Một cách tiếp cận trong việc tự động sinh các biểu diễn tương đương của đoạn văn bản

An approach to automatically generate different presentations of natural language paraphrases

Lê Thanh Hương

Abstract: This paper proposes a system to automatically generate different presentations of a paraphrase. To build such a system, three main tasks need to be done: (1) recognizing the discourse structure of a document; (2) dealing with co-references (optional); and (3) restating sentences. The system has firstly been implemented for English language using two main modules corresponding to the tasks (1) and (3). The experiments have shown promising results. It indicates that the system will be improved if the task (2) is also implemented. We remain this task for future work.

I. Đặt vấn đề

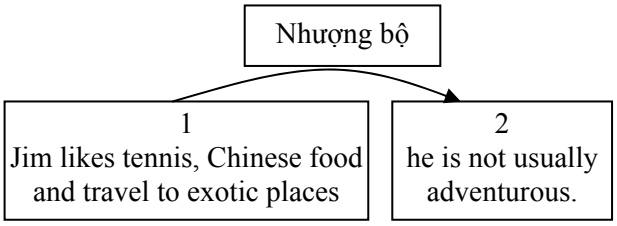
Khi diễn đạt cùng một nội dung, mỗi người có thể trình bày theo những cách khác nhau. Những cách phát biểu khác nhau đó có thể do thói quen, do người phát ngôn muốn nhấn mạnh đến một khía cạnh nào đó của sự việc, hay do người phát ngôn muốn thể hiện lại nội dung theo cách dễ hiểu hơn.

Trên thế giới, hướng nghiên cứu về tự động sinh các cách phát biểu khác nhau cho một đoạn văn bản đang được quan tâm trong thời gian gần đây. Phần lớn các nghiên cứu tập trung vào việc chuyển đổi từ vựng và cú pháp các câu đơn lẻ [1, 3, 4]. Chúng tôi hướng tới việc xây dựng bài toán ở phạm vi lớn hơn: thay đổi cấu trúc toàn bộ văn bản. Nghiên cứu được thực hiện trước tiên cho tiếng Anh. Để thực hiện việc đó, chúng tôi sử dụng cách tiếp cận dựa trên việc phân tích cấu trúc diễn ngôn của văn bản.[1] Ví dụ, xét câu sau:

(1) *Although* Jim likes tennis, Chinese food and travel to exotic places, he is not usually adventurous.

[1] Cấu trúc diễn ngôn của văn bản cho biết mối quan hệ diễn ngôn giữa các thành phần của văn bản. Xem phần III để biết thêm chi tiết.

Cấu trúc diễn ngôn của câu này là:



Hình 1 - Cấu trúc diễn ngôn của câu (1)

Hình 1 thể hiện quan hệ diễn ngôn "Nhượng bộ" (Concession) giữa mệnh đề 1 "*Jim likes tennis, Chinese food and travel to exotic places*" và mệnh đề 2 "*he is not usually adventurous*", trong đó mệnh đề 1 là mệnh đề phụ và mệnh đề 2 là mệnh đề chính trong câu. Hệ thống sinh ra văn bản mới bằng các cách: (i) phát biểu lại các mệnh đề; và/hoặc (ii) đổi vị trí các mệnh đề; và/hoặc (iii) thay đổi từ nối giữa chúng. Một trong các cách phát biểu lại của ví dụ (1) là:

(1a) Jim likes tennis, Chinese food and travel to exotic places. *However*, he is not usually adventurous.

Câu ghép ban đầu được tách thành hai câu đơn trong cách phát biểu mới. Từ nối "*although*" đứng trước mệnh đề thứ nhất đã được thay bằng từ nối "*However*" đứng trước mệnh đề thứ hai. Việc tách các câu ghép thành câu đơn như trong ví dụ này làm cho đoạn văn dễ đọc, dễ hiểu hơn.

Trong bài này, chúng tôi sẽ đề xuất một hệ thống sinh các cách phát biểu tương đương của văn bản, giới thiệu một số cài đặt thử nghiệm và đánh giá kết quả. Phần còn lại của bài báo được trình bày như sau. Mô hình hệ thống được giới thiệu ở phần 2. Phần 3 mô tả việc xây dựng cấu trúc diễn ngôn của văn bản. Vấn đề sinh văn bản từ cấu trúc diễn ngôn được đề cập ở phần 4. Phần 5 đưa ra các kết quả thí

nghiệm dựa trên hệ thống đã xây dựng. Các đánh giá và hướng phát triển của hệ thống sẽ được trình bày ở phần 6.

II. Mô hình hệ thống

Với dữ liệu vào của hệ thống là văn bản do người soạn thảo, hệ thống sẽ sinh ra các văn bản có nội dung tương tự như nội dung đưa vào nhưng với các cách viết khác nhau. Để có một hình dung về các bước cần tiến hành, ta hãy phân tích ví dụ 2 sau:

(2) The child had a fever *because of* hunger and coldness.

Nếu bỏ qua khả năng thay đổi các danh từ và động từ chính trong câu, một số khả năng biến đổi câu trên là:

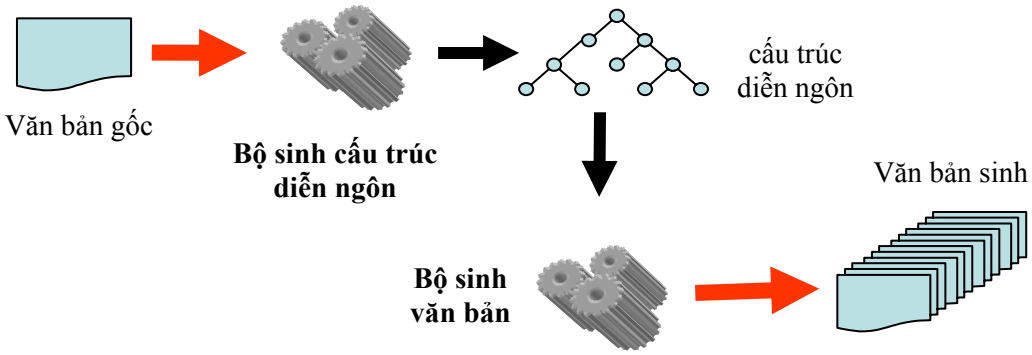
(2a) The child had a fever *since* he had suffered from hunger and coldness.

(2b) *Since* the child had suffered from hunger and coldness, he had a fever.

Nếu so sánh các câu biến đổi trên với câu ban đầu, ta có thể thấy một số vấn đề trong việc biến đổi câu là:

1. **Xác định cấu trúc diễn ngôn** của văn bản. Trong ví dụ 2, ta có một mệnh đề "*the child had a fever*" và một danh ngữ (ta gọi là giả mệnh đề - xem phần 3.1) là "*hunger and coldness*". Trong đó "*hunger and coldness*" là (giả) mệnh đề phụ chỉ nguyên nhân của mệnh đề chính "*the child had a fever*". Từ nối "*because of*" xác định quan hệ "Nguyên nhân" trong ví dụ này.
2. **Thay đổi các từ nối** giữa các thành phần của văn bản. Trong ví dụ 2, mệnh đề và danh ngữ liên hệ với nhau bằng quan hệ "Nguyên nhân". Do vậy, từ nối "*because of*" chỉ quan hệ "Nguyên nhân" có thể thay thế bằng một từ nối chỉ nguyên nhân khác ("*since*" - ví dụ 2a,2b).
3. **Thay đổi cấu trúc văn bản**. Việc thay đổi này phải đảm bảo tính logic và tính hiểu được của văn bản. Một vấn đề khó khăn trong việc chuyển đổi vị trí các thành phần của văn bản là nó thường làm hỏng các liên kết về mặt logic của văn bản. Lấy ví dụ đơn giản về một cách phát biểu khác của ví dụ 2a: "*Since he had suffered from hunger and coldness, the child had a fever*". Thông thường, một đại từ xuất hiện khi danh từ tương ứng với nó đã được nhắc đến trước trong đoạn văn. Trong trường hợp này, đại từ "*he*" đứng trước "*the child*" là sai qui tắc. Do vậy, hệ thống sinh văn bản mới từ văn bản cũ phải có khả năng nhận biết các danh từ và đại từ cùng chỉ đến một đối tượng. Ta gọi những từ này là **từ đồng tham chiếu** (co-reference). Khi thay đổi cấu trúc văn bản, hệ thống phải có khả năng chuyển đại từ thành danh từ nếu danh từ ứng với đại từ đó chưa được nhắc đến trước đó.
4. **Chuyển đổi các đơn vị diễn ngôn**. Danh ngữ "*hunger and coldness*" ở ví dụ 2 được thay thế bằng mệnh đề "*he had suffered from hunger and coldness*" ở ví dụ 2a. Vấn đề này sẽ được thảo luận kỹ hơn ở phần 4.2.

Với các vấn đề đặt ra ở trên, hệ thống của chúng tôi bao gồm hai thành phần chính: bộ sinh cấu trúc diễn ngôn và bộ sinh văn bản. Mô hình hệ thống được giới thiệu ở hình 2.



Hình 2 – Mô hình hệ thống sản sinh các cách phát biểu tương đương của văn bản

Bộ sinh cấu trúc diễn ngôn có nhiệm vụ sinh cấu trúc diễn ngôn của văn bản và xác định các quan hệ đồng tham chiếu. Bộ sinh văn bản có nhiệm vụ xây dựng cấu trúc văn bản mới, thay các từ nối và cách phát biểu các mệnh đề. Hệ thống trước tiên được xây dựng cho ngôn ngữ tiếng Anh.

III. Sinh cấu trúc diễn ngôn [2]

Việc sinh cấu trúc diễn ngôn của văn bản được thực hiện qua các bước:

1. Chia văn bản thành các đơn vị diễn ngôn;
2. Nhận dạng các quan hệ diễn ngôn giữa các đoạn văn bản;
3. Lựa chọn và kết hợp các quan hệ diễn ngôn tạo ra ở bước 2 để tạo ra một cấu trúc diễn ngôn cho toàn bộ văn bản.

III.1 Chia văn bản thành các đơn vị diễn ngôn

Cấu trúc diễn ngôn được xây dựng từ các thành phần diễn ngôn. Thành phần diễn ngôn nhỏ nhất gọi là đơn vị diễn ngôn (Mann and Thompson, 1988). Mỗi đơn vị diễn ngôn thường diễn đạt một ý trọn vẹn. Đơn vị diễn ngôn có thể là câu đơn, mệnh đề (như mệnh đề 2 trong ví dụ 1) hay cách thành phần có vai trò như mệnh đề trong câu, tạm gọi là giả mệnh đề (như danh ngữ "*hunger and coldness*" trong ví dụ 2). Giả mệnh đề được nhận biết bởi một danh ngữ đi kèm với các từ nối đặc biệt như *according to, as a result of, although, because of, but also, despite, despite of, in spite of, irrespective, not only, regardless, without.* Chúng tôi gọi những từ nối đó là các từ nối mạnh, nhằm phân biệt chúng với các từ nối không có tác dụng biến danh ngữ thành các đơn vị diễn ngôn.

Vì một đơn vị diễn ngôn điển hình là mệnh đề hoặc câu đơn, để chia văn bản thành các đơn vị diễn ngôn, trước tiên chúng tôi tiến hành bước phân tách thứ nhất: phân tách văn bản dựa trên cấu trúc cú pháp của câu.[3] Để giải quyết trường hợp giả mệnh đề, chúng tôi tiến hành bước phân tách thứ hai sau bước phân tách thứ nhất. Quá trình phân tách thứ hai này tìm các từ nối mạnh trong các câu đơn và các mệnh đề. Sau đó nó tiếp tục tách các câu đơn hay mệnh đề thành các đơn vị nhỏ hơn nếu từ nối mạnh xuất hiện trong các thành phần đó. Khi từ nối mạnh xuất hiện, câu đơn/mệnh đề được chia làm hai đơn vị diễn ngôn: một là danh ngữ đi kèm với từ nối mạnh, và một là phần còn lại của câu đơn/mệnh đề.

III.2 Các yếu tố xác định quan hệ diễn ngôn

Ba yếu tố quan trọng nhất được sử dụng trong hệ thống này để xác định quan hệ diễn ngôn là các từ nối (cue phrases) và các từ khoá (keywords) trong danh ngữ và động ngữ. Các từ nối (ví dụ, *despite of*, *however*,…) là các từ/ngữ đặc biệt được sử dụng để nối các đơn vị diễn ngôn.[4] Từ nối "*when*" trong ví dụ 3 xác định quan hệ diễn ngôn "Hoàn cảnh" giữa hai mệnh đề "*He was staying at home*" và "*the police arrived*".

(3) [He was staying at home][*when* the police arrived.]

Từ khoá trong danh ngữ, động ngữ là những từ/ngữ phát tín hiệu về quan hệ diễn ngôn như trong các ví dụ (4) và (5).

(4) [New York style pizza meets Californian ingredients,][and the *result* is the pizza from this Church Street pizzeria.]

(5) [By the end of this year, 63-year-old Chairman Silas Cathcart retires to his Lake Forest, Ill., home.][And that *means* 42-year-old Michael Carpenter will for the first time take complete control of Kidder.]

Danh từ "*result*" chỉ quan hệ "Nguyên nhân" trong ví dụ 4. Động từ "*means*" xác định quan hệ "Bổ sung thông tin" giữa hai câu trong ví dụ (5). Ngoài các từ nối và từ khoá nói trên, các yếu tố liên kết văn bản khác cũng được sử dụng để xác định quan hệ diễn ngôn. Các yếu tố đó là cấu trúc cú pháp câu, sự tham chiếu về thời gian, các từ đồng nghĩa và hiện tượng tỉnh lược các thành phần câu.

III.3 Nhận dạng quan hệ diễn ngôn

Chúng tôi sử dụng một tập gồm 13 quan hệ diễn ngôn để biểu diễn cấu trúc diễn ngôn. Các quan hệ này là: nhượng bộ, nguyên nhân, hoàn cảnh, điều kiện, bổ sung thông tin, phát biểu lại, phương tiện, mục đích, liên kết, tách rời, tuần tự, đối lập và kết nối (*concession, cause, circumstance, conditional, elaboration, restatement, means, purpose, disjunction, conjunction, sequence, contrast, joint)*. Kết nối (*joint*) là quan hệ mặc định, được sử dụng khi không tìm được quan hệ diễn ngôn nào khác liên kết hai đoạn văn bản. Quá trình phát hiện các quan hệ diễn ngôn dựa trên sự xuất hiện của các yếu tố xác định quan hệ diễn ngôn (đề cập ở phần 3.2). Chúng tôi đã xây dựng một tập luật để phát hiện các quan hệ diễn ngôn dựa trên các yếu tố đó. Ví dụ:

Nếu trong câu ghép có một mệnh đề chứa từ nối chỉ quan hệ "Nhượng bộ" (ví dụ "although") thì mệnh đề đó là mệnh đề phụ trong mối quan hệ "Nhượng bộ" với mệnh đề còn lại trong câu.

Vì mỗi yếu tố có một ảnh hưởng mạnh/yếu khác nhau trong việc xác định quan hệ diễn ngôn, mỗi luật được gán một trọng số khác nhau trong khoảng

[2] Xem [5, 6] để biết chi tiết hơn về vấn đề phân tích cấu trúc diễn ngôn.

[3] Bộ phân tích cú pháp của Charniak [2] được sử dụng để sinh cấu trúc cú pháp của câu.

[4] Khi xây dựng chương trình, chúng tôi tổ chức các file riêng để lưu các từ nối và từ khoá này.

0 đến 100. Các luật liên quan đến từ nối có trọng số cao nhất (100) vì từ nối là yếu tố mạnh nhất để xác định các quan hệ diễn ngôn. Từ khoá trong danh ngữ và động ngữ là yếu tố mạnh thứ hai sau từ nối nên có trọng số 90. Trọng số của các yếu tố khác nằm trong khoảng 20 đến 80 vì các yếu tố này yếu hơn các từ khoá (xem [5] để có các mô tả chi tiết hơn về các luật này).

Bên cạnh việc gán trọng số cho các luật, chúng tôi còn gán trọng số cho các từ nối và từ khoá. Các luật ứng với từ nối có trọng số 100 nghĩa là hệ thống chắc chắn 100% về quan hệ diễn ngôn được phát hiện dựa trên từ nối. Tuy nhiên, điều này chỉ đúng nếu từ nối chắc chắn xác định quan hệ diễn ngôn đó. Trên thực tế, các từ nối có độ chắc chắn khác nhau trong việc xác định các quan hệ. Ví dụ, từ nối "*although*" luôn chỉ định quan hệ "Nhượng bộ", trong khi từ nối "*and*" có thể chỉ định quan hệ "Liên kết", "Tách rời", hoặc "Bổ sung thông tin". Điều đó có nghĩa là luật ứng với từ nối "*and*" không chắc chắn 100% về quan hệ "Liên kết" giữa hai đoạn văn bản. Nói cách khác, ta cần giảm trọng số của luật khi luật liên quan đến một từ nối yếu. Chúng tôi gán trọng số của một từ nối trong khoảng [0, 1]. Trọng số thực tế của luật ứng với từ nối là:

Actual-score(luật) = Score(luật) * Score(từ nối).

Vì một từ khoá cũng có thể phát tín hiệu về một vài quan hệ diễn ngôn, các từ khoá ứng với danh ngữ và động ngữ cũng được gán trọng số trong khoảng [0, 1]. Trọng số thực tế của luật ứng với từ khoá là:

Actual-score(luật) = Score(luật)* Score(từ khoá).

Trọng số thực tế của luật ứng với các yếu tố còn lại là:

Actual-score(luật) = Score(luật)

Nếu một số luật ứng với một quan hệ diễn ngôn thoả mãn thì trọng số của luật sẽ là tổng trọng số của tất cả các yếu tố góp phần vào quan hệ đó.

Total-heuristic-score = $\sum$ Actual-score (luật)

Hệ thống tìm các yếu tố xác định quan hệ diễn ngôn theo trình tự sau: từ nối, từ khoá, và các yếu tố khác. Một quan hệ diễn ngôn sẽ được gán cho quan hệ giữa hai đoạn văn bản nếu *total-heuristic-score* của quan hệ đó lớn hơn hoặc bằng một giá trị ngưỡng θ. Việc chọn giá trị ngưỡng hợp lý rất quan trọng vì sự thay đổi của giá trị này sẽ ảnh hưởng đến việc xác định các quan hệ diễn ngôn, dẫn đến thay đổi cấu trúc diễn ngôn của văn bản. Hiện tại, chúng tôi gán cho ngưỡng này giá trị 30 (so với 100 là giá trị lớn nhất của một luật). Giá trị này được xác định dựa trên việc thử nghiệm và đánh giá độ chính xác của hệ thống với các giá trị ngưỡng khác nhau.

Đánh trọng số các yếu tố xác định quan hệ diễn ngôn:

Rõ ràng là việc đưa ra các trọng số thích hợp cho từ nối, từ khoá và các luật xác định quan hệ diễn ngôn rất quan trọng trong việc sinh cấu trúc diễn ngôn. Hiện tại, trọng số của các luật được gán dựa trên kinh nghiệm của chuyên gia. Hiện nay, tập các yếu tố xác định quan hệ diễn ngôn cũng như các trọng số của chúng hoạt động tốt với tập dữ liệu thử nghiệm. Trong thời gian tới, chúng tôi dự định sử dụng phương pháp học máy nhằm tối ưu hoá các trọng số này.

III.4 Xây dựng cấu trúc diễn ngôn

Với một văn bản, ta có thể tìm ra nhiều mối quan hệ khác nhau và nhiều cách liên kết khác nhau giữa các mệnh đề, câu và đoạn văn. Ví dụ, một câu có thể có quan hệ "Bổ sung thông tin" cho câu trước, nhưng lại có quan hệ "Tuần tự" với câu sau. Vì vậy, ta cần phải lựa chọn và kết hợp các quan hệ diễn ngôn tạo ra ở các bước trên để tạo ra một cấu trúc diễn ngôn duy nhất cho toàn bộ văn bản. Để tận dụng quan hệ giữa các mệnh đề trong câu (dựa trên cấu trúc cú pháp của câu), chúng tôi tách việc xây dựng cấu trúc diễn ngôn của văn bản thành hai mức: mức câu và mức văn bản. Bộ phân tích mức câu sinh cấu trúc diễn ngôn cho từng câu dựa trên quan hệ cú pháp giữa các mệnh đề. Trong khi đó, bộ phân tích mức văn bản sử dụng thuật toán tìm kiếm kiểu hạt (beam search) trên tập các quan hệ diễn ngôn có thể có giữa các câu và đoạn văn để tìm cách kết hợp các quan hệ diễn ngôn nhằm mô tả cấu trúc diễn ngôn của văn bản một cách hợp lý nhất.

IV. Sinh văn bản từ cấu trúc diễn ngôn

Dựa trên các kết quả nghiên cứu [5] và [8][5], chúng tôi đề xuất hệ thống sinh các cách phát biểu khác nhau của một đoạn văn bản như sau. Với đầu vào là cấu trúc diễn ngôn của văn bản nguồn, bộ sinh văn bản sẽ sinh ra các cách phát biểu khác nhau của văn bản đó. Biện pháp sinh văn bản đơn giản nhất là thay đổi các từ nối giữa các đơn vị diễn ngôn. Một phương pháp ở mức cao hơn là chuyển đổi vị trí các đơn vị diễn ngôn. Ví dụ, với câu ban đầu:

(6) Doctors recommend Elixir *since* it gives quick results and it has few side-effects.

Cấu trúc diễn ngôn của câu này được sinh bởi modul sinh cấu trúc diễn ngôn là:

```
<RhetRep relation=cause>
```

[5] Chúng tôi xin chân thành cảm ơn giáo sư Donia Scott và tiến sĩ Richard Power đã hỗ trợ chúng tôi thực hiện nghiên cứu này.

```
   <SemRep syncat=clause prop="doctors
recommend Elixir"/>
   <RhetRep relation=conjunction>
   <SemRep syncat=clause prop="it gives
quick results"/>
   <SemRep syncat=clause prop="it has
few side-effects"/>
   </RhetRep>
   </RhetRep>
```

trong đó,

- thẻ **RhetRep** và **/RhetRep** (Rhetorical Representation) dùng để mở đầu và kết thúc một biểu diễn cấu trúc lưu trữ của quan hệ diễn ngôn.
- thẻ **SemRep** (Semantic Representation) đánh dấu mở đầu các thông tin về một đoạn văn bản.
- thẻ **syncat** (syntactic catergory) cho biết vai trò ngữ pháp của đoạn văn bản (mệnh đề, câu, đoạn văn)
- thẻ **relation** cho biết tên quan hệ diễn ngôn giữa các đoạn văn bản.
- thẻ **prop** (proposition) nhằm lưu nội dung đoạn văn bản.

Nếu chỉ thay đổi từ nối, ta sẽ có câu (6a) sau:

(6a) Doctors recommend Elixir *because* it gives quick results and it has few side-effects.

Câu này không khác mấy với câu ban đầu. Nếu chỉ chuyển vị trí các mệnh đề, ta sẽ có câu (6b) sau:

(6b) *Since* it gives quick results and it has few side effects, doctors recommend Elixir.

Tuy cách này có thể tạo ra các câu khác nhiều hơn so với câu ban đầu, nó lại thường gây ra sự không mạch lạc. Ở ví dụ (6b), đại từ đi trước danh từ mà nó thay thế. Điều này không đúng với qui tắc ngữ pháp. Để giải quyết vấn đề này, ta phải dùng cơ chế thay đổi các từ đồng tham chiếu. Cơ chế này được giới thiệu ở phần tiếp theo.

IV.1 Thay đổi các từ đồng tham chiếu

Để giải quyết vấn đề câu không mạch lạc nói ở phần trên, ngoài việc xác định các đơn vị diễn ngôn, văn bản được phân tích chi tiết hơn bằng cách xác định các thuộc tính ngữ nghĩa đơn giản của danh từ. Từ đó xác định các từ đồng tham chiếu. Ví dụ, thông tin phân tích từ của ví dụ (6) được thể hiện qua ngôn ngữ đánh dấu như sau:

```
   <edu>
   <np id=1 phrase="doctors"
   class="human" number="plural"/>
   recommend
   <np id=2 phrase="Elixir"
   class="thing" number="singular"/>
   </edu>
   <edu>
   <pronoun id=2 phrase="it"/>
   gives
   <np id=3 phrase="quick results"
   class="thing" number="plural"/>
   </edu>
   <edu>
   <pronoun id=2 phrase="it"/>
   has
   <np id=4 phrase="few side-effects"
   class="thing" number="plural"/>
   </edu>
```

Việc xác định các từ đồng tham chiếu hỗ trợ cho quá trình chuyển đổi các đơn vị diễn ngôn theo ba cách.

1. Biến đổi đại từ thành danh từ, ví dụ "*it gives quick results*" có thể chuyển thành "*Elixir gives quick results*"
2. Biến đổi danh từ thành đại từ, ví dụ "*doctors recommend Elixir*" có thể chuyển thành "*doctors recommend it*"
3. Lược bớt đại từ, ví dụ "*it gives quick results and it has few side-effects*" có thể chuyển thành "*it gives quick results and has few side-effects*".

Như vậy, ví dụ (6) có thể chuyển thành:

(6c) *Since* Elixir gives quick results and has few side effects, doctors recommend it.

Việc chuyển đổi văn bản mới từ văn bản cũ đã trôi chảy hơn nhờ sự đóng góp của cơ chế chuyển đổi từ đồng tham chiếu. Tuy nhiên, chương trình vẫn bị giới hạn vì chưa có khả năng chuyển đổi thời chủ động thành bị động, chuyển đổi danh ngữ thành mệnh đề/câu đơn và ngược lại. Vấn đề này sẽ được phân tích kỹ hơn ở phần tiếp theo.

IV.2 Chuyển đổi các đơn vị diễn ngôn

Chuyển đổi các đơn vị diễn ngôn là một trong những phương pháp phát biểu lại câu. Hệ thống phải có khả năng chuyển danh ngữ thành mệnh đề/câu đơn và ngược lại. Đồng thời, hệ thống phải có khả năng nhận biết và chuyển đổi thời và thể của đơn vị diễn ngôn. Ví dụ, với câu ban đầu là:

(7) He came late *because of* the rain.

Một cách phát biểu khác của câu này là:

(7a) He came late *because* it was raining.

Danh ngữ "*the rain*" trong ví dụ (7) đã được chuyển thành chủ ngữ của mệnh đề phụ "*it was raining*" trong ví dụ (7a). Chủ ngữ này đi với động từ ở thời quá khứ tiếp diễn do sự kiện trong mệnh đề chính diễn ra ở thời quá khứ. Tuy nhiên, không phải

lúc nào ta cũng có thể sử dụng đại từ ở ngôi số 3 "*it*" và động từ "*to be*" để chuyển đổi danh ngữ thành mệnh đề/câu đơn. Ví dụ (8) minh hoạ một tình huống như vậy (ở đây ta chỉ xét đến chuyển đổi câu đầu tiên trong đoạn).

(8) Andy is going to be dangerous this year *because of* his style. He has great strength and power and is such an entertaining player. Andy knows what he wants to do with his career and will step it up to get the win he wants here at Wimbledon.

Nếu ta chuyển "*because of his style*" thành "*because it is his style*" thì câu này sẽ càng khó hiểu hơn câu ban đầu. Thay vào đó, ta có thể nói "*Andy is going to be dangerous this year because he has a powerful style*." Cần phải nhấn mạnh rằng "*because of his style*" không thể hiểu được nếu ta không đọc tiếp các câu sau, do vậy sẽ không thể chuyển danh ngữ này thành mệnh đề tương ứng.

Như vậy ta có thể thấy rằng việc chuyển đổi danh ngữ thành mệnh đề là một vấn đề khá phức tạp. Nếu ta chỉ chuyển đổi cấu trúc ngữ pháp của câu thì đôi khi chưa đủ mà ta còn cần lưu tâm đến ý nghĩa của danh ngữ đó.

Trong một số trường hợp, người viết giả thiết rằng người đọc đã có các hiểu biết về vấn đề đang được nói đến, chẳng hạn như trong ví dụ (9) sau:

(9) It was the year the final got put back to the third Monday *because of* the weather.

Ví dụ (9) có thể viết lại là:

(9a) It was the year the final got put back to the third Monday *because* the weather was too bad.

Khi chuyển đổi ví dụ (9) thành ví dụ (9a), ta phải biết trước với thời tiết như thế nào thì trận chung kết bị huỷ bỏ. Điều này không được nhắc đến trong văn bản.

Tóm lại, việc chuyển đổi các đơn vị diễn ngôn là một vấn đề khá khó vì nó gắn với việc phân tích ngữ nghĩa ở mức sâu. Tuy nhiên, nếu giải quyết được vấn đề này, chúng ta sẽ xây dựng được hệ thống sinh văn bản mạnh hơn và linh động hơn. Chương trình thử nghiệm của chúng tôi hiện nay đã có thể chuyển đổi các danh ngữ đơn giản thành mệnh đề/câu đơn. Việc phân tích ngữ nghĩa ở mức sâu vẫn đang còn là một thách thức lớn. Vấn đề này sẽ được tiếp tục nghiên cứu trong thời gian tới.

IV.3 Lựa chọn các ràng buộc trong việc sinh văn bản

Như đã nói ở phần trên, chúng ta có nhiều cách để sinh văn bản mới từ một văn bản cho trước. Thông qua các ví dụ ta thấy việc kết hợp các phương pháp thường đưa ra kết quả tốt hơn việc áp dụng một phương pháp đơn lẻ. Bên cạnh việc kết hợp các phương pháp nói trên, hệ thống sinh văn bản còn được điều khiển bởi các ràng buộc cứng và các ràng buộc mềm. Các ràng buộc cứng được sử dụng để đảm bảo không có các văn bản sinh dị thường như văn bản sinh (6d) từ ví dụ (6).

(6d) *Since* Elixir gives quick results doctors recommend it, and it has few side effects.

Các ràng buộc cứng được lựa chọn thông qua giao diện người sử dụng. Tất cả các đầu ra của hệ thống đều phải thoả mãn các ràng buộc này. Các ví dụ về ràng buộc cứng là:

- Cho phép sử dụng các gạch đầu dòng để biểu diễn quan hệ diễn ngôn chính-phụ (Câu hỏi có/không)
- Cho phép sử dụng các gạch đầu dòng để biểu diễn quan hệ diễn ngôn chính-chính (Câu hỏi có/không)
- Cho phép sử dụng từ nối để bắt đầu một thành phần của danh sách (Câu hỏi có/không)

Các ràng buộc mềm cho phép đánh giá mức độ trôi chảy của văn bản sinh thông qua trọng số. Các điều kiện của ràng buộc mềm có thể bị vi phạm nhưng khi đó trọng số của văn bản sẽ giảm. Các ví dụ về ràng buộc mềm là:

- Tránh các đoạn chỉ có một câu đơn. Ràng buộc này sẽ giảm trọng số của giải pháp trong đó hai mệnh đề của câu được chuyển thành hai đoạn riêng biệt.
- Tránh các từ tham chiếu rời rạc (từ tham chiếu không tham chiếu đến đối tượng nào cả).
- Tránh sử dụng câu bị động.
- Tránh các câu phức.

Văn bản đầu ra tốt nhất phụ thuộc vào yêu cầu của người sử dụng và do người sử dụng lựa chọn.

V. Một số kết quả đạt được

Hiện nay, chúng tôi đã xây dựng một hệ thống thử nghiệm dựa trên các kết quả nghiên cứu trên. Vì vấn đề xác định và chuyển đổi các từ đồng tham chiếu khá phức tạp, chúng tôi tạm thời chưa cài đặt modul này. Việc chuyển đổi các đơn vị diễn ngôn được thực hiện ở mức: chuyển danh ngữ thành mệnh

đề hoặc câu đơn; chuyển câu chủ động thành bị động và ngược lại. Việc chuyển đổi các đơn vị diễn ngôn dựa trên việc biến đổi các danh từ, động từ chính trong câu chưa được xét đến trong hệ thống hiện tại. Mặc dù vậy, hệ thống đã có thể đưa ra khá nhiều khả năng chuyển đổi văn bản. Chúng tôi sẽ tiếp tục nghiên cứu hoàn thiện tiếp hệ thống trong thời gian tới. Trong phần này, chúng tôi sẽ giới thiệu một số thử nghiệm được thực hiện trên hệ thống đã xây dựng.

V.1 Thử nghiệm 1

Văn bản vào:

```
Although one of the main ingredients is penicillin, the medicine has no significant side-effects. However, some people might suffer a mild allergic reaction.
```

Cấu trúc diễn ngôn của văn bản vào:

```
<RhetRep relation=concession>
    <SemRep syncat=clause prop="some people might suffer a mild allergic reaction"/>
    <RhetRep relation=concession>
        <SemRep syncat=clause prop="the medicine has no significant side-effects"/>
        <SemRep syncat=clause prop="one of the main ingredients is penicillin"/>
    </RhetRep>
</RhetRep>
```

Các văn bản đầu ra:

A. Although one of the main ingredients is penicillin, the medicine has no significant side-effects. However, some people might suffer a mild allergic reaction.

B. The medicine has no significant side-effects although one of the main ingredients is penicillin. However, some people might suffer a mild allergic reaction.

C. One of the main ingredients is penicillin. However, the medicine has no significant side-effects. But some people might suffer a mild allergic reaction.

Văn bản đầu ra A cho thấy hệ thống có thể dựng lại nguyên dạng văn bản đầu vào từ cấu trúc diễn ngôn của nó. Các đầu ra còn lại có thể hiểu được và diễn tả được đúng nội dung của văn bản đầu vào.

V.2 Thử nghiệm 2

Văn bản vào:

```
Press a tablet from a sachet. Then eat the tablet crushed with food, or swallow it with a glass of water.
```

Cấu trúc diễn ngôn của văn bản vào:

```
<RhetRep relation=sequence>
    <SemRep syncat=clause prop="Press a tablet from a sachet"/>
    <RhetRep relation=disjunction>
        <SemRep syncat=clause prop="eat the tablet crushed with food"/>
        <SemRep syncat=clause prop="swallow it with a glass of water"/>
    </RhetRep>
</RhetRep>
```

Các văn bản đầu ra:

A.
- Press a tablet from a sachet.
- Eat the tablet crushed with food or swallow it with a glass of water.

B. Press a tablet from a sachet; then,
- eat the tablet crushed with food
- or swallow it with a glass of water.

C. Press a tablet from a sachet. Then, eat the tablet crushed with food or swallow it with a glass of water.

Trong thử nghiệm này, văn bản đầu ra sáng sủa và dễ hiểu hơn văn bản đầu vào. Với văn bản đầu ra C, ta thấy hệ thống có thể dựng lại gần như nguyên dạng văn bản đầu vào từ cấu trúc diễn ngôn của nó.

V.3 Thử nghiệm 3

Văn bản vào:

```
Although Jim likes tennis, Chinese food and travel to exotic places, he is not usually adventurous.
```

Cấu trúc diễn ngôn của văn bản vào:

```
<RhetRep relation=concession>
    <SemRep syncat=clause prop="he is not usually adventurous"/>
    <SemRep syncat=clause prop="Jim likes tennis , Chinese food and travel to exotic places"/>
</RhetRep>
```

Các văn bản đầu ra:

A. Jim likes tennis , Chinese food and travel to exotic places. However, he is not usually adventurous.

B. Although Jim likes tennis , Chinese food and travel to exotic places, he is not usually adventurous.

C. He is not usually adventurous although Jim likes tennis , Chinese food and travel to exotic places.

Văn bản đầu ra C của ví dụ trên không mạch lạc do đại từ "*he*" được sử dụng trước khi danh từ riêng "*Jim*" được đề cập. Văn bản này sẽ đúng nếu "*he*" được thay bằng "*Jim*" và "*Jim*" được thay bằng "*he*". Vấn đề này sẽ được giải quyết nếu bộ phân tích các từ đồng tham chiếu được cài đặt.

VI. Kết luận và hướng phát triển

Trong bài này, chúng tôi đã giới thiệu một cách tiếp cận trong việc sinh các cách phát biểu khác nhau ứng với một văn bản cho trước. Việc này được thực hiện thông qua việc biến đổi cấu trúc văn bản và phát biểu lại các mệnh đề trong câu. Các ràng buộc cứng và mềm được sử dụng để hạn chế việc sinh các văn bản dị thường và đánh giá độ trôi chảy của văn bản đầu ra. Bước đầu, chúng tôi đã xây dựng một hệ thống thử nghiệm dựa trên một số ý tưởng đã đề xuất. Hệ thống có thể nhận một văn bản vào và sinh nhiều cách phát biểu tương đương. Kết quả thử nghiệm cho thấy việc cài đặt tiếp các đề xuất còn lại là rất cần thiết để tăng độ hoàn thiện về mặt ngữ nghĩa và tính trôi chảy của văn bản.

Acknowledgment

Dr. Le Thanh Huong gratefully acknowledges the receipt of a grant from the Flemish Interuniversity Council for University Development Cooperation (VLIR UOS) which enabled the research team to carry out this work

Tài liệu tham khảo

[1] Barzilay, R. and McKeown, K. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, 2001.

[2] Charniak, E. *A Maximum-Entropy-Inspired Parser*. Proceedings of NAACL-2000.

[3] Inui, K. and Nogami, M. A paraphrase-based exploration of cohesiveness criteria. In *Proceedings of the 8th European Workshopon Natural Language Generation (EWNLG),* 2001.

[4] Kozlowski, R., McCoy, K. F. and Vijay-Shanker, K. Generation of singlesentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *Proceedings of the Second International Workshop on Paraphrasing*, 2003.

[5] Le-Thanh, H. *Investigation into an Approach to Automatic Text Summarisation*, Ph.D. dissertation, Middlesex University, U.K. 2004. (Bản lưu tại Thư viện Quốc gia Việt Nam).

[6] Le-Thanh, H., Abeysinghe, G., and Huyck, C. *Generating Discourse Structures for Written Texts*. In Proceedings of the International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, Aug. 23-27, 2004.

[7] Mann, W.C. and Thompson, S.A. Rhetorical Structure Theory: Toward a Functional Theory of Text Organisation. *Text*, vol. 8(3), 1988, pp.243-281.

[8] Power, R., Scott, D. and Bouayad-Agha, N. Document Structure, *Computational Linguistics, 29(4), 2003,* pp. 211 - 260.

Thông tin về tác giả:

Họ tên: Lê Thanh Hương

Ngày sinh: 12/01/1976

Nơi sinh: Hà Nội

Địa chỉ liên lạc: Phòng 325 C1 trường ĐHBK Hà Nội. Số 1 Đại Cồ Việt

Điện thoại liên hệ: 0904674102

Email: huonglt@it-hut.edu.vn

Nơi công tác: Bộ môn Các Hệ thống thông tin, Khoa Công nghệ thông tin, trường Đại học Bách khoa Hà nội.

Quá trình công tác:

- nhận bằng Tiến sĩ CNTT, trường Đại học tổng hợp Middlesex, Vương quốc Anh năm 2004 (về xử lý ngôn ngữ tự nhiên).
- nhận bằng Thạc sĩ CNTT, Trường Đại học tự do Brussels, Vương quốc Bỉ năm 2001 (về Robotics)
- nhận bằng Thạc sĩ CNTT, trường ĐHBK Hà Nội năm 1999 (về xử lý ngôn ngữ tự nhiên)
- tốt nghiệp Đại học ngành Tin học, trường ĐHBK Hà Nội năm 1997.

Hiện đang giảng dạy tại Khoa Công nghệ Thông tin, trường Đại học Bách khoa Hà Nội.

Lĩnh vực nghiên cứu: Xử lý ngôn ngữ tự nhiên, khai phá dữ liệu và văn bản, các kỹ thuật học máy.

Tóm tắt bài báo bằng tiếng Việt:

Bài này đề xuất việc xây dựng một hệ thống có khả năng tự động sinh các cách phát biểu tương đương của đoạn văn bản. Ba công việc chính trong hệ thống này là: (1) xây dựng cấu trúc diễn ngôn của văn bản; (2) xử lý vấn đề đồng tham chiếu (tuỳ chọn); và (3) phát biểu lại văn bản. Hệ thống được cài đặt thử nghiệm trước tiên cho tiếng Anh, sử dụng hai modul chính ứng với các công việc (1) và (3). Kết quả thử nghiệm cho kết quả tương đối khả quan. Kết quả thử nghiệm cũng cho thấy rằng nếu công việc (2) được cài đặt thì hệ thống sẽ cho kết quả tốt hơn. Chúng tôi sẽ nghiên cứu và cài đặt modul ứng với công việc này trong tương lai.