

Tóm tắt văn bản tiếng Việt sử dụng cấu trúc diễn ngôn

Vietnamese text summarisation using discourse structures

Nguyễn Trọng Phúc, Lê Thanh Hương

Đại học Bách khoa Hà Nội

Abstract

Text summarisation is one of typical problems in Text mining. To solve this problem, most of previous approaches are based on statistical methods such as TFXIPF, Title, Position,... Such approaches often derive output texts with low precisions. In this paper, we present a method to summarise Vietnamese texts using discourse structure, which is a method to represent semantic relations among textual units (e.g., cause-effect relation). The importance of textual units is computed based on the discourse tree of text. The input text is then extracted to generate its summary.

Keywords: text summarisation, discourse structure, rhetorical relation, Vietnamese.

Tóm tắt nội dung

Tóm tắt văn bản là một trong những bài toán điển hình của lĩnh vực Khai phá văn bản. Để giải quyết bài toán này, phần lớn các hướng tiếp cận trước đây đều dựa trên kết quả thống kê như TFXIPF, Title, Position,... Các phương pháp này thường đưa ra các kết quả tóm tắt có độ chính xác thấp. Trong báo cáo này, chúng tôi sẽ trình bày phương pháp tóm tắt văn bản tiếng Việt dựa trên cấu trúc diễn ngôn (CTDN). CTDN là một phương tiện cho phép biểu diễn mối quan hệ diễn ngôn giữa các đoạn văn bản (như quan hệ nhân-quả). Từ cây CTDN, ta có thể đánh giá được tầm quan trọng của các đoạn văn bản và tiến hành trích rút hay tóm tắt nội dung văn bản.

Từ khóa: tóm tắt văn bản, cấu trúc diễn ngôn, quan hệ diễn ngôn, tiếng Việt.

1. Giới thiệu

Tóm tắt văn bản là quá trình rút ra những thông tin quan trọng nhất từ một hay nhiều nguồn văn bản để tạo ra một văn bản gọn hơn phục vụ cho một số nhiệm vụ hay người dùng cụ thể. Bài toán tóm tắt văn bản là một trong những bài toán hữu ích nhất trong lĩnh vực Khai phá văn bản. Một số biến thể của bài toán này như sinh phụ đề tự động (Subtitling), sinh ý chính tài liệu (Document Gisting), sinh tiêu đề văn bản (Header Generating)... Những thuật toán dùng trong tóm tắt văn bản cũng được áp dụng trong các bài toán tương tự như tóm tắt hình ảnh, âm thanh hay nói chung là những dữ liệu đa phương tiện.

Điểm cốt lõi của một hệ thống Tóm tắt văn bản xét theo lĩnh vực Khai phá văn bản là việc tìm ra những thành phần quan trọng trong văn bản cần tóm tắt. Các thành phần này được gọi là các đơn vị văn bản (ĐVVB). Đơn vị văn bản ở đây có thể hiểu là đơn vị nhỏ nhất có nghĩa mà ta chọn để trích rút như mệnh đề hoặc câu. Các ĐVVB quan trọng sẽ có xác suất lớn để chứa ý chính hay nội dung quan trọng của cả đoạn văn hay văn bản. Sau khi chọn được các ĐVVB quan trọng, hệ thống Tóm tắt có thể tóm lược chúng, biến đổi chúng và sau cùng hiển thị ra màn hình, thống kê kết quả hay sinh ra các dữ liệu mới cho các hệ thống khác.

Các phương pháp tóm tắt văn bản khác nhau chủ yếu ở cách đánh giá và xác định các ĐVVB quan trọng. Phần lớn các phương pháp tóm tắt văn bản cổ điển đều dựa trên kết quả thống kê như phương pháp dựa trên tần suất xuất hiện từ (TFxIPF), phương pháp sử dụng tiêu đề (title-based), phương pháp dựa trên vị trí (position) của câu trong đoạn, của đoạn trong văn bản,... Các phương pháp như vậy đều được đưa ra dựa trên ý tưởng: các ĐVVB quan trọng nói nhiều tới nội dung chính của văn bản, do đó trong các ĐVVB này phải xuất hiện nhiều từ ngữ liên quan tới nội dung của văn bản. Tùy theo từng phương pháp, các từ ngữ quan trọng là các từ ngữ xuất hiện nhiều lần trong văn bản, các từ ngữ xuất hiện trong tiêu đề của văn bản, hoặc các từ ngữ xuất hiện ở câu đầu, câu cuối của văn bản,... Tuy nhiên, kết quả tóm tắt của các phương pháp dựa trên thống kê thường có độ chính xác thấp bởi việc chứa nhiều từ quan trọng chưa đủ để nói lên mức độ quan trọng của ĐVVB. Bởi vậy, trong các nghiên cứu gần đây về Tóm tắt văn bản, người ta thường sử dụng các phương pháp dựa trên ngữ nghĩa để nâng cao độ chính xác. Song, trong phạm vi tóm tắt văn bản tiếng Việt, đến nay đa phần các nghiên cứu đều sử dụng hoặc cải tiến các phương pháp dựa trên thống kê.

Trong bài này, chúng tôi sẽ trình bày phương pháp tóm tắt văn bản tiếng Việt dựa trên cấu trúc diễn ngôn. Đây là phương pháp cho phép biểu diễn mối quan hệ diễn ngôn giữa các đoạn văn bản (như các quan hệ *nhân-quả*, *liệt kê*, *diễn giải*,...). Từ cây CTDN, ta có thể đánh giá được độ quan trọng của các đoạn văn bản và tiến hành trích rút, tạo ra tóm tắt nội dung cho văn bản.

2. Lý thuyết cấu trúc diễn ngôn

Lý thuyết cấu trúc diễn ngôn (Rhetorical Structure Theory) [9] là một phương pháp dùng để biểu diễn sự mạch lạc, chặt chẽ của văn bản. Trung tâm của lý thuyết CTDN là các quan hệ diễn ngôn (QHDN) giữa các

đoạn văn bản không gộp lên nhau. Đoạn văn bản quan trọng hơn trong một QHDN được gọi là *nhân* (nuclei - N), còn thành phần ít quan trọng hơn được gọi là *vệ tinh* (satellite - S). Nếu các đoạn văn bản đó có tầm quan trọng ngang nhau (như quan hệ ***Tương phản*** hay quan hệ ***Liệt kê***) thì các đoạn văn bản đó đều đóng vai trò *nhân* trong mỗi QHDN giữa các thành phần đó.

Điểm khác biệt cơ bản giữa các N và S là: N chứa nhiều thông tin quan trọng hơn là S; và trong một quan hệ diễn ngôn, N có tính độc lập cao hơn S.

Ví dụ 1:

Dưới đây là định nghĩa quan hệ ***chứng minh***:

Ràng buộc đối với N : người đọc (R) có thể không thực sự tin thông tin trong N vì người đọc không thực sự tin người viết (W).

Ràng buộc đối với S : người đọc tin các thông tin trong S hoặc cảm thấy thông tin đó đáng tin cậy.

Ràng buộc đối với sự kết hợp N+S : sự hiểu S của R làm tăng sự tin cậy của R đối với N.

Hiệu ứng : sự tin tưởng của R đối với N được tăng lên.

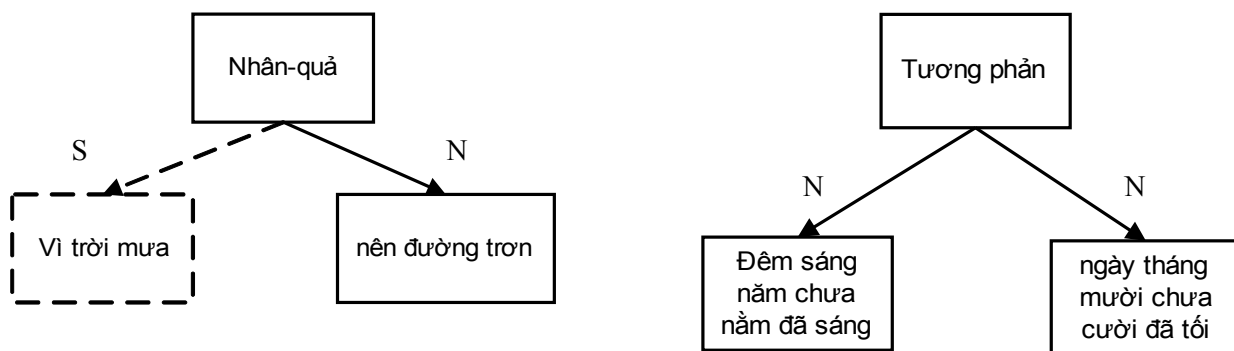
Nơi nhận hiệu ứng : N

Để minh họa cho quan hệ này ta xét 2 câu:

[Sự thật là áp lực dẫn đến việc hút thuốc trong trường PTCS cao hơn trong các giai đoạn khác của cuộc đời.¹][Chúng ta thấy rằng mỗi ngày có thêm 3000 trẻ vị thành niên bắt đầu hút thuốc²]

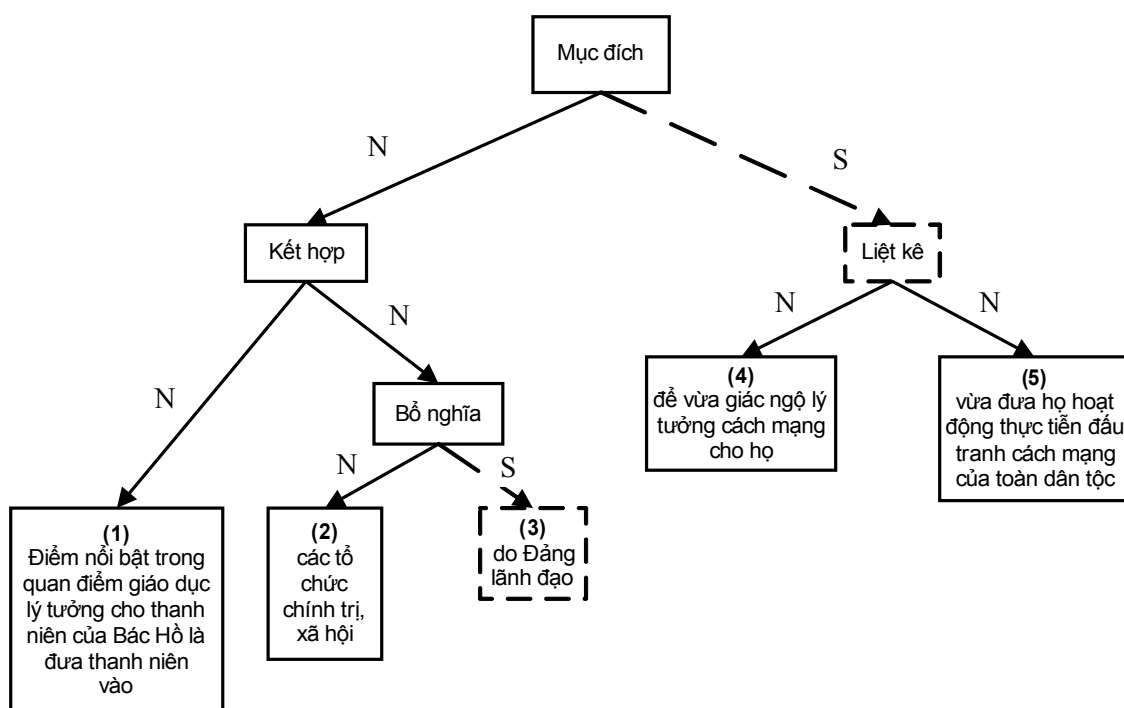
Trong ví dụ trên, câu 2 có tác dụng chứng minh cho phát biểu ở câu 1. Vì vậy hai câu này có quan hệ ***Chứng minh***.

Các quan hệ cấu trúc diễn ngôn có thể được mô tả dưới dạng các cây cấu trúc diễn ngôn như biểu diễn ở hình 1.



Hình 1: Một số quan hệ diễn ngôn được biểu diễn dưới dạng cây

Khi phân tích một văn bản, một đoạn văn hay thậm chí là một câu dài, ta sẽ thu được một cây CTDN gồm nhiều QHDN (hình 2):



Hình 2: Cây cấu trúc diễn ngôn.

Trong hình 2, các mũi tên liền nét trở tới các ĐVVB có vai trò *nhân*, còn các mũi tên đứt trở tới các ĐVVB có vai trò *vệ tinh*.

Dựa vào cây CTDN của văn bản, ta sẽ tiến hành đánh giá độ quan trọng của các ĐVVB. Do đó có thể nói vấn đề lớn nhất của hướng tiếp cận này là xây dựng hệ thống phân tích CTDN tự động cho văn bản tiếng Việt. Chúng tôi sẽ đi sâu vào vấn đề này trong phần 3.

3. Phân tích cấu trúc diễn ngôn của văn bản

Bài toán phân tích cấu trúc diễn ngôn:

Đầu vào : văn bản T.

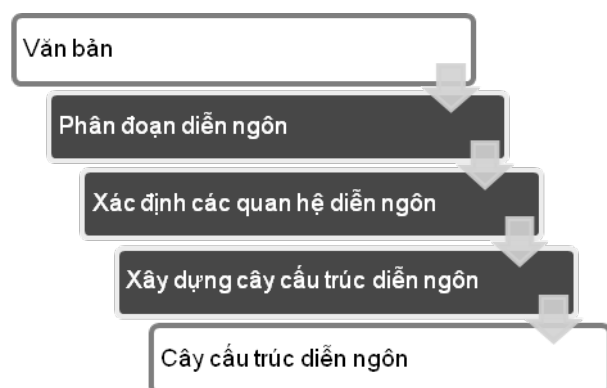
Đầu ra : cây cấu trúc diễn ngôn của T.

Để giải quyết bài toán này, ta sẽ thực hiện quá trình gồm 3 bước :

1. *Phân đoạn diễn ngôn* : tách văn bản thành tập các đơn vị diễn ngôn.

2. *Xác định quan hệ diễn ngôn* : xác định mọi quan hệ có thể có giữa các đơn vị diễn ngôn cơ bản và giữa các đơn vị văn bản cấp cao hơn.

3. *Sinh cấu trúc diễn ngôn phù hợp nhất cho văn bản*: sử dụng các quan hệ diễn ngôn giả định thu được ở bước 2 để xây dựng các cây cấu trúc diễn ngôn của văn bản. Chọn một cây “tốt nhất” theo tiêu chí nào đó.



Hình 3: *Quá trình phân tích diễn ngôn*

3.1. Phân đoạn diễn ngôn

Theo lý thuyết của Mann và Thompson [9], cấu trúc diễn ngôn được tạo nên từ các phân đoạn diễn ngôn nhỏ hơn. Mọi phân đoạn diễn ngôn đều phải mang một ý nghĩa trọn vẹn nào đó, ví dụ như câu đơn hoặc mệnh đề trong câu ghép. Đơn vị diễn ngôn nhỏ nhất được gọi là *đơn vị diễn ngôn cơ bản* (ĐVDNCB).

Quá trình phân tách văn bản thành các ĐVDNCB gọi là quá trình phân đoạn diễn ngôn. Tính chính xác ở bước này có ảnh hưởng đến chất lượng của bước xác định quan hệ diễn ngôn sau này. Ngoài ra, mức độ gọn nhẹ của bản tóm tắt sau này cũng phụ thuộc vào độ chi tiết (kích thước) của các ĐVDNCB.

Quá trình phân đoạn diễn ngôn gồm hai bước:

1. Tách các đoạn văn, câu. Bước này được thực hiện khá đơn giản dựa trên các dấu xuống dòng và dấu chấm câu.
2. Tách các câu thành các ĐVDNCB.

Phần lớn các nghiên cứu trên thế giới về cấu trúc diễn ngôn cho văn bản tiếng Anh đều dựa trên từ hiệu (cue phrase) như *because, but, although...* để phân đoạn diễn ngôn [2,3,9,10]. Khi xuất hiện các từ này, ta có thể xác định được ngay các thành phần diễn ngôn trong câu. Tuy nhiên, quá trình phân đoạn diễn ngôn cho văn bản tiếng Việt không đơn giản như vậy. So với tiếng Anh, quá trình phân đoạn diễn ngôn tiếng Việt có một số đặc điểm riêng, đòi hỏi các thao tác xử lý phức tạp hơn. Các điểm khác biệt đó được nói tới ở phần sau.

a. Điểm khác biệt giữa phân đoạn diễn ngôn trong tiếng Việt và trong tiếng Anh

Điểm khác biệt đầu tiên giữa phân đoạn diễn ngôn tiếng Anh và tiếng Việt là việc xác định các từ hiệu (cue phrase). Do việc cấu tạo một từ trong tiếng Việt có thể gồm nhiều tiếng nên có hiện tượng một số từ hiệu trùng hoặc là một phần của một từ không phải là từ hiệu. Các ví dụ sau minh họa cho vấn đề này.

Ví dụ 2:

a. *Dù trời mưa to nhưng chúng tôi vẫn đá bóng.*

Trong câu này, “dù” là từ hiệu trong cấu trúc “dù ... nhưng ...”. Nó được dùng để xác định hai ĐVDN trong câu này là “trời mưa to” và “chúng tôi vẫn đá bóng”.

b. *Trên các máy bay, các phi công và hành khách luôn được trang bị dù.*

Trong câu này, “dù” là danh từ chứ không phải là từ hiệu.

Ví dụ 3:

a. *Chúng tôi càng chờ, mưa càng nặng hạt.*

Trong câu này, “càng” là từ hiệu trong cấu trúc “... càng ... càng ...”.

b. *Cậu ăn chiếc càng cua này đi !*

Ở đây, “càng” là một phần của danh từ “càng cua”.

Chúng tôi đã sử dụng biện pháp tách từ và gán nhãn từ loại để xác định được chính xác các từ hiệu này.

Khác biệt thứ hai là, với tiếng Việt, ngay cả khi đã xác định được đâu là từ hiệu, nhiều khi chúng ta vẫn cần phải căn cứ thêm vào cấu trúc ngữ pháp của các thành phần bên cạnh từ hiệu đó thì mới có thể biết được từ hiệu đó có chức năng diễn ngôn hay không và chỉ ra vị trí tách phân đoạn.

Ví dụ 4:

a. *Vì trời mưa nên đường trơn.*

Trong câu này, “vì” có chức năng diễn ngôn, thỏa mãn cấu trúc [vì <câu> nên <câu>].

b. *Bác làm mọi việc đều vì nước vì dân.*

Trong câu này, “vì” đóng vai trò tạo nên cụm từ bổ nghĩa cho động từ “làm”, không có chức năng diễn ngôn.

Khác biệt thứ ba là, do các từ trong tiếng Việt ít có sự biến đổi về hình thái và độ dài các cụm từ hiệu khá ngắn nên có các từ hiệu khi thì nằm ở phân đoạn ngữ danh từ hoặc ngữ động từ (giống như một từ hiệu mạnh trong tiếng Anh), khi thì thậm chí không có chức năng diễn ngôn.

Ví dụ 5:

a. *[Khi được dự báo trước đợt rét lạnh,][bà con nông dân sẽ chuẩn bị đối phó hiệu quả, không bị mất mùa.]*

Trong câu này, “khi” đóng vai trò như một từ hiệu.

b. *Mặt trăng khi tỏ khi mờ.*

Trong câu này, “khi” không có chức năng diễn ngôn.

b. Phương pháp phân đoạn diễn ngôn

Từ các đặc điểm kể trên, có thể thấy kết quả phân đoạn diễn ngôn trong tiếng Việt sẽ rất thiếu chính xác nếu như chỉ dựa vào các từ hiệu mà không biết cấu trúc ngữ pháp của các thành phần văn bản xung quanh. Do đó, chúng tôi đề xuất sử dụng kết hợp các dấu hiệu sau để thực hiện thao tác phân đoạn :

❶ Dấu câu, dấu chú thích :

,	;	:	.
(	)	-	“
”			

Bảng 1: Các dấu câu

❷ Các từ hiệu :

Nhân-quả	vì, bởi, nhờ có, cho nên, do, thành ra, sở dĩ, là vì, chính vì, cũng là do, vì lẽ đó, do vậy nên, ...
Nhân-nhượng	tuy, nhưng, dầu sao, mặc dù, song, dầu cho, ...
Điều kiện	nếu, thì, hễ, nếu như, ...
Mục đích	để, cốt cho, để cho, ...

Bảng 2: Một số từ hiệu được phân loại theo quan hệ diễn ngôn

❸ Cấu trúc ngữ pháp: việc sử dụng kết quả phân tích cú pháp thực sự giúp cải thiện hiệu năng của mô-đun Phân đoạn diễn ngôn và mô-đun Xác định quan hệ diễn ngôn.

Một số ví dụ minh họa các trường hợp kết hợp khác nhau của 3 dấu hiệu trên:

-*Dấu câu + Cấu trúc ngữ pháp:*

[Trời mưa,][sân trơn,][bóng ướt.]

Luật tách được sử dụng trong trường hợp này là <câu>{,}!<câu><T>.

-*Dấu câu:*

[Ngày mai [(Trung thu)], mọi người đến nhà tớ chơi nhé !]

Luật tách được sử dụng trong trường hợp này là <T>!{(<T>)}!<T>.

-*Cấu trúc ngữ pháp:*

[Ngôi nhà [tôi mới xây] rất đẹp.]

Luật tách được sử dụng trong trường hợp này là <SUB <NP>!<SBAR>!> <PRE>.

-*Từ hiệu:*

[Vì trời mưa][nên đường trơn.]

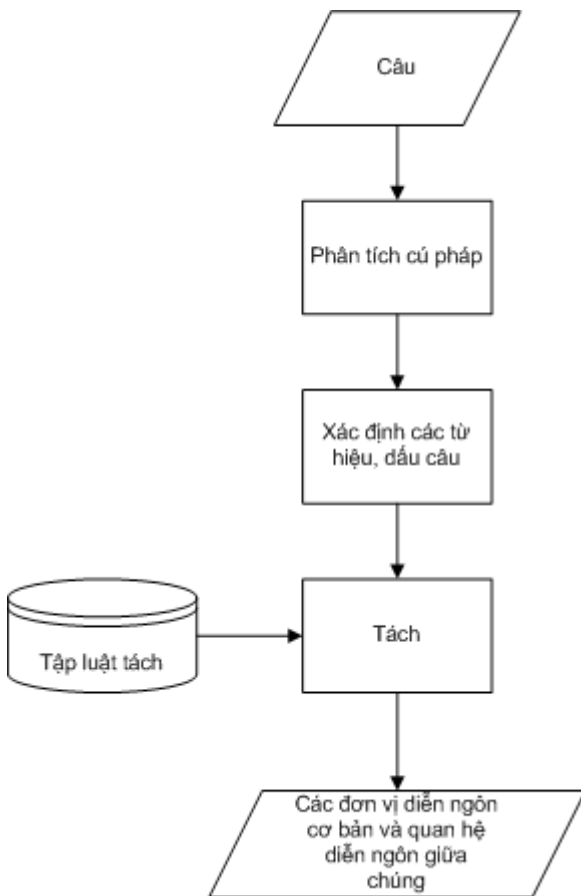
Luật : {<V>}<T>!{<nên>}<T>.

-Từ hiệu + dấu câu + cấu trúc ngữ pháp:

[*Khi* được dự báo trước đợt rét lạnh,][bà con nông dân sẽ chuẩn bị đối phó hiệu quả, không bị mất mùa.]

Luật : {**Khi**}<VP>!{,}<S>.

Qua các ví dụ trên, có thể thấy công việc có độ phức tạp cao nhất trong bước phân đoạn diễn ngôn là tách câu thành các đơn vị diễn ngôn cơ bản. Từ yêu cầu về đầu vào, đầu ra, và các cơ sở để phân đoạn ta có sơ đồ tách câu khái quát như sau:



Hình 4: Sơ đồ các thao tác cần thực hiện trong bước tách câu thành các ĐVDNCB

Tương tự như tập luật ngữ pháp trong hệ thống phân tích cú pháp, ở đây tập luật tách là yếu tố có ảnh hưởng lớn nhất tới độ chính xác của kết quả và khả năng phân tích của hệ thống. Để mô tả các trường hợp xuất hiện của từ hiệu trong câu, chúng tôi đề xuất sử dụng các khuôn dạng (pattern) có dạng:

$$_ \{ \text{vi} \} _ < \text{S} > _ ! _ \{ \text{nên} \} _ < \text{S} > _$$

$$_ < \text{T} > _ ! _ \{ : \} _ < \text{T} _ , _ >$$

Mỗi khuôn dạng là một xâu mô tả đặc điểm của một lớp các câu (hoặc một phần của câu), thỏa mãn một số các điều kiện về mặt cấu trúc ngữ pháp và vị trí xuất hiện của từ.

Ví dụ 6:

$$_ \{ \text{vi} \} _ < \text{S} > _ ! _ \{ \text{nên} \} _ < \text{S} > _$$

Khuôn dạng này mô tả một xâu có 2 vế, mỗi vế là một câu, đầu vế thứ nhất có chứa quan hệ từ “vì”, đầu vế thứ hai có chứa quan hệ từ “nên”. Câu “vì trời mưa nên đường trơn” thỏa mãn khuôn dạng này.

Ví dụ 7:

$$_ < \text{T} > _ ! _ \{ : \} _ < \text{T} _ , _ >$$

Khuôn dạng này mô tả một xâu có 2 thành phần ngữ pháp bất kỳ, ở giữa 2 thành phần ngữ pháp này là một dấu hai chấm “:”, thành phần ngữ pháp thứ hai có chứa dấu phẩy. Câu “*Thanh niên phải thường xuyên rèn luyện, trau dồi đạo đức cách mạng: cần, kiệm, liêm, chính; chí công, vô tư*” thỏa mãn khuôn dạng này.

3.2. Xác định quan hệ diễn ngôn

Tương tự quá trình phân đoạn diễn ngôn, quá trình xác định quan hệ diễn ngôn cũng được thực hiện ở 3 mức văn bản: dưới câu, giữa các câu, và giữa các đoạn văn.

a. Xác định quan hệ diễn ngôn mức dưới câu

Ở mức dưới câu, các phương tiện như từ hiệu và khuôn dạng đủ mạnh để có thể xác định quan hệ diễn ngôn ngay khi tách câu thành các ĐVDNCB. Do đó ta sẽ đưa thêm một số thông tin vào luật tách nhằm gắn quan hệ diễn ngôn với khuôn dạng tương ứng. Một luật tách khi đó sẽ có dạng:

$$\text{Marker} \mid \text{Pos1} \mid \text{Pos2} \mid \text{Pattern} \mid \text{Type} \mid \text{Rel} \mid \text{Score}$$

Trong đó:

- **Marker:** là quan hệ từ hoặc dấu câu (trong một số trường hợp có thể là cụm từ bất kỳ)
- **Pos1:** vị trí xuất hiện của từ hiệu trong về thứ nhất, nhận 1 trong 3 giá trị: B(đầu), M(giữa) hoặc E(cuối).
- **Pos2:** vị trí xuất hiện của từ hiệu trong về thứ hai, nhận 1 trong 3 giá trị: B(đầu), M(giữa) hoặc E(cuối).
- **Pattern:** khuôn dạng.
- **Type:** dùng để chỉ quan hệ diễn ngôn thuộc loại S-N, N-S hay N-N.
- **Rel:** tên quan hệ diễn ngôn.
- **Score:** điểm số phản ánh mức độ chắc chắn của việc xác định quan hệ diễn ngôn, nhận giá trị thực từ 0 đến 1.

Ví dụ 8:

vì | B | B | _{vi}_ <S_> !_{nên}_ <S_> |
S-N | CAUSE_EFFECT | 1

- Luật này được xét tới khi trong xâu xuất hiện từ hiệu “vì”.

- Cả 2 quan hệ từ trong đều đứng ở đầu các vế sau khi tách.

- Khuôn dạng để nhận biết xâu là _{vi}_ <S_> !_{nên}_ <S_>. Nếu xâu đang xét thỏa mãn khuôn dạng này thì tách thành 2 vế. Vị trí tách là ngay trước khi xuất hiện quan hệ từ “nên”.

- Quan hệ diễn ngôn giữa 2 vế sau khi tách là CAUSE_EFFECT (*nhân – quả*), thuộc loại quan hệ S-N.

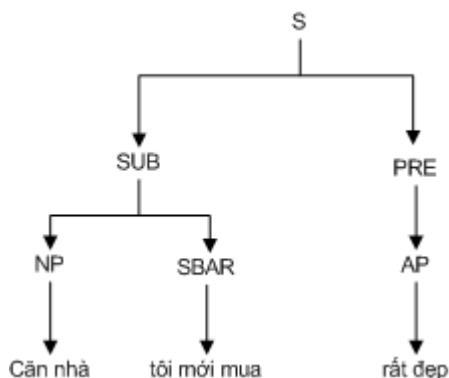
- Độ chắc chắn của quan hệ này là 100%.

Ngoài ra, đối với trường hợp câu có cấu trúc ngữ pháp dạng _<SUB_ <NP_> <SBAR_> >_ <PRE_>, ta sẽ xây dựng cây CTDN cho câu này ngay tại bước tách câu.

Ví dụ 9:

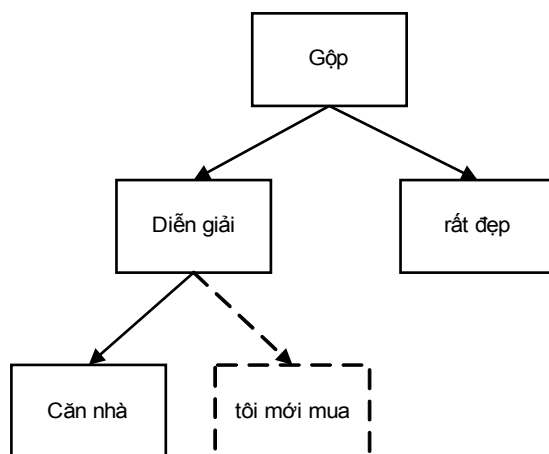
Xét câu “Căn nhà tôi mới mua rất đẹp”.

Cấu trúc ngữ pháp của câu này là:



Các ĐVDNCB của câu sau khi tách là [Căn nhà¹][tôi mới mua²][rất đẹp³].

ĐVDN 2 có quan hệ ELABORATION (*Diễn giải*) với ĐVDN 1. Giữa {1,2} và 3 có quan hệ SAME-UNIT (*gộp*). Đây là một nhãn quan hệ giả, dùng để mô tả quan hệ giữa hai đơn vị văn bản phải kết hợp với nhau mới đủ nghĩa)



Hình 5: Xây dựng cây CTDN cho các ĐVDNCB mức dưới câu ngay ở bước tách.

b. Xác định quan hệ diễn ngôn ở mức câu và đoạn

Khi xác định quan hệ diễn ngôn ở mức câu và đoạn, cấu trúc ngữ pháp sẽ không giúp được gì nhiều. Mặt khác, không phải câu nào, đoạn nào cũng xuất hiện từ hiệu (thậm chí có những đoạn văn mà cả đoạn không hề có một từ hiệu nào). Do đó, ngoài từ hiệu ra, ta cần thêm một số phương tiện khác để có thể liệt kê đủ số quan hệ diễn ngôn cho việc dựng thành cây ở bước tiếp theo. Trong trường hợp không có những từ có chức năng diễn ngôn để

suy đoán quan hệ diễn ngôn, ta sẽ căn cứ vào quan hệ về mặt nội dung giữa các đơn vị văn bản. Dưới đây là một số phương pháp dùng để chỉ ra quan hệ về mặt nội dung giữa hai đơn vị văn bản:

- **Sử dụng từ đồng hiện:** phương pháp này thống kê số từ cùng xuất hiện trong cả hai đơn vị văn bản. Khi số lượng từ đồng hiện lớn hơn một giới hạn nào đó, quan hệ giữa hai đơn vị văn bản sẽ được coi là ELABORATION (**Diễn giải**) và là JOINT (**Kết nối**) trong trường hợp ngược lại. Đây là phương pháp có độ chính xác thấp nhất trong các phương pháp được đề cập ở đây.

- **Sử dụng công thức cô-sin để xác định độ tương đồng:** lấy các từ trong đơn vị văn bản đứng trước làm véc-tơ cơ sở, véc-tơ hóa đơn vị văn bản thứ hai bằng véc-tơ cơ sở này, rồi tính toán độ tương đồng giữa hai đơn vị văn bản bằng công thức cô-sin:

$$\cos(X,Y) = \frac{\sum x_i \times y_i}{\sqrt{\sum (x_i)^2} \times \sqrt{\sum (y_i)^2}}$$

Trong đó (x_i) , (y_i) là các véc-tơ tương ứng với hai đơn vị văn bản.

Khi độ tương đồng giữa hai đơn vị văn bản lớn hơn một giới hạn nào đó, quan hệ giữa hai đơn vị văn bản sẽ được coi là ELABORATION và là JOINT trong trường hợp ngược lại. Phương pháp này có độ chính xác cao hơn đôi chút so với phương pháp trên vì có sử dụng thêm tần suất xuất hiện của từ. Ví dụ: câu thứ nhất có ba từ “thanh niên”, câu thứ hai có năm từ “thanh niên”; quan hệ giữa hai câu này rõ ràng không thể tương đương với quan hệ giữa hai câu mà trong mỗi câu, từ “thanh niên” chỉ xuất hiện có một lần được.

- **Sử dụng các cặp từ có quan hệ về mặt ngữ nghĩa với nhau.** So với hai phương pháp trên, đây là phương pháp có độ chính xác cao hơn và chỉ ra được nhiều loại quan hệ hơn. Ngoài ra phương pháp này còn có thể được sử dụng để xác định quan hệ giữa hai vế trong câu. Khi chưa sử dụng phương pháp này, trong trường hợp hai vế với câu tạo ngữ pháp giống nhau chỉ phân cách nhau bởi một dấu

phẩy, hai vế đó sẽ được coi là có quan hệ LIST (**liệt kê**).

Ví dụ 10:

a. *Hùng bị ngã. Dừng vừa đẩy nó.*

Trong cặp câu trên có xuất hiện cặp từ “đẩy” – “ngã” có quan hệ **Nhân-quả** → nếu không còn manh mối nào khác, có thể coi hai câu này có quan hệ **nhân-quả** với nhau.

b. *[Đêm tháng năm chưa nằm đã sáng,][ngày tháng mười chưa cười đã tối.]*

Trong hai vế câu trên có xuất hiện cặp từ “sáng” – “tối” có quan hệ **Tương phản** → nếu không còn manh mối nào khác, có thể coi hai câu này có quan hệ **Tương phản**.

Nhận xét: Có thể thấy độ chính xác của hai phương pháp đầu tiên tỉ lệ thuận với chiều dài của hai đơn vị văn bản đang xét (do dựa trên kết quả thống kê). Phương pháp thứ ba có độ chính xác khá cao và có thể áp dụng cho rất nhiều trường hợp nhưng đòi hỏi phải xây dựng bộ từ điển trong đó mô tả ràng buộc về mặt ngữ nghĩa giữa các từ với nhau (giống như WordNet). Trong phạm vi bài này, chúng tôi lựa chọn phương pháp thứ nhất để cài đặt cho mô-đun phân tích CTDN.

3.3. Xây dựng cây cấu trúc diễn ngôn

Quá trình xây dựng cây CTDN trải qua ba giai đoạn:

- Xây dựng cây CTDN mức dưới câu.
- Xây dựng cây CTDN mức đoạn văn.
- Xây dựng cây CTDN mức văn bản.

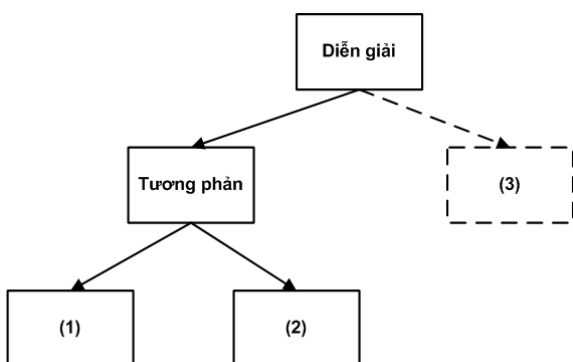
Trong ba bước trên, bước xây dựng cây diễn ngôn mức dưới câu đã được thực hiện ngay tại khâu Phân đoạn diễn ngôn. Đối với mức đoạn văn và mức văn bản, ta sẽ xây dựng cây CTDN từ tập các QHDN, sử dụng giải thuật bottom-up (tương tự giải thuật CYK).

Ví dụ 11:

Với tập các QHDN:

$\{rhet_rel(CONTRAST, 1, 2) \oplus rhet_rel(CONTRAST, 1, 3)$
 $\quad rhet_rel(ELABORATION, 1, 3)$

ta sẽ xây dựng được cây CTDN như sau:



Hình 6: Cây CTDN cho ví dụ 11

4. Tóm tắt văn bản sử dụng cấu trúc diễn ngôn

Trong cách tiếp cận của chúng tôi, văn bản tóm tắt được hình thành từ các thành phần quan trọng của văn bản. Độ quan trọng của một ĐVVB được xác định dựa trên một hàm lượng giá. Hàm lượng giá này được mô tả ở phần 4.1.

4.1. Hàm lượng giá

Chúng tôi sử dụng hàm lượng giá do Marcu [10] đề xuất khi tạo tóm tắt cho văn bản tiếng Anh. Hàm này là công thức đệ quy dùng để tính toán độ quan trọng của một đơn vị u trong một cấu trúc diễn ngôn D có độ sâu d :

$$score(u, D, d) = \begin{cases} 0 & \text{if } D \text{ is NULL} \\ d & \text{if } u \in promotion(D) \\ \max(score(u, left(D), d-1), & \\ score(u, right(D), d-1)) & \end{cases}$$

Trong đó $promotion(D)$ là tập các ĐVDNCB quan trọng nhất nằm dưới D , được xác định như sau:

- Promotion của nút lá tương ứng với ĐVDNCB thứ i là $\{i\}$.
- Promotion của nút D có hai nút con là L và R sẽ chứa $promotion(L)$ nếu L là nhân và không chứa $promotion(L)$ trong trường hợp ngược lại. Tương tự đối với nút con R .

Ví dụ 13:

Xét cây CTDN ở hình 2, áp dụng công thức trên ta thu được bảng giá trị sau:

Đơn vị diễn ngôn cơ bản	1	2	3	4	5
Độ quan trọng	4	4	1	3	3

4.2. Giải thuật tóm tắt văn bản

Đầu vào :

Văn bản T .

Số p : $1 \leq p \leq 100$.

Đầu ra :

$p\%$ các đơn vị văn bản quan trọng nhất của T .

Các bước thực hiện :

1. Xây dựng cây cấu trúc diễn ngôn DS của T bằng giải thuật phân tích diễn ngôn.
2. Xác định thứ tự theo mức độ quan trọng các đơn vị của cây diễn ngôn bằng công thức lượng giá ở trên.
3. Tiến hành phân cụm các đơn vị theo độ quan trọng.
4. Chọn ra n đơn vị đầu tiên trong danh sách để tạo văn bản tóm tắt. Ở đây n được tính toán sao cho độ dài văn bản tóm tắt gần $p\%$ độ dài văn bản ban đầu nhất. Chú ý các đơn vị thuộc cùng một cụm phải được cùng chọn hoặc cùng không được chọn.

Ví dụ 14: Thực hiện sinh văn bản tóm tắt cho đoạn văn bản với CTDN ở hình 2.

Sau khi xác định được giá trị đo độ quan trọng của đơn vị văn bản như trong bảng trên, ta sắp xếp các đơn vị theo thứ tự giảm dần độ quan trọng :

$$1, 2 > 4, 5 > 3$$

Giả sử $p = 40\%$, ta sẽ có văn bản tóm tắt gồm các thành phần 1, 2. Văn bản tóm tắt trong trường hợp này là “Điểm nổi bật trong quan điểm giáo dục lý tưởng cho thanh niên của

Bác Hồ là đưa thanh niên vào các tổ chức chính trị, xã hội”.

5. Kết luận

Bài này trình bày hướng tiếp cận sử dụng cấu trúc diễn ngôn để giải quyết bài toán tóm tắt văn bản mà trọng tâm là phương pháp phân tích cấu trúc diễn ngôn cho văn bản tiếng Việt. Do thời gian thực hiện hạn chế, chúng tôi chưa xây dựng được tập luật đủ lớn để có thể áp dụng cho nhiều trường hợp văn bản cũng như để tiến hành kiểm thử và đánh giá hệ thống. Tuy nhiên, với độ chính xác khá cao đã được chỉ ra trong các nghiên cứu về tóm tắt văn bản tiếng Anh, chúng tôi hy vọng hướng tiếp cận sử dụng cấu trúc diễn ngôn sẽ giúp xây dựng được các hệ thống tóm tắt tự động văn bản tiếng Việt thực sự hiệu quả và hữu ích.

Tài liệu tham khảo

- [1] Báo điện tử Đảng Cộng sản Việt Nam, Tư tưởng Hồ Chí Minh, www.cpv.org.vn.
- [2] Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Aravind Joshi and Bonnie Webber. 2003. D-LTAG System: Discourse Parsing with a Lexicalized Tree-Adjoining Grammar. *Journal of* 12(3), 261-279.
- [3] Barbara J. Grosz and Candace L. Sydner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12:175-204.
- [4] Hoàng Trọng Phiến, Ngữ pháp tiếng Việt – Câu, Nhà xuất bản Đại học và Trung học chuyên nghiệp 1980.
- [5] Lê Thanh Hương, Automatic Discourse Structure Generation Using Rhetorical Structure Theory. 2004. Ph.D. dissertation, Middlesex University, U.K.
- [6] Linguistic Data Consortium, The Rhetorical Structure Theory Discourse Treebank Publication, catalog number LDC2002T07 and ISBN 21-58563-223-6.
- [7] H.P. Luhn, The automatic creation of literature abstracts.
- [8] Inderjeet Mani, Eric Bloedorn, Barbara Gates, Using Cohesion and Coherence Models for Text Summarization.
- [9] William Mann, Sandra Thompson, Rhetorical structure theory: A theory of text organization.
- [10] Daniel Marcu, The Rhetorical Parsing, Summarization and Generation, of Natural Language Texts, Ph.D thesis 1997, Department of Computer Science – University of Toronto.
- [11] Dragomir R. Radev, A tutorial about Text Summarization, ACM SIGIR 2001.
- [12] Stuart Russell, Peter Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall 1995.
- [13] Simone Teufel, Argumentative Zoning: Information Extraction from Scientific Text, University of Edinburgh 1999.