Phân cụm từ Tiếng Việt bằng phương pháp học máy cấu trúc

Nguyễn Lê Minh
Japan Advanced Institute of Science and Technology

Cao Hoàng Trụ
Ho Chi Minh City University of Technology

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. Bài báo 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.

1. Tổng quan

Bài toán phân cụm từ được nghiên cứu và được sử dụng trong nhiều ứng dụng thực tế như các hệ thống chích 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. Việc nghiên cứu bài toán phân cụm trên thế giới đã được thực hiện khá kỹ lưỡng 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. Đối với bài toán phân cụm tiếng Anh, tiếng Trung, etc. 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 ngôn ngữ 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 (Xem 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. Điều này hoàn toàn phù hợp với nhu cầu sử dụng các thông tin về ngữ pháp trong các sản phẩm ứng dụng tiếng Việt đòi hỏi tốc độ nhanh. 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 đã được phối hợp chặt trẽ với nhóm 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 1:

năm trăm
(QP (M năm) (M trăm))

Ví dụ 2:

hơn 200
(QP (R hơn) (M 200))

2. Phương pháp Phân Cụm Từ Tiếng Việt

Bài toán phân 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 (cụm danh từ, cụm động từ). Để giải quyết bài toán này chúng tôi quy về vấn đề học đoán nhận 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 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 khá thông dụng bao gồm: Conditional Random Fields [4] và Online Learning [5]. Cả 2 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. Trước khi đi vào 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, $\boldsymbol{x}$, $\boldsymbol{y}$ 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. Ví dụ:

$$s_i = \begin{cases} 1 \text{ nếu } \boldsymbol{x_i} = \text{“Bill” và } \boldsymbol{y_i} = \text{I\_PER} \\ 0 \text{ nếu ngược lại} \end{cases}$$

$$t_i = \begin{cases} 1 \text{ nếu } \boldsymbol{x_{i-1}} = \text{“Bill”, } \boldsymbol{x_i} = \text{”Clinton” và } \boldsymbol{y_{i-1}} = \text{B\_PER, } \boldsymbol{y_i} = \text{I\_PER} \\ 0 \text{ nếu ngược lại} \end{cases}$$

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. Pha học mô hình trọng số bằng phương pháp MIRA

Trong bài báo này chúng tôi cũng triển khai việc sự dụng mô hình học Online Learning (Voted Perceptron) [5] cho bài toán phân cụm. Lợi điểm 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 số lượng vòng lặp được sử dụng khoảng 10 vòng lặp là thuật toán có thể hội tụ. Thuật toán MIRA là một trong những thuật toán Online Learning phổ biến và 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 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].

[1] Thuật ngữ tiếng Anh là “discriminative learning”

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. Nó 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)
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 **y'** so với **y**, tính bằng số mục từ trên **y'** có cung đi vào khác **y**; $parses(\mathbf{x})$ là không gian tất cả các cây (tập các cụm) có thể ứng với câu ***x***.

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 ***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.

2.3 Thuộc tính

Trong cả 2 mô hình CRFs và Online Learning chúng tôi sử dụng chung một kiểu 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 sủ 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 dưới (Bảng 3).

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

U00:%x[-2,0] : (xét từ trước 2 vị trí và POS hiện tại)
U01:%x[-1,0]: (xét từ trước 1 vị trí hiện tại)
U02:%x[0,0] U03:%x[1,0] (Từ sau vị trí hiện tại)
U04:%x[2,0] từ sau 2 vị trí
U05:%x[-1,0]/%x[0,0]: từ trước và từ hiện tại
U06:%x[0,0]/%x[1,0] từ sau và từ hiện tại
U10:%x[-2,1] : POS từ trước 2 vị trí
U11:%x[-1,1] POS từ trước 1 vị trí
U12:%x[0,1] : POS của từ hiện tại
U13:%x[1,1] : POS của từ sau 1 vị trí
U14:%x[2,1] : POS của từ sau 2 vị trí
U15:%x[-2,1]/%x[-1,1] U16:%x[-1,1]/%x[0,1]
U17:%x[0,1]/%x[1,1] U18:%x[1,1]/%x[2,1]
U20:%x[-2,1]/%x[-1,1]/%x[0,1]
U21:%x[-1,1]/%x[0,1]/%x[1,1]
U22:%x[0,1]/%x[1,1]/%x[2,1]

Bảng 3. Bảng thuộc tính dùng cho việc phân cụm từ Tiếng Việt

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]. 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 cùng trên một tập dữ liệu.

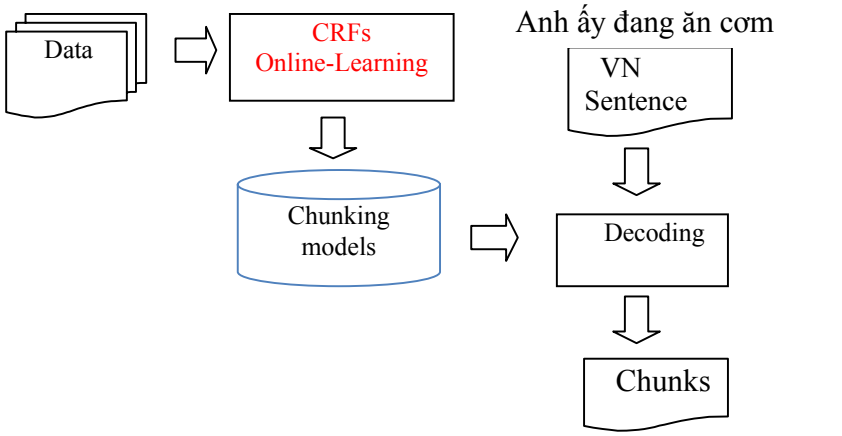
2.4 Thuật toán giải mã

Các mô hình sau khi ước lượng sẽ được sử dụng trong thuật toán giải mã. Thuật toán giải mã cho cả hai mô hình CRFs và Online Learning là như nhau và 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. Phương pháp

Conditional Random Fields 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. Chúng tôi có thể khắc phục nhược điểm này bằng cách sử dụng khả năng tính toán song song của bộ FlexCRFs. Cùng với FlexCRFs [6] nhiều kết quả sử dụng online learning method (Voted Perceptron) cũng cho kết quả tương đương với CRFs. Lợi thế 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. Trong thời gian này chúng tôi đã cài đặt mô hình chung cho cả 2 phương pháp dưới dạng mã nguồn mở.



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 này 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 mang tính hiệu quả cao cho các ngôn ngữ khác nhau Anh, Trung, Nhật, etc [1][3]. Nội dung cụ thể của phương pháp này có thể tóm tắt 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.

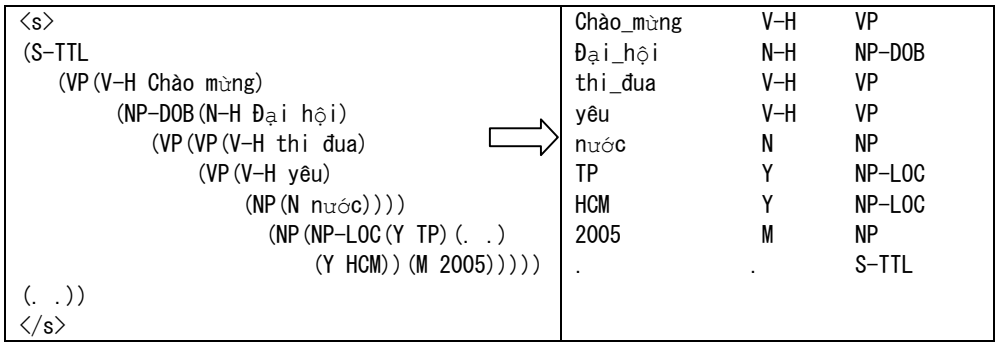
Hiện tại chúng tôi đang cần thêm dữ liệu huấn luyện từ nhóm TreeBank để huấn luyện mô hình gộp nhóm từ Việt. Nhóm dữ liệu Viet-TreeBank sẽ chuyển giao dữ liệu cho chúng tôi trong thời gian tới với số lượng dữ liệu đủ lớn (10,000 câu) cho việc phân cụm từ tiếng Việt. Thêm nữa, các tool về phân đoạn từ,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

4. Kết quả thực nghiệm

Chúng tôi sử dụng dữ liệu từ VTB (Viet Tree Bank) cho bài toán phân cụm sử dụng mô hình CRFs và mô hình học Online Learning. Số lượng dữ liệu không nhiều (trước mắt nhóm VTB mới cung cấp 260 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 chúng tôi sinh dữ liệu chunking từ 1 cây VTB được thực hiện như sau:

Bảng 4. Thuật toán sinh dữ liệu từ VTB

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



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

Để chứng tỏ sự hiểu quả của các phương pháp, chúng tôi chia ngẫu nhiên 215 câu làm dữ liệu huấn luyện và 45 câu được sử dụng như dữ liệu để đánh giá độ chính xác của chương trình.

Sau 45 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 45 câu khi thử nghiệm trên mô hình dùng 215 câu làm dữ liệu huấn luyện. Kết quả thực nghiệm được thể hiện như bảng dưới đây:

Bảng 5 Kết quả trên tập Viet Tree Bank

Thuộc tính	Độ chính xác (CRFs)	Độ chính xác (MIRA)
Toàn bộ features	63.55%	64.78%
Không dùng thuộc tính từ vựng	62.32%	61.82%
Không dùng bigram	65.27%	64.82%

Bảng kết quả thể hiện rằng mặc dù với một số lượng corpus nhỏ nhưng chúng tôi đã thu được kết quả rất đáng khích lệ. Có thể lý giả lý do tại sao kết quả chưa cao là bởi vì quá trình học máy với số lượng dữ liệu bé sẽ xuất hiện trường hợp dữ liệu thưa. Có nhiều hiện tượng ngữ pháp của tập dữ liệu kiểm định sẽ không xuất hiện trong tập huấn luyện. Qua thí nghiệm cũng cho thấy hai phương pháp CRFs và MIRA đều cho kết quả sấp xỉ nhau. Tuy nhiên phương pháp MIRA cho kết quả cao hơn khi sử dụng toàn bộ thuộc tính trong bảng 3. Như vậy phương pháp MIRA có thể thích ứng hơn với số lượng corpus nhỏ. Ngoài ra chúng tôi đánh giá 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 chúng tôi sẽ kiểm định lại cả hai phương pháp này sau khi sử dụng một tập corpus lớn hơn. Bảng 5 cũng thể hiện việc đánh giá các thuộc tính sử dụng trong việc huấn luyện. Củ thể, sử dụng toàn bộ features có nghĩa là sử dụng toàn bộ các thuộc tính đã khai báo trong bảng. Không dùng thuộc tính từ vựng có nghĩa là chúng tôi không xét các từ vựng bao quanh từ cần lấy nhãn. Không dùng bigram có nghĩa là chúng tôi không xét nhãn của cụm đứng trước. Có thể nói thuộc tính toàn bộ từ vựng có ảnh hưởng lớn nhất khi so sánh với các loại thuộc tính khác. Bảng 5 cho thấy độ chính xác giảm khá nhiều khi không thuộc tính này.

Trong giai đoạn tiếp theo, sau khi có một số lượng dữ liệu và các kết quả của các tool như phân đoạn từ, gán nhãn từ loại, chúng tôi có thể thực hiện được các thí nghiệm một cách tốt hơn. Bảng 5 cho thấy với việc sử dụng một số lượng corpus với dung lượng bé, bộ phân cụm đã đạt đến một kết quả rất đáng khích lệ (65.27%). 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.

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 báo cáo 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 VietTreeBank gồm khoảng 260 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 ban đầu không nhiều nhưng kết quả thể hiện mô hình CRFs và Online Learning là các lựa chọn đúng đắn. 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 phân cụm từ tiếng Việt dùng dữ liệu chuẩn VTB cho kết quả khả quan. Chúng tôi hy vọng kết quả sẽ tốt hơn nữa khi thử nghiệm mô hình này với một lượng dữ liệu lớn hơn.

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)