GÁN NHÃN TỪ LOẠI CHO TIẾNG VIỆT DỰA TRÊN VĂN PHONG VÀ TÍNH TOÁN XÁC SUẤT

Nguyễn Quang Châu [(1)], Phan Thị Tươi [(2)], Cao Hoàng Trụ[(2)]
(1) Trường Đại học Công Nghiệp Tp.HCM
(2) Trường Đại học Bách Khoa, ĐHQG- HCM
(Bài nhận ngày 09 tháng 12 năm 2006)

TÓM TẮT: *Xác định từ loại chính xác cho các từ trong văn bản tiếng Việt là vấn đề rất quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên.Việc xác định này sẽ hỗ trợ cho việc phân tích cú pháp các văn bản, góp phần giải quyết tính đa nghĩa của từ, và trợ giúp các hệ thống rút trích thông tin hướng đến ngữ nghĩa, v.v... Bài báo này trình bày một hướng tiếp cận cho bài toán gán nhãn từ loại trong văn bản tiếng Việt trên cơ sở vận dụng các mô hình thống kê dựa vào kho ngữ liệu, từ điển, cú pháp và ngữ cảnh. Đồng thời trong quá trình phát triển hệ thống ứng dụng, do chưa có kho ngữ liệu dành cho mục đích nghiên cứu về xử lý ngôn ngữ tự nhiên tiếng Việt, chúng tôi cũng đã xây dựng có tính kế thừa [1][4] được một kho ngữ liệu gồm gần 75.000 từ tiếng Việt, và một từ điển gồm 80.000 mục từ, để phục vụ cho vấn đề nghiên cứu này.*
Từ khóa : Tiếng Việt, từ loại, gán nhãn từ loại, văn phong, từ điển, kho ngữ liệu, thống kê, mô hình Markov, thuật toán Viterbi, rút trích thông tin.

I. GIỚI THIỆU

Một trong các vấn đề nền tảng của ngôn ngữ tự nhiên là việc phân loại các từ thành các lớp từ loại dựa theo thực tiễn hoạt động ngôn ngữ. Mỗi từ loại tương ứng với một hình thái và giữ một vai trò ngữ pháp nhất định. Các công cụ chú thích từ loại hay công cụ gán từ loại cho từ có thể thay đổi tuỳ theo quan niệm về đơn vị từ vựng và thông tin ngôn ngữ cần khai thác trong các ứng dụng cụ thể. Mỗi từ trong một ngôn ngữ nói chung có thể gắn với nhiều từ loại, và việc giải thích đúng nghĩa một từ phụ thuộc vào việc nó được xác định đúng từ loại hay không. Công việc gán nhãn từ loại cho một văn bản là xác định từ loại của mỗi từ trong phạm vi văn bản đó. Khi hệ thống văn bản đã được gán nhãn, hay nói cách khác là đã được chú thích từ loại thì nó sẽ được ứng dụng rộng rãi trong các hệ thống tìm kiếm thông tin, trong các ứng dụng tổng hợp tiếng nói, các hệ thống nhận dạng tiếng nói cũng như trong các hệ thống dịch máy.

Đối với các văn bản Việt ngữ, việc gán nhãn từ loại có nhiều khó khăn, đặc biệt là bản thân việc phân loại từ tiếng Việt cho đến nay vẫn là một vấn đề còn nhiều tranh cãi, chưa có một chuẩn mực thống nhất. Nghiên cứu của chúng tôi nhằm phục vụ đồng thời hai mục đích: Một mặt thực hiện nỗ lực xây dựng công cụ gán nhãn từ loại cho từ tiếng Việt, phục vụ cho hệ thống rút trích thông tin. Mặt khác, xây dựng một kho ngữ liệu tiếng Việt cho 48 loại từ loại, đặt nền tảng cho việc phát triển các ứng dụng xử lý ngôn ngữ tiếng Việt trên máy tính phục vụ cho các ứng dụng khác.

Để nghiên cứu áp dụng cho vấn đề tự động gán nhãn từ loại cho từ tiếng Việt, chúng tôi đã thực hiện các công việc cụ thể sau:

1.Xác định bộ chú thích 48 từ loại [1] với 10 miền giới hạn:

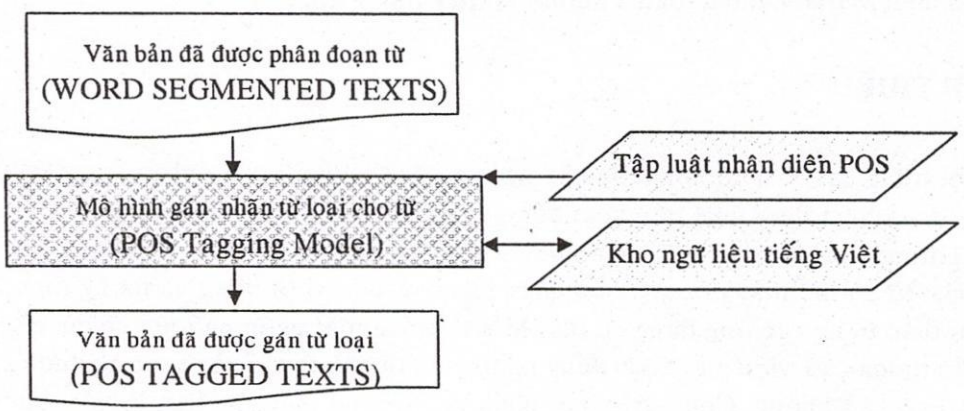
- Lớp thực thể về các nhân vật
- Lớp thực thể về các tổ chức
- Lớp thực thể về các công ty
- Lớp thực thể về các thành phố
- Lớp thực thể về các tỉnh
- Lớp thực thể về các núi non
- Lớp thực thể về các sông ngòi
- Lớp thực thể về các con đường
- Lớp thực thể về các địa điểm đặc biệt (các địa điểm du lịch, thắng cảnh, di tích lịch sử,...)
- Lớp thực thể về tên khác của các thực thể

2. Xây dựng kho ngữ liệu huấn luyện (training corpus) cho tiếng việt đã loại bỏ nhập nhằng từ loại, và gán các nhãn có thể cho mỗi từ với bộ chú thích 48 từ loại [16].

3.Xây dựng tập luật (rule set) để nhận diện 48 nhãn từ loại trong văn bản tiếng Việt trong các trường hợp không có sự nhập nhằng về nhãn từ loại.

4. Xây dựng công cụ gán nhãn từ loại tự động dùng phương pháp xác suất, dựa trên các thông tin, các quy tắc kết hợp từ loại học được từ kho ngữ liệu đã gán nhãn mẫu và văn phong của văn bản.

Mô hình tổng quát cho bài toán gán nhãn từ loại như sau (Hình 1):



Hình 1.Mô hình tổng quát.

2. PHƯƠNG PHÁP LUẬN

Phương pháp gán nhãn từ loại cho từ Tiếng Việt.

Trong phần này bài báo giới thiệu tổng quan về các kỹ thuật gán nhãn từ loại cho văn bản tiếng Việt. Quá trình gán nhãn từ loại có thể chia làm 3 bước [14].

- Phân tách xâu ký tự thành chuỗi các từ. Giai đoạn này có thể đơn giản hay phức tạp tuỳ theo ngôn ngữ và quan niệm về đơn vị từ vựng. Chẳng hạn đối với tiếng Anh hay tiếng Pháp, việc phân tách từ phần lớn là dựa vào các ký hiệu trắng. Tuy nhiên vẫn có những từ ghép hay những cụm từ gây tranh cãi về cách xử lý. Trong khi đó với tiếng Việt thì dấu trắng càng không phải là dấu hiệu để xác định ranh giới các đơn vị từ vựng do tần số xuất hiện từ ghép rất cao.

- Khởi tạo gán nhãn, tức là tìm cho mỗi từ tập tất cả các nhãn từ loại mà nó có thể có. Tập nhãn này có thể thu được từ cơ sở dữ liệu từ điển hoặc kho văn bản đã gán nhãn bằng tay. Đối với một từ mới chưa xuất hiện trong cơ sở ngữ liệu thì có thể dùng một nhãn ngầm định hoặc gắn cho nó tập tất cả các nhãn. Trong các ngôn ngữ biến đổi hình thái người ta cũng dựa vào hình thái từ để đoán nhận lớp từ loại tương ứng của từ đang xét.

- Quyết định kết quả gán nhãn, đó là giai đoạn loại bỏ nhập nhằng, tức là lựa chọn cho mỗi từ một nhãn phù hợp nhất với ngữ cảnh trong tập nhãn khởi tạo. Có nhiều phương pháp để thực hiện việc này, trong đó người ta phân biệt chủ yếu *các phương pháp dựa vào quy tắc ngữ pháp* mà đại diện nổi bật là phương pháp Brill [8] và *các phương pháp xác suất* [14]. Ngoài ra còn có các hệ thống sử dụng mạng nơ-ron [15], các hệ thống lai sử dụng kết hợp tính toán xác suất và ràng buộc ngữ pháp, gán nhãn nhiều tầng [9].

Về mặt ngữ liệu, các phương pháp phân tích từ loại thông dụng hiện nay dùng một trong các loại tài nguyên ngôn ngữ sau:

- Từ điển và các văn phạm loại bỏ nhập nhằng [11].

- Kho văn bản đã gán nhãn [13], có thể kèm theo các quy tắc ngữ pháp xây dựng bằng tay [8].

- Kho văn bản chưa gán nhãn, có kèm theo các thông tin ngôn ngữ như là tập từ loại và các thông tin mô tả quan hệ giữa từ loại và hậu tố [14].

- Kho văn bản chưa gán nhãn, với tập từ loại cũng được xây dựng tự động nhờ các tính toán thống kê [1]. Trong trường hợp này khó có thể dự đoán trước về tập từ loại.

Các công cụ gán nhãn từ loại dùng từ điển và văn phạm gần giống với một công cụ phân tích cú pháp. Các hệ thống học sử dụng kho văn bản để học cách đoán nhận từ loại cho mỗi từ [10]. Từ giữa những năm 1980 các hệ thống này được triển khai rộng rãi vì việc xây dựng kho văn bản mẫu ít tốn kém hơn nhiều so với việc xây dựng một từ điển chất lượng cao và một bộ quy tắc ngữ pháp đầy đủ. Một số hệ thống sử dụng đồng thời từ điển để liệt kê các từ loại có thể cho một từ, và một kho văn bản mẫu để loại bỏ nhập nhằng. Công cụ gán nhãn của chúng tôi kết hợp tính toán xác suất và các đặc thù ràng buộc ngữ pháp cũng như văn phong.

Các công cụ gán nhãn thường được đánh giá bằng độ chính xác của kết quả: ***[số từ được gán nhãn đúng] / [tổng số từ trong văn bản]***. Các công cụ gán nhãn tốt nhất hiện nay có độ chính xác đạt tới 98% [14].

3. CÔNG CỤ GÁN NHÃN

Nghiên cứu áp dụng cho vấn đề tự động gán nhãn từ loại tiếng Việt, chúng tôi đã thực hiện các bước sau:

- Bước thứ nhất: Xác định các nhãn từ loại (bao gồm 48 từ loại như danh từ loại thể, đại từ nhân xưng, phụ từ chỉ thời gian, ..vv.) cho các từ thích hợp dựa trên các luật cú pháp và ngữ cảnh.
- Bước thứ hai: Khởi tạo gán nhãn, tức là tìm cho mỗi từ còn lại tập tất cả các nhãn từ loại mà nó có thể có. Tập nhãn này có thể thu được từ cơ sở dữ liệu từ điển hoặc kho ngữ liệu đã gán nhãn bằng tay. Đối với một từ mới chưa xuất hiện trong cơ sở ngữ liệu thì có thể dùng một nhãn ngầm định hoặc gắn cho nó tập tất cả các nhãn. Trong các ngôn

ngữ biến đổi hình thái người ta cũng dựa vào hình thái từ để đoán nhận lớp từ loại tương ứng của từ đang xét.

Bước thứ ba: Quyết định kết quả gán nhãn, đó là giai đoạn loại bỏ nhập nhằng, tức là lựa chọn cho mỗi từ một nhãn phù hợp nhất với ngữ cảnh trong tập nhãn khởi tạo.

Về mặt ngữ liệu, chúng tôi dùng kết hợp hai loại tài nguyên ngôn ngữ sau:

- Từ điển gồm 80.000 mục từ và các văn phạm loại bỏ nhập nhằng.
- Kho ngữ liệu đã gán nhãn gồm gần 75 000 mục từ, có thể kèm theo các quy tắc ngữ pháp xây dựng bằng tay.

3.1. Phương pháp gán nhãn bằng xác suất

Về ý tưởng của phương pháp gán nhãn từ loại bằng xác suất là xác định phân bố xác suất trong không gian kết hợp giữa dãy các từ S_w và dãy các nhãn từ loại S_t. Sau khi đã có phân bố xác suất này, bài toán loại bỏ nhập nhằng từ loại cho một dãy các từ được đưa về bài toán lựa chọn một dãy từ loại sao cho xác suất điều kiện $P(S_t \mid S_w)$ kết hợp dãy từ loại đó với dãy từ đã cho đạt giá trị lớn nhất.

Theo công thức xác suất Bayes ta có: $P(S_t \mid S_w) = P(S_w \mid S_t).P(S_t)/P(S_w)$. Ở đây dãy các từ S_w đã biết, nên thực tế chỉ cần cực đại hoá xác suất $P(S_w \mid S_t).P(S_t)$.

Với mọi dãy $S_t = t_1t_2 \ldots t_N$ và với mọi dãy $S_w = w_1w_2 \ldots w_N$:

$$P(w_1w_2 \ldots w_N \mid t_1t_2 \ldots t_N) = P(w_1 \mid t_1t_2 \ldots t_N)\, P(w_2 \mid w_1, t_1t_2 \ldots t_N) \ldots P(w_N \mid w1 \ldots w_{N-1}, t_1t_2 \ldots t_N)$$

$$P(t_1t_2 \ldots t_N) = P(t_1)P(t_2 \mid t_1)\, P(t_3 \mid t_1t_2) \ldots P(t_N \mid t_1 \ldots t_{N-1})$$

Người ta đưa ra các giả thiết đơn giản hoá cho phép thu gọn mô hình xác suất về một số hữu hạn các tham biến.

Đối với mỗi $P(w_i \mid w1 \ldots w_{i-1}, t_1t_2 \ldots t_N)$, giả thiết khả năng xuất hiện một từ khi cho một nhãn từ loại là hoàn toàn xác định khi biết nhãn đó, nghĩa là $P(w_i \mid w1 \ldots w_{i-1}, t_1t_2 \ldots t_N) = P(w_i \mid t_i)$.

Như vậy, xác suất $P(w_1w_2 \ldots w_N \mid t_1t_2 \ldots t_N)$ chỉ phụ thuộc vào các xác suất cơ bản có dạng $P(w_i \mid t_i)$:

$$P(w_1w_2 \ldots w_N \mid t_1t_2 \ldots t_N) = P(w_1 \mid t_1)P(w_2 \mid t_2) \ldots P(w_N \mid t_N)$$

Đối với các xác suất $P(t_i \mid t_1 \ldots t_{i-1})$, giả thiết khả năng xuất hiện của một từ loại là hoàn toàn xác định khi biết các nhãn từ loại trong một lân cận có kích thước k cố định, nghĩa là: $P(t_i \mid t_1 \ldots t_{i-1}) = P(t_i \mid t_{i-k} \ldots t_{i-1})$. Nói chung, các công cụ gán nhãn thường sử dụng giả thiết k bằng 1 (bigram) hoặc 2 (trigram).

Như vậy mô hình xác suất này tương đương với một mô hình Markov ẩn [12][5], trong đó các trạng thái ẩn là các nhãn từ loại (hay các dãy gồm k nhãn nếu $k > 1$), và các trạng thái hiện (quan sát được) là các từ trong từ điển. Với một kho văn bản đã gán nhãn mẫu, các tham số của mô hình này dễ dàng được xác định nhờ thuật toán Viterbi [3][12] và được mô tả như sau.

THUẬT TOÁN Viterbi

Cho một chuỗi các từ $W_1, \ldots, W_T$, từ loại $C_1, \ldots, C_N$, xác suất $\Pr(W_i \mid C_i)$ và xác suất Bigram $\Pr(C_i \mid C_j)$, tìm chuỗi từ loại $C_1, \ldots, C_T$ phù hợp nhất cho chuỗi từ $W_1, \ldots, W_T$.

Bước khởi tạo: for i = 1 to K do /*K là số lượng từ loại; <Start>: từ loại rỗng */SeqScore(i,1) = $\Pr(C_1 \mid$ <Start>$)$* $\Pr(W_1 \mid C_i)$

BACKPTR(i,1) = 0;

Bước lặp:

for t = 2 to T do /* T là số lượng từ trong câu cho trước */

for i = 1 to K do

SeqScore (i,t) = Max (SeqScore (j,t -1)* Pr (C_i | C_j))* Pr (W_t | C_i), với j = 1,..K

BACKPTR(i,t) = Chỉ số j cho giá trị Max ở trên.

Bước xác định chuỗi từ loại:

C(T) = i là Max của SeqScore(i,t)

for i = T-1 to 1 do

C(i) = BACKPTR(C(i+1),i+1)

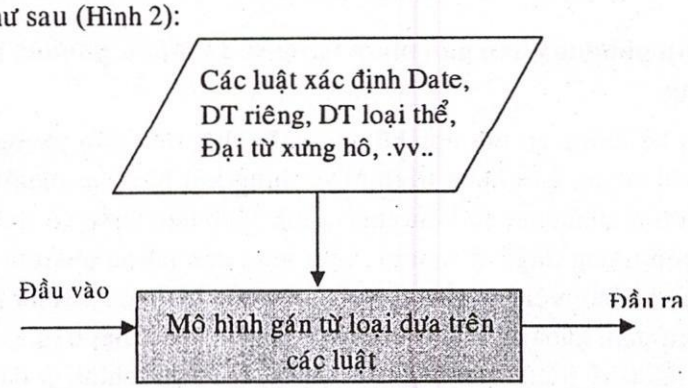
3.2. Phương pháp gán nhãn dựa trên văn phong

Văn phong là cách viết văn của mỗi người, mỗi thể loại văn bản. Phương pháp gán nhãn từ loại dựa trên văn phong thực chất là căn cứ vào cách thể hiện của văn bản trong một ngữ cảnh cụ thể để xác định từ loại cho các từ, điều này bao hàm việc xác định phải đảm bảo các luật văn phạm của các từ trong câu.

Ví dụ: *Trong buổi họp, anh$_1$ Nguyễn Văn Thành đã phát biểu về ... một cách tích cực.Tuy nhiên, anh$_2$ không tập trung vào vấn đề chính của cuộc họp.*

Như vậy, từ ***anh*** trong đoạn văn bản trên đóng hai vai trò ngữ pháp, ***anh$_1$*** là Danh từ loại thể (Nt) và ***anh$_2$*** là Đại từ xưng hô (Pp). Và ***Nguyễn Văn Thành*** đóng vai trò ngữ pháp là Danh từ riêng (Np).

Trên cơ sở dựa vào cách thể hiện của văn bản trong một ngữ cảnh cụ thể như ví dụ trên và ngữ pháp tiếng Việt [6] [7] [16], chúng ta có thể xây dựng một hệ thống các luật mà dựa vào đó chúng ta có thể xác định được từ loại cho các từ trong văn bản trong trường hợp không bị nhập nhằng. Mô hình của phương pháp gán nhãn từ loại dựa trên văn phong được mô phỏng như sau (Hình 2):



Hình 2 . Mô hình của phương pháp gán nhãn từ loại dựa trên văn phong

Về ý tưởng của phương pháp gán nhãn từ loại dựa trên văn phong được diễn đạt thông qua thủ tục như sau:

- Áp dụng các luật xác định danh từ riêng [1].
- Trên cơ sở các danh từ riêng được xác định, tiếp tục áp dụng các luật để xác định 48 nhãn từ loại còn lại.

Như trong ví dụ: *Trong buổi họp, anh$_1$ Nguyễn Văn Thành đã phát biểu về ... một cách tích cực.Tuy nhiên, anh$_2$ không tập trung vào vấn đề chính của cuộc họp.*

Thủ tục nhận diện được diễn đạt như sau:

1. Các luật xác định danh từ riêng → ***Nguyễn Văn Thành***
2. Các luật xác định danh từ loại thể → ***anh$_1$***
3. Các luật xác định đại từ nhân xưng → ***anh$_2$***
4. Các luật xác định từ loại khác → ...

Về phương pháp xây dựng hệ thống luật, chúng tôi dựa vào JAPE (Java Annotation Patterns Engine)[2] để hiện thực được trên 270 luật chính để xác định 48 nhãn từ loại[1]. Do sự giới hạn trình bày, bài báo chỉ minh họa hai luật đơn giản trong trường hợp đoán nhận một từ có nhãn là date như sau:

```
Rule: date1
 (({Token.kind=="number"})((({SpaceToken})*{Token.string=="-"}
({SpaceToken})*|({SpaceToken})*{Token.string=="/"}({SpaceToken})*)
{Token.kind=="number"})+)+ --> date

        Rule: date2
(({Token.string=="ngày"}|{Token.string=="Ngày"})({SpaceToken})+{Token.kind=="num
ber"}({SpaceToken})+({Token.string=="tháng"}|{Token.string=="Tháng"})({SpaceToken
})+{Token.kind=="number"}({SpaceToken})+({Token.string=="năm"}|{Token.string=="N
ăm"})({SpaceToken})+{Token.kind=="number"} ({SpaceToken})+
)+ --> date
```

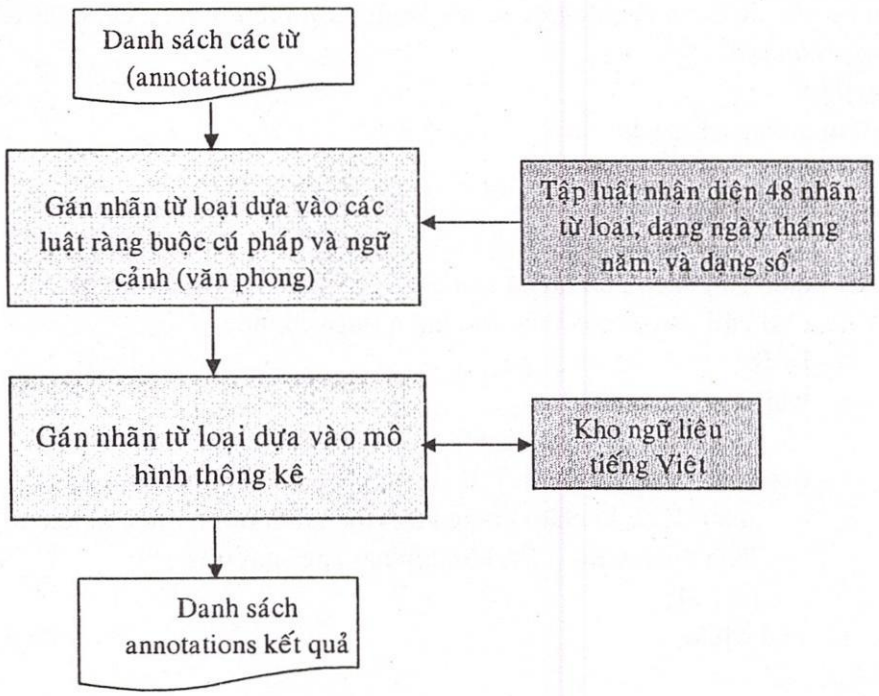
Trong đó: Token.kind – Loại Token, SpaceToken – Ký tự trắng, * có nghĩa là không hoặc nhiều, + có nghĩa là một hoặc nhiều, và | có nghĩa là Hoặc.

Với luật date1 sẽ đoán nhận các dạng date như *1/2/2006, 1-2-2006, .vv...* Và luật date2 sẽ đoán nhận các dạng date như *Ngày 2 Tháng 3 Năm 2006, Ngày 03 tháng 03 năm 2006, .vv...*

3.3. Mô hình kết hợp phương pháp gán nhãn bằng xác suất và phương pháp gán nhãn dựa trên văn phong

Bộ gán nhãn là một hệ thống lai kết hợp bộ gán nhãn dựa trên văn phong và bộ gán nhãn trigram. Trong đó chỉ có bộ gán nhãn trigram sử dụng kết hợp hai nguồn thông tin: Một từ điển khoảng 80.000 từ chứa các từ kèm theo danh sách các nhãn có thể của chúng cùng với tần suất xuất hiện tương ứng; và một ma trận gồm các bộ ba nhãn từ loại có thể xuất hiện liền nhau trong văn bản với các tần số xuất hiện của chúng. Loại dữ liệu này thu được dựa vào kho ngữ liệu mẫu khoảng 78.920 từ đã gán nhãn. Các loại dấu câu và các ký hiệu khác trong văn bản được xử lí như các đơn vị từ vựng, với nhãn chính là dấu câu tương ứng.

Còn ở trong bộ gán nhãn dựa trên văn phong, chúng tôi xây dựng được hệ thống trên 270 luật để xác định cho 48 từ loại (danh từ riêng, đại từ xưng hô, danh từ loại thể, .vv..) và các luật để xác định các kiểu ngày tháng năm (date). Mô hình của bài toán gán nhãn từ loại được trình bày như sau (Hình 3):



Hình 3. Mô hình bài toán gán nhãn từ loại.

Bộ gán nhãn sẽ làm việc với dữ liệu vào là một danh sách các annotations, mỗi một annotation tương ứng với một từ trong văn bản. Bộ gán nhãn có thể gán một dãy gồm bốn nhãn từ loại cùng với thông tin xác suất tương ứng cho mỗi từ trong danh sách, hoặc chỉ gán kết quả cuối cùng - nhãn có khả năng xuất hiện cao nhất. Và chúng ta thu được annotations kết quả có cấu trúc như sau:

annotation.id = chỉ số id // *id của annotation*
annotation.type = "vnWord"; // *loại của annotation*
annotation.fm={ // *các tính chất của annotation*
string = giá trị 1; // *chuỗi ký tự của từ*
kind = giá trị 2; // *loại của từ*
length = giá trị 3; // *chiều dài của từ*
orth= giá trị 4; // *dạng của các ký tự của từ có 4 giá trị:*
- lowercase: các ký tự đều chữ thường
-upperInitial: ký tự đầu tiên là chữ hoa
-allCaps: các ký tự đều chữ hoa
-mixedCaps: các ký tự chữ hoa và chữ hoa xen nhau
pos= giá trị 5; // *nhãn từ loại của từ*
};
annotation.start=startNode; // *vị trí bắt đầu của annotation trong văn bản*
annotation.end = endNode; // *vị trí cuối của annotation trong văn bản*

Về mặt thuật toán, bộ gán nhãn thực hiện thủ tục như sau:

- Đọc tất cả các từ trong văn bản;
- Gán nhãn từ loại cho các từ mà không gây ra sự nhập nhằng;
  + Áp dụng các luật xác định danh từ riêng;

+ Trên cơ sở các danh từ riêng được xác định, tiếp tục áp dụng các luật để xác định 48 nhãn từ loại còn lại;

- Ghi vào bộ đệm
- while(bộ đệm không trống) do

+ Đọc 3 từ từ bộ đệm;
+ for mỗi từ trong 3 từ này do
  - o if từ đó có trong từ điển
  - o then gán cho từ đó tất cả các nhãn (tag) có trong từ điển;
  - o else gán cho từ đó tất cả các nhãn (tag) có thể;
  - o j = 0;
  - o while(j < số nhãn) do
    - Tính Pw = P(tag|token) là xác suất từ token có nhãn tag;
    - Tính Pc = P(tag|t1,t2), là xác suất nhãn tag xuất hiện sau các nhãn t1, t2, là nhãn tương ứng của hai từ đứng trước từ token;
    - Tính Pw,c = Pw * Pc, kết hợp hai xác suất trên.;
    - j = j +1;
  - o end while;

+ end for;

- end while;

Sau đây là một ví dụ kết quả sau khi qua bộ gán nhãn của câu: *"Năm ngoái /, / Ông / Nguyễn Thành Tài / đi / thăm / khu / di tích / lịch sử / Củ Chi."* được thể hiện dưới dạng XML như sau:

<w pos="Jt"> ***Năm ngoái***</w>
<w pos="Nt"> ***Ông*** </w>
< w pos="Np"> ***Nguyễn Thành Tài*** </w>
<w pos=","> , </w>
<w pos="Vm"> ***đi*** </w>
<w pos="Vtim"> ***thăm*** </w>
<w pos="Nt"> ***khu*** </w>
<w pos="Na"> ***di tích*** </w>
<w pos="Na"> ***lịch sử*** </w>
<w pos="Np"> ***Củ Chi*** </w>
<w pos="."> . </w>

Trong đó: Jt – Phụ từ chỉ thời gian , Nt – Danh từ loại thể, Np – Danh từ riêng, Vm – Động từ chuyển động, Vtim – Động từ ngoại động cảm nghĩ, Na – Danh từ trừu tượng

4. ĐÁNH GIÁ

Chương trình được viết bằng ngôn ngữ lập trình Java trên môi trường GATE [2], Mã chương trình đích khoảng 160KB. Mã nguồn mở dễ dàng sửa đổi và tích hợp trong các ứng dụng khác. Thời gian huấn luyện hay gán nhãn với ngữ liệu khoảng 34000 lượt từ đều tốn khoảng 43 giây.

Kết quả thử nghiệm tốt nhất với các tập mẫu đã xây dựng đạt tới độ chính xác ~80% nếu chỉ dùng phương pháp gán nhãn bằng xác suất (**P1**) và đạt ~90% nếu dùng phương pháp gán nhãn dựa trên văn phong kết hợp với phương pháp xác suất (**P2**). Bảng 1 minh hoạ kết quả gán nhãn: Tỉ lệ tương ứng trong mỗi thử nghiệm là độ chính xác.

Bảng 1. Kết quả gán nhãn từ loại

Văn bản / Văn phong	Số đơn vị từ	P1	P2
Chuyện tình1 / Tiểu thuyết VN	16787	80,53%	**90,75%**
Chuyện tình2 / Tiểu thuyết VN	14698	80,78%	**90,39%**
Hoàng tử bé / Truyện nước ngoài	18663	80,90%	**90,48%**
Lược sử thời gian / Sách khoa học	11626	78,44%	**88,20%**
Công nghệ / Báo chí	10662	77,81%	**87,90%**
Độ chính xác trung bình		79,69%	**89,54%**

5. KẾT LUẬN

Trên đây bài báo đã trình bày một phương pháp tiếp cận để giải quyết bài toán gán nhãn từ loại tự động là kết hợp tính toán xác suất và các đặc thù ràng buộc ngữ pháp cũng như văn phong cho các văn bản tiếng Việt. Tuy những kết quả ban đầu có độ chính xác chưa thật cao, nhưng chúng cũng đáp ứng được tốt yêu cầu đặt ra ban đầu của đề tài và đặt nền tảng cho các nghiên cứu tiếp theo. Với các kết quả gán nhãn thu được, chúng tôi sẽ tiếp tục bổ sung kho dữ liệu gồm các văn bản được gán nhãn mẫu, cũng như phát triển phương pháp gán nhãn từ loại dựa trên văn phong cho các từ loại, để làm tăng chất lượng công cụ gán nhãn. Và kho dữ liệu này cũng đặc biệt hữu ích cho việc nghiên cứu văn phạm tiếng Việt.

Việc nghiên cứu văn phạm trên cơ sở các văn bản đã gán nhãn cũng giúp cho chúng tôi điều chỉnh công cụ gán nhãn từ loại, sao cho các từ loại đưa ra đáp ứng được tốt nhất yêu cầu thể hiện các đặc trưng ngữ pháp của các đơn vị từ vựng. Bên cạnh đó, các công cụ tự động gán nhãn từ loại cũng hỗ trợ tích cực cho các nhà ngôn ngữ phát hiện các *hiện tượng ngôn ngữ* cần nghiên cứu.

VIETNAMESE PART-OF-SPEED TAGGING BASED ON STYLE OF TEXTS AND PROBABILITY MODEL

Chau Quang Nguyen [(1)], Tuoi Thi Phan [(2)], Tru Hoang Cao[(2)]
(1) Ho Chi Minh University of Industry
(2) University of Technology, VNU-HCM

ABSTRACT : *Accurate part-of-speech (POS) tagging for words in Vietnamese texts is very important problem. It will support for texts parsing, resolve polysemy, assist with semantic information extraction systems, etc. Therefore, this paper presents an approach to POS tagging for Vietnamese texts. This method used probability model and based on a lexicon with information about possible POS tags for each word, a manually labelled corpus, syntax and context of texts. Concurrently, we also built a corpus with 75,000 entries and a*

lexicon with 80,000 entries for the purpose of Vietnamese language processing research and application development.

Keywords: *Vietnamese, Part-of-Speech (POS), POS Tagging, style of texts, lexicon, corpus, probability, Markov model, Viterbi algorithm, Information Extraction.*

TÀI LIỆU THAM KHẢO

[1]. Chau Quang Nguyen, Tuoi Thi Phan, Tru Hoang Cao, *Vietnamese Proper Noun Recognition*, Proceedings of The Fourth International IEEE Conference on Computer Sciences- RIVF'06, pp.144-151, 2006.

[2]. Hamish, Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Cristian ,Ursu, Marin Dimitrov, Mike Dowman, Niraj Aswani , *Developing Language Processing Components with GATE*, The University of Sheffield 2001-2006 ,http://gate.ac.uk/sale/tao/.

[3]. Nguyễn Chí Hiếu, Phan Thị Tươi, Nguyễn Xuân Dũng, Nguyễn Quang Châu, *Sử Dụng Kỹ Thuật Pruning Vào Bài Toán Xác Định Từ Loại*, Tạp chí Phát triển Khoa học & Công nghệ, Tập 8, Số 11, 14-23, 2005.

[4]. Nguyễn Thị Minh Huyền, Vũ Xuân Lương, Lê Hồng Phong, *Sử Dụng Bộ Gán Nhãn Từ Loại Xác Suất QTAG Cho Văn Bản Tiếng Việt*, Proceedings of ICT.rda'03. Hanoi, Feb 2003.

[5]. Sang-Zhu Lee, Jung-ichi Tsujii, Hae-Chang Rim, *Lexicalized Hidden Markov Models for Part-of-Speech Tagging*, University of Tokyo, Japan, Korea University, Korea, 2000.

[6]. Cao Xuân Hạo, *Tiếng Việt - mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa*, NXB Giáo dục, 2000.

[7]. Nguyễn Tài Cẩn, *Ngữ pháp tiếng Việt*, NXB Đại học quốc gia Hà nội, 1999.

[8]. Brill E., *Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging*, Computational Linguistics, 21(4), pp.543-565, December 1999.

[9]. Tufis D., *Tiered Tagging and combined classifier, In Jelineck F. and North E. (Eds), Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligence 1692, Springer, 1999.

[10]. Abney S.,*Part-of-Speech Tagging and Partial Parsing*, in Young S. and Bloothooft (Eds), Corpus-Based Methods in Language and Speech processing, Kluwer Academic Publishers, Dodreht (The Netherlands), 1997.

[11]. Oflazer K., *Error-tolenrant finite-state recognition with applications to morphological analysis and spelling correction*, Computational Linguistics, 22(1), pp. 73-89, 1996.

[12]. James Allen, *Natural Language Understanding*, Benjamin/Cummings Publishing Company, 1995.

[13]. Dermatas E., Kokkinakis G., *Automatic Stochastic Tagging of Natural Language Texts*, Computational Linguistics 21.2, pp. 137 - 163, 1995.

[14]. Levinger M., Ornan U., Itai A., *Learning morpho-lexical probabilities from an untagged corpus with an application to Hebrew*, *Comutational Linguistics*, 21(3), pp. 383-404, 1995.

[15]. Schmid H., *Part-of-Speech Tagging with Neural networks*, International Conference on Computational Linguistics, Japan, pp. 172-176, Kyoto, 1994.

[16]. Uỷ ban khoa học xã hội Việt Nam, *Ngữ pháp tiếng Việt*, NXB Khoa học Xã hội, Hà nội, 1993.