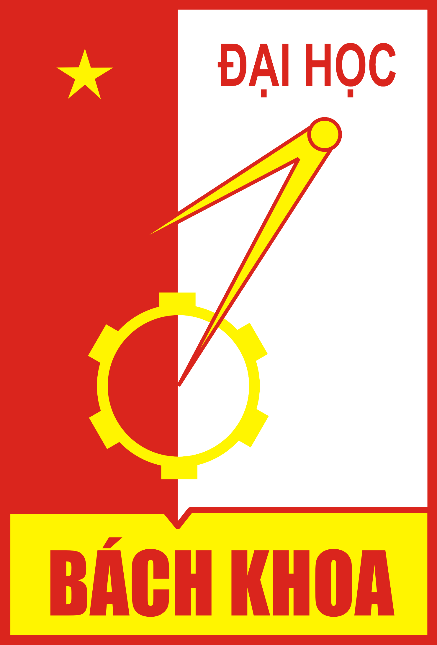
**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

Viện Công nghệ Thông tin và Truyền thông



**BÁO CÁO PROJECT 2**

Đề tài: **Nghiên cứu các giải pháp hiện có trên thế giới và**

**tại Việt Nam cho bài toán nhận diện thực thể có tên**

GVHD: TS. Đào Thành Chung

Lớp: LTU15

Sinh viên: Nguyễn Đức Thiên

MSSV: 20168806

HN 05/2020

# **LỜI NÓI ĐẦU**

Một trong những vấn đề nổi trội của lĩnh vực xử lý ngôn ngữ tự nhiên và các bài toán trích xuất thông tin hiện nay là bài toán nhận diện thực thể có tên. Trong phạm vi Project 2 này, em nghiên cứu và tìm hiểu về các phương pháp từ cổ điển đến hiện đại nhất để tiếp cận và giải quyết bài toán này.

Việc nghiên cứu và tìm hiểu về các phương pháp này giúp em nắm được và hiểu rõ bài toán, từ đó có thể tiếp tục xây dựng và đề xuất các giải pháp để cải tiến các phương pháp đã có cũng như xây dựng các hướng đi mới trong lĩnh vực xử lý ngôn ngữ tự nhiên nói chung và bài toán trích xuất thông tin nói riêng.

**Mục lục**

[LỜI NÓI ĐẦU 2](#_Toc45720166)

[Chương 1: Giới thiệu bài toán nhận diện tên thực thể có tên. 4](#_Toc45720167)

[1. Bài toán nhận diện tên thực thể có tên. 4](#_Toc45720168)

[3. Các vấn đề và thách thức hiện nay. 6](#_Toc45720169)

[Chương 2: Các hướng tiếp cận và giải quyết bài toán NER. 7](#_Toc45720170)

[1. Phương pháp sử dụng tập luật và từ điển. 7](#_Toc45720171)

[2. Phương pháp sử dụng học máy. 7](#_Toc45720172)

[3. Phương pháp sử dụng học sâu. 9](#_Toc45720173)

[Chương 3: Các giải pháp sử dụng phương pháp học sâu hiện nay. 12](#_Toc45720174)

[Chương 4: Thực nghiệm mô hình CNN + LTSM + CRFs 13](#_Toc45720175)

# **Chương 1: Giới thiệu bài toán nhận diện tên thực thể có tên.**

1. Bài toán nhận diện tên thực thể có tên.

Bài toán nhận diện thực thể có tên (Named Entity Regconition - NER) là bài toán nhận dạng được các thực thể xuất hiện trong văn bản. Thực thể là những đối tượng, chủ thể tồn tại một cách cụ thể (vd. “Hà Nội”) hay trừu tượng (vd. “quy định”), vật lý (vd. “Trái Đất”) hay phi vật lý (vd. “phương pháp”) [1]. Thực thể có tên là những thực thể có thể được đặt tên và được xác định là duy nhất trong mọi thế giới mà nó có thể tồn tại. Với đầu vào như trên, cần nhận dạng cả ba thực thể có tên là “*Macron”*, “*Mỹ”*, và “*Đại Tây Dương”*. Một thực thể có tên là “Barack Obama” có thể được nhắc đến trong các hoàn cảnh khác nhau như “*Obama là tổng thống của nước Mỹ.*” hay “*Ông đã đến thăm Việt Nam năm 2016*”. Trong câu đầu tiên, Obama là một thực thế có tên mặc dù ta chưa biết chắc “Obama” này có phải là Barack Obama hay không (thực chất đây là nội dung của một bài toán khác là liên kết thực thể). Trong câu thứ hai, “ông” không phải là một thực thể có tên. Tuy nhiên, ta cần xác định “ông” cùng chỉ đến một thực thể được nhắc đến trong câu đầu tiên bằng “Obama” (nội dung của bài toán phân giải đồng tham chiếu) và cả hai đều cùng chỉ đến một đối tượng “Barack Obama” cụ thể (nội dung của bài toán liên kết thực thể).

Thông thường, các loại thực thể có tên được quan tâm đến là con người (PERSON), tổ chức (ORGANIZATION), và địa điểm (LOCATION) (Sundheim & M. 1995 [2]):

* PERSON: Các cá nhân, các nhân vật hư cấu, các dòng họ…
* ORGANIZATION: Các cơ quan nhà nước, các đảng phái chính trị, các câu lạc bộ chuyên môn, các nhóm nhạc, các tổ chức phi chính phủ…
* LOCATION: Các đất nước, các địa danh hành chính, các cơ sở hạ tầng như nhà ga, bệnh viện, cầu, các địa danh tự nhiên như sông, núi, biển...
* Ngoài ra, các loại thực thể khác cũng được quan tâm đến như sự kiện (EVENT), tiền tệ (CURRENCY) .v.v.

1. Tập dữ liệu

* Vào năm 2018 thì bộ dữ liệu VLSP 2018 [3] được công bố với cấu trúc các nhãn tương tự như trong bộ dữ liệu CoNLL 2003 [4]. Các nhãn được chia làm ba loại nhãn chính là:
  + LOCATION: các thực thể thuộc nhãn LOCATION như đường xá, vùng miền, nơi công cộng, các công trình xây dựng, …
  + ORGANIZATION: các tổ chức, cá nhân, thương hiệu, phòng ban chính phủ, các tập thể, …
  + PERSON: họ, tên đệm, tên, chức danh của người, nhân vật trong văn học, vật nuôi hoặc bí danh.
* Các nhãn được đi kèm với một mô hình gán nhãn, dùng để phân đoạn các cụm từ với nhau. Mô hình đơn giản nhất là mô hình IO, mô tả mỗi nhãn là thuộc (I) hoặc không thuộc (O) thực thể có tên đó.
* Mô hình BIO được biết đến rộng rãi hơn, được công bố trong CoNLL Shared Tasks, bao gồm 3 ký hiệu B (Begin), I (Inside), O (Outside). B là để đánh dấu từ đầu tiên bắt đầu một cụm từ mô tả một thực thể có tên, theo sau đó là các nhãn loại I. Các từ không nằm trong thực thể có tên nào được đánh dấu O. Kết hợp lại ta có 7 loại nhãn được sử dụng trong tập dữ liệu VLSP 2018 là: B-LOC, I-LOC, B-PER, I-PER, B-ORG, I-ORG, O.
* Các câu trong tập dữ liệu được chia thành mỗi từ một dòng, các câu cách nhau bởi 1 dòng trống. Một dòng bao gồm 5 cột tương ứng với từ, loại từ, phân đoạn BIO của loại từ, nhãn của các thực thể có tên đầy đủ, nhãn của các thực thể có tên rút gọn. Ví dụ với câu Anh Thanh là cán bộ Uỷ ban nhân dân Thành phố Hà Nội, ta có bảng gán nhãn sau:

Anh N B-NP O O

Thanh NNP INP B-PER O

là V B-VP O O

cán\_bộ N B-NP O O

Uỷ\_ban N B-NP B-ORG O

nhân\_dân N I-NP I-ORG O

Thành\_phố N I-NP I-ORG B-LOC

Hà\_Nội NPP I-NP I-ORG I-LOC

. . O O O

1. Các vấn đề và thách thức hiện nay.

Không chỉ NER nói riêng mà NLP nói chung hiện nay còn rất nhiều khó khăn và thách thức. Một số khó khăn chung cho NER được kể đến như:

* Sự nhập nhằng về từ vựng: Một từ có thể có nhiều nghĩa, phụ thuộc vào ngữ cảnh mà từ đó được sử dụng, gây khó khăn từ những giai đoạn đầu khi định nghĩa từ ngữ.
* Sự nhập nhằng về ngữ pháp: Sự thay đổi vị trí các từ trong câu là một ví dụ về nhập nhằng ngữ pháp. Ví dụ như câu “Ông già đi nhanh quá”, từng từ “ông” “già” “đi” đều có thể có những nghĩa khác nhau và dẫn đến cách hiểu khác nhau.
* Việc tiền xử lí dữ liệu còn kém: Việc trích xuất dữ liệu từ các nguồn dữ liệu như báo, tạp chí, … thường gặp nhiều lỗi, gây ra sai sót từ nguồn dữ liệu, dẫn đến sai sót trong các bước tiếp theo.
* Tập dữ liệu chưa cân bằng: đây là một khó khăn vô cùng lớn của lĩnh vực NLP. Trong các trường hợp, thường chúng ta chỉ sử dụng một vài nhãn cần thiết trên tập dữ liệu của mình, như PER, LOC, ORG. Và các thành phần khác trong cùng một ngữ cảnh sẽ được gán nhãn tương đương cho các lần xuất hiện tiếp theo. Điều nay gây ra một lượng sai số không nhỏ khi tần số xuất hiện các từ dùng để thay thế các từ khác như “này” “kia” “nó”.
* Có quá nhiều thực thể có tên trong các lĩnh vực khác nhau, với một số lượng không hề nhỏ. Trong các tập dữ liệu khác nhau giữa các lĩnh vực khác nhau, các thực thể có tên lại yêu cầu một phương thức khác nhau để trích xuất.

Đối với ngôn ngữ tiếng Việt, bên cạnh những khó khăn chung của lĩnh vực NLP, còn có những đặc thù riêng của tiếng việt trong lĩnh vực này nói chung và bài toán NER nói riêng, kể đến như:

* Sự thiếu hụt nguồn dữ liệu: Với tiếng Anh, có rất nhiều các tập dữ liệu cho NLP nói chung với quy mô trải rộng các ngành khác nhau. Nhưng trong tiếng Việt chỉ có 2 tập dữ liệu đó là VLSP 2016 và VLSP 2018 đặc tả riêng cho tiếng Việt. Sự thiếu hụt này dẫn đến các nhà nghiên cứu trong các lĩnh vực khác của NLP phải tự xây dựng tập dữ liệu riêng, điều này tiêu tốn nhiều thời gian và công sức.
* Thiếu hụt về phân định ngữ nghĩa: sự đa dạng ngữ nghĩa giữa cách sử dụng các từ ở nhiều vùng miền khác nhau. Từ “viết” có thể hiểu là động từ chỉ việc trình bày nội dung lên văn bản, cũng có thể hiểu là danh từ chỉ cây bút.
* Sự kết hợp phức tạp của từ ghép: tiếng Việt là ngôn ngữ đơn lập nên từ vựng chủ yếu là các từ ghép lại với nhau, tạo ra sự đa dạng trong cách sử dụng đồng thời cũng là khó khăn trong các bài toán NLP.

# **Chương 2: Các hướng tiếp cận và giải quyết bài toán NER.**

1. Phương pháp sử dụng tập luật và từ điển.

Phương pháp dựa trên từ điển tiến hành so khớp để nhận diện các thực thể có tên dựa trên một từ điển được cung cấp sẵn (Nadeau et al. 2006 [5]). Ưu điểm của nó là không phức tạp. Tuy nhiên, phương pháp này đòi hỏi phải có sẵn một từ điển các thực thể với chất lượng tốt và số lượng đầy đủ. Những dữ liệu có sẵn trên Internet với số lượng lớn có thể được sử dụng để tự động lọc ra các ứng cử viên nhằm hỗ trợ cho công việc xây dựng từ điển. Ngoài ra, phương pháp này không thể giải quyết được trường hợp nhập nhằng giữa các loại thực thể khác nhau. Phương pháp này rất có khả năng nhận dạng “nhầm” ‘Đức’ từ thực thể người thành thực thể địa danh trong “*Vì sao hai cổ phiếu của bầu Đức hết thời 'bước chung đôi'?*”

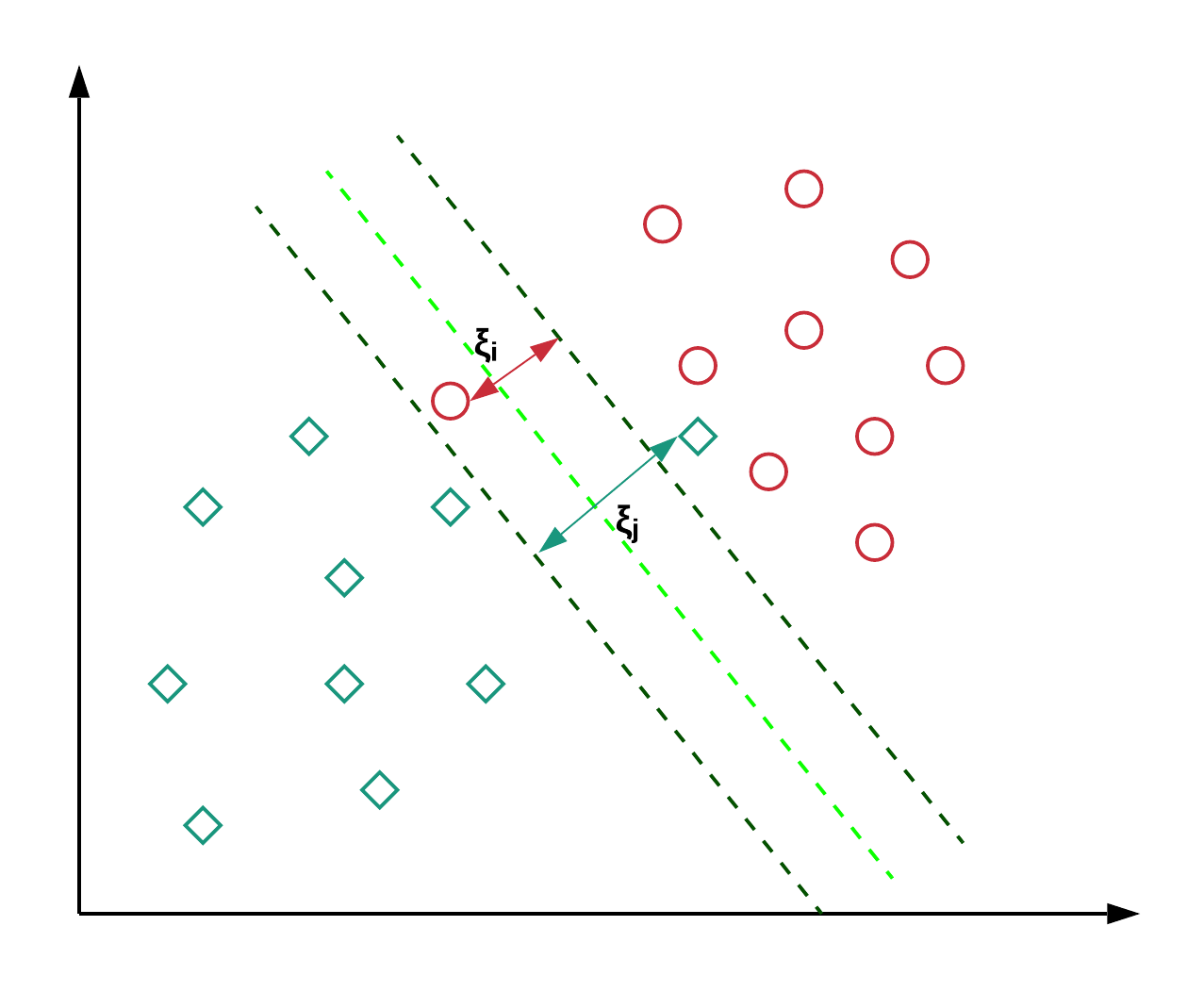
Phương pháp dựa trên tập luật tuy không đòi hỏi từ điển nhưng cần phải xây dựng các luật một cách thủ công (Soderland 1999 [6]). Ví dụ, luật “nước + NNP” (‘nước’ và tiếp theo là một danh từ riêng) sẽ nhận dạng được các địa danh đất nước xuất hiện trong văn bản, hay “tỉnh + NNP” (‘tỉnh’ và tiếp theo là một danh từ riêng) sẽ nhận dạng được các địa đanh hành chính cấp tỉnh trong văn bản. Các luật này có ưu điểm cho độ chính xác cao nhưng độ phủ không cao do tính cứng nhắc của các luật ngôn ngữ. Bên cạnh đó, việc xây dựng các luật này đòi hỏi nhiều công sức và hiểu biết cả về mặt lĩnh vực và cả về mặt ngôn ngữ.

1. Phương pháp sử dụng học máy.

Có nhiều phương pháp học máy có thể sử dụng trong bài toán NER được kể đến như: học có giám sát, học không giám sát hay học nửa giám sát. Các phương pháp học máy cho chất lượng rất tốt về độ chính xác, tuy nhiên nhược điểm của các phương pháp pháp này là phải có dữ liệu huấn luyện do con người gán nhãn. Các phương pháp học máy dựa trên đặc trưng yêu cầu xây dựng các tập đặc trưng phù hợp với bài toán (Nadeau & Sekine 2007 [7]). Việc xây dựng các tập đặc trưng này thường mất thời gian và nhiều khi yêu cầu các đặc trưng về mặt ngôn ngữ được cung cấp từ các bài toán khác như phân tích từ loại (POS tagging) hay phân cụm (chunking).

Tuy nhiên trong phạm vi bài báo cáo em xin được phép đề cập tới các mô hình học máy có giám sát chính. Các mô hình này dựa trên tập dữ liệu đã được gán nhãn chính xác, từ đó “học” được cách gán nhãn các từ loại mới.

* Support Vector Machine (SVM) [8]: mô hình vector hỗ trợ máy được xây dựng dựa trên ý tưởng chia các dữ liệu được gán nhãn thành các điểm, từ đó ánh xạ thành các vùng phân chia dữ liệu tương ứng với các nhãn sao cho khoảng cách giữa các vùng là xa nhất có thể. Các dữ liệu cần được gán nhãn mới sẽ được ánh xạ thành các điểm và từ đó lựa chọn nhãn tương ứng với vùng mà điểm đó rơi vào.



* Hidden Markov Model (HMM) [9]: mô hình Markov ẩn là một mô hình rất đơn giản được ứng dụng vào bài toán gán nhãn, với tập dữ liệu huấn luyện đã được gán nhãn sẵn. HMM tính toán xác suất của chuỗi đầu vào x tương ứng với chuỗi nhãn y, dựa trên các trạng thái trước đó.

A picture containing screen, building, drawing

Description automatically generated

* Maximum Entropy Markov Model (MEMM) [10]: Entropy cực đại là một kỹ thuật dùng để ước lượng xác suất các phân phối từ dữ liệu. Tư tưởng chủ đạo của nguyên lý Entropy cực đại là “mô hình phân phối đối với mỗi tập dữ liệu và tập các ràng buộc đi cùng phải đạt được độ cân bằng / đều nhất có thể”. Tập dữ liệu học (tức là tập gồm các dữ liệu đã được gán nhãn) được sử dụng để tìm ra các ràng buộc cho mô hình, đó là cơ sở để ước lượng phân phối cho từng lớp cụ thể. Những ràng buộc này được thể hiện bởi các giá trị ước lượng được của các đặc trưng. Từ các ràng buộc sinh ra bởi tập dữ liệu này, mô hình sẽ tiến hành tính toán để có được một phân phối cho Entropy cực đại.

A picture containing screen, building, drawing

Description automatically generated

* Conditional Random Fields (CRFs) [11]: trường điều kiện ngẫu nhiên là một mô hình học máy phổ biến cho các bài toán phân loại. CRFs kết hợp điểm mạnh của MEMM với các đặc trưng dữ liệu không phụ thuộc vào ngữ cảnh mà không tạo ra sai số thiên kiến. Điểm khác nhau chính giữa CRFs và HMM/MEMM đó là CRFs là đồ thị vô hướng trong khi HMM và MEMM là đồ thị có hướng. Hơn nữa, mô hình CRFs còn sử dụng đặc trưng từ các nhãn đã được gán và các nhãn sẽ được gán để cải thiện độ chính xác.

A close up of a screen

Description automatically generated

1. Phương pháp sử dụng học sâu.

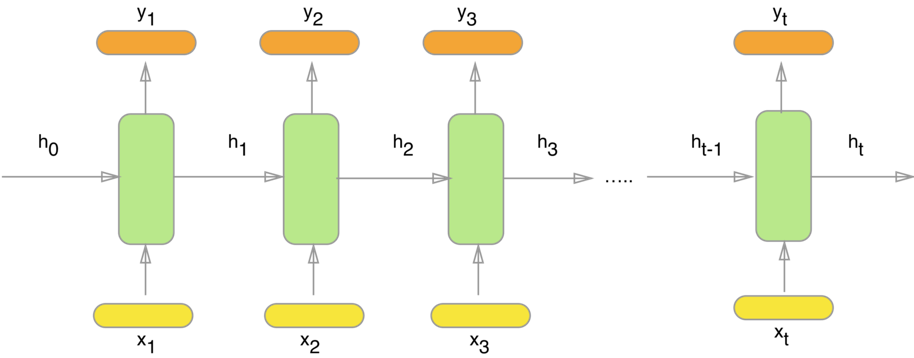
Học sâu là phương pháp kết hợp nhiều lớp học máy thành một mạng thần kinh nhân tạo, xây dựng dựa trên mô phỏng sinh học của não bộ. Các nơ-ron sẽ tập trung vào dữ liệu đầu vào và dữ liệu đầu ra, từ đó xây dựng một mạng để nhận diện các dữ liệu đầu vào và đầu ra có sự liên quan đến nhau. Cách tiếp cận đầu-cuối dựa trên các mô hình mạng nơ-ron có ưu điểm là công việc xây dựng tập đặc trưng được thực hiện tự động thông qua các tham số của mạng nơ-ron (Lample et al. 2016 [12]). Đồng thời, cách tiếp cận đầu-cuối cũng khuyến khích việc học các bài toán liên quan đến nhau một cách kết hợp (ví dụ học đồng thời bài toán nhận diện thực thể có tên với các bài toán trước như phân tích từ loại, phân cụm, hay các bài toán sau như nhận diện quan hệ). Nhược điểm chính của phương pháp dựa trên học sâu là tính diễn giải thấp dẫn đến phải sử dụng nhiều heuristics và kinh nghiệm khi lựa chọn siêu tham số của mô hình và việc học mô hình. Ngoài ra, các mô hình mạng nơ-ron cũng đòi hỏi hiệu năng tính toán cao khi huấn luyện.

* Convolutional Neural Network (CNN) [13]: Mạng tích chập là một mô hình học sâu được sử dụng trong nhiều lĩnh vực, sử dụng phép tích chập ma trận để trích xuất thông tin từ dữ liệu. CNN sử dụng phép tích chập trong mỗi lớp giữa các nút mạng kết hợp với một số hàm kích hoạt như Sigmoid, tanh, ReLU … Trong lĩnh vực NLP, mạng CNN được dùng để trích chọn thông tin từ văn bản đã cho bằng cách biểu diễu văn bản đã cho dưới dạng một ma trận, sau đó tích chập để trích chọn ra thông tin. Mỗi từ được nhúng thành 1 vector với số chiều cố định, từ đó kết hợp thành ma trận đầu vào của CNN.

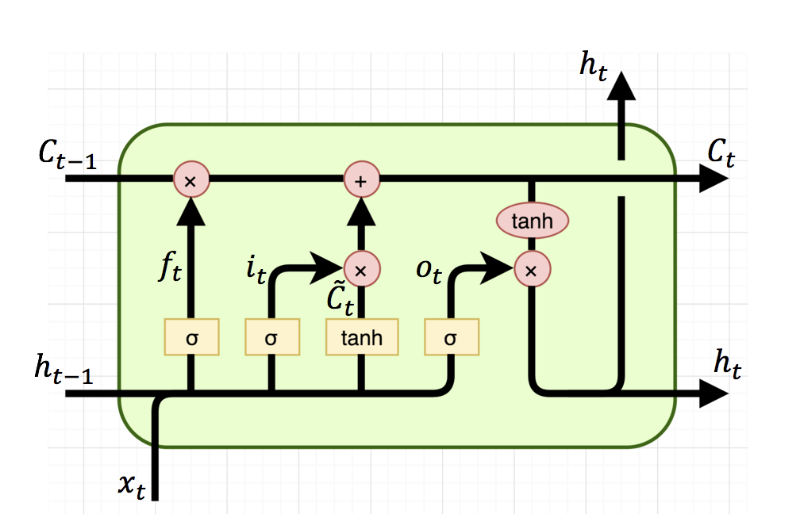
A picture containing drawing

Description automatically generated

* Recurrent Neural Network (RNN) [14]: Mạng hồi quy khắc phục điểm yếu chính của mạng CNN đó là ngữ cảnh. Trong NLP thì ngữ cảnh của một văn bản đóng một vai trò quan trọng khi xử lí các thông tin liên quan. Mạng RNN là một mô hình hồi quy, tại mỗi nơ-ron trong mạng thì ngoài giá trị đầu vào, còn có thêm giá trị từ các nơ-ron trước đó. Và điều này khiến RNN có thể lưu trữ thông tin của các nơ-ron trước, giữ lại được thông tin từ ngữ cảnh cụ thể của từng văn bản.



* Long-short term memory(LSTM) [15]: Mạng RNN được coi là mô hình dành cho NLP, nhưng trong thực tế, rất khó để có thể áp dụng mạng RNN vào để huấn luyện trong NLP bởi vì các trọng số khi qua nhiều lớp mạng bị nhỏ dần dẫn tới không thể thực hiện tính toán. Mô hình LSTM là sự cải tiến của mạng RNN với các cổng quên, loại bỏ các dư thừa dữ liệu không cần thiết trong chuỗi cần gán nhãn bằng các hàm phi tuyến như Sigmoid, tanh, ReLU. Ý tưởng của mạng LSTM là loại bỏ các dữ liệu không cần thiết khi thực hiện các thuật toán lan truyền xuôi và ngược bằng các cổng này.



# **Chương 3: Các giải pháp sử dụng phương pháp học sâu hiện nay.**

Như đã trình bày ở phần tổng quan tình hình trong và ngoài nước, bài toán nhận diện thực thể có tên có thể thực hiện theo ba phương pháp chính: phương pháp dựa trên tập luật và từ điển, phương pháp dựa trên các mô hình học máy và phương pháp dựa trên các mô hình học sâu. Cả ba phương pháp này đều có các ưu nhược điểm riêng như đã đề cập ở trên. Trong khi đó, nhu cầu thực tiễn của đề tài vừa yêu cầu độ chính xác cao của các quan hệ được trích rút, vừa đòi hỏi các quan hệ phải phủ được nhiều miền dữ liệu.

Ở Việt Nam hiện nay, đối với bài toán nhận diện các thực thể có tên nhìn chung đang dành được nhiều sự quan tâm từ cộng đồng nghiên cứu về tiếng Việt do nhu cầu ứng dụng vào các lĩnh vực xử lý nội dung như xử lý tin tức, thương mại điện tử. Trong khi đó, bài toán phát hiện mối liên hệ giữa các thực thể chỉ mới dành được một vài sự quan tâm từ cộng đồng.

Hiện nay, với phương pháp sử dụng mô hình kết hợp BiLSTM + CNN + CRFs đã đem lại một kết quả cho bài toán NER cho tiếng Anh với độ chính xác 91.21% [16]. Còn đối với tiếng việt, theo như bài báo của Anh-Duong Nguyen 2019 [17], với việc sử dụng mô hình BiLSTM + BiLSTM + CRFs đã đem lại độ chính xác cho bài toán NER trong tiếng Việt lên tới 94.88%. Phương pháp học sâu sử dụng mô hình lai này không quan tâm đến các đặc trưng mà chỉ cần cung cấp đầu vào là biểu diễn liên tục của các từ trong câu. Các biểu diễn liên tục này được học trước thông qua các mô hình biểu diễn từ như word2vec, GloVe trên một tập văn bản lớn trên tiếng Việt. Theo như công bố của Anh-Duong Nguyen 2019, kết quả của việc sử dụng mô hình kết hợp nói trên đạt độ chính xác F1 Score lên tới 94.88%, dựa trên tập dữ liệu VLSP2016.

|  |  |  |
| --- | --- | --- |
| **Dữ liệu** | **#sent** | **#word** |
| Train | 14,861 | 325,686 |
| Dev | 2,000 | 43,706 |
| Test | 2,831 | 66,097 |

# **Chương 4: Thực nghiệm mô hình CNN + LTSM + CRFs**

1. Mô hình thực nghiệm.

Mô hình được sử dụng để thực nghiệm trong báo cáo này, em xin chọn mô hình học sâu sử dụng BiLSTM + CNN + CRFs. Mô hình này đã được sử dụng cho bài toán NER trong tiếng Anh và đạt kết quả 91.21% F1 Score.

Tập dữ liệu sử dụng cho mô hình là tập dữ liệu VLSP 2016 đã được đề cập ở trên.

Chi tiết mô hình:

- Mỗi từ sẽ được phân tách thành các ký tự, sau đó sử dụng mạng CNN để trích chọn ra các thông tin về hình thái học của từng từ thành vector biểu diễn ký tự.

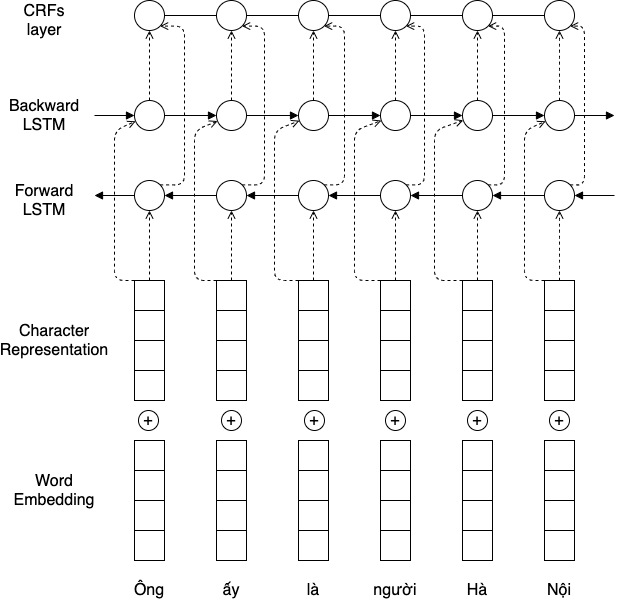
- Các từ sẽ được embedding thành các vector biểu diễn từ, sử dụng tập dữ liệu embedding GloVe.

- Sau đó hai vector biểu diễn này sẽ được kết hợp thành đầu vào cho mô hình mạng BiLSTM để trích chọn ra các thông tin về ngữ cảnh.

- Cuối cùng là tầng CRFs để gán nhãn toàn bộ câu thay vì gán nhãn từng từ.

A close up of text on a white surface

Description automatically generated



Kết quả thu được:

Dựa trên phương pháp chấm điểm F-Score, mô hình thu được kết quả khả quan, từ đó cho thấy đây là một mô hình tốt trong việc nhận diện các thực thể có tên.

|  |  |  |
| --- | --- | --- |
| **Mô hình huấn luyện** | **F1** | **Language** |
| CNN + BiLSTM + CRFs | 88.40 | English |
| CNN + BiLSTM + CRFs | 86.17 | Vietnamese |
| CNN + BiLSTM + CRFs + POS + Chunk | 91.29 | Vietnamese |

So với các mô hình đã công bố trước đó, mô hình huấn luyện này cần được cải tiến thêm để đạt được độ chính xác tốt hơn.

Một số mô hình đã công bố và độ chính xác:

|  |  |  |
| --- | --- | --- |
| **Mô hình đã công bố** | **F1** | **Language** |
| BiLSTM [16] | 87.00 | English |
| BiLSTM + CNN [16] | 89.21 | English |
| BiLSTM + CNN + CRFs [16] | 91.21 | English |
| BiLSTM + BiLSTM + CRFs [17] | 94.88 | Vietnamese |

**Tài liệu tham khảo**

[1]: Grishman, Ralph, and Beth Sundheim. "Design of the MUC-6 evaluation." Proceedings of the 6th conference on Message understanding. Association for Computational Linguistics, 1995.

[2]: Sundheim, Beth M. Overview of results of the MUC-6 evaluation. NAVAL COMMAND CONTROL AND OCEAN SURVEILLANCE CENTER SAN DIEGO CA, 1995.

[3]: Nguyen, Huyen TM, et al. "Vlsp shared task: Named entity recognition." Journal of Computer Science and Cybernetics 34.4 (2018): 283-294.

[4]: E. F. T. K. Sang and F. D. Meulder, “Introduction to the conll-2003 shared task: Language-independent named entity recognition,” in Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003. [Online]. Available: http://www.aclweb.org/anthology/W03-0419

[5]: Nadeau, David, Peter D. Turney, and Stan Matwin. "Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity." Conference of the Canadian society for computational studies of intelligence. Springer, Berlin, Heidelberg, 2006.

[6]: Soderland, Stephen. "Learning information extraction rules for semi-structured and free text." Machine learning 34.1-3 (1999): 233-272.

[7]: Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30.1 (2007): 3-26

[8]: Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their applications 13.4 (1998): 18-28.

[9]: Fine, Shai, Yoram Singer, and Naftali Tishby. "The hierarchical hidden Markov model: Analysis and applications." Machine learning 32.1 (1998): 41-62.

[10]: McCallum, Andrew, Dayne Freitag, and Fernando CN Pereira. "Maximum Entropy Markov Models for Information Extraction and Segmentation." Icml. Vol. 17. No. 2000. 2000.

[11]: Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

[12]: Lample, Guillaume, et al. "Neural architectures for named entity recognition." arXiv preprint arXiv:1603.01360 (2016).

[13]: Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. "A convolutional neural network for modelling sentences." arXiv preprint arXiv:1404.2188 (2014).

[14]: Li, Lishuang, et al. "Biomedical named entity recognition based on extended recurrent neural networks." 2015 IEEE International Conference on bioinformatics and biomedicine (BIBM). IEEE, 2015.

[15]: Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[16]: Ma, Xuezhe, and Eduard Hovy. "End-to-end sequence labeling via bi-directional lstm-cnns-crf." arXiv preprint arXiv:1603.01354 (2016).

[17]: Nguyen, Anh-Duong, Kiem-Hieu Nguyen, and Van-Vi Ngo. "Neural sequence labeling for Vietnamese POS tagging and NER." 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF). IEEE, 2019.