Xây dựng treebank tiếng Việt

Nguyễn Phương Thái[1], Vũ Xuân Lương[2], Nguyễn Thị Minh Huyền[3]

Tóm tắt

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. Treebank thường được dùng để xây dựng các hệ phân tích cú pháp chất lượng cao. Các hệ phân tích cú pháp này lại được sử dụng trong các ứng dụng quan trọng như truy vấn thông tin, dịch máy, v.v. Bài báo này liên quan đến việc xây dựng ngân hàng câu tiếng Việt được chú giải cú pháp. Bài báo trình bày một số kết quả ban đầu mà chúng tôi đã đạt được như: xây dựng tập nhãn từ loại, xây dựng tập nhãn cú pháp, xây dựng công cụ, triển khai gán nhãn. Trong phần đánh giá kết quả gán nhãn, bài báo chỉ ra là độ đồng thuận giữa những người gán nhãn còn chưa cao chứng tỏ còn nhiều vấn đề cần được giải quyết.

1. Giới thiệu

Tiếng Việt là ngôn ngữ mà thứ tự từ khá cố định do đó chúng tôi chọn xây dựng treebank gồm các cây thành phần. Đối với các ngôn ngữ mà thứ tự từ khá tự do như tiếng Nhật, Séc thì cây phụ thuộc 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, một ngôn ngữ thuộc họ Ấn-Âu; tiếng Trung, một họ ngôn ngữ riêng; tiếng Hàn; tiếng Ả-rập.

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. Hiện tại chúng tôi đã tiến hành gán nhãn dữ liệu được khoảng 2 ngàn câu. 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. Chúng tôi thu thập các bài báo của báo Tuổi Trẻ điện tử. Hiện tại chúng tôi đang gán nhãn cho các bài báo thuộc chủ đề Chính trị-Xã hội. Chúng tôi sẽ gán nhãn thêm cho chủ đề Kinh tế hay Tin học nữa.

Cấu trúc của bài báo này như sau. Trước hết, chúng tôi trình bày về tập nhãn từ loại và hướng dẫn gán nhãn từ loại. Thứ hai là phần tập nhãn cú pháp và hướng dẫn gán nhãn cú pháp. Thứ ba là về công cụ hỗ trợ người làm ngữ liệu. Thứ tư là về qui trình gán nhãn cú pháp. Thứ năm là kết quả đạt được cho tới thời điểm hiện tại. Cuối cùng là phần kết luận.

[1] Đại học Công nghệ, Đại học Quốc gia Hà Nội
[2] Trung tâm Từ điển học
[3] Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội

2. Hướng dẫn gán nhãn từ loại và cú pháp

2.1 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 có hai quan điểm:

- Quan điểm không phân từ loại, phủ nhận sự tồn tại của nó (*Lê Quang Trinh, Nguyễn Hiến Lê, Hồ Hữu Tùng*)
- Quan điểm phân từ loại (rất nhiều nhà ngôn ngữ học):
    - dựa vào khả năng kết hợp và chức vụ ngữ pháp (gọi chung là thái độ ngữ pháp). Ngoài ra một số nghiên cứu về đối sánh ngôn ngữ học còn nêu lên hiện tượng "biến đổi hình thái" từ tiếng Việt với sự tham gia của từ chức năng.
    - dựa vào nghĩa khái quát

Chúng tôi theo quan điểm phân từ loại khi xây dựng treebank tiếng Việt. Về nguyên tắc, các thông tin về từ có thể được chứa trong nhãn từ loại bao gồm: từ loại cơ sở (danh từ, động từ, v.v.), thông tin hình thái (số ít, số nhiều, thì, ngôi, v.v.), thông tin về phân loại con (ví dụ động từ đi với danh từ, động từ đi với mệnh đề, v.v.), thông tin ngữ nghĩa, hay một số thông tin cú pháp khác. Chúng tôi xây dựng tập nhãn từ loại chỉ chứa thông tin về từ loại cơ sở mà không bao gồm các thông tin như hình thái, phân loại con, v.v. Tập nhãn từ loại của chúng tôi được liệt kê trong Bảng 1, tổng số nhãn là 17.

STT	Tên	Chú thích
1	N	Danh từ
2	Np	Danh từ riêng
3	Nc	Danh từ chỉ loại
4	Nu	Danh từ đơn vị
5	V	Động từ
6	A	Tính từ
7	P	Đại từ
8	L	Định từ
9	M	Số từ
10	R	Phụ từ
11	E	Giới từ
12	C	Liên từ
13	I	Thán từ
14	T	Trợ từ, tiểu từ, từ tình thái
15	U	Từ đơn lẻ
16	Y	Từ viết tắt
17	X	Các từ không phân loại được

Bảng 1. Tập nhãn từ loại

2.2 Tập nhãn cú pháp

Nhãn thành phần cú pháp

Loại nhãn này mô tả các thành phần cú pháp cơ bản là cụm từ và mệnh đề. Nhãn thành phần cú pháp là thông tin cơ bản nhất trên cây cú pháp, nó tạo thành xương sống của cây

cú pháp[4]. Tập nhãn cú pháp của các ngôn ngữ khác nhau là khác nhau (ở một tỉ lệ nhất định) vì hai nguyên nhân. Nguyên nhân cơ bản nhất là do sự khác biệt về ngôn ngữ. Chẳng hạn như trong tiếng Trung, từ chỉ loại có chức năng làm bổ nghĩa trước cho danh từ. Từ chỉ loại lại có thể được kết hợp với số từ trong phần phụ trước của cụm danh từ. Vì vậy nhóm thiết kế Chinese Treebank (CTB) đã đặt ra nhãn cụm từ chỉ loại. Đây là một điểm khác biệt với treebank tiếng Anh (PTB). Nguyên nhân thứ hai là do kỹ thuật thiết kế tập nhãn. Chẳng hạn như với các cụm từ nghi vấn, PTB có 4 loại nhãn là WHNP, WHPP, WHADJP, WHADVP. Trong khi CTB lại chỉ đặt ra một nhãn chức năng là WH. Nhãn này sẽ được dùng kèm với nhãn cụm từ khi trong cụm từ đó có từ dùng để hỏi. Như vậy vẫn đủ để mô tả các cụm từ nghi vấn (NP-WH, PP-WH, ADJP-WH, ADVP-WH). Bảng 2 liệt kê tập nhãn cụm từ và Bảng 3 là nhãn mệnh đề của chúng tôi.

STT	Tên	Chú thích
	NP	Cụm danh từ
	VP	Cụm động từ
	AP	Cụm tính từ
	RP	Cụm phụ từ
	PP	Cụm giới từ
	QP	Cụm từ chỉ số lượng
	MDP	Cụm từ tình thái
	WHNP	Cụm danh từ nghi vấn (ai, cái gì, con gì, v.v.)
	WHAP	Cụm tính từ nghi vấn (lạnh thế nào, đẹp ra sao, v.v.)
	WHRP	Cụm từ nghi vấn dùng khi hỏi về thời gian, nơi chốn, v.v.
	WHPP	Cụm giới từ nghi vấn (với ai, bằng cách nào, v.v.)

Bảng 2. Tập nhãn cụm từ

STT	Tên	Chú thích
	S	Câu trần thuật (khẳng định hoặc phủ định)
	SQ	Câu hỏi
	SBAR	Mệnh đề phụ (bổ nghĩa cho danh từ, động từ, và tính từ)

Bảng 3. Tập nhãn mệnh đề

[4] Nhiều lý thuyết về cú pháp dựa trên cấu trúc xương sống này.

Nhãn chức năng cú pháp

Nhãn chức năng của một thành phần cú pháp cho biết vai trò của nó trong thành phần cú pháp mức cao hơn. Nhãn chức năng cú pháp được gán cho các thành phần chính trong câu như chủ ngữ, vị ngữ, tân ngữ. Nhờ thông tin do nhãn chức năng cung cấp ta có thể xác định các loại quan hệ ngữ pháp cơ bản sau đây:

- Chủ-vị
- Đề-thuyết
- Phần chêm
- Bổ ngữ
- Phụ ngữ
- Sự kết hợp

STT	Tên	Chú thích
1	SUB	Nhãn chức năng chủ ngữ
2	DOB	Nhãn chức năng tân ngữ trực tiếp
3	IOB	Nhãn chức năng tân ngữ gián tiếp
4	TPC	Nhãn chức năng chủ đề
5	PRD	Nhãn chức năng vị ngữ không phải cụm động từ
6	LGS	Nhãn chức năng chủ ngữ logic của câu ở thể bị động
7	EXT	Nhãn chức năng bổ ngữ chỉ phạm vi hay tần suất của hành động
8	H	Nhãn phần tử trung tâm (của cụm từ hoặc mệnh đề)
9-12	TC, CMD, EXC, SPL	Nhãn phân loại câu: đề-thuyết, mệnh lệnh, cảm thán, đặc biệt
13	TTL	Tít báo hay tiêu đề
14	VOC	Thành phần than gọi

Bảng 4. Nhãn chức năng cú pháp

Ngoài ra nhãn chức năng cũng có thể tương ứng với một loại trạng ngữ nào đó như thời gian, nơi chốn, hay mục đích. Như vậy loại nhãn chức năng này chứa thông tin ngữ nghĩa "nông" của một thành phần cú pháp. Bảng 5 liệt kê các nhãn chức năng trạng ngữ mà chúng tôi sử dụng.

STT	Tên	Chú thích
1	TMP	Nhãn chức năng trạng ngữ chỉ thời gian

2	LOC	Nhãn chức năng trạng ngữ chỉ nơi chốn
3	DIR	Nhãn chức năng trạng ngữ chỉ hướng
4	MNR	Nhãn chức năng trạng ngữ chỉ cách thức
5	PRP	Nhãn chức năng trạng ngữ chỉ mục đích hay lý do
6	ADV	Nhãn chức năng trạng ngữ nói chung (dùng khi trạng ngữ không thuộc một trong các loại cụ thể trên)

Bảng 5. Nhãn chức năng trạng ngữ

Nhãn thành phần rỗng

Đây là một loại thành phần khá đặc biệt. Nó chỉ ra sự tồn tại (được ngầm hiểu) của một thành phần cú pháp cho dù nó không xuất hiện ở vị trí đó. Thông thường thành phần rỗng được gán chỉ số của thành phần mà nó đại diện. Dưới đây là một ví dụ:

Tôi đã mua quyển sách mà thầy giáo giới thiệu .

```
(S (NP-SBJ Tôi)
   (VP đã mua
      (NP (NP-OBJ-1 quyển sách)
         (SBAR mà
            (S (NP-SBJ thầy giáo)
               (VP giới thiệu
                  (NP-OBJ *T*-1))))))
   (. .))
```

Trong ví dụ trên đại từ “Tôi” có nhãn chức năng là SBJ cho biết nó là chủ từ trong câu, còn danh từ “quyển sách” có nhãn chức năng OBJ cho biết nó là danh từ làm tân ngữ.

2.3 Xây dựng tài liệu hướng dẫn gán nhãn

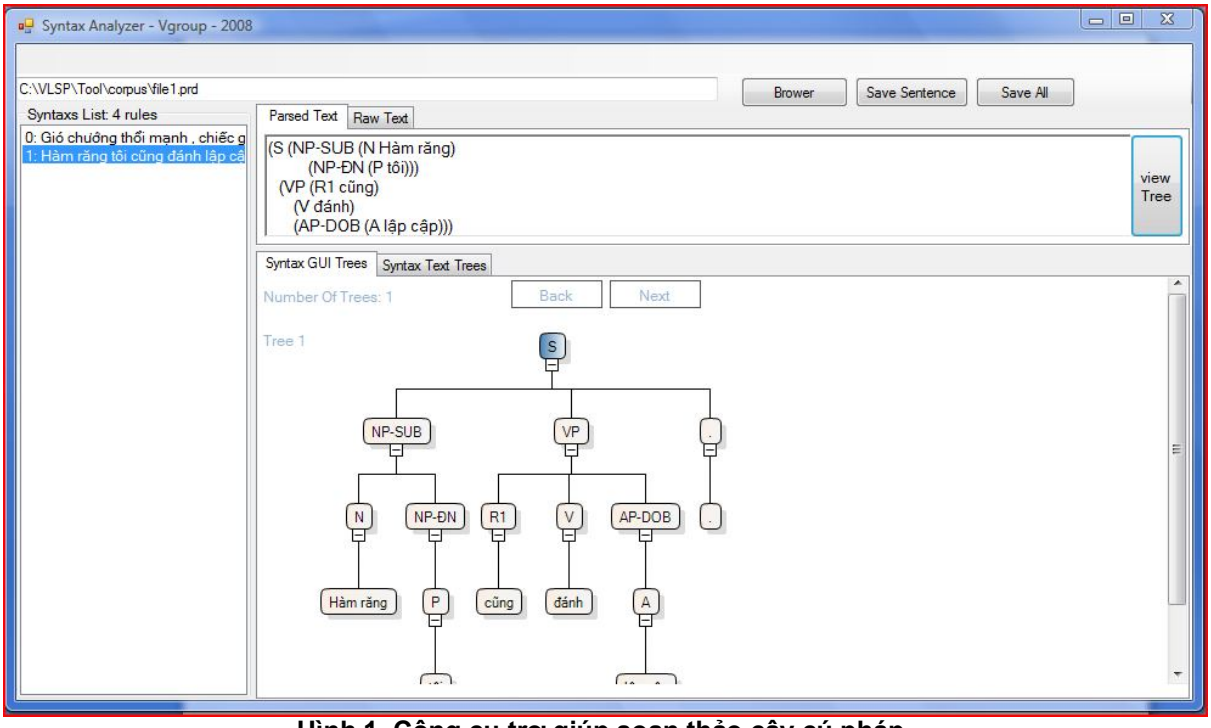
Đây là một tài liệu rất quan trọng bao gồm không chỉ các thông tin về tập nhãn, mà còn hướng dẫn gán nhãn cho các hiện tượng cụ thể với các ví dụ minh họa. Để xây dựng tài liệu này, trước tiên chúng tôi nghiên cứu các tài liệu về ngữ pháp và kinh nghiệm xây dựng treebank đã có. Ngoài ra chúng tôi còn cộng tác chặt chẽ với các nhà ngôn ngữ để xử lý các hiện tượng khó. Khi gặp hiện tượng khó và có một vài lựa chọn, chủ động chọn một cái và khi cần thì chuyển đổi sang cái kia. Những người gán nhãn được khuyến khích đưa ra các câu hỏi trong quá trình làm việc.

Khi xây dựng phiên bản đầu tiên của tài liệu này, nhóm thiết kế đã 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 xung trong quá trình gán nhãn văn bản.

Với mỗi hiện tượng ngữ pháp, chúng tôi trình bày cách nhận diện và cách gán nhãn cùng với các ví dụ cụ thể để minh họa. Các ví dụ được lấy từ sách ngữ pháp hoặc từ ngữ liệu thực tế. Khi có thể, chúng tôi cố gắng trích dẫn tài liệu tham khảo để người đọc có thể nắm được đầy đủ hơn về vấn đề được nêu.

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. Hiện tại chúng tôi chưa sử dụng công cụ gán nhãn tự động nhưng sẽ sớm đưa vào trong thời gian sắp tới.



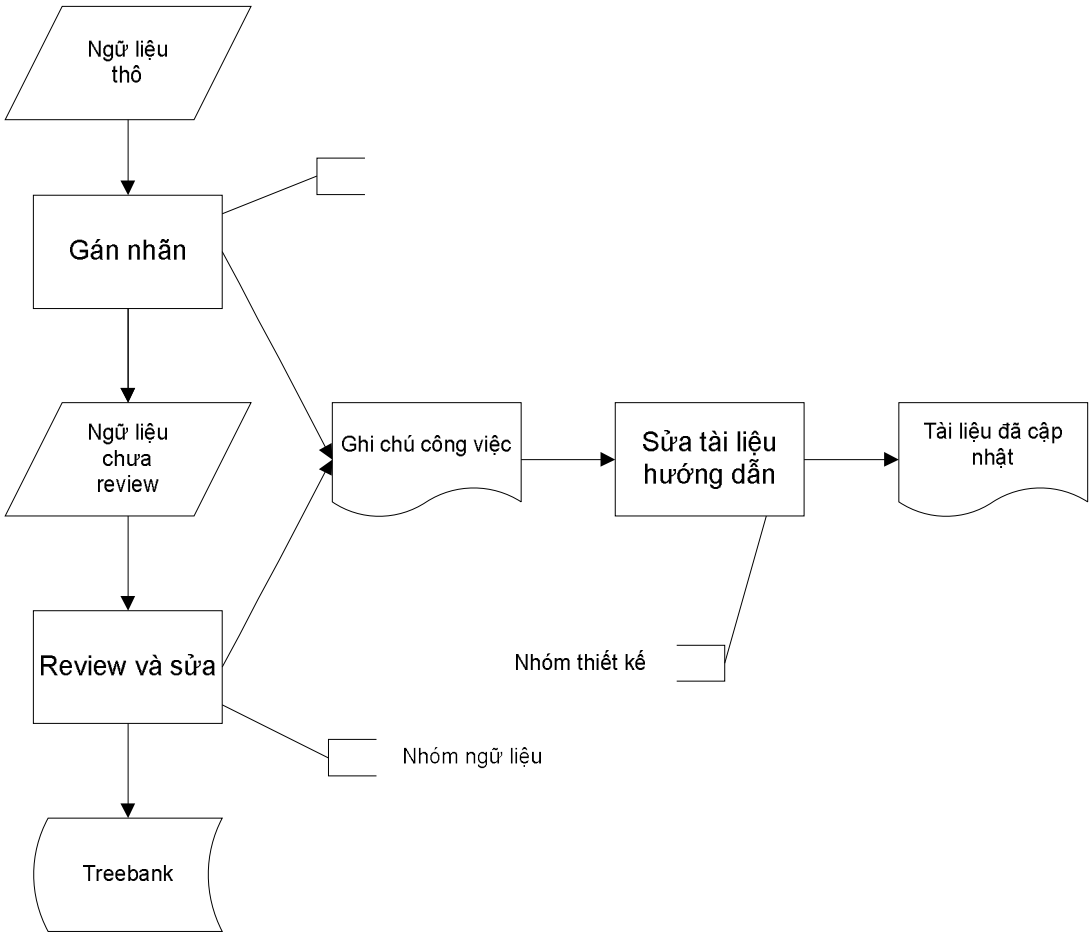
Hình 1. Công cụ trợ giúp soạn thảo cây cú pháp

Chương trình này có ba cửa sổ chính. Thứ nhất là cửa sổ bên trái hiển thị danh sách câu trong file vào. Người làm ngữ liệu click vào câu nào thì các thông tin tương ứng được hiển

thị ở bên phải. Cửa sổ phía trên bên phải (Parsed Text) hiển thị cây cú pháp dạng văn bản và cho phép sửa cây đó. Cửa sổ dưới bên phải (Syntax GUI Trees) hiển thị cây cú pháp dạng đồ họa. Sau khi sửa đổi cây có thể được lưu vào bộ nhớ trong và kết thúc phiên làm việc với file thì ghi ra đĩa cứng.

4. Quá trình gán nhãn

Quá trình gán nhãn một câu gồm ba bước: tách từ, gán nhãn từ loại, và phân tích cú pháp. Qui trình thực hiện gán nhãn là tương tự nhau, tuy nhiên mỗi bước yêu cầu những kiến thức và có những đặc trưng riêng. Trước tiên, những người gán nhãn cần được huấn luyện về cách gán nhãn, tập nhãn, và cách sử dụng công cụ. Sau đó họ sẽ gán nhãn cho từng phần của corpus thô. Quá trình gán nhãn được thể hiện trong Hình 2. Mỗi người làm có 1 người review và sửa lỗi. Những trường hợp không chắc chắn thì ghi lại để thảo luận với nhóm thiết kế. Người review được yêu cầu có con mắt phê phán khi làm việc. Họ có tinh thần làm việc nhóm cao vừa để gán nhãn chính xác vừa để giúp cải tiến tài liệu hướng dẫn.



Hình 2. Sơ đồ quá trình làm ngữ liệu

Khi gán nhãn, người làm dữ liệu cần:

- Hiểu đúng câu trước khi phân tích, nếu cần thì biến đổi câu để hiểu đúng nó (thêm từ, bớt từ, thay thế từ, đổi thứ tự từ)
- Nhận dạng mẫu (đặc biệt là động từ): chẳng hạn nếu ta đã biết các mẫu động từ đi với danh từ, động từ đi với cụm giới từ, động từ đi với mệnh đề thì cũng là căn cứ ra quyết định.

Khi review, người làm dữ liệu cần chú ý kiểm tra các điểm sau:

- Sai tách từ không?
- Sai từ loại không?
- Có lỗi liên kết cụm từ không?
- Có sai nhãn cú pháp nào không?
- Có thiếu gì không? (nhãn H, nhãn chức năng trạng ngữ, v.v.)

5. 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. Chúng tôi đã cài đặt một chương trình bằng C++ thực hiện tự động việc đánh giá này.

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)

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:

	Người 1-Người 2	*Người 2-Người 3*	*Người 3- Người 1*
Nhãn đầy đủ	0.54	0.62	0.59
Bỏ qua nhãn chức năng	0.66	0.69	0.69
Không tính nhãn	0.74	0.75	0.76

Bảng 6. Đánh giá độ đồng thuận

Kết quả này cho thấy độ đồng thuận chưa cao. Cần cải tiến tài liệu hướng dẫn gán nhãn và huấn luyện người gán nhãn kỹ hơn.

6. Kết luận

Trong bài báo này chúng tôi đã trình bày những kết quả ban đầu trong quá trình xây dựng treebank tiếng Việt. Nhiều chi tiết kỹ thuật đã được bỏ qua vì giới hạn khuôn khổ của bài báo. Hiện tại chúng tôi vẫn còn nhiều vấn đề phải giải quyết để có thể đạt các mục tiêu đã đề ra. Trong tương lai, khi có điều kiện thì chúng tôi sẽ mời các chuyên gia nước ngoài cố vấn, trực tiếp có những trao đổi với các nhóm đã xây dựng thành công treebank của nước họ. Chúng tôi cũng sẽ nhanh chóng đưa công cụ gán nhãn tự động vào hỗ trợ người làm dữ liệu. Thêm vào đó là cải tiến công cụ soạn thảo cây cú pháp trực quan giúp người làm dữ liệu sửa cây cú pháp nhanh hơn.

Lời cảm ơn

Bài báo này có được là nhờ sự hỗ trợ kinh phí của đề tài nhánh SP7.3 thuộc đề tài nhà nước "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt", mã số KC01.01/06-10.

Tài liệu tham khảo

[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.

[8] Sabine Brants et al. The TIGER Treebank. 2003. COLING.

[9] Chung-hye Han et al. Development and Evaluation of a Korean Treebank and its Application to NLP. 2002. LREC.

[10] Mitchell P. Marcus et al. Building a Large Annotated Corpus of English: The Penn Treebank. 1993. Computational Linguistics.

[11] Peter Sells. Lectures on Contemporary Syntactic Theories. 1987. CSLI.

[12] Fei Xia et al. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. 2000. COLING.

[13] Nianwen Xue et al. Building a Large-Scale Annotated Chinese Corpus. 2002. COLING.