BÁO CÁO THỰC HIỆN ĐỀ TÀI

Nguyễn Lê Minh
Cao Hoàng Trụ

1. Tên đề tài nhánh:

   SP8.4: **Hệ phân cụm từ Tiếng Việt**

2. Thời gian thực hiện:

   5/2007 - 10/2007

3. Người phụ trách: **Nguyễn Lê Minh**

4. Kế hoạch của đề tài

Nội dung cần thực hiện từ 05/2007 đến 05/2009

Nội dung thực hiện	Thời gian
Nghiên cứu tập nhãn cụm từ cho các ngôn ngữ trên thế giới. SP: 1 báo cáo.	12/2007
Xác định các quy tắc cho tập gán nhãn cụm từ chuẩn tiếng Việt. SP: 1 báo cáo.	12/2007
Xây dựng tài liệu hướng dẫn gán nhãn cụm từ. SP: 1 báo cáo.	03/2008
Nghiên cứu các phương pháp gộp nhóm tự động. SP: 1 báo cáo	03/2008
Xây dựng công cụ hỗ trợ gộp nhóm mẫu. SP phần mềm. Đầu vào: Kho văn bản đã gán nhãn từ loại, có hoặc chưa có gộp nhóm. Đầu ra: Kho văn bản đã được gộp nhóm từ mẫu.	05/2008
Xây dựng tập câu gộp nhóm mẫu dung lượng bé (2000 câu, rút từ kho ngữ liệu đã phân loại từ mẫu ở SP7.3). Gộp nhóm 100 câu cần 0,7 triệu.	10/2008
- Phần mềm gộp nhóm từ Việt tự động (vận dụng các phương pháp CRF's và SVM). Đầu vào: Văn bản đã gán nhãn từ loại. Đầu ra: Văn bản đã gộp nhóm từ.	12/2008
Công cụ đánh giá kết quả gộp nhóm tự động. Đầu vào: Kho văn bản đã gộp nhóm mẫu và tự động. Đầu ra: Độ đo đánh giá kết quả gộp nhóm tự động.	12/2008
Giao diện web cho phép cộng đồng nghiên cứu tra cứu, hiệu chỉnh kho ngữ liệu câu đã được gộp nhóm. Đầu vào: Kho ngữ liệu đã gộp nhóm từ. Giao diện web cho phép tra cứu và cập nhật kho ngữ liệu.	4/2009

5. Nội dung chi tiết

Để Xác định được bộ nhãn cho bài toán gộp nhóm từ Việt, trước hết chúng tôi đưa ra định nghĩa của nó như sau:

Đầu vào: Một dãy các từ tố tiếng Việt đã được gán nhãn từ loại

Đầu ra: Các từ tố được gộp nhóm thành dãy các cụm từ. Chu ý rằng các từ tố liền kề nhau mới có thể được gộp thành nhóm.

Bài toán gộp nhóm từ đã được thực hiện một cách khá chính xác cho ngôn ngữ tiếng Anh, tiếng Pháp và tiếng Trung. Tiếng Anh là ngôn ngữ được nghiên cứu rất kỹ lưỡng, tiếng Trung có dạng ngữ pháp khá tương đồng so với tiếng Việt. Đó là lý do chúng tôi chọn 2 ngôn ngữ này để khảo sát xây dựng hệ gộp nhóm từ Việt.

5.1 Nghiên cứu cụm từ tiếng Anh và tiếng Trung

Sử dụng các tài liệu và kết quả đã được công bố ở hiệp hội ngôn ngữ tự nhiên thế giới (ACL-SIGNLL), chúng tôi quan sát cách phân loại từ tiếng Anh thông qua các bộ nhãn chuẩn như (Xem http://www.cnts.ua.ac.be/conll2000/chunking/) cụm danh từ (NP), cụm động từ (VP), cụm tính từ (ADJP), cụm phó từ (ADVP), cụm giới từ (PP), etc.

Chẳng hạn: 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 [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. 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 cho bài toán phân cụm từ tiếng Việt như ở bảng 2 sau khi tham khảo các tài liệu về ngữ pháp tiếng Việt [3][4][5].

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.

Bảng 2 thể hiện các loại cụm từ chính thường xuất hiện như cụm danh từ (NP), cụm động từ, cụm tính từ, cụm phó từ, ...

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 :

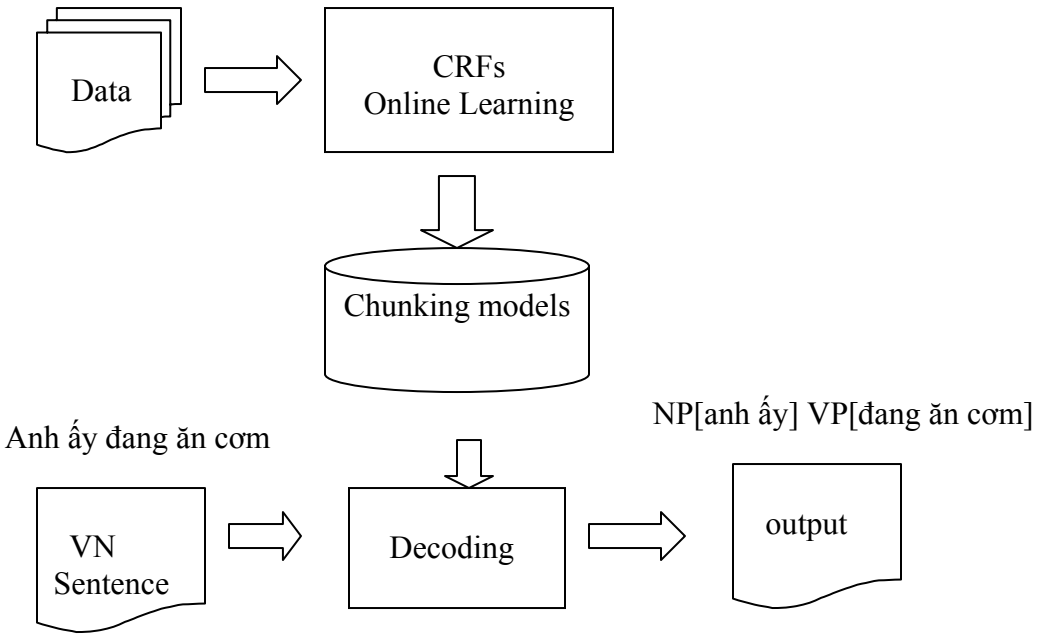
vào Sài Gòn
(PP (S vào) (NP Sài Gòn))

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))

5.2. **Phương pháp**



Hình 1. Mô hình hoạt động của bộ gộp nhóm từ Việt

Hình 1 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 CRFs được sử dụng một cách thông dụng đối với Chunking Tiếng Anh và cho kết quả rất tốt (state of the art), tuy nhiên một trong những nhược điểm của phương pháp này là thời gian chậm. 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 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. 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 có thể đem lại kết quả cao nhất. Chúng tôi đã có sẵn công cụ huấn luyện cho bài toán phân cụm (dưới dạng mã nguồn mở), có thể tham khảo tại địa chỉ (http://flexcrfs.sourceforge.net/). Với lý do độ chính xác giữa CRFs và SVM chênh lệch không nhiều, do đó chúng tôi chọn sử dụng phương pháp CRFs cho việc xây dựng công cụ hỗ trợ việc xây dựng tập dữ liệu gán nhãn cho bộ gộp nhóm mẫu. Công cụ này sẽ được sự dụng để huấn luyện trên một tập các dữ liệu bé khoảng xấp xỉ 1,000 câu, sau đó sẽ được 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, chung 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]. Ví dụ như câu tiếng Anh được phân cụm:

NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September] .

Có thể chuyển sang dạng B-I-O nhu sau:

```
He          PRP   B-NP
reckons     VBZ   B-VP
the         DT    B-NP
current     JJ    I-NP
account     NN    I-NP
deficit     NN    I-NP
will        MD    B-VP
narrow      VB    I-VP
to          TO    B-PP
only        RB    B-NP
#           #     I-NP
1.8         CD    I-NP
billion     CD    I-NP
in          IN    B-PP
September   NNP   B-NP
.           .     O
```

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

Ví dụ: [Cô ấy] thích [một cái áo màu xanh da trời]

Một	A	B-NP
cái áo	N	I-NP
màu xanh	N	I-NP
da trời	N	I-NP

Với định dạng dữ liệu này chúng ta có thể huấn luyện bằng các phương pháp "squence learning" như CRFs [2]. Vấn đề của chúng tôi hiện tại cần phải có bộ phân đoạn từ tiếng Việt và bộ phân loại từ tiếng Việt. Do đó kết quả của hệ thống gộp nhóm từ Việt sẽ phụ thuộc chặt trẽ vào 2 công cụ kể trên.

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. Tiếp đến, chúng tôi cần 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. Phương pháp lựa chọn cho việc huấn luyện bao gồm CRFs và Online Learning. 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][6] cũng đã khẳng định được thế mạnh của việc dùng CRFs cho nhận dạng tên riêng tiếng Việt.

Kế hoạch tiếp theo của chúng tôi là xây dựng công cụ trợ giúp người chuyên gia xây dựng tập dữ liệu huấn luyện cho việc phân cụm từ tiếng Việt.

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 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] Diệp Quang Ban (1989). Ngữ pháp tiếng Việt phổ thông (tập 2). NXB Đại học và Trung học chuyên nghiệp, Hà Nội.

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

[7] http://www.cnts.ua.ac.be/conll2000/chunking/