Tạp chí Tin học và Điều khiển học, T.30, S.1 (2014), 15–27

TRÍCH RÚT QUAN HỆ GIỮA CÁC THỰC THỂ TỪ VĂN BẢN TIẾNG VIỆT SỬ DỤNG PHƯƠNG PHÁP LAN TRUYỀN NHÃN

LÊ THANH HƯƠNG[1], SAM CHANRATHANY[1], NGUYỄN THANH THUỶ[2],
NGUYỄN THÀNH LONG[1], TRỊNH MINH DŨNG[1]

[1] *Viện Công nghệ Thông tin và Truyền thông, DH Bách khoa Hà Nội*
[2] *Khoa CNTT, Trường DH Công nghệ, DHQG Hà Nội*

Tóm tắt. Bài báo đề xuất việc xây dựng hệ thống trích rút quan hệ giữa các thực thể từ văn bản tiếng Việt sử dụng phương pháp lan truyền nhãn. Các đóng góp chính là: (i) đề xuất các phương pháp đo độ tương đồng giữa các câu; và (ii) đề xuất phương pháp giảm ảnh hưởng của các nhãn có tần suất xuất hiện lớn đến quá trình lan truyền nhãn. Thử nghiệm cho thấy phương pháp giảm ảnh hưởng của các nhãn có tần suất xuất hiện lớn cho kết quả tốt hơn đáng kể phương pháp lan truyền nhãn gốc [10]. Ngoài ra, khi sử dụng cùng dữ liệu huấn luyện nhỏ phương pháp lan truyền nhãn tốt hơn phương pháp SVM.

Từ khóa. Trích rút mối quan hệ, lan truyền nhãn, học bán giám sát.

Abstract. This paper presents a relation extraction system for Vietnamese texts using label propagation. In this paper, we propose: (i) a measure of similarities between two sentences; (ii) a method to decrease the effect of high frequency labels in the documents. Our experimental results show that proposed label propagation method achieves a higher accuracy than the ordinary one [10]. Moreover, its accuracy is also higher than the support vector machine method applied.

Key words. Relation extraction, labeled propagation, semi supervised learning.

1. MỞ ĐẦU

Trích rút mối quan hệ giữa các thực thể (Relation Extraction - RE) là công việc xác định quan hệ giữa các cặp thực thể trong văn bản. Ví dụ, quan hệ *sống* ở hai thực thể "*tên người*" và "*tên địa điểm*", quan hệ *họ hang* giữa hai thực thể "*tên người*" và "tên người".

Trong hơn một thập niên qua, đã có nhiều nghiên cứu về trích rút quan hệ giữa các thực thể [1, 3, 6, 9, 12]. Các nghiên cứu được chia thành hai hướng. Đó là cách tiếp cận dựa trên việc xây dựng tập luật trích rút một cách thủ công và cách tiếp cận dựa trên học máy. Trong cách tiếp cận thứ nhất, các luật thủ công được xây dựng dựa trên việc quan sát quy luật của dữ liệu, nên thường có độ chính xác cao. Tuy nhiên, cách tiếp cận này không xử lý hết các trường hợp chưa bao quát được trong tập luật.

Trong khi đó, các kĩ thuật học máy thường sử dụng một tập các dữ liệu đã được gán nhãn cho trước để xây dựng nên một mô hình, phục vụ cho mục đích của bài toán (học có giám sát). Đây là cách tiếp cận tự động, cho phép ta học những luật có xuất hiện trong dữ liệu huấn luyện, nhưng khó có thể phát hiện được bằng quan sát thủ công của con người. Khó

khăn trong học có giám sát là cần một tập dữ liệu đã được gán nhãn có kích cỡ 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 vẫn 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 [4, 8, 11]. Ý tưởng cơ bản của phương pháp học máy bán giám sát là: huấn luyện hệ thống sử dụng cả dữ liệu được gán nhãn (thường có kích cỡ nhỏ) và dữ liệu chưa được gán nhãn (thường có kích cỡ lớn).

Zhang và các cộng sự [11] giải quyết bài toán trích rút mối quan hệ giữa các thực thể bằng cách sử dụng phương pháp Bootstrapping kết hợp với SVM. Đầu tiên, họ biểu diễn câu dưới dạng $(c_{pr}, e_1, c_m, e_2, c_{pt}) \rightarrow r$, trong đó e_1 và e_2 là thực thể đang xét mối quan hệ r, c_{pr}, c_m, c_{pt} lần lượt là ngữ cảnh trước, giữa và sau cặp thực thể. Sau đó, sử dụng phương pháp Bagging Bootstrapping để huấn luyện hệ thống. Ý tưởng của phương pháp này là: Giả sử có L mẫu có nhãn và U mẫu chưa gán nhãn. Đầu tiên, nhân bản các mẫu có nhãn L thành B gói và huấn luyện B bộ phân lớp sử dụng dữ liệu đã nhân bản. B bộ phân lớp này được áp dụng trên dữ liệu chưa có nhãn U. Sau khi đã gán nhãn cho tập dữ liệu U, hệ thống tính độ tin cậy để tìm S câu có độ tin cây cao (độ tin cậy này được tính bằng hàm entropy) và đưa thêm vào dữ liệu huấn luyện. Quá trình này được lặp lại cho đến khi không tìm được dữ liệu nào thỏa mãn nữa.

Tác giả trong [8] sử dụng phương pháp học máy bán giám sát sử dụng phương pháp SVM kết hợp với kỹ thuật bagging bootstrapping để trích rút mối quan hệ trong văn bản tiếng Việt. Đầu tiên, họ biến đổi các câu trong văn bản thành hai hàm nhân. Hai hàm nhân đó là hàm nhân ngữ cảnh toàn cục (thu thập thông tin ngữ cảnh trong câu để suy ra mối quan hệ) và hàm nhân ngữ cảnh cục bộ (để suy ra vai trò của các thực thể trong câu, xác định đâu là tác nhân, đâu là đích). Tiếp theo, họ sử dụng SVM kết hợp với kỹ thuật bagging-bootstrapping để huấn luyện hệ thống.

Chen và các cộng sự [4] đề xuất phương pháp bán giám sát, sử dụng giải thuật lan truyền nhãn (label propagation). Họ biểu diễn các mẫu (có nhãn và chưa có nhãn) dưới dạng các nút, khoảng cách giữa các nút là trọng số các cạnh của đồ thị. Trên cơ sở đó, xây dựng hai ma trận Y và T. Ma trận Y có kích thước $m \times n$, với n là số mẫu có nhãn và chưa có nhãn, m là số nhãn cần xét. Ma trận T, có kích thước $n \times n$, đo độ tương đồng giữa các mẫu. Thực hiện nhân hai ma trận này và lặp lại quá trình đó nhiều lần cho đến khi hội tụ. Kết thúc quá trình, trong ma trận Y, các mẫu sẽ có nhãn tương ứng với phần tử có giá trị lớn nhất. Như vậy, điểm nhấn của phương pháp này là đo mức độ tương đồng giữa các mẫu. Có thể thấy rõ ưu điểm của phương pháp ở chỗ, nhãn quan hệ dựa trên sự tương tự giữa mẫu nên không cần đến bộ dữ liệu lớn.

Trên cơ sở ưu nhược điểm của các phương pháp đó, bài báo đề xuất cải tiến giải thuật lan truyền nhãn của Chen và các cộng sự [4] cho bài toán trích rút quan hệ giữa các thực thể cho văn bản tiếng Việt.

2. TRÍCH RÚT QUAN HỆ GIỮA CÁC THỰC THỂ SỬ DỤNG PHƯƠNG PHÁP LAN TRUYỀN NHÃN

2.1. Phương pháp lan truyền nhãn

Trong phương pháp này, các dữ liệu đã gán nhãn và chưa gán nhãn được biểu diễn dưới dạng các điểm trong không gian. Quá trình lan truyền nhãn sẽ được thực hiện theo kiểu qui

nạp, bằng cách gán nhãn dần các điểm chưa gán nhãn, dựa trên khoảng cách giữa chúng với điểm đã gán nhãn. Cách biểu diễn của dữ liệu này là đồ thị.

Giả sử ta có đồ thị $G = (V, E)$, với $V = \{1, ..., n\}$ là tập các nút và E là tập các cạnh. Trong bài toán trích rút quan hệ giữa các thực thể, mỗi nút là một câu đã gán nhãn hoặc chưa gán nhãn quan hệ. Mỗi cạnh ứng với độ tương đồng giữa các câu đó. Độ tương đồng này được biểu diễn bởi ma trận T khi x_i và x_j là láng giềng thì $T_{ij} \neq 0$. Khi đó, cạnh (i, j) trong E có trọng số là T_{ij}.

Ý tưởng học bán giám sát nhằm lan truyền nhãn trong đồ thị được thể hiện như sau: Tại thời điểm ban đầu, các nút $1, 2, ..., l$ có nhãn và các nút $l + 1, ..., n$ chưa có nhãn. Tiến hành lan truyền nhãn của mỗi nút cho các láng giềng của nó. Quá trình này lặp đi lặp lại cho đến khi không lan truyền nhãn tiếp được nữa hoặc khi đã gán nhãn cho tất cả các đỉnh trong đồ thị.

Trong phương pháp lan truyền nhãn, mỗi mẫu được biểu diễn bằng một nút và khoảng cách giữa hai nút là trọng số cạnh nối của chúng. Sau đó, thông tin nhãn của một nút trong đồ thị được lan truyền cho nút bên cạnh thông qua trọng số của cạnh cho đến khi đạt được trạng thái ổn định. Trọng số của cạnh càng lớn, nhãn đi qua cạnh dễ dàng. Do đó mẫu càng giống nhau thì càng có nhãn giống nhau.

Giải thuật lan truyền nhãn đề xuất bởi các tác giả trong [10], được mô tả trong giải thuật 1. Ma trận Y (biểu diễn mối quan hệ giữa mẫu và nhãn) và ma trận T (đo độ tương đồng giữa các mẫu) được xây dựng. Ma trận Y có n hàng, m cột với n là tổng số mẫu đã gán nhãn và chưa gán nhãn, m là số nhãn cần xét; $Y_{ij} = 1$ nếu mẫu thứ i có nhãn j, và bằng 0 trong trường hợp ngược lại. Ma trận T có kích thước $n \times n$ với n là tổng số mẫu bao gồm cả mẫu đã gán nhãn và chưa gán nhãn; T_{ij} là độ tương tự giữa mẫu thứ i với mẫu thứ j. Sau đó, lặp lại việc nhân ma trận T với ma trận Y nhiều lần đến khi hội tụ. Cuối cùng, các mẫu chưa có nhãn trong ma trận Y sẽ được gán nhãn ứng với phần tử có giá trị lớn nhất trong hàng ứng với mẫu đó.

Trong quá trình lan truyền nhãn, nhãn ban đầu của các mẫu đã được gán bằng tay được giữ lại trong mỗi bước lặp để cung cấp nguồn nhãn, có nghĩa là trong mỗi bước lặp l dòng đầu của ma trận Y sẽ mang giá trị giống hệt ma trận khởi tạo. Các mẫu đã được gán nhãn bằng tay này đóng vai trò như nguồn để sinh nhãn cho các dữ liệu không có nhãn.

Giải thuật 1: Lan truyền nhãn trong [10]

Bước 1: Khởi tạo

$+ t = 0$

$+ Y^0$ khởi tạo nhãn ban đầu kết nối với mỗi nút, trong đó $Y^0_{ij} = 1$ nếu y_i có nhãn r_j và ngược lại bằng 0.

$+ Y^0_L$ là l dòng phía trên của ma trận Y^0 tương ứng với l dữ liệu đã có nhãn và Y^0_U là u dòng còn lại, tương ứng với các dữ liệu chưa có nhãn.

Bước 2: Lan truyền nhãn của các nút nào cho nút láng giềng bằng cách $Y^{t+1} = \overline{T}Y^t$, trong đó $\overline{T}$ là ma trận chuẩn hóa của ma trận T.

Bước 3: Giữ lại phần có nhãn ban đầu, tức là thay l dòng đầu của ma trận Y^{t+1} bằng Y^0_L.

Bước 4: Lặp lại bước 2 cho đến khi thoả mãn điều kiện dừng.

Bước 5: Gán x_h $(l + 1 \leq h \leq n)$ bằng nhãn $y_h = \arg\max_j Y_{hj}$.

Điều kiện dừng ở đây có thể là số vòng lặp lớn hơn tham số Q nào đó hoặc là vòng vặp sẽ dừng khi $Y^t = Y^{t+1}$.

2.2. Đo độ tương đồng giữa các câu dựa trên phương pháp so trùng thuộc tính từ

Mục tiêu của bài toán là tính độ tương đồng giữa các câu có chứa ít nhất hai thực thể. Bài toán được phát biểu như sau: Xét một tài liệu d có n câu: $d = S_1, S_2, ..., S_n$. Mục tiêu của bài toán là tìm các giá trị độ tương đồng giữa các cặp câu (S_i, S_j). Giá trị này càng cao, sự giống nhau về ngữ nghĩa của hai câu càng lớn. Hai câu có độ tương đồng càng lớn, khả năng nó chứa cùng một mối quan hệ càng cao.

Giả sử:

Câu thứ nhất có m từ, $S_1 = A_1A_2A_3...A_m$.

Câu thứ hai có p từ, $S_2 = B_1B_2B_3...B_p$.

Sim$W(A_i, B_j)$ là độ tương đồng giữa từ A_i trong S_1 và từ B_j trong S_2, $i = \overline{1,m}$, $j = \overline{1,p}$.

Sim$WS(A_i, S_2)$ là độ tương đồng giữa từ A_i với tất cả các từ trong câu thứ hai $B_1B_2B_3...B_p$.

Sim$GB(S_1, S_2)$ là độ tương đồng ngữ cảnh toàn cục giữa hai câu.

Sim$LC(S_1, S_2)$ là độ tương đồng ngữ cảnh cục bộ giữa hai câu.

Sim$S(S_1, S_2)$ là độ tương đồng giữa hai câu.

Chúng tôi đề xuất tính độ tương đồng ngữ nghĩa giữa hai câu như sau: Mỗi từ trong câu thứ nhất được so với tất cả các từ trong câu thứ hai về các khía cạnh: từ, từ loại, kiểu thực thể, cây ngữ nghĩa. Độ tương đồng giữa mỗi từ trong câu thứ nhất với tất cả các từ trong câu thứ hai được tính bằng

$$\text{Sim}WS(A_i, S_2) = \max_{1 \le j \le p} \text{Sim}W(A_i, B_j), \tag{1}$$

tức là chỉ giữ lại giá trị độ tương đồng từ lớn nhất của từ A_i trong câu thứ nhất so với tất cả các từ trong câu thứ hai. Cuối cùng, độ tương đồng ngữ nghĩa giữa hai câu được tính bằng

$$\text{Sim}GB(S_1, S_2) = \sum_{i=1}^{m} \text{Sim}WS(A_i, S_2). \tag{2}$$

Ví dụ: câu "Nam hiện nay đang sống ở Sài Gòn với đồng nghiệp" và câu "Thủy sống tại Hà Nội" khi gán thẻ từ loại có dạng sau:

Nam/E1	hiện nay	đang	sống	ở	**Sài Gòn/E2**	với	đồng nghiệp
Np	N	R	V	P	Np	P	N

Thủy/E1	sống	tại	**Hà Nội/E2**
Np	V	P	Np

Trong đó N, R, V, P, N_p tương ứng là danh từ, phụ từ, động từ, giới từ, danh từ riêng. E_1, E_2 là thực thể đang xét mối quan hệ.

Ta sẽ thực hiện tính mức độ tương đồng giữa các từ: Như vậy trong ví dụ này $m = 8, n = 4$. Ta có hai tập từ: { Nam, hiện nay, đang, sống, ở, Sài Gòn, với, đồng nghiệp} và { Thủy, sống, tại, Hà Nội}. Giả sử xét độ tương đồng S_{T1} của từ "Nam" trong câu thứ nhất với tất cả các từ trong câu thứ hai {Thủy, sống, tại, Hà Nội}. Ta sẽ tính độ tương đồng từ giữa các cặp (Nam, Thủy), (Nam, sống), (Nam, tại), (Nam, Hà Nội), và sau đó chọn giá trị lớn nhất giữa các độ tương đồng từ này sẽ được giá trị Sim$WS(A_1, S_2)$.

Ta thấy rằng từ "Nam" và từ "Thủy" là hai từ khác nhau và cùng từ loại là N, có cùng kiểu thực thể là tên người và cùng trong một lớp của cây ngữ nghĩa vậy độ tương đồng của hai từ Sim$W$(Nam,Thủy)= 3. Tương tự Sim$W$(Nam, sống)=1/5, Sim$W$(Nam, tại)=1/5, Sim$W$(Nam, Hà Nội)=7/6. Như vậy SimWS (Nam, { Thủy, sống, tại, Hà Nội})=3. Tiếp tục làm như vậy với từ khác sau đó cộng lại, ta được độ tương đồng ngữ nghĩa giữa hai câu.

Nhược điểm của phương pháp trên và cách giải quyết

Xét hai câu sau:

(a) "*Hiện nay, anh **Nam** đang sống tại **Mỹ Đình** và làm việc cho **công ty FPT** ở Hai Bà Trưng*".

(b) "*Chị **Thủy** hiện nay đang sống ở **Mỹ Đình***".

Dựa trên câu (a), cho ta biết rằng anh Nam đang sống ở Mỹ Đình, nhưng làm việc ở Hai Bà Trưng và làm việc cho công ty FPT. Như vậy, trong câu này có ba mối quan hệ: *sống ở* (Nam, Mỹ Đình), *địa điểm làm việc* (Nam, Hai Bà Trưng), *làm việc cho* (Nam, FPT).

Giả sử là câu (a) đã gán nhãn và câu (b) chưa gán nhãn. Nói cách khác, câu (a) đã gán nằm trong tập dữ liệu đã gán nhãn L và câu (b) nằm trong tập dữ liệu chưa gán nhãn U. Các kiểu quan hệ được xét là *sống ở, làm việc cho, địa điểm làm việc*. Như vậy để đảm bảo có đủ thông tin cả ba mối quan hệ trên thì câu (a) phải xuất hiện ba lần trong L, mỗi lần tương ứng một kiểu quan hệ.

(a1) "*Hiện nay, anh **Nam (A)** đang sống tại **Mỹ Đình (T)** và làm việc cho **công ty FPT** ở **Hai Bà Trưng***". Đây là quan hệ "*sống ở*".

(a2) "*Hiện nay, anh **Nam (A)** đang sống tại **Mỹ Đình** và làm việc cho **công ty FPT(T)** ở **Hai Bà Trưng***". Đây là quan hệ "*làm việc cho*".

(a3) "*Hiện nay, anh **Nam (A)** đang sống tại **Mỹ Đình** và làm việc cho **công ty FPT** ở **Hai Bà Trưng(T)***". Đây là quan hệ "*địa điểm làm việc*".

trong đó A chỉ tới thực thể tác nhân, T chỉ tới thực thể đích. Nói cách khác, A và T cho ta biết đang xét kiểu quan hệ giữa cặp thực thể nào.

Như vậy, khi xây dựng ma trận độ tương đồng T, ta cũng cần đo độ tương đồng giữa (b,a1), (b,a2), (b,a3). Ta thấy, bản chất của câu (b) là kiểu quan hệ sống ở và có phần rất giống với câu (a1). Nhưng khi áp dụng phương pháp đo độ tương đồng ngữ nghĩa giữa hai câu trên thì SimGB(b,a1)=SimGB(b,a2)=SimGB(b,a3). Nghĩa là câu b thuộc cả ba kiểu quan hệ, như vậy tạo ra sự nhập nhằng dẫn đến thuận toán sẽ nhận dạng sai các mối quan hệ.

Độ tương đồng ngữ cảnh cục bộ giữa hai câu: là độ tương đồng so khớp các từ trong cửa số ngữ cảnh xung quanh hai thực thể của hai câu.

Ta thấy rằng khi ta biết thực thể nào đang xét mối quan hệ, thực thể nào là tác nhân và thực thể nào là đích thì chúng ta có thể thu hẹp được phạm vi đo độ tương đồng trong câu. Hơn nữa, với những câu như vậy, các động từ chỉ mối quan hệ thường nằm gần thực thể đích. Dựa trên ý tưởng đó, chúng tôi khắc phục vấn đề trên bằng cách tính độ tương đồng ngữ cảnh cục bộ Sim$LC(S_1, S_2)$ như sau:

- Gán nhãn A và T cho các thực thể trong câu, nhằm chỉ ra đâu là thực thể tác nhân và đâu là thực thể đích đang xét mối quan hệ.
- Tạo cửa sổ ngữ cảnh xung quanh thực thể A và thực thể T kích thước 7 (gồm thực thể đang xét, 3 từ trước và 3 từ sau nó).

- Tính độ tương đồng ngữ cảnh cục bộ xung quanh các thực thể $A, \text{Sim}LCA(S_1, S_2)$; và xung quanh các thực thể $T, \text{Sim}LCT(S_1, S_2)$. Cả hai độ tương đồng này được tính tương tự như phương pháp đo độ tương đồng ngữ cảnh toàn cục giữa hai câu nói trên nhưng chỉ khác ở chỗ thay vì so khớp toàn bộ từ trong câu thứ nhất với toàn bộ từ trong câu thứ hai, chỉ so sánh các từ nằm trong cửa số ngữ cảnh xung quanh thực thể. Ví dụ, đối với $\text{Sim}LCA(S_1, S_2)$ chỉ so khớp tất cả các từ nằm trong cửa số 7 xung quanh thực thể A trong câu thứ nhất với tất cả các từ nằm trong cửa số 7 xung quanh thực thể A trong câu thứ 2.

- Độ tương đồng ngữ cảnh cục bộ giữa hai câu được tính bằng

$$\text{Sim}LC(S_1, S_2) = \text{Sim}LCA(S_1, S_2) + \text{Sim}LCT(S_1, S_2). \quad (3)$$

Độ tương đồng giữa hai câu: là sự kết hợp giữa độ tương đồng ngữ cảnh toàn cục với độ tương đồng ngữ cảnh cục bộ.

$$\text{Sim}S(S_1, S_2) = \text{Sim}GB(S_1, S_2) + \text{Sim}LC(S_1, S_2). \quad (4)$$

Làm như vậy, ta có thể tạo sự khác biệt giữa 3 lần tính độ tương đồng trên, tức là: $\text{Sim}S(\text{b,a1}) \neq \text{Sim}S(\text{b,a2}) \neq \text{Sim}S(\text{b,a3})$, làm cho giải thuật có thể phân lớp tốt hơn.

Giải thuật đo độ tương đồng câu dựa trên phương pháp so trùng thuộc tính từ do nhóm nghiên cứu đề xuất như sau.

Giải thuật 2. Độ tương đồng câu dựa trên phương pháp so trùng thuộc tính từ

Đầu vào: Cặp câu đã gán nhãn thực thể E_1, E_2

Đầu ra: Độ tương đồng giữa các câu

Khởi tạo: Độ tương đồng ngữ cảnh toàn cục giữa hai câu $\text{Sim}GB(S_1, S_2) = 0$

Số từ trong câu thứ nhất là m

Số từ trong câu thứ hai là p

Chuyển câu thành tập các từ và xác định A và T

Phương pháp:

Bước 1: Tính độ tương đồng ngữ cảnh toàn cục giữa hai câu

For $i = 1$ to m do {

- For $j = 1$ to p do *Tính độ tương đồng giữa hai từ* $SimW(A_i, B_j)$

- Tính độ tương đồng của từ thứ i trong câu thứ nhất so với tất cả các từ trong câu thứ hai $\text{Sim}WS(A_i, S_2)$

-Tính độ tương đồng ngữ cảnh toàn cục

$$\text{Sim}GB(S_1, S_2) = \text{Sim}GB(S_1, S_2) + \text{Sim}WS(A_i, S_2)$$

}

Bước 2: Tính độ tương đồng ngữ cảnh cục bộ $\text{Sim}LCA(S_1, S_2),\ \text{Sim}LCT(S_1, S_2)$.

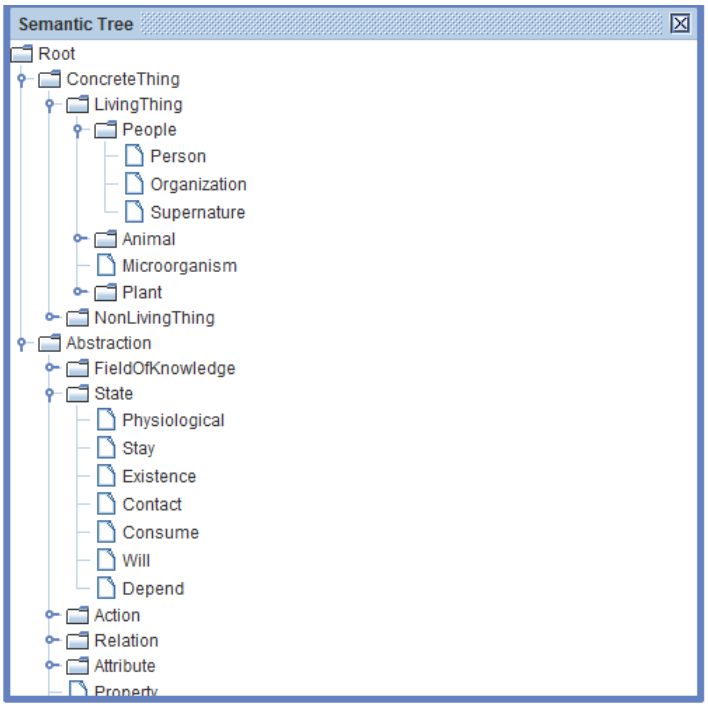
Bước 3: Kết hợp độ tương đồng ngữ nghĩa giữa hai câu và độ tương đồng ngữ cảnh xung quanh hai thực thể thì ta được độ tương đồng giữa hai câu.

Trong giải thuật trên, ta cần đo độ tương đồng từ trong hai câu. Để làm việc này, một phương pháp đo độ tương đồng từ được đề xuất như sau.

Phương pháp đo độ tương đồng từ trên cây ngữ nghĩa

Cây ngữ nghĩa là một cấu trúc phân cấp biểu diễn quan hệ ngữ nghĩa giữa các khái niệm. Trên thế giới có nhiều phương pháp đánh giá độ tương tự giữa các từ dựa trên một mạng ngữ nghĩa biểu diễn quan hệ giữa các từ (ví dụ: Wordnet).

Đối với tiếng Việt, do chưa có một mạng ngữ nghĩa như vậy, nên để giải quyết vấn đề này có thể kết hợp cây phân cấp ngữ nghĩa (hình 1) và từ điển từ do trung tâm từ điển học Việt nam (Vietlex) [14] xây dựng.



Hình 1. Cấu trúc cây phân cấp ngữ nghĩa

Để tính độ tương tự ngữ nghĩa giữa hai từ nào đó, trước tiên ta sẽ tìm hai từ đó trong từ điển từ để tìm lớp khái niệm mà hai từ đó thuộc. Sau đó dựa trên cây phân cấp ngữ nghĩa sẽ tính khoảng cách giữa lớp khái niệm mà hai từ thuộc. Ví dụ:

Đối với hai từ "*con trai*" và "*con mèo*" tìm trong từ điển Vietlex thì từ "*con trai*" thuộc lớp Person và từ "*con mèo*" thuộc lớp Animal. Lớp Person và lớp Animal có khoảng cách là 2 trên cây phân cấp ngữ nghĩa.

Một ví dụ khác:

Khi tìm trong từ điển Vietlex, từ "*nông dân*" thuộc lớp Person và từ "*công dân*" cũng thuộc lớp Person. Vậy hai từ này có khoảng cách là 1 trên cây phâp cấp ngữ nghĩa.

Độ tương tự ngữ nghĩa giữa hai từ $c1$ và $c2$ được tính như sau

$$\text{Sim}(c1, c2) = 1/dist(c1, c2), \tag{5}$$

trong đó, $dist(c1, c2)$ là khoảng cách giữa từ $c1, c2$ trên cây phân cấp ngữ nghĩa. Đối với ví dụ đầu tiên, độ tương tự ngữ nghĩa giữa từ "*con trai*" và "*làm việc*" là $1/3$. Với ví dụ thứ hai, độ tương tự ngữ nghĩa giữa từ "*nông dân*" và từ "*công nhân*" là 1.

2.3. Đo độ tương đồng giữa hai câu dựa trên phương pháp mô hình Dirichlet ẩn

Mô hình Dirichlet ẩn (Latent Dirichlet Allocation – LDA) [5, 7] dựa trên ý tưởng: mỗi tài liệu là sự trộn lẫn của nhiều chủ đề, trong đó mỗi chủ đề là một phân bố trên một tập từ vựng. Cụ thể là, ta có K chủ đề ứng với D tài liệu; mỗi tài liệu liên quan đến các chủ đề này theo các tỷ lệ khác nhau. Về bản chất, LDA là một mô hình Baysian ba cấp trong đó mỗi phần của tập hợp được biểu diễn như một mô hình trộn hữu hạn trên cơ sở tập các xác suất chủ đề. Trong ngữ cảnh của mô hình văn bản, xác suất chủ đề cung cấp một biểu diễn tường minh cho tài liệu. Sự tương tác giữa các tài liệu và chủ đề ẩn được thể hiện trong tiến trình ngẫu nhiên, giả định sinh ra các tài liệu.

Mô hình sinh trong LDA

Cho một tập ngữ liệu có M tài liệu, được biểu diễn bởi $D = \{d_1, d_2, ..., d_M\}$, trong đó mỗi tài liệu có N_m từ w_i rút từ tập từ vựng $\{t_1, t_2, ..., t_v\}$, V là số kích thước của tập từ vựng. Ta có tiến trình sinh xác suất cho một tập tài liệu như sau.

Giải thuật 3. Tiến trình sinh xác suất các tài liệu dạng văn bản trong bộ dữ liệu

Phương pháp

(1) Với mỗi chủ đề,

a. Tính phân bố của các từ trên chủ đề đó $\vec{\varphi_k} \sim Dir(\overline{\beta})$

(2) Với mỗi tài liệu,

a. Tính phân bố chủ đề trên tài liệu đó $\vec{\vartheta}_m \sim Dir(\overline{\alpha})$

b. Với mỗi từ,

Tìm chủ đề được gán với từ đó $Z_{m,n} \sim Mult(\vec{\vartheta}_m),\ Z_{m,n} \in \{1, 2, ..., K\}$

Tìm một từ dựa vào chủ đề được gán với nó

$$W_{d,n} \sim Mult(\vec{\varphi}_{Z_{m,n}}),\ W_{m,n} \in \{1, 2, ..., V\}$$

Ở đây, Dir và $Mult$ lần lượt là các phân phối *Dirichlet*, *Multinomial* (lấy mẫu theo phân phối Dirichlet, Poisson, Multinomial).

Đối với bài toán đo độ tương đồng giữa các câu, giải thuật LDA nhận đầu vào là các câu, đầu ra là kết luận về chủ đề của câu.

Mỗi câu sẽ được gắn với các phân phối xác suất của các chủ đề trên câu và phân phối xác suất của từ trên chủ đề. Nói cách khác, với mỗi câu i, LDA sinh phân phối chủ đề cho câu. Với mỗi từ trong câu, phân phối xác suất chủ đề của từ j trong câu i (Z_{ij}) được lấy mẫu dựa theo phân phối chủ đề trên. Dựa vào Z_{ij}, hệ thống làm giàu các câu bằng cách thêm từ. Vectơ tương ứng với câu thứ i có dạng sau

$$S_i = \{p_1, p_2, ..., p_k,\ q_1, q_2, ..., q_v\}$$

với p_l là trọng số của chủ đề thứ l trong K chủ đề đã được phân tích; q_i là trọng số của từ thứ i trong tập từ vựng V của tất cả các câu. Trường hợp từ j nào đó không có trong câu i, sẽ được gán giá trị 0.

Mỗi câu có thể có nhiều phân phối xác suất chủ đề. Với hai câu s_1 và s_2, ta sử dụng độ

đo cosine để tính độ tương đồng của hai câu.

$$S_s = \frac{S_i \times S_j}{\|S_i\| \times \|S_j\|}. \tag{6}$$

Hay

$$\text{Sim}_{i,j}(\text{chủ đề -Part}) = \frac{\prod_{k=1}^{K} p_{i,k} \times p_{j,k}}{\sqrt{\sum_{k=1}^{K} p_{i,p}^2} \times \sqrt{\sum_{k=1}^{K} p_{j,p}^2}}, \tag{7}$$

$$\text{Sim}_{i,j}(\text{từ -Part}) = \frac{\prod_{p=1}^{|V|} q_{i,p} \times q_{j,p}}{\sqrt{\sum_{i=1}^{|V|} q_{i,p}^2} \times \sqrt{\sum_{p=1}^{|V|} q_{j,p}^2}}, \tag{8}$$

trong đó $\text{Sim}_{i,j}$ (chủ đề-Part) là độ tương đồng của hai câu i và j tính theo vectơ trọng số chủ đề p_i; $\text{Sim}_{i,j}$ (từ-part) là độ tương đồng của hai câu i và j tính theo vectơ trọng số từ q_i;

Độ tương đồng giữa hai câu được tính trên cơ sở tổ hợp của hai độ đo trên

$$\text{Sim}(s_i, s_j) = \lambda \times \text{Sim}(\text{chủ đề -Part}) + (1 - \lambda) \times \text{Sim}(\text{từ -Part}), \tag{9}$$

trong công thức trên λ là hằng số trộn, nằm trong đoạn [0,1].

2.4. Điểm yếu của giải thuật lan truyền nhãn [10] và cách cải tiến

Ta thấy rằng, trong giải thuật lan truyền nhãn có sự ảnh hưởng giữa nhãn này so với nhãn khác, phụ thuộc vào số lượng mỗi nhãn trong tập dữ liệu. Nói cách khác, kết quả đầu ra sẽ bị ảnh hưởng rất lớn bởi các nhãn có tần suất xuất hiện lớn. Ví dụ, giả sử trong tập dữ liệu đã gán nhãn, số nhãn ứng với các quan hệ *sống ở, làm việc cho, chức vụ* ít hơn rất nhiều số nhãn O (không thuộc quan hệ quan tâm).Do vậy, khi áp dụng giải thuật lan truyền nhãn qua các phép nhân ma trận, do số lượng nhãn O quá nhiều nên các giá trị liên quan đến nhãn O lớn hơn nhiều các giá trị liên quan đến các nhãn khác. Điều này dẫn đến trong bước 5 của giải thuật lan truyền nhãn (gán x_h bằng nhãn $y_h = \arg\max_j Y_{hj}$), các nhãn O đã thay thế dần các nhãn khác. Nhiều chỗ kiểu thực thể thực tế phải là tên người hoặc một quan hệ nào đó khác nhưng lại trở thành nhãn O. Đây chính là vấn đề dữ liệu huấn luyện không cân bằng, trong đó có một loại mẫu nhiều hơn hẳn các loại mẫu khác.

Cải tiến thuận toán:

Các giải pháp thường được nghĩ đến trong việc giải quyết vấn đề dữ liệu huấn luyện không cân bằng là: (i) tăng số lượng mẫu chiếm thiểu số hoặc giảm số lượng mẫu chiếm đa số; và (ii) gán trọng số cho các loại mẫu dữ liệu huấn luyện. Do tập ngữ liệu phục vụ bài toán xác định mối quan hệ giữa các thực thể trong văn bản tiếng Việt không có sẵn mà chúng tôi phải tự xây dựng bằng tay nên tập ngữ liệu có được không dư thừa để có thể loại bỏ các mẫu ứng với nhãn O. Hơn nữa, việc bổ sung thêm một tập dữ liệu có các nhãn khác để cân bằng với nhãn O cũng khó khăn. Số lượng mẫu có nhãn O nhiều hơn các nhãn khác là do chúng tôi xuất phát từ việc gán nhãn các quan hệ trong một văn bản; do đó số lượng nhãn thuộc mỗi

loại không thể cân bằng. Vì vậy, ta chọn giải pháp thứ hai: mỗi mẫu huấn luyện thuộc loại mẫu chiếm số đông có trọng số nhỏ hơn các loại mẫu còn lại.

Trong cách tiếp cận của bài báo, trọng số của một loại mẫu sẽ bằng phần bù xác suất xuất hiện của loại mẫu đó trong tập mẫu. Cụ thể là, để giảm ảnh hưởng của các nhãn có tần suất xuất hiện lớn trong ma trận, khi xây dựng ma trận Y^0, đối với nửa trên của ma trận là các từ đã được gán nhãn, thay vì gán các giá trị là 1 ta sẽ gán nó bằng phần bù xác suất xuất hiện của nhãn đó trong phần dữ liệu L đã gán nhãn.

Ví dụ: Tập L có 100 câu trong đó 12 câu có quan hệ *sống ở*, 20 câu có quan hệ *làm việc cho*, 18 câu có quan hệ *chức vụ*, 50 câu có quan hệ O. Với ma trận Y^0, những câu có quan hệ O được gán trọng số $\alpha = 1 - 50/100 = 0,5$, những câu có quan hệ *sống ở* có $\alpha = 1 - 12/100 = 0,88$, những câu có quan hệ *làm việc cho* có trọng số $\alpha = 1 - 20/100 = 0,8$, và câu có quan hệ *chức vụ* có trọng số $\alpha = 1 - 18/100 = 0,82$.

Như vậy, với một nhãn xuất hiện nhiều lần thì giá trị trong Y^0 sẽ nhỏ và ngược lại.

Cách giải quyết này rất hiệu quả, đối với bài toán trích rút thực thể bởi vì số nhãn O luôn luôn xuất hiện nhiều hơn các nhãn khác.

3. THỬ NGHIỆM

3.1. Tập ngữ liệu và phương pháp thử nghiệm

Tập ngữ liệu thử nghiệm được thu thập thủ công từ các trang web tiếng Việt bao gồm cáctrang web cá nhân và các trang tin tức (vnexpress.net, dantri.com, wikipedia) thuộc các lĩnh vực thể thao, khoa học, văn hóa, giáo dục, và kinh tế đã thu thập được 950 văn bản từ các trang web đó. Từ các văn bản trên, ta trích chọn được 1200 câu chứa ít nhất hai thực thể thuộc một trong các loại tên người, tên địa điểm, tên tổ chức, chức vụ. Các câu này đượcgán nhãn bằng tay, 960 câu trong số các câu này được giấu nhãn để làm tập test. Mỗi văn bản có hai người gán (một người gán, một người kiểm tra lại). Như đã nói ở trên, các mối quan hệ được xét trong thử nghiệm là *làm việc tại (tên người-tên tổ chức), sống ở (tên người-tên địa điểm), chức vụ (tên người-chức vụ).*

Ví dụ 1

-<per> Stephen Hawking </per> - nhà vật lý thiên văn số 1 thế giới- <per> đã không thắng được bệnh tật, nhưng nó cũng không quật ngã được ông.

Câu này chỉ chứa một thực thể tên người (per) nên không được chọn.

Ví dụ 2

- Vài ngày trước đó, ở <loc> Paris </loc> chúng tôi đã gặp anh <per> Christophe Galfard </per>, một trong 6 nghiên cứu sinh của <per> Hawking </per> .

Câu này được chọn vì chứa cả thực thể *tên người* (per) và *tên địa điểm* (loc). Câu này dù có chứa thực thể *tên người* và thực thể *tên địa điểm* nhưng không phải là quan hệ *sống ở*, như vậy được gán nhãn là "0" và làm mẫu âm cho quá trình huấn luyện.

Để đánh giá hệ thống, ta khởi tạo u bằng 960 câu chưa có nhãn mối quan hệ, l bằng 240 câu có nhãn mối quan hệ. Vì ba mối quan hệ được xét là *làm việc cho, sống ở, chức vụ*,ma trận Y sẽ có 4 cột và $n = 1200$ dòng là số câu có nhãn và chưa có nhãn quan hệ; $l = 240$ là số câu có nhãn tương ứng với l dòng đầu tiên của ma trận Y, u=960 là số câu chưa có nhãn tương ứng với u dòng còn lại của ma trận Y. Ma trận T kích thước $n \times n$ là ma trận đo mức độ tương tự giữa các câu.

3.2. Phương pháp đánh giá

Kết quả của hệ thống sẽ được đánh giá thông qua 3 độ đo: độ chính xác P, độ phủ R, độ đo trung bình F. Độ chính xác P xác định phần trăm các mẫu (trong thí nghiệm này là câu) đúng được hệ thống tìm thấy so với các mẫu được hệ thống cho là đúng. Độ phủ R xác định phần trăm mẫu đúng được hệ thống tìm thấy so với thực tế hoặc gán bằng tay. Độ đo F là giá trị trung bình giữa độ phủ R và độ chính xác P. Độ đo P, R, F được tính theo công thức sau

$$P = \frac{\text{Số mẫu được hệ thống gán đúng}}{\text{Số mẫu được hệ thống cho là đúng}}, \tag{10}$$

$$R = \frac{\text{Số mẫu được hệ thống gán đúng}}{\text{Số mẫu đúng gán bằng tay}}, \tag{11}$$

$$F = \frac{2 \times P \times R}{P + R}. \tag{12}$$

Như đã nói ở trong Mục 2.2, trong trường hợp câu chứa n mối quan hệ thì câu đó sẽ xuất hiện n lần trong tập ngữ liệu, mỗi lần xuất hiện ứng với một quan hệ. Các độ đo P, R, F vẫn được tính bình thường theo công thức trên. Sau đây là kết quả thử nghiệm của hệ thống đối với bài toán trích rút thực thể và trích rút mối quan hệ giữa các thực thể.

3.3. Kết quả thu được

Bảng 1. So sánh kết quả khi chưa chuẩn hoá ma trận và sau khi chuẩn hoá ma trận

Kiểu mối quan hệ	**Chưa chuẩn hoá ma trận**			**Sau khi chuẩn hoá ma trận**		
	P	**R**	**F**	**P**	**R**	**F**
Chức vụ	90.90	18.51	30.76	81.53	98.14	89.07
Sống ở	100.0	18.75	31.57	74.57	91.66	82.24
Làm việc cho	83.0	78.0	80.42	70.39	97.36	82.62

Bảng 1 cho thấy kết quả của phương pháp lan truyền nhãn được đề xuất (chuẩn hoá ma trận Y^0) tốt hơn phương pháp lan truyền nhãn trong [4] (khi chưa chuẩn hoá ma trận).

Bảng 2. So sánh độ đo F của ba phương pháp đo độ tương đồng từ

	Chức vụ	**Sống ở**	**Làm việc cho**
So trùng thuộc tính từ	89.07	82.24	82.62
LDA	87.25	82.07	82.10
LDA + trùng thuộc tính từ	90.9	85.0	82.65

Bảng 2 cho thấy kết quả phương pháp LDA và so trùng thuộc tính từ gần tương đương nhau. Phương pháp so trùng thuộc tính từ sử dụng rất nhiều các đặc trưng ngữ cảnh như độ tương đồng từ (sử dụng thông tin về từ loại, kiểu thực thể, cây ngữ nghĩa,...), độ tương đồng ngữ cảnh xung quanh cặp thực thể; trong khi phương pháp LDA chỉ sử dụng thuộc tính là từ, trong khi đó, khi kết hợp hai phương pháp, kết quả cho ra là tốt nhất. Việc tích hợp này được thực hiện bằng cách cộng hai ma trận đo độ tương đồng của hai phương pháp trên.

Bài báo cũng thực hiện so sánh phương pháp lan truyền nhãn với phương pháp SVM và phương pháp SVM kết hợp với Bootstrapping [8], sử dụng cùng một bộ dữ liệu.

Bảng 3 tóm tắt kết quả của phương pháp lan truyền nhãn (ở đây sử dụng LDA kết hợp với so trùng thuộc tính từ), phương pháp SVM [8] và phương pháp SVM kết hợp Bootstrapping [8], thông qua độ đo F.

Bảng 3. So sánh độ đo F của ba phương pháp lan truyền nhãn sử dụng so trùng thuộc tính từ, SVM và SVM kết hợp Bootstrapping

	Chức vụ	**Sống ở**	**Làm việc cho**
Lan truyền nhãn	90.9	85.0	82.65
SVM	87.8	59.5	79.8
SVM + Bootstrapping	92.9	87.0	82.7

Bảng 3 cho thấy phương pháp lan truyền nhãn cho kết quả tốt hơn phương pháp SVM có giám sát, nhưng thấp hơn phương pháp SVM bán giám sát (SVM kết hợp với kỹ thuật Bagging-Bootstrapping). Xét về mặt thời gian thì phương pháp lan truyền nhãn nhanh hơn nhiều phương pháp SVM bán giám sát.

4. KẾT LUẬN

Trích rút mối quan hệ giữa các thực thể là bài toán còn mở đối với tiếng Việt. Để được một hệ thống trích rút quan hệ giữa các thực thể có độ chính xác cao, cần có tập dữ liệu huấn luyện lớn. Do tiếng Việt chưa có tập dữ liệu như vậy, bài báo đề xuất hệ thống học bán giám sát kết hợp với các đặc tính của ngôn ngữ Việt cho bài toán trích rút quan hệ giữa các thực thể. Ở đây, ta sử dụng phương pháp lan truyền nhãn. Bài báo đã đề xuất thử nghiệm ba phương pháp đo độ tương đồng giữa các câu bao gồm: phương pháp so trùng thuộc tính từ, phương pháp LDA và kết hợp hai phương pháp. Ngoài ra, bài báo còn đưa ra giải pháp giảm ảnh hưởng của các nhãn có tần suất xuất hiện lớn đến quá trình lan truyền nhãn. Thử nghiệm cho thấy kết quả cải tiến phương pháp lan truyền nhãn sau khi chuẩn hoá ma trận tốt hơn phương pháp cũ (chưa chuẩn hoá ma trận). Ngoài ra, phương pháp so trùng thuộc tính từ và phương pháp LDA cho ra kết quả tương tự nhau. Việc kết hợp giữa hai phương pháp đo độ tương đồng cho kết quả tốt hơn khi chưa kết hợp. Thử nghiệm cũng cho thấy phương pháp lan truyền nhãn cho kết quả tốt hơn phương pháp SVM, nhưng thấp hơn phương pháp bán giám sát trong [8]. Tuy nhiên, phương pháp lan truyền nhãn chạy nhanh hơn phương pháp SVM bán giám sát.

Trong tương lai, chúng tôi sẽ mở rộng thử nghiệm với các kiểu mối quan hệ giữa các thực thể khác để đánh giá tính chính xác của hệ thống đề xuất. Ngoài ra, do cấu trúc ngữ pháp của câu là thông tin quan trọng trong bài toán trích rút mối quan hệ giữa các thực thể, cần phải nghiên cứu cách tích hợp thông tin này vào hệ thống trích rút mối quan hệ giữa các thực thể nhằm tăng độ chính xác của hệ thống.

TÀI LIỆU THAM KHẢO

[1] R.C. Bunescu, R.J. Mooney, Subsequence kernels for relation extraction, *Proceedings of 19th Annual Conference on Neural Information Processing Systems (NIPS' 05)*, Vancouver,

British Columbia, Canada, 2005.

[2] A. Culotta, J. Sorensen, Dependency tree kernels for relation extraction, *Proceedings of the 42nd Meeting of the Association for ComputationalLinguistics (ACL'04), Main Volume*, Barcelona, Spain, July 2004 (423–429).

[3] M. E. Califf, and R. J. Mooney, Relational learning of pattern-match rules for information extraction, *Proceedings of the Sixteenth National Conferenceon Artificial Intelligence (AAAI-99)*, Orlando, Florida, July 1999 (328–334).

[4] J. Chen, D. Ji, L.C. Tan, Z. Niu, Relation extraction using label propagation based semi-supervised learning, *Proceeding of 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistic*, Stroudsburg, PA, USA, 2006 (129–136).

[5] M.B. David, Y.N. Andrew, I.J. Michael, Latent dirichlet allocation, *Journal of Machine Learning Research* **3** (January 2003) 993–1022.

[6] C. Giuliano, A. Lavelli, and L. Romano, Exploiting shallow linguistic information for relation extraction from biomedical literature, *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL '06)*, Trento, Italy, 2006.

[7] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, *Machine Learning* **42** (1-2) (2001) 177–196.

[8] C.R. Sam, T.H. Le, T.T.Nguyen, A.D.Le, and T.M.N. Nguyen, Semi-supervised learning for relation extraction in vietnamese text, *Proceedings of the Second Symposium on Information and Communication Technology (SoICT'2011)*, Hanoi, Vietnam, 2011 (100–105).

[9] M.V. Tran, V.V. Nguyen, T.U. Pham, T.O. Tran, Q.T. Ha, An experimental study of vietnamese question answering system, *Proceedings of the International Conference on Asian Language Processing*, Singapore, 2009 (152–155).

[10] Z. Xiaojin, and G. Zoubin, "Learningfrom Labeled and Unlabeled Data with Label Propagation", CMU CALD tech report CMU-CALD-02-107 (2002).

[11] Z. Zhang, Weakly supervised relation classification for information extraction*Proceedings of Thirteenth International Conference on Information and Knowledge Management*, Washington, DC, 2004.

[12] D. Zelenko, A. Aone, and A. Richardella, Kernel methods for relation extraction *Journal of Machine Learning Research* **3** (2003) 1083–1106.

[13] X. Zhu, "Semi-supervised learning literature survey (2008)", Technical Report 1530, University of Wisconsin Madison, 2008.

[14] Vietlex: http://www.vietlex.com.

Ngày nhận bài 26 - 11 - 2012
Nhận lại sau sửa ngày 12 - 03 - 2014