Tập dữ liệu

Các thực nghiệm của bài toán nhận diện tên thực thể Y Sinh trong khóa luận này sẽ sử dụng 5 bộ dữ liệu có tên BC2GM, BC4CHEMD, BC5CDR, NCBI-disease và JNLPBA theo gợi ý của [**wang2018cross**]. Mỗi bộ dữ liệu đều chia ra thành tập huấn luyện, tập phát triển và tập kiểm tra. Tập huấn luyện và tập phát triển được gộp chung để huấn luyện mô hình như trong các nghiên cứu [**luo2018attention**], [**lu2015chemdner**] hay [**leaman2016taggerone**]. Tất cả các tập dữ liệu đều được công khai. Bộ nhãn được dùng trong bài toán là IOBES. Ví dụ một từ mô tả tên một bực thể được gán nhãn là "B-Gene"khi và chỉ khi từ đó đánh dấu bắt đầu tên một thực thể, gán nhãn là "I-Gene"khi từ đó ở giữa tên một thực thể và "E-Gene"khi từ đó là từ cuối cùng xuất hiện trong tên. Bên cạnh đó, các tên thực thể có 1 từ thì được gán nhãn là S- và đi kèm sau đó là loại thực thể, ví dụ như "S-Gene". Tất cả những từ khác không giúp mô tả tên thực thể ta cần quan tâm sẽ gán nhãn là O. Các tập dữ liệu tuân theo định dạng giống như tập dữ liệu CoNLL 2003 NER. Cụ thể hơn, các dòng trống thể hiện ngăn cách giữa các câu, và ngăn cách giữa các tài liệu có dạng là:

```
-DOCSTART- -X- -X- -X- O
```

Các dòng khác chứa từ và nhãn của mỗi từ. Trường đầu tiên của mỗi dòng phải là từ, và trường cuối cùng trong dòng là nhãn của từ.

Bảng 4.1 mô tả kích cỡ cùng loại dữ liệu của mỗi tập dữ liệu.

Cách đánh giá

Khóa luận sẽ báo cáo trên tập kiểm thử. Mỗi tên thực thể do mô hình dự đoán được xem là đúng khi và chỉ khi loại thực thể và tên thực thể khớp hoàn toàn với dữ liệu thực tế. Đây gọi là đánh giá so khớp trên mức tên thực thể.

Ma trận lỗi của bài toán nhận diện tên thực thể bao gồm:

Bảng 4.1: Các tập dữ liệu Y Sinh sử dụng cho thí nghiệm

Tập dữ liệu	Kích cỡ	Loại thực thể và số lượng
BC2GM	20000 câu	Gene/Protein (24583)
BC4CHEMD	10000 tóm tắt	Chất hóa học (84310)
BC5CDR	1500 bài báo	Chất hóa học (15935), Tên bệnh (12852)
NCBI-Disease	793 tóm tắt	Tên bệnh (6881), Gene/Protein (35336)
JNLPBA	2404 tóm tắt	Dòng tế bào gốc (4330), DNA (10589) Loại tế bào (8649), RNA (1069)

		Nhãn thật sự	
		True	False
Nhãn dự đoán	Positive	True Positive	False Positive
	Negative	True Negative	False Negative

Bảng 4.2: Định nghĩa ma trận lỗi của bài toán nhận diện tên thực thể

Xét trong một thực thể. Khi các từ trong tên của thực thể gán nhãn sai tất cả sẽ tính số từ gán nhãn sai đó vào False Negative. Ngược lại, khi tất cả các từ trong tên một thực thể được gán nhãn đúng, số lượng nhãn đúng tính vào mục True Positive. Thêm vào đó, trong trường hợp mô hình đoán sai tên biên của thực thể, số nhãn sai trong tên đó được tính vào False Negative. Không kể các nhãn được gán đúng, các nhãn gán sai khi gặp trường hợp xác định không đúng biên được tính vào False Positive.

Sau đó báo cáo kết quả sử dụng độ đo F1 trên toàn bộ các tập dữ liệu.

Môi trường và các tham số

Để huấn luyện mô hình, tham số tốc độ học đặt là 0.01. Chiều embedding của từ là 200, chiều embedding của kí tự là 30. Nhằm giúp mô hình tránh gặp tình trạng quá khớp, hệ số dropout đặt là 0.5, đồng nghĩa sẽ lược bỏ 50% đỉnh

trong mạng khi huấn luyện BiLSTM. Bên cạnh đó, mô hình sẽ được chỉ định dừng lại khi số epoch tối đa không cho ra kết quả f1 cải thiện là 30. Kích cỡ của mỗi lô là 10, đồng thời mô hình phải chạy ít nhất 50 epoch cho mỗi lần huấn luyện.

Mô hình cài đặt sử dụng Python và Pytorch. Pytorch là thư viện mã nguồn mở về các thuật toán học máy, thường sử dụng cho các ứng dụng như xử lý ngôn ngữ tự nhiên hoặc thị giác máy bên cạnh các ưu điểm như nhiều tài liệu, cộng đồng hỗ trợ đông (hơn 30000 người).

Các thực nghiệm

- So sánh giữa mô hình học đơn tác vụ và đa tác vụ

Tập dữ liệu	Đơn tác vụ	Đa tác vụ
BC2GM	79.64	79.98
BC4CHEMD	87.51	89.02
BC5CDR	86.95	88.75
NCBI-disease	83.74	86.61
JNLPBA	72.02	73.45

Bảng 4.3: Kết quả khi chạy mô hình học đơn tác vụ và đa tác vụ

Bảng 4.3 ghi lại số đo F1 khi huấn luyện mô hình chỉ sử dụng một bộ dữ liệu duy nhất và sử dụng kết hợp nhiều bộ dữ liệu với nhau. Số liệu cho thấy kết quả đã có sự tiến bộ rõ rệt khi sử dụng mô hình học kết hợp nhiều tác vụ với nhau. Kết quả của tập dữ liệu BC2GM tăng ít nhất với mức tăng 0.34% trong khi kết quả cải tiến được nhiều nhất ở tập NCBI-disease với mức tăng 2.87%. Tổng mức tăng trên cả bộ dữ liệu là 7.95% khi áp dụng việc học đa tác vụ. Từ đó thấy được mô hình học đa tác vụ có thể

tận dụng được thông tin trên các bộ dữ liệu với nhau và đồng thời cùng nhau tăng độ chính xác của các trên các tác vụ.

- Mức độ tương quan giữa các tập dữ liệu

  Để xác định tác động của việc học kết hợp các bộ dữ liệu theo mô hình đa tác vụ, khóa luận sẽ chọn một ra n bộ dữ liệu (với $n \geq 2$ và $n \leq 4$ rồi huấn luyện mô hình trên n tập dữ liệu đó. Mỗi tập dữ liệu tương ứng với một tác vụ. Dấu "thể hiện tập dữ liệu đó không tham gia vào quá trình huấn luyện. Vì giới hạn tránh quá dài dòng, các kết quả quan trọng được thể hiện bằng bảng dưới đây:

Các tập dữ liệu	BC2GM	BC4CHEMD	BC5CDR	NCBI-disease	JNLPBA
BC2GM + BC4CHEMD + BC5CDR	79.75	88.84	88.32	-	-
BC2GM + BC5CDR + NCBI-disease	79.87	-	88.71	86.43	-
BC4CHEMD + BC5CDR	-	88.86	87.93	-	-
BC2GM + NCBI-disease + JNLPBA	79.85	-	-	86.33	73.31
BC4CHEMD + BC2GM	79.63	88.82	-	-	-
BC2GM + BC5CDR + NCBI-disease + JNLPBA	79.89	-	88.63	88.61	73.32
BC4CHEMD + JNLPBA	-	88.84	-	-	72.04

Bảng 4.4: Kết quả của từng bộ dữ liệu khi kết hợp huấn luyện với nhau

Số liệu từ bảng 4.4 cho thấy việc kết hợp học giữa các tập dữ liệu cũng sẽ ảnh hưởng đến kết quả. Nhìn chung, kết quả sẽ có sự cải thiện khi tăng số tác vụ lên theo số tập dữ liệu, đặc biệt với những bộ dữ liệu có đặc điểm chung về tên của thực thể. Cụ thể, việc sử dụng các tập dữ liệu có chung loại thực thể như BC2GM, BC5CDR và NCBI-disease cùng có nhiều thực thể về gene/protein (24583 và 35336 thực thể) sẽ giúp cho mô hình dự đoán được chính xác hơn (thể hiện qua f1 trên tập BC2GM tăng từ 79.80% lên 79.85% khi được huấn luyện cùng NCBI-disease và BC5CDR) thay vì việc kết hợp giữa các bộ dữ liệu không có chung loại thực thể rồi khiến kết quả giảm đi. Điều này có thể lý giải bằng việc khi chọn các tập dữ

liệu không có đặc điểm chung gần như có nghĩa rằng ta đang vừa huấn luyện các mô hình đơn tác vụ, đồng thời vô tình gây nhiễu giữa các tác vụ nhận diện với nhau.

Tuy nhiên với tập dữ liệu BC4CHEMD, kết quả thay đổi rất ít (chênh lệch khoảng 0.02% đến 0.04% dù đã sử dụng với bộ dữ liệu BC5CDR cũng có tên về các chất hóa học. Điều này là do tương quan về dữ liệu giữa 2 tập này không cao khi BC4CHEMD có đến 84310 tên chất hóa học, nhiều hơn gấp 5 lần so với số thực thể trong tập BC5CDR - chỉ có 15935 thực thể. Thêm nữa, việc kết hợp thêm tập BC4CHEMD còn khiến thời gian huấn luyện tăng lên trong khi như đã thực nghiệm ở trên, đôi khi còn gây khó khăn cho việc huấn luyện do kích cỡ của tập này không nhỏ (gồm 10000 tóm tắt chứa 84310 tên chất hóa học).

Kết quả trên tập dữ liệu JNLPBA đạt 73.31% và 73.32% khi huấn luyện cùng tập BC2GM và NCBI-disease. Kết quả này cho thấy cũng không có sự thay đổi nhiều khi được sử dụng để huấn luyện chung với các tập dữ liệu khác, do đặc điểm đây là bộ dữ liệu có nhiều loại thực thể, đồng thời cũng có độ nhiễu nhất định. Ví dụ như cụm từ "truncated RARalpha"thì "trucated"(dịch: đã cắt bớt) không mang lại bất kì định danh nào có ý nghĩa cho từ "RARalpha"phía sau. Tuy nhiên khi đặt vào huấn luyện chung với tập BC4CHEMD, F1 trên tập JNLPBA chỉ đạt 72.04%, cho thấy để đạt kết quả trên 73% trên JNLPBA cũng cần thiết có sự kết hợp việc học được các tên thực thể về gene/protein trên tập BC2GM và NBCI-disease.

- Ảnh hưởng của word embedding

  Để có được biểu diễn vector của các từ, mô hình trong khóa luận thử nghiệm vector lấy từ 2 mô hình là Word2Vec và BioWordVec. Mô hình Word2Vec đã được huấn luyện sử dụng skip-gram trên bộ dữ liệu Pubmed và Wikipedia, trong khi BioWordVec là mô hình fastText và được huấn luyện trên Pubmed và MeSH. Kết quả từ bảng 4.5 cho thấy việc sử dụng

vector biểu diễn từ sinh ra bởi mô hình BioWordVec giúp cho kết quả tốt hơn từ 0.03% cho đến 0.36%. Kết quả có được nhờ việc huấn luyện riêng từng tập dữ liệu. Lý giải cho cải tiến này là vì fastText có thể sinh ra được biểu diễn của những từ không nằm trong bộ từ vựng đã có dựa trên biểu diễn của những kí tự hoặc cụm kí tự cấu tạo nên từ, trong khi Word2Vec vẫn tồn tại những từ chưa có biểu diễn vector. Các từ mà chưa biết đó sẽ được thay bằng từ "UNK"và vector biểu diễn cho từ "UNK"sẽ được khởi tạo ngẫu nhiên. Kết quả trong bảng là số đo f1 khi chạy riêng từng tác vụ, chỉ thay đổi word embedding sử dụng cho từ là Word2Vec hay BioWordVec.

Tập dữ liệu	Word2Vec	BioWordVec
BC2GM	79.64	79.72
BC4CHEMD	87.48	87.51
BC5CDR	83.38	83.74
NCBI-disease	83.51	83.74
JNLPBA	71.89	72.02

Bảng 4.5: Kết quả mô hình đơn tác vụ khi sử dụng Word2Vec và BioWordVec

- Ảnh hưởng của việc lựa chọn tham số để chia sẻ giữa các tác vụ

  Như trong mô hình đề xuất, giữa các tác vụ sẽ sử dụng chung các tham số cho mô hình BiLSTM trên kí tự và trên các từ. Thử nghiệm với việc lần lượt chỉ cho các tác vụ chia sẻ tham số về BiLSTM ở mức kí tự và BiLSTM ở mức từ và để mô hình tự học ra các tham số còn lại cho từng tác vụ. Kết quả thu được như sau:

  MTL-C thể hiện việc huấn luyện 5 tập dữ liệu và chỉ chia sẻ tham số về BiLSTM mức kí tự. MTL-W thể hiện việc huấn luyện 5 tập dữ liệu và chỉ chia sẻ tham số về BiLSTM mức từ ngữ. MTC-WC là kết quả mô hình

Tập dữ liệu	MTL-C	MTL-W	MTL-CW
BC2GM	77.88	78.73	79.98
BC4CHEMD	88.46	88.65	89.02
BC5CDR	86.57	88.16	88.75
NCBI-disease	82.37	84.52	88.61
JNLPBA	71.31	71.69	73.45

Bảng 4.6: Kết quả khi chạy mô hình với các tham số dùng chia sẻ giữa các tụ được thay đổi

khi đề xuất trong khóa luận. Kết quả cho thấy mô hình MTL-CW cho kết quả tốt nhất, thường cao hơn cách chọn tham số là từ hay kí tự từ khoảng hơn 1% đến gần 6%. Điều này giúp củng cố rằng việc học được thông tin về hình thái từ bởi BiLSTM mức kí tự và thông tin về từ vựng cũng như ngữ cảnh ở mức từ là cần thiết để cải thiện kết quả việc nhận diện tên thực thể.