BỘ KHOA HỌC VÀ CÔNG NGHỆ
CHƯƠNG TRÌNH KH&CN CẤP NHÀ NƯỚC KC01/06-10
Đề tài KC01.01/06-10
"Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt"

BÁO CÁO KẾT QUẢ SẢN PHẨM

SP7.3 - KHO NGỮ LIỆU CÂU TIẾNG VIỆT CÓ CHÚ GIẢI (VIETREEBANK)

(Quyển 1)

Thời gian thực hiện: 5/2007- 5/2009

Chủ trì thực hiện: **TS. Nguyễn Phương Thái**
Đơn vị công tác: **Khoa Công nghệ thông tin,**
Đại học Công nghệ,
Đại học Quốc gia Hà nội

HÀ NỘI, 05/2009

MỤC LỤC

NỘI DUNG ĐĂNG KÝ

SP1.5: Kho ngữ liệu câu tiếng Việt có chú giải (VieTreeBank)

Chỉ tiêu chất lượng cần đạt theo đăng ký:

- Tài liệu mô tả tập nhãn cú pháp và tập qui tắc gán nhãn cú pháp
- Kho ngữ liệu 10.000 câu được gán nhãn cú pháp đầy đủ, khuôn dạng như các TreeBank khác.
- Tài liệu hướng dẫn.

NHỮNG THÀNH VIÊN THỰC HIỆN CHÍNH

TT	Họ và tên	Cơ quan	Nhiệm vụ chính
1	TS. Nguyễn Phương Thái	Đại học Công nghệ, ĐHQG Hà Nội	Phụ trách chung, thiết kế, xây dựng công cụ
2	TS. Vũ Xuân Lương	Trung tâm Từ điển học	Thiết kế
3	TS. Nguyễn Thị Minh Huyền	Đại học Khoa học Tự nhiên, ĐHQG Hà Nội	Thiết kế
4	NCS. Lê Hồng Phương	Đại học Khoa học Tự nhiên, ĐHQG Hà Nội	Xây dựng công cụ
5	THS. Đào Minh Thu	Trung tâm Từ điển học	Gán nhãn cú pháp
6	CN. Nguyễn Thị Minh Ngọc	Trung tâm Từ điển học	Gán nhãn cú pháp
7	CN. Lê Kim Ngân	Trung tâm Từ điển học	Gán nhãn cú pháp
8	CN. Nguyễn Mai Vân	Trung tâm Từ điển học	Gán nhãn cú pháp

Ngoài ra còn một số người khác :

PGS. TS. Nguyễn Văn Hiệp : cố vấn về các vấn đề cú pháp

PGS. TS. Đinh Điền : xây dựng guideline tách câu

TS. Lê Anh Cường : xây dựng phiên bản đầu tiên của công cụ

1. Giới thiệu

Ngân hàng câu được chú giải cú pháp (treebank) là kho ngữ liệu rất quan trọng trong nghiên cứu và xây dựng ứng dụng xử lý ngôn ngữ tự nhiên. Tiếng Việt là ngôn ngữ còn thiếu nhiều tài nguyên trong đó có treebank. Bài báo này trình bày các kết quả của dự án xây dựng treebank tiếng Việt. Tiếng Việt là ngôn ngữ đơn lập và không có ký tự tách từ cho nên việc phân tích câu có nhiều nhập nhằng. Để giải quyết nhập nhằng đó chúng tôi phải vận dụng nhiều thủ thuật ngôn ngữ học (mô tả trong tài liệu hướng dẫn gán nhãn). Quá trình gán nhãn được hỗ trợ bằng các công cụ gán nhãn tự động và công cụ soạn thảo. Nguồn ngữ liệu thô được lấy từ báo Tuổi Trẻ điện tử. Độ đồng thuận mà chúng tôi đạt được là khá cao, khoảng hơn 90%.

Treebank thường được dùng để xây dựng các hệ phân tích cú pháp, gán nhãn từ loại, tách từ. Các hệ đó lại có thể được dùng cho các ứng dụng như trích rút thông tin, dịch tự động, hỏi đáp, và tóm tắt văn bản. Ngoài ra treebank còn có thể được dùng cho các nghiên cứu ngôn ngữ học, chẳng hạn như khảo sát hiện tượng ngôn ngữ đặc thù nào đó. Gần đây cùng với sự ra đời của các phương pháp thống kê trên dữ liệu lớn thì treebank và các kho ngữ liệu khác càng đóng vai trò quan trọng hơn.

Tiếng Việt là ngôn ngữ đơn lập và không có ký tự phân tách từ. Đơn vị nhỏ nhất cấu tạo nên từ là âm tiết. Từ tiếng Việt thì có thể là đơn âm tiết hoặc đa âm tiết. Chữ viết tiếng Việt được sáng tạo ra dựa trên bộ chữ cái Latin "mở rộng" trong đó phần mở rộng ứng với các ký tự có dấu (ă, â, ê, v.v.) và các dấu trọng âm.

Tiếng Việt là ngôn ngữ có thứ tự từ khá cố định do đó chúng tôi sẽ biểu diễn cây cú pháp bằng cấu trúc thành phần (constituency structure). Đối với các ngôn ngữ mà thứ tự từ khá tự do như tiếng Nhật hay tiếng Séc thì cấu trúc phụ thuộc (dependency structure) thích hợp hơn. Chúng tôi áp dụng tiếp cận xây dựng treebank của Marcus và cộng sự (1993). Đây là một tiếp cận đã được kiểm chứng qua việc áp dụng cho nhiều ngôn ngữ khác nhau như: tiếng Anh, tiếng Trung, tiếng Hàn, tiếng Ả-rập, v.v.

Với tiếng Việt, có ba mức độ gán nhãn là tách từ[1], gán nhãn từ loại, và gán nhãn cú pháp. Bước tách từ có nhiệm vụ xác định xem trong câu có những từ nào. Bước gán nhãn từ loại xác định từ loại cho các từ trong câu. Bước cuối cùng là gán nhãn cú pháp, bao gồm cả nhãn thành phần và nhãn chức năng. Mục tiêu chính của chúng tôi là nghiên cứu xây dựng kho ngữ liệu gồm 10 ngàn câu tiếng Việt được chú giải cú pháp. Quá trình xây dựng treebank có một số bước cơ bản là: tìm hiểu, thiết kế, xây dựng công cụ, thu thập ngữ liệu thô, và gán nhãn dữ liệu. Thực chất quá trình này là xoáy trôn ốc, vừa gán dữ liệu vừa hoàn thiện thêm tài liệu hướng dẫn gán nhãn (thiết kế) hay cải tiến công cụ. Chúng tôi chọn văn bản báo chí để gán nhãn, mà cụ thể là các bài báo thuộc chủ đề Chính trị-Xã hội của báo Tuổi Trẻ điện tử.

Vì tiếng Việt là ngôn ngữ đơn lập và không có ký hiệu phân tách từ nên nhập nhằng xuất hiện trọng tất cả các mức độ phân tích. Do đó trong guideline gán nhãn chúng tôi thường mô tả thủ thuật phân tích đi kèm với các hiện tượng ngữ pháp. Các thủ thuật này có thể được thu thập từ các sách ngữ pháp hoặc do chúng tôi đưa ra. Chúng được mô tả kèm với ví dụ minh họa và biện luận sự phù hợp cũng như đưa ra thông tin về lựa chọn phân tích khác.

Trong công cụ gán nhãn tự động, chúng tôi sử dụng các mô hình học máy mạnh như CRFs cho gán nhãn từ loại hay LPCFGs cho phân tích cú pháp. Các công cụ này giúp cải thiện tốc độ gán nhãn rất nhiều. Thêm vào đó công cụ soạn thảo cây cũng có nhiều tính năng hữu dụng với người gán nhãn.

2. Phương pháp nghiên cứu và kết quả

[1] Các ngôn ngữ mà dấu cách là ký tự phân tách từ như tiếng Anh hay tiếng Pháp thì không cần pha tách từ

2.1 Phương pháp nghiên cứu

a) Qui trình xây dựng treebank

Qui trình xây dựng treebank thường tuân theo các bước:

- Tìm hiểu: Xác định tiếp cận xây dựng treebank phù hợp với ngôn ngữ đang được xem xét. Định ra các công việc cần làm và lường trước các khó khăn sẽ gặp phải. Bước này liên quan đến các lĩnh vực như ngôn ngữ học, treebank, và ngôn ngữ học tính toán.
- Thiết kế guideline: là một bước rất quan trọng, quyết định chất lượng treebank được làm ra. Guideline là tài liệu mô tả tập nhãn và hướng dẫn gán nhãn cho các hiện tượng cụ thể với các ví dụ minh họa.
- Xây dựng công cụ
- Thu thập văn bản thô: thường được lấy từ báo chí hoặc sách. Chẳng hạn treebank tiếng Anh chọn báo Wall Street Journal, treebank tiếng Trung chọn báo XinHua.
- Gán nhãn dữ liệu

b) Lược đồ chú giải

Lược đồ chú giải có thể được chia ra hai loại lớn là chú giải theo cấu trúc thành phần và chú giải theo cấu trúc phụ thuộc. Loại thứ nhất quan tâm đến cấu trúc ngữ đoạn của câu trong khi loại thứ hai quan tâm chủ yếu đến sự phụ thuộc ngữ pháp giữa các từ trong câu. Rambow và Joshi (1997) đã chỉ ra rằng ta có thể chuyển đổi cây thành phần thành cây phụ thuộc chứ không thể làm ngược lại. Ngoài ra tùy đặc điểm ngôn ngữ học mà lược đồ chú giải của các ngôn ngữ cũng khác nhau. Có ngôn ngữ thì cần pha tách từ, có ngôn ngữ thì cần mã hóa thông tin hình thái trong tập nhãn từ loại, có ngôn ngữ thì cần mô hình hóa nhiều phụ thuộc xa trong biểu diễn cây cú pháp. Các lược đồ chú giải theo cấu trúc thành phần thường có các đặc điểm sau:

- Mức độ gán nhãn: tách từ, từ loại, cú pháp. Trong đó mức tách từ chỉ dành cho các ngôn ngữ không có ký hiệu phân tách từ như tiếng Việt
- Nhãn cú pháp gồm cả nhãn chức năng (trong đó có nhãn vai nghĩa)
- Cấu trúc sâu được chỉ ra bằng cách sử dụng nhãn phần tử rỗng: hữu ích đối với một số hiện tượng ngữ pháp như chuyển vị wh, bị động, đề ngữ
- Đảm bảo sự nhận diện các quan hệ ngữ pháp cơ bản: quan hệ phụ thuộc, quan hệ đẳng lập, chủ-vị, bổ ngữ, phụ ngữ, đề ngữ, v.v.

c) Thiết kế guideline

Có một số yêu cầu với guideline. Thứ nhất, mức độ cụ thể của nhãn cần hợp lý, không cần quá chi tiết và cũng không được quá sơ sài. Thứ hai, phải đảm bảo tính nhất quán của guideline. Một yêu cầu khác là guideline cần trung lập với lý thuyết ngôn ngữ[2]. Và cuối cùng là thiết kế cần có tính tổng quát, dễ mở rộng về sau.

Khi xây dựng phiên bản đầu tiên của guideline, nhóm thiết kế cần tự tay phân tích tập câu mẫu lấy từ sách ngữ pháp, vừa phân tích vừa viết tài liệu. Kết quả sẽ bao trùm các cấu trúc và hiện tượng ngữ pháp cơ bản nhất. Bước kế tiếp là phân tích các câu lấy từ ngữ liệu thực tế (kết quả của bước chọn văn bản thô). Việc này rất quan trọng, nó giúp nhóm thiết kế đưa ra được tài liệu sát với thực tế hơn là chỉ dựa vào các câu mẫu trong sách. Kinh nghiệm cho thấy các vấn đề ngôn ngữ phát sinh khi xây dựng treebank đa dạng và phức tạp hơn nhiều so với những hiện tượng cơ bản được chỉ ra trong các sách ngữ pháp (Han và cộng sự, 2002). Do đó tài liệu hướng dẫn còn được chỉnh sửa, nâng cấp, và bổ sung trong quá trình gán nhãn văn bản.

[2] Thường thì các treebank không giải thích rõ điểm này

2.2 Kết quả

Đề tài đã đạt các chỉ tiêu đề ra:

- Ba kho ngữ liệu[3] (xem Bảng 1)
  - Kho ngữ liệu 10000 cây cú pháp
  - Kho ngữ liệu 10000 câu đã được gán nhãn từ loại
  - Kho ngữ liệu 2 triệu âm tiết đã được tách từ
- Tài liệu thiết kế
  - Hướng dẫn tách từ
  - Hướng dẫn gán nhãn từ loại
  - Hướng dẫn gán nhãn cú pháp
- Công cụ cho phép soạn thảo (thêm, sửa, xóa, v.v.) cây cú pháp và quản lý dữ liệu (thống kê, đánh giá, so sánh, v.v.)
- Giao diện web cho phép tra cứu và cập nhật kho ngữ liệu

Kho ngữ liệu	**Số câu**	**Số từ**	**Số âm tiết**
Được tách từ	69,482	1,541,869	2,000,409
Được gán nhãn từ loại	10,120	206,991	248,445
Được gán nhãn cú pháp đầy đủ	10,471	225,085	271,268

Bảng 1. Thống kê các kho ngữ liệu

3. Tài liệu thiết kế, qui trình sản phẩm

3.1 Tách từ

Có nhiều quan điểm nhận diện từ, chẳng hạn như dựa vào hình thái, dựa vào cú pháp, dựa vào ngữ nghĩa, hay dựa vào đối sánh ngôn ngữ học. Chúng tôi theo quan điểm của Sciullo và Williams (1987) coi từ là *nguyên tử cú pháp* theo nghĩa không thể áp dụng các qui tắc cú pháp để phân tích cấu tạo từ hay nói cách khác từ là đơn vị nhỏ nhất có khả năng hoạt động độc lập về cú pháp. Chúng tôi chọn tiêu chí này một phần là vì mục tiêu chính của đề tài là xây dựng tập cây cú pháp tiếng Việt, do đó tách từ trước hết là phục vụ cho mục tiêu này.

Xét về quan điểm ứng dụng, người làm dịch tự động Anh-Việt có thể lý luận là từ tiếng Việt nên có sự tương ứng với từ tiếng Anh thì sẽ thuận tiện cho dịch hơn. Tuy nhiên nếu tách từ chỉ đáp ứng cặp ngôn ngữ này thôi thì các cặp ngôn ngữ khác sẽ thế nào. Người làm từ điển thuật ngữ có thể thích đưa vào từ điển các cụm từ mà họ cho là cần được lý giải nghĩa chuyên ngành. Tuy nhiên nếu xét vai trò công cụ của phân tích cú pháp với từ điển thuật ngữ thì chỉ cần khả năng nhận dạng cụm từ và đưa ra danh sách từ (hay cụm từ) tiềm năng cho người làm từ điển thuật ngữ là hữu ích với họ rồi. Để có hệ nhận dạng cụm từ như thế thì không cần coi các cụm từ theo chuyên ngành là từ.

Các loại đơn vị từ vựng sau đây đã được xét trong quá trình làm dữ liệu:

- Từ đơn
- Từ ghép đẳng lập
- Từ ghép chính phụ
- Từ láy, dạng láy
- Từ ghép giữa yếu tố cấu tạo từ (vô, hoá, gia, nhà) với từ nguyên thuỷ
- Thành ngữ
- Quán ngữ
- Tên riêng (*)

[3] Các tập câu là phân biệt nhau

- Ngày tháng, số (*)
- Các loại dấu câu, ngoặc (*)
- Các cụm từ tiếng nước ngoài (*)
- Các chữ viết tắt (*)

Nhập nhằng tách từ là một trong những vấn đề chính mà người gán nhãn văn bản cần xử lý. Ta xét các từ "nhà cửa", "sắc đẹp", "hiệu sách". Họ cần xác định những từ đó là từ trong những câu như:

a. Nhà cửa bề bộn quá
b. Cô ấy giữ gìn sắc đẹp.
c. Ngoài hiệu sách có bán cuốn này

Và không phải là từ trong những câu như:

a. Ở nhà cửa ngõ chẳng đóng gì cả.
b. Bức này màu sắc đẹp hơn.
c. Ngoài cửa hiệu sách báo bày la liệt.

Trong quá trình tách từ, chúng tôi sử dụng từ điển như một tài liệu tham khảo quan trọng. Cụ thể hơn, chúng tôi chấp nhận hầu hết các từ từ điển là từ khi chúng xuất hiện trong ngữ cảnh phù hợp.

Trong các loại đơn vị từ vựng trên thì từ ghép chính phụ là loại mà tiêu chí nhận diện gây nhiều tranh cãi nhất. Một nguyên nhân cơ bản là qui tắc cấu tạo từ giống với qui tắc cú pháp. Vì thế nên ranh giới giữa luật từ vựng và luật cú pháp khó được xác lập. Các loại còn lại thì dễ nhận diện và ít nhập nhằng với cụm từ hơn do luật cấu tạo của chúng là riêng biệt. – thêm qđiểm NTCẩn&CXHạo – Đối với tiếng Trung, một số tác giả đưa ra các test có phải từ hay không (wordhood test) để nhận diện ranh giới từ. Chúng tôi cũng áp dụng các qui tắc này cho từ ghép chính phụ.
Trong nghiên cứu về cú pháp (mức cụm từ, mệnh đề), việc áp dụng test như trên không phải là vấn đề mới mẻ. Vậy đâu là sự khác biệt khi áp dụng ở mức tách từ và mức trên tách từ? Đối với mức tách từ, ta cần kiểm tra là từ ứng cử viên có phải là đơn vị được dùng ổn định hay không, nghĩa của nó không thể tổ hợp từ nghĩa thành phần hay không, thêm vào đó, không thể tạo ra cụm từ nghĩa tương đương bằng các qui tắc cú pháp (tạo lập cụm từ). Nói một cách nôm na, những đơn vị đúng (xứng đáng) là từ thì cần được đưa vào từ điển làm thành tri thức nền tảng mà dựa vào đó người ta có thể xây dựng các câu với nghĩa bất kỳ. Việc này cần được tối ưu theo nghĩa kích thước từ vựng là nhỏ nhất có thể.

3.2 Hướng dẫn gán nhãn từ loại và cú pháp

a) Tập nhãn từ loại

Trong các ngôn ngữ Châu Âu, khái niệm từ loại gắn với các phạm trù hình thái học như giống, số, cách, v.v. Trong tiếng Việt thì quan điểm phân từ loại thường dựa vào khả năng kết hợp và chức vụ ngữ pháp (gọi chung là thái độ ngữ pháp) và dựa vào nghĩa khái quát. Chúng tôi theo quan điểm phân từ loại dựa vào khả năng kết hợp và chức vụ ngữ pháp khi xây dựng treebank tiếng Việt. Như vậy nhãn từ loại của chúng tôi sẽ không có các thông tin hình thái (số ít, số nhiều, thì, ngôi, v.v.), thông tin về phân loại con[4] (ví dụ động từ đi với danh từ, động từ đi với mệnh đề, v.v.), và thông tin ngữ nghĩa. Tập nhãn từ loại của chúng tôi được liệt kê trong Bảng x, tổng số nhãn là 17.

[4] Subcategorization

b) Tập nhãn cú pháp

Tập nhãn cú pháp bao gồm ba loại nhãn là nhãn thành phần cú pháp (cụm từ và mệnh đề), nhãn chức năng, và nhãn phần tử rỗng. Các tập nhãn này được liệt kê trong phụ lục của bài báo. Chúng tôi sử dụng nhãn phần tử trung tâm H cho cụm từ Việt. Nếu cụm từ có nhiều từ trung tâm nối với nhau bởi liên từ đẳng lập (hay dấu phảy) thì tất cả các từ đó đều được gán nhãn H. Các treebank khác thường không chỉ rõ phần tử trung tâm của cụm từ. Do vậy những nghiên cứu về phân tích cú pháp (Collins, 1999) thường phải sử dụng các luật kinh nghiệm (heuristic rules) để xác định từ trung tâm hoặc dùng phương pháp học máy (Chiang and Bikel, 2002). Nhãn phần tử rỗng được dùng chủ yếu cho các hiện tượng tỉnh lược, bị động, khởi ngữ.

c) Thủ thuật phân tích câu và cụm từ

Ngôn ngữ đơn lập và không có ký tự phân tách từ như tiếng Việt thường có nhiều nhập nhằng trong phân tích cấu trúc câu và cụm từ. Do đó chúng tôi phải vận dụng nhiều thủ thuật phân tích để giải quyết các vấn đề như: xác định từ trung tâm của cụm từ, phân biệt giữa các loại bổ ngữ, phân biệt trạng ngữ với các thành phần khác, v.v. Các thủ thuật phân tích câu hay được dùng là: lược, thế, bổ sung, cải biến, thử câu hỏi. Các thủ thuật này dựa trên văn cảnh, khả năng kết hợp của từ, trật tự từ, và hư từ để giúp khử nhập nhằng giữa các cấu trúc có thể có.

STT	Vấn đề	Thủ thuật	Ví dụ
1	Xác định phần tử trung tâm	Lược	một đàn gà con mới nở... → một đàn gà con → một đàn gà → đàn gà con → đàn gà
2	Phân biệt mẫu động từ NP-VP và S	Thử bị động, nếu câu có dạng bị động thì phân tích là NP-VP, trái lại là S	Tôi bầu ông ấy làm hiệu trưởng. → Ông ấy được tôi bầu làm hiệu trưởng. (NP-VP) Anh ấy kể công ty đang làm ăn phát đạt. ~~→ Công ty được anh ấy kể đang làm ăn phát đạt.~~ (S)
3	Phân biệt "là" NP hay "là" S	Chèn "mà"	Tôi là người đặt cược ít nhất → Tôi là người mà đặt cược ít nhất
4	Phân biệt "vì" NP hay "vì" S	Thế đại từ, chèn "ấy", "đó", "này", v.v.	Vì chuyện quá lạ nên anh Bình vẫn chưa tin. ("chuyện quá lạ" là S)
5	Phân biệt bổ ngữ và vị ngữ	Chèn phó từ ("đã", đang", "rồi", v.v.)	Cô gái tên Phạm Thị Thanh Th. , 19 tuổi , quê ở Cần Thơ học chưa hết lớp 7. → Cô gái tên Phạm Thị Thanh Th. , 19 tuổi , quê ở Cần Thơ mới chỉ học chưa hết lớp 7. Ngữ đoạn "tên Phạm Thị … Cần Thơ" chỉ là định ngữ của "cô gái" do không thể chèn các từ chỉ thời thể

			(“đã”, “đang”, v.v.) vào giữa nó và “cô gái”
6	Có gán nhãn PRP hay không	Chèn kết từ chính phụ “để”	Nó đi làm kiếm tiền nuôi em. → Nó đi làm để kiếm tiền nuôi em.
7	Phân biệt trạng ngữ với các thành phần khác	Đảo vị trí, nguyên nhân hóa, v.v.	

Bảng 2. Một số thủ thuật phân tích câu và cụm từ

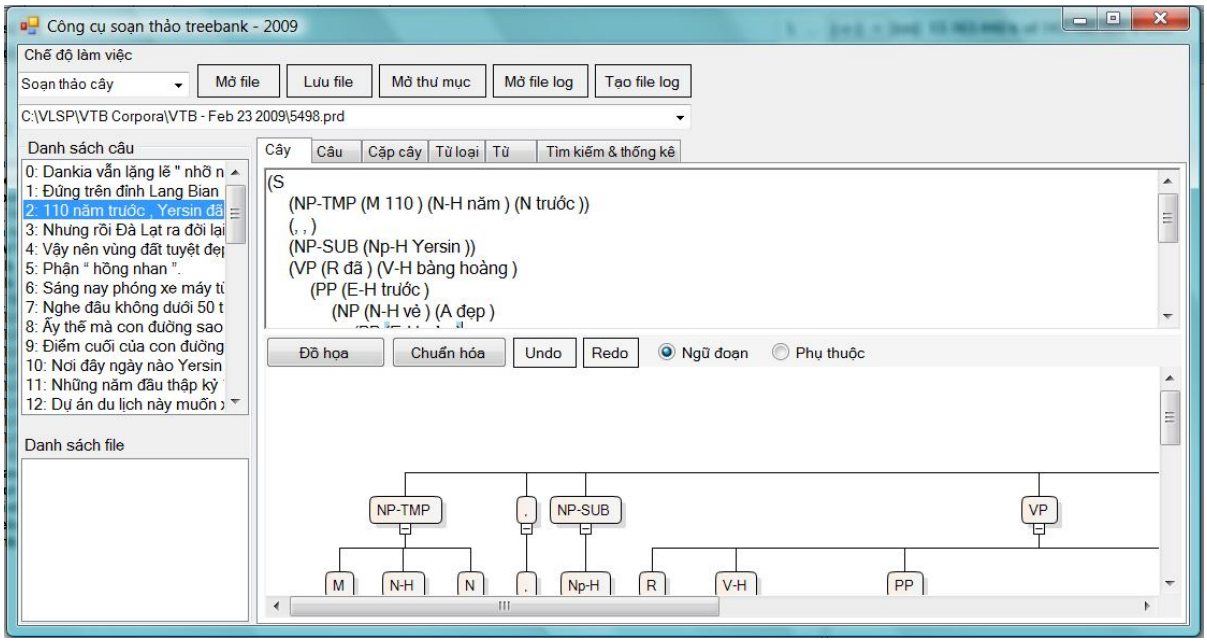
d) Các vấn đề ngôn ngữ

Về cơ bản treebank có thể coi như bài toán ứng dụng của ngôn ngữ học mặc dù trên thế giới người ta cũng hay sử dụng treebank cho nghiên cứu ngôn ngữ học[5]. Tuy nhiên trong những kết quả ngôn ngữ học được áp dụng vào treebank, không phải kết quả nào cũng được giới nghiên cứu ngôn ngữ học thừa nhận tuyệt đối. Chẳng hạn vấn đề trung tâm của danh ngữ, người thì coi danh từ chỉ loại là trung tâm, nhà ngôn ngữ khác thì cho đó là phần phụ trước. Hoặc có người coi chủ-vị là cấu trúc chủ đạo của câu tiếng Việt trong khi người khác lại coi đề-thuyết là cấu trúc chủ đạo. Cũng có nhà ngôn ngữ đề xuất việc kết hợp hai tiếp cận này, trong đó chủ-vị là cấu trúc ngữ pháp của cú (mệnh đề) còn đề-thuyết là cấu trúc cú pháp của câu. Treebank của chúng tôi thiên về cấu trúc chủ-vị hơn. Chúng tôi lựa chọn giải pháp ngôn ngữ phù hợp nhất với thiết kế của mình.

3.3 Công cụ hỗ trợ

Công cụ hỗ trợ người gán nhãn làm việc hiệu quả hơn. Có hai nội dung chính là hỗ trợ soạn thảo cây cú pháp và gán nhãn tự động (sau đó người sẽ sửa lại). Kinh nghiệm xây dựng treebank đã cho thấy là công cụ giúp tăng tốc độ gán nhãn lên rất nhiều. Hình 1 cho thấy công cụ soạn thảo cây cú pháp mà chúng tôi đang sử dụng.

[5] Corpus linguistics



Hình 1. Công cụ trợ giúp soạn thảo cây cú pháp

Các chức năng chính của công cụ soạn thảo:

- Soạn thảo và hiển thị cây ở chế độ văn bản và đồ họa
- Hiển thị file log, chỉ ra các thay đổi
- Tìm kiếm theo từ, mẫu cú pháp
- Tính độ đồng thuận
- Tự động đoán lỗi dữ liệu
- Thống kê dữ liệu

Các công cụ gán nhãn tự động mà chúng tôi sử dụng là:

- Hệ tách từ (Lê Hồng Phương và cộng sự, 2008)
- Hệ gán nhãn từ loại sử dụng mô hình CRFs (conditional random fields)
- Hệ phân tích cú pháp sử dụng mô hình LPCFGs (lexicalized probabilistic context-free grammars) (Lê Anh Cường và cộng sự, 2009)

3.4 Qui trình gán nhãn

Quá trình gán nhãn dữ liệu gồm ba công đoạn tách từ, gán nhãn từ loại, gán nhãn cú pháp. Vì ngay từ đầu đã có công cụ tách từ tự động nên nó được sử dụng ngay. Đối với gán nhãn từ loại và cú pháp thì trong giai đoạn đầu việc làm dữ liệu là hoàn toàn thủ công. Chỉ sau khi đã có một lượng dữ liệu nhất định chúng tôi mới dùng nó để huấn luyện các hệ phân tích tự động. Mỗi cây cú pháp sẽ phải qua tay ít nhất hai người làm dữ liệu (không kể tool). Người thứ nhất gán nhãn thô hoặc sửa lại kết quả máy đưa ra. Người thứ hai sẽ kiểm tra lại và sửa nếu cần. Thêm vào đó chúng tôi còn có qui trình kiểm tra dữ liệu theo các hiện tượng ngữ pháp, ví dụ nhóm từ chỉ hướng, câu hỏi, v.v. Tất nhiên là công cụ sẽ trích ra tập câu tương ứng để người làm dữ liệu kiểm tra. Như vậy là có những cây sẽ được kiểm tra ba lần trở lên.

4. Kết quả đánh giá sản phẩm

4.1 Đánh giá theo các tiêu chí kỹ thuật

4.1.1 Đánh giá độ đồng thuận

Độ đồng thuận được hiểu là mức độ giống nhau của kết quả gán nhãn cú pháp do hai người thực hiện độc lập trên cùng một văn bản. Vấn đề này tương tự như bài toán so sánh cây cú

pháp trong đánh giá chất lượng hệ phân tích cú pháp. Chúng tôi sử dụng cách so sánh thành phần cú pháp. Các cây cú pháp sẽ được chuyển thành dạng:

{(i, j, nhãn)}

trước khi được so sánh với nhau. Dựa vào đó ta sẽ tính được: tỉ lệ các thành phần giống nhau hoàn toàn (cả nhãn thành phần và nhãn chức năng), tỉ lệ các thành phần giống nhau bỏ qua nhãn chức năng, và tỉ lệ các thành phần chỉ giống nhau về cặp (i,j). Theo cách này, ta có thể đánh giá được độ đồng thuận cho từng thành phần cú pháp cụ thể như S, NP, VP, v.v.

Ví dụ: Hằng ngắm mưa trong công viên.

Người 1	**Người 2**
(S (NP (Np Hằng)) (VP (V ngắm) (NP (N mưa)) (PP (E trong) (NP (N công viên)))) (. .))	(S (NP (Np Hằng)) (VP (V ngắm) (NP (NP (N mưa)) (PP (E trong) (NP (N công viên))))) (. .))
(1,6,S); (1,1,NP); (2,5,VP); (3,3,NP); (4,5, PP); (5,5,NP)	(1,6,S); (1,1,NP); (2,5,VP); (3,3,NP); *(3,5,NP)*; (4,5, PP); (5,5,NP)

Bảng 3. Ví dụ về tính độ đồng thuận

Độ đồng thuận A giữa hai người gán nhãn sẽ được tính như sau:

$$A = \frac{2 * C}{C1 + C2}$$

Trong đó:

- C1 là số thành phần cú pháp trong kết quả gán nhãn của người thứ nhất
- C2 là số thành phần cú pháp trong kết quả gán nhãn của người thứ hai
- C là số thành phần cú pháp giống nhau

Trong ví dụ trên: C1=6; C2=7; C=6. Do đó A=12/13=0.92

Chúng tôi thực hiện một test với ba người làm ngữ liệu gán nhãn cho 100 câu. Các câu này được thu thập từ hai nguồn báo Tuổi Trẻ điện tử và sách ngữ pháp (tỉ lệ 50/50). Ba người đã tiến hành gán nhãn độc lập sau đó kết quả được chương trình đánh giá như sau:

Test	**A1-A2**	**A2-A3**	**A3-A1**
Full tags	90.32%	91.26%	90.71%
Constituent tags	92.40%	93.57%	91.92%
No tags	95.24%	96.33%	95.48%

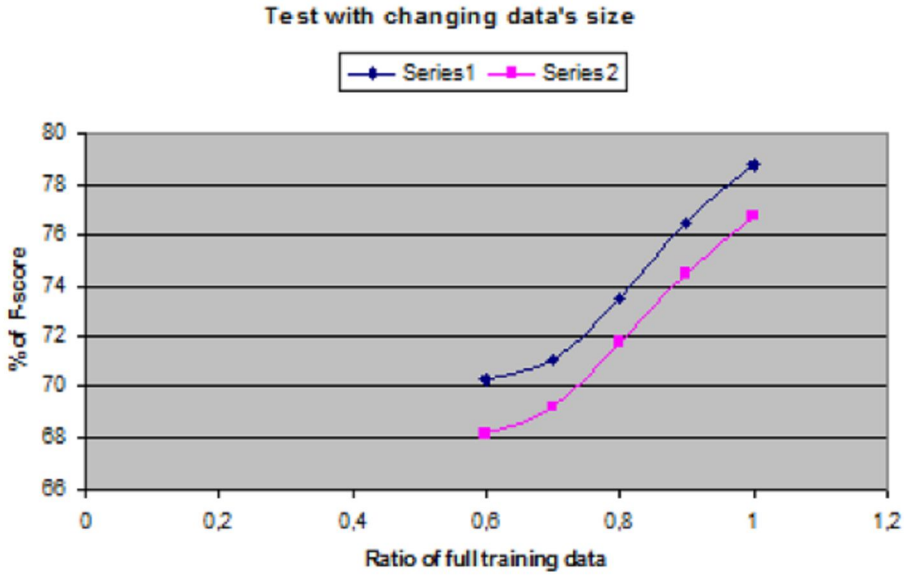
Bảng 4. Đánh giá độ đồng thuận

Kết quả này cho thấy độ đồng thuận khá cao, xấp xỉ độ đồng thuận của các kho ngữ liệu tiếng Trung và tiếng Anh.

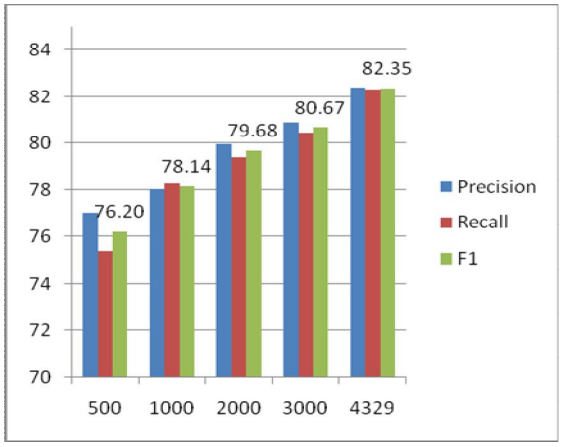
4.1.2 Đánh giá gián tiếp

Chất lượng kho ngữ liệu có thể được đánh giá gián tiếp dựa vào các công cụ được huấn luyện và test trên đó. Việc đánh giá thường dựa vào đường cong học (learning curve) thể hiện mối quan hệ giữa kích thước dữ liệu huấn luyện và độ chính xác. Nếu một công cụ đạt độ chính xác cao và mối tương quan giữa hai yếu tố đó được thể hiện rõ thì kho ngữ liệu cũng là đạt yêu cầu. Hình 2 thể hiện đường cong học của hệ phân tích cú pháp tiếng Việt (Lê Anh Cường và cộng sự, 2009). F-score đạt 78% nếu sử dụng toàn bộ dữ liệu huấn luyện và đường cong đi lên rất rõ. Tương tự, Hình 3 thể hiện đường cong học của hệ nhận biết

cụm danh từ tiếng Việt (Nguyễn Thị Hương Thảo và cộng sự, 2009). F-score đạt 82.35% nếu dùng toàn bộ dữ liệu huấn luyện và đồ thị thể hiện rõ rằng khi tăng dữ liệu huấn luyện thì F-score tăng.



Hình 2. Mối tương quan kích thước dữ liệu huấn luyện và F-score của hệ phân tích cú pháp tiếng Việt



Hình 3. Mối tương quan kích thước dữ liệu huấn luyện và F-score của hệ NP chunker tiếng Việt

4.2 Đánh giá chủ quan

Xem trong tài liệu đi kèm báo cáo đề tài.

5. Tài liệu tham khảo

Tiếng Việt

[1] Diệp Quang Ban. 2005. Ngữ pháp tiếng Việt (2 tập). *NXB Giáo dục*.

[2] Vũ Tiến Dũng. Tiếng Việt và ngôn ngữ học hiện đại sơ khảo về cú pháp. 2003. VIET Stuttgart – Germany.

[3] Cao Xuân Hạo. 2006. Tiếng Việt sơ thảo ngữ pháp chức năng. *NXB Khoa học Xã hội*.

[4] Nguyễn Văn Hiệp. Vài nét về lịch sử nghiên cứu cú pháp tiếng Việt. Tạp chí Ngôn ngữ, Hà Nội, số 10/2002.

[5] Nguyễn Kim Thản. 2008. Cơ sở ngữ pháp tiếng Việt. *NXB Khoa học Xã hội*.
[6] Nguyễn Minh Thuyết và Nguyễn Văn Hiệp. 1999. Thành phần câu tiếng Việt. *NXB ĐHQG Hà Nội*.
[7] Ủy ban Khoa học Xã hội Việt Nam. 1983. Ngữ pháp tiếng Việt. NXB Khoa học Xã hội.

Tiếng Anh

[8] Sabine Brants et al. The TIGER Treebank. 2003. *COLING*.
[9] David Chiang and Daniel M. Bikel. 2002. Recovering Latent Information in Treebanks. *COLING*.
[10] Michael Collins. 1999. Head-Driven Statistical Models for Natural Language Parsing. PhD thesis, University of Pennsylvania.
[11] Anh-Cuong Le, Phuong-Thai Nguyen, Hoai-Thu Vuong, Minh-Thu Pham, Tu-Bao Ho. 2009. An Experimental on Lexicalized Statistical Parsing for Vietnamese. *KSE (to appear)*.
[12] Chung-hye Han et al. Development and Evaluation of a Korean Treebank and its Application to NLP. 2002. *LREC*.
[13] Mitchell P. Marcus et al. Building a Large Annotated Corpus of English: The Penn Treebank. 1993. *Computational Linguistics*.
[14] L. H. Phuong, N. T. M. Huyen, R. Azim, H. T. Vinh. A hybrid approach to word segmentation of Vietnamese texts. *Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, Springer LNCS 5196, Tarragona, Spain, 2008.*
[15] Peter Sells. Lectures on Contemporary Syntactic Theories. 1987. CSLI.
[16]Phuong-Thai Nguyen, Xuan-Luong Vu, Thi-Minh-Huyen Nguyen, Van-Hiep Nguyen, Hong-Phuong Le. Building a Large Syntactically-Annotated Corpus of Vietnamese. *Proceedings of the 3rd Linguistic Annotation Workshop (LAW) at ACL-IJCNLP 2009 (to appear)*.
[17] Huong-Thao Nguyen, Phuong-Thai Nguyen, Quang-Thuy Ha, and Le-Minh Nguyen. 2009. Vietnamese Noun Phrase Chunking based on Conditional Random Fields. *KSE (to appear)*.
[18] Fei Xia et al. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. 2000. *COLING*.
[19] Nianwen Xue et al. Building a Large-Scale Annotated Chinese Corpus. 2002. *COLING*.