Kết quả đạt được từ khóa luận

Khóa luận này đã sử dụng hướng tiệp cận bài toán nhận diện tên các thực thể Y Sinh trong một văn bản bằng cách học nhiều tác vụ liên quan, làm giàu thông tin từ các tác vụ đó. Quá trình nghiên cứu, tìm hiểu, xây dựng và thử nghiệm mô hình theo hướng tiếp cận này đã đem đến những kết quả như:

- Đem đến cái nhìn khác cho bài toán nhận diện tên thực thể. Bài toán này giờ đây được tiếp cận bằng cách gán nhãn chuỗi bằng các định danh chỉ một từ nằm trong, bắt đầu, kết thúc, chính từ đó là tên thực thể hoặc không thuộc về bất kì tên thực thể nào.
- Mô hình tương đối đơn giản chỉ với 2 tầng BiLSTM và 1 tầng CRF đặc trưng cho mỗi tác vụ. Tuy mô hình không phức tạp nhưng lại đem lại những cải tiến cho bài toán khi so sánh với việc chỉ sử dụng một tác vụ. Điều này có được là nhờ việc chia sẻ thông tin về từ cũng như về kí tự giữa các loại thực thể Y Sinh.
- Nghiên cứu về mức độ tương quan giữa các tập dữ liệu về Y Sinh. Việc kết hợp các tập dữ liệu phù hợp sẽ giúp cho việc nhận diện tên thực thể hiệu quả hơn. Ngược lại, việc chọn không đúng thậm chí có thể gây nhiễu cho tác vụ mà mô hình muốn cải tiến.
- Tìm hiểu về mô hình tạo ra biểu diễn vector của các từ sử dụng công cụ Word2Vec và fastText (BioWordVec) và thực nghiệm áp dụng 2 công cụ vào mô hình của khóa luận để quyết định bộ embedding nào là phù hợp hơn. Qua thực nghiệm đã thấy được BioWordVec đem lại hiệu quả tốt hơn nhờ việc biểu diễn được cả những từ không nằm trong từ điển so với Word2Vec đã huấn luyện trên bộ dữ liệu Y Sinh.

Phương hướng tiếp cận trong tương lai

Tuy nhiên mô hình này vẫn có thể cải tiến thêm theo những hướng đi như:

- Chỉnh sửa tham số về mức độ tham gia của mỗi tập dữ liệu thay vì để

mặc định mỗi bộ dữ liệu đều có vai trò tương đương nhau tham gia vào kết quả chung.

- Sử dụng những phép biểu diễn từ khác như BERT, Elmo, ... là những phép biểu diễn từ dựa trên cả ngữ nghĩa của những từ xung quanh. Việc sử dụng được các biểu diễn mang nhiều thông tin có thể sẽ giúp việc nắm bắt thông tin về ngữ nghĩa được tốt hơn, làm cơ sở cho việc kết hợp thông tin về từ và về kí tự một từ một cách hiệu quả.
- Kết hợp học đơn tác vụ và đa tác vụ cũng có thể là cách để giúp mô hình nhẹ hơn, đồng thời vẫn cải tiến được độ chính xác của mô hình.
- Tìm hiểu và thử nghiệm thêm các bộ dữ liệu đã được gán nhãn khác dựa trên sự tương quan về loại thực thể cũng là một hướng cần nghiên cứu thêm.