Tạp chí Tin học và Điều khiển học, T.28, S.2 (2012), 115–128

TRÍCH RÚT THÔNG TIN TỰ ĐỘNG TỪ VĂN BẢN TIẾNG VIỆT*

SAM CHANRATHANY[1], LÊ THANH HƯƠNG[1], NGUYỄN THANH THỦY[2], NGUYỄN HỮU THIỆN[1]

[1] *Viện Công nghệ Thông tin và Truyền thông, Trường Đại học Bách khoa Hà Nội*
[2] *Trường Đại Công nghệ, Đại học Quốc gia Hà Nội*

Tóm tắt. Bài báo đề xuất các hướng tiếp cận học bán giám sát trong việc xây dựng hệ thống trích rút thông tin tự động từ văn bản tiếng Việt. Với trích rút thực thể, mở rộng phương pháp của Liao [7] bằng cách sử dụng các luật đồng tham chiếu về tên và các luật nhóm 2 để tìm các thực thể mới. Thử nghiệm cho thấy, hệ thống đề xuất có độ chính xác cao hơn hệ thống của Liao [7]. Với trích rút mối quan hệ cải tiến hàm nhân mức nông SLK của Giuliano [6] bằng cách bổ sung thêm các đặc trưng cho việc biểu diễn câu bao gồm từ loại, loại thực thể, từ điển động từ và thay đổi kích cỡ cửa số của hàm nhân. Kết quả thử nghiệm cho thấy phương pháp học có giám sát sử dụng SLK cải tiến tốt hơn phương pháp học có giám sát sử dụng SLK của Giuliano [6]. Và khi áp dụng phương pháp học bán giám sát, hệ thống thu được kết quả tốt hơn học có giám sát.

Abstract. This paper presents semi-supervised approaches to construct a Vietnamese information extraction system. Our approach in named entity extraction inherits the idea of Liao [7] and extends it by using proper name coreference rules to find new entities. The new entities are put into the training set to learn new context features for the extracting module. The experimental results show that our method achieves higher accuracy than Liao's [7]. In relation extraction, we improve the Shallow Linguistic Kernel (SLK) of Giuliano et al.'s [6] by modifying the window size of the kernel and using additional features to present sentences, including part of speech, another entity types, and a dictionary of compound verbs. Our experimental results also show that the supervised method using our SLK achieves higher accuracy than one used by Giuliano et al. [6]. Moreover, its accuracy when applying the semi-supervised method is higher than that when using the supervised one.

1. MỞ ĐẦU

Trích rút thông tin (Information Extraction - IE) là quá trình tự động trích rút các thông tin có cấu trúc như thực thể (ví dụ, *tên người*, *tên địa điểm*, *tên tổ chức*) và mối quan hệ giữa các thực thể (ví dụ, quan hệ *sống ở* giữa *tên người* và *tên địa điểm*) từ dữ liệu phi cấu trúc.Trích rút thông tin mở rộng khả năng tìm kiếm thông tin trên dữ liệu phi cấu trúc so với cách tìm kiếm dựa trên từ khóa truyền thống. Ngoài ra, trích rút thông tin còn có nhiều

*Nghiên cứu này được hoàn thành dưới sự hỗ trợ từ Quỹ phát triển khoa học và Công nghệ quốc gia (NAFOSTED) mã số 102.01-2011.08 và Đề tài khoa học và công nghệ cấp Bộ mã số B2012-01-24

ứng dụng rộng rãi và hữu ích khác, như lấy thông tin về tên của các công ty, tên người điều hành công ty, theo dõi thông tin về các dịch bệnh, theo dõi các sự kiện khủng bố,... Trong hơn một thập niên qua, đã có nhiều nghiên cứu về trích rút thực thể [2, 11, 12, 14] và trích rút mối quan hệ giữa các thực thể [6, 13]. Phần lớn các nghiên cứu thường tập trung vào cách tiếp cận học có giám sát. Khó khăn trong học có giám sát là cần một tập dữ liệu đã được gán nhãn với kích thước lớn để phục vụ cho việc huấn luyện mô hình trích rút. Việc xây dựng tập dữ liệu huấn luyện lớn như vậy đòi hỏi phải đầu tư nhiều thời gian và công sức. Đối với tiếng Việt, chưa có tập dữ liệu đã được gán nhãn với kích thước lớn như vậy. Để giải quyết vấn đề này, cách tiếp cận học máy bán giám sát đã được đề xuất trong những năm gần đây [7, 13]. Ý tưởng cơ bản của phương pháp học máy bán giám sát là: bắt đầu huấn luyện mô hình với một tập dữ liệu đã gán nhãn có kích cỡ nhỏ; sau đó sử dụng mô hình vừa học, kết hợp với các tri thức chuyên gia bên ngoài để tự động sinh ra các dữ liệu gán nhãn mới từ tập dữ liệu chưa gán nhãn cho trước, và bổ sung những dữ liệu gán nhãn mới này vào tập dữ liệu đã gán nhãn. Phương pháp này chỉ đòi hỏi một tập dữ liệu huấn luyện ban đầu nhỏ để định hướng cho quá trình trích rút, đồng thời tận dụng được các tri thức chuyên gia sẵn có, cũng như tập dữ liệu chưa gán nhãn phong phú bên ngoài, để nâng cao hiệu năng trích rút thực thể và các mối quan hệ giữa chúng.

Đối với hướng tiếp cận học máy, nhiều kỹ thuật đã được áp dụng cho bài toán trích rút thông tin như mô hình trường ngẫu nhiên có điều kiện (Conditional Random Fields – CRF) [7], máy vectơ hỗ trợ (Support Vector Machine – SVM) [12], mô hình markov ẩn (Hidden Markov Model –HMM) [14], mô hình markov entropy cực đại (Maximum Entropy Markov Model- MEMM) [2],.... Bản chất của trích rút thực thể là gán nhãn các từ, cụm từ trong văn bản với loại thực thể tương ứng (như *tên người*, *tên tổ chức*). Vì vậy, có thể coi bài toán trích rút thực thể là bài toán phân loại dữ liệu, tức là phân loại mỗi từ thành kiểu thực thể mà nó thuộc vào. SVM là phương pháp phân loại dữ liệu, nên được coi là một giải pháp cho bài toán này. Vấn đề khó khăn là việc gán nhãn kiểu thực thể cho một từ phụ thuộc vào nhãn của các từ xung quanh nó. Ví dụ, từ *"phát triển"* trong cụm từ *"công ty phát triển phần mềm FPT"* có từ bên trái và từ bên phải được gán nhãn là tên tổ chức nên từ *"phát triển"* cũng được gán nhãn là tên tổ chức. Hạn chế của SVM là không giải quyết được vấn đề phụ thuộc nhãn giữa các từ. Vì vậy ta cần đến một mô hình khác có thể giải quyết được vấn đề này, đó là mô hình CRF. CRF cho phép ta tích hợp nhiều đặc trưng của bản thân từ, cũng nhưng các từ xung quanh của nó, để làm cơ sở cho việc xây dựng mô hình nên thích hợp hơn cho bài toán này. SVM phù hợp với bài toán trích rút mối quan hệ giữa các thực thể hơn do bài toán trích rút mối quan hệ giữa các thực thể không gán nhãn cho chuỗi từ mà chỉ quan tâm đến xác định mối quan hệ giữa các thực thể, cụ thể là đi xác định xem một câu có thuộc mối quan hệ đang xét hay không. Do SVM và CRF đều là các phương pháp học có giám sát ta sẽ sử dụng các phương pháp này dưới dạng học bán giám sát, kết hợp với kỹ thuật khác như kỹ thuật bootstrapping, do các tác giả trong [1] đề xuất.

Với tiếng Việt, các phần mềm tiền xử lý văn bản như tách câu, tách từ, phân tích từ loại đã đạt được độ chính xác khá cao (>93%). Hiện nay đã có một số nghiên cứu về trích rút thực thể cho tiếng Việt nhưng đều sử dụng phương pháp học máy có giám sát [11, 12]. Hiện chưa có nghiên cứu nào về trích rút mối quan hệ giữa các thực thể trong văn bản tiếng Việt. Vì vậy, bài báo này đề xuất phương pháp học bán giám sát cho bài toán trích rút thực thể và trích rút mối quan hệ giữa các thực thể trong văn bản tiếng Việt, tập trung vào trích rút mối quan hệ giữa các thực thể trong cùng một câu.

2. ĐẶC ĐIỂM TIẾNG VIỆT

2.1. Đặc điểm của tiếng Việt ảnh hưởng đến nhận dạng thực thể

Qua nghiên cứu các văn bản tiếng Việt cho thấy một thực thể có thể xuất hiện nhiều lần trong một văn bản dưới các hình thức khác nhau thông qua hiện tượng đồng tham chiếu. Như vậy, các tên này có thể có cùng một kiểu thực thể do chúng cùng tham chiếu đến một thực thể chung duy nhất. Trên cơ sở tập luật đồng tham chiếu về tên của các tác giả trong [3], Nguyễn và Cao [10] đã xây dựng tập 11 luật đồng tham chiếu nhằm giải quyết nhập nhằng thực thể trong văn bản tiếng Việt. Áp dụng tập luật này cho việc phát hiện thực thể và gọi các luật này là luật nhóm 1.

Bên cạnh các hiện tượng đồng tham chiếu nói trên, qua nghiên cứu đặc điểm của tên người, tên địa điểm, tên tổ chức trong tiếng Việt, một tập luật nhận dạng thực thể được đề xuất như sau (gọi là các luật nhóm 2)(xem trong Bảng 1).

Bảng 1. Các luật nhận dạng thực thể (nhóm 2)

Luật	**Định nghĩa**
1	Nếu NP có tiền tố thuộc vào một trong bốn loại *tên người*, *tên địa điểm*, *tên tổ chức* thì gán kiểu thực thể tương ứng với tiền tố của chúng.
2	Nếu cụm từ đang xét nằm trong 1 trong 4 từ điển về tên người Việt Nam, lĩnh vực nghiên cứu, địa điểm, tổ chức, cụm từ đó được gán nhãn dựa trên từ điển tương ứng.
Các luật sau xử lý các NP thỏa mãn 2 điều kiện sau: + Không có tiền tố thuộc vào một trong ba loại từ điển tiền tố *tên người*, *tên địa điểm*, *tên tổ chức*. + Chỉ chứa duy nhất một từ và tất cả các chữ của từ này đều được viết hoa chữ cái đầu, ví dụ: *Hà Lan*, *Bồ Đào Nha*	
3	Nếu sau một NP có 1 chuỗi các từ dưới dạng sau: [**trợ từ**][**từ dùng để định nghĩa**][**số từ**] [**từ thuộc một trong ba từ điển**] thì NP này sẽ được gán nhãn dựa trên từ điển tương ứng với **từ thuộc một trong ba từ điển**. Với trợ từ như: đã, đang, vẫn, ...; từ định nghĩa như: là, làm, chỉ,...; số từ như: các, một, ...Từ định nghĩa là bắt buộc, còn hai loại từ kia có thể có hoặc không. Chẳng hạn như trong ví dụ dưới đây, *"Andrew Grove"* là tên người, còn *"Hồ Chí Minh"* là tên địa điểm. (a) **Andrew Grove** là một giám đốc công ty. (b) **Hồ Chí Minh** là con đường huyền thoại.
4	Nếu đứng trước NP là 1 từ thuộc một trong hai loại: động từ thường đi kèm với từ chỉ nơi chốn (*đến*, *đi*, ...) hoặc trạng từ chỉ nơi chốn (*tại*, *ở*, ...) thì NP sẽ được gán nhãn là *tên địa điểm*.
5	Nếu một NP đứng trước một chuỗi có dạng: [**dấu câu định nghĩa**][**số từ**] [**từ thuộc một trong ba từ điển**] Các dấu câu định nghĩa gồm có ",", "-", "(" .... thì NP đó sẽ được gán nhãn dựa trên từ điển tương ứng với **từ thuộc một trong ba từ điển**. Ví dụ:**Vinamilk**, công ty sữa lớn nhất Việt Nam, được thành lập năm 1976. Trong ví dụ này, *"Vinamilk"* là tên một tổ chức.
6	Nếu một NP đứng trước một chuỗi có dạng: [**số từ**][**từ thuộc một trong ba từ điển**][**từ bổ sung ý nghĩa cho từ thuộc một trong ba từ điển**][**dấu hai chấm ":" hoặc các từ liệt kê**] Trong đó: **Từ liệt kê** thường là *như*, *gồm*, *gồm có* ...; **Từ bổ sung ý nghĩa** thường là tính từ. Khi đó NP và tất cả các từ theo sau NP này sẽ được gán nhãn dựa trên từ điển tương ứng với **từ thuộc một trong ba từ điển**. (các từ theo sau NP chỉ được chứa các chữ được viết hoa chữ đầu và phải tạo thành một dãy từ liên tiếp mà NP là từ đầu tiên). Ví dụ: Các nước tiên tiến như: **Mỹ, Nhật, Pháp**,... đều quan tâm đến vấn đề này. Trong ví dụ trên, các từ *"Mỹ"*, *"Nhật"*, *"Pháp"* đều là *tên địa điểm* vì từ *"nước"* đi trước chúng thuộc từ điển tiền tố địa điểm.

2.2. Đặc điểm của tiếng Việt ảnh hưởng đến nhận dạng quan hệ

Trên cơ sở nghiên cứu các đặc điểm của ngôn ngữ tiếng Việt, cho thấy các mối quan hệ giữa các thực thể trong câu tiếng Việt có thể biểu diễn theo ba trường hợp chính dưới đây:

Trường hợp 1: Các từ giữa hai thực thể ứng cử viên (là các thực thể đang xét mối quan hệ) cho ta biết mối quan hệ giữa chúng. Mẫu này được gọi là "mẫu ngữ cảnh giữa" và có dạng : E_1 <*các từ*> E_2.

Ví dụ 2.1. Ông **Nguyễn Tất Đắc** <*đang làm việc cho*> **công ty FPT**.

Cụm từ *"làm việc cho"* trong ví dụ trên xác định mối quan hệ *con người-tổ chức* giữa các thực thể trong câu.

Trường hợp 2: Các từ trước và giữa hai thực thể ứng cứ viên xác định mối quan hệ giữa các thực thể. Mẫu này được gọi là "mẫu ngữ cảnh trước - giữa" và có dạng: <*các từ*> E_1 <*các từ*> E_2.

Ví dụ 2.2. Chúng tôi <*đã tổ chức đám cưới cho*> ông **Nam** <*và*> cô **Lê**.

Trong ví dụ này, sự kết hợp giữa cụm từ *"đã tổ chức đám cưới cho"* và từ *"và"* cho biết hai thực thể *con người-con người "**Nam**"* và *"**Lê**"* có mối quan hệ vợ chồng.

Trường hợp 3: Các từ giữa và sau hai thực thể ứng cứ viên xác định mối quan hệ giữa chúng. Mẫu này được gọi là "mẫu ngữ cảnh giữa - sau" và có dạng: E_1 <*các từ*> E_2 <*các từ*>.

Ví dụ 2.3. **Ngọc** <*và*> **Nam** <*đã ly hôn*>.

Trong câu này, sự kết hợp giữa cụm từ *"và"* và cụm từ *"đã ly hôn"* cho ta thấy hai thực thể *con người-con người* đã từng có mối quan hệ vợ chồng.
Ngoài các trường hợp trên, còn có một số trường hợp đặc biệt khác như:

Trường hợp 4: Hai thực thể liền kề có mẫu dạng: E_1 E_2 <*phần còn lại của câu*>.

Ví dụ 2.4. **Thủ tướng** (E_1) **Nguyễn Tấn Dũng** (E_2) đến thăm hỏi các bà con nông dân.

Trong ví dụ này, thực thể *chức vụ "**Thủ tướng**"* và thực thể *con người "**Nguyễn Tấn Dũng**"* kề nhau. Hai thực thể liên kết bởi quan hệ *chức vụ* (*tên người-chức vụ*).

Trường hợp 5: Hai thực thể phân cách nhau bằng một thực thể khác và có mẫu dạng: E_1 E_2 E_3 <*phần còn lại của câu*>.

Ví dụ 2.5. **Bộ trưởng** (E_1) **Bộ Giáo dục và Đào tạo** (E_2) **Nguyễn Thiện Nhân** (E_3) đến thăm ĐHBK.

Trong ví dụ này, hai thực thể *chức vụ "**Bộ trưởng**"* và *con người "**Nguyễn Thiện Nhân**"* có quan hệ *chức vụ*. Các thực thể này phân cách nhau bằng thực thể thứ ba có kiểu *tổ chức* là *"**Bộ Giáo dục và Đào tạo**"*.
Trường hợp 4 và trường hợp 5 là trường hợp đặc biệt của ba trường hợp trên, khi không có từ nào giữa hai thực thể ứng cứ viên hoặc các từ giữa hai thực thể ứng cứ viên thuộc kiểu thực thể khác.

Các kết luận dưới đây được rút ra từ các trường hợp trên:

* Các từ và cụm từ trước, giữa và sau hai thực thể ứng cứ viên, đặc biệt cụm động từ, là các yếu tố quan trọng trong việc xác định quan hệ giữa hai thực thể.

* Khi nhận dạng quan hệ giữa hai thực thể trong câu, thông tin về các thực thể khác trong câu cũng đóng vai trò quan trọng.

Phần dưới đây sẽ trình bày phương pháp giải bài toán trích rút thực thể và trích rút mối quan hệ giữa các thực thể sử dụng phương pháp học máy. Các hệ thống sử dụng học máy gồm hai quá trình: quá trình huấn luyện hệ thống và quá trình sử dụng mô hình sau khi học. Vì huấn luyện là vấn đề quan trọng nhất trong học máy, nên ở đây, ta chỉ tập trung trình bầy về quá trình huấn luyện cho hai bài toán nói trên.

3. TRÍCH RÚT THỰC THỂ

Hệ thống trích rút thực thể sử dụng phương pháp học máy bán giám sát được đề xuất kế thừa và mở rộng phương pháp trong [7]. Các tác giả trong [7] đã áp dụng mô hình ban đầu M để gán nhãn (trích rút thực thể) cho một tập dữ liệu chưa gán nhãn cho trước U, sau đó sử dụng các thực thể E trong U được nhận biết bởi M với độ tin cậy thấp, nhưng được nhận biết bởi các tri thức chuyên gia bên ngoài với độ tin cậy cao, để bổ sung vào tập dữ liệu huấn luyện L ban đầu. Ta gọi tập dữ liệu L khi đã được bổ sung thêm dữ liệu huấn luyện mới là L'. Với việc các thực thể E có độ tin cậy thấp trong M được bổ sung vào tập dữ liệu huấn luyện L, mô hình mới M' được học ra từ tập dữ liệu huấn luyện mới L' sẽ tốt hơn mô hình ban đầu M. Trong [7], các tri thức bên ngoài được sử dụng là 2 giả định ngôn ngữ sau đây: (i) những cụm từ (viết hoa) giống hệt nhau cùng xuất hiện trong một văn bản thường có kiểu thực thể giống nhau (giả định xuất hiện nhiều lần); (ii) các thực thể như con người, tổ chức, địa điểm thường có các tiền tố (ví dụ: Mr., CEO) hoặc hậu tố (ví dụ: Inc, Co) (giả định ngữ cảnh).

Kế thừa hai nguồn tri thức trên từ [7] và mở rộng chúng bằng cách áp dụng thêm các luật đồng tham chiếu về tên cũng như các luật thủ công được xây dựng dựa trên các kiến thức chuyên gia. Cụ thể, đối với giả định xuất hiện nhiều lần, ta thấy rằng không chỉ những cụm từ giống hệt nhau xuất hiện nhiều lần trong cùng một văn bản thì mới thường có kiểu thực thể giống nhau mà cả những cụm từ tham chiếu đến nhau thông qua các luật đồng tham chiếu về tên (xem Mục 2.1) và xuất hiện trong cùng một văn bản cũng thường có các kiểu thực thể giống nhau. Chẳng hạn, hai cụm từ *"Sài Gòn"* và *"TP Hồ Chí Minh"* là hai cụm từ cùng tham chiếu đến nhau theo luật 3 trong [10](một tên là bí danh của tên khác). Khi đó, trong cùng một văn bản, hai cụm từ này sẽ thường có cùng một kiểu thực thể, trong trường hợp này của ta là kiểu thực thể địa điểm. Do đó, nếu hai cụm từ này cùng xuất hiện trong một văn bản và bằng cách nào đó ta biết được rằng cụm từ *"TP Hồ Chí Minh"* là một tên địa điểm thì ta có thể suy ra rằng cụm từ "Sài Gòn" cũng là tên một địa điểm. Đối với giả định ngữ cảnh, qua nghiên cứu các văn bản tiếng Việt cho thấy, bên cạnh những thông tin ngữ cảnh về tiền tố đi ngay trước một tên có thể giúp nhận biết được kiểu thực thể, ta có thể có thêm những qui luật khác giúp nhận biết các thực thể một cách tương đối chính xác và đơn giản. Các luật này đã được trình bày trong Bảng 1. Từ các sự mở rộng đó, giải thuật học máy bán giám sát cho bài toán trích rút thực thể được giới thiệu trong Hình 1.

Hình 1. Giải thuật học bán giám sát dựa trên luật đồng tham chiếu về tên

Đầu vào:

L - tập dữ liệu được gán nhãn kích thước nhỏ

U - tập dữ liệu chưa gán nhãn kích thước lớn

T_1 - ngưỡng độ tin cậy cao; T_2 - ngưỡng độ tin cậy thấp

Đầu ra: Mô hình đã được huấn luyện M

Giải thuật:

Lặp lại K lần các bước từ 1 đến 4 sau đây:

Bước 1: Luyện mô hình M_k bằng phương pháp CRF trên tập dữ liệu L

Bước 2: Trích rút dữ liệu mới E dựa trên M_k

- Sử dụng M_k gán nhãn thực thể cho tập dữ liệu U và tính độ tin cậy của các thực thể được phát hiện
- Tìm các thực thể có độ tin cậy cao ($> T_1$) và tìm các đồng tham chiếu đến các thực thể đó
- Sử dụng các đồng tham chiếu để tìm kiếm dữ liệu đã gán nhãn mới trong U
- Hậu xử lý dữ liệu đã gán nhãn mới và thêm dữ liệu này vào tập dữ liệu mới E

Bước 3: Tìm các thực thể mới dựa trên các luật nhóm 2

- Áp dụng các luật nhóm 2 để tìm thêm các dữ liệu mới từ U
- Thêm các dữ liệu mới tìm được từ các luật nhóm 2 này vào tập dữ liệu mới E

Bước 4: Thêm tập dữ liệu gán nhãn mới E vào vào tập dữ liệu huấn luyện ban đầu L

Trong giải thuật học bán giám sát được đề xuất, ở bước 1, hệ thống thực hiện công việc tách từ và gán nhãn từ loại cho các văn bản trong tập dữ liệu đã gán nhãn L. Sau đó, hệ thống tính giá trị cho các đặc trưng sử dụng trong mô hình CRF và sử dụng các đặc trưng này để huấn luyện ra một mô hình gán nhãn M_k từ L. Các đặc trưng được sử dụng là các đặc trưng được áp dụng rộng rãi trong các hệ thống trích rút thực thể khác bao gồm từ, định dạng từ, từ loại và từ điển tên riêng (từ điển tên thành phố, từ điển tên người, từ điển tên tổ chức). Qua việc thử nghiệm với các kích thước cửa sổ khác nhau sử dụng trong tính các đặc trưng, ta chọn kích thước cho kết quả tốt nhất là 5. Ở bước 2, các văn bản chưa được gán nhãn ở U sẽ được đưa qua bước tiền xử lý và phân tích đặc trưng tương tự ở bước 1. Sau đó, các văn bản này được gán nhãn thực thể sử dụng mô hình M_k đã học được ở bước 1. Đồng thời trong quá trình gán nhãn đó, những thực thể nào được M_k phát hiện cũng sẽ được hệ thống tính một độ tin cậy tương ứng. Độ tin cậy này thể hiện mức độ tin tưởng của mô hình M_k khi gán nhãn cho thực thể tương ứng. Trong bài báo này, cũng giống như [7], sử dụng thuật toán tính độ tin cậy được đề xuất trong [4]. Tiếp đến, những thực thể được trích rút bởi mô hình M_k cũ có độ tin cậy cao ($> T_1$) sẽ được kết hợp với các luật đồng tham chiếu về tên được mô tả ở Mục 2.1 để tìm các thực thể ứng cử viên (có độ tin cậy thấp $< T_2$, hoặc chưa được gán nhãn) trong văn bản đang xét của U. Các thực thể ứng cử viên này sẽ có nhãn tương ứng với nhãn thực thể đồng tham chiếu với nó, và sau đó bổ sung vào tập dữ liệu huấn luyện L ban đầu để làm dữ liệu huấn luyện mới. Ở bước 3, để mở rộng thêm tập dữ liệu huấn luyện và tận dụng các văn bản chưa gán nhãn trong U triệt để hơn, chúng tôi áp dụng tập luật nhóm 2 (xem trong phần 2.1) để tìm thêm những tên thực thể mới mà không thể phát hiện được bởi mô hình cũ M_k cũng như phương pháp sử dụng các luật đồng tham chiếu ở trên. Ngoài ra, để

cân bằng giữa số mẫu luyện dương (thực thể quan tâm) như tên người, tên địa điểm,... và số mẫu luyện âm (không thuộc các thực thể quan tâm, có nhãn O), các thực thể được mô hình cũ gán nhãn là O với độ tin cậy cao và không trùng với những thực thể mới đã được trích rút ở trên cũng sẽ được sử dụng như những mẫu âm để thêm vào dữ liệu huấn luyện.

Các vấn đề đối với thuật toán và các cải tiến đề xuất

Một vấn đề cần đặt ra là những ứng cử viên thực thể có thể chưa hẳn đã là những tên thực thể thật sự. Hiện tượng này là do những sự nhập nhằng có thể có trong cách sử dụng ngôn ngữ. Qua nghiên cứu, ta thấy có hai hình thức nhập nhằng phổ biến sau:

Hậu xử lý 1: Một ứng cử viên thực thể có thể là một phần của tên một thực thể khác. Cụ thể hơn, một ứng cử viên thực thể có thể là phần giữa hoặc phần cuối của một thực thể khác. Trong trường hợp này, ta cần gán nhãn cho tên thực thể phủ phần văn bản lớn hơn chứ không phải cho ứng cử viên thực thể đang xét. Để giải quyết vấn đề này, cần tìm cụm NP nhỏ nhất chứa ứng cử viên thực thể và kiểm tra xem phần đầu của nó có nằm trong từ điển tiền tố *"tên người"*, *"tên tổ chức"* hoặc *"tên địa điểm"* hay không. Nếu đúng, cụm NP đó sẽ được gán nhãn theo từ điển tiền tố tương ứng và thay thế cho ứng cử viên thực thể của ta.

Ví dụ 3.1.

(*a*) Hôm nay, anh **Toàn Thắng** đi Sài Gòn.

(*b*) Hôm nay, **Công ty Toàn Thắng** sẽ mở cửa.

Giả sử trong câu (a), ta tìm thấy cụm từ *"Toàn Thắng"* là tên người có độ tin cậy cao. Sau đó trong câu (b) ta tìm cụm NP nhỏ nhất chứa ứng cử viên thực thể *"Toàn Thắng"* (chính là cụm từ *"Công ty Toàn Thắng"*) sau đó kiểm tra tiền tố của NP này. Ta thấy rằng tiền tố trong trường hợp này là *"Công ty"* và nó nằm trong từ điển tiền tố về tổ chức. Từ đó, ta sẽ thay ứng cử viên thực thể *"Toàn Thắng"* ban đầu của ta bằng cả cụm NP *"Công ty Toàn Thắng"* với kiểu tổ chức và đẩy cả cụm này vào tập dữ liệu huấn luyện thay vì đẩy ứng cử viên thực thể *"Toàn Thắng"* ban đầu.

Hậu xử lý 2: Việc nhận dạng tên thực thể phụ thuộc vào ngữ cảnh của nó:

Ví dụ 3.2.

(*a*) Hôm nay, **công ty FPT** đã tổ chức liên hoan cho các thành viên.

(*b*) Hôm nay, chúng tôi sẽ tổ chức liên hoan ở **công ty FPT**.

Trong câu đầu tiên, *"công ty FPT"* là *tên tổ chức*, trong khi nó lại là *tên địa điểm* trong ví dụ thứ hai. Ta giải quyết bằng cách, nếu một cụm từ được gán nhãn là *tên tổ chức* lại có trạng từ chỉ nơi chốn (trong, cạnh, ...) đi liền đằng trước, thì cụm từ này sẽ được gán lại nhãn là *tên địa điểm* trước khi được đẩy vào dữ liệu huấn luyện một cách thực sự.

4. TRÍCH RÚT QUAN HỆ GIỮA CÁC THỰC THỂ

Quá trình huấn luyện hệ thống trích rút mối quan hệ giữa các thực thể nhận đầu vào là văn bản có đánh dấu mối quan hệ đang xét là mối quan hệ *sống ở* (*tên người - tên địa điểm*), *chức vụ* (*tên người-chức vụ*), *làm việc cho* (*tên người- tên tổ chức*). Đầu ra là mô hình gán nhãn kiểu quan hệ. Hệ thống hướng tới việc trích rút mối quan hệ giữa hai thực thể cùng nằm

trong một câu. Thông tin về cấu trúc ngữ pháp của câu sẽ giúp ích rất nhiều cho việc trích rút này. Tuy nhiên, do các công cụ phân tích cú pháp tiếng Việt cho kết quả chưa cao nên ta sẽ cải tiến phương pháp hàm nhân (kernel) sử dụng ngôn ngữ mức nông trong [6] để biểu diễn dữ liệu và tính hàm nhân. Ưu điểm của phương pháp này là không cần đến công cụ phân tích cú pháp, chỉ cần phân tích từ loại, vẫn đảm bảo cho kết quả tương đối cao so với phương pháp khác. Hơn nữa, cách biểu diễn câu thành dạng ngữ cảnh toàn cục trong [6] phù hợp với dạng mối quan hệ trong câu của tiếng Việt (xem trong Mục 2.2).

4.1. Mô hình có giám sát dựa trên phương pháp ngôn ngữ mức nông

Như đã nói ở trên, ta sử dụng phương pháp hàm nhân và áp dụng phương pháp SVM để phân loại quan hệ. Hàm nhân trong hệ thống ở đây được định nghĩa như sau:

Hàm nhân K trên không gian đối tượng X là hàm K: $X \times X \to [0\ \infty]$, ánh xạ cặp đối tượng x, y thuộc X thành hàm đo mức độ giống nhau $K(x, y)$. Trong hệ thống này, hàm nhân được tạo ra bằng cách cải tiến hàm nhân dựa trên ngôn ngữ mức nông trong [6]. Hàm nhân ngôn ngữ mức nông là sự kết hợp giữa hàm nhân ngữ cảnh toàn cục với hàm nhân ngữ cảnh cục bộ.

Hàm nhân ngữ cảnh toàn cục

Như đã trình bày ở trong Mục 2.2, mối quan hệ trong văn bản tiếng Việt có thể được biểu diễn bằng ba trường hợp chính: mẫu giữa (E_1 <*các từ*> E_2), mẫu trước - giữa (<*các từ*> E_1<*các từ*> E_2), mẫu giữa - sau (E_1 <*các từ*> E_2 <*các từ*>).

Mục đích của hàm nhân ngữ cảnh toàn cục là thu được thông tin ngữ cảnh của toàn bộ câu chứa mối quan hệ. Để làm được việc này, mỗi mối quan hệ sẽ được biểu diễn dạng các mẫu ngữ cảnh tương ứng với ba mẫu đã nói ở trên:

Ví dụ 4.1. Năm 1988, **John** (E_1) có chức vụ là **tổng giám đốc** (E_2) của ngân hàng HSBC

Tương ứng với ví dụ trên, ta có ba mẫu ngữ cảnh toàn cục tương ứng là:

Mẫu ngữ cảnh giữa

Năm 1988, **John** (E_1) | có chức vụ là **tổng giám đốc** | (E_2) của ngân hàng HSBC

Mẫu ngữ cảnh giữa - sau

Năm 1988, | **John** (E_1) có chức vụ là **tổng giám đốc** (E_2) của ngân hàng HSBC |

Mẫu ngữ cảnh trước - giữa

| Năm 1988, **John** (E_1) có chức vụ là **tổng giám đốc** (E_2) | của ngân hàng HSBC

Dựa trên đặc tính của tiếng Việt như đã đề cập trong Mục 2.2, ta cải tiến hàm nhân ngữ cảnh toàn cục trong [6] bằng cách bổ sung thêm các đặc trưng vào trong quá trình học và dự đoán quan hệ trong câu để đo mức độ tương tự giữa các mẫu. Các đặc trưng này gồm có: từ, từ loại, loại thực thể, từ điển động từ (ví dụ từ điển động từ cho kiểu quan hệ *làm việc cho* bao gồm các từ: *làm việc ở*, *công tác tại* ...). Sử dụng $n - gram = 3$ để xét các tổ hợp từ có thể có trong ngữ cảnh. Ví dụ: câu *"Tôi đi học"* sẽ có các $n - gram = \{$*tôi*, *tôi đi*, *tôi đi học*$\}$. Cuối cùng hàm nhân ngữ cảnh toàn cục được định nghĩa là:

$$K_{GC}(R_1, R_2) = K_{FB}(R_1, R_2) + K_B(R_1, R_2) + K_{BA}(R_1, R_2),$$

trong đó, K_{FB}, K_B và K_{BA} lần lượt là các hàm nhân $n-gram$ thao tác trên các mẫu ngữ cảnh trước-giữa, giữa và giữa-sau.

Hàm nhân ngữ cảnh cục bộ

Ngữ cảnh cục bộ được sử dụng để phát hiện thực thể nào là thực thể đầu tiên (tác nhân), thực thể nào là thực thể thứ hai (đích) khi xét mối quan hệ. Để tìm thông tin này, 2 ngữ cảnh cục bộ xung quanh hai thực thể ứng cử viên được sử dụng và gọi là ngữ cảnh cục bộ bên trái và ngữ cảnh cục bộ bên phải. Ngữ cảnh cục bộ của thực thể bao gồm các đặc trưng: từ, từ loại, định dạng từ như chữ hoa, chữ thường, số, ký tự la mã. Trong quá trình xét ngữ cảnh cục bộ, cửa sổ ngữ cảnh của thực thể bên trái bắt đầu từ đầu câu đến từ đứng ngay trước thực thể thứ hai, còn cửa số ngữ cảnh thực thể bên phái được bắt đầu từ từ đứng ngay sau thực thể thứ nhất cho đến cuối câu. Đây cũng là một điểm khác biệt giữa hệ thống được nghiên cứu trong bài báo so với hệ thống trong [6]. Khi đó hàm nhân ngữ cảnh cục bộ được tính bằng:

$$K_{LC}(R_1, R_2) = K_{Right}(R_1, R_2) + K_{Left}(R_1, R_2),$$

trong đó, K_{LC} là hàm nhân ngữ cảnh cục bộ, K_{Right} là hàm nhân ngữ cảnh bên phải, K_{Left} là hàm nhân ngữ cảnh bên trái.

Hàm nhân ngữ cảnh tổng hợp K_{SL}

Hàm nhân ngữ cảnh tổng hợp là sự kết hợp của hàm nhân ngữ cảnh toàn cục với hàm nhân ngữ cảnh cục bộ như sau:

$$K_{SL}(R_1, R_2) = K_{GC}(R_1, R_2) + K_{LC}(R_1, R_2).$$

4.2. Mô hình học bán giám sát dựa trên phương pháp ngôn ngữ mức nông theo kiểu Bagging Bootstrapping

Đầu vào của hệ thống bao gồm tập dữ liệu nhỏ đã gán nhãn L và tập dữ liệu lớn chưa gán nhãn U. Trước tiên, từ bộ dữ liệu này, hệ thống đưa ra được dạng mẫu ngữ cảnh. Áp dụng phương pháp bagging bằng cách nhân bản tập dữ liệu L thành B tập. B tập dữ liệu này được sử dụng để huấn luyện B mô hình trích rút. Sau khi được huấn luyện, mỗi mô hình thực hiện gán nhãn cho dữ liệu chưa có nhãn U. Hệ thống sẽ tính mức độ thỏa thuận giữa các mô hình, để chọn S câu có mức độ thỏa thuận giữa các mô hình có giá trị cao nhất và gán nhãn tốt nhất (là nhãn được gán bởi nhiều mô hình nhất) cho dữ liệu chưa gán nhãn. S câu đó được đưa thêm vào tập dữ liệu huấn luyện để huấn luyện lại hệ thống. Quá trình này sẽ lặp đi lặp lại, cho tới khi không còn dữ liệu nào để thêm vào tập dữ liệu huấn luyện hoặc không đạt được ngưỡng mức độ thỏa thuận (quá trình bootstrapping).

Hình 2. Giải thuật học bán giám sát sử dụng phương pháp Bagging Bootstrapping trong học mối quan hệ

Đầu vào:

L - tập dữ liệu có kích cỡ nhỏ đã được gán nhãn, được biểu diễn dưới dạng mẫu ngữ cảnh

U - tập dữ liệu chưa gán nhãn, được biểu diễn dưới dạng mẫu ngữ cảnh

S - số mẫu mới đưa vào tập dữ liệu huấn luyện trong một lần lặp

B - số mẫu bootstrap được tạo từ L

Đầu ra: Mô hình phân loại quan hệ đã được huấn luyện

Giải thuật:

Repeat

1. *For* $i = 1$ to B do

- Tạo ra dữ liệu bootstrap L_i từ L sử dụng phương pháp Bagging
- Huấn luyện bộ phân loại quan hệ M_i trên L_i sử dụng phương pháp SVM
- Chạy bộ phân loại quan hệ M_i trên U

2. Tìm S mẫu trong U có mức độ thỏa thuận cao nhất giữa B bộ phân loại quan hệ và gán nhãn phù hợp nhất S mẫu đó

3. Thêm S vào L

Until không có dữ liệu mới nào phù hợp

Ở đây mức độ thỏa thuận tương ứng với mỗi mẫu (câu) được tính bằng hàm entropy sau (hàm entropy càng nhỏ thì mức độ thỏa thuận càng cao):

$$H = -\sum_{i}^{M} \frac{||r_i||}{B} log \frac{||r_i||}{B},$$

trong đó, M là số lớp nhãn, B là số tập dữ liệu được nhân bản hoặc số bộ phân loại quan hệ được tạo ra, $||r_i||$ là số bộ phân loại quan hệ gán nhãn r_i cho mẫu đang xét.

5. THỬ NGHIỆM

5.1. Tập dữ liệu

Tập dữ liệu thử nghiệm ở đây được thu thập thủ công từ các trang Web tiếng Việt bao gồm các trang web cá nhân và các trang tin tức (*vnexpress.net*, *dantri.com*, *wikipedia*) trong các lĩnh vực: thể thao, khoa học, văn hoa, giáo dục, kinh tế. Với bài toán trích rút thực thể, ta xây dựng tập dữ liệu gồm 950 văn bản, trung bình mỗi văn bản có 750 từ. Các văn bản này có chứa các thông tin về *tên người*, *tên tổ chức* và *tên địa điểm*. Với tập dữ liệu dùng trong trích rút mối quan hệ, ta lựa chọn những câu chứa ít nhất hai thực thể và các thực thể đó phải là *tên người*, *tên địa điểm*, *tên tổ chức* hoặc *chức vụ* từ các văn bản trên. Và quan hệ mà ta đang xét ở đây là *sống ở* (giữa *tên người* và *tên địa điểm*), *làm việc cho* (giữa *tên người* và tên tổ chức) và *chức vụ* (giữa *tên người* và *tên chức vụ*). Tổng số câu trong tập dữ liệu dùng trong trích rút mối quan hệ là 1200 câu. Trong quá trình gán nhãn thủ công, để làm cho việc gán nhãn chính xác hơn, mỗi văn bản được gán nhãn bởi hai người (một người gán nhãn, một người kiểm tra lại).

5.2. Phương pháp thử nghiệm

5.2.1. Bài toán trích rút thực thể

Kịch bản thử nghiệm như sau: 900 văn bản chưa gán nhãn sẽ được chia làm 9 phần, mỗi phần có 100 văn bản. 50 văn bản đã được gán nhãn sẽ được sử dụng để học ra một mô hình ban đầu sử dụng phương pháp CRF đã mô tả ở trên. Sau đó, hệ thống tiến hành 9 lần lặp. Ở mỗi lần lặp, 1 trong số 9 phần văn bản chưa gán nhãn trên được kết hợp với mô hình đã học được ở bước lặp trước để khai thác thêm dữ liệu huấn luyện mới. Ở đây, ta chọn ngưỡng độ tin cậy cao $T_1 = 0.95$ và ngưỡng độ tin cậy thấp $T_2 = 0.85$ dựa trên các kinh nghiệm nghiên cứu thực tế.

5.2.2. Bài toán trích rút quan hệ giữa các thực thể

Để đánh giá hiệu quả của thuật toán đề xuất, ta cài đặt cả hai phương pháp học có giám sát sử dụng hàm nhân trong [6] và hàm nhân vừa xây dựng. Trong thử nghiệm này, 240 câu đã gán nhãn được sử dụng làm dữ liệu huấn luyện và 960 câu chưa gán nhãn được sử dụng làm dữ liệu đánh giá. Với phương pháp bán giám sát, 240 câu có nhãn được sử dụng làm dữ liệu huấn luyện khởi tạo ban đầu L và 960 câu còn lại sử dụng làm dữ liệu chưa được gán nhãn U. 960 câu này cũng được sử dụng để làm dữ liệu đánh giá tổng thể. Số bộ dữ liệu nhân bản B là 5, số câu được đưa vào dữ liệu huấn luyện mỗi lần S là 100, ngưỡng là 3/5.

5.3. Kết quả

Kết quả của hệ thống sẽ được đánh giá thông qua 3 độ đo: độ chính xác P, độ phủ R và độ đo trung bình F, trong đó độ đo F được tính theo công thức:

$$F = \frac{}{2PR}P + R.$$

Sau đây, là kết quả thử nghiệm của hệ thống đề xuất đối với bài toán trích rút thực thể và trích rút mối quan hệ giữa các thực thể.

5.3.1. Bài toán trích rút thực thể

Từ Bảng 2 ta thấy, bắt đầu từ một mô hình với độ đo F cho các kiểu thực thể tên người, tên địa điểm, tên tổ chức lần lượt là 71.65%, 55.74% và 49.16%. Thông qua các lần lặp, hệ thống đã cải thiện được các độ đo F cho các kiểu thực thể này lên thành 93.13%, 88.15% và 79.35% tương ứng. Qua mỗi lần lặp, nhờ các dữ liệu huấn luyện mới có chất lượng được bổ sung tập dữ liệu huấn luyện, hiệu năng của hệ thống được cải thiện dần.

Để so sánh độ chính xác của hệ thống đề xuất với độ chính xác trong hệ thống trong [7], ta tiến hành thử nghiệm theo ba cách: (*i*) chỉ sử dụng 2 heuristic trong [7] (*ii*) sử dụng luật đồng tham chung về tên (luật nhóm 1), (*iii*) sử dụng luật nhóm 1 và luật nhóm 2 (Bảng 1)(kết quả trong Bảng 3).

Bảng 3 cho thấy độ đo F của các kiểu thực thể *tên người*, *tên địa điểm*, *tên tổ chức* khi sử dụng luật đồng tham chiếu về tên tốt hơn hệ thống trong [7]. Khi sử dụng thêm các luật nhóm 2 bài báo đề xuất, ta thu được kết quả tốt nhất cho các độ đo F.

Bảng 2. Kết quả thực nghiệm của hệ thống trích rút thực thể sử dụng phương pháp học bán giám sát

Lần	Tên người			Tên địa điểm			Tên tổ chức		
	$P(\%)$	$R(\%)$	$F(\%)$	$P(\%)$	$R(\%)$	$F(\%)$	$P(\%)$	$R(\%)$	$F(\%)$
1	**69.88**	**73.51**	**71.65**	**48.27**	**65.95**	**55.74**	**46.03**	**53.75**	**49.16**
2	73.06	77.38	75.16	58.49	69.85	63.67	58.22	65.33	61.57
3	76.13	79.21	77.64	63.34	75.07	68.71	65.66	71.07	68.26
4	80.17	80.89	80.53	69.11	76.97	72.83	67.67	73.60	70.51
5	85.65	84.30	84.97	71.20	78.31	74.59	69.55	75.13	72.23
6	87.45	87.25	87.35	74.83	81.28	77.92	71.83	75.92	73.82
7	91.31	88.57	89.92	75.42	83.78	79.38	75.20	78.85	76.98
8	91.80	91.14	91.47	79.18	86.63	82.74	76.01	80.16	78.03
9	**93.53**	**92.73**	**93.13**	**85.32**	**91.17**	**88.15**	**77.10**	**81.74**	**79.35**

*B*ảng 3. So sánh kết quả lần lặp thứ 9 của ba phương pháp nói trên

Phương pháp	Tên người			Tên địa điểm			Tên tổ chức		
	$P(\%)$	$R(\%)$	$F(\%)$	$P(\%)$	$R(\%)$	$F(\%)$	$P(\%)$	$R(\%)$	$F(\%)$
Heuristic [7]	79.61	87.48	83.36	65.39	74.23	69.53	67.35	64.15	65.71
Nhóm 1	86.62	90.82	88.67	78.80	85.32	81.93	72.62	80.59	76.40
Nhóm1+Nhóm2	93.53	92.73	93.13	85.32	91.17	88.15	77.10	81.74	79.35

5.3.2. Bài toán trích rút quan hệ giữa các thực thể

*B*ảng 4. So sánh các hệ thống trích rút mối quan hệ có giám sát sử dụng các hàm nhân ngữ cảnh tổng hợp khác nhau

Kiểu quan hệ	SLK trong [6]			SLK được đề xuất		
	$P(\%)$	$R(\%)$	$F(\%)$	$P(\%)$	$R(\%)$	$F(\%)$
Chức vụ	81.80	85.70	83.70	83.70	92.60	87.80
Sống ở	55.00	56.20	55.29	55.10	64.80	59.50
Làm việc cho	68.00	67.60	67.79	68.90	94.80	79.80

Bảng 4 cho thấy độ đo F của các kiểu quan hệ *chức vụ*, *sống ở*, *làm việc cho* khi sử dụng hàm nhân tổng hợp của chúng tôi tốt hơn khi sử dụng hàm nhân tổng hợp trong [6]. Kết quả so sánh ở trên đã cho thấy việc đưa thêm các đặc trưng vào để quá trình xét ngữ cảnh toàn cục và mở rộng cửa sổ của quá trình xét ngữ cảnh cục bộ phù hợp để trích rút quan hệ trong văn bản tiếng Việt.

Bảng 5 cho thấy độ đo F của các hệ thống cho các kiểu quan hệ *chức vụ*, *sống ở*, *làm việc cho* khi sử dụng cùng một tập dữ liệu gán nhãn và chưa gán nhãn về mối quan hệ. Hệ thống học bán giám sát tốt hơn hệ thống học có giám sát.

*B*ảng 5. So sánh kết quả của hệ thống có giám sát và hệ thống bán giám sát sử dụng hàm hàm nhân ngữ cảnh tổng hợp SLK đề xuất

Kiểu quan hệ	Hệ thống có giám sát			Hệ thống bán giám sát		
	$P(\%)$	$R(\%)$	$F(\%)$	$P(\%)$	$R(\%)$	$F(\%)$
Chức vụ	83.70	92.60	87.80	92.90	92.90	92.9
Sống ở	55.10	64.80	59.50	81.60	93.20	87.00
Làm việc cho	68.90	94.80	79.80	88.70	77.50	82.70

6. KẾT LUẬN

Trích rút thông tin là bài toán còn mở đối với tiếng Việt, đặc biệt đối với bài toán trích rút thực thể và trích rút mối quan hệ giữa các thực thể. Để có được một hệ thống trích rút thông tin có độ chính xác cao thì cần có một tập dữ liệu huấn luyện lớn. Do tiếng Việt chưa có tập dữ liệu như vậy nên có thể đề xuất hệ thống học bán giám sát sử dụng các đặc tính của ngôn ngữ Việt cho việc trích rút thực thể và quan hệ giữa các thực thể. Đối với trích rút thực thể, ta sử dụng các luật đồng tham chiếu về tên và xử lý nhập nhằng để tìm thêm các thực thể mà hệ thống cũ chưa phát hiện được hoặc phát hiện nhưng được hệ thống đánh giá với độ tin cậy thấp. Các thực thể này được đưa vào tập dữ liệu ban đầu để huấn luyện lại hệ thống sử dụng phương pháp trường ngẫu nhiên có điều kiện CRF. Ngoài ra còn sử dụng thêm các luật có xác suất xuất hiện cao trong các văn bản tiếng Việt để tìm thêm các thực thể huấn luyện mới. Đối với bài toán trích rút quan hệ, ta sử dụng phương pháp hàm nhân để biểu diễn các câu và sử dụng kỹ thuật bootstrapping kết hợp với máy vectơ hỗ trợ SVM để huấn luyện hệ thống bán giám sát. Đối với bài toán trích rút thực thể, các kết quả thực nghiệm trên các kiểu thực thể *con người*, *địa điểm*, *tổ chức* của hệ thống áp dụng phương pháp cải tiến bằng luật đồng tham chiếu về tên cao hơn các kết quả tương ứng của hệ thống trong [7]. Đối với bài toán trích rút quan hệ giữa các thực thể, các kết quả thực nghiệm đối với các quan hệ *chức vụ*, *sống ở*, *làm việc cho* của phương pháp học có giám sát cải tiến sử dụng SLK đề xuất tốt hơn các kết quả tương ứng của phương pháp học có giám sát trong [6]. Hơn nữa, khi áp dụng phương pháp bagging-bootstrapping cho hệ thống sử dụng hàm nhân mức nông cải tiến thì ta thu được kết quả tốt nhất.

Những nghiên cứu tiếp theo sẽ mở rộng các thử nghiệm đối với các kiểu thực thể khác, cũng như các kiểu mối quan hệ giữa các thực thể khác. Ngoài ra, do cấu trúc ngữ pháp của câu là thông tin quan trọng trong bài toán trích rút mối quan hệ việc nghiên cứu cách tích hợp thông tin này vào hệ thống trích rút mối quan hệ giữa các thực thể nhằm tăng thêm độ chính xác của hệ thống là rất cần thiết.

TÀI LIỆU THAM KHẢO

[1] L. Breiman, *Bagging Predictors* **1** (4) (1966) 437–442.

[2] A. Borthwick, "Maximum entropy approach to named entity recognition", Ph.D. thesis, New York University, 1999.

[3] K. Bontcheva, M. Dimitrov, D. Maynard, V. Tablan, H. Cunningham, Shallow methods for named entity conference resolution, *Proc. of TALN 2002 Workshop*, Nancy, France, 2002.

[4] A. Culotta, A. McCallum, Confidence Estimation for Information Extraction, *Proceeding of HLT-NAACL*, 2004 (109–112).

[5] C.C. Chang, C.J. Lin, *"LIBSVM: a Library for Support Vector Machines"*, (2009).

[6] C. Giuliano, A. Lavelli, and L. Romano, Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature, *Proc. EACL.*, 2006.

[7] W. Liao, S. Veeramachaneni, A simple semi-supervised algorithm for named entity recognition, *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, 2009 (28–36).

[8] J. Lafferty, A. McCallum, and F. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. ICML,* pages 282-290.

[9] A. McCallum, W. Li, Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons, *Proceedings of CoNLL*, Canada, 2003 (188-191).

[10] T.H. Nguyen, H.T. Cao, An Approach to Entity Coreference and Ambiguity Resolution in Vietnamese Texts, *Vietnamese Journal of Post and Telecommunication* (19) (2008) 74–83.

[11] C.T. Nguyen, T.O. Tran, X.H. Phan, Q.T. Ha, Named entity recognition in Vietnamese free-text and Web documents using conditional random fields, *The* 8^{th} *Conference on Some selection problems of Information Technology and Telecommunication*, Hai Phong, Vietnam, 2005.

[12] Q.T. Tran, T.X.T. Pham, Q.H. Ngo, D. Dinh, and N. COLLIER, Named entity recognition in Vietnamese using classifier voting, *ACM Transactions on Asian Language Information Processing (TALIP)*, 2007.

[13] Z.Zhang, Weakly supervised relation classification for information extraction, *Proc. of CIKM',* 2004.

[14] G.D. Zhou, J. Su, Named entity recognition using an HMM-based chunk tagger, *Proceedings of the* 40^{th} *Annual Meeting of the ACL*, 2011.

Ngày nhận bài 29 - 3 - 2012