} \quad (7)$$

Trong đó, $\overrightarrow{cv}$ là vector one-hot chỉ ra các đặc trưng bị làm nhiễu trong $\widetilde{x}$, và $|\overrightarrow{cv}|$ là số lượng số 1 trong $\overrightarrow{cv}$. Việc gán trọng số này giúp ước lượng giá trị Shapley ϕ_i của đặc trưng thứ i khi huấn luyện mô hình hồi quy tuyến tính:

$$g(\widetilde{x}) = \phi_0 + \sum_{i=1}^{n} \phi_i \widetilde{x}_i \quad (8)$$

Làm nhiễu dữ liệu cũng được sử dụng để so sánh khả năng giải thích của các mô hình NLP bằng cách lần lượt ẩn đi các từ có độ quan trọng cao trong đầu vào, và mô hình NLP nào có nhiều thay đổi đầu ra hơn được coi là dễ giải thích hơn[33],[34].

Giải thích thông qua đặc trưng từ vựng

Từ vựng là loại đặc trưng dễ nhận biết nhất trong các tác vụ NLP, dễ hiểu với người dùng và vì vậy thu hút sự chú ý đáng kể trong cộng đồng nghiên cứu X-NLP. Tầm quan trọng của mỗi từ có thể được định lượng bằng huấn luyện có giám sát. Trong đó, một tập ngữ liệu có gán nhãn sẵn độ quan trọng của mỗi từ được dùng để huấn luyện việc sinh ra các lời giải thích mới[35]. Trong các lĩnh vực không có được dữ liệu huấn luyện như vậy, một giải pháp là dùng huấn luyện đối kháng không giám sát[36]. Cụ thể, các tác giả phát triển hai mô hình lựa chọn từ vựng, một mô hình P cố gắng chọn những từ có sức ảnh hưởng cao, và một mô hình đối kháng $\widetilde{P}$ chọn các từ ngẫu nhiên, và mô hình NLP chính M sẽ đưa ra các dự đoán riêng biệt từ mỗi bộ từ được chọn bởi P và $\widetilde{P}$. Mục tiêu là làm cho các dự đoán của M từ bộ từ của P có F1 cao hơn so với các dự đoán từ bộ từ của $\widetilde{P}$.

Nhận xét

Các kỹ thuật làm nhiễu dữ liệu hiện tại gặp nhiều hạn chế đáng kể. Cách làm nhiễu dữ liệu phổ biến nhất trong X-NLP là cố ý che giấu (masking) hoặc xóa bỏ (deleting) các từ trong câu nhằm đánh giá tác động của chúng đối với dự đoán của mô hình và ước tính tầm quan trọng của chúng[31],[32]. Tuy nhiên, các cách làm này có nguy cơ cao sẽ làm sai lệch lời giải thích. Việc che giấu các từ vựng trong các mô hình ngôn ngữ thường dẫn đến việc đánh giá thấp tầm quan trọng của đặc trưng, bởi vì các mô hình ngôn ngữ, vốn được huấn luyện cho các tác vụ như dự đoán từ tiếp theo hoặc điền vào chỗ trống, vẫn có thể dự đoán các từ bị che, dẫn đến ít thay đổi trong dự đoán đầu ra và do đó gây hiểu lầm rằng từ bị che ít quan trọng. Ngược lại, việc loại bỏ một khái niệm có thể làm gián đoạn cấu trúc câu và ngữ pháp, gây thêm nhiễu ngoài ý muốn. Những gián đoạn này có thể dẫn đến nhiều thay đổi hơn trong dự đoán, không chỉ đến từ sự vắng mặt của từ bị xóa mà còn do lỗi ngữ pháp và sự không hoàn chỉnh của câu, dẫn đến việc đánh giá quá cao tầm quan trọng của các từ bị xóa.

Ngoài ra, như đã đề cập, hiện trạng nghiên cứu về X-NLP tập trung giải thích những kết quả dự đoán ở mức từ, mức câu, và mức văn bản. Do đó, cần có sự nghiên cứu tỉ mỉ hơn về giải pháp tính toán tầm quan trọng của đặc trưng đối với những tác vụ dự đoán ở mức chuỗi từ như SRL.

Từ những nhận xét trên, bài báo này đề xuất một giải pháp làm nhiễu dữ liệu mới nhằm khắc phục hạn chế của che giấu hoặc xóa bỏ từ vựng. Bên cạnh đó, chúng tôi cũng phát triển một kỹ thuật tính toán tầm quan trọng đặc trưng dành riêng cho dự đoán NLP ở mức chuỗi từ.

PHƯƠNG PHÁP THỰC HIỆN

Phương pháp làm nhiễu dữ liệu

Để khắc phục việc đánh giá quá cao hoặc quá thấp tầm quan trọng của một đặc trưng, một vấn đề thường gặp phải trong làm nhiễu dữ liệu bằng xóa từ hoặc che từ (Phần Nhận xét), chúng tôi đề xuất phương pháp Smart Substitution. Phương pháp này không che từ hay xóa từ, mà sẽ thay từ cần đánh giá bằng một từ khác, lựa chọn từ thay thế một cách khôn ngoan. Các tiêu chí lựa chọn từ thay thế:

- TC1: Từ dùng để thay thế phải có cùng từ loại với từ gốc để hạn chế xáo trộn ngữ pháp trong câu. Do đó, tập ứng viên chính là tập hợp những thực từ có cùng từ loại với từ gốc cần được thay thế.

- Ý nghĩa: TC1 giúp khắc phục mặt hạn chế của kỹ thuật xóa bỏ (deleting). Việc loại bỏ từ cần đánh

giá bằng cách xóa bỏ sẽ gây phá vỡ cấu trúc câu, dẫn tới những biến động trong dự đoán đầu ra của mô hình không chỉ đến từ sự vắng mặt của từ bị xóa bỏ mà còn đến từ dữ liệu ngoài phân phối do ngữ pháp sai và cấu trúc câu không hoàn chỉnh. TC1 giúp đảm bảo ngữ pháp đúng đắn trong câu và nhờ đó giảm thiểu rủi ro này.

- TC2: Từ dùng để thay thế phải có ngữ nghĩa đủ khác biệt với từ gốc để đảm bảo loại bỏ tối đa đóng góp của từ gốc vào dự đoán của mô hình, tránh đánh giá quá thấp vai trò của nó.

- Ý nghĩa: TC2 giúp khắc phục mặt hạn chế của kỹ thuật che giấu (masking). Vì tiêu chí này yêu cầu thay từ cần đánh giá bằng từ có ngữ nghĩa khác biệt nên đã áp đặt được các mô hình ngôn ngữ và không cho phép chúng suy đoán ra từ gốc. Điều này giúp loại bỏ từ cần đánh giá ra khỏi câu một cách hiệu quả, từ đó đo lường được tầm quan trọng của từ một cách trung thực hơn.

Để đạt TC2, từ gốc và ứng viên thay thế được so sánh với nhau bằng độ tương đồng *cosine*. Tuy nhiên, độ tương đồng *cosine* nguyên thủy chỉ quan tâm chênh lệch góc xoay giữa hai vec-tơ biểu diễn từ, không quan tâm đến chênh lệch về độ lớn vec-tơ. Điều này có thể dẫn đến không đánh giá trọn vẹn được độ tương đồng giữa hai từ. Do đó, chúng tôi đề xuất một phiên bản mới của độ tương đồng *cosine*, gọi là $cosine_{module}$, có quan tâm cả khác biệt về độ lớn của vec-tơ.

$$cosine_{module}(v_1, v_2) = \left(1 - \frac{||v_1| - |v_2||}{|v_1| + |v_2|}\right) .cosine(v_1, v_2) \quad (9)$$

$$cosine(v_1, v_2) = \frac{v_1 . v_2}{|v_1| . |v_2|} \quad (10)$$

Liên quan đến việc lựa chọn ứng viên thay thế từ tập ứng viên C dựa trên độ tương đồng, chúng tôi đặt ra hai giả thuyết (GT) về giá trị độ tương đồng giữa ứng viên (ký hiệu là *cand*) và từ gốc (ký hiệu là *org*). Hai giả thuyết này sẽ được xác thực qua thực nghiệm:
- GT1: Từ dùng để thay thế phải trái nghĩa với từ gốc. Theo đó, độ tương đồng phải càng gần -1 càng tốt, ứng viên thay thế được chọn sẽ là $\underset{cand \in C}{argmin}\left(1 + cosine\left(v_{org}, v_{cand}\right)\right|$.
- GT2: Từ dùng để thay thế phải xa nghĩa với từ gốc. Theo đó, độ tương đồng phải càng gần 0 càng tốt, ứng viên thay thế được chọn sẽ là $\underset{cand \in C}{argmin}\left(cosine\left(v_{org}, v_{cand}\right)\right|$.
Ngoài ra, quá trình tách từ của NLP có thể tách một từ thành nhiều hơn một token, mỗi token sẽ có vec-tơ biểu diễn riêng, nên để tổng hợp được một vec-tơ biểu diễn duy nhất của từ, phục vụ cho TC2, chúng tôi cũng đặt ra hai phương án cần được đánh giá bằng thực nghiệm:
- PA1: Vec-tơ biểu diễn từ là tổng trên từng chiều của các vec-tơ token trong từ.
- PA2: Vec-tơ biểu diễn từ là trung bình cộng trên từng chiều của các vec-tơ token trong từ.
Như vậy, khi tìm từ thay thế cho một từ gốc, phương pháp Smart Substitution sẽ tổ hợp hai công thức tính độ tương đồng (*cosine* và $cosine_{module}$), hai giả thuyết về giá trị độ tương đồng (GT1 và GT2), và hai phương án về tổng hợp vec-tơ biểu diễn từ (PA1 và PA2), tất cả tạo nên 8 trường hợp riêng biệt, cho ra 8 từ thay thế cho một từ gốc. Trong 8 trường hợp này, trường hợp nào cho lời giải thích trung thực nhất là điều sẽ được kiểm tra qua thực nghiệm (Phần Kết quả và thảo luận).
Ngoài ra, trong ngôn ngữ luôn tồn tại một số lượng lớn các mạo từ, giới từ vốn chỉ đóng vai trò ngữ pháp (function word), không mang nhiều ngữ nghĩa. Để hạn chế việc bùng nổ dữ liệu cần xử lý một cách vô ích, chúng tôi chỉ đánh giá tầm quan trọng của những thực từ (content word) trong câu, bao gồm động từ chính, danh từ, tính từ và các trạng từ.

Định lượng tầm quan trọng của đặc trưng từ vựng cho dự đoán NLP mức chuỗi từ

Xét câu s, mỗi vị ngữ hoặc đối số mà mô hình SRL nhận biết trong s gọi là một dự đoán $s.p$. Điểm khác biệt của dự đoán mức chuỗi từ $s.p$ là nó bao gồm nhiều nhãn đầu ra cho nhiều từ t nằm trong chuỗi từ của dự đoán $s.p$. Một ví dụ phổ biến là các nhãn B, I, O trong các dự đoán mức chuỗi từ như thực thể của tác vụ nhận biết thực thể có tên (NER), đối số của tác vụ SRL… Vì vậy, tầm quan trọng của một từ w đối với $s.p$ được tổng hợp từ tầm quan trọng của w đối với từng từ t trong nằm trong $s.p$. Mỗi từ này được mô hình SRL dự đoán một nhãn đầu ra $t.y$ với xác suất sự đoán $P(t.y)$. Tầm quan trọng này bao gồm hai khía cạnh: Sức ảnh hưởng (Impact – ký hiệu: Imp) và độ hữu ích (Usefulness – ký hiệu: Usf).
Sức ảnh hưởng của từ w lên dự đoán p, ký hiệu là $w.Imp(p)$ đo lường tỷ lệ đóng góp của w vào dự đoán p. Nếu $w.Imp(p)$ là một đại lượng dương thì sự hiện diện của từ w ủng hộ cho dự đoán p. Ngược lại, nếu $w.Imp(p)$ âm thì w chống lại dự đoán p, nghĩa là w khi ấy có xu hướng kéo mô hình sang một kết quả dự đoán nào đó khác chứ không phải p. Trong quá trình tổng hợp $w.Imp(p)$ từ sức ảnh hưởng của w lên mỗi từ trong p, các từ này sẽ có trọng số khác nhau. Cụ thể, từ nào không bị w làm cho thay đổi nhãn phân loại đầu ra thì chỉ nhận trọng số bằng 1 vì những từ

ấy thuộc vùng ảnh hưởng yếu của w. Ngược lại, từ nào bị w làm cho thay đổi nhãn phân loại đầu ra, chứng tỏ thuộc vùng ảnh hưởng trọng yếu của w, sẽ có trọng số bằng 2. Sức ảnh hưởng của w lên mỗi từ t được ước lượng dựa vào độ chênh lệch xác suất của nhãn đầu ra dành cho t giữa lúc có mặt w và lúc vắng mặt w.

Gọi $\widetilde{s}$ là phiên bản của s sau khi thay thế w bằng $\widetilde{w}$ lấy từ kỹ thuật SmartSub. Khi đó, $w.Imp(p)$ sẽ được định lượng bằng giải thuật 4.1 sau đây.

Giải thuật 4.1: Định lượng sức ảnh hưởng của đặc trưng từ vựng w lên dự đoán p trong câu s

function $Imp(s, w, s.p)$

$\widetilde{s} = SmartSub(s, w)$; // tạo phiên bản nhiễu $\widetilde{s}$ của s để so sánh hai kết quả dự đoán đầu ra của s và $\widetilde{s}$

$M(\widetilde{s})$; // và cho mô hình SRL thực hiện việc dự đoán trên $\widetilde{s}$

$Imp = 0; sumWeight = 0$; //Khởi gán sức ảnh hưởng của w và tổng trọng số của các từ trong p = 0

For $t \in s.p \cup \widetilde{s}.p$; // Tính toán Imp chỉ quan tâm những từ nằm trong dự đoán gốc và dự đoán nhiễu

$l = s.t.y$; $\widetilde{l} = \widetilde{s}.t.y$; // Ở s và $\widetilde{s}$, cùng một từ có thể nhận nhãn dự đoán khác nhau

If $l = \widetilde{l}$ // Nếu việc làm nhiễu không thay đổi nhãn của từ t (trước sau t vẫn nhận nhãn l)

//Tầm ảnh hưởng của w lên t là tỷ lệ biến thiên xác suất nhãn l giữa câu gốc và câu nhiễu

$Imp = Imp + (P(s.t.l) - P(\widetilde{s}.t.l))$;

$sumWeight += 1$; // khi nhãn của t không đổi thì trọng số của t chỉ là 1

Else // Nếu việc làm nhiễu có gây thay đổi nhãn của từ t từ l thành $\widetilde{l}$

//Tầm ảnh hưởng của w lên t là tổng tỷ lệ biến thiên xác suất của cả hai nhãn l và $\widetilde{l}$

$Imp = Imp + (P(s.t.l) - P(\widetilde{s}.t.l))$;

$Imp = Imp + (P\left(\widetilde{s}.t.\widetilde{l}\right) - P\left(s.t.\widetilde{l}\right))$;

$sumWeight += 2$; // khi nhãn của t thay đổi thì trọng số của t là 2

End if

End For

// Tầm ảnh hưởng của w lên $s.p$ là trung bình cộng có trọng số các ảnh hưởng của w lên từng trong p

$Imp = Imp/sumWeight$;

return Imp;

end function

Độ hữu ích của từ w lên dự đoán p, ký hiệu là $w.Usf(p)$ đo lường tỷ lệ đóng góp của w vào việc làm cho dự đoán p tiến gần hơn với dự đoán chuẩn vàng p_{gold}. Nếu $w.Usf(p)$ là một đại lượng âm thì w là một từ có hại cho dự đoán của mô hình, và ngược lại. Tương tự như với $w.Imp(p)$, việc định lượng $w.Usf(p)$ cũng dành cho mỗi từ trong dự đoán của mô hình và dự đoán chuẩn vàng những trọng số khác nhau. Những từ mà tính chính xác trong dự đoán về nhãn của chúng bị đảo ngược bởi w (đúng thành sai, hoặc sai thành đúng) thể hiện rõ rệt sự hữu ích hay có hại của w nên có trọng số bằng 2. Những từ còn lại không bị w làm thay đổi tính đúng sai trong dự đoán sẽ có trọng số bằng 1. Độ hữu ích của w đối với một từ t được tính dựa trên độ biến thiên tăng của xác suất nhãn đúng và sự biến thiên giảm của xác suất nhãn sai dành cho t. Cụ thể, $w.Usf(p)$ sẽ được định lượng bằng giải thuật 4.2 sau đây.

Giải thuật 4.2: Định lượng độ hữu ích của đặc trưng từ vựng w lên dự đoán p trong câu s

function $Usf(s, w, s.p)$

$\widetilde{s} = SmartSub(s, w)$; // tạo phiên bản nhiễu $\widetilde{s}$ của s bằng SmartSub để so sánh kết quả dự đoán

$M(\widetilde{s})$; // và cho mô hình SRL thực hiện việc dự đoán trên $\widetilde{s}$

$Usf = 0; sumWeight = 0$; //Khởi gán độ hữu ích của từ w và tổng trọng số của các từ trong p = 0

For $t \in s.p \cup \widetilde{s}.p \cup p_{gold}$; // Chỉ quan tâm các từ trong dự đoán gốc, dự đoán nhiễu và dự đoán đúng

$l = s.t.y$; $\widetilde{l} = \widetilde{s}.t.y$; $l_{gold} = t.y_{gold}$;

If $l = \widetilde{l} = l_{gold}$ // Nếu khi có hay không có w thì nhãn đoán cho t vẫn đúng

//Độ hữu ích của w lên t là độ tăng xác suất nhãn l khi có sự hiện diện của w

$Usf = Usf + (P(s.t.l) - P(\widetilde{s}.t.l))$;

$sumWeight += 1$; // khi độ chính xác cho nhãn của t không đổi thì trọng số của t chỉ là 1

Else if $l \neq l_{gold} \wedge \widetilde{l} \neq l_{gold}$ // Nếu khi có hay không có w thì nhãn đoán cho t vẫn sai

// Độ hữu ích của w lên t là trung bình cộng mức độ mà w làm giảm xác suất hai nhãn sai l và $\widetilde{l}$

// và mức độ mà w làm tăng xác suất nhãn đúng l_{gold}

$Usf = Usf + (P(\widetilde{s}.t.l) - P(s.t.l))$;

$Usf = Usf + (P\left(\widetilde{s}.t.\widetilde{l}\right) - P\left(s.t.\widetilde{l}\right))$;

$Usf = Usf + (P\left(s.t.l_{gold}\right) - P\left(\widetilde{s}.t.l_{gold}\right))$;

$Usf = Usf/3$;

$sumWeight += 1$; // khi độ chính xác cho nhãn của t không đổi thì trọng số của t chỉ là 1

Else if $l \neq l_{gold} \wedge \widetilde{l} = l_{gold}$ // Nếu khi không có w thì nhãn đoán đúng, có w thì đoán sai

// Độ hữu ích của w lên t là trung bình cộng mức độ mà w làm giảm xác suất nhãn sai l

// và mức độ mà w làm tăng xác suất nhãn đúng $\widetilde{l}$ (cả hai tỷ lệ này đều sẽ âm vì khi này w có hại)

$Usf = Usf + (P(\widetilde{s}.t.l) - P(s.t.l))$;

$Usf = Usf + (P\left(s.t.\widetilde{l}\right) - P\left(\widetilde{s}.t.\widetilde{l}\right))$;

$Usf = Usf/2$;

$sumWeight+ = 2$; // khi độ chính xác cho nhãn của t có thay đổi thì trọng số của t sẽ là 2

$Else\ if\ l = l_{gold} \wedge \widetilde{l} \neq l_{gold}$ // Nếu khi không có w thì nhãn đoán sai, có w thì đoán đúng

// Độ hữu ích của w lên t là trung bình cộng tỷ lệ mà w làm giảm xác suất nhãn sai $\widetilde{l}$

// và tỷ lệ mà w làm tăng xác suất nhãn đúng l (cả hai tỷ lệ này đều sẽ dương vì khi này w có ích)

$Usf = Usf + (P(s.t.l) - P(\widetilde{s}.t.l))$;

$Usf = Usf + (P(\widetilde{s}.t.\widetilde{l}) - P(s.t.\widetilde{l}))$;

$Usf = Usf/2$;

$sumWeight+ = 2$; // khi độ chính xác cho nhãn của t có thay đổi thì trọng số của t sẽ là 2

End if

End For

// Độ hữu ích của w lên $s.p$ là trung bình cộng có trọng số các độ hữu ích của w lên từng từ trong p

$Usf = Usf/sumWeight$;

return Usf;

end function

Trong hầu hết các nghiên cứu về giải thích bằng tầm quan trọng đặc trưng, các phương pháp thường ước lượng tầm quan trọng của đặc trưng dựa trên biến động trong dự đoán cuối cùng của mô hình. Điều này chỉ nhận biết được những ảnh hưởng đủ sức làm thay đổi dự đoán cuối cùng, không thể nhận biết những ảnh hưởng tinh vi lên từng phân lớp đầu ra. Giải thuật 4.1 và 4.2 giúp khắc phục điều này, quan sát từng thay đổi nhỏ trong phân bố xác suất, không bỏ lỡ vai trò của bất kỳ từ nào dù là mạnh hay yếu.

Nhận xét

Kỹ thuật tính toán sức ảnh hưởng và độ hữu ích của đặc trưng mà chúng tôi đề xuất tiết kiệm chi phí tính toán khá đáng kể so với LIME và đặc biệt là với SHAP. Cả LIME và SHAP đều đòi hỏi huấn luyện lại mô hình hồi quy tuyến tính cho mỗi điểm dữ liệu.

Ngoài ra, giải thuật 4.1 và 4.2 chỉ có độ phức tạp O(n) với n là số từ trong $s.p \cup \widetilde{s}.p$ (với giải thuật 4.1) và trong $s.p \cup \widetilde{s}.p \cup p_{gold}$ (với giải thuật 4.2), và n luôn nhỏ hơn tổng số từ trong câu. Trong khi đó, độ phức tạp của LIME, do liên quan đến việc tính trọng số cho mỗi mẫu nhiễu, có độ phức tạp O(n d) với n là toàn bộ số lượng từ trong câu và d là số chiều của vec-tơ biểu diễn. SHAP, có độ phức tạp tính toán rất cao vì nó tính toán các giá trị Shapley, vốn dựa trên lý thuyết trò chơi. Độ phức tạp của SHAP có thể đạt mức $O(2^n)$ trong trường hợp tính toán chính xác hoặc O(n log n) với các phương pháp xấp xỉ bằng kỹ thuật Kernel.

Không chỉ vậy, LIME sử dụng kỹ thuật làm nhiễu xóa bỏ và SHAP sử dụng kỹ thuật làm nhiễu che giấu nên đều chịu ảnh hưởng từ hạn chế của hai kỹ thuật làm nhiễu này khi áp dụng vào giải thích dự đoán NLP, nhất là dự đoán của những mô hình dựa trên mô hình ngôn ngữ tiền huấn luyện.

KẾT QUẢ VÀ THẢO LUẬN

Dữ liệu và kịch bản thử nghiệm

Dữ liệu thử nghiệm

Về dữ liệu, chúng tôi sử dụng bộ ngữ liệu PASBio+[37], một nguồn dữ liệu PAS cho lĩnh vực Y Sinh được gán nhãn chi tiết để làm bối cảnh cho kỹ thuật giải thích của mình. PASBio+ là một nguồn dữ liệu phù hợp cho nghiên cứu của chúng tôi vì:

- PASBio+ cung cấp bối cảnh văn bản Y Sinh.
- Kích thước của PASBio+ là trên 7000 câu, đủ để huấn luyện mô hình SRL. Các bộ dữ liệu khác như BioVerbNet[38] hoặc GREC[14] chỉ có dưới 1500 câu, đem đến rủi ro vấn đề dữ liệu thưa, gây nhiễu cho quá trình ước lượng tầm quan trọng của đặc trưng.
- Bộ đối số của PASBio+ là kế thừa từ PASBio[15], được các chuyên gia viết thủ công cho từng động từ nên có độ chuyên biệt về Y Sinh cao hơn rất nhiều so với các bộ dữ liệu khác cũng sử dụng văn bản Y Sinh nhưng lại vay mượn khung đối số của lĩnh vực tổng quát như BioProp[6], BioVerbNet[38]. Điều ấy là quan trọng vì các đối số này chính là đối tượng được giải thích trong nghiên cứu của chúng tôi.

Vì những lý do trên, PASBio+ là bộ dữ liệu phù hợp nhất mà chúng tôi có thể tiếp cận được cho nghiên cứu của mình.

Kịch bản thử nghiệm

Thách thức về dữ liệu chuẩn vàng (benchmark)

X-NLP đối mặt với những thách thức phức tạp hơn nhiều so với NLP thông thường. Trong NLP, mỗi tác vụ với một bộ ngữ liệu đều có sẵn một bộ giải pháp chuẩn vàng thống nhất để đánh giá hiệu quả của tất cả các mô hình. Tuy nhiên, X-NLP không có được sự đơn giản này. Để đánh giá các lời giải thích do X-NLP tạo ra, nhiều nghiên cứu đã cố gắng xây dựng thủ công các bộ dữ liệu giải thích chuẩn vàng chung[39],[40],[41]. Dù vậy, các bộ dữ liệu này chỉ có thể đo lường mức độ dễ hiểu của lời giải thích bằng cách so sánh nó với cách suy nghĩ chủ quan của con người.

Một thách thức lớn hơn đối với X-NLP là việc đánh giá tính trung thực của lời giải thích. Đánh giá tính trung thực yêu cầu lời giải thích phải phản ánh chính xác hành vi bên trong của mô hình NLP, tức là giải thích được chính xác cách thức mà mô hình đưa ra kết quả. Điều này có nghĩa là mỗi mô hình NLP đòi hỏi một chuẩn vàng riêng, dựa trên tư duy nội tại của chính mô hình đó, do đó không thể tạo ra một bộ dữ liệu giải thích chuẩn vàng chung cho mọi mô hình. Việc xây dựng thủ công các bộ dữ liệu riêng để đánh

giá tính trung thực cho từng mô hình NLP là không khả thi trong thực tế.

Giải pháp đánh giá không cần dữ liệu chuẩn vàng

Do đó, chúng tôi đề xuất một giải pháp có thể đánh giá gián tiếp tính trung thực của một kỹ thuật giải thích ε bất kỳ khi giải thích một mô hình NLP M bất kỳ mà không cần thông qua dữ liệu giải thích chuẩn vàng:

- Trước tiên, ta dùng kỹ thuật giải thích ε để ước lượng được mức độ thông minh của mô hình NLP. Mô hình càng thông minh khi nó càng để cho những đặc trưng hữu ích (giá trị Usf cao) phát huy sức ảnh hưởng cao (giá trị Imp cao) trong khi kìm hãm xuống tối thiểu sức ảnh hưởng của những đặc trưng có hại, và ngược lại. Cụ thể, độ thông minh của mô hình M khi thực hiện dự đoán trên tập hợp câu S được ước lượng thông qua kỹ thuật giải thích ε được tính như sau:

$$Smart_{\varepsilon}(M,S) = \underset{s\in S; w\in s.W; p\in s.P}{Spearman} (M.Imp(s,w,p); M.Usf(s,w,p)) \quad (11)$$

Với $s.W$ là tập hợp các từ trong câu s, và $s.P$ là tập hợp các dự đoán trong câu s.

- Sau đó, ta so sánh độ thông minh ước lượng này với hiệu quả dự đoán thực tế của mô hình NLP thể hiện qua các độ đo kinh điển (như độ chính xác, độ bao phủ, điểm F1, điểm Brier…). Độ thông minh càng tương quan mạnh với hiệu quả dự đoán thực tế cho thấy phương pháp giải thích ε càng phản ánh trung thực năng lực của mô hình NLP. Sự tương quan này, ký hiệu là $Honesty$, cũng được tính bằng hệ số tương quan Spearman giữa độ thông minh (được ước lượng từ kết quả giải thích) và hiệu quả dự đoán (được đo từ thực nghiệm) khi xét trên một chuỗi n các tập câu S_i khác nhau trong toàn bộ ngữ liệu. Do sức ảnh hưởng và độ hữu ích được tính dựa trên phân bố xác suất nhãn dự đoán đầu ra nên chúng tôi chọn đại lượng đo hiệu quả dự đoán là (1 – điểm Brier).

$$\varepsilon.Honesty(M) = \underset{i=1..n}{Spearman}(Smart_{\varepsilon}(M, S_i); (1 - M.BrierScore(S_i))) \quad (12)$$

Giá trị $\varepsilon.Honesty$ càng cao cho thấy kỹ thuật giải thích ε càng trung thực và ngược lại.

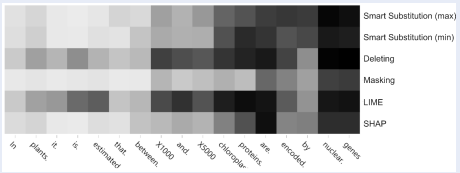
Kỹ thuật giải thích đối chứng (Baseline)

Chúng tôi so sánh kết quả giải thích của mình với các kỹ thuật đối chứng sau:

- Kỹ thuật tạo dữ liệu nhiễu bằng xóa bỏ (deleting) và che giấu (masking) kết hợp với giải thuật 4.1 và 4.2 để ước lượng tầm quan trọng đặc trưng (Bảng 2).
- Hai phương pháp kinh điển tạo dữ liệu nhiễu bằng xóa bỏ (LIME) và che giấu (SHAP) (Bảng 3).

Kết quả thực nghiệm và thảo luận

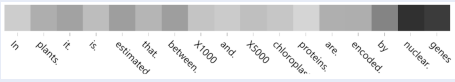
Dưới đây là một ví dụ cụ thể để so sánh cách các phương pháp khác nhau ước lượng sức ảnh hưởng của đặc trưng. Xét câu: *"In plants it is estimated that between 1000 and 5000 chloroplast proteins are **encoded** by nuclear genes"*. Trong đó, *"encoded"* là vị ngữ, *"nuclear genes"* là đối số A0 với vai trò ngữ nghĩa là *gene or RNA*, và *"between 1000 and 5000 chloroplast proteins"* là đối số A1 với vai trò ngữ nghĩa là *gene product*. Khi cho các phương pháp khác nhau ước lượng sức ảnh hưởng của từng từ khi mô hình SRL dự đoán đối số thứ nhất *"nuclear genes"*, kết quả được trực quan hóa bằng bản đồ nhiệt trong Hình 1 (màu đậm thể hiện sức ảnh hưởng cao).



Hình 1: Một ví dụ minh họa kết quả ước lượng sức ảnh hưởng của từ bởi các phương pháp khác nhau.

Có thể thấy cả sáu phương pháp đều nhận biết sức ảnh hưởng cao nhất của chính bản thân đối số *"nuclear genes"*, đối tượng trung tâm của việc giải thích. Tuy nhiên, Smart Substitution còn nhận biết được tầm quan trọng của hai thành phần còn lại trong PAS là vị ngữ *"encoded"* và đối số A1. LIME và phương pháp xóa bỏ (deleting) cũng nhận biết được điều này. Dù vậy, hai phương pháp này lại không làm được điều mà Smart Substitution làm được: (i) Phân biệt rõ độ quan trọng của A0 cao hơn hẳn A1 do A0 mới chính là dự đoán đang được giải thích; (ii) Phân biệt độ quan trọng của đầu tố cao hơn thành phần bổ nghĩa trong A1. Ngoài ra, LIME và phương pháp xóa bỏ, với cùng một hạn chế về cách tạo dữ liệu nhiễu, đã đánh giá quá cao sức ảnh hưởng của các thành phần nòng cốt trong ngữ pháp câu (như các danh từ và đại từ: *"it"*, *"plants"*; hoặc động từ: *"is"*, *"estimated"*) dù cho chúng không liên quan mật thiết với PAS trong câu này. Phương pháp che giấu (masking) cho kết quả nhạt nhất trên bản đồ nhiệt, điều này là vì những vị trí che giấu của phương pháp này đã bị đoán biết bởi mô hình SRL do mô hình này có cốt lõi là BioBERT, vốn là một mô

hình ngôn ngữ có khả năng điền vào chỗ trống. Một điểm nổi bật nữa của Smart Substitution trong ví dụ này là sự nhận biết tầm quan trọng của *"are"* và *"by"*, hai từ này giữ vai trò then chốt trong việc phân loại *"nuclear genes"* là A0. Trong khi đó, bốn phương pháp còn lại đều không nhận biết *"by"* là quan trọng.



Hình 2: Điểm attention của đối số A0 dành cho các từ còn lại trong câu.

Ngoài ra, khi chúng tôi lấy trung bình cộng điểm attention của hai từ trong dự đoán *"nuclear genes"* dành cho các từ còn lại trong câu, kết quả trực quan hóa trên bản đồ nhiệt không đem lại lời giải thích có ý nghĩa nào (Hình 2). Ngoại trừ việc hai từ này chú ý mạnh nhất vào chính mình, sự chú ý của chúng vào các từ còn lại trong câu rất mờ nhạt và hầu như ngẫu nhiên. Điều này tương đồng với nhận định của nhiều nghiên cứu cho rằng điểm attention không đem lại lời giải thích mà con người có thể hiểu được[25,42]. Một ưu điểm khác của phương pháp chúng tôi đề xuất so với các phương pháp dựa trên attention là có thể ước lượng cả sức ảnh hưởng và độ hữu ích của mỗi từ trong câu, trong khi cơ chế attention, dù cho có thực sự là một cách tiếp cận giải thích hiệu quả, cũng chỉ ước lượng được sức ảnh hưởng.

Để đánh giá tính trung thực của tám phương pháp làm nhiễu dữ liệu đã được đề xuất (Phần 4.1), chúng tôi so sánh các giá trị *Honesty* tương ứng. Kết quả thực nghiệm được trình bày trong Bảng 1.

Bảng 1: Giá trị *Honesty* của tám phương pháp Smart Sunstitution

	PA1		PA2	
	cosine	$cosine_{module}$	*cosine*	$cosine_{module}$
GT1	0.715	0.724	0.727	0.738
GT2	0.778	0.782	0.776	0.780

Kết quả thực nghiệm trên đây cho thấy khi so sánh phương pháp chọn từ thay thế xa nghĩa (GT2) với phương pháp chọn từ thay thế trái nghĩa (GT1), dù cho vector biểu diễn được lấy tổng hay lấy giá trị trung bình, và bất kể sử dụng độ tương đồng *cosine* hay $cosine_{module}$, thì đều cho giá trị *Honesty* cao hơn 5-6% một cách ổn định. Điều này cho thấy GT2 là một giả thuyết phù hợp với thực nghiệm hơn GT1, qua đó có thể kết luận rằng ứng viên thay thế xa nghĩa sẽ cho hiệu quả giải thích trung thực hơn ứng viên thay thế trái nghĩa. Điều đó thể hiện ở chỗ lời giải thích này cho phép ước lượng năng lực của mô hình có tương quan mạnh hơn với hiệu quả dự đoán thực tế. Kết quả này cũng phù hợp với tư duy của con người, vì với một từ trái nghĩa, ta có thể dễ dàng đoán biết từ gốc ban đầu, nhưng với một từ xa nghĩa, việc suy đoán từ gốc ban đầu sẽ khó khăn hơn rất nhiều.

Về phương diện đo lường độ tương đồng giữa từ gốc với từ ứng viên, khi so sánh việc sử dụng độ tương đồng *cosine* với $cosine_{module}$, ta nhận thấy $cosine_{module}$ cho lời giải thích có giá trị *Honesty* cao hơn *cosine* trong mọi phương án và giả thuyết. Điều này cho thấy $cosine_{module}$ đem lại hiệu quả đo lường độ tương đồng giữa hai vec-tơ biểu diễn tốt hơn nhờ có xem xét đến độ lớn vec-tơ, thay vì chỉ quan tâm góc lệch vec-tơ như độ tương đồng *cosine*. Tuy nhiên, có thể thấy rằng mức cải thiện này chỉ tầm trên dưới 1%, không cải thiện rõ rệt bằng việc sử dụng ứng viên xa nghĩa thay cho ứng viên trái nghĩa.

Ngược lại với các so sánh trên, việc so sánh giữa PA1 và PA2 trong việc tổng hợp các vec-tơ biểu diễn token thành vec-tơ biểu diễn từ không cho ta thấy một xu hướng rõ rệt nào. Qua đó, ta có thể nhận xét rằng cả hai phương án đều cho hiệu quả tương đương nhau trong việc giải thích tầm quan trọng của một từ. Tuy nhiên, về phương diện chi phí tính toán thì tính tổng (PA1) hiệu quả hơn tính trung bình cộng (PA2). Dù cho khác biệt này là nhỏ, nhưng khi đối diện với khối lượng ngữ liệu lớn, với hàng ngàn câu và mỗi câu có vài chục từ, thì đây cũng là một yếu tố đáng xem xét.

Ngoài ra, chúng tôi cũng so sánh giá trị *Honesty* đạt được trong kỹ thuật làm nhiễu dữ liệu của chúng tôi với giá trị đạt được trong các kỹ thuật làm nhiễu dữ liệu truyền thống là che giấu (masking) và xóa bỏ (deleting), cũng như với LIME và SHAP. Kết quả thực nghiệm được trình bày trong Bảng 2 và Bảng 3.

Bảng 2: Giá trị Honesty của phương pháp truyền thống (deleting, masking) so với phương pháp Smart Sunstitution

Deleting	Masking	Smart Substitution (max)	Smart Substitution (min)
0.583	0.576	0.782	0.715

Số liệu trong Bảng 2 và Bảng 3 cho thấy tất cả phương án của Smart Substitution, dù là phương án tốt nhất (*Honesty* = 0.782) hay kém nhất (*Honesty* = 0.715), đều cho kết quả vượt trội hơn hẳn so với các phương pháp làm nhiễu dữ liệu truyền thống.

LIME cho *Honesty* thấp nhất. Điều này phù hợp với nhận định của một khảo sát về giải thích hậu nghiệm

Bảng 3: Giá trị Honesty của phương pháp LIME và SHAP so với phương pháp Smart Sunstitutio

LIME	SHAP	Smart Substitution (max)	Smart Substitution (min)
0.574	0.686	0.782	0.715

rằng LIME không có thế mạnh trong việc giải thích các mô hình NLP[5]. Nguyên nhân nằm ở cách LIME xử lý câu văn như một túi từ (bag of words) nên đã bỏ qua trật tự từ, một yếu tố rất quan trọng trong xử lý ngôn ngữ tự nhiên. SHAP dựa trên nền tảng là LIME nên cũng không tránh khỏi hạn chế này. Tuy nhiên, việc xấp xỉ giá trị Shapley giúp SHAP ước lượng toàn diện tầm quan trọng của mỗi từ khi có quan tâm đến tương tác của nó với các từ còn lại. Do đó, *Honesty* của SHAP cao hơn đáng kể so với LIME, Deleting và Masking.

Về mặt nguyên lý, phương pháp Smart Substitution mà chúng tôi đề xuất dựa trên vec-tơ biểu diễn từ. Điều này khắc phục được hạn chế của túi từ do có quan tâm đến trật tự từ trong câu (cùng một từ nhưng đứng ở những vị trí khác nhau trong câu sẽ có vec-tơ biểu diễn khác nhau). Cũng như SHAP, phương pháp của chúng tôi có quan tâm đến sự tương tác giữa các từ trong câu vì vec-tơ biểu diễn từ được mã hóa bởi BioBERT, vốn là một mô hình transformer nên mỗi vec-tơ biểu diễn từ đều được tính toán dựa trên tương quan với các từ còn lại.

Về mặt thực nghiệm, phương pháp Smart Substitution cải thiện đáng kể hiệu quả làm nhiễu dữ liệu, đem lại lời giải thích có độ trung thực cao hơn khoảng 20% so với các phương pháp làm nhiễu dữ liệu hiện đang được sử dụng rộng rãi.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong bài báo này, chúng tôi giới thiệu một góc nhìn mới về tầm quan trọng của đặc trưng từ vựng trong việc giải thích các mô hình hộp đen, qua đó cho thấy tầm quan trọng của mỗi từ trong câu đối với dự đoán đầu ra của mô hình NLP không chỉ thể hiện ở tầm ảnh hưởng mà còn thể hiện ở độ hữu ích của nó trong việc đem lại một dự đoán chính xác. Chúng tôi cũng phát triển những giải thuật nhằm tính toán các đại lượng này, đặc biệt hướng đến các dự đoán NLP mức chuỗi từ vốn chưa được quan tâm đúng mức trong các nghiên cứu về X-NLP hiện có. Các giải thuật này, nhờ vào việc sử dụng vec-tơ phân bố xác suất trên các nhãn đầu ra thay vì dự đoán cuối cùng của mô hình, có thể nhận biết những ảnh hưởng dù là nhỏ nhất của từng đặc trưng từ vựng lên việc ra quyết định của mô hình NLP, giúp ích thiết thực cho hoạt động gỡ lỗi và tinh chỉnh mô hình.

Chúng tôi cũng đề xuất một phương pháp mới trong việc làm nhiễu dữ liệu, là công đoạn rất quan trọng của hướng tiếp cận giải thích hậu nghiệm dựa trên tầm quan trọng của đặc trưng. Phương pháp làm nhiễu dữ liệu của chúng tôi giúp khắc phục các hạn chế về tính trung thực của lời giải thích mà các phương pháp làm nhiễu dữ liệu phổ biến hiện nay đang gặp phải.

Kết quả thực nghiệm cho tác vụ SRL trên văn bản Y Sinh cho thấy các phương pháp được đề xuất trong bài báo này cho ra lời giải thích phù hợp với hiệu quả dự đoán thực tế của mô hình và đạt được độ trung thực vượt trội hơn các phương pháp làm nhiễu dữ liệu hiện có.

Hiện tại, phương pháp làm nhiễu dữ liệu của chúng tôi chỉ xem xét từ loại và ngữ nghĩa của từ khi chọn ứng viên thay thế. Hướng phát triển tương lai sẽ quan tâm thêm các khía cạnh chuyên sâu hơn về ngữ pháp như sự tương hợp giữa chủ từ và vị từ, thì (tense) và cách (voice) của động từ… Ngoài ra, việc ứng dụng giá trị Shapley trong giải thích tầm quan trọng của đặc trưng cũng là một hướng nghiên cứu đáng quan tâm[19]. Trong tương lai, chúng tôi sẽ nghiên cứu để tích hợp việc ước lượng Shapley vào các giải thuật đã được đề xuất trong bài báo này. Kỹ thuật hiện tại được đề xuất và thử nghiệm cho tác vụ SRL trên văn bản Y Sinh, việc nghiên cứu nhằm tổng quát hóa cho các tác vụ NLP khác và trên những lĩnh vực văn bản khác cũng là một hướng phát triển tương lai có nhiều ý nghĩa.

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 2024-13.

DANH MỤC TỪ VIẾT TẮT

PAS: Predicate Argument Structure
SRL: Semantic Role Labelling
X-NLP: Explainable Natural Language Processing

XUNG ĐỘT LỢI ÍCH TÁC GIẢ

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

ĐÓNG GÓP CỦA TÁC GIẢ

Bài báo do tác giả thực hiện tất cả các bước.

TÀI LIỆU THAM KHẢO

1. Bender EM, Gebru T, Mcmillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM; 2021. p. 610–623. Available from: https://doi.org/10.1145/3442188.3445922.
2. Garrido-Muñoz I, Montejo-Ráez A, Martínez-Santiago F, Ureña-López LA. A Survey on Bias in Deep NLP. Applied Sciences. 2021;11(7):3184–3184. Available from: https://doi.org/10.3390/app11073184.
3. Danilevsky M, Qian K, Aharonov R, Katsis Y, Sen P. A Survey of the State of Explainable AI for Natural Language Processing. Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. 2020;p. 447–459. Available from: https://xainlp2020.github.io/xainlp/.
4. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. ACM Comput Surv. 2022;54(6):1–35. Available from: https://doi.org/10.1145/3457607.
5. Madsen A, Reddy S, Chandar S. Post-hoc Interpretability for Neural NLP: A Survey. ACM Comput Surv. 2023;55(8):1–42. Available from: https://doi.org/10.1145/3546577.
6. Chou WC, Tsai RTH, Ying-Shan WK, Su TY, Sung WL, Hsu. A Semi-Automatic Method for Annotating a Biomedical Proposition Bank. In: Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora. Association for Computational Linguistics; 2006. p. 5–12.
7. Pezeshkpour P, Tian Y, Singh S. Investigating Robustness and Interpretability of Link Prediction via Adversarial Modifications. Proceedings of NAACL-HLT 2019. 2019;p. 3336–3347. Available from: http://dx.doi.org/10.18653/v1/N19-1337.
8. Bhan M, Achache N, Legrand V, Blangero A, Chesneau N. Evaluating self-attention interpretability through human-grounded experimental protocol. 2023;Available from: https://arxiv.org/abs/2303.15190.
9. Agnew E, Qiu M, Zhu L, Wiseman S, Rudin C. The Mechanical Bard: An Interpretable Machine Learning Approach to Shakespearean Sonnet Generation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. vol. 2. Association for Computational Linguistics; 2023. p. 1627–1638. Available from: https://doi.org/10.18653/v1/2023.acl-short.140.
10. Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable Prediction of Medical Codes from Clinical Text. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. vol. 1. Association for Computational Linguistics; 2018. p. 1101–1111. Available from: https://doi.org/10.18653/v1/N18-1100.
11. Kipper K, Dang H, Trang P, Martha. Class-Based Construction of a Verb Lexicon. Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence. 2000;p. 691–696.
12. Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet Project. In: Proceedings of the 36th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics; 1998. p. 86–90. Available from: http://doi.org/10.3115/980845.980860.
13. Pradhan S. PropBank Comes of Age-Larger, Smarter, and more Diverse. In: Proceedings of the 11th Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics; 2022. p. 278–288. Available from: https://doi.org/10.18653/v1/2022.starsem-1.24.
14. Thompson P. Building a Bio-Event Annotated Corpus for the Acquisition of Semantic Frames from Biomedical Corpora. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08); 2008. p. 123.
15. Wattarujeekrit T, Shah PK, Collier N. PASBio: predicate-argument structures for event extraction in molecular biology. BMC Bioinformatics. 2004;5(1):155. Available from: https://doi.org/10.1186/1471-2105-5-155.
16. Jacovi A, Goldberg Y. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In: " in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020. p. 4198–4205. Available from: https://doi.org/10.18653/v1/2020.acl-main.386.
17. Madsen A, Meade N, Adlakha V, Reddy S. Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining. In: Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics; 2022. p. 1731–1751. Available from: https://arxiv.org/abs/2110.08412.
18. Ribeiro M, Singh S, Guestrin C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Association for Computational Linguistics; 2016. p. 97–101. Available from: https://doi.org/10.18653/v1/N16-3020.
19. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017;p. 4768–4777. Available from: https://dl.acm.org/doi/10.5555/3295222.3295230.
20. Kennedy B, Jin X, Davani A, Dehghani M, Ren X. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020. p. 5435–5442. Available from: https://doi.org/10.18653/v1/2020.acl-main.483.
21. Jin X, Wei Z, Du J, Xue X, Ren X. Towards Hierarchical Importance Attribution: Explaining Compositional Semantics for Neural Sequence Models. Proceedings of the International Conference on Learning Representations. 2020;Available from: https://arxiv.org/abs/1911.06194.
22. Wich M, Bauer J, Groh G. Impact of Politically Biased Data on Hate Speech Classification. In: Proceedings of the Fourth Workshop on Online Abuse and Harms. Association for Computational Linguistics; 2020. p. 54–64. Available from: https://doi.org/10.18653/v1/2020.alw-1.7.
23. Li J, Chen X, Hovy E, Jurafsky D. Visualizing and Understanding Neural Models in NLP. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics; 2016. p. 681–691. Available from: https://doi.org/10.18653/v1/N16-1082.
24. Kindermans PJ. The (Un)reliability of Saliency Methods. 2019;p. 267–280. Available from: https://doi.org/10.1007/978-3-030-28954-6_14.
25. Jain S, Wallace BC. Attention is not Explanation. In: Proceedings of the 2019 Conference of the North. Association for Computational Linguistics; 2019. p. 3543–3556. Available from: https://doi.org/10.18653/v1/N19-1357.
26. Wiegreffe S, Pinter Y. Attention is not not Explanation. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019;18:11–20. Available from: http://dx.doi.org/10.18653/v1/D19-1002.
27. Ghaeini R, Fern X, Tadepalli P. Interpreting Recurrent and Attention-Based Neural Models: a Case Study on Natural Language Inference. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2018. p. 4952–4957. Available from: https://doi.org/10.18653/v1/D18-1537.
28. Li D, Hu B, Chen Q, He S. Towards Faithful Explanations for Text Classification with Robustness Improvement and Explanation Guided Training. In: Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing. Association for Computational Linguistics; 2023. p. 1–14.

29. Aksenov D, Bourgonje P, Zaczynska K, Ostendorff M, Moreno-Schneider J, Rehm G. Fine-grained Classification of Political Bias in German News: A Data Set and Initial Experiments. In: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021). Association for Computational Linguistics; 2021. p. 121–131. Available from: https://doi.org/10.18653/v1/2021.woah-1.13.
30. Tang R. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. vol. 1. Association for Computational Linguistics; 2023. p. 5644–5659. Available from: http://doi.org/10.18653/v1/2023.acl-long.310.
31. Clouatre L, Parthasarathi P, Zouaq A, Chandar S. Local Structure Matters Most: Perturbation Study in NLU. In: Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics; 2022. p. 3712–3731. Available from: http://doi.org/10.18653/v1/2022.findings-acl.293.
32. Sinha K, Parthasarathi P, Pineau J, Williams A. UnNatural Language Inference. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. vol. 1. Association for Computational Linguistics; 2021. p. 7329–7346. Available from: https://doi.org/10.18653/v1/2021.acl-long.569.
33. Zafar MB, Donini M, Slack D, Archambeau C, Das S, Kenthapadi K. On the Lack of Robust Interpretability of Neural Text Classifiers. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics; 2021. p. 3730–3740.
34. Yoo JY, Qi Y. Towards Improving Adversarial Training of NLP Models. Findings of the Association for Computational Linguistics: EMNLP 2021. 2021;p. 945–956. Available from: http://dx.doi.org/10.18653/v1/2021.findings-emnlp.81.
35. Ross AS, Hughes MC, Doshi-Velez F. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence; 2017. p. 2662–2670. Available from: https://www.ijcai.org/proceedings/2017/371.
36. Hsu ST, Moon C, Jones P, Samatova N. An Interpretable Generative Adversarial Approach to Classification of Latent Entity Relations in Unstructured Sentences. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. 2018;p. 5181–5188. Available from: https://doi.org/10.1609/aaai.v32i1.11972.
37. Tuan-Nguyen HSHD, Van-Thuc Pham, Hoang. A semi-automatic approach to biomedical semantic role corpus construction. Science and Technology Development Journal - Natural Sciences. 2022;6(2):2083–2094. Available from: https://doi.org/10.32508/stdjns.v6i2.1151.
38. Majewska O. BioVerbNet: a large semantic-syntactic classification of verbs in biomedicine. J Biomed Semantics. 2021;12(1). Available from: https://doi.org/10.1186/s13326-021-00247-z.
39. Voskarides N, Meij E, Tsagkias M, De Rijke M, Weerkamp W. Learning to Explain Entity Relationships in Knowledge Graphs. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. vol. 1. Association for Computational Linguistics; 2015. p. 564–574. Available from: https://doi.org/10.3115/v1/P15-1055.
40. Carton S, Mei Q, Resnick P. Extractive Adversarial Networks: High-Recall Explanations for Identifying Personal Attacks in Social Media Posts. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2018. p. 3497–3507. Available from: https://doi.org/10.18653/v1/D18-1386.
41. Rajani NF, Mccann B, Xiong C, Socher R. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2019. p. 4932–4942. Available from: https://doi.org/10.18653/v1/P19-1487.
42. Subramanian S. Obtaining Faithful Interpretations from Compositional Neural Networks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020. p. 5594–5608. Available from: https://doi.org/10.18653/v1/2020.acl-main.495.

Science & Technology Development Journal – Natural Sciences 2025, 9(3):3402-3415

Open Access Full Text Article

Research Article

Explaining semantic role labelling outputs with word feature importance

Hoai-Duc Tuan-Nguyen[1,2,*]

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

ABSTRACT

While deep learning models show impressive performance in natural language processing (NLP) tasks, they raise concerns about reliability and ethical implications due to their black-box nature. Consequently, a burgeoning area of research focuses on elucidating NLP models by clarifying the importance of input features to the models' output predictions. Among the various levels of NLP predictions, sequence-level predictions are crucial for tasks such as named entity recognition, semantic role labeling, and event extraction. However, current explanation techniques have largely overlooked sequence-level predictions. This paper presents an approach to explaining the importance of word features in sequence-level predictions within the semantic role labeling task. Our method evaluates both the impact and usefulness of each word in a sentence regarding model predictions. Additionally, we propose a novel data perturbation mechanism that strategically selects substitution words to more effectively mask word feature information, addressing the limitations of existing data perturbation techniques. Through experiments on biomedical texts, we demonstrate the effectiveness of our explanation method in providing explanations that are both comprehensible to humans and faithful to the actual processing within NLP models.

Key words: predicate-argument structure, semantic role labelling, explainability, interpretability, explainable NLP

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

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

Correspondence

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

Vietnam National University Ho Chi Minh City, Vietnam

Email: tnhduc@fit.hcmus.edu.vn

History

- Received: 16-07-2024
- Revised: 14-07-2025
- Accepted: 14-07-2025
- Published Online:12-09-2025

DOI :
https://doi.org/10.32508/stdjns.v9i3.1397

Cite this article : Tuan-Nguyen H. **Explaining semantic role labelling outputs with word feature importance**. *Sci. Tech. Dev. J. - Nat. Sci.* 2025; 9(3):3402-3415.