BÁO CÁO KỸ THUẬT SP8.4

Nguyễn Lê Minh

Việc phân nhóm các cụm từ tiếng Việt đóng một vai trò hết sức quan trọng cho các ứng dụng như tìm kiếm thông tin, trích chọn thông tin, và dịch máy. Để thực hiện tốt công việc này, chúng tôi đã tìm hiểu các phương pháp áp dụng thành công cho các ngôn ngữ tương tự tiếng Việt bao gồm tiếng Trung, tiếng Thái, và tiếng Anh. Sau khi khảo sát các phương pháp này chúng tôi đã lựa chọn phương pháp học Conditional Random Fileds và Online Learning, và ứng dụng cho tiếng bài toán phân cụm Việt. 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. Phần 2 trình bày các kỹ thuật dùng trong bài toán phân cụm từ tiếng Anh. Phần 3 trình bày mô hình của hệ thống. Phần 4 trình bày công cụ xây dựng dữ liệu. Phần 5 mô tả các kết quả đạt được.

1. Nghiên cứu cụm từ tiếng Anh và tiếng Trung

Sử dụng các tài liệu và kết quả đã được công bố ở SIGNL 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 của chúng ta trong các sản phẩm ứng dụng tiếng Việt. 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 và khá gần gũi đối với tiếng Việt. Cụ thể chúng tôi khảo sát chi tiết các hệ thống chunking tiếng Trung, dữ liệu, cũng như các loại nhãn. Chúng tôi tập trung vào tài liệu tham khảo [3].

Bảng 1. Các nhãn của Chiness chunking
(copy từ bài báo [3])

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. 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 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.

Một số giải nghĩa các nhãn cụm từ [Tham khảo chi tiết hơn nhóm VTB]

Cấu trúc cơ bản của một cụm danh từ như sau [1, trg24]:

<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

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. 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).

2.1 Mô hình học

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 và Online Learning. 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. 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 = 1 nếu x_{i-1}= "Bill", x_i="Clinton" và y_{i-1}=B_PER, y_i=I_PER

0 nếu ngược lại

Thừa số chuẩn hóa $Z(x)$ được tính như sau:

$$Z(x) = \sum_y \exp\left(\sum_i \sum_k \lambda_k t_k(y_{i-1}, y_i, x) + \sum_i \sum_k \mu_k s_k(y_i, x) \right)$$

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

Trong đề tài này chúng tôi cũng triển khai việc sự dụng mô hình học Online Learning (Voted Perceptron) 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.

Nội dung thuật toán Online Learning (voted Perceptron) có thể được trình bày một cách tóm tắt như hình 1 dưới đây:

Inputs:

- Một tập huấn luyện gồm các câu đã được gán nhãn ($w^i_{[1:n]}$, $t^i_{[1:n]}$), với $i = 1 \ldots n$.

- Tham số T là số lần lặp trên tập huấn luyện

- Mỗi đặc trưng cục bộ ϕ là một hàm ánh xạ một cặp history/tag đến một vector đặc trưng d chiều. Một biến toàn cục Φ được xác định thông qua ϕ theo công thức

$$\Phi_s(w_{[1:n]}, t_{[1:n]}) = \sum_{i=1}^{n} \phi_s(h_i, t_i)$$

Initialization: khởi tạo vectơ tham số $\bar{\alpha} = 0$.

Thuật toán:

Với $t = 1 \ldots T$, $i = 1 \ldots n$.

Dùng thuật toán Viterbi để tìm đầu ra của mô hình trên câu huấn luyện thứ *i* với tham số hiện thời:

$$z_{[1:n_i]} = \arg\max_{u_{[1:n_i]} \in \mathcal{T}^{n_i}} \sum_s \alpha_s \Phi_s(w^i_{[1:n_i]}, u_{[1:n_i]})$$

Với $\mathcal{T}^{n_i}$ là một tập tất cả các chuỗi nhãn có độ dài n_i.

Nếu $z_{[1..n]} \neq t^i_{[1:n]}$ thì ta sẽ cập nhật các tham số như sau:

$$\alpha_s = \alpha_s + \Phi_s(w^i_{[1:n_i]}, t^i_{[1:n_i]}) - \Phi_s(w^i_{[1:n_i]}, z_{[1:n_i]})$$

Ouput: Vector tham số $\overline{\alpha}$

Hình 1. Thuật toán Online Learning: Voted Perceptron

Thông thường số lượng vòng lặp T được sử dụng khoảng 10 vòng lặp là thuật toán có thể hội tụ. Thuật tóan Voted Perceptron là thuật toán Online Learning phổ biến nhất và cho kết quả tương đương với CRFs trên nhiều bài toán khác nhau.

2.2 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:

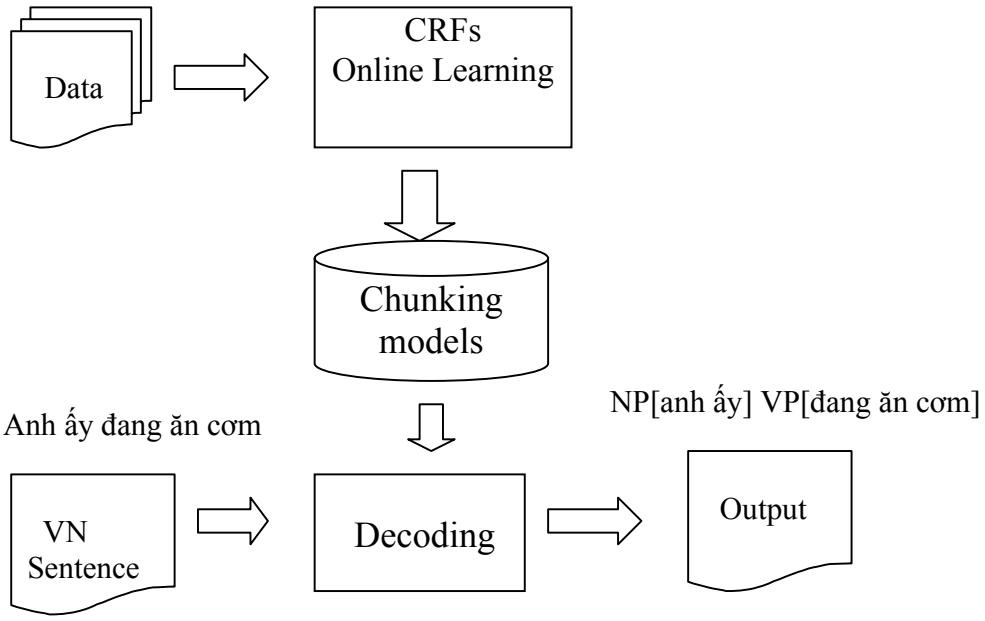
```
U00:%x[-2,0] :  ( xét từ trước 2 vị trí và POS hiện tại)
U01:%x[-1,0]: (xét từ trước 1 vị trí và POS hiện tại)
U02:%x[0,0]
U03:%x[1,0]
```

```
U04:%x[2,0]

U05:%x[-1,0]/%x[0,0]:

U06:%x[0,0]/%x[1,0]

U10:%x[-2,1]

U11:%x[-1,1]

U12:%x[0,1]q

U13:%x[1,1]

U14:%x[2,1]

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]
```

Chúng tôi sử dụng các template này để sinh ra tập các thuộc tính dùng trong mô hình CRFs và Online Learning. 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. Chúng tôi hy vọng bộ thuộc tính này sẽ tương thích đối với bài toán gộp nhóm từ Việt.

3. Sơ đồ hệ thống

Hình 2 mô tả mô hình của bộ gộp nhóm từ Việt. Bộ gộp nhóm gồm hai thành phần chính. Thành phần huấn luyện, từ tập dữ liệu có sẵn và thành phần gộp nhóm. 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 CRFs được sử dụng một cách thông dụng đối với Chunking Tiếng Anh và cho kết quả rất tốt, tuy nhiên nhược điểm của phương pháp này là thời gian tính toán tương đối chậm khi số lượng dữ liệu huấn luyện lớn. Chúng tôi có thể khắc phục nhược điểm này bằng khả năng tính toán song song của bộ FlexCRFs. Cùng với FlexCRFs [2] nhiều kết quả sử dụng online learning method (Voted Perceptron) cũng cho kết quả tương đương với CRFs. Lợi thế của phương pháp này là thời gian huấn luyện khá nhanh và không cần sử dụng đến tính toán song song. Trong thời gian này chúng tôi đã cài đặt mô hình chung cho cả 2 phương pháp dưới dạng mã nguồn mở. Quá trình cài đặt tiếp tục hoàn thiện hơn trong thời gian tới.



Hình 2. 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: Huấn luyện sử dụng CRFs. 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 đợi 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. 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 dừng ở dạng khuôn mẫu.

4. Xây dựng công cụ hỗ trợ làm dữ liệu

Để tiện cho việc xây dựng dữ liệu gán nhãn, chúng tôi đã tiến hành xây dựng bộ công cụ cho phép người dùng soạn thảo và gán các nhãn CHUNKING. Công cụ được viết bằng ngôn ngữ C++ và C.NET. Người dùng có thể đánh dấu nhãn bằng các thao tác đồ họa đơn giản, dữ liệu được biểu diễn dưới dạng XML. Văn bản dạng XML sẽ được chuyển thành dạng B-I-O bằng một chương trình đơn giản. Ngược lại, văn bản dạng B-I-O cũng được chuyển đổi sang dạng XML. Công cụ có thể sử dụng cho việc gán nhãn đối với các bài toán về phân đoạn từ hay nhận dạng tên riêng. Bộ xây dựng dữ liệu gán nhãn cũng được tích hợp với bộ đoán nhận CHUNK.

5. Kết quả

Trong giai đoạn hiện tại chúng tôi đã thực hiện được những nội dung sau đây:

① Hoàn thành bộ công cụ gán nhãn từ loại:

Chúng tôi đã xây dựng một bộ công cụ cho phép người dùng soạn và nhập dữ liệu. Bộ công cụ có thể áp dụng cho các bài toán gán nhãn từ loại.

② Xây dựng mô hình mẫu cho việc phân cụm

Chúng tôi xây dựng một mô hình mẫu cho việc phân cụm, mô hình dựa trên phương pháp học máy CRFs và Perceptron. Cả hai mô hình đều có thể tiến hành với số lượng dữ liệu lớn trong khuôn khổ thời gian cho phép.

③ Các tài liệu kỹ thuật về phương pháp

④ Xây dựng tập dữ liệu test: Quá trình đang tiến hành

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.

6. 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. Chúng tôi đã xây dựng một bộ công cụ hỗ trợ người làm dự liệu. Bộ công cụ này sẽ được huấn luyện trên một tập nhỏ các dữ liệu mẫu, sau đó sinh ra các dữ liệu gán nhãn tự động trước khi đưa cho người chuyên gia hiệu chỉnh. Giao diện của bộ công cụ đơn giản và dễ dùng và có thể cho phép chuyển đổi từ dạng B-I-O sang XML.

Phương pháp lựa chọn cho việc huấn luyện bao gồm CRFs và Online Learning (Perceptropn Structued). Đây là hai phương pháp kinh tế, đảm bảo cả về mặt thời gian lẫn độ chính xác. Các kết quả đối với gộp nhóm tiếng Anh và tiếng Trung đã khẳng định điều này. Thêm nữa, các kết quả các việc tương tự khác cho tiếng Việt [2][5][6] cũng đã khẳng định được thế mạnh của việc dùng CRFs cho việc nhận dạng tên riêng tiếng Việt.

Hiện tại những điều thiết yếu chúng tôi cần bao hàm như sau:

1. Cần dữ liệu huấn luyện
2. Cần các công cụ cho việc phân đoạn từ và gán nhãn từ loại

Chúng tôi hy vọng sẽ có sự giao tiếp chung giữa các tools này trong thời gian tới.

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 and LLL-2000*, Lisbon, Portugal, 2000.

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

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

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

[4] Cao Xuân Hạo:"**Tiếng Việt: Sơ Thảo; Ngữ pháp chức năng"**, Nhà Xuất Bản Khoa Học Xã Hội, 1991

[5] Tri Tran Q, et al . "**Named Entity Recognition in Vietnamese document"**, Progress in informatics No 4, pp 5-13 (2007)

[6] Pham Thi Xuan Thao, Tran Quoc Tri, Dinh Dien, Nigel Collier, "**Named entity recognition in Vietnamese using classifier voting",** ACM Transactions on Asian Language Information Processing (TALIP), Volume 6 , Issue 4 (December 2007)