ĐẠI HỌC QUỐC GIA TP HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

TRẦN TRUNG

TÓM TẮT ĐOẠN VĂN BẢN TIẾNG VIỆT DỰA TRÊN CÁCH TIẾP CẬN TẠO SINH

Chuyên ngành: **Khoa học máy tính**
Mã số: **62 48 01 01**

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

TP. HỒ CHÍ MINH – Năm 2019

Công trình được hoàn thành tại: **Khoa Khoa học máy tính, Trường Đại học Công nghệ thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh**.

Người hướng dẫn khoa học:

1. **PGS. TS. NGUYỄN TUẤN ĐĂNG**
2. **PGS. TS. PHẠM HỮU ĐỨC**

Phản biện 1: **PGS. TS. NGUYỄN LÊ MINH**
Phản biện 2: **TS. ĐẶNG TRƯỜNG SƠN**

Luận án sẽ/đã được bảo vệ trước
Hội đồng chấm luận án cấp Trường tại:
Trường Đại học Công nghệ thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh
vào lúc 08 giờ 30 ngày 08 tháng 01 năm 2020

Có thể tìm hiểu luận án tại:
- Thư viện Quốc gia Việt Nam
- Thư viện Trường Đại học Công nghệ Thông tin

MỤC LỤC

MỞ ĐẦU

Đặt vấn đề và lý do lựa chọn đề tài

Tạo sinh ra được một văn bản tóm tắt mạch lạc và tự nhiên từ một văn bản cho trước là mục tiêu quan trọng nhất của lĩnh vực Tóm tắt văn bản. Các nghiên cứu trong cả hai hướng tiếp cận chính hiện nay là *tóm tắt trích xuất* ("extractive") và *trừu tượng* ("abstractive") [Das và Martins 2007; Fattah và Ren 2008; Jezek và Steinberger 2008; Jones 1999, 2007; Lloret 2008; Mani và Maybury 1999; Mani 2001b] đều tập trung vào vấn đề cải thiện chất lượng của văn bản tóm tắt. Thách thức đầu tiên trong việc nâng chất lượng văn bản tóm tắt là đảm bảo được mối liên hệ giữa từng yếu tố hồi chỉ với yếu tố tiền ngữ tương ứng. Kế tiếp, đó là vấn đề tạo dựng được một mô hình biểu diễn ngữ nghĩa cho văn bản gốc để thực hiện việc tóm tắt trên mô hình này. Nhiệm vụ đầu tiên trở nên khó khăn hơn trong tiếng Việt do có rất nhiều dạng yếu tố hồi chỉ khác nhau, đặc biệt là những đại từ hồi chỉ. Hầu hết những hệ thống tóm tắt theo hướng trích xuất đều không bao gồm các cơ chế xử lý yếu tố hồi chỉ [Das và Martins 2007; Fattah và Ren 2008; Jezek và Steinberger 2008; Jones 1999, 2007; Lloret 2008; Mani và Maybury 1999; Mani 2001b] vì mục tiêu chính của những hệ thống này là gom nhóm những câu hoặc những cụm từ có điểm đánh giá cao nhất để tạo thành tóm tắt. Bên cạnh đó, các nghiên cứu theo hướng trừu tượng [Das và Martins 2007; Kasture và cộng sự 2014; Khan và Salim 2014] cũng nỗ lực tìm kiếm các giải pháp để tạo dựng mô hình biểu diễn ngữ nghĩa cho văn bản gốc nhưng chưa hoàn chỉnh. Cuối cùng là cơ chế tạo sinh

câu và văn bản tóm tắt. Việc đề xuất được những cơ chế tạo sinh câu và đoạn văn bản hoàn chỉnh vẫn đang là thách thức không nhỏ ngay cả trong lĩnh vực Tạo sinh ngôn ngữ tự nhiên. Một điểm quan trọng nữa là văn bản tóm tắt cần có được tính đúng đắn ngữ pháp trong khi đảm bảo về mặt ngữ nghĩa.

Dựa trên những khảo sát bên trên về Tóm tắt văn bản, luận án xác định theo hướng tiếp cận tóm tắt trừu tượng với sự kết hợp giữa những kỹ thuật về khoa học máy tính như Hiểu và biểu diễn văn bản, Tạo sinh ngôn ngữ tự nhiên với kiến thức ngôn ngữ học phù hợp.

Mục tiêu và nội dung nghiên cứu

Mục tiêu của luận án là đề xuất giải pháp tạo sinh câu và đoạn văn bản tóm tắt nhằm tóm tắt nội dung thông tin của đoạn văn bản tiếng Việt cho trước. Để thực hiện mục tiêu này, luận án đề ra những nội dung cụ thể:

1. Đề xuất các phương pháp tạo dựng *Cấu trúc biểu diễn ngữ nghĩa trừu tượng* (CT-BD-NN-TT) cho đoạn văn bản tiếng Việt đầu vào. Nội dung này bao gồm việc giải quyết hai bài toán con: (a) Tạo dựng *Cấu trúc biểu diễn cấp độ bề mặt* (CT-BD-CĐ-BM) cho đoạn văn bản tiếng Việt đầu vào; (b) Chuyển đổi CT-BD-CĐ-BM về CT-BD-NN-TT.
2. Đề xuất các phương pháp tạo sinh câu và đoạn văn bản tóm tắt dựa trên CT-BD-NN-TT.

Phạm vi và đối tượng nghiên cứu

Phạm vi nghiên cứu

Đề xuất mô hình giải pháp tạo sinh câu và đoạn văn bản tóm tắt nhằm tóm tắt nội dung thông tin đoạn văn bản tiếng Việt cho trước theo hướng tiếp cận tóm tắt trừu tượng, với sự kết hợp những kỹ thuật tạo sinh ngôn ngữ tự nhiên và kiến thức ngôn ngữ học trong Ngữ pháp chức năng [Cao 2006; Halliday và Matthiessen 2004].

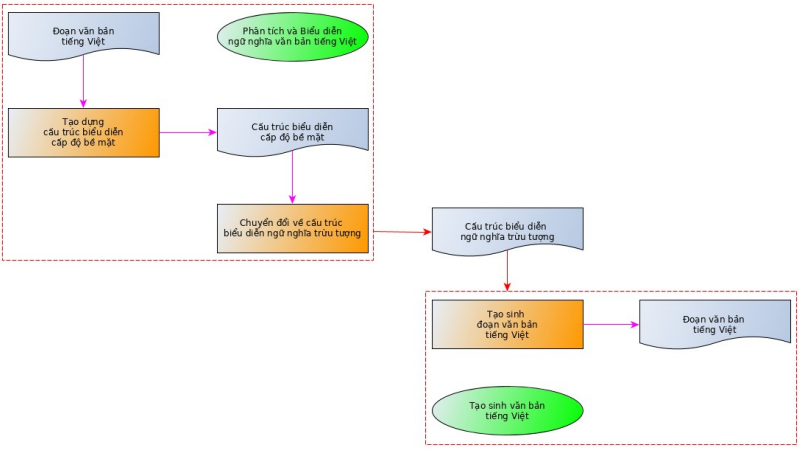
Đối tượng nghiên cứu

Đối tượng nghiên cứu thứ nhất là những cặp câu tiếng Việt có cấu trúc đơn giản. Mối quan hệ giữa hai câu được thể hiện bởi một hoặc hai đại từ hồi chỉ trong câu thứ hai. *Đối tượng nghiên cứu thứ hai* là những đoạn văn bản ngắn gồm vài câu tiếng Việt có cấu trúc đơn giản. Các câu có sự xuất hiện của một hoặc hai đại từ hồi chỉ. *Đối tượng nghiên cứu thứ ba* là những đoạn văn bản tiếng Việt gồm từ 2 đến 5 câu ở thể trần thuật. Từng câu có số lượng không quá 25 từ vựng tiếng Việt. Trong từng đoạn văn bản có sự xuất hiện của các yếu tố hồi chỉ. *Đối tượng nghiên cứu thứ tư* là những dạng yếu tố hồi chỉ trong tiếng Việt, dựa trên sự phân loại trong Ngữ pháp chức năng [Cao 2006].

Phương thức tiếp cận

Để thực hiện những nội dung nghiên cứu được xác định bên trên, phương thức tiếp cận của luận án như sau. *Giai đoạn 1*, luận án phân tích các đối tượng nghiên cứu là những dạng đoạn văn bản tiếng Việt khác nhau và những yếu tố hồi chỉ xuất hiện trong từng

đoạn văn bản. *Giai đoạn 2*, luận án đề xuất các quy tắc và giải thuật để tạo dựng CT-BD-CĐ-BM giúp biểu diễn toàn bộ nội dung thông tin của đoạn văn bản đầu vào. *Giai đoạn 3*, luận án đề xuất các quy tắc và giải thuật để chuyển đổi CT-BD-CĐ-BM về CT-BD-NN-TT giúp biểu diễn nội dung thông tin chính của đoạn. *Giai đoạn 4*, luận án đề xuất các cơ chế tạo sinh các câu và đoạn văn bản tóm tắt từ CT-BD-NN-TT. Phương thức tiếp cận của luận án được thể hiện qua mô hình giải pháp trong Hình 0.1.



Hình 0.1. Mô hình giải pháp tổng thể của luận án.

Đóng góp khoa học của luận án

Luận án có những đóng góp khoa học chính:

1. Đề xuất mô hình giải pháp tạo sinh đoạn văn bản tóm tắt.
2. Đề xuất những phương pháp tạo dựng CT-BD-NN-TT. *Phương pháp sử dụng cấu trúc ngữ đoạn bề mặt* gồm hai giai đoạn: (i) tạo dựng một CT-BD-CĐ-BM được gọi là cấu

trúc ngữ đoạn bề mặt và chuyển đổi thành đoạn văn bản bao gồm các câu tiếng Việt có cấu trúc đơn giản; (ii) xác định mối liên hệ giữa từng đại từ hồi chỉ với yếu tố tiền ngữ tương ứng đồng thời tạo dựng CT-BD-NN-TT. *Phương pháp sử dụng cấu trúc đồ thị ngữ đoạn được gán nhãn* gồm hai giai đoạn: (i) tạo dựng một CT-BD-CĐ-BM được gọi là cấu trúc đồ thị ngữ đoạn được gán nhãn, đồng thời xác định mối liên hệ giữa từng yếu tố hồi chỉ với yếu tố tiền ngữ tương ứng; (ii) chuyển đổi CT-BD-CĐ-BM thành CT-BD-NN-TT.

3. Đề xuất những phương pháp xác định yếu tố tiền ngữ cho những dạng yếu tố hồi chỉ khác nhau trong đoạn văn bản tiếng Việt nguồn.
4. Đề xuất những phương pháp tạo sinh câu và đoạn văn bản tóm tắt tiếng Việt dựa trên một dạng CT-BD-NN-TT.

Bố cục của luận án

Luận án được bố cục gồm các Chương, Mục như sau. *Chương Mở đầu* giới thiệu: vấn đề nghiên cứu; mục tiêu nghiên cứu; phạm vi và đối tượng nghiên cứu; phương pháp nghiên cứu và cách tiếp cận; các đóng góp khoa học của luận án; bố cục của luận án. *Chương 1* trình bày tổng quan về những nghiên cứu liên quan trong lĩnh vực Tóm tắt văn bản. *Chương 2* trình bày những kiến thức nền tảng trong Khoa học máy tính là cơ sở để đề xuất những phương pháp, cơ chế xử lý trong luận án. *Chương 3* trình bày các phương pháp được đề xuất để hiện thực thành phần Phân tích và Biểu diễn ngữ nghĩa văn bản tiếng Việt trong mô hình giải pháp ở Hình 0.1. *Chương 4* trình bày các phương pháp được đề xuất để hiện thực

thành phần Tạo sinh văn bản tiếng Việt trong mô hình giải pháp ở Hình 0.1. *Chương Kết luận và hướng phát triển* trình bày tóm tắt lại những đóng góp chính và hướng phát triển tiếp theo.

CHƯƠNG 1. TỔNG QUAN VỀ TÓM TẮT VĂN BẢN

Dựa trên cách thức xây dựng và tạo sinh văn bản tóm tắt, lĩnh vực tóm tắt văn bản được phân loại thành những hướng tiếp cận: (i) *tóm tắt dựa trên trích xuất* ("extractive summarization"); (ii) *tóm tắt trừu tượng* ("abstractive summarization").

1.1. Hướng tiếp cận tóm tắt dựa trên trích xuất

Với nền tảng là những giải thuật về máy học và trích xuất thông tin, những nghiên cứu theo hướng trích xuất [Das và Martins 2007; Fattah và Ren 2008; Jezek và Steinberger 2008; Jones 1999, 2007; Lloret 2008; Mani và Maybury 1999; Mani 2001b; Saranyamol và Sindhu, 2014] tập hợp những câu được xác định có điểm đánh giá cao nhất để tạo thành tóm tắt. Ý tưởng chính là phân tích thống kê những yếu tố ở cấp độ bề mặt như từ khóa, từ tiêu đề, vị trí hay độ dài câu. Với việc không cần hiểu sâu ngữ nghĩa ban đầu, những phương pháp theo hướng này trở nên ít phức tạp và có thể áp dụng cho nhiều dạng văn bản khác nhau. Tuy nhiên, vấn đề còn tồn tại của hướng này là đảm bảo tính mạch lạc trong văn bản tóm tắt. *Lý do thứ nhất* là những câu được trích xuất không hoàn toàn kết nối dựa theo luồng dữ liệu ban đầu. *Lý do thứ hai* là những mối quan hệ giữa các yếu tố hồi chỉ với yếu tố tiền ngữ tương ứng có thể bị phá vỡ. Một vấn đề khác cần được xem xét sâu hơn là đảm bảo rằng tất cả những câu có điểm đánh giá cao nhất thì chứa đựng những thông tin quan trọng.

1.2. Hướng tiếp cận tóm tắt trừu tượng

Trong hướng tiếp cận trừu tượng, nhiều phương pháp [Kasture và cộng sự 2014; Khan và Salim 2014; Lloret 2008] được đề xuất với ý tưởng chính là chuyển đổi văn bản nguồn thành một mô hình biểu diễn, xác định ngữ nghĩa chính và tạo sinh một tóm tắt từ mô hình này. Ý tưởng này dẫn đến sự phát triển của những hướng tiếp cận thứ cấp: *dựa trên cấu trúc* ("structure-based") [Harabagiu và Lacatusu 2002; Lee và cộng sự 2005; Tanaka và cộng sự 2009; Genest và Lapalme 2012] trong đó các tác giả tập trung vào việc biểu diễn ngữ cảnh của văn bản đầu vào trong những dạng cấu trúc khác nhau; *dựa trên ngữ nghĩa* ("semantic-based") với những kỹ thuật trong lĩnh vực *tạo sinh ngôn ngữ tự nhiên* ("natural language generation") để biểu diễn ngữ nghĩa văn bản nguồn và tóm tắt [Greenbacker 2011; Genest và Lapalme 2011; Moawad và Aref 2012]. Một số điểm còn tồn tại cần được nghiên cứu sâu hơn cho hướng tiếp cận này: (a) chưa có một cơ chế hoàn chỉnh để hiểu được chính xác ngữ nghĩa văn bản nguồn; (b) biểu diễn trừu tượng của tóm tắt chưa được hoàn chỉnh theo những kỹ thuật hiện tại trong tạo sinh ngôn ngữ tự nhiên; (c) sự kết hợp những kiến thức trong các lý thuyết ngôn ngữ học.

Những phương pháp mới trong hướng tiếp cận thứ cấp *nén và hợp nhất câu* ("sentence fusion and compression") [Barzilay và McKeown 2005; Krahmer và cộng sự 2008; Filippova và Strube 2008a; Filippova 2010; Boudin và Morin 2013] cố gắng vượt qua những vấn đề trên. Những nghiên cứu này nén những câu liên quan và hợp nhất thông tin.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. Giới thiệu

Chương này trình bày những kiến thức nền tảng trong Khoa học máy tính, là cơ sở để đề xuất những phương pháp, cơ chế xử lý trong luận án.

2.2. Phân tích và biểu diễn ngữ nghĩa

2.2.1. Lý thuyết biểu diễn diễn ngôn

Lý thuyết biểu diễn diễn ngôn ("Discourse Representation Theory" – DRT) được giới thiệu trong [Blackburn và Bos 1999; Covington và cộng sự 1988, 1989; Kamp 1981] với ý tưởng cơ bản: một đoạn văn bản ngôn ngữ tự nhiên sẽ được biểu diễn trong một ngữ cảnh của một *cấu trúc biểu diễn diễn ngôn* (CT-BD-DN) ("Discourse Representation Structures" – DRS). Một CT-BD-DN bao gồm một cặp danh sách có thứ tự: (i) U là một danh sách những đánh dấu văn bản, hay còn có thể hiểu là những đối tượng của văn bản; (ii) Con là danh sách những điều kiện, hay có thể hiểu là những vị từ hay công thức mà những đối tượng trong U phải thỏa. **Ví dụ 2.1** "*Nhân thông minh. Nó viết chương trình.*" có CT-BD-DN như sau:

```
[1, 2]
```

```
nhân(1), thông_minh(1), chương_trình(2), viết(1,2)
```

Hình 2.1. CT-BD-DN của "*Nhân thông minh. Nó viết chương trình.*"

2.2.2. Ngữ pháp dựa trên sự hợp nhất

Ngữ pháp dựa trên sự hợp nhất ("Unification-based Grammar" – UBG) được giới thiệu trong [Covington 2007; Shieber 2003] là một hình thức trong đó những lý thuyết về ngữ pháp có thể được biểu diễn, với vai trò nổi bật của việc hợp nhất những cấu trúc đặc điểm. Trong phân tích cấu trúc cú pháp của câu, ở từng ngữ đoạn hoặc từ vựng, có thể mô tả thêm cấu trúc đặc điểm của ngữ đoạn hay từ vựng này. Những thông tin đặc điểm này có thể truyền lên xuống giữa các ngữ đoạn, và tạo nên cấu trúc đặc điểm từ những thông tin được truyền đến.

2.3. Tạo sinh ngôn ngữ tự nhiên

Tạo sinh ngôn ngữ tự nhiên ("Natural Language Generation" – NLG) là một lĩnh vực con của *Ngôn ngữ học máy tính* ("Computational Linguistic") mà tập trung vào việc tạo sinh những văn bản có thể hiểu được bằng ngôn ngữ của con người [Reiter và Dale 1997a, 1997b]. Thông thường, đầu vào của một hệ thống NLG là một dạng biểu diễn thông tin phi ngôn ngữ nào đó. Hệ thống NLG sẽ áp dụng kiến thức về ngôn ngữ và miền ứng dụng để tạo sinh văn bản hướng con người có chất lượng và tự nhiên.

Kiến trúc truyền thống tổng quát của một hệ thống NLG bao gồm những mô-đun chính [Reiter và Dale 1997a, 1997b]. Mô-đun *Chuẩn bị văn bản* ("Document Planning") chịu trách nhiệm xác định (a) thông tin nào nên được hiển thị trong văn bản đầu ra và (b) làm thế nào những đoạn nội dung khác nhau nên được gom nhóm lại và liên hệ trong những mẫu tu từ. Mô-đun *Vi chuẩn bị*

(“Microplanning”) chịu trách nhiệm xác định (a) những từ vựng hay ngữ đoạn nên được sử dụng để biểu đạt những thông tin được lựa chọn, (b) những dạng biểu hiện nên được sử dụng để liên hệ đến những thực thể, và (c) làm thế nào những cấu trúc tu từ được tạo dựng có thể ánh xạ vào những cấu trúc ngôn ngữ học. Mô-đun *Hiện thực hóa* (“Realisation”) chịu trách nhiệm chuyển đổi (a) biểu diễn trừu tượng thành văn bản thực và (b) cấu trúc trừu tượng thành những biểu tượng đánh dấu dễ hiểu.

CHƯƠNG 3. PHÂN TÍCH VÀ BIỂU DIỄN NGỮ NGHĨA VĂN BẢN TIẾNG VIỆT

3.1. Giới thiệu

Chương này trình bày cơ chế thực hiện thành phần Phân tích và Biểu diễn ngữ nghĩa văn bản tiếng Việt trong Hình 0.1.

3.2. Phương pháp sử dụng cấu trúc ngữ đoạn bề mặt

3.2.1. Tạo dựng cấu trúc biểu diễn cấp độ bề mặt

Trước tiên, luận án thực hiện tạo dựng cấu trúc bề mặt của từng câu tiếng Việt đầu vào, trong đó từng từ vựng và ngữ đoạn được phân tách và gán nhãn phù hợp với mục tiêu nghiên cứu của luận án. Luận án xây dựng tập nhãn F-POS Tagset, định nghĩa nhãn ngữ đoạn OP để gán nhãn bề mặt đối tượng, nhãn ngữ đoạn FP để gán nhãn những ngữ đoạn chức năng mà có chứa nhãn từ vựng thuộc các từ loại "hành động", "quá trình" hay "trạng thái", nhãn từ vựng cho tất cả các loại đối tượng được biểu diễn bởi danh từ riêng hay danh từ chung trong câu, các loại hành động, quá trình hay trạng thái được biểu diễn bởi động từ hay tính từ trong câu, các loại đại từ hỏi chỉ. **Ví dụ 3.1** cấu trúc bề mặt cho "*Người lính đến bên cái bàn lấy cây dù.*":

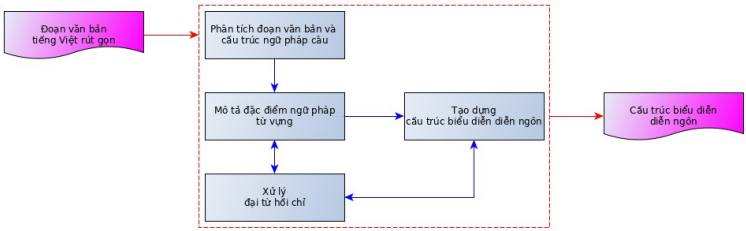
```
[OP người_lính/HUMC] [FP đến_bên/ITMO] [OP cái_bàn/
NANIC] [FP lấy/TTPO] [OP cây_dù/NANIC] ./..
```

Luận án thiết lập những quy tắc chuyển đổi F-ConvRules dựa trên kinh nghiệm thực tế khi sử dụng ngôn ngữ Tiếng Việt trong giao tiếp thông thường. *Quy tắc F-Conv-1*. Chỉ lựa chọn những ngữ đoạn thuộc: `[OP]`; `[FP]`. *Quy tắc F-Conv-2*. Lựa chọn ngữ đoạn `[OP]`

thứ nhất nếu có nhiều ngữ đoạn [OP] liên tiếp cùng giữ vai trò chủ thể hoặc khách thể. *Quy tắc F-Conv-3*. Lựa chọn ngữ đoạn [FP] cuối cùng nếu: (a) Những ngữ đoạn [FP] này liên tiếp; (b) Tất cả ngữ đoạn [FP] đều chứa nhãn từ vựng thuộc từ loại "hành động" / "quá trình" / "trạng thái"; (c) Từng cặp [FP] không được phân tách bởi dấu "," hay một từ nối. *Quy tắc F-Conv-4*. Tách ra và lựa chọn tất cả ngữ đoạn [FP] nếu: (a) Những ngữ đoạn [FP] này là liên tiếp; (b) Tất cả ngữ đoạn [FP] đều chứa nhãn từ vựng thuộc từ loại "hành động"; (c) Từng cặp [FP] được phân tách bởi dấu "," hay một từ nối. *Quy tắc F-Conv-5*. Tách ra và lựa chọn tất cả ngữ đoạn [FP] nếu: (a) Những ngữ đoạn [FP] này là liên tiếp; (b) Tất cả ngữ đoạn [FP] đều chứa nhãn từ vựng thuộc từ loại "quá trình" / "trạng thái"; (c) Từng cặp [FP] được phân tách bởi dấu "," hay một từ nối. *Quy tắc F-Conv-6*. Tách ra và lựa chọn tất cả cặp ngữ đoạn [FP] [OP] nếu những cặp ngữ đoạn này liên tiếp. *Quy tắc F-Conv-7*. Tách ra và lựa chọn cả hai ngữ đoạn trong cặp [FP] [FP] nếu: (a) Ngữ đoạn [FP] thứ nhất chứa nhãn từ vựng thuộc từ loại "trạng thái" hay "quá trình"; (b) Ngữ đoạn [FP] thứ hai chứa nhãn từ vựng thuộc từ loại "hành động". *Quy tắc F-Conv-8*. Lựa chọn ngữ đoạn [FP] thứ nhất trong cặp [FP] [FP] nếu: (a) Ngữ đoạn [FP] thứ nhất chứa nhãn từ vựng thuộc từ loại "hành động"; (b) Ngữ đoạn [FP] thứ hai chứa nhãn từ vựng thuộc từ loại "trạng thái" hay "quá trình". *Quy tắc F-Conv-9*. Lựa chọn ngữ đoạn [FP] thứ nhất trong câu cả khi không có ngữ đoạn [OP] giữ vai trò chủ thể.

3.2.2. Tạo dựng cấu trúc biểu diễn diễn ngôn

Sơ đồ luồng xử lý của cơ chế được minh họa trong Hình 3.3.



Hình 3.3. Sơ đồ luồng xử lý cơ chế tạo dựng CT-BD-DN từ đoạn văn bản tiếng Việt rút gọn.

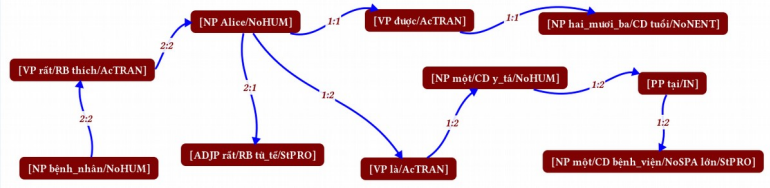
Luận án phân tách rõ ràng thành từng câu và đánh chỉ số vị trí để phân biệt thứ tự các câu trong đoạn văn bản. Sau khi phân tích cấu trúc ngữ đoạn từng câu xuống đến mức từ vựng, sẽ mô tả đặc điểm cú pháp và ngữ nghĩa của từng từ vựng tùy thuộc vào từ loại: đối tượng, hành động, trạng thái, quá trình. Những đặc điểm này được lần lượt thêm vào danh sách U và Con của CT-BD-DN.

Thành phần tìm kiếm yếu tố tiền ngữ tương ứng của từng đại từ hồi chỉ dựa trên những chỉ số riêng biệt và điều kiện đặc điểm của các đối tượng đã có trong danh sách U và danh sách Con của CT-BD-DN. Luận án đề xuất những chiến lược tìm kiếm dựa trên: (a) những ràng buộc là những điều kiện trong đó những ứng viên yếu tố tiền ngữ hay đại từ hồi chỉ phải thỏa mãn; (b) những heuristic với độ ưu tiên cho phép tách biệt những ứng viên yếu tố tiền ngữ và những đại từ hồi chỉ cùng thỏa những ràng buộc.

3.3. Phương pháp sử dụng cấu trúc đồ thị ngữ đoạn được gán nhãn

3.3.1. Tạo dựng cấu trúc biểu diễn cấp độ bề mặt

Đồ thị ngữ đoạn được gán nhãn được tạo dựng bằng cách lần lượt thêm cấu trúc bề mặt của từng câu nguồn. **Ví dụ 3.7** "*Alice được hai mươi ba tuổi. Cô là một y tá tại một bệnh viện lớn. Cô rất tử tế. Bệnh nhân rất thích cô.*" có đồ thị ngữ đoạn được gán nhãn:



Hình 3.4. Cấu trúc đồ thị cho đoạn văn bản trong Ví dụ 3.7.

Luận án đề xuất các chiến lược xử lý yếu tố hồi chỉ dựa trên hai yếu tố. *Ràng buộc* là những đặc điểm của từng từ vựng: ngữ pháp, ngữ nghĩa, ngữ dụng, vai trò ngữ pháp dựa trên vị trí, thói quen giao tiếp của người Việt trong những ngữ cảnh khác nhau. *Heuristic* là những giả định ưu tiên khi xuất hiện nhiều tiền ngữ ứng viên cùng thỏa những đặc điểm về ràng buộc.

3.3.2. Tạo dựng cấu trúc biểu diễn diễn ngôn

Luận án đề xuất giải thuật chuyển đổi về CT-BD-DN:

Giải thuật 3.3. Chuyển đổi về CT-BD-DN.

Đầu vào. G = (V, E). B = Danh sách nhánh; Bi ∈ B = Danh sách nút trong nhánh.

Đầu ra. CT-BD-DN = <U, Con>.

1: for Bi ∈ B do

2: if (CheckRemBranch(Bi)) then RemBranch(Bi);

3: else

4: for Noj ∈ Bi do

5: if (IsMainInfo(Noj)) then

6: if (IsObject(Noj)) then inC ← GenIndex(Noj); U ← U ∪ {inC}; Con ← Con ∪ {CrPredicate(Noj / [inC])};

7: if (IsFunction(Noj)) then

8: noL ← LeftNearesObject(Noj); inL ← Index(noL);

9: if (Type(Noj) == “transitive”) then noR ← RightNearesObject(Noj); inR ← Index(noR);

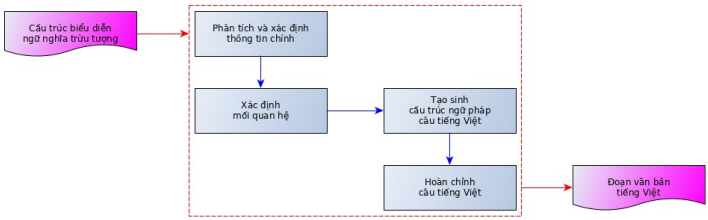
10: Con ← Con ∪ {CrPredicate(Noj / [inL – inR])};

11: else RemNode(Noj);

CHƯƠNG 4. TẠO SINH VĂN BẢN TIẾNG VIỆT

4.1. Giới thiệu

Chương này trình bày cơ chế thực hiện thành phần Tạo sinh văn bản tiếng Việt trong Hình 0.1. Sơ đồ luồng xử lý chung:



Hình 4.1. Sơ đồ kiến trúc cơ chế tạo sinh câu và đoạn văn bản tiếng Việt.

4.2. Từ biểu diễn của các cặp câu có quan hệ hệ quả

Luận án phân nhóm cặp câu dựa theo mối quan hệ: động từ ở câu thứ nhất có độ ưu tiên cao hơn giữ vai trò là hệ quả của động từ ở câu thứ hai. **Nhóm HQ-1**: trạng thái tình trạng – trạng thái tính chất. **Ví dụ 4.1** "*Mai hãnh diện. Cô ấy xinh đẹp.*". **Nhóm HQ-2**: hành động vô tác – trạng thái tính chất. **Ví dụ 4.2** "*Trí tính toán. Anh ta gian giảo.*". **Nhóm HQ-3**: hành động chuyển tác – trạng thái tính chất. Nhóm này được chia thành hai nhóm con. **Nhóm HQ-3.1**. Đại từ trong câu thứ hai đứng một mình chỉ đến đối tượng giữ vai trò chủ thể trong câu thứ nhất. **Ví dụ 4.3** "*Lan học võ. Cô ấy mạnh mẽ.*". **Nhóm HQ-3.2**. Đại từ trong câu thứ hai đứng cùng tính từ chỉ thị ["ta" / "ấy" / "này"] chỉ đến đối tượng giữ vai trò khách thể trong câu thứ nhất. **Ví dụ 4.4** "*Nghĩa ghét Tín. Anh ta keo kiệt.*". **Nhóm HQ-4**: trạng thái tình trạng – hành động vô tác. **Ví dụ 4.5** "*Lễ thư thái. Anh*

khiêu vũ.". **Nhóm HQ-5**: trạng thái tình trạng – hành động chuyển tác. Nhóm này được chia thành hai nhóm con. **Nhóm HQ-5.1**. Đại từ trong câu thứ hai giữ vai trò chủ thể. **Ví dụ 4.6** "*Nhân thoải mái. Anh mặc đồ thể thao.*". **Nhóm HQ-5.2**. Đại từ trong câu thứ hai giữ vai trò khách thể. **Ví dụ 4.7** "*Trúc hạnh phúc. Tín cầu hôn cô.*". **Nhóm HQ-6**: hành động chuyển tác – quá trình chuyển thái. **Ví dụ 4.8** "*Lễ vá lốp xe. Nó bị thủng.*". **Nhóm HQ-7**: hành động chuyển tác – quá trình chuyển vị. **Ví dụ 4.9** "*Cúc nhặt chiếc bình. Nó bị rơi.*". **Nhóm HQ-8**: hành động chuyển tác – quá trình tác động. **Ví dụ 4.10** "*Nghĩa sửa angten. Sét đánh nó.*".

Đối với từng nhóm cặp câu, luận án phân tích CT-BD-DN và xác định hai dạng mối quan hệ: (a) mối quan hệ nội tại giữa đối tượng và hành động / trạng thái / quá trình bên trong một câu; (b) mối quan hệ liên câu giữa hai đối tượng hay giữa đối tượng với hành động / trạng thái / quá trình ở hai câu khác nhau. Xét CT-BD-DN của cặp câu trong Ví dụ 4.10 thuộc nhóm HQ-8:

`[1,2,3]`
`nghĩa(1), angten(2), sửa(1,2), sét(3), đánh(3,2)`

Hình 4.2. CT-BD-DN của "*Nghĩa sửa angten. Sét đánh nó.*".

▶ Diễn giải. Quan hệ nội tại: (a) Câu thứ nhất: `[1] ↦ sửa(1,2) ↦ [2]`; (b) Câu thứ hai: `[3] ↦ đánh(3,2) ↦ [2]`. Quan hệ liên câu: `sửa(1,2) ↦ <QH-HQ> ↦ đánh(3,2).`

Giải thuật 4.1. Tạo sinh cấu trúc cú pháp của câu tóm tắt tiếng Việt.

Đầu vào. Các cấu trúc mối quan hệ nội tại và liên câu.

Đầu ra. Cấu trúc câu tóm tắt tiếng Việt.

1: Bước 1: Với mối quan hệ nội tại thứ nhất

2: Nếu (nhóm 1 / 2 / 4 / 5.1 / 5.2): Thêm [1] ↦ vị_từ_thứ_nhất(1);

3: Ngược lại Nếu (nhóm 3.1 / 3.2 / 6 / 7 / 8): Thêm [1] ↦ vị_từ_thứ_nhất(1,2);

4: Bước 2: Với mối quan hệ liên câu: Thêm: <QH-HQ>;

5: Bước 3: Với mối quan hệ nội tại thứ hai

6: Nếu (nhóm 1 / 2 / 3.1 / 4): Thêm vị_từ_thứ_hai(1);

7: Nếu (nhóm 3.2): Thêm [2] ↦ vị_từ_thứ_hai(2);

8: Nếu (nhóm 5.1): Thêm vị_từ_thứ_hai(1,2) ↦ [2];

9: Nếu (nhóm 5.2) là Bị động cách: Thêm vị_từ_thứ_hai(2,1) ↤ [2];

10: Nếu (nhóm 6 / 7): Thêm [2] ↦ vị_từ_thứ_hai(2);

11: Nếu (nhóm 8) là Bị động cách: Thêm [2] ↤ vị_từ_thứ_hai(3,2) ↤ [3];

Thực thi Giải thuật 4.1 cho CT-BD-DN trong Hình 4.2: `{[1] ↦ sửa(1,2) ↦ [2]} + <QH-HQ> + {[2] ↤ đánh(3,2) ↤ [3]}.`

4.3. Từ biểu diễn của các cặp câu chỉ quá trình

Dựa trên giả thiết về thứ tự thời gian xảy ra các quá trình, luận án phân những cặp câu được xem xét thành ba nhóm lớn. **Nhóm QT-1**. Quá trình ở câu thứ nhất xảy ra trước quá trình ở câu thứ hai. Luận án giả định rằng quá trình ở câu thứ nhất là nguyên nhân của quá trình ở câu thứ hai. **Ví dụ 4.11** "*Sét đánh cành cây. Nó bị gãy.*". **Nhóm QT-2**. Quá trình ở câu thứ nhất xảy ra sau quá trình ở câu thứ hai. Luận án giả định rằng quá trình ở câu thứ nhất là hệ quả của quá trình ở câu thứ hai. **Ví dụ 4.12** "*Cái bình bị nứt. Nó bị rơi.*". **Nhóm**

QT-3. Quá trình ở câu thứ nhất xảy ra đồng thời quá trình ở câu thứ hai. **Ví dụ 4.13** "*Chiếc lá bị úa. Nó bị héo.*".

Sau khi tạo dựng được CT-BD-DN, luận án xác định các yếu tố quan hệ và tạo sinh cấu trúc cú pháp của câu tiếng Việt mới theo các bước sau. *Bước 1*: Xác định vị từ ngữ nghĩa của đối tượng tĩnh vật làm trung tâm. Thêm vị từ này vào cấu trúc cú pháp ở vị trí đầu tiên. *Bước 2*: Thêm <bị> vào cấu trúc cú pháp. *Bước 3*: Thêm các vị từ ngữ nghĩa của quá trình thứ nhất vào cấu trúc cú pháp. *Bước 4*: Thêm yếu tố quan hệ thứ tự thời gian vào cấu trúc cú pháp. *Bước 5*: Thêm <bị> vào cấu trúc cú pháp. *Bước 6*: Thêm các vị từ ngữ nghĩa của quá trình thứ hai vào cấu trúc cú pháp.

Để thử nghiệm và đánh giá, luận án xây dựng được bộ ngữ liệu thử nghiệm bao gồm 1035 cặp câu tiếng Việt. Luận án tiến hành so sánh với các câu tiếng Việt được tạo sinh bởi [Boudin và Morin 2013] và [Filippova 2010].

Bảng 4.3. Kết quả thử nghiệm các cặp câu chỉ quá trình

Hệ thống	**Xử lý "nó"**	**Uni-gram**	**Bi-gram**	**TB Recall**	**TB Precision**	**TB F-score**
Gen		∨		0.8986	0.8695	0.8800
Fi		∨		0.379	0.9177	0.5133
Bo		∨		0.379	0.9177	0.5133
Fi	∨	∨		0.5605	0.9042	0.6812
Bo	∨	∨		0.5605	0.9042	0.6812
Gen			∨	0.7334	0.7191	0.7241

Fi			∨	0.1788	0.4266	0.244
Bo			∨	0.1788	0.4266	0.244
Fi	∨		∨	0.3303	0.5934	0.4126
Bo	∨		∨	0.3303	0.5934	0.4126

4.4. Từ biểu diễn của những đoạn văn bản có nhiều hơn hai câu

Trong giải thuật tạo sinh, luận án xem xét ba vị từ chức năng liên tiếp (F_i-1, F_i, F_i+1): (i) (F_i-1, F_i) có độ ưu tiên xem xét cao hơn hay bằng so với (F_i, F_i+1), luận án tạo sinh câu tiếng Việt mới dựa theo F_i-1 và F_i; (ii) (F_i, F_i+1) có độ ưu tiên xem xét cao hơn, luận án tái tạo câu tiếng Việt đơn giản dựa trên F_i-1.

Giải thuật 4.4. Tạo sinh danh sách cấu trúc câu tiếng Việt mới.

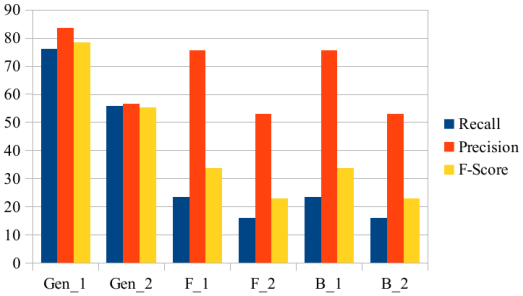
Đầu vào. FP = Danh sách vị từ chức năng.

Đầu ra. GS = Danh sách cấu trúc được tạo sinh.

curF_ ← 2; n = |FP|;
while curF_ < n do
    F_i-1, F_i, F_i+1 ∈ FP; reL ← IsRel(F_i-1, F_i); reR ← IsRel(F_i, F_i+1);
    if (reL && reR) then pr ← Pri(F_i-1, F_i, F_i+1);
        if (pr == -1 || pr == 0) then GS ← GS ∪ Fuse(F_i, F_i+1); curF_ ← curF_ + 2;
        if (pr == 1) then GS ← GS ∪ ReCr(F_i-1); curF_ ← curF_ + 1;
    if (reL) then GS ← GS ∪ Fuse(F_i, F_i+1); curF_ ← curF_ + 2;
    if (reR) then GS ← GS ∪ ReCr(F_i-1); curF_ ← curF_ + 1;

9: else GS ← GS ∪ ReCr(F_i-1); GS ← GS ∪ ReCr(F_i); curF_ ← curF_ + 2;

Để thử nghiệm và đánh giá, luận án xây dựng tập thử nghiệm là danh sách 385 đoạn tiếng Việt với tổng số 1561 câu. Luận án so sánh với hai phương pháp cơ sở dựa trên đồ thị là [Boudin và Morin 2013] và [Filippova 2010], trong đó phương pháp tạo sinh trong Mục 4.4 này là Gen_1 biểu thị ROUGE-1 và Gen_2 biểu thị ROUGE-2.



Hình 4.4. Kết quả so sánh hiệu năng giữa những phương pháp tạo sinh những đoạn văn bản tóm tắt tiếng Việt

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Tóm tắt văn bản theo hướng tiếp cận Tóm tắt trừu tượng đang ngày càng khẳng định được vai trò quan trọng trong các hệ thống thông minh nhân tạo, đặc biệt là trong thời đại bùng nổ thông tin hiện nay. Vấn đề hiểu và biểu diễn được ngữ nghĩa của văn bản cho trước, từ đó tạo sinh một văn bản mới ngắn gọn đúng ngữ pháp và mang tính tự nhiên đối với sự tri nhận của con người càng làm tăng thêm năng lực biểu đạt ngôn ngữ và trí thông minh nhân tạo cho máy tính. Với mục đích đặt ra ban đầu là giải quyết phần nào vấn đề này, luận án tiến sĩ với tiêu đề "**Tóm tắt đoạn văn bản tiếng Việt dựa trên cách tiếp cận tạo sinh**" đã đạt được một số kết quả như sau. *Thứ nhất*, đề xuất các chiến lược và phương pháp, kỹ thuật xử lý đại từ hồi chỉ và biểu diễn ngữ nghĩa cho một số dạng đoạn văn bản ngắn gồm các câu tiếng Việt có cấu trúc xác định. *Thứ hai*, đề xuất các chiến lược và phương pháp, kỹ thuật xử lý đồng tham chiếu và mô hình biểu diễn cho một số dạng đoạn văn bản tiếng Việt ngắn. *Thứ ba*, đề xuất phương pháp chuyển đổi câu tiếng Việt có cấu trúc thông thường về các dạng cấu trúc nông. *Thứ tư*, đề xuất các phương pháp tạo sinh đoạn văn bản tiếng Việt dựa trên việc tạo sinh câu tiếng Việt từ biểu diễn CT-BD-DN.

5.2. Hướng phát triển

Từ những kết quả đạt được trong luận án, các vấn đề đặt ra cần quan tâm nghiên cứu trong thời gian tới như sau. *Thứ nhất*, tiếp tục nghiên cứu áp dụng kiến thức trong các lý thuyết ngôn ngữ học

để xem xét thêm những mối quan hệ giữa các dạng câu, góp phần nâng cao chất lượng câu tiếng Việt mới được tạo sinh. *Thứ hai*, nghiên cứu thêm những kinh nghiệm thực tế khi sử dụng tiếng Việt trong giao tiếp thông thường để đề xuất thêm mở rộng bộ quy tắc chuyển đổi F-ConvRules. *Thứ ba*, cải tiến các mô hình tóm tắt văn bản theo hướng độc lập ngôn ngữ để áp dụng cho tiếng Việt. Điều này cho phép tóm tắt các văn bản dài và xử lý các kho ngữ liệu văn bản lớn.

Luận án thực hiện các thành phần trong mô hình kiến trúc ở Hình 4.1 theo định hướng độc lập ngôn ngữ. Dựa trên kiến trúc này, tác giả luận án tiếp tục tham gia một cuộc thi nghiên cứu quốc tế về tạo sinh ngôn ngữ tự nhiên là End-to-End (E2E) Natural Language Generation (NLG) Challenge 20176. (Phụ lục PL.2).

Một định hướng phát triển trong thời gian tới là mở rộng nghiên cứu trong cuộc thi E2E NLG Challenge 20176 để áp dụng cho ngôn ngữ tiếng Việt: (i) Nghiên cứu áp dụng *biểu diễn ngữ nghĩa phẳng* ("Flat Meaning Representation" – Flat MR) vào biểu diễn văn bản tiếng Việt; (ii) Nghiên cứu áp dụng phương pháp tạo sinh đoạn văn bản tiếng Việt từ biểu diễn Flat MR.

DANH MỤC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN ĐẾN LUẬN ÁN

[CT. 1] Trung Tran, Dang Tuan Nguyen (2013), "A Solution for Resolving Inter- sentential Anaphoric Pronouns for Vietnamese Paragraphs Composing Two Single Sentences", *Proceedings of The 5th IEEE International Conference of Soft Computing and Pattern Recognition (SoCPaR 2013)*, Hanoi, Vietnam, pp. 172–177. (ISBN: 978-1-4799-3400-3).

[CT. 2] Trung Tran, Dang Tuan Nguyen (2013), "Improve Effectiveness Resolving some Inter-sentential Anaphoric Pronouns Indicating Human Objects in Vietnamese Paragraphs using Finding Heuristics with Priority", *Proceedings of The 10th RIVF International Conference on Computing and Communication Technologies – Research, Innovation, and Vision for the Future (RIVF 2013)*, Hanoi, Vietnam, pp. 109–114. (ISBN: 978-1-4799-1350-3).

[CT. 3] Trần Trung, Nguyễn Tuấn Đăng (2014), "Merging Two Vietnamese Sentences Related by Inter-sentential Anaphoric Pronouns for Summarizing", *Kỷ yếu Hội nghị 1st NAFOSTED Conference on Information and Computer Science (NICS 2014)*, Hanoi, Vietnam, pp. 371–381. (ISBN: 978- 604-67-0228-3).

[CT. 4] Trung Tran, Dang Tuan Nguyen (2014), "Specification Model of Paragraph Summarization by Verbal Relationships: Objective, Cause, Consequence, Concurrence", *Proceedings of The 2nd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS 2014)*, Madrid, Tây Ban Nha, pp. 205–210. (ISBN 978-1-4799-7599-0).

[CT. 5] Trung Tran, Dang Tuan Nguyen (2015), "Semantic Predicative Analysis for Resolving Some Cases of Ambiguous Referents of Pronoun "Nó" in Summarizing Meaning of Two Vietnamese Sentences", *Proceedings of The 17th UKSIM-AMSS International Conference on Modelling and Simulation (UKSIM 2015)*, Cambridge, United Kingdom, pp. 340–345. (ISBN 978-1-4799-8712-2).

[CT. 6] Trung Tran, Dang Tuan Nguyen (2015), "Combined Method of Analyzing Anaphoric Pronouns and Inter-sentential Relationships between Transitive Verbs for Enhancing Pairs of Sentences Summarization", *Proceedings of The 4th Computer Science On-line Conference (CSOC 2015) – Vol 1: Artificial Intelligence Perspectives and Applications*, in: R. Silhavy et al. (eds), Advances in Intelligent Systems and Computing – Vol. 347, pp. 67–77. (ISBN 978-3-319-18476- 0).

[CT. 7] Trung Tran, Dang Tuan Nguyen (2015), "Modelling Consequence Relationships between Two Action, State or Process Vietnamese Sentences for Improving the Quality of New Meaning-Summarizing Sentence", *International Journal of Pervasive Computing and Communications*, Vol.11 (2), pp. 169–190. Emerald Group Publishing Limited. (ISBN: 1742-7371). (SCOPUS Index; ISI Index).

[CT. 8] Trần Trung, Nguyễn Tuấn Đăng (2016), "Xác định thứ tự thời gian giữa hai câu tiếng Việt chỉ quá trình để tóm lược" [Determining The Temporal Order between Two Vietnamese Process Sentences for Summarizing], *Chuyên san "Các công trình nghiên cứu, phát triển và ứng dụng Công nghệ thông tin và Truyền thông"* [Vietnam Journal on Information Technology and Communication – Research, Development and

Application on Information and Communication Technology], Tập V-1, Số 15 (35), trang 38–54. (ISSN: 1859-3526).

[CT. 9] Trung Tran, Dang Tuan Nguyen (2016), "Algorithm of Computing Verbal Relationships for Generating Vietnamese Paragraph of Summarization from The Logical Expression of Discourse Representation Structure", *Vietnam Journal of Computer Science*, Vol. 3 (1), pp. 35–46. (ISSN: 2196-8888 (Print) 2196-8896 (Online)).

[CT. 10] Trần Trung, Nguyễn Tuấn Đăng (2016), "Một Phương Pháp Dựa Trên Luật để Chuyển Đổi Văn Bản Tiếng Việt về DRS (Discourse Representation Structure)" [A Rule-based Method for Transforming Vietnamese Paragraphs into DRS (Discourse Representation Structure)], *Chuyên san Công nghệ Thông tin và Truyền thông, Tạp chí Khoa học và Kỹ thuật, Học viện Kỹ thuật quân sự* [Journal of Science and Technology: The Section on Information and Communication Technology (LQDTU-JICT)], Số 9, trang 61–83. (ISSN: 1859-0209).

[CT. 11] Trần Trung, Nguyễn Tuấn Đăng (2017), "Co-Reference Resolution in Graph Model for Enhancing Vietnamese Paragraph Compression", *Kỷ yếu Hội nghị 4th NAFOSTED Conference on Information and Computer Science (NICS 2017)*, Hà Nội, Việt Nam, trang 258–263. (ISBN: 978-1- 5386-3210-9).

[CT. 12] Dang Tuan Nguyen, Trung Tran (2017), "Phrasal Graph-based Method for Abstractive Vietnamese Paragraph Compression", *Proceedings of The 8th International Symposium on Information and Communication Technology (SoICT 2017)*, Nha Trang, Vietnam, pp. 143–150. (ISBN: 978-1-4503-5328-1).

[CT. 13] Trần Trung, Nguyễn Tuấn Đăng (2018), "Anaphoric Pronoun Resolution for Improving Fusion Based Summarization of Simple Vietnamese Texts", *Hội nghị khoa học quốc gia lần thứ XI về "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin"* [Fundamental and Applied IT Research (FAIR 2018)], Hà Nội, Việt Nam, trang 42-50.

[CT. 14] Trần Trung, Nguyễn Tuấn Đăng (2018), "Text Generation from Abstract Semantic Representation for Summarizing Vietnamese Paragraphs Having Co-references", *Kỷ yếu Hội nghị 5th NAFOSTED Conference on Information and Computer Science (NICS 2018)*, Thành phố Hồ Chí Minh, Việt Nam, trang 94–99.

[CT. 15] Dang Tuan Nguyen, Trung Tran (2018), "Structure-based Generation System for E2E NLG Challenge", *Proceedings of E2E NLG Challenge System Descriptions*. Un-official Publication.

[CT. 16] Dang Tuan Nguyen, Trung Tran (2019), "Graph Transformation System from Stanford Basic Dependencies to Universal Conceptual Cognitive Annotation (UCCA)", *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval-2019)*, Minneapolis, Minnesota, USA, pp. 97–101.