Hoàn cảnh nghiên cứu

Bài toán nhận diện tên thực thể Y Sinh

Nhận diện tên thực thể là một trong những bài toán quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên với nhiệm vụ là nhận diện được tên các thực thể trong một đoạn văn. Các tên thực thể có thể kể đến là tên người, tên địa điểm, ngày tháng, thời gian khi tiếp cận với các văn bản trong các lĩnh vực chung. Trong khi đó, tên của các loại gene, protein, tên bệnh, tế bào, tên thuốc, ... là các thực thể trong các văn bản Y Sinh, hay còn gọi là tên các thực thể Y Sinh.

Ví dụ	Loại thực thể
PuB1	Gene
CD28	Protein
Interferon	Thuốc
methylmercury, docosahexaenoic acid	Chất hóa học
Haemoglobin C disease	Tên bệnh

Bảng 1.1: Ví dụ một vài tên thực thể Y Sinh và phân loại thực thể

Bảng 1.1 chỉ ra một vài tên thực thể Y Sinh và loại thực thể tương ứng. Nhiệm vụ của bài toán nhận diện tên thực thể Y Sinh là tìm ra được tất cả những thực thể như vậy trong một đoạn văn.

Tầm quan trọng của bài toán nhận diện tên thực thể Y Sinh nằm ở việc đó là một trong các bước tiền xử lý cho việc khai phá dữ liệu hướng lĩnh vực Y Sinh. Tầm quan trọng này càng được nâng lên khi số bài nghiên cứu trong lĩnh vực Y Sinh ngày càng tăng lên nhanh chóng. Lấy ví dụ một cơ sở dữ liệu nổi tiếng chưa các thông tin về đời sống khóa học cũng như các thông tin về Y Sinh như kháng sinh, dược, chăm sóc sức khỏe, ... là MEDLINE. MEDLINE

được soạn thảo với thư viện quốc gia về Y tế của Mỹ. Cơ sở dữ liệu này là miễn phí và còn có thể truy cập bởi PubMed. Vào năm 2008, số lượng bài báo đã là hơn 17 triệu từ khoảng 5000 tạp chí về Y Sinh. Đến năm 2017, con số này đã là hơn 24 triệu trích dẫn. Tên của các thực thể có thể được sử dụng để tạo ra các chú thích về ngữ nghĩa của một văn bản, xây dựng ra từ khóa đặc trưng cho việc tra cứu. Hay việc nhận diện cũng giúp xây dựng bộ dữ liệu về mối quan hệ thực thể bệnh, thuốc hoặc các triệu chứng đi kèm. Ngoài ra, một ứng dụng khác của bài toán nhận diện tên thực thể là hệ thống trả lời câu hỏi. Các tên thực thể vừa được sử dụng để xử lý truy vấn cũng như tìm kiếm câu trả lời. Hoặc với bài toán trả lời câu hỏi, các kĩ thuật nhận diện tên thực thể được sử dụng như một công cụ giúp chọn ra các câu trả lời phù hợp. Tại hội nghị TREC-8 (Text REtrieval Conference) đã chỉ là 80% các truy vấn dùng để đánh giá các hệ thống này là các câu hỏi "who", "where", "when"[**Srihari2000**] - "ai", "ở đâu", "khi nào vốn có thể trả lời với tên người, tổ chức, địa điểm hoặc ngày tháng.

Với sự phát triển nhanh như vậy, việc luôn tiếp cận mới những kết quả nghiên cứu mới là một thách thức. Chỉ có thể bằng sự trợ giúp của các kĩ thuật khai phá dữ liệu để tiếp thu tri thức và thông tin từ các nghiên cứu đó và góp phần bổ sung cho chính nghiên cứu đang thực hiện. Với các nhà nghiên cứu lĩnh vực này, tên của các thực thể là đơn vị cơ bản nhất nhưng chưa nhiều tri thức và thông tin. Để hiểu được bài viết thì bắt buộc phải hiểu tên của các thực thể và ý nghĩa của chúng. Thêm nữa, việc nhận diện còn giúp trích xuất mối quan hệ hay các thông tin khác bằng việc xác định nội dung chính, biểu diễn những nội dung đó một cách nhất quán và theo định dạng chuẩn.

Tuy nhiên, bài toán nhận diện tên thực thể Y Sinh là không dễ. Điều đầu tiên làm nên độ khó của bài toán là ta không có từ điển nào hoàn chỉnh cho đa phần các loại tên thực thể Y Sinh. Vì thế việc chỉ sử dụng thuật toán so khớp xâu không thể đủ để giải quyết. Đã có một vài dữ liệu được gán nhãn một cách thủ công, tuy nhiên các dữ liệu đó lại quá tập trung vào một miền con và không thể bao phủ hết các tên thực thể ở những miền con đó chứ chưa

nói đến những cái tên liên tục được tạo ra và chưa kịp đưa vào bộ từ điển ngay lập tức.

Thứ hai, tên của các thực thể Y Sinh thường gây mơ hồ, hay xuất hiện từ đồng nghĩa và các tên biển thể vì không có quy tắc cố định về việc đặt tên mặc dù đã có những hướng dẫn. Cùng một từ hay một cụm từ có thể nhắc đến các loại thực thể khác dựa trên ngữ cảnh. Ví dụ "ferritin"có thể vừa là chất sinh học hoặc vừa là tên một loại thí nghiệm. Rồi nhiều thực thể có chung tên như "PTEN"hay "MMAC1"đều nhắc đến cùng 1 gene. Thông thường các tên giống nhau thường có một ít biển đổi về chính tả (NF kappa B còn có thể viết là NF-kappa B hoặc NF kappa-B). Cho dù khi các nhà nghiên cứu chỉ sử dụng những cái tên được chuẩn hóa và chấp nhận thì vẫn còn một số lượng lớn các tài liệu trước chứa tên thực thể mà cần suy luận.

Một đặc điểm khác của tên các thực thể Y Sinh là thường bao gồm nhiều từ gộp lại (ví dụ CD28 surface receptor), và những tên thực thể ngắn hơn có thể gộp lại với nhau tạo thành một cái tên dài hơn. Đặc điểm này đặt ra bài toán con rằng làm sao để xác định biên của tên và xử lý tên các thực thể lồng nhau. Nhất là khi trước khi xử lý dữ liệu thường phải đưa qua bước tách từ thì tên những thực thể như "an alpha galactosyl-1,4-beta-galactosyl-specific adhension"bị phân mảnh quá nhiều khiến cho việc nhận diện tên gặp khó khăn. Hay tên của 2 tế bào là "NB4"và "APL"đôi khi còn xuất hiện cùng nhau tạo thành "NB4-APL". Thực tế thì NB4 là dòng tế bào đi ra từ tế bào acute promyelocytic leukemia (APL). Có đến 90% tên các thực thể chứa từ 2 đến 3 từ trong bộ dữ liệu GENIA, trong khi tên các thực thể gồm 6 từ trở lên là hiếm dù đúng là chúng có tồn lại.

Trong các văn bản Y Sinh còn thường gặp các tên viết tắt. Gần một nửa các tóm tắt trên MEDLINE chưa các từ viết tắt như đã đề cập trong [**chang2004gapscore**]. Trong năm 2004, có 64262 tên viết tắt được đưa ra, và trung bình có từ một từ viết tắt mới trong khoảng từ 5 đến 10 tóm tắt. Nhiều từ viết tắt như vậy nhưng chỉ mới một phần nhỏ được đưa vào từ điển để nhận biết. Việc viết tắt cũng là nguyên nhân cho sự thiếu rành mạnh và rõ ràng.

Ví dụ ACE có thể là angiotensin converting enzyme nhưng cũng có thể hiểu là affinity capillary electrophresis...

Tên các thực thể Y Sinh được kì vọng là sẽ rất lớn, và các tên mới được phát minh ra hàng ngày. Vì thế, bài toán phải đảm bảo có thể mở rộng hay nói cách khác là nhận diện được đa phần tên các thực thể trong thời gian chấp nhận được. Đồng thời, bài toán còn phải phát hiện được những tên mà chưa từng thấy.

Cuối cùng, các kĩ thuật dùng cho nhận diện tên thực thể trong văn bản thường ngày không thể áp dụng trực tiếp cho bài toán nhận diện tên thực thể Y Sinh vì tên các thực thể trong miền dữ liệu có những đặc điểm rất khác với tên các thực thể thường thấy. Ví dụ tên các thực thể trong văn bản thường ngày thường là danh từ riêng với các chữ cái được viết hóa, trong khi rất nhiều tên thực thể Y Sinh không phải danh từ riêng (hơn 60% từ trong các thực thể Y Sinh là chữ cái viết thường [**zhou2004recognizing**]).

Nhìn chung, tên của các thực thể có thể xuất hiện ở nhiều dạng khác nhau trong các văn bản về Y Sinh. Và bài toán nhận diện tên thực thể này phải xử lý được nhiều nhất các trường hợp đó, tạo tiền đề cho các bài toán phía sau cũng dựa trên kết quả của quá trình này.

Các nghiên cứu liên quan

Các phương pháp tiếp cận bài toán thường rơi vào một trong các nhóm: dựa trên từ điển, dựa trên luật hoặc dùng mạng nơ-ron. Việc dựa vào từ điển sẽ tìm ra các tên trong một đoạn văn dựa vào một từ điển đã định nghĩa sẵn một cách thủ công hoặc tự động. Phương pháp sử dụng luật sẽ tận dụng những luật hoặc thành phần được định nghĩa từ trước để so khớp với tên thực thể. Phương pháp sử dụng mạng nơ-ron sẽ sử dụng các kĩ thuật để tạo thành mô hình dự đoán tên các thực thể đó.

- Phương pháp dựa trên từ điển

Phương pháp này cần trước một danh sách tên các thực thể Y Sinh đã biết, thường là lấy từ các cơ sở dữ liệu. Trong miền dữ liệu Y Sinh có một vài cơ sở dữ liệu như vậy cho gene (GeneBank), protein (UniProt) hay chất hóa học (ChemIDplus). Việc tìm kiếm tên sẽ dựa trên so khớp một phần hoặc so khớp toàn phần. Thế mạnh của phương pháp này là đơn giản và hiệu quả vì thuật toán so khớp xâu đã được nghiên cứu rất nhiều trong lĩnh vực khoa học máy tính. Vấn đề lớn nhất là độ chính xác không cao, lý giải là vì các cơ sở dữ liệu bao phủ không nhiều, thường là chuyên biệt về một vài loại thực thể và thường không cập nhật thường xuyên. Lý do khác có thể do sự biến đổi về mặt chính tả khi thay đổi thứ tự từ trong một thực thể hoặc các từ viết tắt. Một vài mô hình đã sử dụng phương pháp này và đạt kết quả bước đầu như [**tsuruoka2003probabilistic**], [**tanabe2002tagging**], [**egorov2004simple**]

- Tiếp cận dựa trên luật

  Phương pháp tiếp cận này sẽ phát hiện ra tên các thực thể dựa trên luật đã được định nghĩa từ trước. Các luật thường mô tả cấu trúc tên sử dụng hình thái từ các các đặc trưng khác ví dụ như việc kết hợp từ và số, sự xuất hiện của kí tự đặc biệt, các danh từ hay tính từ lạ. Có được những luật như vậy yêu cầu phải có kiến thức sâu về miền, về ngôn ngữ và cả ngôn ngữ lập trình. Điểm mạnh của phương pháp này là các thuật được thiết kế cẩn thận để giải quyết được đặc trưng về ngôn ngữ.

- Phương pháp tiếp cận học máy

  Việc nhận diện tên thực thể Y Sinh thường được đưa về bài toán gán nhãn chuỗi với mục tiêu là gán một nhãn cho mỗi từ trong một chuỗi. Các hệ thống nhận diện tên thực thể Y Sinh đạt kết quả tốt nhất thường yêu cầu các đặc trưng định nghĩa thủ công như đặc trưng về chữ hoa, tiền tố hay hậu tố ... Các đặc trưng này sẽ được thiết kế riêng cho từng loại thực thể. Một vài mô hình như vậy đã được đề cập trong [**leaman2016taggerone**], [**huang2016community**] hay [**zhou2004exploring**]. Quá trình tạo ra

các đặc trưng này chiếm phần lớn và thời gian và chi phí khi phát triển một hệ thống nhận diện tên thực thể [**leser2005makes**], đồng thời hướng đến các hệ thống đặc thù mà không thể trực tiếp sử dụng để nhận diện các loại thực thể khác.

Một vài nghiên cứu gần đây sử dụng mạng nơ-ron để tự đồng tạo ra các đặc trưng có giá trị như [**chiu2016named**], [**ma2016end**], [**lample2016neural**], [**liu2018empower**]. Mô hình của Crichton lấy mỗi từ và các từ xung quanh làm ngữ cảnh thành đầu vào cho mạng nơ-ron tích chập (CNN). Habibi sử dụng mô hình giống với Lample, đồng thời thêm vào các word embedding làm đầu vào cho mạng BiLSTM-CRF. Các mô hình này giúp cho ta không phải tạo ra các đặc trưng một cách thủ công. Tuy nhiên, cũng các mô hình này yêu cầu hàng triệu tham số và cần dữ liệu rất lớn để ước lượng được các tham số đó. Điều này là thách thức với lĩnh vực Y Sinh khi những dữ liệu thô thì có nhiều nhưng những dữ liệu đã được gán nhãn thì lại rất mất công để thực hiện, đi kèm đó là tốn chi phí. Thế nên mặc dù các mạng nơ-ron cho thấy kết quả vượt trội với các phương pháp gán nhãn chuỗi truyền thống như mô hình CRF [**lafferty2001conditional**], các kết quả vẫn không vượt qua được những mô hình dựa trên đặc trưng thủ công.

Cấu trúc khóa luận

- Đầu tiên, khóa luận đề cập đến bài toán nhận diện tên thực thể Y sinh, bối cảnh cùng các đặc điểm và thách thức của nghiên cứu.
- Chương thứ 2 nêu ra các kiến thức nền tảng. Chương này tập trung vào các lý thuyết chung, các thành phần tạo nên mô hình để giải quyết bài toán nhận diện tên thực thể trong khóa luận bao gồm word embedding, các mạng nơ-ron đặc biệt, mô hình CRF và phương pháp học đa tác vụ.
- Chương tiếp theo mô tả phương pháp tiếp cận để giải bài toán bằng cách áp dụng việc học đa tác vụ dựa trên một mô hình đơn lẻ ban đầu nhằm

cải thiện độ chính xác của bài toán.

- Chương thứ 4 ghi lại môi trường, các tham số và đặc biệt là kết quả thực nghiệm sử dụng mô hình trong khóa luận, phân tích nhận xét về kết quả.
- Kết luận đánh giá những kết quả mà qua khóa luận đạt được trong việc nghiên cứu áp dụng học sâu đa tác vụ vào bài toán chính, đồng thời đưa ra phương hướng để cải tiến trong tương lai