Tạp chí Tin học và Điều khiển học, T.28, S.4 (2012), 297–309

KẾT HỢP PHÂN ĐOẠN DIỄN NGÔN VỚI BỘ PHÂN TÍCH CÚ PHÁP LIÊN KẾT ĐỂ PHÂN TÍCH CÂU GHÉP NHIỀU MỆNH ĐỀ TIẾNG VIỆT

NGUYỄN THỊ THU HƯƠNG[1], NGUYỄN THÚC HẢI[1], NGUYỄN THANH THỦY[2]

[1] *Trường Đại học Bách khoa Hà Nội*
[2] *Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội*

Tóm tắt. Văn phạm liên kết là lý thuyết về cú pháp mà phân tích của mỗi câu là một tập các mối liên kết giữa các từ. Văn phạm liên kết tương tự như văn phạm phụ thuộc nhưng không định hướng mối quan hệ giữa các từ. Loại văn phạm này bao quát được hầu hết các đặc trưng cú pháp và từ pháp trong tiếng Việt. Bộ phân tích cú pháp liên kết cho phép phân tích câu đơn và câu ghép hai mệnh đề được xây dựng đã cho kết quả phân tích khá tốt. Bài báo trình bày các kết quả đạt được khi mở rộng chức năng của bộ phân tích cú pháp liên kết tiếng Việt để phân tích các dạng câu ghép gồm nhiều mệnh đề. Các mệnh đề được phân tách dựa trên giải thuật phân đoạn diễn ngôn mức câu. Việc phân tích cú pháp riêng biệt từng mệnh đề rồi kết hợp lại thành phân tích tổng thể cho phép khử nhập nhằng liên hợp, đồng thời làm giảm độ phức tạp tính toán.

Abstract. Link grammar is a theory of syntax which builds relations between pairs of words, rather than constructing constituents in a tree-like hierarchy. Link grammar is similar to dependency grammar, but dependency grammar includes a head-dependent relationship, as well as lacking directionality in the relations between words. A link parser has been built for Vietnamese with acceptable results. In this paper, we propose an extended link parser with a new function to parse complex, compound and complex - compound sentences, except sentences with nested clauses. Sentences are segmented into clauses using discourse segmentation algorithm of sentence level. Parsing clauses separately is useful for coordination disambiguating and decreasing time complexity.

Keywords. link grammar, link parser, complex sentence, compound sentence.

1. GIỚI THIỆU

Trong văn bản tiếng Việt, câu chứa hai nòng cốt trở lên chiếm tỷ lệ rất cao. Việc phân tích cú pháp câu nhiều nòng cốt phức tạp hơn nhiều so với câu đơn. Với những loại câu gồm hai nòng cốt trở lên, tiếng Anh phân loại theo mối quan hệ giữa hai mệnh đề. Nếu mối quan hệ là song song (dùng các từ nối "and", "or", "not only... but also"...), câu được gọi là "câu ghép" (compound sentence). Nếu các mối liên hệ có tính chất chính-phụ (dùng các từ nối "if","then", "because"...), câu được gọi là "câu phức hợp" (complex sentence). Câu ghép phức hợp (complex-compound sentence) phức tạp hơn nhiều khi chứa ít nhất hai mệnh đề song song và ít nhất một mệnh đề phụ. Phân loại câu tiếng Việt có chút khác biệt so với tiếng Anh. Diệp Quang Ban[1] phân biệt câu ghép là câu chứa từ hai nòng cốt trở lên, trong đó không

nòng cốt nào bao nhau và câu phức chứa hai nòng cốt trở lên nhưng tồn tại một nòng cốt bao các nòng cốt còn lại. Ví dụ, câu "Tôi đang đứng chờ xe thì một cậu bạn chạy đến" được xếp vào loại câu ghép trong khi câu "Con mèo tôi mua chạy mất rồi" được xếp vào loại câu phức. Việc phân định ranh giới mệnh đề trong câu phức có thể đòi hỏi một bộ ngữ liệu lớn với phương pháp học máy nên chưa được đề cập đến trong bài báo này.

Theo quan điểm của Nguyễn Chí Hòa [3], Trần Ngọc Thêm [13], mệnh đề là đơn vị nhỏ nhất của văn bản, và câu ghép được xây dựng nên từ các "khối", mỗi "khối" là một mệnh đề. Nòng cốt ghép có thể là song song với hai hay nhiều vế, cũng có thể là chính phụ với đúng hai vế [13,15].

Đối với mô hình văn phạm phi ngữ cảnh truyền thống, mệnh đề phụ trong câu ghép có thể được sản sinh từ ký hiệu không kết thúc đặc biệt SBAR của văn phạm. Với một tập luật rất lớn, việc nhập nhằng về giới hạn của mệnh đề rất thường xảy ra. Cũng do tập ký hiệu không kết thúc lớn, cây phân tích cho câu ghép nhiều mệnh đề rất phức tạp. Điều đó sẽ ảnh hưởng đến tốc độ và kết quả của các xử lý khác như phân loại văn bản, tóm tắt văn bản, dịch máy - những bài toán xử lý dựa trên cấu trúc cú pháp của câu.

Mô hình văn phạm phụ thuộc hiện đang rất phổ biến trong phân tích cú pháp vì nhiều lý do: cây phân tích đơn giản (không có tập ký hiệu không kết thúc), biểu diễn dễ dàng các phụ thuộc không lân cận (long distance dependency), biểu diễn được các quan hệ về hình thái hay ngữ nghĩa [10]...

Mô hình văn phạm liên kết được D.Sleator và D.Temperley [12] đưa ra là mô hình theo hướng tiếp cận phụ thuộc. Điểm đặc biệt của bộ phân tích cú pháp liên kết là có thể phân tích một số dạng câu ghép chính phụ thông qua một số liên kết đặc biệt như CO (liên kết giữa thành phần gợi mở và chủ ngữ của mệnh đề đứng sau), CC(liên kết các mệnh đề với liên từ kết hợp)... được xác lập cho các từ nối như "because", "although", "but"... Bộ phân tích cú pháp tiếng Việt do chúng tôi xây dựng [6] cũng nhận được kết quả tương tự cho tiếng Việt. Tuy nhiên với loại câu ghép có nhiều mệnh đề, quan hệ phức tạp như "Nếu cán bộ, công chức được tuyển dụng lại vào làm việc ở cơ quan, đơn vị cũ, thì thời gian thực tế học tập theo chương trình đào tạo (ghi trên chứng chỉ hoặc bằng đào tạo được cấp) được tính vào thời gian xét nâng bậc lương thường xuyên", bộ phân tích cú pháp liên kết không thực hiện được. Hơn nữa, việc chỉ sử dụng liên kết đơn thuần của từ nối sẽ đòi hỏi thời gian tính toán rất lớn. Nếu phân tích riêng từng mệnh đề của câu ghép rồi tổ hợp lại thành một phân tích tổng thể, những vấn đề nói trên có thể giải quyết được.

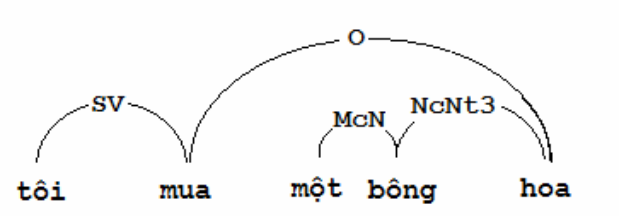
Xuất phát từ đặc điểm của tiếng Việt là hầu hết các giới hạn mệnh đề trong câu ghép có thể phát hiện nhờ dấu hiệu diễn ngôn, kết hợp với một số đặc trưng cú pháp, chúng tôi đã cải tiến giải thuật phân đoạn diễn ngôn [9] ở mức câu để xây dựng cây diễn ngôn của câu. Từ cây diễn ngôn, một phân tích hoàn chỉnh cho toàn bộ câu được xây dựng nhờ kết hợp phân tích liên kết của từng mệnh đề với các kết nối lớn thể hiện quan hệ giữa các mệnh đề với nhau.

Sau đây là phần trình bày về mô hình văn phạm liên kết, bộ phân tích cú pháp liên kết mở rộng với sự kết hợp của lý thuyết cấu trúc diễn ngôn trong phân tách mệnh đề cũng như xây dựng cây diễn ngôn cho câu.

2. VĂN PHẠM LIÊN KẾT

2.1. Một số định nghĩa

Văn phạm liên kết bao gồm một tập các từ, mỗi từ có một yêu cầu liên kết. Một câu được định nghĩa bởi văn phạm nếu tồn tại một cách để vẽ các cung (liên kết) phía trên các từ thoả mãn những điều kiện sau:
+ Tính phẳng (planarity): các liên kết không giao nhau khi được vẽ phía trên các từ.
+ Tính liên thông (connectivity): các liên kết nối tất cả các từ trong câu với nhau.
+ Tính thoả mãn (satisfaction): các liên kết thoả mãn các yêu cầu liên kết của mỗi từ trong câu.
+ Tính thứ tự (ordering): khi các kết nối của một công thức (xem bảng 1) được duyệt từ trái qua phải, các từ mà nó kết nối tới tiến từ gần ra xa.
+ Tính loại trừ (exclusion): không có hai liên kết có thể kết nối cùng một cặp từ. Ví dụ, phân tích của câu "Tôi mua một bông hoa" được thể hiện trong hình 1 dưới đây:



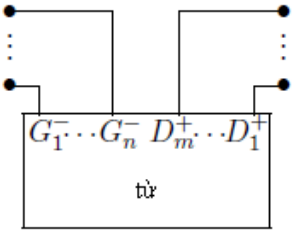
Hình 1. Phân tích câu "*tôi mua một bông hoa*" trong văn phạm liên kết

Ý nghĩa các kết nối trong hình 1 như sau:
SV: Kết nối chủ ngữ (là danh từ hoặc đại từ xưng hô) với động từ chính trong câu.
O: Kết nối vị ngữ và bổ ngữ trực tiếp.
McN: Kết nối số từ và danh từ.
NcNt3: Kết nối danh từ chỉ loại (bông, con, quyển...) với danh từ cụ thể.
Một cách hình thức, một câu hay cụm từ thỏa mãn các yêu cầu của văn phạm liên kết là một mạng liên kết [2].

Định nghĩa 1. Cho Σ là một bảng chữ , Pr là tập các kiểu nguyên thủy và $(\nu, <)$ là một tập sắp thứ tự hoàn toàn. Mạng liên kết là cấu trúc (V, w, E, t) trong đó
+ Tập đỉnh $V \subseteq \nu$ là tập con hữu hạn không rỗng của ν, ký hiệu tập đỉnh V là $(v_1, ...v_n)$;
+ n là số phần tử của V, và $v_1 < ... < v_n$;
+ $w : V \to \Sigma$ ánh xạ mỗi đỉnh với một từ;
+ Tập cung $E \subseteq V \times V$ là tập con đối xứng và phản phản xạ của $V \times V$;
+ $t : E \to Pr$ ánh xạ mỗi cung tới một kiểu nguyên thủy;
+ Các cung không giao nhau: nếu $(a, b) \in E$ và $(c, d) \in E$ sao cho $a < b$ và $c < d$ thì không xảy ra $a < c < b < d$ hay $c < a < d < b$;
+ Đồ thị (V, E) liên thông;

Khi sử dụng khái niệm mạng liên kết để biểu diễn quan hệ liên kết trong câu,, mỗi đỉnh của mạng liên kết biểu diễn một từ, quan hệ $v_1 < ... < v_n$ thể hiện thứ tự của các từ trong câu. Trong ví dụ ở hình 1, mạng liên kết sẽ gồm 5 đỉnh v_1(tôi), v_2(mua), v_3(một) v_4(bông), v_5(hoa), tập Pr các kiểu nguyên thủy chứa $\{SV, O, McN, NcNt3\}$.

Định nghĩa 2. Nút liên kết trên Pr là cặp hai danh sách hữu hạn của Pr. Mỗi nút liên kết X có một danh sách trái các cổng ký hiệu là $t_n^-, ..., t_1^-$ và một danh sách phải các cổng ký hiệu là $t_1^+, ..., t_m^+$.



Hình 2. Nút liên kết

Tập các nút liên kết trên Pr được ký hiệu là Tp.

Với mỗi đỉnh ν của mạng liên kết $N = (V, w, E, t)$, các cung liên quan đến ν ở bên trái là $(x_n, v), ..., (x_1, v)$ và bên phải là $(v, y_1), ..., (v, y_m)$, với $x_n < x_{n-1} <, ..., < x_1 < v < y_1 < ... < y_{m-1} < y_m$. Ta nói ν liên hệ với nút liên kết $node(\nu) = t(x_n, v)^- ... t(x_1, v)^- t(v, y_1)^+ ... + t(v, y_m)^+$. Trong ví dụ ở hình 1, $node(v_2) = SV(v_1, v_2)^-, O(v_2, v_5)^+$.

Định nghĩa 3. Văn phạm liên kết là cấu trúc $G = (\Sigma, I)$ với $I : \Sigma \to P^f(Tp)$ là hàm ánh xạ mỗi phần tử của Σ vào một nút liên kết.

Trong thực tế, để biểu diễn một văn phạm liên kết, tức là toàn bộ các yêu cầu liên kết của tất cả các từ thuộc ngôn ngữ người ta dùng một từ điển, mỗi từ được đi kèm với một công thức. Bảng 1 dưới đây chứa các công thức của một văn phạm liên kết mini chỉ chứa 5 từ. Trong các công thức này, ký hiệu "+" đòi hỏi một kết nối về bên phải, còn ký hiệu "–" đòi hỏi một kết nối về bên trái, {} cho biết một kết nối có thể xuất hiện hoặc không.

Những câu "tôi mua một bông hoa", "tôi mua hoa", "tôi mua bông hoa" hay cụm từ "một bông hoa" đều được biểu diễn bởi văn phạm ở bảng 1.

*B*ảng 1. Công thức liên kết của các từ

Từ	**Công thức**
tôi	SV^+
mua	$SV^- \& \{O^+\}$
một	$McNt^+$
bông	$NcNt3^+$
hoa	$O - \& \{McNt^- \& NcNt3^-\}$

Theo [12], để dễ dàng xử lý tự động, công thức được chuyển thành các dạng tuyển (disjunctive form). Một dạng tuyển bao gồm hai danh sách có thứ tự của các tên liên kết: danh sách liên kết bên trái $(L_1, L_2, ..., L_m)$ và danh sách liên kết bên phải $(R_n, R_{n-1}, ..., R_1)$, trong đó $L_1, L_2, ..., L_m$ là các kết nối về phía trái và $R_n, R_{n-1}, ..., R_1$ là các kết nối về phía phải. Dạng tuyển có thể coi là một biến thể của nút liên kết trong Định nghĩa 2.

Ví dụ, $((O, McNt, NcNt3)())$ là một dạng tuyển suy từ công thức của từ "hoa" trong bảng 1, trong khi nút liên kết của nó là $(O^-, McNt^-, NcNt3^-)()$.

Văn phạm liên kết thuộc dòng văn phạm phụ thuộc. Phân tích liên kết đơn giản hơn nhiều so với cây ngữ cấu do văn phạm liên kết không chứa tập ký hiệu không kết thúc. Văn phạm liên kết thể hiện rất tốt mối liên hệ giữa các từ lân cận, thể hiện được hầu hết mối liên hệ giữa các từ không lân cận.

Khác với văn phạm phụ thuộc, văn phạm liên kết là hoàn toàn từ vựng hóa nên có thể biểu diễn các công thức liên kết cho riêng từng từ, qua đó thể hiện nhiều hiện tượng đặc biệt của tiếng Việt. Ví dụ, kết nối SHA giữa danh từ chỉ bộ phận cơ thể "tay" và danh từ chỉ người "cô giáo" đã biểu diễn được quan hệ sở hữu ẩn(không chứa giới từ sở hữu "của") trong cụm từ "tay cô giáo" hay công thức liên kết $McNt^{-}\&NcNt3^{-}$ đảm bảo số từ khi đi với các danh từ chỉ thực vật như "hoa", "lúa"... phải đi kèm danh từ chỉ đơn vị của danh từ chỉ thực vật "cây", "bông"...

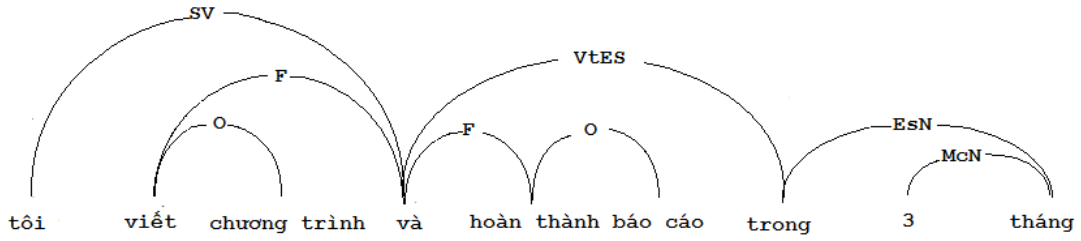
Phân tích cú pháp dựa trên văn phạm liên kết có đặc điểm là không gán nhãn từ trước mà bộ phân tích sẽ xác định chức năng cú pháp của từ thông qua các kết nối, do vậy tránh được nhiều lỗi do bộ gán nhãn từ. Bộ phân tích cú pháp liên kết [6] dựa trên giải thuật của [12] làm việc hiệu quả trên câu đơn và câu ghép hai mệnh đề, đồng thời giải quyết một phần vấn đề nhập nhằng liên hợp khi nhiều từ cùng đóng một vai trò.

Do cung biểu diễn liên kết không định hướng, mạng liên kết có thể chứa chu trình. Đặc điểm này cho phép biểu diễn sự phụ thuộc về ngữ nghĩa giữa các từ trong câu mà không cần phân chia thành nhiều tầng như văn phạm phụ thuộc. Cũng nhờ đặc điểm này mà việc biểu diễn liên kết giữa các mệnh đề dễ dàng hơn biểu diễn quan hệ phụ thuộc giữa hai mệnh đề, đặc biệt trong câu ghép song song.

2.2. Kết nối lớn

Để giải quyết vấn đề nhập nhằng liên hợp, xảy ra với từ "and","or" ("và", "hoặc" trong tiếng Việt) hay dấu phảy, mô hình văn phạm liên kết [12] chấp nhận kết nối giữa các cụm từ gọi là kết nối lớn (fat connector).

Ví dụ, khi phân tích câu "Tôi viết chương trình và hoàn thành báo cáo trong 3 tháng", kết nối lớn F được vẽ trên động từ "viết" của cụm từ "viết chương trình" và động từ "hoàn thành" của cụm từ "hoàn thành báo cáo" (xem hình 3). Dù có quan hệ với cả hai động từ, từ "tôi" được vẽ liên kết sang từ "và" để tránh sự giao nhau của các liên kết.



Hình 3. Phân tích câu "Tôi viết chương trình và hoàn thành báo cáo trong 3 tháng" có sử dụng kết nối lớn

3. PHÂN TÍCH CÚ PHÁP LIÊN KẾT CHO CÂU GHÉP

Một trong những lý do để mô hình văn phạm liên kết được sử dụng trong nhiều ứng dụng trên nhiều ngôn ngữ khác nhau là khả năng phân tích câu ghép. Với các quan hệ liên kết được đặt cho từ nối, bộ phân tích của [12] đã đưa ra phân tích cho hầu hết các câu ghép chính phụ chứa hai mệnh đề. Tuy nhiên, chỉ với liên kết của từ nối, bộ phân tích liên kết không thể thực hiện phân tích câu ghép với nhiều mệnh đề hay câu ghép song song chỉ gồm hai mệnh đề.

Với mô hình văn phạm kiểu phụ thuộc, quá trình phân tích cú pháp cho câu ghép bao gồm các giai đoạn: phân chia thành các mệnh đề, phân tích riêng từng mệnh đề, ghép kết quả phân tích cho các mệnh đề thành phân tích tổng thể cho toàn câu.

Việc tách câu thành các mệnh đề trong văn phạm phụ thuộc dựa trên các từ gợi ý, cấu trúc mệnh đề hoặc phương pháp học máy. Trong tiếng Nhật [14], tiếng Hàn [11] mệnh đề được tách dựa trên việc tính xác suất của một từ có thể nằm ở biên. Với các phương pháp này, có thể nhận biết giới hạn mệnh đề trong câu phức với các nòng cốt bao nhau. Hiện tiếng Việt chưa có bộ ngữ liệu mẫu lớn cho việc phân tách mệnh đề và phân tích liên kết nên chưa thể áp dụng phương pháp này.

Lý thuyết cấu trúc diễn ngôn (Rhetorical Structure Theory) do Mann và Thompson [8] đưa ra, cho phép biểu diễn mối liên hệ giữa các thành phần trong một văn bản dưới dạng cây với lá là các mệnh đề cơ bản. Trọng tâm của lý thuyết cấu trúc diễn ngôn là khái niệm quan hệ diễn ngôn. Đó là những quan hệ giữa hai thành phần không giao nhau của văn bản được gọi là: Hạt nhân (Nucleus–N) và Vệ tinh (Satellite–S). Các mối quan hệ diễn ngôn có thể sắp đặt thành cây cấu trúc diễn ngôn (Rhetorical Structure Tree) dựa vào việc sơ đồ hóa các dạng cơ bản nhất của các quan hệ diễn ngôn. Để xây dựng cây cấu trúc diễn ngôn, trước hết cần phát hiện mối quan hệ giữa các đoạn của văn bản. Đoạn văn bản nhỏ nhất mà giữa chúng còn tồn tại các quan hệ được gọi là đơn vị diễn ngôn nguyên tố (Elementary Discourse Units - EDU). EDU có thể là một mệnh đề hoặc tựa mệnh đề. Việc phân tích văn bản để cho ra các EDU gọi là phân đoạn diễn ngôn. Quá trình từ EDU tìm ra các quan hệ diễn ngôn của các EDU và xây dựng cây diễn ngôn được gọi là phân tích diễn ngôn.

Trừ những câu có một số thành phần ẩn và câu phức với các nòng cốt bao nhau, giới hạn giữa các mệnh đề tiếng Việt thể hiện qua các dấu hiệu diễn ngôn là khá rõ ràng. Giải thuật phân đoạn diễn ngôn sẽ tìm ra giới hạn mệnh đề dựa theo các dấu hiệu diễn ngôn. Kết quả của quá trình phân đoạn diễn ngôn là câu đưa vào được điền thêm các cặp [] để chỉ giới hạn của mệnh đề. Trong [9] xử lý này dựa trên 10 hành động: NORMAL, NORMAL_THEN_COMMA, END,... Mỗi hành động cho phép điền một giới hạn văn bản có dạng][vào vị trí thích hợp. Chẳng hạn,

• Hành động NORMAL ra lệnh cho bộ phân tích thêm một giới hạn văn bản ngay trước xuất hiện của dấu hiệu diễn ngôn.

• Hành động COMMA ra lệnh cho bộ phân tích thêm một giới hạn văn bản ngay sau xuất hiện của dấu phẩy đầu tiên của xâu vào. Nếu dấu phẩy đầu tiên có "và" (hay "hoặc", sau đây gọi tắt là "và") đi ngay sau, biên của văn bản được đặt sau xuất hiện của dấu phẩy tiếp sau. Nếu không tìm thấy dấu phẩy nào trước khi kết thúc câu, một giới hạn văn bản được thiết lập tại điểm cuối của câu.

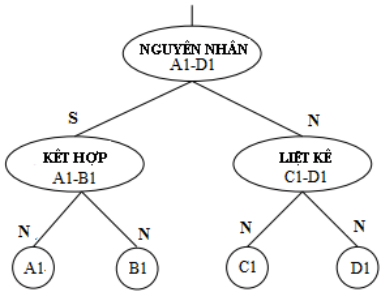
• Hành động NORMAL_THEN_COMMA ra lệnh cho bộ phân tích thêm một giới hạn văn bản ngay trước xuất hiện của dấu hiệu "và" và một giới hạn văn bản khác ngay sau xuất hiện của dấu phẩy đầu tiên sau từ "và" đó. Nếu dấu phẩy đầu tiên được nối tiếp bởi "và", việc xử lý cũng như trong hành động COMMA.

Cây phân tích diễn ngôn được xây dựng từ dưới lên theo phương pháp proof-theoretic mô tả trong [9]. Hình 4 mô tả cây phân tích diễn ngôn của câu ví dụ đã nói. Bộ phân tích diễn ngôn chúng tôi đã xây dựng cho tiếng Việt [5] cho kết quả phân tích diễn ngôn ở mức câu với tỷ lệ đúng đến 78%.

Trong bài báo này đã sử dụng 18 mối quan hệ diễn ngôn giữa các mệnh đề được [3],[1] nêu ra kết hợp với một số quan hệ được nói đến trong [9] làm tên kết nối. Các kết nối này mang

tính chất kết nối lớn vì chúng liên kết các cụm từ với nhau. Các kết nối lớn được xây dựng giữa các cặp mệnh đề dựa theo cây diễn ngôn của câu.

Ví dụ, câu được mô tả bởi cây diễn ngôn ở hình 4. Sau khi phân đoạn diễn ngôn có 4 mệnh đề ký hiệu A1, B1, C1, D1.



Hình 4. Cây phân tích diễn ngôn của câu [trời mưa rất to và $_{A1}$] [gió rất mạnh nên$_{B1}$] [tôi phải nghỉ học,$_{C1}$] [mẹ tôi phải nghỉ làm.$_{D1}$]

Khi chuyển từ quan hệ diễn ngôn sang liên kết, cần chú ý rằng liên kết giữa các mệnh đề phải thỏa mãn các yêu cầu: mỗi liên kết phải nối hai từ và các liên kết phải thỏa mãn các tính chất của văn phạm liên kết: tính phẳng, tính liên thông, tinh thứ tự, tính thỏa mãn, tính loại trừ.

Để đảm bảo tính phẳng, nghĩa là các liên kết không được giao nhau khi vẽ bên trên các từ, cần chọn ra trong mỗi mệnh đề một từ đại diện để liên kết. Mỗi từ trong mệnh đề sẽ được gắn với một trọng số. Từ có trọng số nhỏ nhất sẽ được chọn đại diện cho mệnh đề.

Như vậy, quá trình phân tích cú pháp cho câu ghép cần qua những bước sau:

- Phân đoạn diễn ngôn.
- Phân tích cú pháp cho từng mệnh đề, thêm các liên kết nhận được vào liên kết tổng thể.
- Xây dựng cây phân tích diễn ngôn cho câu.
- Duyệt cây phân tích diễn ngôn theo thứ tự sau, thêm các kết nối ứng với từng quan hệ diễn ngôn.

3.1. Giải thuật phân tích cú pháp cho câu ghép

Dưới đây là giải thuật tổng thể để phân tích một câu thể hiện dưới dạng một dãy các từ.

Vào: Câu tiếng Việt s đã tách từ

Ra: Kết quả phân tích câu bao gồm danh sách các kết nối

Phương pháp:

$U[N]$:= DISCOURSE_SEGMENT (s);// U chứa các đơn vị diễn ngôn của câu s
    root := RS_PARSE(); // Cây phân tích diễn ngôn của s có gốc là root
    for $i := 1$ to N
    if IS_UNIT ($U[i]$)
{ PARSE($U[i], LinkTemp$);
    Lnk.Add LinkTemp;
}
    INSERT_LINK_FROM_RST_TREE (root);
     AFTER_INSERT();

Trong giải thuật này, biến *Lnk* chứa toàn bộ liên kết cho cả câu ghép. Biến *LinkTemp* chứa các liên kết cho từng mệnh đề. Hàm DISCOURSE_SEGMENT thực hiện phân đoạn

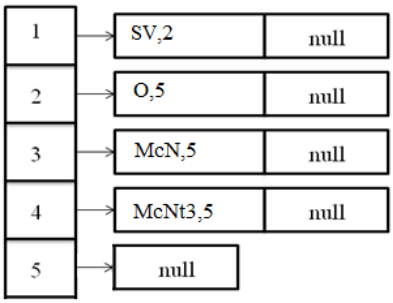
diễn ngôn cho câu s. Hàm RS_PARSE cho phép dựng cây phân tích diễn ngôn của câu. Hàm IS_UNIT trả về giá trị true nếu đơn vị diễn ngôn được xem xét chứa từ hai từ trở lên. Hàm PARSE là hàm phân tích cú pháp, trả 1 nếu câu đúng cú pháp, 0 nếu ngược lại. Kết quả được lưu trong *lnk.lnk* có cấu trúc như hình 5. Mỗi danh sách ứng với mỗi từ là một *linklist*. Hàm AFTER_INSERT cho phép xử lý và tạo liên kết với các mệnh đề phụ trạng ngữ: "hôm qua,", "trong khi đó"... Hàm INSERT_LINK_FROM_RST_TREE thực hiện việc duyệt cây diễn ngôn của câu, thêm các liên kết ứng với từng quan hệ diễn ngôn.

3.2. Tìm từ để kết nối mệnh đề

Nếu trong mô hình văn phạm phụ thuộc, từ đại diện cho mệnh đề chính là từ trung tâm của mệnh đề thì trong mô hình văn phạm liên kết, cần phải chọn từ đại diện cho mệnh đề.

Bậc của kết nối

Việc chọn từ đại diện cho mệnh đề phải đảm bảo cầu về tính phẳng của liên kết. Sau khi phân tích cú pháp cho các mệnh đề, các kết nối được lưu trữ lại dưới dạng danh sách liên kết. Hình 5 dưới đây thể hiện việc lưu trữ phân tích liên kết của câu "Tôi mua một bông hoa". 1, 2. .5 là số thứ tự của từ. Mỗi từ có một danh sách liên kết các kết nối với các từ nằm bên phải nó. Thông tin về mỗi kết nối bao gồm (kiểu, đích, bậc). Ví dụ $(SV, 2)$ chỉ liên kết của từ đầu tiên (tôi) và từ thứ hai (mua).



Hình 5. Minh họa cách lưu trữ phân tích liên kết của câu "Tôi mua một bông hoa"

Bậc của liên kết được tính như sau:

Theo giải thuật phân tích cú pháp, liên kết được vẽ đầu tiên sẽ có bậc 0. Đó là liên kết SV và O. Sau đó, trong quá trình thực hiện giải thuật phân tích trong [12] một cách đệ quy với các từ bên trái và bên phải từ được xét, bậc của McN sẽ là 1, của McNt3 là 2. Nếu câu này đóng vai trò mệnh đề trong liên kết với mệnh đề khác, thì kết nối được chọn để liên kết sẽ là kết nối trên cùng, tức là kết nối bậc 0 (trong ví dụ này là SV hoặc O).

Chọn từ để liên kết

Sau khi tìm được kết nối thích hợp, vấn đề đặt ra chọn từ bên trái hay bên phải. Tiêu chí đưa ra là chọn từ quan trọng hơn. Thông tin về từ bên trái hay bên phải được chọn sẽ được lưu trữ trong danh mục các kết nối. Ví dụ, trong phân tích cụm từ: "một bông hoa", liên kết được chọn là McN giữa từ "một" và từ "hoa" . Với liên kết này, từ được chọn là từ bên phải, đó là từ "hoa".

4. GIẢI QUYẾT VẤN ĐỀ NHẬP NHẰNG VỚI TỪ "VÀ", "HOẶC" VÀ DẤU PHẨY

Như đã nêu trong [6], các từ "và", "hoặc" và dấu phẩy đóng một vai trò đặc biệt khi phân tích một câu trên mô hình văn phạm liên kết vì nó có thể chứa kết nối lớn. Theo lý thuyết cấu trúc diễn ngôn, bản thân từ "và" cũng là một dấu hiệu diễn ngôn. Do vậy cần phân biệt trường hợp từ "và" là dấu hiệu diễn ngôn và từ "và" chỉ nối hai từ hoặc hai cụm từ có cùng vai trò cú pháp. Vấn đề này cũng đã được Lê Thanh Hương đề cập trong [4].

Bộ phân tích diễn ngôn trong [5,9] suy ra quan hệ và dựng cây diễn ngôn chỉ bằng phân tích "nông" tức là hoàn toàn không phân tích cú pháp các câu trong văn bản. Từ "và" chỉ được coi là dấu hiệu diễn ngôn khi đứng ngay trước một số dấu hiệu diễn ngôn, chẳng hạn trong câu "[Mặc dù trời mưa lớn][và mặc dù mọi người đều ngăn cản,] [nó cứ đi]".

Theo [1], các loại từ có thể đóng vai trò vị tố (thành phần chính của vị ngữ) là động từ (đặc biệt, động từ quan hệ "là"), tính từ, danh từ, giới từ hay số từ. Vị tố loại danh từ, số từ, giới từ chỉ xuất hiện trong những câu hội thoại, rất hiếm khi xuất hiện trong các mệnh đề của câu ghép với liên từ "và" nên ta chỉ xét ba loại vị tố chính: động từ "là", động từ khác, tính từ. Một cụm từ là mệnh đề trong câu ghép song song nếu trong phân tích liên kết của nó tồn tại ít nhất một trong các kết nối sau: SV (liên kết giữa chủ ngữ và động từ), SA (liên kết chủ ngữ với tính từ), DT_LA (liên kết của chủ ngữ là danh từ hay đại từ xưng hô với động từ quan hệ "là").

Ta thay đổi giải thuật phân đoạn diễn ngôn của Marcu [9] như sau: thêm phần phân tích cú pháp vào hành động ứng với từ "và" – trong hệ thống của chúng tôi là hành động NORMAL_THEN_COMMA. Trạng thái của hành động NORMAL_THEN_COMMA cũng không còn là COMMA như trong [9] mà là NORMAL_THEN_COMMA. Hành động PH cũng được thêm vào để xử lý dấu phẩy. Hành động này hoàn toàn khác với hành động COMMA được gắn với các từ nối như "mặc dù", "với"... ở xử lý: Khi gặp dấu phẩy, dù cụm từ đang xét là mệnh đề đúng cú pháp chưa chắc giới hạn văn bản đã được thêm ngay sau dấu phẩy. Cần xem xét xem dấu hiệu ở sau cụm đó có là dấu phẩy không. Nếu là dấu phẩy thì giới hạn văn bản sẽ được điền sau dấu hiệu đầu tiên khác dấu phẩy và từ "và" bởi hành động của dấu hiệu diễn ngôn đó. Ví dụ trong câu"tôi mua nhiêu đồ chơi, bánh, kẹo để con tôi tặng các bạn", giới hạn văn bản phải được thêm vào sau từ "kẹo" thay vì thêm sau từ "đồ chơi", dù cụm từ "tôi mua nhiều đồ chơi" đã là một mệnh đề hoàn chỉnh.

Do khuôn khổ của bài báo, ta chỉ trình bày các hành động đã thay đổi trong giải thuật phân đoạn diễn ngôn của [9]. Các hành động khác được xử lý theo [9].

Vào: Câu S; Mảng của n dấu hiệu diễn ngôn tiềm tàng có thể xuất hiện trong S : $marker[n]$
Ra: Các đơn vị diễn ngôn của S
Phương pháp:

{status := nil; clauses := nil; parentheticals := nil;
currClauseStart := 1; currParentStart := 1;
for i from 1 to n
// Xử lý trường hợp có lưu lại status
//Xử lý khi status chứa các giá trị khác (COMMA, SET_AND,MATCH_DASH...)
if NORMAL_THEN_COMMA ∈ status
if not markerTextEqual($i, 1, 1$)

{clauses := clauses ∪
textFromTo(currClauseStart, offset(i), parentheticals);
status:=status{NORMAL_THEN_COMMA}
parentheticals := nil; currParentStart := -1;}
if PH ∈ status ∧ not markerTextEqual(i, ȷ, ȷ)
{if not markerTextEqual(i,"và")
if (isClause(textFromTo(offset(i), offset($i+1$))
{clauses := clauses ∪
textFromTo(currClauseStart, offset(i), parentheticals);
currClauseStart:=$i+1$;}
else
{clauses:=clauses ∪ textFromTo(currClauseStart, offset(i),parentheticals);
status:=status\ {PH};}
swithch getActionType(i))
{...//Xử lý khi getActionType chứa các giá trị khác (COMMA, NOTHING, DUAL...)
case NORMAL_THEN_COMMA
if
isClause(textFromTo(currClauseStart,offset(i))∧ isClause(textFromTo(offset(i),
offset($i+1$))
clauses := clauses ∪ textFromTo(currClauseStart,offset(i),parentheticals);
status:=status ∪ {getActionType(i);
currClauseStart := offset(i); parentheticals := nil; setDiscourse(i,yes);
case PH:
if isClause(textFromTo(currClauseStart, offset(i))
status:=status ∪ getActionType(i);
}
End For

Trong giải thuật này, biến *status* ghi lại những dấu hiệu đã được xử lý trước mà có thể còn ảnh hưởng đến việc xác định ranh giới các mệnh đề và những đơn vị trong dấu ngoặc đơn. Ban đầu, giá trị của nó đặt bằng NIL.

$textFromTo(i, j)$ là hàm trả ra xâu con chứa từ thứ i đến từ thứ j của câu đưa vào.

$offset(i)$ là thứ tự của dấu hiệu diễn ngôn thứ I trong câu.

$getActionType(i)$ cho hành động xử lý ứng với dấu hiệu diễn ngôn thứ i.

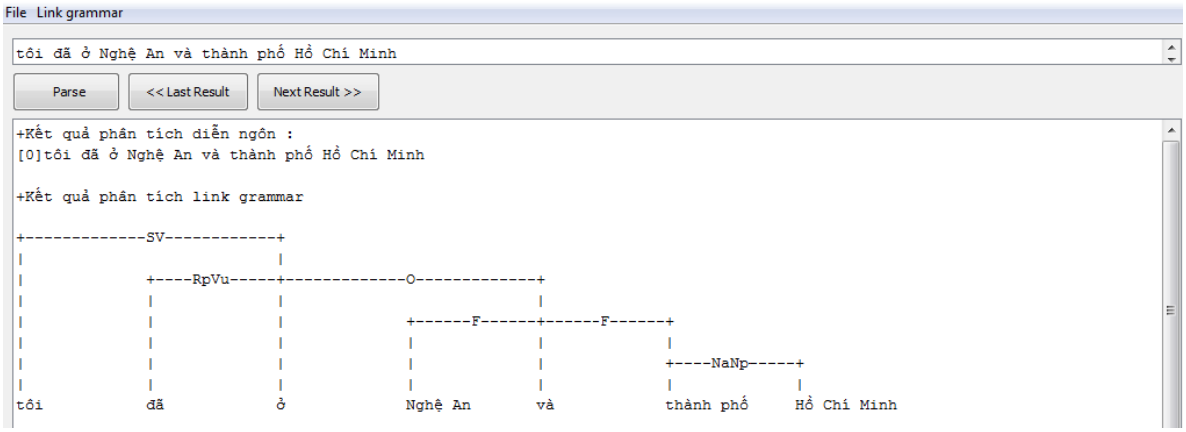
parentheticals chứa tập hợp các mệnh đề được viết trong ngoặc

$isClause(s)$ là hàm kiểm tra xem đoạn văn bản s có là một mệnh đề không. Chi tiết như hàm *isClause* như sau:

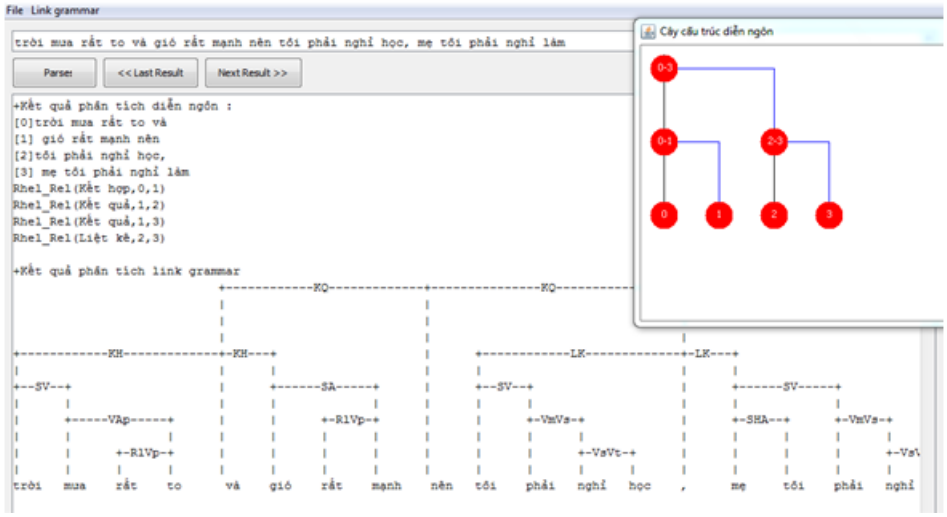
boolean isClause (s)
{linkage lnk;int n;connection c;
n=NumberOfWord(s)
if (PARSE(s,lnk)!=0) //s đúng cú pháp
{for($i = 1; i <= n; i++$)
for each c in lnk.linklist(i)
{ if(c.type="SV" or c.type="DT_ LA" or c.type="SA")//s chứa nòng cốt
{return true;
break;}

```
}
            return false;
        }
return false;//s sai cú pháp
}
```

Hình 6 dưới đây minh họa kết quả phân tích câu "Tôi đã ở Nghệ An và thành phố Hồ Chí Minh" , cụm từ "tôi đã ở Nghệ An" là một mệnh đề, tuy nhiên cụm từ "thành phố Hồ Chí Minh" không phải là mệnh đề nên từ "và" không là dấu hiệu diễn ngôn.



Hình 6. Kết quả phân tích câu "tôi đã ở Nghệ An và thành phố Hồ Chí Minh"



Hình 7. Kết quả phân tích câu
"Trời mưa rất to và gió rất mạnh nên tôi phải nghỉ học, mẹ tôi phải nghỉ làm"

Hình 7 là kết quả thực hiện của bộ phân tích câu ghép với câu "trời mưa rất to và gió rất mạnh nên tôi phải nghỉ học, mẹ tôi phải nghỉ làm". Trong đó, từ "và" là dấu hiệu diễn ngôn (kết nối KH), dấu phảy cũng là dấu hiệu diễn ngôn nên không được liên kết bằng kết nối F thông thường.

5. ĐỘ PHỨC TẠP TÍNH TOÁN

Theo [12], chi phí thời gian của giải thuật phân tích liên kết với một văn phạm xác định là $O(n^3)$ với n là độ dài câu (số từ trong câu). Chi phí này cũng tương đương với chi phí thời gian của các giải thuật phân tích sử dụng văn phạm phi ngữ cảnh.

Nếu quá trình phân đoạn diễn ngôn chia câu thành k mệnh đề, độ dài trung bình mỗi mệnh đề còn n/k, chi phí thời gian trung bình còn $O(n^3/k^2)$.

Thực nghiệm với tập câu mẫu thứ nhất ở bảng 3, cho thấy thời gian để phân tích tập mẫu theo kiểu liên kết từ nối là 296.153 mili giây, trong khi thời gian phân tích câu đó bằng cách phân tích riêng từng mệnh đề là 217.324 mili giây, giảm đáng kể so với phân tích kiểu liên kết từ nối.

6. KẾT QUẢ THỰC NGHIỆM

Để kiểm chứng cho giải thuật phân tích câu ghép, ta tạo bộ mẫu gồm 100 câu đã phân tích và chú giải, chi tiết như trong bảng 2. Nguồn dữ liệu được chọn từ các bài báo trên mạng:

http://www.mediafire.com/?6ajt9btbrtxidr9;

http://www.vietnamtourism.com/v_ pages/tourist/destination.asp?mt=8420& uid=533;

http://dantri.com.vn/c26/s26-484690/barcelona-mu-giac-mo-noi-thien-duong.htm;

Bảng 2. Các tập mẫu của bộ phân tích cú pháp

STT	Tập mẫu	Số lượng câu	Số từ trung bình trong câu
1	Ngữ liệu tiếng Việt phổ quát	50	12.8
2	Thể thao	25	24
3	Du lịch	25	26.5

Các bộ câu mẫu này chủ yếu chứa câu ghép, một số câu đơn có thể gây nhập nhằng với từ "và" , hay chứa các mệnh đề phụ trạng ngữ cũng được thử nghiệm. So sánh kết quả thu được của bộ phân tích cú pháp mở rộng cho câu ghép với kết quả của bộ phân tích cú pháp [6] chỉ đơn thuần dùng liên kết cho thấy kết quả đạt được cao hơn hẳn so với bộ phân tích cũ [6] (xem bảng 3 dưới đây).

Bảng 3. Kết quả phân tích trên tập mẫu

Tập mẫu	Độ chính xác (bộ PT cũ)	Độ phủ (bộ PT cũ)	Độ chính xác (bộ PT mới)	Độ phủ (bộ PT mới)
1	42.5%	35.7%	75.1%	65.7%
2	9.5%	6.1%	33.5%	21.6%
3	28.3%	20.5%	47.4%	58.5%

Trong số các bộ ngữ liệu, bộ ngữ liệu tiếng Việt phổ quát chủ yếu chứa các câu ghép hai mệnh đề và cấu trúc từng mệnh đề cũng khá đơn giản có tỷ lệ câu phân tích diễn ngôn đúng là 100%, và tỷ lệ câu phân tích đúng là cao nhất. Ngữ liệu về du lịch cũng gồm những câu giới thiệu quảng bá du lịch, nhiều câu có trên 3 mệnh đề nhưng cấu trúc vẫn theo đúng luật cú pháp thông thường, không chứa các yếu tố ẩn. Bộ ngữ liệu về thể thao với nhiều dạng thức đặc biệt của câu ghép đạt tỷ lệ thấp nhất.

7. KẾT LUẬN

Bộ phân tích cú pháp liên kết được đề xuất đã đạt được kết quả khá tốt trên những câu ghép gồm nhiều mệnh đề, không bao nhau, có thể xuất hiện những đoạn giải thích với cặp ngoặc hoặc dấu gạch ngang (-).

Vẫn còn một số dạng câu ghép mà bộ phân tích của chúng tôi chưa xử lý được, chẳng hạn câu ghép bị ẩn một số thành phần, câu ghép không có từ nối, câu ghép có chứa những phụ thuộc không lân cận. Những loại câu này đòi hỏi phải nghiên cứu sâu hơn về các thành phần ẩn, các thành phần có vị trí tự do trong câu tiếng Việt và thêm những mối liên kết đặc biệt cho văn phạm liên kết tiếng Việt.

Khi đã xây dựng được bộ ngữ liệu mẫu đủ lớn, sẽ xét trường hợp câu phức với các nòng cốt bao nhau. Thật ra một số trường hợp đã có thể phân tích với bộ phân tích câu ghép của chúng tôi, như câu "Nó bảo rằng nó không đi nữa.". Tuy nhiên một số trường hợp cần dùng phương pháp học máy để nhận ra giới hạn mệnh đề.

Một hướng phát triển khác cũng được quan tâm là tích hợp những mối liên kết về ngữ nghĩa trong văn phạm liên kết tiếng Việt. Điều này là khả thi với mô hình văn phạm liên kết cho phép biểu diễn phân tích câu bằng đồ thị liên kết có chu trình.

TÀI LIỆU THAM KHẢO

[1] Diệp Quang Ban, *Ngữ pháp tiếng Việt* (hai tập), NXB Giáo dục, 2001.

[2] D. Béchet, k-valued link grammars are learnable from strings, *Proceedings of the 8th Conference on Formal Grammars*, Vienna, Austria,2007 (9–18).

[3] Nguyễn Chí Hòa, *Các phương tiện liên kết và tổ chức văn bản*, NXB Đại học Quốc gia Hà Nội, 2005.

[4] Le Thanh Huong, "Automatic Discourse Structure Generation Using Rhetorical Structure Theory", Ph.D. dissertation, Middlesex University. U.K, 2004.

[5] Nguyễn Thị Thu Hương, Lê Văn Chương, Phân tích diễn ngôn cho văn bản tiếng Việt, *Kỷ yếu hội thảo ICT-RDA*, Hà Nội, 8/2008 (227–245).

[6] Nguyen Thi Thu Huong, Pham Nguyen Quang Anh, A link grammar for Vietnamese, *Journal on Information and Communicationn Technology*, (8/2011) 18–27.

[7] Sang-Soo Kim et al., Resolving dependency ambiguity of subordinate clauses using support vector machines world academy of science, *Engineering and Technology* (25/2007).

[8] W. Mann, S. Thompson, Rhetorical structure theory. A Framework for the Analysis of Texts. Information Sciences Institute Research Report (1988).

[9] D. Marcu, "The Rhetorical Parsing. Summarization and Generation of Natural Language Texts", Ph.D. dissertation. University of Toronto, 1997.

[10] I. Mel'čuk, *Dependency Syntax: Theory and Practice*, State University of New York Press, 1988.

[11] Tomohiro Ohno et al., Dependency Parsing of Japanese Monologue Using Clause Boundaries. *Languages Resources and Evaluation*, Springer, 2006.

[12] D. Sleator, D. Temperley, Parsing English with Link Grammar,CMU-CS-91-96. 10/1991.

[13] Trần Ngọc Thêm, *Hệ thống liên kết văn bản tiếng Việt*, NXB Khoa học Xã hội, 2008.

[14] Takehito Utsuro et al., Analyzing Dependencies of Japanese Surbodinate Clauses based on Statistics of Scope Embedding Preference, *Proc. 1st NAACL*, 2000 (169–176).

[15] http://www.abisource.com/projects/link-grammar/.

Ngày nhận bài 04 - 9 - 2012
Nhận lại sau sửa ngày 10 - 2 - 2013