Gán nhãn từ loại tiếng Việt dựa trên các phương pháp học máy thống kê

Phan Xuân Hiếu[1], Lê Minh Hoàng[2], Nguyễn Cẩm Tú[3]
(1) Trường Khoa học thông tin, Đại học Tohoku, Nhật Bản
(2) Đại học Sư Phạm Hà Nội
(3) Đại học Công nghệ, Đại học Quốc gia Hà Nội

Tóm tắt
Trong những năm gần đây, do nhu cầu lớn về tìm kiếm, khai phá và xử lý thông tin tiếng Việt, các vấn đề xử lý tiếng Việt ngày càng nhận được nhiều quan tâm từ cộng đồng nghiên cứu trong và ngoài nước [Socbay, Bamboo, Xalo, VLSP, Biocaster, …]. Gán nhãn từ loại là một trong những bước quan trọng trong xử lý và khai phá dữ liệu tiếng Việt. Báo cáo này tổng kết một số kết quả nghiên cứu về gán nhãn tiếng Việt trong những năm gần đây. Bên cạnh đó, báo cáo còn đưa ra những so sánh, đánh giá chất lượng gán nhãn với hai phương pháp học máy thống kê là phương pháp cực đại hóa entropy (MaxEnt) và Conditional Random Fields. Những kết quả này sẽ góp phần định hướng cho việc xây dựng một hệ gán nhãn từ loại hiệu quả cho cộng đồng khai phá thông tin tiếng Việt nói chung và xử lý tiếng Việt nói riêng.

Từ khóa: Gán nhãn từ loại, tiếng Việt, học máy, Maximum Entropy, Conditional Random Fields, POS Tagging

1) Giới thiệu

Gắn nhãn từ loại là việc xác định các chức năng ngữ pháp của từ trong câu. Đây là bước cơ bản trước khi phân tích sâu văn phạm hay các vấn đề xử lý ngôn ngữ phức tạp khác. Thông thường, một từ có thể có nhiều chức năng ngữ pháp, ví dụ: trong câu "con ngựa đá đá con ngựa đá", cùng một từ "đá" nhưng từ thứ nhất và thứ ba giữ chức năng ngữ pháp là danh từ, nhưng từ thứ hai lại là động từ trong câu.

Một số hướng tiếp cận chính trong gắn nhãn từ loại tiếng Anh [Đinh Điền] bao gồm: gắn nhãn dựa trên mô hình Markov ẩn (HMM); các mô hình dựa trên bộ nhớ (Daelemans, 1996) ; mô hình dựa trên luật (Transformation Based Learning, Brill, 1995); Maximum Entropy; cây quyết định (Schmid, 1994a); mạng nơ-ron(Schmid, 1994b), v.v. Trong các hướng tiếp cận đó, phương pháp dựa trên học máy được đánh giá rất tốt.

Vấn đề gắn nhãn từ loại tiếng Việt có nhiều khó khăn [Nguyễn Huyền, Vũ Lương]. Ngoài khó khăn về đặc trưng riêng về ngôn ngữ, gắn nhãn từ loại tiếng Việt hiện còn rất thiếu các kho dữ liệu chuẩn như Brown hay Penn Treebank trong tiếng Anh cho quá trình so sánh đánh giá. Nghiên cứu này của nhóm chúng tôi hướng tới một số mục đích chính bao gồm: (1) khảo sát các công trình gắn nhãn từ loại tiếng Việt lien quan; (2) đánh giá khả năng áp dụng hướng tiếp cận gán nhãn từ loại tiếng Việt dựa trên 2 phương pháp học máy thống kê (Maximum Entropy và CRFs) - hướng tiếp cận được đánh giá rất tốt trong tiếng Anh; và (3) đánh giá mức độ ảnh hưởng của phân phối các nhãn trong kho dữ liệu đến chất lượng gán nhãn.

Phần còn lại của bài báo được tổ chức như sau: phần 2 tổng hợp một số công trình lien quan đến gắn nhãn từ loại tiếng Việt; phần 3 trình bày những tư tưởng chính của các phương pháp Maximum Entropy và CRFs; phần 4 là một số thử nghiệm và phân tích kết quả thử nghiệm; một số kết luận được rút ra trong phần 5 cũng là phần cuối của bài báo.

2) Gán nhãn từ loại tiếng Việt: các công trình liên quan

Trong nghiên cứu này, chúng tôi tập trung khảo sát hai công trình tách từ tiêu biểu: một của nhóm Đinh Điền và cộng sự; và hai là nhóm Nguyễn Huyền, Vũ Lương và cộng sự. Nhóm thứ nhất [Đinh Điền] xây dựng hệ thống gắn nhãn từ loại cho tiếng Việt dựa trên việc chuyển đổi và ánh xạ từ thông tin từ loại từ tiếng Anh. Cơ sở của hướng tiếp cận này nằm ở hai ý: (1) gắn nhãn từ loại trong tiếng Anh đã đạt độ chính xác cao (trên 97% cho độ chính xác ở mức từ) và (2) những thành công gần đây của các phương pháp gióng hàng từ (word alignment methods) giữa các cặp ngôn ngữ. Cụ thể, nhóm này đã xây dựng một tập ngữ liệu song ngữ Anh – Việt lên đến 5 triệu từ (cả Anh lẫn Việt). Sau đó thực hiện gắn nhãn từ loại cho bên tiếng Anh (dựa trên Transformation-based Learning – TBL [Brill 1995]) và thực hiện gióng hàng giữa hai ngôn ngữ (độ chính xác khoảng 87%) để chuyển chuyển thông tin về nhãn từ loại từ tiếng Anh sang tiếng Việt. Cuối cùng, dữ liệu tiếng Việt với thông tin từ loại mới thu được sẽ đuợc hiệu chỉnh bằng tay để làm dữ liệu huấn luyện cho bộ gắn nhãn từ loại tiếng Việt. Ưu điểm của phương pháp này là tránh được việc gắn nhãn từ loại bằng tay nhờ tận dụng thông tin từ loại ở một ngôn ngữ khác. Tuy vậy mức độ thành công của phương pháp này còn cần phải xem xét kỹ càng hơn. Ở đây, chúng tôi nêu ra vài nhận định chủ quan về những khó khăn mà phương pháp này gặp phải.

1) Sự khác biệt về tính chất ngôn ngữ giữa tiếng Anh và tiếng Việt rất đáng kể: sự khác biệt về cấu tạo từ, trật tự và chức năng ngữ pháp của từ trong câu làm cho việc gióng hàng trở nên khó khăn.

2) Lỗi tích lũy qua hai giai đoạn: (a) gắn nhãn từ loại cho tiếng Anh và (b) gióng hàng giữa hai ngôn ngữ: lỗi tích lũy cả hai giai đoạn này sẽ ảnh hưởng đáng kể tới độ chính xác cuối cùng.

3) Tập nhãn được chuyển đổi trực tiếp từ tiếng Anh sang tiếng Việt thiếu linh động và khó có thể là một tập nhãn điển hình cho từ loại tiếng Việt: do tính chất ngôn ngữ khác nhau, việc chuyển đổi nhãn từ loại của tiếng Anh sang tiếng Việt có phần áp đặt và sẽ không nhất quán hoàn toàn với tập nhãn được xây dựng dựa trên tính chất ngôn ngữ của tiếng Việt.

Do tác giả chỉ công bố kết quả dưới dạng ấn phẩm khoa học và không chia sẻ dữ liệu cụ thể nên chúng tôi không thể tìm hiểu kỹ hơn ở phần nội dung thực hiện và kết quả đạt được. Đây cũng là một khó khăn trong việc học tập, thừa kế lẫn nhau, và đi đến thống nhất một chuẩn chung, tạo tiền đề cho xử lý tiếng Việt sau này.

Nhóm thứ hai [Nguyen Huyen, Vu Luong] tiếp cận vấn đề này dựa trên nền tảng và tính chất ngôn ngữ của tiếng Việt. Nhóm này đề xuất xây dựng tập từ loại (tagset) cho tiếng Việt dựa trên chuẩn mô tả khá tổng quát của các ngôn ngữ Tây Âu, MULTEXT, nhằm

mô đun hóa tập nhãn ở hai mức: (1) mức cơ bản/cốt lõi (kernel layer) và (2) mức tính chất riêng (private layer). Mức cơ bản nhằm đặc tả chung nhất cho các ngôn ngữ trong khi mức thứ hai mở rộng và chi tiết hóa cho một ngôn ngữ cụ thể dựa trên tính chất của ngôn ngữ đó. Cụ thể, mức cơ bản của từ loại do nhóm này đề xuất bao gồm: danh từ (noun – N), động từ (verb – V), tính từ (adjective – A), đại từ (pronoun – P), mạo từ (determine – D), trạng từ (adverb – R), tiền-hậu giới từ (adposition – S), liên từ (conjunction – C), số từ (numeral – M), tình thái từ (interjection – I), và từ ngoại Việt (residual – X, như foreign words, ...). Mức thứ hai được triển khai tùy theo các dạng từ loại trên như danh từ đếm được/không đếm được đối với danh từ, giống đực/cái đối với đại từ, .v.v. Với cách phân loại này, chúng ta có thể co giãn hệ phân loại từ ở mức chung (cơ bản) hoặc cụ thể (chi tiết hóa) tương đối dễ dàng.

Tuy vậy, tập nhãn mà nhóm tác giả thứ hai đưa ra vẫn chưa thực sự tối ưu cho ngôn ngữ tiếng Việt. Hiện nay, hai tác giả chính của nhóm đang là thành viên chính trong việc xây dựng VietTreeBank trong khuôn khổ dự án VLSP. Qua trao đổi với nhóm xây dựng Viet Treebank, chúng tôi được biết các thành viên của nhóm này tiếp tục trao đổi để đưa ra một thiết kế tốt hơn, có hệ thống hơn với sự tham gia của nhiều nhóm liên quan. Những kết quả thống nhất về bộ thẻ và dữ liệu kết hợp với những nghiên cứu về phương pháp và ngôn ngữ sẽ là nền tảng cho xử lý và khai phá dữ liệu trên tiếng Việt.

3) Phương pháp Cực đại hóa Entropy (Maxent) và Conditional Random Fields (CRFs)

a) Phương pháp Maximum Entropy

Tư tưởng chính của Maximum Entropy là *"ngoài việc thỏa mãn một số rang buộc nào đó thì mô hình càng đồng đều càng tốt"*. Để rõ hơn về vấn đề này, ta hãy cùng xem xét bài toán phân lớp gồm có 4 lớp. Ràng buộc duy nhất mà chúng ta chỉ biết là trung bình 40% các tài liệu chứa từ "professor" thì nằm trong lớp ***faculty***. Trực quan cho thấy nếu có một tài liệu chứa từ "professor" chúng ta có thể nói có 40% khả năng tài liệu này thuộc lớp ***faculty***, và 20% khả năng cho các khả năng còn lại (thuộc một trong 3 lớp còn lại).
Mặc dù maximum entropy có thể được dùng để ượng lượng bất kì một phân phối xác suất nào, chúng ta xem xét khả năng maximum entropy cho việc gán nhãn dữ liệu chuỗi. Nói cách khác, ta tập trung vào việc học ra phân phối điều kiện của chuỗi nhãn tương ứng với chuỗi (xâu) đầu vào cho trước.

Các Ràng buộc và Đặc trưng

Trong maximum entropy, người ta dùng dữ liệu huấn luyện để xác định các ràng buộc trên phân phối điều kiện. Mỗi ràng buộc thể hiện một đặc trưng nào đó của dữ liệu huấn luyện. Mọi hàm thực trên chuỗi đầu vào và chuỗi nhãn có thể được xem như là đặc trưng $f_i(o,s)$. Maximum Entropy cho phép chúng ta giới hạn các phân phối mô hình lý thuyết gần giống nhất các giá trị kì vọng cho các đặc trưng này trong dữ liệu huấn luyện D. Vì thế người ta đã mô hình hóa xác suất $P(o \mid s)$ như sau (ở đây, o là chuỗi đầu vào và s là chuỗi nhãn đầu ra)

$$P(o \mid s) = \frac{1}{Z(o)} \exp\left(\sum_i \lambda_i f_i(o,s) \right) \qquad (2.1)$$

Ở đây $f_i(o,s)$ là một đặc trưng, λ_i là một tham số cần phải ước lượng và $Z(o)$ là thừa số chuẩn hóa đơn giản nhằm đảm bảo tính đúng đắn của định nghĩa xác suất (tổng xác suất trên toàn bộ không gian bằng 1) $Z(o) = \sum_c \exp \sum_c \lambda_i f_i(o,s)$

Một số phương pháp huấn luyện mô hình từ dữ liệu học bao gồm: IIS (improved iterative scaling), GIS, L-BFGS, and so forth.

b) Phương pháp Conditional Random Fields

CRFs là mô hình trạng thái tuyến tính vô hướng (máy trạng thái hữu hạn được huấn luyện có điều kiên) và tuân theo tính chất Markov thứ nhất. CRFs đã được chứng minh rất thành công cho các bài toán gán nhãn cho chuỗi như tách từ, gán nhãn cụm từ, xác định thực thể, gán nhãn cụm danh từ, etc.

Gọi $\mathbf{o} = (\mathbf{o_1}, \mathbf{o_2}, \ldots, \mathbf{o_T})$ là một chuỗi dữ liệu quan sát cần được gán nhãn. Gọi S là tập trạng thái, mỗi trạng thái liên kết với một nhãn $l \in L$. Đặt $\mathbf{s} = (\mathbf{s_1}, \mathbf{s_2}, \ldots, \mathbf{s_T})$ là một chuỗi trạng thái nào đó, CRFs xác định xác suất điều kiện của một chuỗi trạng thái khi biết chuỗi quan sát như sau:

$$p_\theta(\mathbf{s} \mid \mathbf{o}) = \frac{1}{Z(\mathbf{o})} \exp\left[\sum_{t=1}^{T} \sum_k \lambda_k f_k(s_{t-1}, s_t, \mathbf{o}, t)\right]. \qquad (1)$$

Gọi $Z(\mathbf{o}) = \sum_{s'} \exp\left(\sum_{t=1}^{T} \sum_k \lambda_k f_k(s'_{t-1}, s'_t, \mathbf{o}, t)\right)$ là thừa số chuẩn hóa trên toàn bộ các chuỗi nhãn có thể. f_k xác định một hàm đặc trưng và λ_k là trọng số liên kết với mỗi đặc trưng f_k. Mục đích của việc học máy với CRFs là ước lượng các trọng số này. Ở đây, ta có hai loại đặc trưng f_k : đặc trưng trạng thái (per-state) và đặc trưng chuyển (transition).

$$f_k^{(per-state)}(s_t, \mathbf{o}, t) = \delta(s_t, l) x_k(\mathbf{o}, t). \qquad (2)$$

$$f_k^{(transition)}(s_{t-1}, s_t, t) = \delta(s_{t-1}, l)\delta(s_t, l). \qquad (3)$$

Ở đây δ là Kronecker-δ . Mỗi đặc trưng trạng thái (2) kết hợp nhãn l của trạng thái hiện tại s_t và một vị từ ngữ cảnh - một hàm nhị phân $x_k(\mathbf{o},t)$ xác định các ngữ cảnh quan trọng của quan sát **o** tại vị trí t. Một đặc trưng chuyển (3) biểu diễn sự phụ thuộc chuỗi bằng cách kết hợp nhãn l' của trạng thái trước s_{t-1} và nhãn l của trạng thái hiện tại s_t.

Người ta thường huấn luyện CRFs bằng cách làm cực đại hóa hàm likelihood theo dữ liệu huấn luyện sử dụng các kĩ thuật tối ưu như L-BFGS. Việc lập luận (dựa trên mô hình đã học) là tìm ra chuỗi nhãn tương ứng của một chuỗi quan sát đầu vào. Đối với CRFs, người ta thường sử dụng thuật toán qui hoạch động điển hình là Viterbi để thực hiện lập luận với dữ liệu mới.

4) Thử nghiệm

a) Dữ liệu thử nghiệm

Để xây dựng các hệ thử nghiệm prototype, chúng tôi sử dụng cùng một tập dữ liệu được sử dụng trong [Nguyen Huyen, Vu Luong]. Tập dữ liêu này gồm khoảng 6400 câu và được gắn nhãn ở hai mức: mức 1 gồm 11 nhãn cơ bản và mức 2 gồm tập nhãn được chi tiết hóa. Từ tập nhãn chi tiết ở mức 2 có thể thu gọn về tập nhãn cơ bản ở mức 1 dễ dàng.

Các nhãn cơ bản bao gồm: N – danh từ; A – tính từ; V – động từ; P – đại từ; Cc – liên từ; Cm – giới từ; J – phụ từ (adverb); E – cảm từ; I – tình thái từ; Nn – số từ; X – không được phân loại. Ngoài ra còn 11 nhãn cho các dấu câu, ký tự đặc biệt, các dấu mở đóng ngoặc được gắn nhãn chính là ký tự đó. Tập nhãn mức cụ thể (mức 2) gồm 49 nhãn và 11 nhãn cho các dấu câu, ký tự đặc biệt như trên.

Để thử nghiệm và đánh giá, chúng tôi chia tập dữ liệu ra thành 4 phần bằng nhau (4 folds) và thực hiện huấn luyện lần lượt trên 3 phần và kiểm thử độ chính xác trên phần còn lại (thuật ngữ gọi là 4-fold cross validation test).

b) Lựa chọn đặc trưng

Để huấn luyện cho các hệ thống phân loại, chúng tôi trích chọn các đặc trưng từ dữ liệu như sau. Để phân lớp từ loại cho mỗi từ trong câu, chúng tôi sử dụng một cửa sổ trượt (sliding window) trải rộng từ 2 từ đi phía trước đến 2 từ đi phía sau của từ hiện tại. Và trong cửa sổ đó, các đặc trưng sau được lựa chọn:

1. Các từ trong cửa sổ từ vị trí -2, -1, 0 (vị trí hiện tại), +1, +2
2. Kết hợp của hai từ phía trước từ hiện tại: -2-1
3. Kết hợp của hai từ phía sau từ hiện tại: +1+2
4. Kết hợp từ phía trước và từ hiện tại: -10
5. Kết hợp của từ hiện tại và từ phía sau: 0+1
6. Từ hiện tại có gồm toàn chữ số hay không?
7. Từ hiện tại có chứa chữ số hay không?
8. Từ hiện tại có chứa ký tự "-" hay không?
9. Từ hiện tại có được viết hoa toàn bộ hay không?
10. Từ hiện tại có được viết hoa ký tự đầu tiên hay không?
11. Từ hiện tại có phải là một trong các dấu câu hay ký tự đặc biệt hay không? (nghĩa là các ký tự .,!,?,;,/,...)

Tập đặc trưng trên đây còn ở mức rất đơn giản do chúng tôi mới bắt đầu quá trình thử nghiệm. Đặc biệt là chúng tôi hoàn toàn chưa sử dụng đến thông tin tra cứu về nhãn từ loại từ từ điển. Trong thời gian tới chúng tôi sẽ thử nghiệm nhiều hơn nhằm tìm ra được những tập đặc trưng khả dĩ nhất.

c) Các thiết lập thử nghiệm

Nhóm thử nghiệm gán nhãn từ loại sử dụng hai công cụ FlexCRF và Jmaxent. Với mỗi phương pháp (Maxent hay CRFs), chúng tôi tiến hành 2 mức thử nghiệm: (1) gán nhãn mức 1 với 9 nhãn từ vựng tổng quát (N, V, J, ...) và 10 nhãn cho các loại kí hiệu; (2) gán nhãn mức 2 với 48 nhãn từ vựng chi tiết (Nt, Vtn, ...) và 10 nhãn cho các loại kí hiệu.

Các thiết lập tham số đối với FlexCRF và Jmaxent được cho như trong bảng sau:

FlexCRF	
order = 1	Thử nghiệm trên CRF bậc 1
f_rare_threshold=1	Bỏ các đặc trưng với tần xuất xuất hiện nhỏ hơn 1
Cp_rare_threshold=1	Bỏ các ngữ cảnh với tần xuất nhỏ hơn 1
init_lamda_val=0.5	Khởi tạo các tham số mô hình bằng 0.5
Jmaxent	
cpRareThreshold=3	Bỏ các ngữ cảnh với tần xuất xuất hiện nhỏ hơn 2
fRareThreshold=2	Bỏ các đặc trưng với tần xuất nhỏ hơn 3

d) Kết quả và đánh giá

Tổng hợp kết quả thực nghiệm gán nhãn từ vựng với Maxent và CRF

Table 4.1. Kết quả gán nhãn từ vựng mức tổng quát (11 nhãn từ vựng và 11 dấu câu) và mức cụ thể (48 nhãn từ vựng và 11 dấu câu)

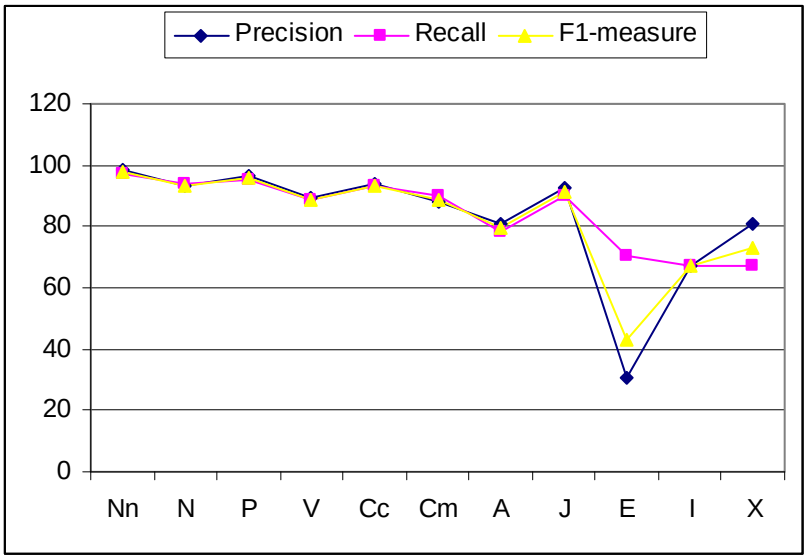
	F1-measure (tổng quát)		**F1-measure (cụ thể)**	
	Maxent	**CRFs**	**Maxent**	**CRFs**
Fold 1	91.33	91.55	83.82	***84.21***
Fold 2	91.18	91.56	83.82	84.12
Fold 3	90.22	***91.98***	82.04	84.01
Fold 4	91.00	91.59	83.70	83.84
Trung bình	**90.93**	**91.67**	**83.35**	**84.05**

Table 4.2. So sánh về thời gian giữa Maximum Entropy và Conditional Random Fields

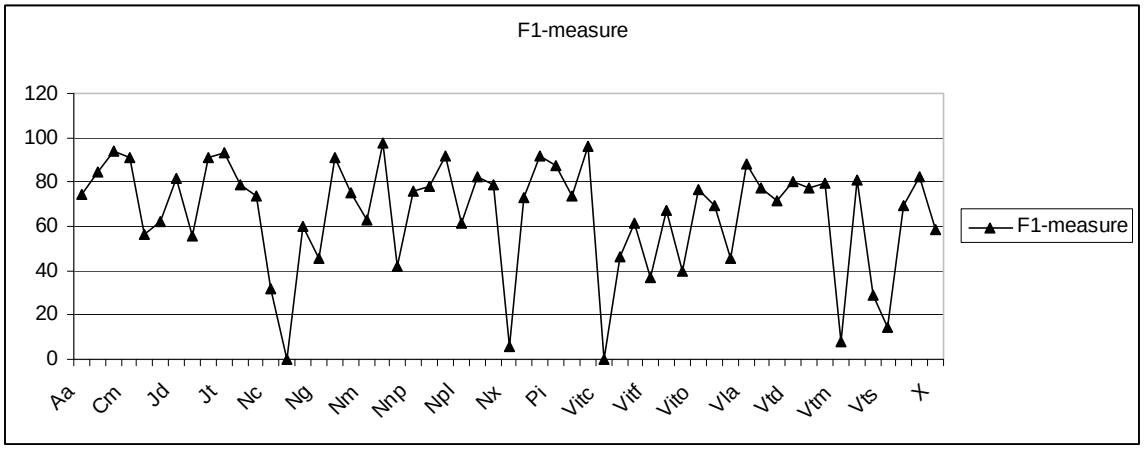
	Thời gian trung bình (s) (trên môt vòng lặp)		**Tối ưu ở vòng lặp thứ (trung bình)**	
	Mức tổng quát	**Mức cụ thể**	**Mức tổng quát**	**Mức cụ thể**
Maxent	~3	~8	~35	~40
CRFs	~48	~353	~36	~40

Table 4.3. So sánh về chất lượng gán nhãn với các nhãn từ loại khác nhau trong trường hợp tổng quát (thử nghiệm với fold3, mức tổng quát và CRFs)

Nhãn	**Độ chính xác**	**Độ hồi tưởng**	**F1-measure**
Nn	98.41	97.01	97.7
N	93.09	94	93.54
P	96.48	95.48	95.98
V	89.13	88.74	88.94
Cc	93.59	93.2	93.4
Cm	87.97	90.01	88.98
A	81.09	78.15	79.59
J	92.44	90.22	91.32
E	30.77	70.59	42.98
I	67.07	67.07	67.07
X	81	66.94	73.3



Hình 1. So sánh về chất lượng gán nhãn với các nhãn từ loại khác nhau trong trường hợp tổng quát (thử nghiệm với fold3, mức tổng quát và CRFs)



Hình 2. So sánh chất lượng gán nhãn với các nhãn từ loại trong trường hợp cụ thể (thử nghiệm với fold 1, mức cụ thể với CRFs)

e) Nhận xét

Thực nghiệm cho thấy tính khả quan của các hướng tiếp cận dựa trên CRFs và Maxent đối với bài toán gán nhãn từ vựng trong tiếng Việt. Dù CRFs mất nhiều thời gian hơn cho việc huấn luyện và gán nhãn nhưng nó đem lại cải thiện đáng kể chất lượng gán nhãn (trung bình tốt hơn Maxent 0.7%). Ưu điểm của cả 2 phương pháp trên là ta có thể tích hợp rất nhiều các đặc trưng phong phú, hữu ích từ dữ liệu. Dù chỉ với một số đặc trưng đơn giản (chưa tích hợp từ điển từ vựng, chưa dùng đến các biểu thức chính qui, ...), kết quả đạt được vẫn rất đáng chú ý (tốt nhất đạt 91.98% với mức tổng quát và CRFs). Thực nghiệm cũng khẳng định những nhận xét trong [Nguyen Huyen, Vu Luong], đó là việc gán nhãn ở mức cụ thể thường không tốt bằng gán nhãn ở mức tổng quát. Hình 1, và 2 so sánh chất lượng gán nhãn đối vời các nhãn trong hai mức tổng quát và cụ thể. Hình 1 cho thấy việc gán với các nhãn từ vựng quan trọng như N, V, P, A đạt được kết quả rất tốt so với các nhãn ít phổ biến hơn như E và I. Chúng tôi tin rằng với việc xây dựng một kho dữ

liệu có độ phủ lớn và cân bằng giữa các nhãn thì sự khác biệt này có thể được cải thiện đáng kể.

5) Kết luận

Tuy chưa thể tối ưu tập đặc trưng cho việc gắn nhãn từ vựng tiếng Việt dựa trên học máy. Chúng tôi thực sự hi vọng những nghiên cứu này sẽ đem lại lợi ích cho cộng đồng xử lý ngôn ngữ tiếng Việt. Những đóng góp của chúng tôi gồm 3 điểm chính: (1) tổng hợp lại một số công trình điển hình về gắn nhãn từ loại tiếng Việt; (2) khẳng định phương pháp CRFs đem lại chất lượng gắn nhãn tốt hơn so với Maxent; và (3) các nhãn có chất lượng gắn nhãn thấp thường là các nhãn ít phổ biến trong tập dữ liệu, từ đó rút ra được tầm quan trọng của việc xây dựng một kho dữ liệu có độ phủ tốt và có phân phối không quá lệch trên tất cả các nhãn từ vựng.

Lời cảm ơn

Nghiên này là một phần của dự án "*Xây dựng các sản phẩm tiêu biểu và thiết yếu về xử lý tiếng nói và văn bản tiếng Việt*" – một đề tài nghiên cứu khoa học và phát triển công nghệ được đầu tư bởi Bộ Khoa học & Công nghệ, Việt Nam. Chúng tôi xin gửi lời cảm ơn tới chủ nhiệm dự án, các bên liên quan, và các cấp quản lý đã hỗ trợ và tạo điều kiện cho chúng tôi thực hiện nghiên cứu này.

Tài liệu tham khảo

Dien Dinh and Kiem Hoang, POS-tagger for English-Vietnamese bilingual corpus. HLT-NAACL Workshop on Building and using parallel texts: data driven machine translation and beyond, 2003.

Thi Minh Huyen Nguyen, Laurent Romary, and Xuan Luong Vu, A Case Study in POS Tagging of Vietnamese Texts. The 10th annual conference TALN 2003.

Thi Minh Huyen Nguyen, Laurent Romary, Mathias Rossignol, and Xuan Luong Vu, A lexicon for Vietnamese language processing. Language Resources and Evaluation, 2007.

Nguyễn Thị Minh Huyền, Vũ Xuân Lương, Lê Hồng Phương, "Sử dụng bộ gán nhãn từ loại xác suất QTAG cho văn bản tiếng Việt", ICT 2003

Nguyễn Quang Châu, Phan Thị Tươi, Cao Hoàng Trụ, Gán nhãn Từ loại cho tiếng Việt dựa trên văn phong và tính toán xác suất, Tạp chí phát triển KH&CN, Tập 9, số 2 năm 2006

Phan, X.H, "JTextPro: A Java-based Text Processing Toolkit", http://jtextpro.sourceforge.net/

Xuan-Hieu Phan, Le-Minh Nguyen, and Cam-Tu Nguyen, "FlexCRFs: Flexible Conditional Random Field Toolkit", http://flexcrfs.sourceforge.net, 2005.