ỨNG DỤNG CRF NHẬN DẠNG THỰC THỂ ĐỊNH DANH TRONG VĂN BẢN TIẾNG VIỆT

APPLICATIONS OF CRF FOR NAMED ENTITY RECOGNITION IN VIETNAMESE DOCUMENTS

Võ Trung Hùng[1], Lâm Tùng Giang[1], Trần Thị Liên[2]

[1]Đại học Đà Nẵng; Email: vthung@dut.ud.vn, gianglt@gmail.com
[2]Học viên Cao học tại Đại học Đà Nẵng; Email: lientranha@gmail.com

Tóm tắt - Nhận dạng các thực thể định danh là một lĩnh vực đang nhận được sự quan tâm rộng rãi của các nhà nghiên cứu. Đã có nhiều kết quả nghiên cứu trong lĩnh vực này ở một số ngôn ngữ như Anh, Ý, Trung Quốc,… nhưng với Tiếng Việt thì còn hạn chế. Mục đích nghiên cứu này là xây dựng một hệ thống nhận dạng thực thể cho phép nhận dạng các thực thể có tên trong văn bản Tiếng Việt như tên người, địa điểm, tổ chức, thời gian,… được phát triển dựa trên công cụ CRF++. Nhiệm vụ chính của bài báo là xây dựng một tập dữ liệu tốt, đầy đủ, chính xác nhằm hỗ trợ cho việc nhận dạng thực thể và xây dựng một hệ thống huấn luyện, kiểm thử và ứng dụng. Hệ thống nhận dạng thực thể ban đầu đã thu thập 300 bài báo với nhiều lĩnh vực khác nhau và hoạt động có tính khả thi với độ đo F1 trung bình qua 10 lần thực nghiệm đạt 84,8%.

Từ khóa - nhận dạng thực thể có tên; mô hình CRF; công cụ CRF++; tên các thực thể trong tiếng Việt; hệ thống nhận dạng thực thể.

Abstract - Named Entity Recognition, a subfield of Information Extraction, is gaining wide attention from researchers in the field. There have been relevant researches published in English, Italian or Chinese, but not many works have been conducted in Vietnamese. The purpose of this study is to build a named entity recognition system that enables the identification of named entities, such as names of people, locations, organizations, or time, in Vietnamese texts by using the CRF + + tool. This paper mainly aims at creating the tools and training data for building a named entity recognition model to facilitate the identification of entities in Vietnamese documents. The Entity Recognition system was evaluated 10 times on over 300 empirical articles and then showed the average F1 measure of 84,8%.

Key words - named entity recognition; CRF model; CRF++ toolkit; names of entities in Vietnamese text; entity recognition system.

1. Giới thiệu

Nhận dạng thực thể định danh (Named Entity Recognition-NER) [1] là một nhiệm vụ con của lĩnh vực trích chọn thông tin (Information Extraction - IE). Mục đích của nó là nhận dạng và phân loại các thực thể trong văn bản cho các đối tượng xác định trước như tên người, tổ chức, địa điểm, thời gian,… Nhận dạng thực thể định danh được ứng dụng trong nhiều lĩnh vực xử lý ngôn ngữ tự nhiên như hệ thống đặt câu hỏi trả lời, hệ thống dịch máy, truy vấn thông tin. Hiện tại, việc nhận dạng đối với tiếng Anh đã có độ chính xác cao do có nguồn dữ liệu tra cứu, cú pháp rõ ràng [2], nhưng đối với tiếng Việt vẫn còn là một thách thức. Bài báo này trình bày tổng quan về công việc nhận dạng thực thể định danh trong văn bản tiếng Việt và sử dụng mô hình CRF (Condition Random Field), cụ thể là công cụ CRF++ phiên bản 0.58 1, để nhận dạng thực thể.

Nội dung bài báo được tổ chức như sau: phần 2 trình bày các nghiên cứu tổng quan về nhận dạng thực thể và mô hình CRF, phần 3 giới thiệu giải pháp đề xuất về hệ thống nhận dạng, phần 4 đánh giá kết quả và xác định hướng nghiên cứu trong tương lai.

2. Nghiên cứu tổng quan

2.1. Nhận dạng thực thể

2.1.1. Trích chọn thông tin

Trích chọn thông tin là tên gọi cho các kỹ thuật trích chọn các thông tin có cấu trúc từ văn bản không có cấu trúc và kết xuất ra những thông tin đã được định nghĩa trước về các thực thể và mối quan hệ giữa chúng từ văn bản [3]. Một số mức độ trích chọn thông tin từ văn bản bao gồm trích chọn các thực thể (Entity Extraction), trích chọn quan hệ giữa các thực thể (Relation Extraction), xác định đồng tham chiếu (Co-reference Resolution). Phạm vi trích chọn không chỉ trong phạm vi các từ trong văn bản mà có thể là âm thanh, hình ảnh,... Các kỹ thuật sử dụng trong trích chọn thông tin gồm: phân đoạn, phân lớp, kết hợp và phân cụm [4].

2.1.2. Bài toán nhận dạng thực thể

Thông thường, mỗi văn bản đều chứa các đối tượng như tên người, tổ chức, địa điểm, ngày, số,... Những đối tượng đó được gọi chung là các thực thể định danh. Mục đích của bài toán nhận dạng thực thể là nhận biết các loại thực thể này để giúp chúng ta trong việc hiểu văn bản. Đây là bài toán cơ bản nhất phải xét đến trước khi giải quyết các bài toán phức tạp hơn trong trích chọn thông tin.

2.2. Các hướng tiếp cận bài toán nhận dạng thực thể

2.2.1. Tiếp cận dựa trên tri thức

Hướng tiếp cận dựa trên tri thức (còn gọi là thủ công) có đặc điểm là hệ thống luật được xây dựng bằng tay hoàn toàn phụ thuộc vào kinh nghiệm riêng của chuyên gia trong từng lĩnh vực [5]. Các luật luôn luôn phát sinh và nó được cập nhật liên tục và đưa vào kho dữ liệu dưới sự kiểm duyệt và sửa chữa chặt chẽ của chuyên gia nhằm có được một hệ thống nhận dạng thực thể hoàn chỉnh. Ví dụ điển hình là hệ thống nhận biết loại thực thể Proteous của đại học New York tham gia hội thảo MUC-6 [6] được hỗ trợ bởi một số lượng lớn các luật.

Để xây dựng một hệ thống như mô hình trên yêu cầu chuyên gia phải có kinh nghiệm về ngôn ngữ học và một quỹ thời gian tương đối lớn để thực hiện việc liên tục cập nhật các luật mới phát sinh.

2.2.2. Tiếp cận dựa trên học máy

Với các hạn chế của hướng tiếp cận tri thức thì vấn đề đặt ra phải xây dựng được hệ thống có thể "tự học" để hệ

[1] http://crfpp.googlecode.com/svn/trunk/doc/index.html

thống trở nên linh hoạt hơn. Có một số phương pháp học máy được sử dụng rộng rãi và hiệu quả như mô hình HMM, MEMM và CRF.

Mô hình HMM (Hidden Markov Model) [7] được giới thiệu và nghiên cứu vào cuối năm 1960 và đầu năm 1970. Đây là mô hình máy trạng thái hữu hạn với các tham số biểu diễn xác suất chuyển trạng thái và xác suất sinh dữ liệu quan sát tại mỗi trạng thái, là mô hình thống kê trong đó hệ thống mô hình hóa được cho là một quá trình Markov với các tham số không biết trước và nhiệm vụ là xác định các tham số ẩn từ các tham số quan sát được dựa trên sự thừa nhận này. Quá trình sinh ra chuỗi dữ liệu quan sát trong HMM thông qua một loạt các bước chuyển trạng thái, xuất phát từ một trong các trạng thái bắt đầu và dừng lại ở một trạng thái kết thúc. Các tham số của mô hình được rút ra sau đó có thể sử dụng để thực hiện các phân tích kế tiếp. Với bài toán nhận dạng thực thể, có thể xem mỗi trạng thái tương ứng một trong các nhãn `B-LOC, I-LOC, B-TIME, B-PER,…` và dữ liệu quan sát là các từ trong câu. Khi đó có thể tìm được chuỗi các trạng thái mô tả tốt nhất cho chuỗi dữ liệu quan sát được bằng cách tính.

$$P(S|O) = P(S,O)/P(O) \quad (1)$$

Trong (1), `S` là chuỗi trạng thái ẩn, `O` là chuỗi dữ liệu quan sát đã biết. Việc tìm chuỗi `S*` với xác suất `P(S|O)` đạt giá trị cực đại tương đương với việc tìm `S*` làm cực đại `P(S,O)`.

Hạn chế của mô hình Markov nằm ở việc để tính được xác suất `P(S,O)` thông thường ta phải liệt kê hết các trường hợp có thể của chuỗi `S` và chuỗi `O`. Thực tế thì chuỗi `Y` là hữu hạn có thể liệt kê được, còn `O` (các dữ liệu quan sát) là rất phong phú. Bên cạnh đó, với một số bài toán thì việc sử dụng xác suất điều kiện `P(S|O)` cho kết quả tốt hơn.

Mô hình MEMM (Maximum Entropy Markov Models) [8] cho rằng các quan sát đã được cho trước và chúng ta không cần quan tâm đến xác suất sinh ra chúng. Điều cần quan tâm ở đây là các xác suất chuyển trạng thái. Đối với mô hình này thì quan sát hiện tại không tồn tại độc lập mà gắn liền với quá trình chuyển trạng thái, nghĩa là nó còn phụ thuộc vào trạng thái trước đó.

Xác suất `P(S|O)` có thể tính như sau:

$$P(S|O) = P(S_1, O_1) * \prod_{t=1}^{n} P(S_t|S_{t-1}, O) \quad (2)$$

MEMM coi dữ liệu quan sát là điều kiện cho trước thay vì coi chúng như các thành phần được sinh ra bởi mô hình như HMM, vì thế xác suất chuyển trạng thái có thể phụ thuộc vào thuộc tính đa dạng của chuỗi dữ liệu quan sát. Những thuộc tính này giữ vai trò quan trọng trong việc xác định trạng thái kế tiếp.

Mô hình CRF (Conditional Random Fields) [9] được giới thiệu lần đầu vào năm 2001. Đây là một mô hình xác suất thực hiện việc gán nhãn và phân đoạn dữ liệu tuần tự.

CRF được xem như một đồ thị vô hướng có điều kiện, `X` là biến ngẫu nhiên nhận giá trị, là chuỗi dữ liệu cần gán. `Y` là biến ngẫu nhiên nhận giá trị, là chuỗi nhãn tương ứng. Trong bài toán nhận dạng thực thể, `X` có thể nhận giá trị là các từ trong văn bản, `Y` là một chuỗi ngẫu nhiên các nhãn tên thực thể (<LOCATION>, <ORGANIZE>,…).

Gọi `G=(V,E)` là đồ thị vô hướng không có chu trình và có các đỉnh `v ∈ V` tương ứng với mỗi biến ngẫu nhiên đại diện cho Y_v của `Y`. Nếu mỗi biến ngẫu nhiên Y_v tuân theo tính chất Markov với đồ thị `G` thì `(Y,X)` là trường ngẫu nhiên điều kiện CRF.

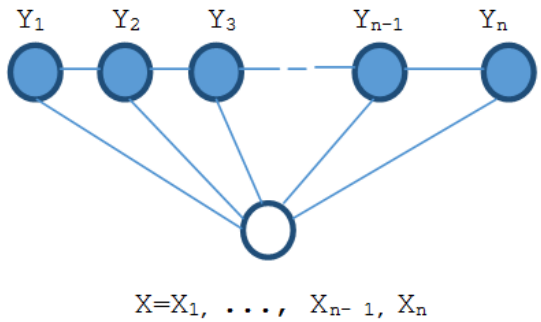
$$P(Y_v|X, Y_w, \omega \neq v) = P(Y_v|X, Y_w, w \in N(v)) \quad (3)$$

Trong đó, `N(v)` là tập hợp các đỉnh láng giềng của `v`.

Trong trường hợp đơn giản nhưng cũng rất quan trọng, khi mô hình hóa các chuỗi tuần tự (sequence), đồ thị `G` được biểu diễn dưới dạng:

$$G = (V = \{1,2,\dots,n\}, E = \{(i, i+1)\} \quad (4)$$

và có thể được minh họa ở hình sau:



Hình 1. *Đồ thị vô hướng mô tả CRF*

Áp dụng [10] cho các trường hợp ngẫu nhiên Markov thì phân phối của chuỗi nhãn Y với chuỗi quan sát X cho trước có dạng:

$$p(y|x) \propto \exp\left(\sum_{e\in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{e\in E,k} \mu_k g_k(v, y|_v, x)\right) \quad (5)$$

trong đó, x là chuỗi quan sát, y là chuỗi trạng thái, $y|_S$ là tập hợp các phần tử của y tương ứng với các đỉnh của đồ thị con S; f_k và g_k là các hàm thuộc tính được tự định nghĩa, λ_k và μ_k là các tham số.

2.3. Công cụ CRF++ Toolkit

Được phát triển trên nền tảng mô hình CRF, CRF ++ là một công cụ mã nguồn mở viết bằng ngôn ngữ C++ và có thể phục vụ cho việc phân đoạn, gán nhãn dữ liệu tuần tự.. Phiên bản 0.58 (CRF++-0.58), chạy trên hệ điều hành Windows được sử dụng trong bài báo này bao gồm các công cụ phục vụ huấn luyện và kiểm thử.

Trong giai đoạn huấn luyện, một tập tin huấn luyện có định dạng riêng của CRF++ được tạo lập và sử dụng. Với mỗi từ trong chuỗi văn bản, các thẻ được xác định, chứa bản thân từ, một số thuộc tính và nhãn được gán. Mỗi thẻ sẽ nằm trên một dòng của tập tin huấn luyện. Các thuộc tính tại vị trí `i` trong chuỗi văn bản quan sát gồm hai phần: thông tin ngữ cảnh tại vị trí `i` và thông tin về nhãn. Lựa chọn thuộc tính là việc chọn ra các mẫu ngữ cảnh thể hiện thông tin cần quan tâm tại vị trí bất kỳ trong chuỗi dữ liệu quan sát. Có thể sử dụng các mẫu ngữ cảnh về đặc điểm

của từ như viết hoa, viết thường, có phải chữ số, dấu câu; sử dụng mẫu ngữ cảnh dạng biểu thức chính quy (ví dụ áp dụng để xác định biểu thức thời gian); sử dụng ngữ cảnh từ điển cho phép tra cứu các từ trong một số danh sách cho trước.

Bên cạnh tập tin huấn luyện, một tập tin mẫu (template) được sử dụng, xác định cách thức quan sát trong quá trình huấn luyện và kiểm tra. Mỗi một dòng trong tập tin mẫu này chỉ ra một mẫu dùng để định nghĩa dữ liệu đầu vào.

Kết quả của quá trình huấn luyện là một tập tin mô hình. Tập tin này được sử dụng để phục vụ việc kiểm thử hoặc trong các ứng dụng. Tập tin kiểm thử gần giống với tập tin huấn luyện, chứa các thẻ. Tại tập tin kiểm thử, nhãn có thể được gán thủ công nhằm mục đích đánh giá mô hình.

3. Xây dựng hệ thống nhận dạng thực thể định danh trong văn bản tiếng Việt

Hiện nay, đã có một số hệ thống nhận dạng thực thể định danh trong văn bản tiếng Việt được xây dựng như "Hệ thống nhận dạng thực thể trong văn bản tiếng Việt sử dụng mô hình CRF" của tác giả Nguyễn Cẩm Tú [4], "Hệ thống nhận dạng thực thể trong văn bản tiếng Việt phát triển trên mã nguồn mở Gate" của tác giả Nguyễn Bá Đạt [11],… Tuy nhiên, các hệ thống này chỉ công bố các mô hình sử dụng và kết quả thu được của hệ thống, không thể hiện rõ các công cụ cũng như các bước cụ thể để xây dựng một hệ thống.

Trong bài báo này, hệ thống nhận dạng tên riêng trong các văn bản tiếng Việt được xây dựng, bao gồm 2 thành phần: hệ thống huấn luyện và ứng dụng nhận dạng thực thể. Các mô-đun phần mềm được viết bằng ngôn ngữ Java.

3.1. Huấn luyện

Trong hệ thống huấn luyện, chúng tôi sử dụng bộ công cụ CRF++ và tạo lập các dữ liệu phục vụ huấn luyện bao gồm các bước sau:

- Đầu tiên cơ sở dữ liệu từ điển được xây dựng, bao gồm các tập tin văn bản chứa từ điển họ người, địa điểm, các từ đứng trước tên người, tổ chức, thời gian.

- Để tạo lập bộ dữ liệu huấn luyện, đầu tiên các bài báo được thu thập thủ công và lưu vào các tập tin văn bản. Chúng tôi sử dụng công cụ vnTagger 4.22 để gán nhãn từ loại cho văn bản và cho kết quả là một tập tin chứa các từ khóa.

- Việc xác định thuộc tính cho các từ trong văn bản được thực hiện bằng các mô-đun phần mềm. Trên mỗi dòng, cột đầu tiên là bản thân từ, cột tiếp theo là nhãn từ loại. Tiếp theo, chúng tôi tạo lập các cột thuộc tính `Is_Cap` (chữ hoa), `Is_Num` (chữ số), `Is_Mark` (dấu câu), `Is_Num` (số), `Is_4_Digit` (4 số), `Is_Date` (giá trị ngày tháng), `Is_Family` (họ người), `Is_Location` (địa điểm), `Is_BeforePER` (từ trước tên người), `Is_BeforeORG` (từ trước tên tổ chức), `Is_BeforeTime` (từ trước thời gian) ở các cột tiếp theo.

- Thực hiện việc dán nhãn thủ công tại cột cuối cùng: các nhãn được định nghĩa trong hệ thống được trình bày trong **Bảng 1**.

Với các từ đa âm tiết (multi-syllable), các tiền tố được sử dụng để xác định vị trí của âm (sylabble) trong từ (B: bắt đầu, I: bên trong và O: kết thúc từ). Ví dụ từ "Thừa Thiên Huế" sẽ tương ứng với 3 thẻ sau:

```
Thừa    B-LOC
Thiên   I-LOC
Huế     E-LOC
```

- Do công cụ CRF++ không hỗ trợ tốt cho bảng mã tiếng Việt, tập tin văn bản kết quả chứa các cột thuộc tính thẻ và nhãn được chuyển đổi sang dạng tiếng Việt mã hóa Telex (ví dụ chữ Việt được mã hóa thành Vieejt). Kết quả, tập tin *train.data* được tạo lập để sử dụng với công cụ huấn luyện *crf_learn.exe* để tạo lập tập tin mô hình `model.data`.

Bảng 1.

Nhãn	Ý nghĩa
LOC	Tên địa danh
PER	Tên người
ORG	Tên tổ chức
NUM	Số
CUR	Tiền tệ
TIME	Thời gian
PCT	Phần trăm
MISC	Các thực thể khác
O	Không phải thực thể

3.2. Mở rộng dữ liệu huấn luyện

Sau khi đã tạo lập mô hình nhận dạng thực thể đầu tiên, dữ liệu thử nghiệm được thu thập, bao gồm 300 bài báo từ các website tin tức *http://vnexpress.net* và *http://vietnamnet.vn;* xác định thuộc tính tự động và chuyển sang dạng mã Telex tương tự như dữ liệu thử nghiệm để tạo tập tin `test.data`, tuy nhiên bước dán nhãn thủ công không được thực hiện. Thay vào đó, chúng tôi sử dụng công cụ kiểm thử *crf_test.exe* của CRF++ để gán nhãn tự động vào cột cuối cùng. Tiếp theo tập tin này được kiểm tra thủ công và chỉnh sửa lỗi để đảm bảo chính xác. Dữ liệu tập tin `test.data` sau đó được bổ sung vào `train.data` để lặp lại quá trình huấn luyện. Quá trình thử nghiệm - bổ sung dữ liệu huấn luyện này được thực hiện lặp lại một số lần nhằm làm tăng độ tin cậy của mô hình.

3.3. Kiểm thử

Để đánh giá hiệu suất của hệ thống nhận dạng thực thể 3 thông số độ chính xác (precision), độ hồi tưởng (recall) và F1 (f-measure) được sử dụng.

Độ chính xác đo bằng tỉ lệ phần trăm số thực thể được gán nhãn chính xác (giá trị t_1) trên tổng số tên thực thể được gán nhãn (giá trị t_2).

$$\text{Độ chính xác} = \frac{t_1}{t_2} \quad (6)$$

Độ hồi tưởng đo bằng tỉ lệ phần trăm số thực thể được gán nhãn chính xác (giá trị t_1) trên tổng số thực thể được gán nhãn của công cụ CRF++ trong tập `test.data` (giá trị t_3)

[2] https://github.com/stnguyen/vnTagger

$$\text{Độ hồi tưởng} = \frac{t_1}{t_3} \quad (7)$$

F1 là đại lượng được tính bởi sự kết hợp giữa độ chính xác và độ hồi tưởng theo công thức sau:

$$F_1 = \frac{2 * \text{Độ chính xác} * \text{Độ hồi tưởng}}{\text{Độ chính xác} + \text{Độ hồi tưởng}} \quad (8)$$

Hệ thống thực nghiệm sử dụng phương pháp "10-fold cross validation". Dữ liệu được chia thành 10 phần bằng nhau, lần lượt lấy 9 phần để huấn luyện và một phần còn lại để kiểm tra, kết quả sau 10 lần thực nghiệm được ghi lại và đánh giá tổng thể được trình bày tại **Bảng 2**.

Bảng 2.

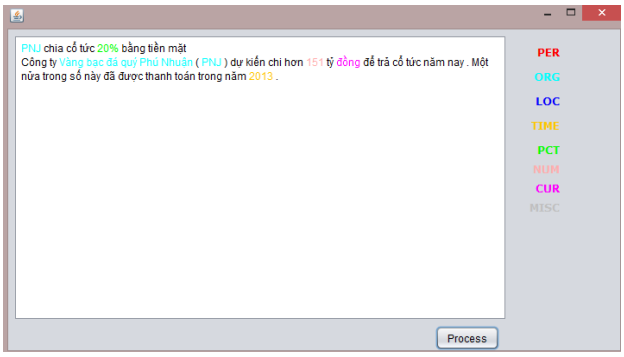
Lần thử nghiệm	Độ chính xác	Độ hồi tưởng	F1
1	71.90%	88.82%	79.47%
2	83.27%	88.31%	85.71%
3	83.48%	93.03%	88.00%
4	81.23%	87.50%	84.25%
5	85.83%	84.20%	85.01%
6	82.59%	94.53%	88.16%
7	79.69%	87.93%	83.61%
8	77.72%	84.03%	80.75%
9	82.08%	93.11%	87.25%
10	82.87%	88.85%	85.76%
Trung bình	81.07%	89.03%	84.80%

Bên cạnh đó, kết quả thử nghiệm cũng được xem xét cho từng loại nhãn với kết quả tại **Bảng 3**.

Bảng 3.

Tên thực thể	Độ chính xác	Độ hồi tưởng	F_1
CUR	81.25%	81.25%	81.25%
LOC	59.09%	100.00%	74.29%
NUM	100.00%	99.08%	99.54%
ORG	52.94%	75.00%	62.07%
PCT	100.00%	91.30%	95.45%
PER	92.00%	92.00%	92.00%
TIME	67.44%	100.00%	80.56%

3.4. Xây dựng ứng dụng



Hình 2. *Ứng dụng nhận dạng thực thể*

Trên cơ sở mô hình đã được xây dựng và kiểm thử, qua 10 lần thực nghiệm và chọn ra mô hình tốt nhất trong 10 lần này, một ứng dụng được xây dựng, áp dụng mô hình CRF để nhận dạng các thực thể trong văn bản tiếng Việt. Với đầu vào là một tập tin văn bản, ứng dụng phân tích nội dung văn bản, nhận dạng các thực thể định danh trong văn bản và thay đổi màu sắc cho các cụm từ tương ứng với các nhãn khác nhau. Ví dụ: những thực thể được nhận dạng có nhãn B-PER, I-PER thì đổi màu sắc thành màu đỏ, nhãn là B-LOC, I-LOC đổi màu sắc thành màu xanh,... Kết quả được trình bày tại **Hình 2**.

4. Kết luận

4.1. Kết quả đạt được

Kết quả chính được trình bày trong bài báo là một hệ thống ứng dụng mã nguồn mở, cho phép huấn luyện mô hình nhận dạng thực thể định danh dựa trên mô hình CRF. Hệ thống này bao gồm các mô-đun huấn luyện, kiểm thử và ứng dụng nhận dạng thực thể định danh trong văn bản tiếng Việt. Độ đo F1 của hệ thống đạt giá trị 84,8% trên tập dữ liệu kiểm thử. Với quy trình được trình bày tại mục 3.1, hệ thống có thể tiếp nhận các dữ liệu huấn luyện tùy biến khác nhau (ví dụ thuộc các lĩnh vực khác nhau) tùy thuộc nhu cầu sử dụng nhằm tạo lập các mô hình phù hợp phục vụ việc nhận dạng thực thể định danh trong các văn bản tiếng Việt.

4.2. Hướng phát triển

Để tăng độ chính xác cho việc nhận dạng thực thể trong hệ thống thì nguồn dữ liệu huấn luyện cần phải lớn và chính xác. Chúng tôi sẽ tiếp tục khai thác và thu thập thêm nguồn dữ liệu mới và mở rộng các loại thực thể cần nhận dạng, bổ sung các luật mới nhằm tạo lập các thuộc tính hỗ trợ cho quá trình huấn luyện nhằm tăng độ chính xác của mô hình.

TÀI LIỆU THAM KHẢO

[1] Nancy Chinchor and Patty Robinson, MUC-7 Named Entity Task Definition, *Proc. Sixth Messag. Underst. Conf. MUC6*, p. 21, 1997.

[2] Alireza Mansouri, Lilly Suriani Affendey, and Ali Mamat, Named Entity Recognition Approaches, *J. Comput. Sci.*, vol. 8, pp. 339–344, 2008.

[3] Sunita Sarawagi, *Information Extraction*, vol. 1, no. 3, pp. 261–377, 2008.

[4] Nguyễn Cẩm Tú, *Nhận biết các loại thực thể trong văn bản tiếng Việt nhằm hỗ trợ Web ngữ nghĩa và tìm kiếm hướng thực thể*, Luận văn tốt nghiệp ĐHCN, 2005.

[5] Nguyễn Thị Loan, *Tìm hiểu mô hình CRF và ứng dụng trong trích chọn thông tin trong tiếng Việt*, Luận văn tốt nghiệp ĐHCN, 2005.

[6] Douglas E. Appelt, Jerry R. Hobbs, John Bear, and David Israel, SRI International FASTUS system MUC-6 test results and analysis, in *MUC-6, NIST*, 1995.

[7] Phil Blunsom, Hidden Markov Models, *Lect. notes*, 2004.

[8] A. McCallum, D. Freitag, and F. Pereia, Maximum entropy markov models for information extraction and segmentation, in *International Conference on Machine Learning*, 2000.

[9] John Lafferty, Andrew Mccallum, and FCN Fernando C. N. Pereira, Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data, in *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, vol. 2001, pp. 282–289.

[10] John M. Hammersley and Peter Clifford, *Markov fields on finite graphs and lattices*, 1971.

[11] Nguyễn Bá Đạt, *Nhận dạng thực thể trong văn bản tiếng Việt*, Luận văn tốt nghiệp ĐHCN-ĐH Quốc gia Hà Nội, 2009.

(BBT nhận bài: 23/06/2014, phản biện xong: 22/07/2014)