SỬ DỤNG BỘ GÁN NHÃN TỪ LOẠI XÁC SUẤT QTAG CHO VĂN BẢN TIẾNG VIỆT

A case study of the probabilistic tagger QTAG for Tagging Vietnamese Texts

Nguyễn Thị Minh Huyền, Vũ Xuân Lương, Lê Hồng Phương

Tóm tắt

Trong bài báo này chúng tôi trình bày chi tiết các thử nghiệm về gán nhãn từ loại cho các văn bản tiếng Việt bằng cách áp dụng bộ gán nhãn QTAG, một bộ gán nhãn xác suất độc lập với ngôn ngữ. Chúng tôi sử dụng hai bộ nhãn từ loại với độ mịn khác nhau. Việc gán nhãn tự động dựa trên một bộ từ vựng có thông tin từ loại cho mỗi từ và một tập văn bản đã được gán nhãn bằng tay. Chúng tôi cũng trình bày khâu tiền xử lí cho việc gán nhãn: phân tách các đơn vị từ trong văn bản.

Từ khoá: từ loại, từ vựng, kho văn bản, phân tách từ, gán nhãn xác suất, QTAG

Abstract

In this paper we describe in detail our experiments on tagging Vietnamese texts using QTAG, a language independent probabilistic tagger with two part-of-speech (POS) sets at two different levels of finesse, based on a lexicon with information about possible POS tags for each word and a manually labeled corpus. We also describe the pre-processing for POS tagging, saying text tokenization.

Keywords: POS, lexicon, corpus, tokenization, probabilistic tagging, QTAG

GIỚI THIỆU

Một trong các vấn đề nền tảng của phân tích ngôn ngữ 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 một vai trò ngữ pháp nhất định. Các bộ chú thích từ loại 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 [19]. 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 [3], [5], [8], [13], [18]. Nghiên cứu của nhóm chúng tôi

phục vụ đồng thời hai mục đích: một mặt thực hiện nỗ lực nhằm xây dựng các công cụ cho việc xử lí văn bản tiếng Việt trên máy tính phục vụ cho các ứng dụng công nghệ, mặt khác các công cụ này cũng hỗ trợ tích cực cho các nhà ngôn ngữ nghiên cứu tiếng Việt.

Trong báo cáo này chúng tôi sẽ trình bày phương pháp tiếp cận và kết quả thu được của nhóm nghiên cứu trong bước thử nghiệm đầu tiên với một công cụ gán nhãn tự động thuần tuý xác suất.

BÀI TOÁN GÁN NHÃN TỪ LOẠI

Trong phần này chúng tôi giới thiệu tổng quan về các kĩ thuật gán nhãn từ loại và các bước giải quyết bài toán gán nhãn từ loại cho văn bản tiếng Việt.

Quá trình gán nhãn từ loại có thể chia làm 3 bước [15].

- Phân tách xâu kí 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 kí hiệu trắng. Tuy nhiên vẫn có những từ ghép hay những cụm từ công cụ gây tranh cãi về cách xử lí. 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.

- Gán nhãn tiên nghiệm, 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 tiên nghiệm. 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 ([2]) và *các phương pháp xác suất* ([4]). Ngoài ra còn có các hệ thống sử dụng mạng nơ-ron ([16]), các hệ thống lai sử dụng kết hợp tính toán xác suất và ràng buộc ngữ pháp [6], gán nhãn nhiều tầng [17].

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 [14].

- Kho văn bản đã gán nhãn [4], có thể kèm theo các quy tắc ngữ pháp xây dựng bằng tay [2].

- 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 [10].

- Kho văn bản chưa gán nhãn, với tập từ loại cũng được xây dựng tự động nhờ các tính toán thống kê [11]. Trong trường hợp này khó có thể dự đoán trước về tập từ loại.

Các bộ gán nhãn từ loại dùng từ điển và văn phạm gần giống với một bộ phân tích cú pháp. Các hệ thống *học* sử dụng kho văn bản để *học* cách đoán nhận từ loại cho mỗi từ [1]. 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. Bộ gán nhãn của chúng tôi nằm trong số các hệ thống này.

Các bộ 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 bộ gán nhãn tốt nhất hiện nay có độ chính xác đạt tới 98% [15].

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

1. Xây dựng từ điển từ vựng, lựa chọn tiêu chí xác định từ loại trong quá trình phân tích từ vựng. Hầu hết các mục từ trong từ điển đều có thông tin từ loại đi kèm.

2. Xây dựng công cụ phân tách các đơn vị từ vựng trong văn bản.

3. Xây dựng kho văn bản đã loại bỏ nhập nhằng từ loại bằng tay, sau khi tự động gán tất cả các nhãn có thể cho mỗi từ.

4. Xây dựng bộ gán nhãn từ loại tự động, dựa trên các thông tin từ loại trong từ điển từ vựng và các quy tắc kết hợp từ loại *học được* từ kho văn bản đã gán nhãn mẫu.

Trong phần tiếp theo của báo cáo, chúng tôi sẽ lần lượt trình bày các bước 1, 2 và 4.

XÂY DỰNG TỪ ĐIỂN TỪ VỰNG, XÁC ĐỊNH BỘ CHÚ THÍCH TỪ LOẠI TIẾNG VIỆT

Trong khuôn khổ đề tài cấp Nhà nước KC01 "Nghiên cứu phát triển công nghệ nhận dạng, tổng hợp và xử lí ngôn ngữ tiếng Việt", nhóm nghiên cứu đã triển khai các công việc xây dựng kho ngữ liệu tiếng Việt bao gồm từ điển từ vựng và kho văn bản có kèm theo mô tả từ loại của các đơn vị từ vựng với chất lượng cao, tuân theo các chuẩn quốc tế về biểu diễn dữ liệu[1], cho phép cập nhật và mở rộng dễ dàng.

[1] cf. ISO TC37/SC4 http://www.tc37sc4.org

Từ điển từ vựng

Trong tiếng Việt, bên cạnh những đơn vị rõ ràng là từ, là ngữ cố định như thành ngữ (*sơn cùng thuỷ tận, tay xách nách mang...*), quán ngữ (*lên lớp, lên mặt, ra vẻ*), còn tồn tại những đơn vị có người cho là từ, có người cho là ngữ cố định (như *xe lăn đường, máy quay đĩa, làm ruộng, lạnh ngắt*, *suy cho cùng, ...*). Ranh giới của từ trong tiếng Việt là một vấn đề phức tạp, trong nhiều trường hợp còn có những ý kiến khác nhau [8].

Chúng tôi lựa chọn quan niệm đơn vị từ vựng theo cuốn *Từ điển tiếng Việt* [7] (do Viện Ngôn Ngữ Học biên soạn) để xây dựng cơ sở ngữ liệu. Trong toàn bộ cuốn từ điển này, quan điểm về việc thu thập từ vựng, về chuẩn hoá chính tả, về chú thích từ loại là rõ ràng và thống nhất.

Ngoài ra, chúng tôi có đưa thêm các đơn vị từ vựng ít dùng, gặp trong kho văn bản nhưng không được thu thập trong từ điển vào *Từ điển từ vựng*. Mặt khác, chúng tôi cũng đưa thêm các đơn vị từ vựng mới xuất hiện (mà từ điển chưa thu thập) vào *Từ điển từ vựng* cùng với những đơn vị là tên người, tên địa danh, tên tổ chức thường gặp để tiện cho chương trình xử lí.

Chính tả trong [7] "theo đúng các *Quy định về chính tả tiếng Việt và về thuật ngữ tiếng Việt* trong các sách giáo khoa, được ban hành theo Quyết định số 240/QĐ ngày 5-3-1984 của Bộ trưởng Bộ Giáo dục" (chẳng hạn vấn đề viết nguyên âm "-i", viết "-uy", cách ghi dấu thanh, cách viết thuật ngữ khoa học, sử dụng con chữ f, j, w, z cho các từ mượn tiếng nước ngoài, v.v.).

Trên thực tế, trong các văn bản tiếng Việt vẫn không có sự thống nhất trong cách ghi dấu thanh ở những âm tiết có âm đệm, vì vậy mà trước khi áp dụng cho chương trình tách từ và gán nhãn từ loại, văn bản đã được chúng tôi xử lí lại cho nhất quán với từ điển.

Xây dựng bộ chú thích từ loại

Từ loại phản ánh vị trí khác nhau của các từ trong hệ thống ngữ pháp. Để phản ánh được chính xác tất cả các quan hệ ngữ pháp thì cần có một bộ từ loại rất lớn. Nhưng càng nhiều chú thích từ loại thì công việc gán nhãn càng khó khăn. Bởi vậy cần phải có một sự thoả hiệp để đạt được một bộ chú thích từ loại không quá lớn và có chất lượng.

Chúng tôi chọn làm việc với hai bộ từ loại. Trước hết là sử dụng bộ chú thích 8 từ loại (danh từ, động từ, tính từ, đại từ, phụ từ, kết từ, trợ từ, cảm từ) được cộng đồng ngôn ngữ học thoả hiệp tương đối, trình bày trong cuốn Ngữ pháp tiếng Việt [18] và được chú thích cụ thể cho từng mục từ trong [7].

Bộ từ loại thứ hai được xây dựng bằng cách phân nhỏ mỗi từ loại trên thành các tiểu từ loại. Ban đầu chúng tôi dùng ngay cách chia thành tiểu loại trong [18].

Những chú thích từ loại được chọn như trên sau đó được phản ánh đầy đủ trong *Từ điển từ vựng*, làm cơ sở dữ liệu cho chương trình tự động xác định ý nghĩa danh từ, động từ...,

động từ nội động hay động từ ngoại động... của mỗi từ khi phân xuất trực tiếp trong văn bản. Cùng với từ điển này là kho văn bản đã được chúng tôi gán nhãn bằng tay sau khi đã chạy chương trình tách từ và xác định tất cả các nhãn có thể tìm được trong từ điển cho mỗi từ.

Trong quá trình xác định nhãn cho từng từ trong văn bản cụ thể, chúng tôi nhận thấy sự cần thiết phải bổ sung thêm một số nhãn từ loại để tránh trường hợp một từ mang cùng một lúc nhiều nhãn từ loại (chẳng hạn động từ ngoại động chỉ cảm nghĩ hay động từ nội động chỉ cảm nghĩ). Như vậy quá trình xây dựng tập mẫu cũng đồng thời là quá trình điều chỉnh việc phân chia từ loại hợp lí hơn. Hiện tại chúng tôi làm việc với bộ nhãn từ loại ở mức mịn hơn gồm 47 từ loại và bổ sung một nhãn cho các từ chưa xác định được từ loại.

PHÂN TÁCH TỪ TRONG VĂN BẢN TIẾNG VIỆT

Đặt bài toán.

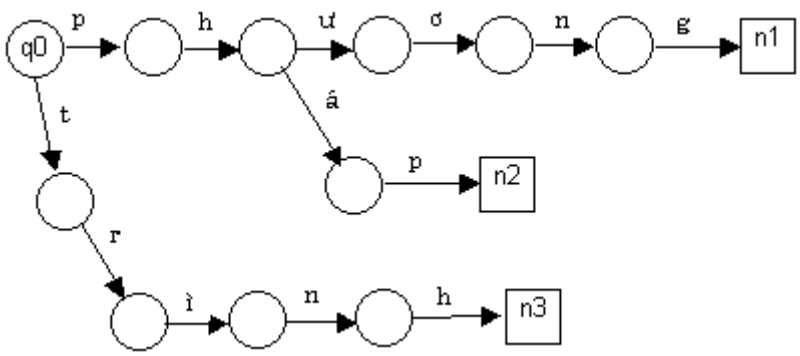
> Cho một câu tiếng Việt bất kỳ, hãy tách câu đó thành những đơn vị từ vựng (từ), hoặc chỉ ra những âm tiết nào không có trong từ điển (phát hiện đơn vị từ vựng mới).

Để giải quyết bài toán đặt ra, chúng tôi sử dụng tập dữ liệu gồm bảng âm tiết tiếng Việt (khoảng 6700 âm tiết) và từ điển từ vựng tiếng Việt (khoảng 30.000 từ). Các từ điển được lưu dưới dạng các tệp văn bản có định dạng mã TCVN hoặc Unicode dựng sẵn (UTF-8). Chương trình xây dựng bằng Java, mã nguồn mở (liên hệ nhóm tác giả).

Các bước giải quyết

1. Xây dựng ôtômát âm tiết đoán nhận tất cả các âm tiết tiếng Việt
2. Xây dựng ôtômát từ vựng đoán nhận tất cả các từ vựng tiếng Việt.
3. Dựa trên các ôtômát nêu trên, xây dựng đồ thị tương ứng với câu cần phân tích và sử dụng thuật toán tìm kiếm trên đồ thị để liệt kê các cách phân tích có thể.

Bảng chữ cái của ôtômát âm tiết là bảng chữ cái tiếng Việt, mỗi cung chuyển được ghi trên đó một ký tự. Ví dụ, với ba âm tiết *phương*, *pháp*, *trình* ta sẽ có ôtômát đoán nhận âm tiết như Hình 1.



Hình 1. Xây dựng ôtômát âm tiết

Thuật toán xây dựng ôtômát âm tiết

Input: Từ điển âm tiết

Output: Ôtômát âm tiết.

Thuật toán:

1. Lập trạng thái khởi đầu q_0;
2. Vòng lặp đọc cho tới khi hết tệp dữ liệu, lấy ra từng âm tiết. Gọi các ký tự của âm tiết đó là $c_0, c_1, ..., c_{n-1}$.

a. $p := q_0; i := 0;$

b. Vòng lặp trong khi ($i \leq n - 1$)

i. Lấy ra ký tự c_i;

ii. Tìm trong các cung chuyển từ trạng thái p cung trên đó ghi ký tự c_i. Nếu có cung (p,q) như thế:

1. $i := i + 1;$
2. $p := q;$

iii. Nếu không có cung (p,q) nào như thế thì thoát khỏi vòng lặp b.

c. Với j từ i đến $n - 1$

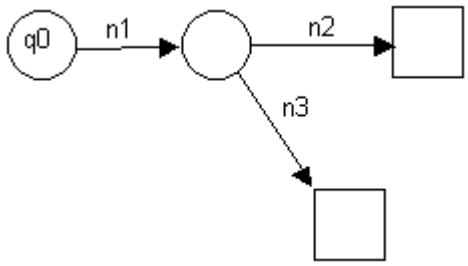
i. Tạo mới trạng thái q, ghi nhận q là trạng thái không kết;

ii. Thêm cung chuyển (p,q) trên đó ghi ký tự c_j;

iii. $p := q;$

d. Ghi nhận q là trạng thái kết;

Ôtômát từ vựng được xây dựng tương tự, với điểm khác như sau: thay vì ghi trên mỗi cung chuyển một âm tiết, ta ghi số hiệu của trạng thái (kết) của ôtômát âm tiết tại đó đoán nhận mỗi âm tiết của từ nhằm giảm kích thước của ôtômát từ vựng. Ví dụ, với hai từ *phương pháp* và *phương trình,* giả sử khi đưa lần lượt các âm tiết *phương*, *pháp*, *trình* qua ôtômát âm tiết, ta đến được các trạng thái kết ghi các số n_1, n_2, n_3 thì trên các cung chuyển tương ứng ta ghi các số n_1, n_2, n_3 (Hình 2).



Hình 2. Xây dựng ôtômát từ vựng

Thuật toán xây dựng ôtômát từ vựng

Input: Từ điển từ vựng, ôtômát âm tiết

Output: Ôtômát từ vựng.

Thuật toán:

1. Lập trạng thái khởi đầu q_0;
2. Vòng lặp đọc cho tới khi hết tệp dữ liệu, lấy ra từng mục từ *word*. Gọi các âm tiết của *word* là $s_0, s_1, ..., s_{n-1}$;
3. Sử dụng ôtômát âm tiết để đoán nhận các âm tiết trên, được các số hiệu của trạng thái (kết) tương ứng là $m_0, m_1, ..., m_{n-1}$

a. $p := q_0; i := 0;$

b. Vòng lặp trong khi ($i \le n - 1$)

i. Lấy ra số m_i;

ii. Tìm trong các cung chuyển từ trạng thái p cung trên đó ghi số m_i. Nếu có cung (p,q) như thế

1. $i := i + 1;$
2. $p := q;$

iii.Nếu không có cung (p,q) nào như thế thì thoát khỏi vòng lặp b.

c. Với j từ i đến $n - 1$

i. Tạo mới trạng thái q, ghi nhận q là trạng thái không kết;

ii. Thêm cung chuyển (p,q) trên đó ghi số m_j;

iii. $p := q;$

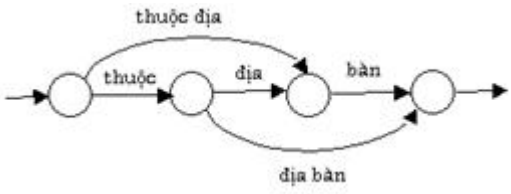
d. Ghi nhận q là trạng thái kết

Sau khi đã xây dựng xong hai ôtômát, ta ghi chúng vào hai tệp định kiểu để dùng trong bước phân tách từ vựng. Nếu mỗi ký tự (char) được ghi vào tệp với kích thước 2 byte (mã Unicode), mỗi số nguyên (int) có kích thước 4 byte thì tệp lưu ôtômát âm tiết có kích thước 146KB, tệp ôtômát từ vựng có kích thước 1MB.

Tư tưởng của thuật toán phân tách từ vựng là quy việc phân tách câu về việc tìm đường đi trên một đồ thị có hướng, không có trọng số.

Giả sử câu ban đầu là một dãy gồm $n+1$ âm tiết s_0, s_1, ..., s_n. Ta xây dựng một đồ thị có $n+2$ đỉnh v_0, v_1, ..., v_n, v_{n+1}, sắp thứ tự trên một đường thẳng từ trái sang phải; trong đó, từ đỉnh v_i đến đỉnh v_j có cung ($i < j$) nếu các âm tiết s_i, s_{i+1}, ..., s_{j-1} theo thứ tự lập thành một từ. Khi đó mỗi cách phân tách câu khác nhau tương ứng với một đường đi trên đồ thị từ đỉnh đầu v_0 đến đỉnh cuối v_{n+1}. Trong thực tế, cách phân tích câu đúng đắn nhất thường ứng với đường đi qua ít cung nhất trên đồ thị.

Trong trường hợp câu có sự nhập nhằng thì đồ thị sẽ có nhiều hơn một đường đi ngắn nhất từ đỉnh đầu đến đỉnh cuối, ta liệt kê toàn bộ các đường đi ngắn nhất trên đồ thị, từ đó đưa ra tất cả các phương án tách câu có thể và để người dùng quyết định sẽ chọn phương án nào, tuỳ thuộc vào ngữ nghĩa hoặc văn cảnh. Ví dụ, xét một câu có cụm "*thuộc địa bàn*", ta có đồ thị như sau (Hình 3)



Hình 3. Một tình huống nhập nhằng

Cụm này có sự nhập nhằng giữa *thuộc địa* và *địa bàn* và ta sẽ có hai kết quả phân tách là "*thuộc địa / bàn*" và "*thuộc / địa bàn*". Ta có thể chỉ ra rất nhiều những cụm nhập nhằng trong tiếng Việt, chẳng hạn "tổ hợp âm tiết", "bằng chứng cớ",...

Trường hợp trong câu có âm tiết không nằm trong từ điển thì rõ ràng ôtômát âm tiết không đoán nhận được âm tiết này. Kết quả là đồ thị ta xây dựng từ câu đó là *không liên thông*. Dựa vào tính chất này, ta thấy rằng nếu đồ thị không liên thông thì dễ dàng phát hiện ra rằng đơn vị âm tiết không đoán nhận được không nằm trong từ điển âm tiết, tức nó bị viết sai chính tả hoặc là một đơn vị âm tiết (từ vựng) mới.

Đánh giá kết quả

Với cách tiếp cận như trên, bài toán phân tách từ vựng trong câu tiếng Việt về cơ bản đã được giải quyết, đặc biệt là vấn đề tách các tổ hợp từ tương đương với một đơn vị từ vựng, thường là các cụm từ cố định, ngữ cố định hoặc các thành ngữ trong tiếng Việt. Với những câu nhập vào có sự nhập nhằng từ vựng, tức có nhiều hơn một cách phân tách thì chương trình liệt kê toàn bộ các phương án tách từ có thể và giành quyền lựa chọn kết quả cho người sử dụng. Trong tất cả các phương án phân tách đó bao giờ cũng tồn tại phương án đúng.

Dưới đây là một số câu nhập vào và kết quả tách từ tương ứng.

1. Nó | là | một | bản | tuyên ngôn | đặc sắc | của | chủ nghĩa nhân đạo | , một | tiếng | chuông | cảnh tỉnh | trước | hiểm họa | lớn lao | của | hành tinh | trước | sự | điên rồ | của | những | kẻ | cuồng tín

2. Trong khi | các | thành phần | tư bản chủ nghĩa | có | những | bước | phát triển | mạnh | hơn | thời kì | trước | thì | thế lực | của | giai cấp | địa chủ | vẫn | không hề | suy giảm.

Như vậy, còn một số vấn đề khó khăn cần phải tiếp tục nghiên cứu giải quyết:

Thứ nhất là vấn đề giải quyết nhập nhằng phân tách. Cần phải chọn một phương án đúng giữa nhiều phương án. Các hướng tiếp cận khả thi cho vấn đề này có thể là:

- Dùng các quy tắc ngữ pháp do chuyên gia ngôn ngữ xây dựng. Tiến hành phân tích cú pháp của câu với những phương án tách từ vựng có thể, từ đó loại ra những phương án sai cú pháp.

- Dùng phương pháp xác suất - thống kê. Phải thống kê trong kho văn bản tương đối lớn của tiếng Việt để tìm ra xác suất của các bộ đôi hay bộ ba từ loại hoặc từ vựng đi cạnh nhau. Từ đó lựa chọn phương án phân tách có xác suất sai ít nhất.

Chương trình phân tích cú pháp tiếng Việt chúng tôi hiện có cũng đã có khả năng nhận biết được một số câu nhập nhằng từ vựng. Ví dụ, với câu "*bản sao chụp mờ*" thì có thể có hai cách phân tích có thể là "bản | sao chụp" và "bản sao | chụp", trình phân tích nhận thấy cả hai cách tách từ này đều đúng cú pháp và đưa ra hai cây phân tích tương ứng. Với câu "*anh ấy rất thuộc địa bàn*" thì mặc dù cụm "thuộc địa bàn" có hai cách phân tách từ vựng là "thuộc | địa bàn" và "thuộc địa | bàn" nhưng trình phân tích chỉ đoán nhận được một và đưa ra cách phân tích tương ứng với cách tách từ đó. Do đó, cách tách từ còn lại là sai.

Thứ hai là vấn đề giải quyết tên riêng, tên viết tắt và tên có nguồn gốc nước ngoài có mặt trong câu. Hiện tại chương trình phân tách chưa nhận ra được các cụm từ dạng "Nguyễn Văn A", "Đại học Khoa học Tự nhiên", hoặc "ĐT. 8.20.20.20", "1.000$", "0,05%"...

THỬ NGHIỆM BỘ GÁN NHÃN QTAG CHO TIẾNG VIỆT

QTAG là một bộ gán nhãn như vậy, do nhóm nghiên cứu Corpus Research thuộc trường đại học tổng hợp Birmingham phát triển, cung cấp miễn phí cho mục đích nghiên cứu[2]. Chúng tôi đã sửa đổi phần mềm này để thích nghi với việc thao tác trên văn bản tiếng Việt, cũng như cho phép sử dụng từ điển từ vựng có thông tin từ loại bên cạnh việc sử dụng kho văn bản đã gán nhãn. Với sự đồng ý của tác giả O. Mason, chúng tôi công bố phiên bản QTAG cho tiếng Việt cùng với kho ngữ liệu (vnQTAG) tại địa chỉ: http://www.loria.fr/equipes/led/outils.php.

Phương pháp gán nhãn xác suất

Ý tưởng của phương pháp gán nhãn từ loại 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 $S_t = t_1t_2... t_N$ và với mọi dãy $S_w = w_1w_2... w_N$:

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

[2] http://www.clg.bham.ac.uk/staff/oliver/software/tagger/

$P(t_1t_2...t_N) = P(t_1)P(t_2 | t_1)\ P(t_3 | t_1t_2) \ ...\ P(t_N | t_1...t_{N-1})$

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 $P(w_i | w1...\ w_{i-1,}\ t_1t_2...t_N)$, 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 $P(w_i | w1...\ w_{i-1,}\ t_1t_2...t_N) = P(w_i | t_i)$.

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

$$P(w_1w_2...\ w_N | t_1t_2...t_N) = P(w_1 | t_1)P(w_2 | t_2) \ ...\ P(w_N | t_N)$$

Đối với các xác suất $P(t_i | t_1...t_{i-1})$, 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: $P(t_i | t_1...t_{i-1}) = P(t_i | t_{i-k}...t_{i-1})$. Nói chung, các bộ 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, 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.

Bộ gán nhãn QTAG

1.1.1 Dữ liệu mẫu

Bộ gán nhãn QTAG là một bộ gán nhãn trigram. QTAG sử dụng kết hợp hai nguồn thông tin: một từ điển 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. Cả hai loại dữ liệu này thu được dễ dàng dựa vào kho văn bản mẫu đã gán nhãn. Các loại dấu câu và các kí hiệu khác trong văn bản được xử lí như các đơn vị từ vựng, với nhãn chính là dấu câu tương ứng.

1.1.2 Thuật toán gán nhãn từ loại

Về mặt thuật toán, QTAG làm việc trên một cửa sổ chứa 3 từ, sau khi đã bổ sung thêm 2 từ giả ở đầu và cuối văn bản. Các từ được lần lượt đọc và thêm vào cửa sổ mỗi khi cửa sổ di chuyển từ trái sang phải, mỗi lần một vị trí. Nhãn được gán cho mỗi từ đã lọt ra ngoài cửa sổ là nhãn kết quả cuối cùng. Thủ tục gán nhãn như sau:

1. Đọc từ (token) tiếp theo
2. Tìm từ đó trong từ điển
3. Nếu không tìm thấy, gán cho từ đó tất cả các nhãn (tag) có thể

4. Với mỗi nhãn có thể
5. tính Pw = P(tag|token) là xác suất từ token có nhãn tag
6. 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.
7. tính Pw,c = Pw * Pc, kết hợp hai xác suất trên.
8. Lặp lại phép tính cho hai nhãn khác trong cửa sổ

Sau mỗi lần tính lại (3 lần cho mỗi từ), các xác suất kết quả được kết hợp để cho ra xác suất toàn thể của nhãn được gán cho từ. Vì các giá trị này thường nhỏ, nên chúng được tính trong biểu thức logarit cơ số 10. Giá trị xác suất tính được cho mỗi nhãn tương ứng với một từ thể hiện độ tin cậy của phép gán nhãn này cho từ đang xét.

1.1.3 Thực hiện gán nhãn

Sau khi đã xây dựng từ điển từ vựng và ma trận xác suất chuyển giữa các từ loại từ dữ liệu mẫu, QTAG làm việc với dữ liệu vào là một văn bản đã được tách từ, mỗi từ nằm trên một dòng. Chương trình có thể in ra dãy các nhãn từ loại cùng với thông tin xác suất tương ứng cho mỗi từ trong văn bản, hoặc chỉ in ra kết quả cuối cùng - nhãn có khả năng xuất hiện cao nhất.

Sử dụng QTAG cho tiếng Việt

1.1.4 Dữ liệu mẫu

Nhóm nghiên cứu ngôn ngữ của Trung tâm Từ điển học xây dựng cơ sở dữ liệu mẫu bao gồm:

- Từ điển từ vựng gồm 37454 mục từ, mỗi mục từ có kèm theo dãy tất cả các từ loại mà nó có thể có, những đơn vị chưa xác định được từ loại thì gắn nhãn **X**.
- Các văn bản thuộc một số thể loại khác nhau (văn học Việt Nam/nước ngoài, khoa học, báo chí) được gán nhãn bằng tay, bao gồm 63732 lượt từ với 48 nhãn từ loại cùng với một số nhãn tương ứng với các dấu câu và một số kí hiệu khác.

1.1.5 Thử nghiệm

Như đã trình bày, bộ gán nhãn QTAG ban đầu chỉ làm việc với một kho văn bản đã được gán nhãn mẫu để "huấn luyện" cho mô hình xác suất. Trong quá trình gán nhãn, nếu gặp một đơn vị mới (có thể là từ, con số, các kí hiệu toán học...) chưa thấy xuất hiện trong tập mẫu, QTAG giả thiết đơn vị đó có thể có một nhãn từ loại bất kì nằm trong tập tất cả các nhãn đã xuất hiện trong tập huấn luyện.

Cơ sở dữ liệu của chúng tôi có từ điển từ vựng độc lập nên chúng tôi đã thực hiện một số thay đổi sau:

1. Đưa vào kho từ vựng của bộ gán nhãn tất cả các mục từ có trong từ điển từ vựng của chúng tôi và các mục từ có trong tập huấn luyện

2. Khi gặp một đơn vị mới trong tập văn bản cần gán nhãn, kiểm tra nếu đơn vị đó là số hay tên riêng thì gán nhãn số hay tên riêng

3. Ngoài ra, một môđun đoán nhận từ loại cho một từ mới dựa vào hậu tố của từ đó - không áp dụng được cho tiếng Việt - cũng được lược bỏ.

Phương pháp thử nghiệm của chúng tôi là lấy một phần kho văn bản đã gán nhãn làm tập huấn luyện cho mô hình xác suất. Sau đó chúng tôi áp dụng mô hình này để tự động gán nhãn cho phần các văn bản còn lại rồi so sánh kết quả thu được với dữ liệu mẫu. Các thử nghiệm được thực hiện đối với 2 bộ chú thích từ loại trình bày trong mục 3. Với mỗi mức trên chúng tôi đã thực hiện các thử nghiệm, tương ứng với các tập mẫu khác nhau về kích thước và văn phong.

1.1.6 Đánh giá kết quả

Chương trình được cài đặt bằng ngôn ngữ lập trình Java, chạy trong mọi môi trường, có thể dùng mã tiếng Việt Unicode (dựng sẵn) hoặc TCVN. Mã chương trình đích khoảng 16KB. Mã nguồn dễ dàng sửa đổi và dùng lại. Thời gian huấn luyện hay gán nhãn với ngữ liệu khoảng 32000 lượt từ đều tốn khoảng 30 giây. Kết quả gán nhãn một câu nếu chọn định dạng XML như ví dụ sau:

<w pos="Nc">hồi</w> <w pos="Vto">lên</w> <w pos="Nn">sáu</w> <w pos=",">, </w> <w pos="Vs">có</w> <w pos="Nu">lần</w> <w pos="Pp">tôi</w> <w pos="Jt">đã</w> <w pos="Vt">nhìn</w> <w pos="Vt">thấy</w> <w pos="Nn">một</w> <w pos="Nt">bức</w> <w pos="Nc">tranh</w> <w pos="Jd">tuyệt</w> <w pos="Aa">đẹp</w>

trong đó: Nc - danh từ đơn thể, Vto - ngoại động từ chỉ hướng, Nn - danh từ số lượng, Vs - động từ tồn tại, Nu - danh từ đơn vị, Pp - đại từ nhân xưng, Jt - phụ từ thời gian, Vt - ngoại động từ, Nt - danh từ loại thể, Jd - phụ từ chỉ mức độ, Aa - tính từ hàm chất.

Kết quả thử nghiệm tốt nhất với các tập mẫu đã xây dựng đạt tới độ chính xác ~94% đối với bộ nhãn thứ nhất (9 nhãn từ vựng và 10 nhãn cho các loại kí hiệu), trong khi với bộ nhãn thứ hai chỉ đạt tới ~85% (48 nhãn từ vựng và 10 nhãn cho các loại kí hiệu). Bảng 1 minh hoạ kết quả gán nhãn với bộ nhãn thứ nhất: tỉ lệ tương ứng trong mỗi thử nghiệm là độ chính xác. Nếu không dùng đến từ điển từ vựng (chỉ sử dụng kho văn bản đã gán nhãn mẫu) thì các kết quả chỉ đạt được tương ứng là ~80% và ~60%.

Kết quả của các thử nghiệm ban đầu cũng cho chúng tôi một số nhận xét sau:

4. Với kích thước tập mẫu ban đầu như nhau, do tập nhãn từ loại ở mức 2 lớn hơn nhiều so với mức 1, nên tỉ lệ lỗi ở mức 2 cao hơn mức 1 khá nhiều.

5. Đúng như mong đợi, khi xử lí các văn bản cùng một văn phong, tập mẫu càng lớn thì tỉ lệ lỗi càng giảm
6. Tập mẫu với các văn bản có văn phong khác nhau có ảnh hưởng tới kết quả gán nhãn.

Bảng 1. Kết quả gán nhãn từ loại mức 1

Văn bản / Văn phong	Số đơn vị từ	Test 1	Test 2	Test 3	Test 4
Chuyện tình1 / Tiểu thuyết VN	16787	91,53%	89,75%	tập mẫu	tập mẫu
Chuyện tình2 / Tiểu thuyết VN	14698	91,78%	90,39%	94,28%	93,82%
Hoàng tử bé / Truyện nước ngoài	18663	tập mẫu	10,48%	tập mẫu	tập mẫu
Lược sử thời gian / Sách khoa học	11626	90,44%	tập mẫu	91,42%	tập mẫu
Muốii của rừng / Truyện ngắn VN	3573	90,68%	11,42%	91,04%	91,32%
Những bài học / Truyện ngắn VN	8244	91,45%	10,24%	92,90%	92,89%
Công nghệ / Báo chí	1162	88,81%	9,90%	89,24%	89,67%
Độ chính xác trung bình		91,25%	89,77%	92,70%	93,04%

KẾT LUẬN

Trên đây chúng tôi đã trình bày một phương pháp tiếp cận để giải quyết bài toán gán nhãn từ loại tự động cho các văn bản tiếng Việt. Tuy những kết quả ban đầu có độ chính xác chưa thật cao, nhưng chúng hứa hẹn triển vọng tốt 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, làm tăng chất lượng bộ gán nhãn. 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 bộ 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 tách từ và gán nhãn từ loại tự động cũng hỗ trợ tích cực cho các nhà ngôn ngữ phát hiện các *hiện tượng ngôn ngữ* cần nghiên cứu. Với mong muốn mở rộng sự quan tâm nghiên cứu của mọi người, chúng tôi sẵn sàng cung cấp tất cả các tài nguyên và công cụ đã xây dựng trong cộng đồng nghiên cứu xử lí tiếng Việt.

TÀI LIỆU THAM KHẢO

1. 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.
2. Brill E., "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging", Computational Linguistics, 21(4), December 199, p.543-565.

3. Cao Xuân Hạo, Tiếng Việt - mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa, NXB Giáo dục, 2000.
4. Dermatas E., Kokkinakis G., "Automatic Stochastic Tagging of Natural Language Texts", Computational Linguistics 21.2, 1995, p. 137 - 163.
5. Diệp Quang Ban, Hoàng Văn Thung, Ngữ pháp tiếng Việt (2 tập), NXB Giáo dục, 1999.
6. El-Bèze M, Spriet T., "Etiquetage probabiliste et contraintes syntaxiques", Actes de la conférence sur le Traitement Automatique du Langage Naturel (TALN95), Marseille, France,14-16/6/1995.
7. Hoàng Phê (chủ biên), Từ điển tiếng Việt 2002, Nhà xuất bản Đà Nẵng - Trung Tâm Từ Điển Học.
8. Hữu Đạt, Trần Trí Dõi, Đào Thanh Lan, Cơ sở tiếng Việt, NXB Giáo dục, 1998.
9. Kuipec J., "Robust Part-of-Speech Tagging Using a Hidden Markov Model", Computer Speech and Language, vol. 6, 1992, p. 225-242.
10. Levinger M., Ornan U., Itai A., "Learning morpho-lexical probabilities from an untagged corpus with an application to Hebrew", Comutational Linguistics, 21(3), 1995, p. 383-404.
11. MacMahon J.G., Smith F.J., "Improving statistical language model performance with automatically generated word hierarchies", Computational Linguistics, 19(2), 1993, p. 313-330.
12. Mason O., Tufis D., "Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger", 1st International Conference on Language Ressources and Evaluation (LREC98), Granada (Spain), 28-30 May 1998, p. 589-596.
13. Nguyễn Tài Cẩn, Ngữ pháp tiếng Việt, NXB Đại học Quốc gia Hà Nội, 1998.
14. Oflazer K., "Error-tolenrant finite-state recognition with applications to morphological analysis and spelling correction", Computational Linguistics, 22(1), 1996, p. 73-89.
15. Paroubek P., Rajman M., "Etiquetage morpho-syntaxique", Ingénierie des langues, chapitre 5, Hermes Science Europe, 2000.
16. Schmid H., "Part-of-Speech Tagging with Neural networks", International Conference on Computational Linguistics, Japan, 1994, p. 172-176, Kyoto.
17. Tufis D., "Tiered Tagging and combined classifier", In Jelineck F. and Nörth E. (Eds), Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, Springer, 1999.
18. Uỷ ban khoa học xã hội Việt Nam, Ngữ pháp tiếng Việt, NXB Khoa học Xã hội, Hà nội, 1993.
19. Vergnes J., Giguet E., "Regards théoriques sur le tagging", 5e conférence sur le Traitement Automatique du Langage Naturel (TALN98), Paris, 10-12 juin, 1998.