BÁO CÁO KỸ THUẬT SP8.4: XÂY DỰNG BỘ XÁC ĐỊNH NHÓM CỤM TỪ TIẾNG VIỆT

Thực hiện: Nguyễn Lê Minh, Cao Hoàng Trụ, Nguyễn Phương Thảo

Cộng Tác: Nguyễn Phương Thái, Trần Mạnh Kế

Tóm tắt

Việc phân nhóm các cụm từ tiếng Việt đóng một vai trò hết sức quan trọng trong các ứng dụng thực tế như tìm kiếm thông tin, trích chọn thông tin, và dịch máy. Để thực hiện tốt công việc này, chúng tôi đã khảo sát các phương pháp học máy được áp dụng thành công cho các ngôn ngữ bao gồm tiếng Trung, tiếng Nhật, và tiếng Anh. Sau khi khảo sát các phương pháp này chúng tôi đã lựa chọn phương pháp Conditional Random Fileds và Online Learning như là công cụ chính trong việc xây dựng một bộ phân cụm từ Tiếng Việt. Nghiên cứu về phân cụm từ tiếng Việt là khá mới mẻ đối với bài toán tiếng Việt. Do đó bài báo này không những trình bày việc thiết kế mô hình mà còn trình bày những nét cơ bản nhất hay những yếu tố chính liên quan đến khía cạnh ngôn ngữ trong bài toán phân cụm. Chúng tôi cũng đã khảo sát và xây dựng một tập các nhãn cũng như một bộ dữ liệu thử nghiệm để thực hiện việc đánh giá mô hình phân cụm rõ ràng hơn. Ngoài ra chúng tôi cũng trình bày các đánh giá dựa trên việc lựa chọn các thuộc tính phù hợp cho bài toán huấn luyện dãy. Bản báo cáo này bao gồm các phần: Phần 1 trình bày sự khảo sát bài toàn gộp nhóm (Chunking) cho tiếng Anh và tiếng Trung. Chúng tôi cũng trình bày các đặc thù của ngôn ngữ tiếng Việt. Phần 2 trình bày các kỹ thuật thông dụng được sử dụng trong bài toán phân cụm. Phần 3 trình bày mô hình của hệ thống. Phần 4 mô tả các thí nghiệm ban đầu khi thử nghiệm trên tập Vietnamese TreeBank (VTB). Phần 5 trình bày một số quan điểm của tác giả về định hướng nghiên cứu trong tương lai cũng như những nhận định về bài toán phân cụm từ Tiếng Việt.

Từ khóa: *Cụm từ, Phân tích cú pháp, Học máy cấu trúc*

1. Tổng quan

Bài toán phân nhóm cụm từ được nghiên cứu và được sử dụng trong nhiều ứng dụng thực tế như các hệ thống trích trọn thông tin, dịch máy, và tóm tắt văn bản. Bài toán phân cụm có thể hiểu là việc gộp một dãy liên tiếp các từ trong câu để gán nhãn cú pháp (ví dụ: cụm danh từ, cụm động từ). Việc nghiên cứu bài toán xác định nhóm cụm trên thế giới đã được thực hiện khá kỹ lưỡng và thành công cho nhiều ngôn ngữ bao gồm: tiếng Anh, tiếng trung, tiếng Nhật, tiếng Pháp. Gần đây các phương pháp học máy đã chứng tỏ sức mạnh và tính hiệu quả khi sử dụng cho bài toán xử lý ngôn ngữ tự nhiên bao gồm cả bài toán phân cụm. Đối với bài toán phân cụm tiếng Anh, tiếng Trung, phương pháp học máy đã cho kết quả rất tốt [1][2]. Với những lý do đó, chúng tôi đã nghiên cứu và vận dụng phương pháp học máy cho bài toán phân cụm tiếng Việt. Trước khi đi sâu và trình bày mô hình cụ thể, chúng tôi sẽ tóm tắt các nghiên cứu phận cụm cho tiếng Anh và tiếng Trung.

1.1. Nghiên cứu cụm từ tiếng Anh và tiếng Trung

Theo các kết quả đã được công bố ở SIGNL2001, các nhãn cụm được chia thành như sau (http://www.cnts.ua.ac.be/conll2000/ chunking/).
Ví dụ sau đây mô tả kết quả của bộ chunking tiếng Anh.
NP **He**] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September] .
Chúng ta có thể thấy các nhãn cụm từ bao gồm:

a) Noun Phrase (NP) Mô tả một cụm danh từ ví dụ Anh ấy là [“người bạn tốt của tôi”]
b) Verb Phrase (VP)

Mô tả một cụm động từ, là một dãy các từ bao gồm các động từ và các từ bổ trợ
Ví dụ: Chim [bay lên cao]

c) ADVP and ADJP

Tương đương với tiếng việt: cụm tính từ và cụm phó từ.

d) PP and SBAR

Tương đương với tiếng Việt: Cụm phó từ

e) CONJC

Tương đương với tiếng Việt: Cụm liên từ.

Quan sát các tập nhãn này chúng ta thấy rằng chúng hoàn toàn tương đồng với các khái niệm về tập nhãn trong tiếng Việt. Thêm nữa, hầu hết các ứng dụng như dịch máy, tóm tắt văn bản, trích lọc thông tin đều chủ yếu sự dụng các loại nhãn này. Để tìm hiểu một cách đúng đắn hơn chúng tôi cũng tham khảo thêm các nhãn của tiếng Trung bởi vì đây là ngôn ngữ châu Á có đặc tính cú pháp khá gần gũi đối với tiếng Việt. Cụ thể chúng tôi khảo sát chi tiết các hệ thống phân cụm từ tiếng Trung, dữ liệu, cũng như các loại nhãn. Chúng tôi tập trung vào tài liệu tham khảo [2].

Bảng 1. Các nhãn của Chiness chunking [2]

Kiểu nhãn	Khai báo
ADJP	Adjective Phrase
ADVP	Adverbial Phrase
CLP	Classifier Phrase
DNP	DEG Phrase
DP	Determiner Phrase
DVP	DEV Phrase
LCP	Localizer Phráe
LST	List Marker
NP	Noun Phrase
PP	Prepositional Phrase
QP	Quantifier Phrase
VP	Verb Phrase

Bảng 1 chỉ ra một số khác biệt của tiếng Trung, chẳng hạn LST, DEG, CLP. DP và QP. Chúng tôi khảo sát thêm đối với văn bản tiếng Việt cho các loại nhãn này thì thấy rằng không cần thiết có các tập nhãn đó.

1.2 Nhãn cụm từ

Sau khi nghiên cứu khảo sát ngôn ngữ tiếng Việt, chúng tôi xác định những tập nhãn cho việc phân cụm là hữu ích đối với bài toán này. Chúng tôi chỉ đưa ra những tập nhãn chuẩn và xuất hiện nhiều trong câu văn tiếng Việt. Từ đó, chúng tôi đưa ra bộ nhãn của việc phân cụm từ tiếng Việt bao gồm như sau:

Bảng 2. Nhãn cụm từ cho hệ phân cụm từ Việt

Tên	Chú thích
NP	Cụm danh từ
VP	Cụm động từ
ADJP	Cụm tính từ
ADVP	Cụm phó từ
PP	Cụm giới từ
QP	Cụm từ chỉ số lượng
WHNP	Cụm danh từ nghi vấn (ai, cái gì, con gì, v.v.)

WHADJP	Cụm tính từ nghi vấn (lạnh thế nào, đẹp ra sao, v.v.)
WHADVP	Cụm từ nghi vấn dùng khi hỏi về thời gian, nơi chốn, v.v.
WHPP	Cụm giới từ nghi vấn (với ai, bằng cách nào, v.v.)

Chú ý rằng bộ nhãn này đồng nhất với bộ nhãn trong Vietnames TreeBank (VTB) và sẽ còn được hiệu chỉnh trong tương lai. Cấu trúc cơ bản của một cụm danh từ như sau [8]:
<phần phụ trước> <danh từ trung tâm> <phần phụ sau>
Ví dụ: "mái tóc đẹp" thì danh từ "tóc" là phần trung tâm, định từ "mái" là phần phụ trước, còn tính từ "đẹp" là phần phụ sau.
(NP (D mái) (N tóc) (J đẹp))
Một cụm danh từ có thể thiếu phần phụ trước hay phần phụ sau nhưng không thể thiếu phần trung tâm.

Ký hiệu: VP
Cấu trúc chung:
Giống như cụm danh từ, cấu tạo một cụm động từ về cơ bản như sau:
<bổ ngữ trước> <động từ trung tâm> <bổ ngữ sau>
Bổ ngữ trước:
Phần phụ trước của cụm động từ thường là phụ từ.
Ví dụ:
"đang ăn cơm"
(VP (R đang) (V ăn) (NP cơm))

Ký hiệu: ADJP
Cấu trúc chung: Cấu tạo một cụm tính từ về cơ bản như sau:
<bổ ngữ trước> <tính từ trung tâm> <bổ ngữ sau>
Bổ ngữ trước:
Bổ ngữ trước của tính từ thường là phụ từ chỉ mức độ.
Ví dụ:
rất đẹp
(ADJP (R rất) (J đẹp))

Ký hiệu: PP
Cấu trúc chung :
<giới từ> <cụm danh từ>
Ví dụ :
vào Sài Gòn
(PP (S vào) (NP Sài Gòn))

Ký hiệu : QP

Cấu trúc chung :
Thành phần chính của QP là các số từ. Có thể là số từ xác định, số từ không xác định, hay phân số. Ngoài ra còn có thể có phụ từ như "khoảng", "hơn", v.v. QP đóng vai trò là thành phần phụ trước trong cụm danh từ (vị trí -2).
Ví dụ:

năm trăm
(QP (M năm) (M trăm))

Ví dụ:

hơn 200
(QP (R hơn) (M 200))

2. Phương pháp Phân Cụm Từ Tiếng Việt dùng CRFs và MIRA

Bài toán xác định nhóm cụm tiếng Việt được phát biểu như sau: Gọi X là câu đầu vào tiếng Việt bao gồm một dãy các từ tố kí hiệu $X=(X_1, X_2,\ldots, X_n)$. Chúng ta cần xác định $Y=(Y_1, Y_2, ..., Y_n)$ là một dãy các nhãn cụm từ (ví dụ: cụm danh từ, cụm động từ). Để giải quyết bài toán này chúng tôi quy về vấn đề học đoán nhận cấu trúc (ở đây là cấu trúc dãy) (có thể được thực hiện qua việc sử dụng các mô hình học máy [4][5]). Quy trình học được thực hiện bằng cách sử dụng một tập các câu đã được gán nhãn để huấn luyện mô hình học, và sử dụng mô hình này cho việc gán nhãn câu mới (không thuộc tập huấn luyện). Để thực hiện việc gán nhãn cụm cho câu tiếng Việt, chúng tôi sử dụng hai mô hình học máy câu trúc khá thông dụng bao gồm: CRFs [4] và Online Learning [5]. Cả hai phương pháp đối với bài toán này đều dựa trên giả thuyết các từ tố trong câu $X=(X_1, X_2,\ldots, X_n)$ tuân theo quan hệ của chuỗi Markov. Ở đây chúng tôi sử dụng mô hình Makov bậc 1. Về mặt lý thuyết chúng ta có thể dùng mô hình bậc cao hơn, tuy nhiên trong khuôn khổ dữ liệu hạn chế chúng tôi chỉ tập trung vào mô hình bậc 1, còn các bậc cao hơn sẽ được thí nghiệm ở công việc tương lai. Trước khi mô tả chi tiết mô hình phân cụm, chúng tôi giới thiệu mô hình học CRFs và Online Learning sau đây.

2.1 Mô hình học bằng CRFs

Mô hình CRFs cho phép các quan sát trên toàn bộ X, nhờ đó chúng ta có thể sử dụng nhiều thuộc tính hơn phương pháp Hidden Markov Model (HMM). Một cách hình thức chúng ta có thể xác định được quan hệ giữa một dãy các nhãn y và câu đầu vào x qua công thức dưới đây:

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_i \sum_k \lambda_k t_k(\boldsymbol{y}_{i-1}, \boldsymbol{y}_i, \boldsymbol{x}) + \sum_i \sum_k \mu_k s_k(\boldsymbol{y}_i, \boldsymbol{x}) \right) \quad (1)$$

Ở đây, x, y là chuỗi dữ liệu quan sát và chuỗi trạng thái tương ứng; t_k là thuộc tính của toàn bộ chuỗi quan sát và các trạng thái tại ví trí i-1, i trong chuỗi trạng thái; s_k là thuộc tính của toàn bộ chuỗi quan sát và trạng thái tại ví trí i trong chuỗi trạng thái. Thừa số chuẩn hóa $Z(\boldsymbol{x})$ được tính như sau:

$$Z(\boldsymbol{x}) = \sum_{y} \exp\left(\sum_{i} \sum_{k} \lambda_k t_k(\boldsymbol{y}_{i-1}, \boldsymbol{y}_i, \boldsymbol{x}) + \sum_{i} \sum_{k} \mu_k s_k(\boldsymbol{y}_i, \boldsymbol{x}) \right)$$

$\theta(\lambda_1, \lambda_2, ..., \mu_1, \mu_2 ..)$ là các vector các tham số của mô hình. Giá trị các tham số được ước lượng nhờ các phương pháp tối ưu LBFGS.

2.2. Huấn luyện mô hình trọng số bằng phương pháp MIRA

Trong bài báo này chúng tôi cũng triển khai việc sử dụng mô hình học Online Learning (Voted Perceptron) [5] cho bài toán phân cụm. Điểm mạnh của phương pháp này là tốc độ nhanh, dễ cài đặt, và cho hiệu quả khá cao đối với các bài toán đoán nhận cấu trúc, đặc biệt là dạng cấu trúc dãy như trong bài toán phân cụm. Thông thường chỉ sau khoảng 10 vòng lặp là thuật toán MIRA có thể hội tụ. Thuật toán MIRA là một trong những thuật toán Online Learning phổ biến và độ chính xác cho kết quả tương đương với CRFs trên nhiều bài toán khác nhau [5]. Do sự hiệu quả của phương pháp này, chúng tôi sẽ xem xét sử dụng thuật toán MIRA trong bài toán phân cụm từ Việt một cách hiệu quả nhất.

Lý do chọn MIRA

Các đặc tính của MIRA khiến nó phù hợp với bài toán phân cụm tiếng Việt sau đây:

1) Nó là phương pháp học máy phân biệt[1].
2) Phân lớp được chia thành nhiều bài toán con, trong số đó có bài toán học có cấu trúc bằng phân lớp tuyến tính. Phân tích phụ thuộc là bài toán học có cấu trúc, MIRA nằm trong số ít các phương pháp học máy giải quyết hiệu quả bài toán này.
3) Khi đã có mô hình, bước suy luận của MIRA dựa trên giải thuật Hildreth [5] giải bài toán quy hoạch bậc hai; không cần tới các giải thuật forward-backward, inside-outside phức tạp như CRFs hay các tính toán về phân phối và tối ưu phức tạp của CRFs [4].

Cách tiếp cận của MIRA

MIRA là online SVMs[2] nhờ dùng phép xấp xỉ. Chúng ta có thể so sánh phương pháp MIRA với phương pháp SVM một cách tóm tắt như hình 1.

SVMs cho bài toán học có cấu trúc	**MIRA** (mỗi lần cập nhật **w** ta chọn vectơ trọng số mới gần với vectơ cũ nhất)

[1] Thuật ngữ tiếng Anh là "discriminative learning"

[2] SVMs là viết tắt của "Support Vector Machines"

tìm min$\|\|\mathbf{w}\|\|$ với những $s(\mathbf{x},\mathbf{y}) - s(\mathbf{x},\mathbf{y}') \geq L(\mathbf{y},\mathbf{y}')$ cho $\forall (\mathbf{x},\mathbf{y}) \in T, \mathbf{y}' \in chunker(\mathbf{x})$	$\mathbf{w}^{(i+1)} = \text{argmin}_{\mathbf{w}^*}\|\|\mathbf{w}^* - \mathbf{w}^{(i)}\|\|$ với những $s(\mathbf{x}_t,\mathbf{y}_t) - s(\mathbf{x}_t,\mathbf{y}') \geq L(\mathbf{y}_t,\mathbf{y}')$ ứng với $\mathbf{w}^*$ cho $\forall \mathbf{y}' \in chunker(\mathbf{x}_t)$

Hình 1 So sánh MIRA và SVMs

Trong đó $L(\mathbf{y},\mathbf{y}')$ là hàm xác định độ sai sót của $\mathbf{y}'$ so với $\mathbf{y}$, tính bằng số mục từ trên $\mathbf{y}'$ có cung đi vào khác $\mathbf{y}$; $parses(\mathbf{x})$ là không gian tất cả các cây (tập các cụm) có thể ứng với câu $\mathbf{x}$. Chú ý $\mathbf{w}$ là vector trọng số tương ứng đối với mỗi thuộc tính trong không gian thuộc tính. Mỗi một giá trị trong $\mathbf{w}$ chỉ ra mức độ ảnh hưởng của thuộc tính tương ứng đối với tập dữ liệu huấn luyện. Mục tiêu của bài toán là tìm vector $\mathbf{w}$ phù hợp nhất để giảm thiểu độ sai sót khi dùng $\mathbf{w}$ cho việc phân cụm lại các câu trên tập huấn luyện khi so sánh chúng với cây phân tích chuẩn (cụm).

Dùng k-best MIRA xấp xỉ MIRA để tránh số nhãn tăng theo hàm mũ

Chỉ áp dụng ràng buộc về lề cho k c $\mathbf{y}'$ có $s(\mathbf{x},\mathbf{y}')$ cao nhất.

$\mathbf{w}^{(i+1)} = \text{argmin}_{\mathbf{w}^*}\|\|\mathbf{w}^* - \mathbf{w}^{(i)}\|\|$ với những $s(\mathbf{x}_t,\mathbf{y}_t) - s(\mathbf{x}_t,\mathbf{y}') \geq L(\mathbf{y}_t,\mathbf{y}')$ ứng với $\mathbf{w}^*$ cho những $\mathbf{y}' \in best_k(\mathbf{x}_t, \mathbf{w}^{(i)})$

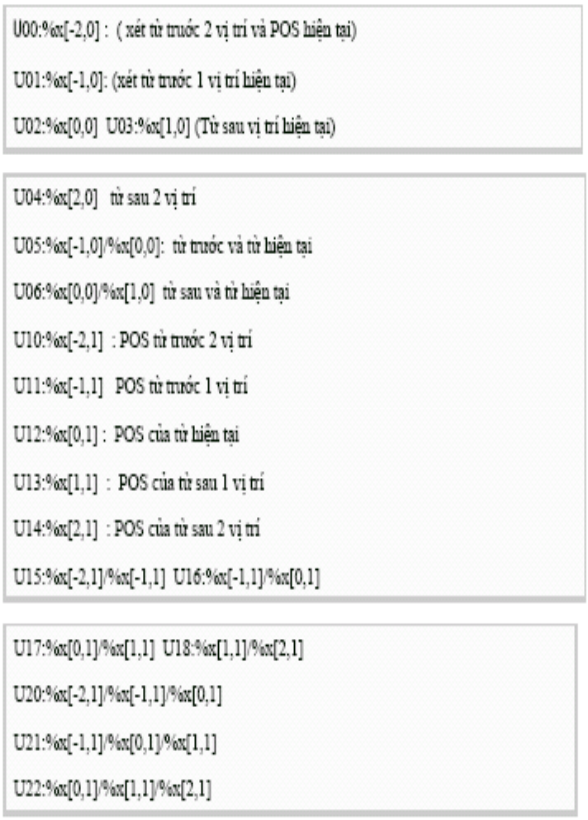
Hình 2 k-best MIRA

Hình 2 là k-best MIRA tổng quát, trong MST tác giả chỉ sử dụng k=1. Trong hệ thống hiện tại chúng tôi sử dụng k=1, khi dữ liệu lớn hơn chúng tôi sé thử nghiệm đối với các giá trị k khác nhau. Thông thường các kết quả nghiên cứu cho thấy k=5 hay k=10 thường đạt kết quả tốt nhất.

2.3 Thuộc tính

Trong cả 2 mô hình CRFs và Online Learning chúng tôi sử dụng chung một tập thuộc tính. Chúng tôi sử dụng các "template" sau đây để sinh ra các thuộc tính cho bài toán phân cụm từ: Các template được sử dụng để lấy các thông tin từ vựng (lexical), thông tin về từ loại (part of speech tagging) và thông tin về nhãn cụm từ. Ở trong bảng U00 là loại thuộc tính từ vựng (xét từ vựng ở trước 2 vị trí và POS hiện tại). Có thể xem chi tiết ở Bảng 3). Chúng tôi sử dụng các "template" này để sinh ra tập các thuộc tính dùng trong mô hình CRFs [4] và Online Learning [5]. Hiện tại thí nghiệm trên tập dữ liệu CONLL-2000 cho kết quả tương đương với các kết quả đã được công bố đối với bài toán phân cụm từ tiếng Anh [9] (cỡ vào khoảng 94% độ chính xác). Chúng tôi hy vọng tập thuộc tính này sẽ tương thích đối với bài toán gộp nhóm từ Việt. Trong phần thực nghiệm chúng tôi sẽ mô tả sự so sánh của hai phương pháp này trên cùng một tập dữ liệu chuẩn (i.e VTB corpus).

Bảng 3. Bảng thuộc tính của bài toán phân cụm từ Tiếng Việt

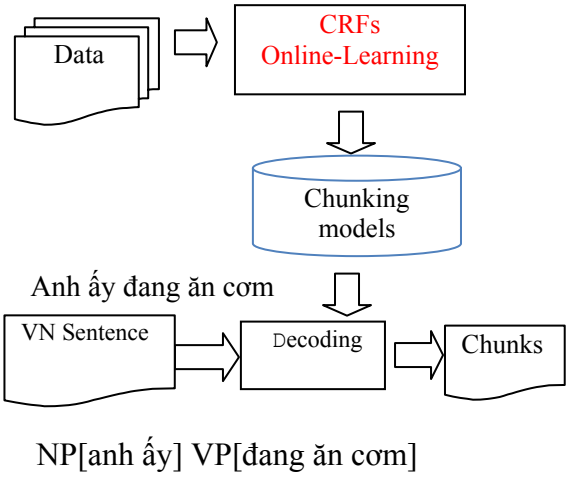


2.4 Thuật toán giải mã

Các mô hình sau khi ước lượng sẽ được sử dụng trong thuật toán giải mã. Thuật toán giải mã (decoding) cho cả hai mô hình CRFs và Online Learning đều dựa trên thuật toán quy hoạch động (dynamic programming), hay còn gọi là thuật toán Viterbi.

3. Sơ đồ hệ thống

Hình 3 mô tả mô hình của bộ gộp nhóm từ Việt, gồm hai thành phần chính. Thành phần huấn luyện từ tập dữ liệu có sẵn và thành phần gộp nhóm (decoding). Để huấn luyện chúng, tôi tập trung vào phương pháp CRFs và Online Learning. Mô hình CRFs được sử dụng khá thông dụng ở các bài toán phân cụm cho các ngôn ngữ khác. Phương pháp CRF cho Chunking Tiếng Anh đã thể hiện kết quả rất tốt [9], tuy nhiên nhược điểm của phương pháp này là thời gian tính toán tương đối chậm trong trường hợp số lượng dữ liệu huấn luyện lớn. Một mặt, ưu điểm của phương pháp Online Learning là thời gian huấn luyện khá nhanh và có thể áp dụng cho một số lượng dữ liệu huấn luyện lớn bởi vì bản chất của mô hình này là học tăng cường.



Hình 3. Mô hình hoạt động của bộ gộp nhóm từ Việt

Chúng tôi cũng khảo sát thêm các phương pháp học máy sử dụng trong việc gán nhãn tiếng Trung [3]. Kết quả cho thấy CRFs tốt hơn SVMs, tuy nhiên việc kết hợp các phương pháp khác nhau (kết hợp CRFs và SVMs) đem lại kết quả cao nhất. Trước hết chúng tôi chọn sử dụng phương pháp CRFs cho việc xây dựng công cụ hỗ trợ gộp nhóm mẫu. Công cụ này sẽ được sự dụng để huấn luyện trên một tập các dữ liệu bé sau đó dùng phương pháp học nửa giám sát (semi-supervised learning) để làm tăng số lượng của mẫu huấn luyện gộp nhóm từ trước khi đưa cho người dùng gán nhãn.

Để thực hiện được việc gán nhãn này, chúng tôi áp dụng mô hình chuyển đổi nhãn B-I-O trong bài toán chunking. Phương pháp này đã được khẳng định tính hiệu quả cao khi áp dụng với các ngôn ngữ khác nhau như Anh, Trung, Nhật, … [1][3]. Nội dung cụ thể của phương pháp này có thể được tóm tắt một cách hết sức đơn giản như sau: Với mỗi một từ trong một cụm, ta chia làm hai loại B-Chunk và I-Chunk. B-Chunk là từ đầu tiên của cụm từ đó và I-Chunk là các từ tiếp theo trong cụm. Ví dụ: (NP (N máy tính) IBM (PP của cơ quan))

Ta có thể chuyển thành dạng chuẩn như sau

Máy tính	N	B-NP
IBM	N	I-NP
của	-	B-PP
cơ quan	N	I-PP

Phương pháp học nửa giám sát (semi-supervised learning) được thực hiện bằng cách hết sức đơn giản dựa trên mô hình Boostraping. Gồm các bước sau đây:

Bước 1: Tạo bộ dữ liệu huấn luyện bé. Bước này được thực hiện bằng việc nhập liệu từ người chuyên gia Bước 2: Sử dụng mô hình CRFs để huấn luyện trên tập dữ liệu này. Bước 3: Cho tập test và sự dụng CRFs để gán nhãn Bước 4: Tạo bộ dữ liệu mới. Bộ dữ liệu mới được bổ sung kết quả từ việc gán nhãn tập test.

Hiện tại chúng tôi đang cần thêm dữ liệu huấn luyện từ nhóm TreeBank để huấn luyện mô hình gộp nhóm từ Việt. Nhóm dữ liệu Vietnamese TreeBank (VTB) sẽ chuyển giao dữ liệu cho chúng tôi trong thời gian tới với số lượng dữ liệu đủ lớn (10,000 câu) cho việc phân cụm từ tiếng Việt. Thêm nữa, các công cụ về phân đoạn từ, gán nhãn từ loại, cũng như từ điển sẽ hết sức cần thiết để xây dựng bộ phân cụm chuẩn. Hiện tại các tài nguyên này chưa hoàn toàn có sẵn. Bởi vậy, trong giai đoạn hiên nay, hệ thống của chúng tôi mới thử nghiệm bộ phân cụm từ tiếng Việt trên tập dữ liệu tương đối nhỏ do nhóm VTB cung cấp.

4. Kết quả thực nghiệm

4.1 Thử nghiệm phân cụm toàn bộ

Chúng tôi sử dụng dữ liệu từ VTB (Vietnamese Tree Bank) cho bài toán phân cụm sử dụng mô hình CRFs và mô hình học MIRA (Online Learning). Số lượng dữ liệu không nhiều (trước mắt nhóm VTB mới cung cấp xấp xỉ 2,000 câu được gán nhãn) nhưng kết quả thực nghiệm rất khích lệ. Trước hết nhiệm vụ của chúng tôi là trích lọc dữ liệu từ tập corpus VTB hiện có. Cách sinh dữ liệu chunking từ 1 cây VTB được mô tả như sau (bảng 4)

Bước 1. Lấy một cây trong VTB Bước 2. Duyệt đến nút lá trong cây và sinh ra các thành phần [Word, POS, Chunk] (Nhãn POS là nhãn của nút cha và nhãn Chunk là nhãn của nút "ông". Bước 3. Chuẩn hóa dữ liệu dưới dạng B-I-O. Chú ý rằng các nhãn ở mức chi tiết có thể được thay thế ở mức cao hơn (ví dụ NP-DOP có thể thay đổi thành NP)

<s> (S-TTL (VP(V-H Chào mừng) (NP-DOB(N-H Đại hội) (VP(VP(V-H thi đua) (VP(V-H yêu) (NP(N nước)))) (NP(NP-LOC(Y TP)(. .) (Y HCM))(M 2005))))) (. .)) </s>	Chào_mừng V-H VP Đại_hội N-H NP-DOB thi_đua V-H VP yêu V-H VP nước N NP TP Y NP-LOC HCM Y NP-LOC 2005 M NP . . O

Hình 4. Mô tả quá trình sinh ra dạng dữ liệu phân cụm dùng thuật toán ở Bảng 4.

Để chứng tỏ sự hiệu quả của các phương pháp, chúng tôi chia ngẫu nhiên 1,996 câu cho dữ liệu huấn luyện và 300 câu còn lại được dùng để đánh giá độ chính xác của chương trình. Sau 50 vòng lặp, mô hình CRFs cho kết quả hội tụ. Chúng tôi bước đầu đánh giá độ chính xác của phương pháp phân cụm đối với 300 câu khi thử nghiệm trên mô hình dùng 2,000 câu làm dữ liệu huấn luyện. Chúng tôi đánh giá dựa vào độ chính xác tương tự như phương pháp đánh giá của CONLL-2000 cho bài toán phân cụm từ tiếng Anh. Kết quả thực nghiệm được thể hiện như bảng dưới đây.

Bảng 5 Kết quả trên tập Vietnamese Tree Bank

Thuộc tính	**Độ chính xác (CRFs)**	**Độ chính xác (MIRA)**
Toàn bộ thuộc tính	91.33%	90.96%
Không dùng thuộc tính từ vựng	90.88%	89.02%
Không dùng bigram	91.72%	89.76%

Bảng kết quả thể hiện rằng mặc dù với một corpus nhỏ, nhưng chúng tôi đã thu được kết quả tương đối tốt (91.72%). Có thể lý giải tại sao hệ thống phân cụm từ của chúng tôi cho kết quả khá hoàn chỉnh. Đó là nhờ khả năng của các phương pháp học máy cấu trúc, và cách chọn thuộc tính của chúng tôi. Qua thí nghiệm cho thấy, hai phương pháp CRFs và MIRA đều cho độ chính xác với kết quả xấp xỉ nhau.

Tuy nhiên phương pháp CRFs cho kết quả cao hơn khi sử dụng toàn bộ thuộc tính mô tả trong bảng 3. Điều đó cho thấy cả MIRA lẫn CRFs đều có thể thích ứng với bài toán phân cụm tiếng Việt. Ngoài ra chúng tôi cũng so sánh thời gian huấn luyện của MIRA và CRF, kết quả cho thấy thời gian hội tụ của MIRA là nhanh hơn 30% so với phương pháp CRFs. Trong tương lai gần, chúng tôi sẽ kiểm định lại cả hai phương pháp CRF và MIRA trong đó một tập dữ liệu huấn luyện lớn hơn nhiều sẽ được sử dụng. Bảng 5 cũng thể hiện việc đánh giá sự ảnh hưởng các loại thuộc tính sử dụng trong việc huấn luyện. Cụ thể, thuật ngữ sử dụng toàn bộ thuộc tính có nghĩa là sử dụng toàn bộ thuộc tính đã khai báo trong bảng và "không dùng thuộc tính từ vựng" tương đương với việc chúng tôi không xét các từ vựng bao quanh từ cần lấy nhãn. Ở dòng thứ 3 trong bảng 5, "không dùng bigram" có nghĩa là chúng tôi không xét nhãn của cụm đứng trước vị trí cần xét. Kết quả thí nghiệm cho thấy "sử dụng toàn bộ thuộc tính", nhung không sử dụng bigram cho kết quả tốt nhất khi so sánh với các loại thuộc tính khác. Bảng 5 cũng cho thấy độ chính xác giảm khá nhiều khi chúng tôi không sử dụng thuộc tính này.

Bảng 6 thể hiện độ chính xác của phương pháp huấn luyện thay đổi theo số vòng lặp của phương pháp CRFs. Kết quả từ bảng 6 cho thấy cho đến vòng lặp thứ 10, thuật toán CRFs đã cho độ chính xác tương đương với các vòng lặp nhiều hơn. Điều đó cho thấy trong thực tế, để tiết kiệm thời gian huấn luyện, chúng ta chỉ cần huấn luyện trong 10 vòng lặp.

Bảng 6. Độ chính xác của CRFs thay đổi theo vòng lặp

Vòng lặp	Không dùng bigram	Không dùng từ vựng	Toàn bộ thuộc tính
10	91.35	90.69	91.16
20	91.25	90.91	91.21
30	91.45	90.84	91.09
40	91.52	90.74	91.52
50	91.72	90.88	91.33

Trong giai đoạn tiếp theo, sau khi có một số lượng dữ liệu và các kết quả của các công cụ như phân đoạn từ (SP8.2), gán nhãn từ loại (SP8.3), chúng tôi có thể thực hiện được các thí nghiệm một cách tốt hơn. Bảng 5 và 6 cho thấy với việc chỉ sử dụng một số lượng corpus với dung lượng bé, bộ phân cụm đã đạt đến một kết quả rất đáng khích lệ (91.92%). Trong tương lai gần chúng tôi sẽ thực hiện việc huấn luyện lại mô hình phân cụm sau khi có thêm corpus bổ sung từ nhóm dữ liệu VTB.

4.2 Thử nghiệm phân cụm danh từ

Với dữ liệu hiện tại để xây dựng một bộ công cụ gán nhãn từ loại đủ tốt cho việc sử dụng thực tế là chưa được. Do đó, chúng tôi tiến hành nghiên cứu xây dựng bộ nhận dạng cụm danh từ.

Chia tập dữ liệu ngẫu nhiên theo tỉ lệ 2:1, hai phần cho huấn luyện và một phần để kiểm thử mô hình. Thống kê về tập dữ liệu như sau:

	Số câu	Số từ	Số cụm danh từ
Dữ liệu huấn luyện	1812	38679	9912
Dữ liệu kiểm tra	905	19427	5021
Tổng	2717	58106	14933

kết quả tốt nhất thu được cho bài toán đóan nhận cụm danh từ khi chúng tôi sử dụng các mẫu thuộc tính sau:

Bảng 7: Các mẫu thuộc tính liên quan đến nhãn cú pháp

Mẫu thuộc tính unigram		Mẫu bigram
%x[0,1]	%x[-1,1]	%x[0,1]/%x[1,1]
%x[-2,1]	%x[-3,1]	%x[0,1]/%x[-1,1]
%x[1,1]	%x[2,1]	
%x[3,1]	%x[0,0]/%x[1,0]	
%x[1,0]/%x[2,0]	%x[-1,1]/%x[0,1]	
%x[-2,1]/%x[-1,1]	%x[-2,1]/%x[-1,1]/%x[0,1]	
%x[1,1]/%x[0,1]/%x[1,1]	%x[0,1]/%x[1,1]/%x[2,1]	

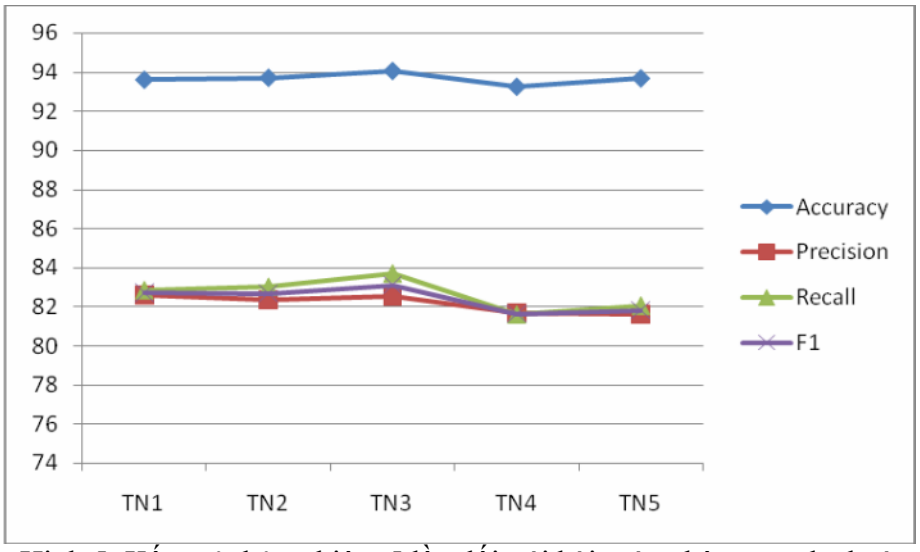
Bảng 8: Các mẫu thuộc tính kết hợp giữa từ vựng và nhãn cú pháp sử dụng cho bài toán gán nhãn cụm từ

Mẫu thuộc tính unigram		Mẫu bigram
%x[0,0]/%x[0,1]	%x[-1,0]/%x[-1,1]	%x[0,0]/%x[0,1]
%x[1,0]/%x[1,1]	%x[0,1]/%x[0,0]/%x[1,1]	%x[-1,0]/%x[-1,1]
%x[0,1]/%x[0,0]/%x[-1,1]		

Kết quả 5 lần thử nghiệm lựa chọn ngẫu nhiên tập dữ liệu học và kiểm tra được mô tả trong bảng và biểu đồ dưới:

	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
TN1	93.61	82.59	82.83	82.71
TN2	93.71	82.34	83.00	82.67
TN3	94.05	82.50	83.68	83.09
TN4	93.26	81.68	81.60	81.64
TN5	93.68	81.59	82.05	81.82

Bảng 9: Kết quả thử nghiệm 5 lần đối với bài toán phân cụm danh từ



Hinh 5: Kết quả thử nghiệm 5 lần đối với bài toán phân cụm danh từ

Từ kết quả trên có thể thấy, độ chính xác ở mức từ khá cao, trong cả 5 thực nghiệm đều > 93%, tuy nhiên ở mức cụm từ vẫn chưa tốt lắm. Sự chênh lệch kết quả này được hiểu như sau: giả sử một cụm danh từ gồm m từ, hệ thống gán nhãn m-1 nhãn đúng và chỉ một nhãn sai thì cụm từ này vẫn bị coi là gán nhãn sai. Chúng tôi tin rằng những nhược điểm này sẽ được khắc phục khi dữ liệu huấn luyện lớn.

5. Thảo luận

Quan sát tập dữ liệu tiếng Anh từ CONLL-2000 shared task và tiếng Trung (Chiness Tree Bank), chúng tôi nhận thấy các khái niệm về gán nhãn hầu như tương đồng với tiếng Việt. Dựa trên cơ sở đó và trên cơ sở tham khảo nhóm VTB (Viet Tree Bank) chúng tôi chọn tập nhãn như trình bày trong bài báo này.

Đồng thời, chúng tôi cũng đã xây dựng một bộ công cụ phân cụm từ tiếng Việt, sử dụng hai phương pháp học máy cấu trúc bao gồm CRFs và MIRA. Công cụ này đã được huấn luyện trên một tập dữ liệu VTB gồm khoảng 2,000 câu. Quá trình thử nghiệm cho thấy mô hình đề ra hoàn toàn tương thích với dữ liệu VTB. Mặc dầu với số lượng dữ liệu khoảng 2,000 câu, nhưng những kết quả của nó thể hiện rằng mô hình CRFs và Online Learning là các lựa chọn đúng đắn cho bài toán gộp nhóm cụm từ tiếng Việt. Đây là hai phương pháp kinh tế, đảm bảo cả về mặt thời gian lẫn độ chính xác. Các kết quả thu được đối với hệ thống gộp nhóm cụm từ tiếng Việt dùng dữ liệu chuẩn VTB thể hiện sự phù hợp khi vận dụng các phương pháp học máy cấu trúc (CRF++, MIRA). Chúng tôi hy vọng kết quả sẽ tốt hơn nữa khi thử nghiệm mô hình này với một lượng dữ liệu lớn hơn. Trong tương lai chúng tôi sẽ thử nghiệm bộ gộp nhóm từ Việt với dữ liệu cỡ 10,000 câu.

Lời cảm ơn

Nghiên cứu này được thực hiện trong khuôn khổ Đề tài Nhà nước "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt" mã số KC01.01/06-10.

TÀI LIỆU THAM KHẢO

[1] Erik F. Tjong Kim Sang and Sabine Buchholz, **Introduction to the CoNLL-2000 Shared Task: Chunking**. In: *Proceedings of CoNLL-2000* , Lisbon, Portugal, 2000.

[2] W. Chen, Y. Zhang, and H. Ishihara**. "An empirical study of Chinese chunking**", in Proceedings COLING/ACL 2006.

[3] Diệp Quang Ban (2005). *Ngữ pháp tiếng Việt*. NXB Giáo Dục.

[4] J. Lafferty, A. McCallum, and F. Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In the proceedings of International Conference on Machine Learning (ICML), pp.282-289, 2001

[5] Koby Crammer et al, "Online Pasive-Aggressive Algorithm", Journal of Machine Learning Research, 2006

[6] X.H. Phan, M.L. Nguyen, C.T. Nguyen, **"FlexCRFs: Flexible Conditional Random Field Toolkit**", http://flexcrfs.sourceforge.net, 2005

[7] Thi Minh Huyen Nguyen, Laurent Romary, Mathias Rossignol, Xuan Luong Vu, **"A lexicon for Vietnamese language processing**", Language Reseourse & Evaluation (2006) 40:291-309.

[8] Cao Xuân Hạo:"**Tiếng Việt: Sơ Thảo; Ngữ pháp chức năng"**, Nhà Xuất Bản Khoa Học

Xã
Hội, 1991

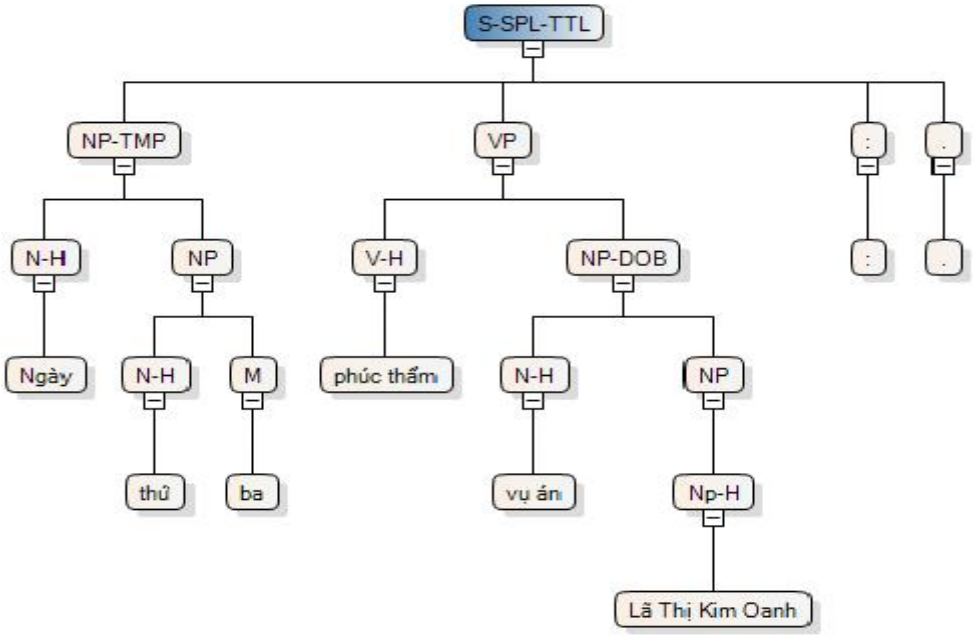
[9] F. Sha and F. Pereira **"Shallow Parsing with Conditional Random Fields",** *Proceedings of*
HLT-NAACL 2003 213-220 (2003)

Phụ lục 1: Phương pháp xây dựng dữ liệu gán nhãn từ loại cho cụm danh từ (NP)

Ví dụ về một câu trong tập dữ liệu đã được phân tích cú pháp: "**Ngày thứ ba phúc thẩm vụ án Lã Thị Kim Oanh : .**"

```
(S-SPL-TTL
    (NP-TMP(N-H Ngày)
            (NP(N-H thứ )(M ba)))
    (VP(V-H phúc thẩm)
           (NP-DOB(N-H vụ án)(NP(Np-H Lã Thị Kim Oanh))))
(: :)
(. .))
```

Biểu diễn dạng cây của ví dụ trên như hình vẽ sau:



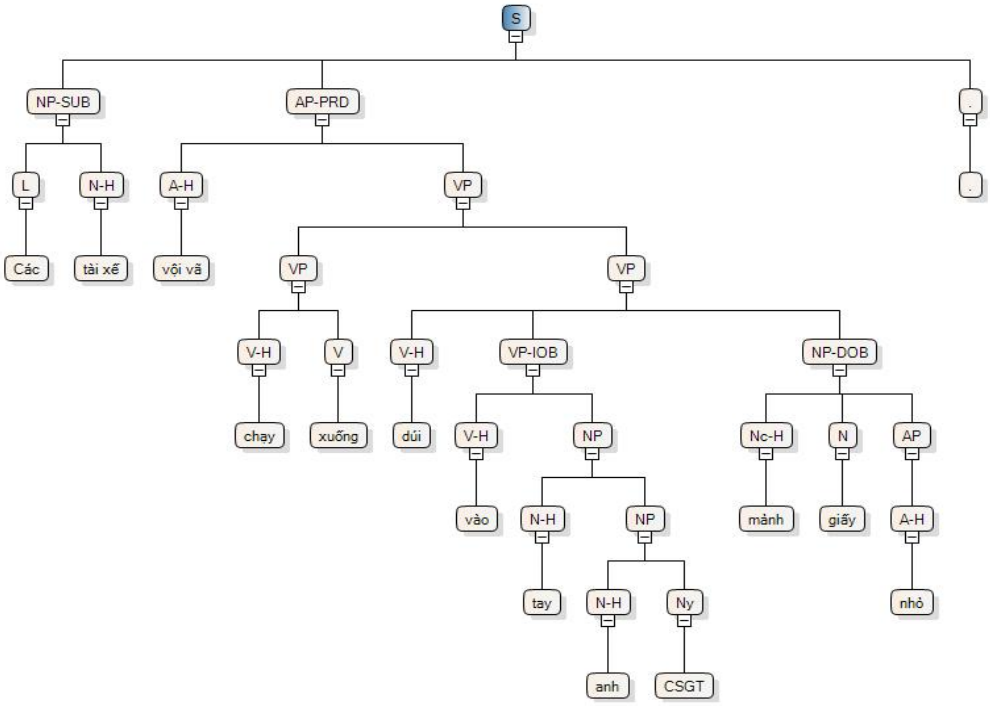
Các cụm danh từ được trích ra ở ví dụ trên là: "Ngày thứ ba", "vụ án Lã Thị Kim Oanh".

Để trích chọn được những cụm danh từ này, chúng tôi xây dựng một số quy tắc trích chọn dựa vào cây phân tích cú pháp. Về cơ bản sẽ lựa chọn các nhánh có nhãn là NP và các NP này là không lồng nhau. Tuy nhiên, do tính chất phức tạp của tiếng Việt, một số NP chứa phần bổ nghĩa sau nên ta cần xem xét thêm những tiêu chí khác. Ví dụ câu trên, nhánh NP-DOB chứa cụm NP bên trong, tuy nhiên nếu tách thành hai cụm là "vụ án" và "Lã Thị Kim Oanh" thì sẽ làm mất đi một phần ý nghĩa. Hơn nữa, theo cấu trúc của cụm danh từ tiếng Việt thì "vụ án Lã Thị Kim Oanh" cũng là một cụm danh từ, trong đó "vụ án" là danh từ trung tâm, "Lã Thị Kim Oanh" là danh từ bổ nghĩa. Với những trường

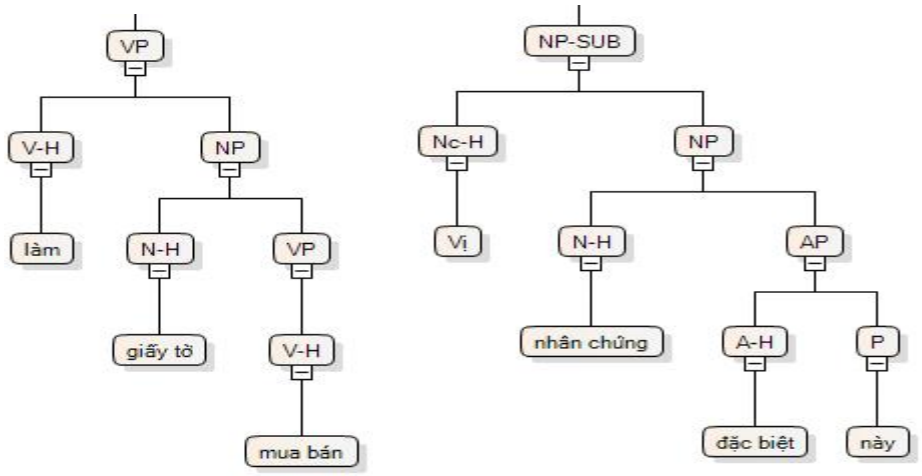
hợp này, chúng tôi bổ sung thêm các quy tắc khác liên quan đến độ sâu của nhánh, nhãn của các nhánh anh em với danh từ trung tâm.

Các tiêu chí để trích rút các cụm danh từ cơ sở một cách tự động từ tập dữ liệu đã phân tích cú pháp Viet Treebank như sau:

- Nếu nhánh NP có độ sâu là 1 thì cụm danh từ sẽ là toàn bộ nhánh NP đó. Ví dụ câu: "Các tài xế vội vã chạy xuống dúi vào tay anh CSGT mảnh giấy nhỏ". Trong câu này, nhánh NP đầu tiên có độ sâu bằng 1 nên cụm danh từ sẽ là toàn bộ nhánh NP này: "Các tài xế"



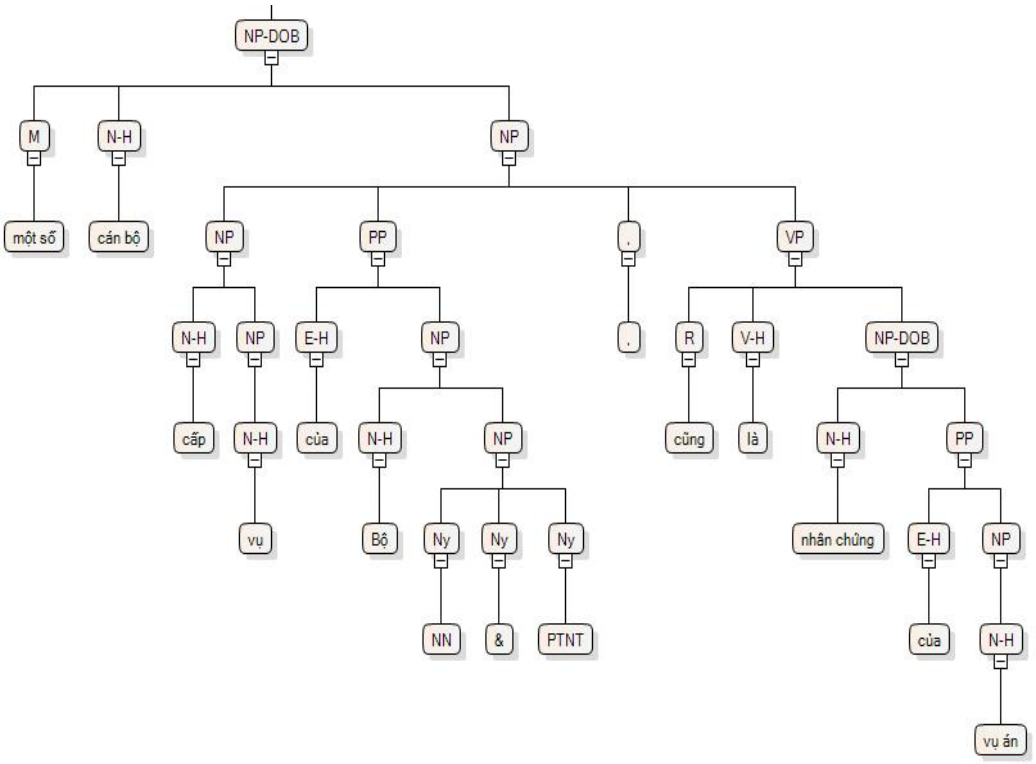
- Nếu nhánh NP có độ sâu bằng 2 thì cụm NP sẽ gồm phần đầu, danh từ trung tâm và phần sau, trong đó phần sau là các nhánh có nhãn khác PP (cụm giới từ) và SBAR. Vẫn xét ví dụ trên, hai nhánh NP có độ sâu bằng 2 được lựa chọn là "tay anh CSGT" và "mảnh giấy nhỏ". Hai cụm này chứa danh từ trung tâm, theo sau là cụm danh từ hoặc cụm tính từ (AP). Điều này phù hợp với cấu trúc cụm danh từ như đã trình bày ở phần trên. Một trường hợp khác cũng khá phổ biến trong tập dữ liệu là cụm danh từ cơ sở chứa cụm động từ (VP) đứng sau danh từ trung tâm:



Trong ví dụ này, "giấy tờ mua bán" là một cụm danh từ.

- Một số trường hợp đặc biệt đối với các nhánh NP có độ sâu bằng 3 nhưng vẫn được xét như một cụm danh từ cơ sở. Những trường hợp này, chúng tôi lựa chọn các nhánh NP có độ sâu bằng 3, chỉ gồm danh từ trung tâm và theo sau là một NP có độ sâu bằng 2. Ví dụ hình trên, "vị nhân chứng đặc biệt này" là một cụm danh từ cơ sở.

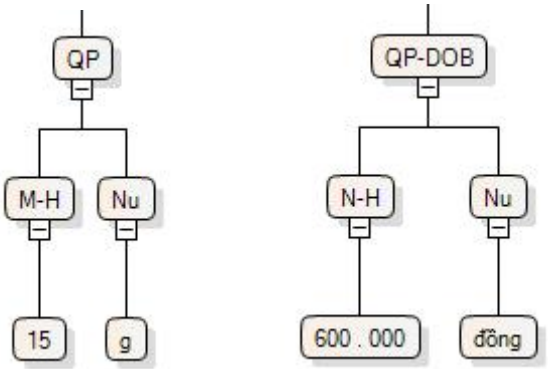
- Các trường hợp nhánh NP còn lại chúng tôi chỉ lựa chọn những từ ở độ sâu bằng 1 thuộc cụm danh từ cơ sở , ví dụ:



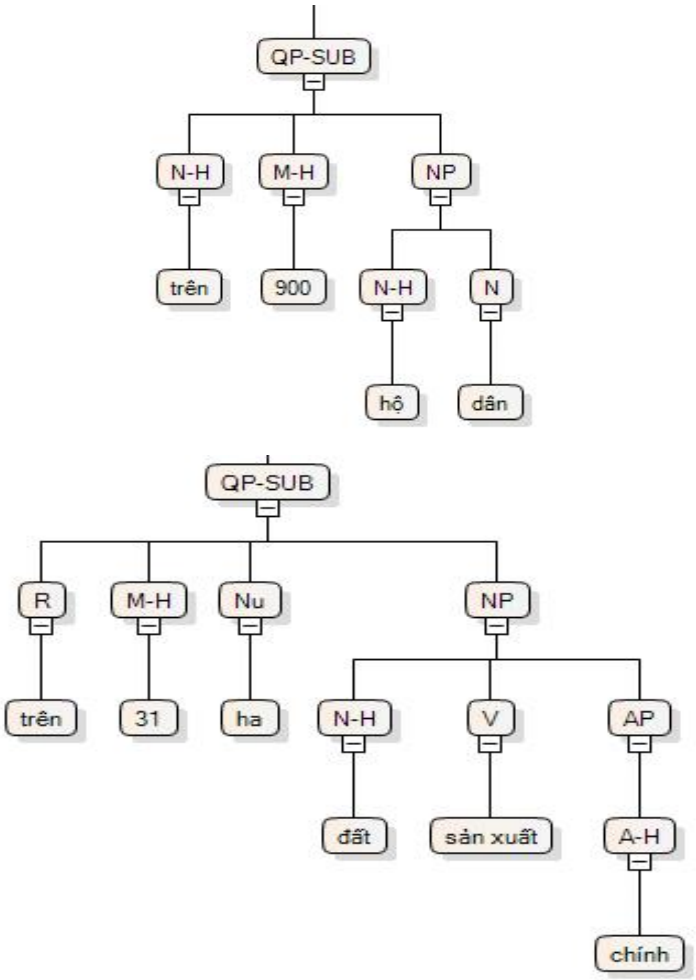
Trong trường hợp này, nhánh NP tương đối phức tạp. Lựa chọn các từ ở độ sâu bằng 1 của nhánh NP-DOB ta thu được cụm danh từ cơ sở "một số cán bộ", "nhân chứng".

- Ngoài ra, chúng tôi cũng xem xét các nhánh có nhãn QP là các cụm danh từ chỉ số lượng:

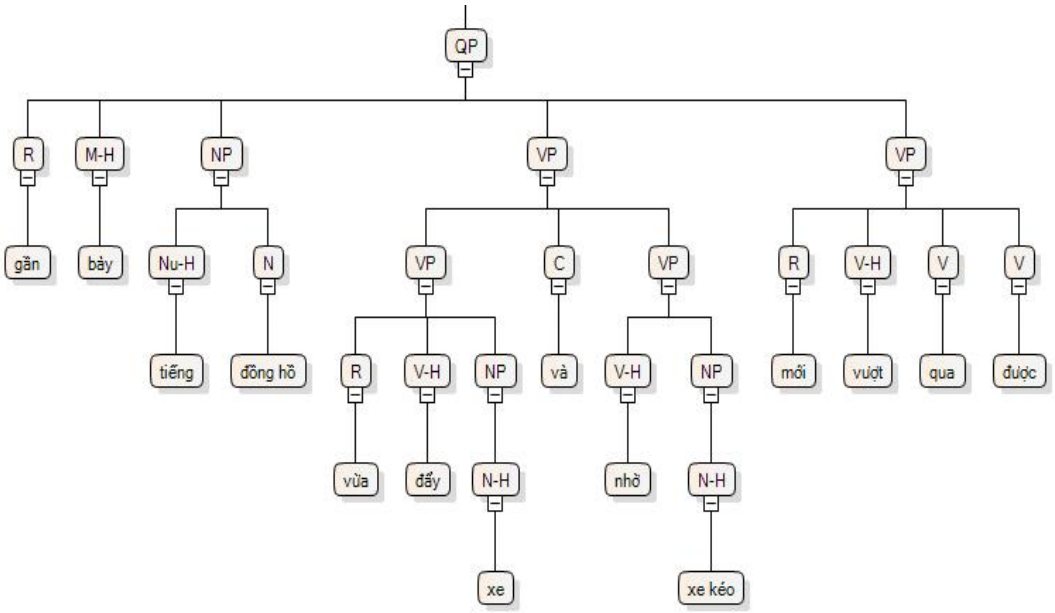
    - Nếu QP có độ sâu bằng 1 và chứa danh từ thì cụm danh từ sẽ là toàn bộ nhánh QP đó. Ví dụ: "15 g", "600.000 đồng"



- Nếu nhánh có nhãn QP và chứa NP có độ sâu nhỏ hơn hoặc bằng 2 thì cụm NP sẽ toàn bộ nhánh QP. Ví dụ: “trên 900 hộ dân” và “trên 31 ha đất sản xuất chính” là một cụm danh từ.



- Nếu nhánh có nhãn QP có độ sâu lớn hơn 3 nhưng chứa NP có độ sâu bằng 1 thì cụm danh từ sẽ gồm các từ thuộc nhánh QP có độ sâu là 1 và nhánh NP đó. Ví dụ, cụm danh từ được trích chọn trong hình vẽ dưới là “gần bảy tiếng đồng hồ”



Đối với các cụm chứa dấu phảy, hoặc liên từ "và", tùy từng trường hợp sẽ được phân tách thành hai cụm hoặc mở rộng thành một cụm. Ví dụ trong câu: "Khói mịt mù, Nguyên phải dùng tay và chân đập bể các cửa kính cho khói thoát ra ngoài …" thì "tay và chân" được coi là một cụm danh từ. Tuy nhiên, trong ví dụ sau: "Tuy nhiên, HĐXX vẫn quyết định công khai băng ghi hình và cuộc đối chất này để các luật sư, cử toạn theo dõi" thì "băng ghi hình và cuộc đối chất này" được tách thành hai cụm danh từ: "băng ghi hình", "cuộc đối chất này". Cụm "các luật sư, cử tọa" được coi là một cụm mặc dù có dấu phảy ở giữa.

Một số trường hợp đặc biệt, cụm danh từ chứa dấu nháy kép, ví dụ trong câu: "*Nhưng chứng cứ thu thập được vẫn còn mỏng, phải làm sao để bắt quả tang việc giao nhận tiền giữa các "trùm" trong đường dây này.",* các "trùm" được coi là một cụm danh từ cơ sở.

Để rút ra được những tiêu chí trên, chúng tôi đã phải nghiên cứu và tìm hiểu kỹ lưỡng về tập dữ liệu, đồng thời kết hợp với kết quả thực nghiệm để chỉnh sửa dần những tiêu chí này sao cho phù hợp và chính xác. Tuy nhiên, vì quá trình này được thực hiện tự động nên có nhiều trường hợp các cụm danh từ được trích chọn chưa chính xác. Do đó chúng tôi thực hiện rà soát lại tập dữ liệu này và sửa lại những trường hợp chưa chính xác một cách thủ công.