TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

KHÓA LUẬN NGHIÊN CỨU KHOA HỌC

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP PHÁT SINH CÂU HỎI TỰ ĐỘNG TỪ CÂU TRẢ LỜI CÓ SẴN VÀ ỨNG DỤNG

Người hướng dẫn: **TS. TRẦN THANH PHƯỚC**

Người thực hiện: **NGUYỄN DUY KHANH – 518H0519**

ĐINH NGUYỄN NHẬT TÙNG – 518H0584

Lớp : **18H50201**

Khoá : **22**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

KHÓA LUẬN NGHIÊN CỨU KHOA HỌC

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP PHÁT SINH CÂU HỎI TỰ ĐỘNG TỪ CÂU TRẢ LỜI CÓ SẴN VÀ ỨNG DỤNG

Người hướng dẫn: **TS. TRẦN THANH PHƯỚC**
Người thực hiện: **NGUYỄN DUY KHANH – 518H0519**
ĐINH NGUYỄN NHẬT TÙNG – 518H0584
Lớp **: 18H50201**
Khoá **: 22**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021

LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn tới tất cả quý thầy cô giáo, cán bộ khoa Công Nghệ Thông Tin trường Đại học Tôn Đức Thắng vì công tình dạy dỗ, giúp đỡ chúng em trong suốt khoảng thời gian ngồi trên ghế nhà trường.

Trong quá trình thực hiện bài nghiên cứu này cũng như trong suốt năm học vừa qua, chúng em nhận được sự hỗ trợ và chỉ bảo đầy nhiệt tình của TS. Trần Thanh Phước. Chúng em rất cảm ơn Thầy vì đã bên cạnh, hướng dẫn, định hướng con đường và kiến thức cho chúng em, đó là những điều vô cùng đáng trân trọng. Và một lần nữa chúng em muốn gửi Thầy một lời cảm ơn chân thành nhất.

Chúng em cũng xin gửi lời cảm ơn tới bậc cha mẹ đã luôn ủng hộ, động viên, tạo điều kiện cho chúng em trong quá trình học tập và hoàn thành bài nghiên cứu này.

Mặc dù đã cố gắng hoàn thành nhưng do hạn chế về kinh nghiệm nên bài nghiên cứu sẽ không tránh khỏi những sai sót. Chúng em mong nhận được sự cảm thông và những ý kiến đóng góp của các thầy cô và các bạn.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan những nội dung trình bày trong báo cáo này là kết qảu nghiên cứu, tìm hiểu của riêng chúng tôi dưới sự hướng dẫn của TS Trần Thanh Phước;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào chúng tôi xin hoàn toàn chịu trách nhiệm về nội dung báo cáo của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày 28 tháng 5 năm 2021

Tác giả

Nguyễn Duy Khanh

Đinh Nguyễn Nhật Tùng

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của giảng viên hướng dẫn

_____

_____

_____

_____

_____

_____

TP. Hồ Chí Minh, ngày tháng năm 2021

Giảng viên hướng dẫn

TS. Trần Thanh Phước

Phần đánh giá của giảng viên chấm bài

_____

_____

_____

_____

_____

TP. Hồ Chí Minh, ngày tháng năm 2021

Giảng viên chấm bài

TÓM TẮT

Câu hỏi là một trong những thước đo chính xác nhất trong vấn đề đánh giá kiến thức của chúng ta, vì lẽ thường tình đó mà các câu hỏi luôn tồn tại trong hầu hết các bài kiểm tra đánh giá năng lực trong trường học từ trước đến nay. Như vậy chúng ta có thể cảm thấy rằng tầm ảnh hưởng của câu hỏi trong đời sống là quan trọng như thế nào. Với một khía cạnh của người trong ngành công nghệ, chúng tôi nhận thấy một mô hình khởi tạo câu hỏi tự động từ việc trích xuất các thông tin trong văn bản là một mô hình cực kỳ hữu ích, đem lại một chất xúc tác rất mạnh đến những việc tự học, tự nghiên cứu của các nghiên cứu sinh, học sinh, sinh viên tự học trên Internet. Vì họ có thể tự đánh giá được trình độ của mình sau khi đọc một bài báo hoặc một trang blog kiến thức trên mạng, điều mà trước đây họ chưa được trải nghiệm.

Tại sao chúng tôi lại đánh giá cao mô hình khởi tạo câu hỏi tự động này, tại sao lại trả lời câu hỏi giúp chúng ta củng cố được kiến thức. Trong tháp Phân loại tư duy của Bloom, sau khi đọc một kiến thức nào đó, chúng ta chỉ mới dừng lại ở mức độ nhớ (Remember), vậy làm thế nào để chúng ta có thể biết được mình hiểu kiến thức đấy sâu, rộng đến mức nào. Muốn biết được điều này, ta không cách nào khác ngoài việc trả lời được các câu hỏi về kiến thức đã được học, nhưng thật không may, đa số các trang web thường không tích hợp các câu hỏi giúp ta củng cố tư duy, vẫn có một số nhưng lượng này không nhiều. Giả định rằng có một mô hình có thể đánh giá được trình độ hiểu của bản thân về một kiến thức nào đó, thì luồng tư duy và nhận thức của chúng ta sẽ thay đổi, chúng ta không còn lầm tưởng rằng ta đã hiểu được vấn đề nữa, hoặc cứ học đi học lại một kiến thức nhất định và nghĩ ta chưa thật sự hiểu về nó. Như vậy đối với chúng tôi, việc tạo được một mô hình phát sinh câu hỏi tự động là một công việc vô cùng ý nghĩa, nó thúc đẩy rất mạnh quá trình tự học, tự tìm hiểu của các học sinh, sinh viên, đặc biệt là trong kỷ nguyên công nghệ số, mạng Internet là nguồn tri thức vô tận của nhân loại và dễ dàng được truy cập khi ngồi tại nhà mà không cần tốn sức đi tìm.

Bên cạnh đó, việc tạo câu hỏi là một vấn đề nhưng câu hỏi có sâu, có rộng, có dẫn dắt được người trả lời vào trọng tâm vấn đề không thì cái đó còn là một vấn đề khác. Chúng tôi sẽ cố gắng tìm hiểu các phương pháp khởi tạo mô hình phát sinh câu hỏi tự động, và đầu ra là những câu hỏi có ý nghĩa sâu và rộng, thúc đẩy được tư duy và giúp học sinh, sinh viên, những người tự học có thể đánh giá được năng lực thật sự của bản thân mà có những chiến lược học tập hợp lý.

Trong bài nghiên cứu này, chúng tôi sẽ trình bày lý do chọn đề tài, các khái niệm tổng quan về hệ thống khởi tạo câu hỏi tự động, các cơ sở lý thuyết của việc trích xuất thông tin, xử lý văn bản, các công cụ cơ bản của xử lý ngôn ngữ tự nhiên và đề xuất hướng xây dựng mô hình phát sinh câu hỏi tự động từ văn bản.

MỤC LỤC

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC KÝ HIỆU

CÁC CHỮ VIẾT TẮT

QR:	Natural Language Processing
NLG:	Natural Language Generation
NLU:	Natural Language Understanding
FAQ:	Frequently Asked Question
NXB:	Nhà xuất bản
POS:	Part of Speech
NER:	Named Entity Recognition
ITS2010:	10th International Conference Intelligent Tutoring Systems

DANH MỤC CÁC HÌNH VẼ

DANH MỤC CÁC BẢNG

CHƯƠNG 1 – TỔNG QUAN ĐỀ TÀI

1.1 Bối cảnh thực hiện

Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ – công cụ hoàn hảo nhất của tư duy và giao tiếp. Trong NLP, tiếp tục được chia thành hai nhánh, Natural Language Generation (NLG) và Natural Language Understanding (NLU).

NLG đảm nhận nhiệm vụ chuyển đổi ngôn ngữ của máy, bằng một cách nào đó sang ngôn ngữ tự nhiên của con người. Cụ thể hơn, để có thể dịch ngôn ngữ máy sang ngôn ngữ của chúng ta thì NLG phải hoàn thành các nhiệm vụ như xác định được nội dung – thông tin trình bày trong văn bản, sắp xếp từ ngữ đúng ngữ pháp và ngữ cảnh, sắp xếp nội dung trình bày một cách hợp lý, có ý nghĩa. Trái ngược với NLG, NLU có chức năng hiểu ngôn ngữ con người và chuyển những kiến thức đó thành ngôn ngữ máy.

Phát sinh câu hỏi (Question Generation – QG) từ một văn bản có sẵn là một trong những nhánh con của NLG. Cho trước một số nội dung văn bản, mục tiêu của QG là tự động khởi tạo một tập các câu hỏi và câu trả lời cho các câu hỏi này chính là những nội dung trong văn bản. Cũng giống như con người, để đặt được câu hỏi từ một bài báo, luận văn, chúng ta phải hiểu được nó, nếu không hoàn toàn hiểu thì vẫn có thể đặt được câu hỏi tương ứng. Một vài ứng dụng có thể có của QG:

1. Trong giáo dục, đặt câu hỏi giúp đánh giá khả năng đọc hiểu. Học sinh có thể sử dụng nó để tự đánh giá. Nó có thể được sử dụng để tạo các câu đố và bài kiểm tra trực tuyến mà không cần tạo theo cách thủ công từ giáo viên. Bên cạnh đó, đối với những nghiên cứu sinh, những người thích tự học, tự nghiên cứu tài liệu thì họ rất cần những câu hỏi mang tính đánh giá cao, và tự nhận thức được trình độ của bản thân

2. Đối với Chatbot, thì chúng ta có thể ứng dụng QG để tạo ra các cặp câu hỏi-câu trả lời (QA), và sử dụng kết quả đó như một bộ dữ liệu lớn để huấn luyện mô hình học sâu cho hệ thống trả lời tự động

3. Trích xuất các câu hỏi thường gặp (FAQ – Frequently Asked Question)

4. Gợi ý câu hỏi cho bệnh nhân và người chăm sóc trong y học

Nghiên cứu này của chúng tôi sẽ tập trung vào bài toán phát sinh câu hỏi giúp đánh giá khả năng tiếp thu, thúc đẩy mạnh và nâng cao chất lượng quá trình tự học của học sinh, sinh viên, nghiên cứu sinh và đặc biệt là những anh/chị thích tự học và không có người hướng dẫn. Rất nhiều nghiên cứu đã chỉ ra rằng, học sinh/sinh viên thường chưa đủ kinh nghiệm để có thể nhận thức mức độ hiểu của mình về văn bản, tài liệu sau khi đọc chúng, và không thường xuyên đặt câu hỏi, nếu có cũng chỉ là những câu hỏi ở mức độ nông, không mang tính đúc kết kiến thức cao. Những nguồn học tập trên Internet thường sẽ không đính kèm theo câu hỏi giúp thúc đẩy việc học. Tuy tài liệu học tập dồi dào nhưng không đính kèm câu hỏi khiến cho QG từ văn bản trở thành một ý tưởng vô cùng đáng giá trong nền giáo dục 4.0 hiện nay.

Để bài toán QG có thể đánh giá chất lượng học tập tốt, thì chúng ta ko chỉ quan tâm đến câu hỏi có thể phát sinh được, mà còn phải tìm hiểu những câu hỏi nên phát sinh. Vậy những câu hỏi như thế nào là nên được khởi tạo, để trả lời cho câu hỏi này thì chúng ta sẽ dựa vào một hệ thống cấp bậc phân loại tư duy của Bloom, được công bố vào năm 1956. Tuy nhiên, thế giới ngày nay đã khác so với những điều mà phương pháp phân loại tư duy của Bloom phản ánh. Sự hiểu biết về cách thức học tập của học sinh, cũng như cách thức dạy học của giáo viên đã được tăng lên rất nhiều và các nhà giáo dục đã nhận ra rằng dạy và học chứa đựng nhiều điều hơn là chỉ có phát triển tư duy. Đó chính là tình cảm, lòng tin của học sinh, của giáo viên cũng như của môi trường văn hóa và xã hội trong lớp học. Như vậy là vào năm 1999, Tiến sĩ Lorin Anderson cùng những đồng nghiệp của mình đã xuất bản phiên bản mới được cập nhật về Phân loại tư duy của Bloom.

Ông lưu tâm tới những nhân tố ảnh hưởng tới việc dạy và học trong phạm vi rộng hơn. Phiên bản Phân loại tư duy mới này đã cố gắng chỉnh sửa một số vấn đề có trong bản gốc. Đối với mục đích bài nghiên cứu, chúng tôi chỉ quan tâm đến phiên bản mới của Bảng phân loại tư duy của Bloom, nó bao gồm 6 bậc chính, càng lên mức độ cao thì độ phức tạp càng tăng, nhưng không vì vậy mà chúng tạo thành một lớp phân cấp cứng nhắc, khái niệm mỗi lớp đôi khi bị chồng chéo lên nhau.

Level	**Category**	**Cognitive Processes**
1	Remember	recognizing, recalling
2	Understand	interpreting, exemplifying, classifying, summarizing, inferring, comparing, explaining
3	Apply	executing, implementing
4	Analyze	differentiating, organizing, attributing
5	Evaluate	checking, critiquing
6	Create	generating, planning, producing

Bảng 1. 1 Bảng phân loại nhận thức, phiên bản mới của Bloom

Những câu hỏi tốt, có chiều sâu, thường là những câu hỏi trên level 1. Những câu hỏi nằm trong độ này thường là những câu hỏi nông, chỉ cần người đọc nhớ chính xác thông tin trong văn bản, mặc dù không hiểu nhưng vẫn có thể trả lời được, những câu hỏi dạng nông cạn như thể này được biết đến với tên gọi câu hỏi factoid. Như vậy, nếu chiến lược của chúng ta chỉ tập trung tạo câu hỏi dựa trên các câu đơn, thì khả năng tạo ra câu hỏi factoid là rất cao. Với những phương pháp đơn giản liên quan đến syntax (cú pháp), semantic (ngữ nghĩa) và thay thế các key word thành từ để hỏi (ai, ở đâu, khi nào, cái gì) chúng ta đều có thể tạo ra các câu hỏi kiểu factoid. Hãy xem xét câu trong ví dụ (1.1) .

Trái Đất nóng lên là do việc thải lượng lớn khí metan quá mức cho phép ở ngưỡng giới hạn nhất định từ Bắc Cực và các vùng đất ẩm ướt. (1.1)

Chúng ta có thể sinh ra các câu hỏi kiểu như sau:

Cái gì nóng lên là do việc thải lượng lớn khí metan quá mức cho phép ở ngưỡng giới hạn nhất định từ Bắc Cực và các vùng đất ẩm ướt? (1.2)

Trái Đất nóng lên là do việc thải lượng lớn khí gì quá mức cho phép ở ngưỡng giới hạn nhất định từ Bắc Cực và các vùng đất ẩm ướt? (1.3)

Trái Đất nóng lên là do việc thải lượng lớn khí metan quá mức cho phép ở ngưỡng giới hạn nhất định từ đâu? (1.4)

Trái Đất nóng lên là do đâu? (1.5)

Dễ dàng thấy được rằng câu hỏi (1.5) là câu thông dụng và có giá trị nhất, 3 câu còn lại quá dài, và dễ dàng để trả lời được bằng việc nhớ thông tin trong bài, không đọng lại nhiều kiến thức sau khi trả lời. Một câu hỏi tốt và nên được khởi tạo phải có dạng như Nguyên nhân vì sao Trái Đất nóng lên?

Mục tiêu của chúng tôi là chứng minh rằng vẫn còn giá trị khi khởi tạo câu hỏi từ một câu đơn và tìm ra cách tiếp cận phù hợp, chúng tôi có thể thoát khỏi rào cản của câu hỏi factoid và đặt ra những câu hỏi không chỉ tốt về mặt syntax và semantic mà còn hữu ích hơn về mặt giáo dục. Chúng tôi muốn chứng minh rằng bằng cách tận dụng nội dung ngữ nghĩa của tài liệu học tập, chúng tôi có thể tạo ra các câu hỏi hữu ích về mặt sư phạm hơn là các câu hỏi factoid.

1.2 Mục tiêu đề tài

Nghiên cứu kỹ thuật khởi tạo câu hỏi tự động dựa trên câu đơn, cấu trúc các nhánh câu đơn và chia ra nhiều trường hợp xử lý phù hợp cho từng nhánh. Áp dụng các luật cú pháp (syntax) và ngữ nghĩa ở mức nông của ngôn ngữ tiếng Việt để trích xuất thông tin cần được hỏi và chuyển câu thành câu hỏi tương ứng trong một văn bản bất kỳ;

Xây dựng mô hình đặt câu hỏi tự động nhằm hỗ trợ học sinh, sinh viên tự học, có thể áp dụng mô hình vào thực tiễn.

1.3 Đối tượng và phạm vi nghiên cứu

1.3.1 Đối tượng nghiên cứu

Luận văn tập trung nghiên cứu các đối tượng sau:

- Kỹ thuật trích xuất câu đơn, câu chứa thông tin có thể hỏi được từ văn bản;
- Kỹ thuật tách các vế trong câu ghép;
- Kỹ thuật và phương pháp tạo câu hỏi trên từng cấu trúc câu đơn khác nhau;
- Các vấn đề tổng quan và đặc điểm của hệ thống phát sinh câu hỏi tự động;
- Nghiên cứu xây dựng mô hình phát sinh câu hỏi tự động bằng phương pháp syntax và ngữ nghĩa nông.

1.3.2 Phạm vi nghiên cứu

Nội dung luận văn này tập trung nghiên cứu các kỹ thuật trích xuất thông tin và xử lý ngôn ngữ từ văn bản. Xử lý và đặt câu hỏi ở phạm vi câu đơn, nhưng vì dữ liệu trong thực tế tồn tại rất ít câu đơn để xử lý, điều này đặc biệt đúng với văn phong của người Việt Nam, nên chúng tôi quyết định nghiên cứu thêm kỹ thuật tách các vế trong câu ghép thành các câu đơn tương ứng. Phương pháp chúng tôi tiếp cận để nghiên cứu sẽ chủ yếu dựa trên cú pháp và ngữ nghĩa mức nông của câu.

1.4 Cấu trúc đề tài

Toàn bộ nội dung bài báo cáo được trình bày trong 6 chương:

Chương 1: TỔNG QUAN ĐỀ TÀI

- Trình bày bối cảnh lựa chọn đề tài, giới thiệu các khái niệm cơ bản của đối tượng nghiên cứu, mục đích, bố cục bài báo cáo;
- Trình bày tổng quan và nhu cầu cần thiết của việc tự học, tự nghiên cứu trong thời đại số hoá hiện nay, kiến thức rất dễ tìm thấy trên mạng. Kiến

thức tuy đã có sẵn, nhưng việc cảm thấu được chúng thì không dễ. Phải có người kiểm chứng, hoặc một bài đánh giá, bài kiểm tra để đánh giá mức độ hiểu của người học, nhưng hầu hết các trang báo, trang tài liệu đều không được tích hợp các câu hỏi đi kèm, nếu có thì nó không được đa dạng và phong phú. Trong Chương 1, chúng tôi sẽ trình bày khái quát một số phương pháp tiếp cận và ý nghĩa của đề tài khi áp dụng vào thực tiễn.

Chương 2: CƠ SỞ LÝ THUYẾT

- Các kiến thức về cú pháp ngôn ngữ tiếng Việt, cơ sở lý thuyết của các thành phần trong câu đơn;
- Trình bày cơ sở lý thuyết về các kỹ thuật cơ bản trong phân tích và xử lý ngôn ngữ tự nhiên;
- Trình bày cơ sở lý thuyết về các kỹ thuật tách câu, tiêu chí chọn câu đạt chỉ tiêu để tạo câu hỏi.

Chương 3: PHƯƠNG PHÁP PHÁT SINH CÂU HỎI TRÊN CÂU ĐƠN

- Trình bày các kỹ thuật chung để khởi tạo câu hỏi;
- Các phương pháp hay để phát sinh câu hỏi như phân tích cú pháp, ngữ nghĩa của câu, tạo khuôn mẫu.

Chương 4: PHƯƠNG PHÁP CỦA CHÚNG TÔI

- Áp dụng kết quả đã nghiên cứu được, xây dựng mô hình khởi tạo câu hỏi tự động dựa trên câu đơn với miền tự do;
- Xây dựng mô hình phân tách các vế trong câu ghép đẳng lập.

Chương 5: THỬ NGHIỆM VÀ ĐÁNH GIÁ MÔ HÌNH

- Thử nghiệm mô hình trên các tài liệu corpus;
- Đề ra phương pháp đánh giá mô hình.

Chương 6: KẾT LUẬN

- Tóm tắt và đánh giá kết quả nghiên cứu;

- Đề xuất định hướng nghiên cứu tiếp theo.

CHƯƠNG 2 – CƠ SỞ LÝ THUYẾT

2.1 Quy tắc cú pháp tiếng Việt

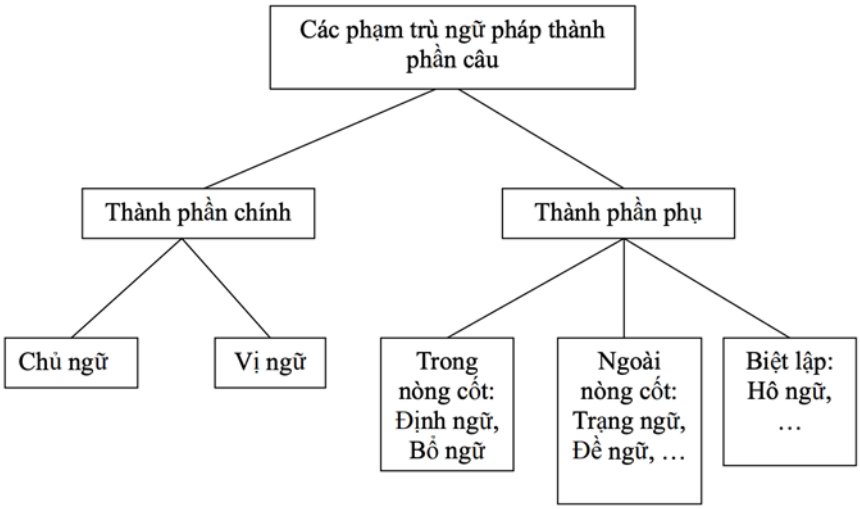
2.1.1 Tập quy tắc cú pháp tiếng Việt cho thành phần câu

Trong mô hình khởi tạo câu hỏi tự động, với phương pháp phân tích cấu trúc cú pháp của câu (syntax), chúng tôi sử dụng chủ ngữ, vị ngữ, trạng ngữ để trích xuất thông tin cần được hỏi. Chủ ngữ sẽ trả lời cho câu hỏi Ai, vị ngữ sẽ là thông tin để trả lời cho các câu hỏi làm gì, là gì, thế nào. Còn trạng ngữ sẽ có chức năng trả lời cho câu hỏi ở đâu, khi nào. Như vậy sẽ rất hữu ích nếu chúng ta nắm được mối quan hệ và nguyên tắc ngữ pháp giữa các thành phần từ đó có cái nhìn tổng thể vấn đề và đưa ra những chiến lược phù hợp để phát sinh câu hỏi trên từng loại câu. Vì lý do đó, trong đề mục này, chúng tôi đã nghiên cứu về quy tắc cú pháp cho các thành phần trong câu đây là mục 4 của bài báo khoa học [1] được soạn bởi hai nhà nghiên cứu Lê Thanh Hương và Đỗ Bá Lâm. Bên cạnh việc tái sử dụng những kiến thức này ở trong chương sau, nó còn giúp chúng ta hình thành những nhận định rõ ràng, chặt chẽ về cấu trúc ngữ pháp trong câu.

2.1.1.1 Thành phần câu

Từ và ngữ sẽ tạo nên thành phần câu – những yếu tố cấu thành nên một câu hoàn chỉnh. Phạm trù ngữ pháp thành phần câu trong hệ thống phân loại lấy cấu trúc chủ - vị làm cơ sở phân biệt các thành phần câu: thành phần chính và thành phần phụ. Thành phần phụ có loại chứa trong thành phần nòng cốt, có loại đứng ngoài hay biệt lập với nòng cốt câu.

- Các thành phần phụ chứa trong thành phần của nòng cốt: định ngữ, bổ ngữ,...
- Các thành phần phụ ngoài nòng cốt: trạng ngữ, đề ngữ, phụ ngữ câu,...
- Các thành phần phụ biệt lập với nòng cốt câu: hô ngữ, liên ngữ, chú ngữ,...



Hình 2.1 Các phạm trù ngữ pháp thành phần câu

Hệ thống thành phần câu được phân thành ba cấp: các thành phần chính, các thành phần thứ, và các thành phần phụ thuộc. Thành phần chính trùng với hai vế của kết cấu chủ - vị. Các thành phần thứ độc lập với nhau khi tham gia vào chỉnh thể câu. Giá trị của mỗi thành phần đối với câu là khác nhau. Chẳng hạn, trạng ngữ có liên quan đến toàn câu và biểu hiện nhiều ý nghĩa khác nhau hơn so với các thành phần phụ thuộc hoặc thành phần xen. So sánh:

a) Cô gái nhà bên, "có ai ngờ", cũng vào du kích. (thành phần xen)

b) Gần sáng, lạnh. (định ngữ cho cả câu)

c) Vào mùa này, cây trái tốt tươi. (trạng ngữ)

Ý nghĩa của thành phần xen trong (a) không phụ thuộc vào từ nào trong câu cả. Ý nghĩa của định ngữ trong (b) bổ sung, thuyết minh cho toàn câu. Ý nghĩa của trạng ngữ trong (c) làm bối cảnh cho vị ngữ được thể hiện. Về cấu trúc câu thì trạng ngữ trong (c) tách khỏi toàn câu, còn thành phần xen trong (a) và thành phần định ngữ trong (b) có vị trí linh động. Chúng có thể ở đầu câu, cũng có thể ở giữa câu.

Các thành phần phụ thuộc thường nằm trong một nhóm nào đó trong mối quan hệ với một yếu tố nào đó của nhóm. Chẳng hạn, bổ ngữ nằm trong nhóm vị ngữ, định ngữ

nằm trong cả nhóm chủ ngữ và nhóm vị ngữ, v.v... Đến lượt mình, các thành phần phụ thuộc cũng có nét khác biệt nhau. Định ngữ khác bổ ngữ, bổ ngữ có khả năng chuyển đổi vị trí, còn định ngữ thì không có khả năng đó. Định ngữ bị chi phối bởi quan hệ thuộc tính (attribute), bổ ngữ thì bị chi phối bởi quan hệ bổ túc (completive). Quan hệ bổ túc này có liên quan đến thuộc tính từ vựng, ngữ pháp của từ. Và do đó, bổ ngữ được xem như là một thành phần phụ thuộc có liên đới đến cấu trúc câu trực tiếp hơn là định ngữ. So sánh:

a) Tôi đọc quyển sách hay.

b) Tôi đọc quyển sách.

c) Tôi đọc.

Ở đây, (a), (b) đều có bổ ngữ, trong khi đó (c) không có bổ ngữ. Bổ ngữ của (a) có định ngữ trong khi đó bổ ngữ của (b) không có định ngữ. Sự hiện diện của định ngữ ở (a) chỉ mang thêm nét nghĩa thuyết minh cho bổ ngữ mà thôi. Như vậy có vấn đề lồng nhau giữa các thành phần theo quan hệ tầng bậc: định ngữ nằm trong nhóm bổ ngữ (nếu bổ ngữ có định ngữ), nhóm bổ ngữ nằm trong nhóm vị ngữ. Do đó, phân tích đúng thành phần câu tiếng Việt phải biết tìm các quan hệ cú pháp tầng bậc này.

Vị trí của chủ ngữ luôn đứng trước vị ngữ. Tuy nhiên, ở vị trí này không chỉ có chủ ngữ, mà còn có các thành phần khác. Cho nên trong các câu dài, mở rộng cấu trúc, việc xác định ranh giới nhóm chủ ngữ và nhóm vị ngữ có lúc gặp khó khăn.

Sau đây chúng ta sẽ nghiên cứu quy tắc cú pháp của các thành phần câu, đó là: chủ ngữ, vị ngữ, bổ ngữ, trạng ngữ, và định ngữ.

2.1.1.2 Chủ ngữ

Chủ ngữ phần lớn các trường hợp mang ý nghĩa chỉ người và sự vật nhưng trong một số trường hợp chúng sẽ mang ý nghĩa khác. Chủ ngữ có thể là danh từ, danh ngữ, đại từ, tính từ, tính ngữ, động từ, động ngữ, số từ, hoặc trong trường hợp câu phức, chủ ngữ có thể là một cụm chủ - vị.

- *Trường hợp 1*: Chủ ngữ là danh ngữ/danh từ

  Ví dụ:

  Các bác nông dân ngoài ruộng quyết định nghỉ ngơi dùng bữa cơm.

  Cả hai chúng tôi đều cảm thấy mệt.

- *Trường hợp 2*: Chủ ngữ là cụm chủ - vị

  Ví dụ:

  Lan học giỏi làm bố mẹ rất vui.

  Cách mạng tháng tám thành công đem lại độc lập, tự do cho dân tộc.

- *Trường hợp 3*: Chủ ngữ là kiến trúc: **"<Từ phủ định> <Danh từ/ngữ> <Đại từ phiếm định>"**

  Ví dụ:

  Không đế quốc nào có thể quay lại bóp chết đời sống các em.

- *Trường hợp 4*: Chủ ngữ là kiến trúc: **"Có (Phiếm định) <Danh từ/ngữ>"**

  Ví dụ:

  Có những điều anh hỏi nghe thật buồn cười.

- *Trường hợp 5*: Chủ ngữ là kiến trúc: **"<Kết từ> <Danh từ/ngữ>"**

  Ví dụ:

  Gần sáng là lúc người ta hay ngủ say.

- *Trường hợp 6*: Chủ ngữ là kiến trúc song hành chỉ khoảng cách không gian và thời gian: **"Từ <Danh từ/ngữ> đến <Danh từ/ngữ>"**

  Ví dụ:

  Từ Hà Nội đến Hải Phòng là 105 km.

- *Trường hợp 7*: Chủ ngữ là ngữ cố định

  Ví dụ:

  Chỉ tay năm ngón thường làm hỏng việc.

- *Trường hợp 8*: Tỉnh lược chủ ngữ

Trong hoạt động ngôn ngữ, chủ ngữ là thành phần dễ bị tỉnh lược hơn vị ngữ. Điều này đưa ta đến hai hệ quả là chủ ngữ hiểu ngầm và chủ ngữ zero. Nhưng trong bài nghiên cứu này, chúng tôi chỉ xét các câu có chủ ngữ, vị ngữ rõ ràng để khởi tạo câu hỏi, nên phần này chúng tôi chỉ dừng ở mức giới thiệu trên.

2.1.1.3 Vị ngữ

Vị ngữ là một thành phần chính của câu và có tác động đến toàn câu. Nó là trung tâm tổ chức của câu và do vậy, vị ngữ có nhiều vấn đề phức tạp hơn chủ ngữ. Là trung tâm của tổ chức câu nên rất hiếm câu có vị ngữ bị lược bỏ. Dựa vào hệ từ và từ loại trong vai trò vị ngữ để chia thành hai kiểu: vị ngữ có hệ từ và vị ngữ không có hệ từ. Chẳng hạn:

a) Nhân dân ta rất anh hùng.

b) Anh ấy ngoài 30 tuổi.

c) Đây là giờ sinh tử.

Các ví dụ (a), (b) có vị ngữ không hệ từ. Ví dụ (c) có vị ngữ có hệ từ cùng với các tổ hợp danh từ, kết cấu chủ - vị.

Về ý nghĩa, vị ngữ biểu hiện sự hoạt động, tính chất, trạng thái của người, hiện tượng, sự vật được nêu ở chủ ngữ. Nghĩa của vị ngữ bao giờ cũng ở trong mối quan hệ với nghĩa của chủ ngữ. Đó là qua hệ đề - thuyết. Tuy nhiên, nghĩa của vị ngữ đa dạng hơn và tùy thuộc vào các kiểu câu khác nhau mà có những vị ngữ khác nhau.

Về tổ chức, vị ngữ được tổ chức bằng các động từ đơn, tính từ đơn hoặc nhóm động từ, nhóm tính từ và một số từ loại khác nhau như đại từ, số từ, danh từ,...

- *Trường hợp 1*: Vị ngữ động từ/ngữ

  Ví dụ:

  Tôi trông cây ở ông.

- *Trường hợp 2*: Vị ngữ với động từ đặc biệt *"là"*: **"là <Danh từ/ngữ>"; "là <Tính từ/ngữ>"**

  Ví dụ:

  Anh ta là chiến sĩ thi đua.

  Chỉ có anh ta là thông minh thôi.

- *Trường hợp 3*: Vị ngữ là tính từ/ngữ

  Ví dụ:

  Cô ta thông minh.

- *Trường hợp 4*: Vị ngữ danh ngữ.

  Ví dụ:

  Đồng hồ này ba kim.

  Cả nước một lòng.

Loại câu với vị ngữ là danh ngữ thường biểu thị ý nghĩa địa điểm, sự kiện, hiện tượng, bản chất. Chúng có mô hình tổng quát như sau:

<Vị ngữ> = <Số từ> <Danh từ>

Nhà này năm tầng

<Vị ngữ> = <Từ so sánh> <Danh từ>

Thân em như tấm lụa đào

<Vị ngữ> = <Đại từ>

Ai *đấy?*

<Vị ngữ> = <Loại từ> <Danh từ>

Mỗi người một phòng

- *Trường hợp 5*: Vị ngữ là ngữ cố định

  Ví dụ:

  Anh ấy ba voi không được bát nước xáo.

- *Trường hợp 6*: Vị ngữ mở rộng là cụm chủ vị

  Ví dụ:

  Sông Thương // nước chảy / đôi dòng.

Trong trường hợp này, vị ngữ bản thân nó là một cụm chủ vị với vị ngữ là "đôi dòng".

- *Trường hợp 7*: Động từ **"có"** gắn liền với các sự kiện tồn tại trong một không gian nhất định. Vì vậy, ở câu có ý nghĩa tồn tại, một khi hiện diện các từ không gian, thì **"có"** vắng mặt nhưng được hiểu như có mặt.

  Ví dụ:

  Bố tôi tóc đã bạc. (có thể chuyển thành "Tóc bố tôi đã bạc")

  Tôi tên Mai.

  Vải này khổ hẹp.

  Xe này máy hỏng.

  Mô hình tổng quát:

<Vị ngữ> = <Danh từ/ngữ> <Tính từ/ngữ>

<Vị ngữ> = <Danh từ/ngữ> <Danh từ/ngữ>

- *Trường hợp 8*: Vị ngữ cũng được tổ chức thành chuỗi theo các mối quan hệ nhất định giữa các động từ trong chuỗi. Trong trường hợp này ta sẽ chia ra hai loại chính đó là vị ngữ đồng loại và vị ngữ phức tạp khởi - thuyết.

Trong loại đầu tiên – vị ngữ đồng loại, chúng biểu thị các hành động liên tục, tiếp nối của một chủ thể hành động. Xét ví dụ sau:

Bấy giờ, My ngồi xuống giường, trông ra các cửa sổ lỗ vuông mờ mờ trăng trắng.

Rõ ràng trong ví dụ này, ta có mô hình câu là:

<Câu> = <Trạng ngữ> <Chủ ngữ> <Vị ngữ 1> <Vị ngữ 2>

Các phương tiện biểu hiện mối liên hệ của chuỗi vị ngữ đồng loại này là các từ nối: *"và", "không chỉ ... mà còn", "vừa ... vừa ...", "hoặc ... hoặc", "nếu không ... thì ...",v.v*

Ví dụ:

Cuộc sống của Bác giản dị mà cao thượng.

Với ví dụ này, ta có mô hình câu là:

<Câu> = <Chủ ngữ> <Vị ngữ 1> <Liên từ> <Vị ngữ 2>

- Ở loại vị ngữ tiếp theo là vị ngữ phức tạp khởi - thuyết.

Loại vị ngữ này biểu hiện hoạt động và kết quả của hành động. Vị ngữ do hai bộ phận cấu thành. Bộ phận đầu nêu ra hành động tình trạng, bộ phận sau nêu hệ quả biến hóa liên đới với bộ phận đầu. Chẳng hạn: *"tìm được", "nổi bùng", "bóp nát", "ngồi dậy", v.v.*

Kết cấu phần khởi thông thường chỉ có một yếu tố. Còn phần thuyết có thể có hơn hai yếu tố tạo thành. Đó là một động từ hay tính từ, ví dụ: *"nói nhỏ", "gào thét", "học giỏi", v.v.* Đó là hai động từ hay hai tính từ, ví dụ: *"đứng vùng dậy", "thấp lè tè", v.v.*

Bộ phận thuyết của vị ngữ phức tạp biểu hiện các loại ý nghĩa: ý nghĩa di chuyển trong không gian, ý nghĩa quy kết mục đích, ý nghĩa xu thế, … Các yếu tố của bộ phận này cũng có khả năng độc lập làm vị ngữ. Khi độc lập làm vị ngữ thì quan hệ liên đới khởi – thuyết bị mờ đi. Giữa khởi – thuyết có thể xen thêm bổ ngữ để dễ tách chúng ra. Cho dù mỗi bộ phận có thể có bổ ngữ, nhưng quan hệ giữa chúng là một khối vị ngữ của toàn câu. Trong quá trình từ vựng hóa, các đơn vị trên đây có khả năng thành một trong các kiểu các động từ ghép.

Mô hình tổng quát:

<Vị ngữ> = <Động từ> <Động ngữ>

2.1.1.4 Bổ ngữ

Thành phần phụ đứng trước hay sau một động từ hay tính từ, bổ nghĩa cho động hay tính từ đó, tạo nên cụm từ làm thành phần câu, gọi là bổ ngữ. Sau đây là một số loại bổ ngữ thường gặp:

- **Bổ ngữ tình thái**: thường đứng trước động từ hay tính từ, biểu thị các tình thái khẳng định, thời gian, thể thức diễn biến của hành động và của trạng thái, tính chất, quan hệ,... được nêu ở động từ hay tính từ trung tâm đó.

Bổ ngữ tình thái do các tiểu loại phụ từ tạo thành. Khi cụm từ có bổ ngữ tình thái làm vị ngữ, thì các phụ từ bổ ngữ đồng thời biểu thị các ý nghĩa tình thái vị ngữ, có tác dụng đánh dấu vị ngữ.

Ví dụ:

Hồi còn đi học, Hải // rất say mê âm nhạc.

- **Bổ ngữ đối tượng**: biểu thị các sự vật có quan hệ với động từ hay tính từ trung tâm.

Bổ ngữ đối tượng xuất hiện trong câu do yêu cầu diễn đạt “cái thông báo” và do ý nghĩa của từ trung tâm đòi hỏi hoặc chi phối. Bổ ngữ đối tượng thường do danh từ,

danh ngữ, đại từ tạo thành. Bổ ngữ đối tượng có thể kết nối với động từ hoặc tính từ theo lối trực tiếp (không dùng quan hệ từ) hoặc gián tiếp (có dùng quan hệ từ). Trong phần này, các nhà nghiên cứu tiếp tục chia thành ba loại là bổ ngữ trực tiếp, bổ ngữ gián tiếp và bổ ngữ miêu tả.

Bổ ngữ trực tiếp trả lời cho câu hỏi *"ai?", "cái gì?"*. Nó thường được sử dụng không có giới từ, thường đứng trực tiếp sau vị ngữ và được phản ánh bằng:

- Danh từ, danh ngữ

  Tôi // đã đọc những tờ báo này
- Đại từ

  Tôi // đọc chúng vào buổi sáng
- Mệnh đề

  Cô ta // nói rằng anh ta có thể đến lúc 5 giờ
- Bổ ngữ gián tiếp được phản ánh bằng danh từ hoặc đại từ, trả lời cho câu hỏi kiểu cho *"ai?"*, *"cho cái gì?"*

  Tôi // định đi mua ít đồ cho gia đình.

- **Bổ ngữ miêu tả**: đứng sau động từ, biểu thị cách thức, trạng thái, tính chất, mục đích, nơi chốn,... bổ nghĩa cho động từ hay tính từ trung tâm.

Ví dụ:

Cỏ dại // cao lút đầu.

Bổ ngữ miêu tả do từ hay cụm từ tạo thành. Bổ ngữ miêu tả có thể nối với từ trung tâm bằng quan hệ từ hoặc không dùng quan hệ từ.

2.1.1.5 Các loại ngữ khác trong câu

- **Trạng ngữ**

Trạng ngữ là thành phần của câu được xét trong chỉnh thể của câu nói chung. Trạng ngữ là thành phần phụ biểu thị hoàn cảnh được nêu ở nòng cốt câu. Trạng ngữ do

từ, cụm từ hay kết cấu chủ vị tạo thành. Có thể có các loại trạng ngữ sau: trạng ngữ thời gian, trạng ngữ nơi chốn, trạng ngữ nguyên nhân, trạng ngữ mục đích, trạng ngữ cách thức.

Ví dụ:

Buổi tối anh ấy mới học.

Để học tiếng Anh giỏi bạn phải học chăm chỉ.

Họ, những người nông dân ấy, đang làm việc ở ngoài đồng.

- **Trạng ngữ thời gian**

Phân biệt trạng ngữ thời gian và bổ ngữ thời hạn: Trạng ngữ có nhiều khả năng tự do về vị trí trong câu, còn bổ ngữ thời hạn thì phụ thuộc vào vị từ và chỉ có một số vị trí cố định sau vị từ.

Ví dụ bổ ngữ thời hạn "*năm phút*" trong "*Tôi đợi hắn năm phút*" nếu đảo lên đầu câu thì ta sẽ được kết cấu mới, với "*năm phút*" là từ trung tâm, "*tôi đợi*" là định ngữ, và câu sẽ trở thành không hoàn chỉnh.

Trạng ngữ thời gian thường đặt ở đầu câu và thường có sự tham gia của các giới từ *"vào", "đến", "từ ... đến", "từ ... sang", "hết ... sang", v.v.*

Trạng ngữ thời điểm thường do danh từ hay trạng từ biểu thị nên có thể không cần sự tham gia của giới từ. Ví dụ, *"chiều hôm nay", "ban nãy", "từ sáng đến giờ",...*

- **Trạng ngữ chỉ địa điểm**

Trạng ngữ địa điểm thường đặt ở đầu câu hoặc cuối câu. Khi đặt ở cuối câu, trạng ngữ này thường bắt đầu bằng giới từ *"ở", "về",...*

- **Trạng ngữ chỉ nguyên nhân**

Những trạng ngữ này phần lớn đặt sau vị ngữ.

Ví dụ:

Tôi thất vọng về ông.

Song chúng có thể đảo lên đầu câu hay trước vị ngữ và khi đó thường có trợ từ *"mà"*.

- **Trạng ngữ chỉ mục đích**

Trạng ngữ chỉ mục đích thường bắt đầu với *"để", "vì", "cho*". Trạng ngữ có *"cho"* bao giờ cũng đặt ở cuối câu.

- **Trạng ngữ chỉ phương tiện**

Loại trạng ngữ này bắt đầu với những từ "*bằng", "với", "nhờ", "theo*" và thường đặt ở cuối câu, nhưng cũng có khả năng đảo lên đầu câu và đôi khi xen vào giữa chủ và vị ngữ.

Ví dụ:

Khách toàn đến bằng xe hơi.

- **Trạng ngữ chỉ tình thái**

Trạng ngữ này thường là ngữ động từ. Nó thường đứng trước chủ ngữ và vị ngữ.

Ví dụ:

Ăn cơm xong, San xếp sách vở đi học ngay.

Rồi nghĩ thế nào, nó đứng dậy.

- **Định ngữ:** là thành phần phụ trong câu. Định ngữ được nhận diện thông qua từ mà nó hạn định. Những từ này có thể là thành phần chính (chủ ngữ, vị ngữ) cũng có thể làm thành phần thứ (bổ ngữ). Có loại định ngữ cho cả câu. Quan hệ giữa định ngữ và đối tượng được định ngữ là quan hệ hạn định. Trong câu, danh từ thường có các định ngữ sau:

  - **Định ngữ chỉ lượng** do số từ, đại từ chỉ định, phụ từ tạo thành

    Ví dụ:

    Mười tám cây vạn tuế tượng trưng cho một hàng quân danh dự.

Cả bầy hăng máu phóng như bay.

- **Định ngữ chỉ loại** do danh từ vật thể (danh từ trung tâm có định ngữ là một danh từ chỉ đơn vị tự nhiên hay quy ước) tạo thành

  Ví dụ:

  Mười tám cây vạn tuế tượng trưng cho một hàng quân danh dự.

Định ngữ chỉ loại kết hợp chặt chẽ với danh từ trung tâm, biểu thị sự vật được nêu trong câu.

- **Định ngữ miêu tả** đứng sau danh từ trung tâm (hoặc sau danh từ trung tâm và định ngữ chỉ loại) chỉ các đặc điểm riêng của vật được quy chiếu nêu ở cụm danh từ. Định ngữ miêu tả do từ, cụm từ chính phụ, cụm từ đẳng lập hay cụm chủ vị và các cấu trúc ngữ pháp tương đương tạo thành. Định ngữ miêu tả kết nối trực tiếp hoặc gián tiếp (với danh từ trung tâm) bằng quan hệ từ.

  Ví dụ:

  Những người chủ vườn tốt bụng và hào phóng thấy thế chỉ cười, ánh mắt thích thú nhìn khách.

  Tinh thần thượng võ của cha ông được nung đúc và lưu truyền.

- **Định ngữ chỉ xuất** đứng ở cuối cụm danh từ, kết thúc cụm danh từ. Định ngữ chỉ xuất thường do đại từ chỉ định hoặc danh từ riêng tạo thành. Một số định ngữ miêu tả cũng có thể có tác dụng chỉ xuất sự vật do danh từ trung tâm biểu thị.

  Ví dụ:

  Những em bé H'mông mắt một mí đang chơi đùa trước cửa hàng mậu dịch.

2.1.2 Tập quy tắc cú pháp tiếng Việt cho câu đơn thông thường

Mô hình khởi tạo câu hỏi tự động của chúng tôi có phạm vi trên câu đơn và nếu là câu ghép thì chúng tôi sẽ sử dụng phương pháp tách các vế câu trong câu ghép đẳng lập để đưa về các câu đơn tương ứng, phù hợp với phạm vi nghiên cứu. Như đã biết câu là đơn vị độc lập nhỏ nhất của lời nói, là đơn vị hiện thực của giao tiếp được tạo từ từ và ngữ theo quy luật ngữ pháp và ngữ điệu của một ngôn ngữ. Để nhận biết câu trong một bài viết, chúng ta sẽ thấy là nó bắt đầu bằng một từ được viết hoa, và kết thúc bằng một dấu chấm câu, có thể là chấm, là chấm hỏi, chấm than v.v… Về mặt âm thanh, câu nằm giữa hai khoảng im lặng tương đối dài. Về mặt nghĩa, câu bao giờ cũng diễn đạt một ý trọn vẹn.

Câu có rất nhiều loại, có loại câu bình thường và những loại câu đặc biệt, có khi câu chỉ là một từ (thí dụ: Đi!). Câu được phân loại theo những xu hướng quan niệm và căn cứ lý thuyết khác nhau, chứa từng quan điểm riêng của từng nhà ngôn ngữ, trong tập quy tắc cú pháp tiếng Việt [1] cũng như bài nghiên cứu này thì chúng tôi sẽ lấy cấu trúc chủ - vị làm cơ sở phân loại, bởi vì cách tiếp cận này là thông dụng nhất, phù hợp với việc xây dựng tự động các hệ thống phân tích cú pháp.

Cấu trúc chủ - vị là cấu trúc do thành phần chủ ngữ và vị ngữ tạo nên. Quan hệ ngữ pháp giữa chủ ngữ và vị ngữ là quan hệ chủ - vị. Cấu trúc chủ - vị có chức năng làm thành phần của cụm từ hay câu, hoặc làm nòng cốt câu.

Cấu trúc chủ - vị làm nòng cốt (câu cơ sở) có dạng cô đọng (câu hạt nhân, trong đó mỗi thành phần chủ ngữ, vị ngữ do một từ tạo thành) và dạng mở rộng (câu cơ sở có thành phần mở rộng, mỗi thành phần chủ ngữ, vị ngữ do cụm từ chính - phụ tạo thành, hay thêm thành phần phụ ngoài nòng cốt).

Câu có nòng cốt do hai hay nhiều cấu trúc chủ - vị tạo thành gọi là câu ghép. Trong câu ghép, cấu trúc chủ - vị có tính tự lập về nghĩa và về ngữ pháp, không lệ thuộc

vào nhau. Câu ghép chia thành các kiểu nhỏ gọi tên theo quan hệ giữa các cấu trúc chủ - vị trong nòng cốt.

Ví dụ:

Trận này chưa qua, trận khác đã tới. (câu ghép đẳng lập)

Tuy cuộc đời có thay đổi, nhưng cái lòng ái quốc vẫn còn. (câu ghép chính phụ)

Câu chứa hai hay nhiều cấu trúc chủ - vị nhưng trong đó có một kết cấu chủ - vị làm nòng cốt. Những cấu trúc chủ - vị khác lệ thuộc vào nòng cốt (làm thành phần phụ mở rộng các thành phần chính trong nòng cốt). Loại câu này gọi là câu phức.

Ví dụ:

Con ngựa mà anh nói tới hôm nọ, hôm nay thi được giải nhất.

Tôi nói để mọi người đều biết.

Trong ví dụ trên, thành phần gạch dưới là cấu trúc chủ - vị làm định ngữ (trong cụm danh từ - chủ ngữ), làm bổ ngữ (trong động ngữ làm vị ngữ). Định ngữ và bổ ngữ lệ thuộc vào thành phần nòng cốt câu (cụm danh từ có "*con ngựa*" là thành phần chính, trung tâm động từ có từ "*nói"* (*"để*") làm thành phần chính.

Ngoài phương pháp phân loại theo cấu tạo ngữ pháp cơ bản chia các loại câu thành câu đơn, câu phức và câu ghép, còn có các cách phân loại khác là:

- Phân loại theo mục đích nói: Câu tường thuật, câu nghi vấn, câu mệnh lệnh và câu cảm thán.
- Phân loại theo cấu tạo dạng phủ định: câu phủ định, câu khẳng định.

2.1.2.1 Câu đơn

Trong tiếng Việt, câu đơn là loại câu cơ sở của ngôn ngữ. Câu đơn, ngoài kết cấu chủ - vị hạt nhân, còn được xây dựng bằng những đơn vị khác, bằng các kết cấu khác.

Đó là các câu đơn một tiếng: "*Mưa!*"; câu đơn một từ đa tiết: "*Hải đảo*"; câu đơn một đoản ngữ: "*Một buổi sáng mùa thu*", "*Đêm trắng*"; câu đơn một kết cấu cố định: "*Ý chí kiên cường và phẩm chất cao cả*", v.v.

Câu đơn có thể là câu kể, câu hỏi, câu cầu khiến và trong mỗi thể câu như vậy lại có thể là khẳng định và phủ định.

Việc nhận diện câu đơn phức tạp hơn việc nhận diện câu ghép. Tính phức tạp này trước hết do tính đa dạng của các dấu hiệu bên trong và bên ngoài của câu đơn quyết định. Về dấu hiệu bên ngoài, câu đơn vừa giống các đơn vị nhỏ hơn nó, tham gia cấu tạo nên nó. Trong trường hợp này, các dấu hiệu ngắt về hơi, về các từ tình thái có vai trò quan trọng.

So sánh:

a) *Đêm tháng bảy*. (tổ hợp từ làm một câu)

b) *Trăng sáng*. (tổ hợp từ làm một câu)

c) *Trăng sáng quá!* (câu)

d) *Đêm tháng bảy, trăng sáng quá*. (câu)

Khi đi vào cấu tạo câu với tư cách là phương tiện biểu diễn cấu trúc cú pháp, các đơn vị này làm chức năng thành phần câu. Vì vậy, nghiên cứu tổ chức cú pháp của câu đơn cũng chính là nghiên cứu các thành phần câu (phần mục 2.1.1).

Căn cứ vào chức năng cú pháp, tức là căn cứ vào ý nghĩa cú pháp của các yếu tố cấu trúc nội tại của câu, và căn cứ vào vị trí của chúng mà phân loại thành phần câu. Từ đó xác lập hệ thống thành phần câu.

2.1.2.2 Cấu trúc câu đơn

Ví dụ về câu đơn:

Mẹ về. (câu cơ sở, hạt nhân, câu không mở rộng)

Mẹ anh Nam đã về thành phố. (câu cơ sở, mở rộng)

Hôm qua, mẹ anh Nam đã về thành phố. (câu cơ sở, mở rộng)

- **Mô hình tổng quát câu đơn**
  - Mô hình cơ sở:

    <Câu> = <Chủ ngữ> <Vị ngữ>
  - Mô hình tổng quát:

    <Câu> = <thành phần phụ*1> <Chủ ngữ> <Vị ngữ> <Bổ ngữ*>

    (dấu * có nghĩa là có hoặc không)

- **Câu hai thành phần có vị ngữ danh từ**

Nhóm này có hệ từ cùng với danh từ làm vị ngữ. Mô hình tổng quát của nhóm này là:

<Câu> = <Chủ ngữ> <Vị ngữ> (là <danh từ>)

Ví dụ:

Tôi <u>là sinh viên</u>.

Nhưng, đó cũng <u>chỉ là một cuộc gặp gỡ bất ngờ trong chiến đấu</u>.

Bộ phận vị ngữ có thể mở rộng tùy thuộc vào mối quan hệ của từ trung tâm trong nhóm danh từ làm vị ngữ.

Từ là trong kết cấu vị ngữ nhóm này có chức năng nhấn mạnh vị ngữ, đôi khi có thể bỏ được. Ví dụ "*Cô Hằng là người Hà Nội*", "*là*" trong trường hợp này có thể bỏ được.

- **Câu hai thành phần có vị ngữ tính từ**

Cấu trúc của nó gồm: chủ ngữ là danh từ, đại từ, còn vị ngữ là tính từ có hoặc không có hệ từ.

Mô hình tổng quát của nhóm này là:

<Câu> = <Chủ ngữ> <Vị ngữ> (là <tính từ>)

Ví dụ:

Cô ta <u>thông minh</u>.

Chỉ cô ta là thông minh thôi.

Thuộc vào kiểu này còn có vị ngữ tính từ kết hợp với một số động từ hoặc tổ hợp động từ làm chức năng công cụ cú pháp như kiểu thứ nhất: *"hóa ra", "trở nên", v.v.*

Ví dụ:

Dạo này bà ta đâm ra khó chịu.

Nó trở nên lanh lợi hơn trước.

- **Câu hai thành phần có vị ngữ là danh từ hoặc tổ hợp danh từ không có hệ từ:**

Loại câu này thường biểu thị ý nghĩa địa điểm, sự kiện, hiện tượng, bản chất. Vị ngữ của nó là tổ hợp danh từ với một số từ loại khác. Về nguồn gốc và chức năng, loại này có thể là biến thể của vị ngữ động từ.

Mô hình tổng quát của nhóm này là:

<Câu> = <Chủ ngữ> <Vị ngữ> (<danh từ>)

Ví dụ:

Đồng hồ này ba kim.

Cả nước một lòng.

Các câu trên đây có khả năng ứng với câu mà bộ phận vị ngữ thêm yếu tố có. Loại này có khả năng chấp nhận những biến thể sau đây:

<Câu> = <Chủ ngữ> <Số từ> <Danh từ>

Ví dụ:

Điện cao thế 3 pha

<Câu> = <Chủ ngữ> <Từ so sánh> <Danh từ>

Ví dụ:

Thân em như tấm lụa đào.

<Câu> = <Chủ ngữ> <Đại từ>

Ví dụ:

Ai đấy nhỉ?

<Câu> = <Chủ ngữ> <Loại từ> <Danh từ>

Ví dụ:

Em mấy tuổi rồi?

Đêm mỗi lúc một khuya.

- **Câu hai thành phần có vị ngữ động từ**

Mô hình tổng quát của nhóm này là:

<Câu> = <Chủ ngữ> <Vị ngữ> (<Động từ>)

Nếu động từ vị ngữ là ngoại động thì mô hình có thêm bổ ngữ:

<Câu> = <Chủ ngữ> <Vị ngữ> <Bổ ngữ>

Ví dụ:

Chúng tôi thường viết thư cho nhau.

Câu với động từ sai khiến có thêm mô hình:

<Câu> = <Chủ ngữ 1> <Vị ngữ 1> <Chủ ngữ 2> <Vị ngữ 2>

Ví dụ:

Tôi khuyên nó học tập.

Câu có động từ biểu hiện ý nghĩa cảm nghĩ, hứa hẹn, mong ước. Mô hình:

<Câu> = <Chủ ngữ 1> <Vị ngữ 1> <Vị ngữ 2> <Bổ ngữ>

Ví dụ:

Tôi hứa giúp đỡ anh.

- **Câu bị động**

Là loại câu có động từ với ý nghĩa may rủi, nguyên nhân. Người ta thường dùng các từ "*bị", "được", "do", "bởi", "phải", "mắc*", nhưng thông dụng nhất là các từ "*bị", "được", "phải".*

Ví dụ:

Tôi bị phạt

Tôi được khen.

Tôi được thầy khen.

Bây giờ nhiều việc phải làm lắm.

Tôi sai nó làm việc này. → Nó làm việc này do tôi sai.

- **Câu với phạm trù tồn tại**

Động từ có gắn liền với các sự kiện tồn tại trong một không gian nhất định. Vì vậy, ở câu có ý nghĩa tồn tại, một khi hiện diện các từ không gian, thì có có thể vắng mặt nhưng vẫn được hiểu là có mặt.

Chẳng hạn: "*Bố tôi tóc đã bạc*."

Mô hình:

<Câu> = <Chủ ngữ> <Vị ngữ> (<Danh từ> <Tính từ>)

Vị ngữ của câu là do danh từ và tính từ tạo thành. Nhưng danh từ ("*tóc*") ở vị ngữ là sự vật gắn liền, tồn tại trong chủ ngữ ("*bố tôi*") và là yếu tố sở hữu bất khả ly.

Câu này có thể cải biến thành: "*Tóc bố tôi đã bạc*."

Cho nên câu tồn tại tiếng Việt không những được biểu hiện bằng vị ngữ động từ có, ở mà còn bằng tổ hợp danh từ với tính từ mang ý nghĩa tồn tại.

Ví dụ:

Mẹ tôi tính tình hiền lành.

Em gái tôi tóc đen và dài.

Các câu này có các câu tương ứng:

Tính tình của mẹ tôi hiền lành.

Tóc em gái tôi đen và dài.

- **Câu hai thành phần có vị ngữ tổ hợp từ cố định**

Chủ ngữ của loại câu này có thể là danh từ, động từ và tổ hợp từ khác nhau, còn vị ngữ là tổ hợp từ cố định có tính thành ngữ hoặc quán ngữ. Câu đơn loại này biểu hiện ý nghĩa so sánh đặc trưng, tính chất vốn có của hiện tượng, sự vật được nêu ra ở chủ ngữ.

Mô hình:

<Câu> = <Chủ ngữ> <Vị ngữ> (<Thành ngữ>)

Ví dụ:

Tình hình nghìn cân treo sợi tóc.

Lão ta đục nước béo cò.

Tuy hoàn cảnh rất khó khăn, Đảng và Chính phủ đã trông xa thấy rộng.

Thành ngữ làm vị ngữ là tổ hợp có danh từ hoặc động từ làm trung tâm. Khi mang nghĩa định danh, tính chất thì dùng tổ hợp thành ngữ danh từ, khi mang nghĩa hành vi, hoạt động, quá trình thì dùng tổ hợp thành ngữ động từ.

2.1.3 Tập quy tắc cú pháp tiếng Việt cho câu đơn đơn đặc biệt

Trong tiếng Việt, có những câu chúng ta không thể xác định được thành phần của nó, đâu là chủ ngữ, đâu là vị ngữ, vì chúng được cấu tạo bằng một từ hay một cụm từ. Những câu như vậy gọi chung là câu đặc biệt. Câu đơn đặc biệt được làm thành từ một từ, một cụm từ, cũng có thể có trung tâm cú pháp phụ đi kèm làm thành phần phụ của câu. Câu đơn đặc biệt chỉ có một thành phần (không phải là chủ ngữ hoặc vị ngữ) làm trung tâm cú pháp chính. Vì không thể xác định rõ ràng cấu trúc chủ - vị, nòng cốt trong câu đơn đặc biệt không bao gồm hai thành phần nên chúng ta sẽ khó có thể xác nhận thông tin có thể hỏi được bằng phương pháp phân tích cú pháp của câu. Phương pháp dựa trên phân tích cú pháp yêu cầu một câu có nòng cốt rõ ràng, minh bạch, tiện cho

nhiệm vụ xác định chủ ngữ, vị ngữ trong câu. Nên trong phạm vi bài nghiên cứu này, chúng tôi sẽ không phân tích cú pháp của câu đơn đặc biệt để tạo câu hỏi, thay vào đó sẽ dựa vào các đặc điểm nhận dạng của câu đơn đặc biệt và tránh tạo câu hỏi trên chúng.

Câu đơn đặc biệt có thể phân loại theo mục đích sử dụng:

2.1.3.1 Câu gọi, đáp

Câu gọi đáp dùng làm lời gọi hay lời đáp. Câu gọi đáp do thán từ gọi đáp, danh từ chỉ người, vật, … hoặc kết hợp danh từ + thán từ để gọi đáp (phân biệt với thành phần phụ gọi, đáp trong câu đơn bình thường và các kiểu câu khác, khi câu gọi đáp đứng độc lập, riêng rẽ trong một ngữ cảnh có dấu hiệu tách biệt với câu khác).

Ví dụ:

Mẹ!

Bà ơi?

Vâng.

Nam thân mến!

Phân biệt với câu cảm thán dùng để biểu thị hay bộc lộ cảm xúc. Câu cảm thán do thán từ, từ ngữ biểu thị cảm xúc, hay kết hợp từ ngữ với thán từ biểu thị cảm xúc … tạo thành (phân biệt với thành phần phụ cảm thán đứng độc lập trong một ngữ cảnh có dấu hiệu tách biệt với câu khác).

Ví dụ:

Ối!

Chết rồi!

Mô hình tổng quát:

<Câu> = <Thán từ gọi đáp> !

<Câu> = <Danh từ> !

<Câu> = <Danh từ> <Thán từ gọi đáp> !

2.1.3.2 Câu tồn tại

Câu tồn tại gồm hai kiểu nhỏ:

- **Câu tồn tại danh từ** có trung tâm cú pháp chính là một danh từ hay một cụm danh từ, một đại từ biểu thị sự vật tồn tại.

Ví dụ:

Tháng giêng. Mạc Tư Khoa tuyết trắng.

Chân đèo Mã Phục.

Nước!

Bộ Giáo dục và Đào tạo.

Mô hình tổng quát:

<Câu> = <Thành phần phụ*> <Danh từ/ngữ>

- **Câu tồn tại - động từ (tính từ)**

Câu tồn tại động từ/tính từ có trung tâm cú pháp chính là động từ, tính từ hay động ngữ, tính ngữ.

Ví dụ:

Chửi. Kêu. Đấm. Đá. Thụi. Bịch.

Còn gạo.

Đã có xe.

Hết cảnh đầu.

Sao mà lâu thế.

Lâu quá!

Mô hình tổng quát:

<Câu> = <Thành phần phụ*> <Động từ/ngữ>

<Câu> = <Thành phần phụ*> <Tính từ/ngữ>

2.1.3.3 Câu rút gọn

Câu rút gọn hay còn gọi là câu tỉnh lược là câu có một hay một số thành phần chính (chủ ngữ hoặc vị ngữ, hoặc cả hai) được được rút gọn. Câu rút gọn dựa trên một tiêu chí cơ bản là ngữ cảnh. Khi ngữ cảnh giao tiếp cho phép, chúng ta có thể rút gọn một phần của câu mà không ảnh hưởng đến việc hiểu nghĩa của câu. Câu rút gọn khác với câu một thành phần ở chỗ người ta có thể dựa vào hoàn cảnh ngôn ngữ mà điền vào đó thành phần đã bị bớt đi, và khôi phục lại câu hoàn chỉnh. Trái lại, câu đơn phần thì hoặc là không tiếp nhận một yếu tố nào khác, hoặc là chỉ tiếp nhận những yếu tố có ý nghĩa mơ hồ, không xác định.

Ví dụ:

Buồn ngủ quá! Đi ngủ nào! (Câu rút gọn chủ ngữ)

Mời chị vào công an với tôi. (Câu rút gọn chủ ngữ)

Anh ấy đói còn tôi thì không. (Câu rút gọn vị ngữ)

Họ chẳng có một tí gì. Đồ đạc không. Hòm xiểng không. (Câu rút gọn vị ngữ)

Mô hình tổng quát:

<Câu> = <Vị ngữ> (Câu rút gọn chủ ngữ)

<Câu> = <Chủ ngữ> (Câu rút gọn vị ngữ)

2.2 Các kỹ thuật cơ bản trong xử lý ngôn ngữ tự nhiên

2.2.1 Sentence Segmentation (Phân tách câu văn)

Sentence segmentation là quá trình xác định và phân tách 1 đoạn văn bản bao gồm 1 hoặc nhiều từ thuộc một ngôn ngữ nào đó thành các câu văn bằng cách xác định điểm đầu và điểm kết thúc của câu văn trong đoạn văn bản đó

Quá trình này liên quan tới việc xác định giới hạn của câu (*sentence boundaries*) giữa các từ có trong đoạn văn bản đó [4]. Thông thường, ở hầu hết các ngôn ngữ, các câu văn này có thể được nhận dạng và xác định nhờ vào các dấu câu cơ bản như *".", ",", "!", "?", v.v*, đánh dấu giới hạn của 1 câu, chẳng hạn như ví dụ sau:

Ví dụ:

Mưa lâm thâm, gió trở lạnh, bầu trời u ám

=> *Mưa lâm thâm*

Gió trở lạnh

Bầu trời u ám

Ông nội đến. Mọi người ra đón ông

=> *Ông nội đến*

Mọi người ra đón ông

Việc nhận dạng câu thông qua các dấu câu khiến cho kỹ thuật này thường được biết đến với các tên gọi khác nhau như *Sentence boundary disambiguation (SBD), sentence boundary detection* hay *sentence boundary recognition* [4].

Trên thực tế, việc phân tách các câu văn còn phải dựa trên các yếu tố khác như ngữ cảnh, cách diễn đạt cũng như cách sử dụng các dấu câu, để đảm bảo độ chính xác về mặt ngữ nghĩa của câu. Tuy nhiên, kỹ thuật phân tách câu văn đóng vai trò quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên, là nền tảng cho các kỹ thuật xử lý khác như *gán*

nhãn từ vựng (part of speech), tóm tắt văn bản (text summarization), nhận dạng và phân loại thực thể (named entity recognition), v.v

2.2.2 Word Tokenization (Tách từ)

Word Tokenization (hay còn được gọi với cái tên *Word Segmentation*) là quá trình xác định giới hạn của các từ trong câu văn cũng như xác định loại từ (từ đơn, từ ghép, v.v) có trong câu. Đây là kỹ thuật cơ sở trong lĩnh vực xử lý ngôn ngữ tự nhiên. Khác với tiếng Anh, một từ tiếng Việt với ngữ nghĩa riêng có thể được tạo bởi nhiều hơn 1 âm, điều đặc biệt là các âm này có thể đứng độc lập tạo thành 1 từ đơn, có ngữ nghĩa riêng của nó. Ví dụ: từ "*cá nhân*" được tạo bởi 2 âm "*cá*" và "*nhân*", 2 âm này cũng là 2 từ đơn mang 1 ý nghĩa khác

Ngoài ra, theo Nghia Ticy và Hung Le [6], với ngôn ngữ tiếng Việt, thuộc loại hình đơn lập, mang đặc điểm là không biến đổi hình thái, ranh giới từ không được xác định mặc nhiên ngoài khoảng trắng. Tiếng Việt có đặc điểm là ý nghĩa ngữ pháp nằm ở ngoài từ, phương thức ngữ pháp chủ yếu là trật tự từ và từ hư. Cho nên có trường hợp một câu có thể có nhiều ngữ nghĩa khác nhau tùy vào cách ta tách từ như thế nào, gây nhập nhằng về ngữ nghĩa của câu.

Ví dụ:

Xoài phun thuốc sâu không ăn

=> *Xoài / phun thuốc / sâu / không / ăn.*

Xoài / phun / thuốc sâu / không / ăn.

Ăn cơm không được uống rượu

=> *Ăn / cơm / không / được / uống / rượu.*

Ăn / cơm không / được / uống / rượu.

Mẹ vào ca ba con ngủ với dì

=> *Mẹ / vào / ca ba / con / ngủ / với / dì.*

Mẹ / vào ca / ba con / ngủ / với / dì.

Vấn đề này tưởng chừng đơn giản với con người nhưng đối với máy tính, đây là bài toán rất khó giải quyết. Vì thế để giải quyết bài toán này, cho đến nay đã có nhiều phương pháp, hướng tiếp cận khác nhau. Các phương pháp và hướng tiếp cận này đã được áp dụng thành công cho các ngôn như như tiếng Anh, tiếng Trung, tiếng Nhật, tiếng Thái, v.v. Trong bài báo cáo "*Sự ảnh hưởng của phương pháp tách từ trong bài toán phân lớp văn bản tiếng Việt*" của Phạm Nguyên Khang và các cộng sự, có đề cập tới 3 hướng tiếp cận sau:

- **Tiếp cận dựa trên bộ từ điển cố định (*dictionary-based approach*):**

Một cách đơn giản để tách từ đó là scan từng từ một trong văn bản từ trái sang phải và xác thực các từ này bằng cách tìm kiếm chúng này trong từ điển. Tuy nhiên sẽ có nhiều vấn đề xảy ra, chẳng hạn như ví dụ ở trên, từ "*cá nhân*" có thể được ghép bởi 2 từ đơn "*cá*" và "*nhân*", khiến cho việc xác định cũng như quá trình tách từ không đạt được độ chính xác cao. Vì thế một số giải thuật được sử dụng để giải quyết vấn đề này, trong đó có thể kể đến 2 giải thuật đó chính là thuật toán *so khớp từ dài nhất (longest matching)* và *so khớp từ cực đại (maximum matching)*

  - **Phương pháp so khớp từ dài nhất (*longest matching*)** : với mỗi câu, duyệt từ trái qua phải các âm tiết trong câu, kiểm tra xem có nhóm các âm tiết có tồn tại từ trong từ điển hay không. Chuỗi dài nhất các âm tiết được xác định là từ sẽ được chọn ra. Tiếp tục thực hiện việc so khớp cho đến hết câu [5]

Ví dụ:

Học sinh học sinh vật học

=> Xét từ trái qua phải, âm tiết đầu tiên là "*học*", "*học*" cũng có thể là 1 từ đơn, nhưng "*học*" cũng có thể kết hợp với âm tiết "*sinh*" để tạo nên từ ghép "*học sinh*", ta

được từ đầu tiên là "*học sinh*", xét tiếp các âm tiết còn lại cho đến khi hết câu ta có các từ sau: "*học sinh*", "*học sinh*", "*vật*", "*học*". Tuy nhiên, ta có thể thấy rằng phương pháp này không đem lại kết quả như mong muốn

- **Phương pháp so khớp cực đại (*maximum matching*):** ứng với mỗi câu dữ liệu đầu vào, tìm tất cả các trường hợp mà các âm tiết có thể kết hợp lại để tạo nên các từ có nghĩa. Ứng với mỗi loại ngôn ngữ khác nhau thì sự lựa chọn các nhóm âm tiết này có thể khác nhau. Phương pháp này là so khớp toàn diện cho một câu thay vì so khớp cục bộ âm tiết đang được xét. [5]

Ví dụ:

Học sinh học sinh vật học

=> Các trường hợp kết hợp của các âm tiết có thể có "*sinh vật học*", "*học sinh*", "*học*", từ được tách trong câu sẽ chính xác hơn phương pháp so khớp từ dài nhất.

- **Tiếp cận thống kê sử dụng học máy (machine learning approach)**: với cách tiếp cận này, các phương pháp cho việc tách từ thông thường dựa trên mô hình ngôn ngữ (language model - LM). Một số phương pháp có thể kể đến như:
  - **Phương pháp tách từ sử dụng mô hình Markov ẩn (*Hidden Markov Models- HMM*)**
  - **Phương pháp tách từ sử dụng mô hình trường xác suất có điều kiện (*CRFs*) và độ hỗn loạn cực đại (*Maximum Entropy*)**
  - **Phương pháp tách từ sử dụng mô hình *Pointwise***

- **Tiếp cận lai (tiếp cận dựa trên cả 2 phương pháp trên, hay Hybrid approach):** để có thể tận dụng được những ưu nhược điểm của 2 hướng tiếp

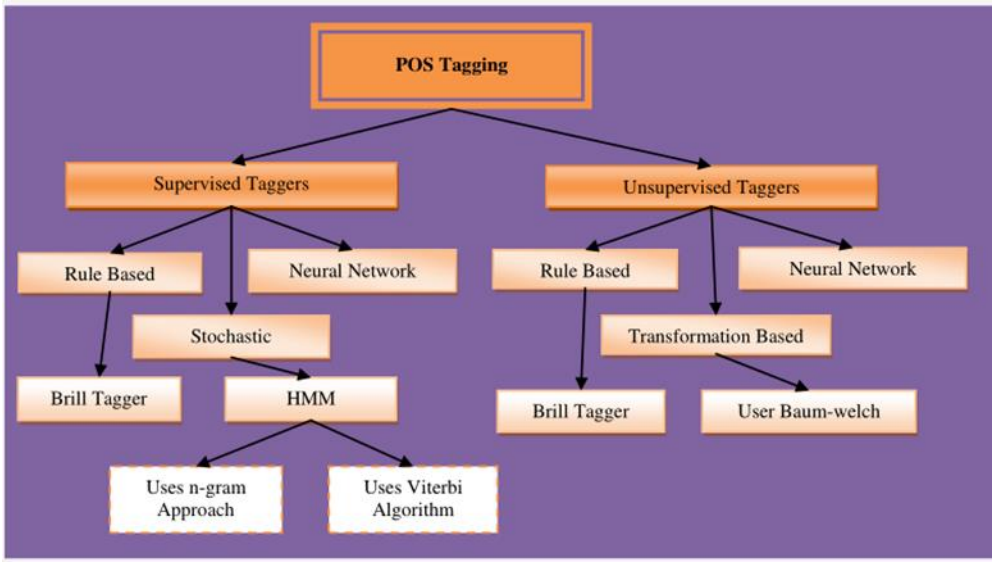
cận kể trên, phương pháp tiếp cận lai đã được đề xuất, kết hợp giữa tiếp cận dựa trên từ điển và tiếp cận thống kê sử dụng học máy.

Hầu hết các kỹ thuật xử lý trong lĩnh vực xử lý ngôn ngữ tự nhiên hiện nay đều dựa trên thành phần cơ bản là từ, vì thế có thể nói việc phân chia từ trong văn bản rất quan trọng đối với 1 ngôn ngữ. Kỹ thuật này hiện nay trở thành 1 bước tiền xử lý văn bản quan trong của hầu hết các hệ thống xử lý ngôn ngữ tự nhiên.

2.2.3 Part of Speech (Gán nhãn từ vựng)

Gán nhãn từ vựng (hay Part of Speech, viết tắt là POS) là việc xác định các chức năng ngữ pháp của từ trong câu hay là quá trình gán từng từ trong đoạn văn bản với các đánh dấu từ loại hoặc cấu trúc ngữ pháp [7]. Thông thường, một từ có thể có nhiều chức năng ngữ pháp. Chẳng hạn: trong câu "con ngựa đá đá con ngựa đá", cùng 1 từ "đá" nhưng từ thứ nhất và từ thứ ba giữ chức năng ngữ pháp là danh từ, nhưng từ thứ 2 lại là động từ trong câu. Trong đề tài

Để gán nhãn POS cho các từ trong 1 câu thông thường người ta sẽ sử dụng 1 bộ phân tích được gọi là POS tagger, là 1 dạng phần mềm hỗ trợ việc gán nhãn từ vựng cho từng từ theo 1 ngôn ngữ nào đó [8]. Các bộ phân tích này có thể sử dụng các kỹ thuật POS Tagging khác nhau. Theo Deepika Kumawat và Vinesh Jain, các kỹ thuật được chia chủ yếu thành 3 hướng tiếp cận chính, thuộc vào 2 phương pháp gán nhãn giám sát (*Supervised Tagger*) và gán nhãn không giám sát (*Unsupervised Tagger*):



Hình 2.2 Các phương pháp gán nhãn cũng như hướng tiếp cận

- **Gán nhãn dựa trên luật (*Rule-based tagging*):**

Gán nhãn dựa trên luật sử dụng từ điển để tìm các từ loại có thể cho động từ, sử dụng các luật làm thành 1 nghĩa. Hướng tiếp cận này sử dụng thông tin ngữ cảnh để gán các nhãn cho các từ chưa biết hoặc các từ nhập nhằng [7]. Ví dụ, trong tiếng Anh, nếu từ đứng trước là một mạo từ, thì từ được đề cập phải là 1 danh từ. Một trong số các kĩ thuật áp dụng hướng tiếp cận dựa trên luật có thể kể đến là *Brill's tagger*.

- **Gán nhãn sử dụng trường xác suất ngẫu nhiên (*Stochastic tagging*):**

Phương pháp này gán nhãn POS dựa trên tần suất, xác suất và thống kê xảy ra của một chuỗi nhãn cụ thể. 2 mô hình phổ biến sử dụng hướng tiếp cận này đó là *Conditional Random Fields (CRFs)* và *Hidden Markov Models (HMMs)*

- **Gán nhãn sử dụng phương pháp học máy**:

Áp dụng mô hình học máy sử dụng mạng neuron để gán nhãn từ vựng

2.2.4 Chunking (Xác định cụm từ)

Chunking (hay còn gọi là shallow parsing) là quá trình trích xuất các cụm từ (chunk) từ một văn bản không có cấu trúc (*unstructured text*). Hoạt động dựa trên *POS*

tagging, nhưng thay vì chỉ hỗ trợ xác định và phân loại từ (danh từ, động từ, tính từ, v.v...), các chunk cung cấp manh mối về cấu trúc của câu văn hoặc cấu trúc của một cụm từ trong câu văn (cụm danh từ *NP*, cụm động từ *VP*, cụm tính từ *AP*, v.v).

Ví dụ:

Bác sĩ bây giờ	có thể	thản nhiên	báo tin bệnh nhân	bị ung thư?
NP	*VP*		*NP*	*VP*

Ngoài ra, các chunk này còn có thể cung cấp thông tin một cách rõ ràng hơn thông qua định dạng IOB, biểu thị phần bắt đầu, phần bên trong của một chunk, cũng như kết thúc của 1 câu văn.

Lấy ví dụ ở trên:

Bác sĩ	bây giờ	có thể	thản nhiên	báo tin	bệnh nhân	bị	ung thư	?
B-NP	*I-NP*	*B-VP*	*I-VP*	*B-NP*	*I-NP*	*B-VP*	*I-VP*	*O*

2.2.5 Named Entity Recognition (Nhận dạng thực thể có tên)

Nhận dạng thực thể có tên (Named Entity Recognition, viết tắt là NER) đôi khi còn được biết đến là công cụ xác định, trích xuất thực thể, là quá trình xác định và phân loại thực thể có tên trong văn bản. Một thực thế có thể là 1 hoặc 1 chuỗi từ cùng đề cập tới 1 đối tượng nào đó (mốc thời gian, tên địa danh, tên người, v.v). NER đóng vai trò quan trọng vì nó là bài toán đơn giản nhất để giải quyết các bài toán phức tạp hơn trong lĩnh vực xử lý ngôn ngữ tự nhiên.

Ví dụ:

Bệnh viện bệnh Nhiệt đới TP.HCM	xét nghiệm dương tính lần một đêm	12/3
ORG		*DAT,*

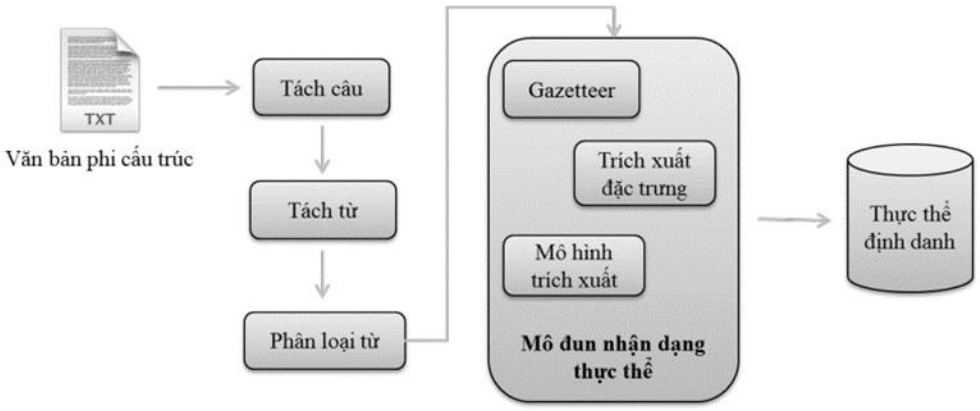
Với *ORG - ORGANIZATION* (tổ chức), *DAT - DATE* (mốc thời gian) [11]

Bệnh nhân “669” là bác sĩ làm việc tại Bệnh viện Đa khoa Đồng Nai

PAT *OCC* *LOC* ,

Với *PAT* - *PATIENT* (bệnh nhân), *OCC* - *OCCUPATION* (nghề nghiệp), *LOC* - *LOCATION* (địa điểm) [11]

Bài toán nhận dạng thực thể thường được chia thành hai quy trình liên tiếp: nhận dạng thực thể và phân loại thực thể. “Nhận dạng thực thể” là quá trình tìm kiếm các đối tượng được đề cập tới trong văn bản trong khi “Phân loại thực thể” là việc gán nhãn cho các đối tượng đó [9]. Mô hình mô tả quy trình nhận dạng thực thể có tên được Phạm Thị Thu Trang [9] trình bày như sau:



Hình 2.3 Quy trình nhận dạng thực thể có tên

Theo Lhioui Chahira, Zouaghi Anis, Zrigui Mounir [10], có 3 hướng tiếp cận chính để giải quyết bài toán này, trong đó bao gồm:

- **Tiếp cận dựa trên luật (Rule-based approach):**

Cách tiếp cận này thường được sử dụng bởi hầu hết các hệ thống phân loại thực thể, sử dụng *internal evidence* - danh sách các tên đã được phân thành các loại thực thể từ trước bởi các chuyên gia ngôn ngữ học (hay còn gọi là các *gazetteers*), kết hợp với

external evidence - các yếu tố ngữ cảnh trong câu văn có quan hệ cú pháp với thực thể chỉ định để nhận diện và phân loại các thực thể định danh

- **Tiếp cận thống kê áp dụng phương pháp học máy (Learning approach):**

Các hệ thống dựa trên hướng tiếp cận này sử dụng phương pháp trường xác suất ngẫu nhiên cũng như học trên 1 dữ liệu huấn luyện cho trước. Giải thuật học máy tiếp đến sẽ áp dụng các thực thể định danh có sẵn này để phát triển nên 1 thực thể mới sử dụng các mô hình thống kê, bao gồm: *Hidden Markov Model (HMM), Support Vector Machine (SVM), Conditional Random Fields (CRFs), v.v*

- **Tiếp cận lai (Hybrid approach):**

Hướng tiếp cận này tận dụng cả 2 phương pháp trước đó nhằm khắc phục nhược điểm của mỗi phương pháp, sử dụng các luật cũng như thuật toán học máy để huấn luyện dữ liệu nhằm cải thiện độ chính xác trong việc nhận dạng và phân loại thực thể

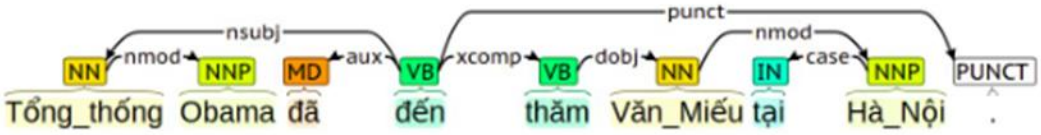
2.2.6 Dependency parsing (Phân tích cú pháp phụ thuộc)

Theo Luong Nguyen Thi và các cộng sự [12], Phân tích cú pháp phụ thuộc là quá trình xây dựng biểu đồ phụ thuộc (*dependency graph*) cho 1 câu văn đầu vào được cho trước. Đầu vào của bài toán là một câu văn đã được phân tách từ cũng như gán nhãn từ vựng. Hầu hết các hệ thống và nghiên cứu về phân tích cú pháp phụ thuộc đều sử dụng phương pháp học máy là chủ yếu, phân tích quan hệ phụ thuộc giữa các từ trong câu với nhau thay vì phân tích chủ ngữ, vị ngữ, các cụm danh từ, cụm động từ, v.v.

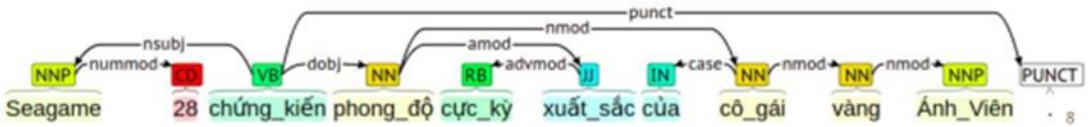
Cấu trúc phụ thuộc là một quan hệ phụ thuộc thể hiện bằng 1 mũi tên có hướng, trong đó:

- Phần có mũi tên là *dependent* (có thể là *modifier, subordinate*, …)
- Phần còn lại là *head* (*governor, regent*, v.v)
- Nhãn phụ thuộc tương ứng giữa 2 từ

Ví dụ:



Hình 2.4 Cấu trúc quan hệ phụ thuộc

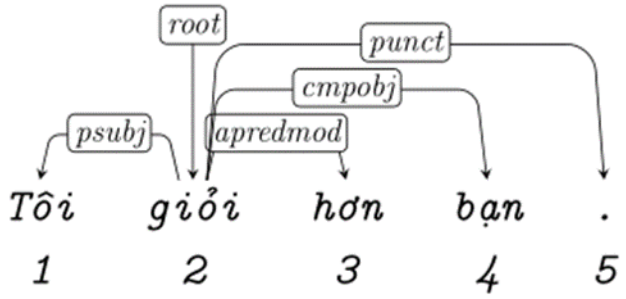


Hình 2.5 Cấu trúc quan hệ phụ thuộc

Nhãn quan hệ phụ thuộc (hay nhãn phụ thuộc) thể hiện sự phụ thuộc giữa hai từ trong câu với nhau. Mỗi cặp từ loại khác nhau, ở những vị trí khác nhau thì sẽ có tên quan hệ phụ thuộc là khác nhau. Đây là cách làm tốt nhất và hiệu quả nhất để hiểu được mối quan hệ giữa 2 từ. Có nhiều bộ nhãn quan hệ dùng cho một ngôn ngữ và độ chi tiết giữa các bộ nhãn là khác nhau. Chẳng hạn như tập nhãn phụ thuộc Stanford (*Stanford Dependencies*), được xây dựng dựa trên tập nhãn phụ thuộc đa ngôn ngữ (*Universal Dependencies*). Theo Hà Mỹ Linh, tác giả của luận văn "*Phân tích cú pháp phụ thuộc tiếng việt*" [13], tất cả các quan hệ phụ thuộc đó đều là quan hệ 2 ngôi: giữa một từ trung tâm và từ phụ thuộc của nó. Mỗi mối quan hệ được đưa ra bởi 3 thành phần chính: tên quan hệ phụ thuộc, từ trung tâm và từ phụ thuộc. Ngoài ra, cùng với *Viettreebank*, tác giả và các cộng sự cũng đã xây dựng một bộ nhãn phụ thuộc dành riêng cho tiếng Việt [13], bao gồm 46 nhãn với các nhãn trùng với các nhãn phụ thuộc đa ngôn ngữ và 1 số nhãn mới như:

- **smpobj**: quan hệ so sánh, mô tả liên hệ so sánh "*hơn", "kém", "nhất",* v.v với các danh từ đi sau

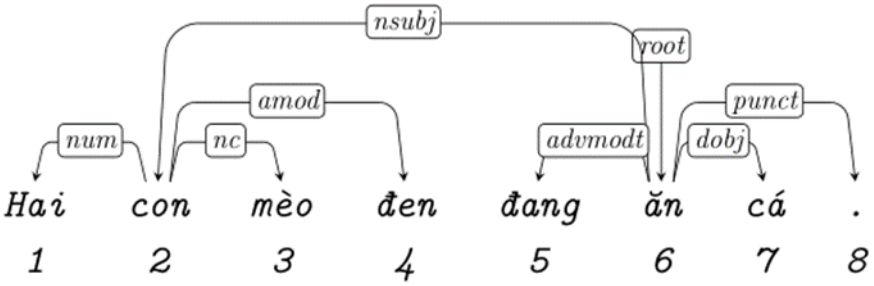
– *Tôi giỏi hơn bạn* → **cmpobj**(*giỏi, bạn*)



Hình 2.6 Nhãn smpobj

- **nc**: bổ nghĩa cho danh từ chỉ loại, các danh từ này luôn đứng trước danh từ chung chẳng hạn như "*cái", "con*", v.v.

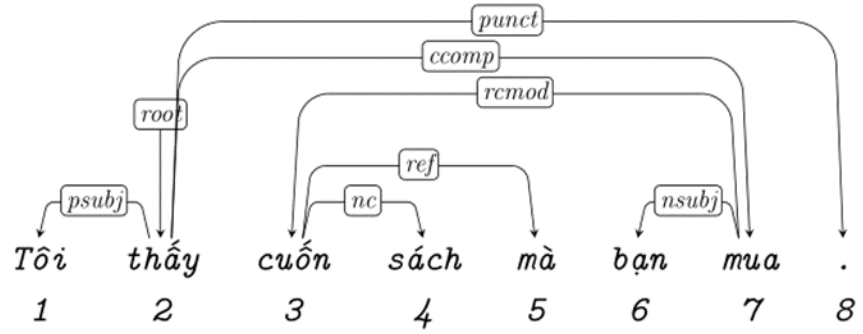
– *Hai con mèo đen đang ăn cá.* → **nc**(*con, mèo*)



Hình 2.7 Nhãn nc

- **ref**: tham chiếu - từ quan hệ liên kết mệnh đề quan hệ bổ nghĩa cho cụm danh từ được, chỉ được

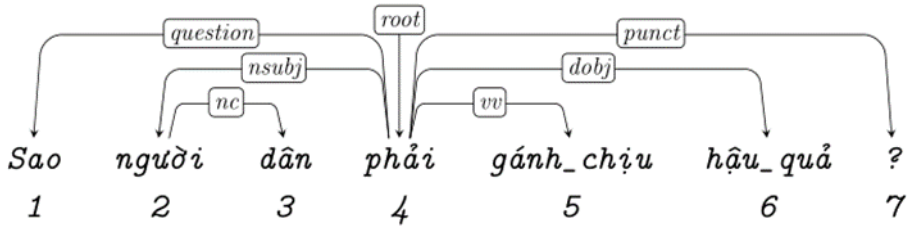
– *Tôi nhìn thấy cuốn sách mà bạn mua.* → **ref**(*cuốn, mà*)



Hình 2.8 Nhãn ref

- **question**: từ để hỏi

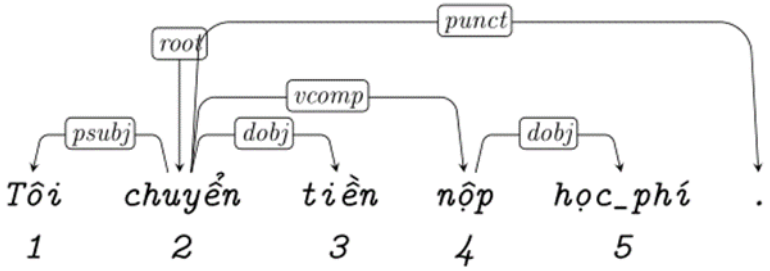
– *Sao người dân phải gánh chịu hậu quả* → **question**(*phải, sao*).



Hình 2.9 Nhãn question

- **vcomp**: bổ ngữ động từ của động từ, được sử dụng để chỉ định quan hệ giữa động từ chính và động từ phụ

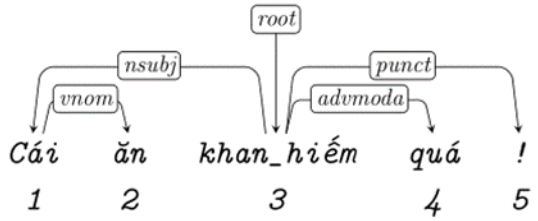
– *Tôi chuyển tiền nộp học phí.* → **vcomp**(*chuyển, nộp*)



Hình 2.10 Nhãn vcomp

- **vnom**: danh từ hóa động từ - thường là 1 từ chỉ loại đứng trước đó (“*cái”, “sự”, “việc”, v.v)*

– *Cái ăn khan hiếm quá!* → **vnom**(*cái, ăn*)

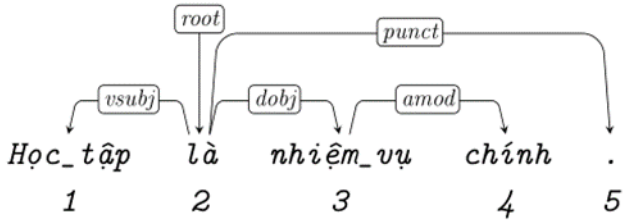


– *Sự học ngày càng phát triển trên quê hương tôi.* → **vnom**(*sự, học*)

Hình 2.11 Nhãn vnom

- **vsubj**: chủ ngữ động từ, mô tả hiện tượng động từ làm chủ ngữ

– *Học tập là nhiệm vụ chính* → **vsubj**(*là, học tập*)



– *Nói chuyện với họ chán phè* → **vsubj**(*chán phè, Nói chuyện*)

– *Viết tiểu thuyết đã trở thành hẳn một nghề riêng* → **vsubj**(*trở thành, viết*)

Hình 2.12 Nhãn vsubj

- **xsubj**: chủ ngữ kiểm soát, mô tả quan hệ giữa chủ ngữ của câu và 1 vị từ

– *Tôi thích ăn thịt* → **xsubj**(*ăn, tôi*)

– *Tôi phải đi ngay đây* → **xsubj**(*đi, tôi*)

Hình 2.13 Nhãn xsubj

Bảng mô tả đầy đủ các nhãn trong tập nhãn phụ thuộc tiếng Việt cũng như so sánh giữa chúng với tập nhãn phụ thuộc đa ngôn ngữ (*Universal Dependencies*) và tập nhãn phụ thuộc tiếng Anh (*Stanford Dependencies*) cũng được tác giả Hà Mỹ Linh trình bày dưới đây [13]:

UD (2015)	SD (2015)	Tiếng Việt (2015)	Ý nghĩa
nsubj	nsubj	nsubj, asubj, psubj	Chủ ngữ của câu là các cụm danh từ, tính từ, đại từ
csubj	csubj	csubj	Mệnh đề là chủ ngữ của câu
dobj	dobj	dobj	Tân ngữ trực tiếp của câu
iobj	iobj	iobj	Tân ngữ gián tiếp của câu
ccomp	ccomp	ccomp	Một mệnh đề bổ ngữ cho động từ
-	acomp	acomp	Bổ ngữ tính từ
-	attr	question	Từ để hỏi
advmod	advmod quantmod tmod	advmoda advmodb advmodt quantmod tmod	Bổ ngữ là trạng từ
neg	neg	neg	Phủ định
det	det, predet	det, predet	Từ hạn định của danh từ
amod	amod	amod, acomp, apredmod	Bổ nghĩa tính từ của danh từ
nummod	num	num	Từ chỉ số lượng
appos	appos	appos	Phần chêm vào của câu
acl/relcl	rcmod	rcmod	Bổ nghĩa là mệnh đề quan hệ
root	root	root	Gốc của câu
punct	punct	punct	Dấu câu
auxpass	auxpass	auxpass	Động từ chỉ nghĩa bị động
dep	dep	dep	Quan hệ tổng quát
case	prep	prep	Bổ nghĩa là giới từ
nmod	pobj	pobj	Tân ngữ của giới từ
ncmod	pcomp	pcomp	Bổ ngữ của giới từ là động từ hoặc một mệnh đề
compound	nn, number	nn, number	Bổ nghĩa cho danh từ
conj	conj	conj	Giới từ liên hợp
cc	cc	cc	Phần liên hợp
parataxis	parataxis	parataxis	Liên kết giữa các thành phần đẳng lập
mark	mark	mark	Từ giới thiệu một mệnh đề phụ
nsubjpass	nsubjpass	nsubjpass	Chủ ngữ danh từ bị động
xcomp	xcomp	vv	Bổ ngữ động từ của vị từ
csubjpass	csubjpass	-	Mệnh đề làm chủ ngữ bị động của câu
advcl	advcl	-	Mệnh đề trạng ngữ bổ nghĩa cho câu
aux	aux	-	Trợ động từ trong câu
cop	cop	-	Quan hệ giữa hệ từ và root trong câu
expl	expl	-	Đại từ phản thân
discourse	discourse	parataxis?	Phần nói thêm vào, chào hỏi.
vocative	vocative	-	Quan hệ về xưng hô
name	mwe	-	Quan hệ giữa các từ trong tên riêng
goeswith	goeswith	-	Các cụm từ, các từ thường đi cùng nhau
foreign	-	-	Từ gốc là từ nước ngoài
list	-	-	Danh sách liệt kê
remnant	-	-	Quan hệ tỉnh lược
reparandum	-	-	Quan hệ sửa sai
dislocated	-	-	
-	-	cmpobj	Quan hệ so sánh
-	-	nc	Danh từ chỉ loại
-	ref	ref	Nhãn tham chiếu
-	-	vcomp	Bổ ngữ động từ của động từ
-	-	vsubj	Động từ làm chủ ngữ của câu
-	xsubj	xsubj	Chủ ngữ kiểm soát

Bảng 2.1 Bảng so sánh các nhãn giữa tập nhãn phụ thuộc tiếng Việt so với tập nhãn phụ thuộc đa ngôn ngữ (Universal Dependencies) và tập nhãn phụ thuộc tiếng Anh (Stanford Dependencies)

Để có thể tạo được những câu hỏi tốt ở mức câu thì việc trích xuất câu phù hợp trong văn bản là rất quan trọng. Thứ nhất, câu được chọn để tạo câu hỏi phải chứa thông tin có thể hỏi được, bên cạnh đó câu được trích xuất phải chuẩn về mặt cú pháp – tuân theo các tập quy tắc cú pháp của tiếng Việt [1]. Như vậy về mặt thông tin của câu, tiêu chí của chúng tôi là chọn ra các loại như câu tả, câu trần thuật, câu kể để đặt câu hỏi. Tránh các câu như câu cảm, câu cầu khiến, câu hỏi vì hầu hết những câu thuộc kiểu câu này không chứa các thông tin để hỏi. Xét một văn bản đúng cú pháp tiếng Việt thì khi kết thúc câu tả, câu kể, và câu trần thuật dấu chấm câu là dấu chấm, còn câu cảm, cầu khiến sẽ là dấu chấm than, câu hỏi là dấu chấm hỏi. Như vậy bằng cách này chúng ta có khả năng phân loại các câu trong đoạn văn, lọc ra câu như câu kể, tả, trần thuật một cách dễ dàng. Tiêu chí tiếp theo là ta chỉ xét những câu đơn và câu ghép có nòng cốt rõ ràng. Theo bài báo cáo VLSP [2] thì một câu đơn cơ bản gồm có một nòng cốt đơn. Nòng cốt đơn gồm có hai phần, phần đề và phần thuyết (theo quan điểm ngữ pháp chức năng) mà quan điểm ngữ pháp truyền thống gọi là chủ ngữ và vị ngữ.

Ví dụ:

Bão Lekima cấp 11 / đang hướng vào Nghệ An - Hà Tĩnh.

Mọi chuyện / rồi sẽ qua đi.

Trong cấu tạo câu đơn có thể có những thành phần ngoài nòng cốt như thành phần than gọi, thành phần chuyển tiếp, thành phần chú thích, thành phần tình huống, thành phần khởi ý.

Ví dụ:

Nhiều lúc, tôi cũng muốn gào thét thật to, đập tung, phá vỡ tất cả…

Con người, đó là cái vốn quý nhất.

Chao, đường còn xa lắm!

Riêng với thành phần than gọi thì ta chỉ xét nó thuộc nòng cốt câu khi nó đứng ở cuối hoặc ở giữa câu.

Ví dụ:

Chúng ta đi về đi, bà con ơi!

Nhưng trong bài nghiên cứu này, chúng tôi sẽ bỏ qua trường hợp câu đơn có thành phần ngoài nòng cốt là thành phần than gọi. Bởi vì hầu hết chúng đều không chứa những thông tin hỏi được. Và một lưu ý đối với tiêu chí số hai khi xét trên câu đơn là chúng ta chỉ tạo câu hỏi dựa trên câu có nòng cốt gồm hai phần, chủ ngữ và vị ngữ. Đối với trường hợp câu đơn đặc biệt – câu mà nòng cốt đơn chỉ có một thành phần, vì không ẩn chứa đầy đủ thông tin nên ta sẽ không chọn loại câu này để đặt câu hỏi. Trong tập quy tắc cú pháp tiếng Việt [1], các nhà nghiên cứu nhóm VLSP cũng đã thống nhất rằng câu đơn đặc biệt bao gồm câu rút gọn, trong câu rút gọn thì tùy vào ngữ cảnh, câu sẽ được rút gọn chủ ngữ hoặc vị ngữ hoặc rút gọn cả hai thành phần, trên mặt lý thuyết ta vẫn có thể tạo câu hỏi ở dạng câu này thông qua việc tìm được các thành phần bị rút gọn, sau đó chuyển về câu đơn bình thường, nhưng trong phạm vi bài nghiên cứu này, chúng tôi chỉ tập trung xét các câu đơn có nòng cốt đầy đủ hai thành phần là chủ ngữ, vị ngữ. Trong ví dụ 4 dưới đây là trường hợp câu đơn đặc biệt.

Ví dụ:

Chỉ còn lại những ngày cuối cùng…

Điều chỉnh lại mình đi!

Đi!

Dựa vào nòng cốt trong câu đơn bình thường, ta có thể dễ chọn thành phần chứa thông tin để đặt câu hỏi, nó ít phức tạp hơn câu có nhiều hơn một nòng cốt - câu ghép và câu phức. Nhưng không phải lúc nào cũng có nhiều câu đơn cho ta trích xuất thông tin, xét văn phong của người Việt Nam thì có xu hướng sử dụng câu ghép và câu phức nhiều hơn câu đơn, nguyên nhân là khi viết, các câu văn có nhiều hơn một nòng cốt làm cho lối hành văn trở nên lưu loát, có nhiều cách kết hợp phong phú, tránh lặp từ, không cứng

nhắc như viết nhiều câu đơn một lúc, việc này cũng ảnh hưởng đến giọng điệu của văn bản. Đó là mặt ưu điểm của câu ghép trong lối hành văn, nhưng nếu ta xét tính chặt chẽ của văn bản thì việc tách các vế trong câu ghép thành câu đơn sẽ khiến cho nội dung của đoạn văn tuy nghe không hay nhưng trở nên dễ hiểu hơn hẳn. Một ưu điểm mạnh mẽ của việc tách các vế trong câu ghép sẽ giảm độ phức tạp khi xử lý văn bản và trích xuất thông tin, không giống như câu có một nòng cốt, các câu văn có nhiều nòng cốt sẽ tồn tại nhiều thông tin, công việc vẫn còn nhẹ nhàng nếu câu ghép và câu phức chỉ có hai nòng cốt, nhưng nếu cấu trúc của chúng nhiều hơn ba nòng cốt thì sẽ rất khó để phân tích chính xác được phụ thuộc các thành phần trong câu (*Dependency Tree*), và việc phát sinh câu hỏi sẽ khó khăn hơn gấp bội, giả sử câu hỏi được sinh ra sau khi chọn được vế để hỏi, thì kết quả cũng rất rườm rà, không được đánh giá cao.

Có hai lưu ý cho vấn đề tách câu trong câu ghép và câu phức, theo một bài báo khoa học ngôn ngữ [2] trên VLSP năm 2008, tác giả có nhấn mạnh câu ghép có hai cấu trúc cơ bản là câu ghép song song (câu ghép đẳng lập) và câu ghép qua lại (câu ghép chính phụ), và chúng ta chỉ có thể tách các vế trong câu ghép đẳng lập, vì các nòng cốt trong mỗi vế là độc lập, không phụ thuộc lẫn nhau. Còn câu ghép qua lại hay còn gọi là câu ghép chính phụ, ta không thể thực hiện việc tách các vế câu vì chúng phụ thuộc lẫn nhau. Có thể nhận biết câu ghép chính phụ qua các cặp từ quan hệ như: nếu thì, tuy nhưng, do mà,…

Ví dụ:

Đa số bà con ủng hộ chủ trương xây dựng khu đô thị mới Thủ Thiêm và họ sẵn sàng giao đất để thực hiện dự án, nhưng họ muốn phải được đảm bảo quyền lợi và cuộc sống sau khi di dời.

→ Theo ngữ nghĩa thì câu này có thể tách:

Đa số bà con ủng hộ chủ trương xây dựng khu đô thị mới Thủ Thiêm.

Họ sẵn sàng giao đất để thực hiện dự án, nhưng họ muốn phải được đảm bảo quyền lợi và cuộc sống sau khi di dời.

Tuy nhiên ta nên hạn chế việc tách câu này, đặc biệt là với những câu ghép đẳng lập mà các vế câu được nối với nhau bằng liên từ (*"và", "rồi", "hay", "còn"*). Vì việc tách câu này có thể làm cho câu cú gọn gàng nhưng ý nghĩa tự nhiên của ngữ liệu ít nhiều đã bị thay đổi. Nhưng nếu cấu trúc câu ghép song song có hơn hai vế và quá phức tạp (gồm nhiều nòng cốt đơn) thì ta có thể tách thành những câu đơn. Bởi vì quan hệ giữa các vế trong câu ghép song song không thật chặt chẽ và tách ra càng đơn giản thì việc xử lý dữ liệu sẽ càng dễ dàng.

Ví dụ:

Giọng của cháu đôi lúc đã nghẹn lại trong quá trình phiên dịch cho tổng thống và Chủ tịch nước, cháu đã cố kiềm chế những giọt nước mắt của mình vì quá xúc động.

→ Ta không tách câu trường hợp này.

Trường hợp câu phức thì càng không thể tách ra thành các câu đơn, biết rằng cơ bản là chúng chứa nhiều hơn một nòng cốt nhưng khi xét cả câu thì chỉ có một nòng cốt đóng vai trò chính, vì thế để giữ được ý nghĩa của câu thì chúng ta không nên tách các vế trong câu thành các thành phần biệt lập.

Một lưu ý thứ hai khi tách các vế trong câu ghép sẽ tạo ra một vấn đề nhập nhằng về đại từ của vế sau, trỏ đến danh từ ở vế trước.

Ví dụ:

An đi Hà Nội, anh ấy đi bằng máy bay.

→ Sau khi tách câu trên ta sẽ được hai câu đơn như sau:

An đi Hà Nội.

Anh ấy đi bằng máy bay.

Từ hai câu trên, chúng ta có thể phát sinh ra câu hỏi tương ứng với câu "Anh ấy đi bằng máy bay" là "*Anh ấy đi bằng phương tiện gì?",* trong trường hợp này đại từ "*Anh*

ấy" chỉ đến "*An*", nó gây ra hiện tượng nhập nhằng, không rõ ràng trong ngôn ngữ, nếu mô hình phát sinh câu hỏi sinh ra câu này thì người trả lời sẽ không biết "a*nh ấy*" là ai. Để tránh được vấn đề này thì khi tách các vế câu, ta cần phải có một bước xử lý đại từ, thay đổi "*Anh ấy*" thành danh từ tương ứng với chúng "*An*", như vậy sau khi tách câu ta sẽ được hai câu đơn là "*An đi Hà Nội*", "*An đi bằng máy bay*". Việc xử lý tính mơ hồ của các đại từ này còn có ý nghĩa rất lớn khi chúng ta xét các câu ở mức đoạn, vì trong một văn bản thì câu không độc lập, mà chúng liên kết và được nối với nhau bằng các phép nối, một trong số chúng là sử dụng đại từ để chỉ danh từ trong câu trước đó, tránh được hiện tượng lặp từ.

Ví dụ:

Tí đi sở thú. Anh ấy nhìn thấy con voi.

Trong ví dụ này, ta cũng phải xử lý đại từ "*Anh ấy*" chuyển thành danh từ riêng "*Tí*", sau đó mới tiến hành phương pháp đặt câu hỏi.

Tóm lại, trong chương này chúng tôi đã đưa ra các tiêu chí như chọn câu phù hợp, bao gồm các câu như câu tả, câu trần thuật, câu kể, và nếu là câu đơn thì phải chứa đầy đủ thông tin, có nòng cốt rõ ràng, gồm hai thành phần chủ ngữ, vị ngữ (câu đơn bình thường). Nếu câu được trích xuất có nhiều hơn một nòng cốt thì sẽ là câu ghép và câu phức, trường hợp câu phức và câu ghép chính phụ thì chúng ta không thể tách các vế để đơn giản cấu trúc câu vì các nòng cốt phụ thuộc lẫn nhau, trường hợp này ta chỉ có thể tách các vế trong câu ghép đẳng lập vì các vế của nó độc lập, không phụ thuộc lẫn nhau, một lưu ý nhỏ là ta chỉ nên tách khi câu ghép đẳng lập đang xét có cấu trúc quá phức tạp để hỗ trợ việc xử lý trở nên dễ dàng hơn, còn nếu cấu trúc đơn giản thì ta hạn chế việc tách các vế bởi vì phần nào ý nghĩa tự nhiên của ngữ liệu sẽ bị thay đổi. Trong chương tiếp theo, chúng tôi sẽ tập trung phân tích cú pháp câu đơn tiếng Việt, theo bài nghiên cứu [1], tổng quát hoá tiếng Việt có bao nhiêu cấu trúc cho câu đơn, từ đó có những hướng tiếp cận phù hợp cho từng cấu trúc câu đơn (không xét câu đơn đặc biệt).

CHƯƠNG 3 – PHƯƠNG PHÁP PHÁT SINH CÂU HỎI TRÊN CÂU ĐƠN

Trong *Hội nghị Quốc tế lần thứ 10 về Hệ thống Gia sư Thông minh (ITS2010)*, các nhà nghiên cứu khoa học ngôn ngữ khi tìm hiểu về phương pháp phát sinh câu hỏi tự động (thông thường ở mức câu) đều đưa ra những phương pháp rất đa dạng, cách tiếp cận khác nhau, mỗi phương pháp đều có những ưu và nhược riêng. Những nỗ lực trước đây trong phát sinh câu hỏi từ văn bản có thể được chia thành ba tiếp cận: dựa trên cú pháp (*syntax-based*), dựa trên ngữ nghĩa (*semantic-based*) và dựa trên khuôn mẫu (*template-based*). Ba loại tiếp cận này chắc chắn không rời rạc nhau, các hệ thống dựa trên phạm trù phân tích cú pháp thường được bổ trợ bởi ngữ nghĩa của văn bản, ít nhất là ngữ nghĩa nông và ngược lại. Và tất nhiên, một hệ thống được xây dựng dựa trên hướng khuôn mẫu luôn phải sử dụng thông tin của cú pháp và ngữ nghĩa văn bản. Nhưng dù là cách tiếp cận nào, các hệ thống đều phải thực hiện được ít nhất bốn bước sau:

1. **Lựa chọn nội dung**

Trong bước này, nhiệm vụ của chúng ta lựa chọn các đoạn văn bản có khả năng xử lý được (thông thường là câu đơn) và chứa nội dung có thể hỏi được như là chọn các loại câu như câu kể, câu tả, câu trần thuật, bỏ các kiểu câu như câu hỏi, câu cảm thán

2. **Xác định mục tiêu**

Xác định từ hoặc cụm từ cụ thể nào có thể được hỏi

3. **Xác định loại câu hỏi**

Sau khi chọn được từ, cụm từ để hỏi thì nhiệm vụ tiếp theo là xác định được loại câu hỏi phù hợp như: ai, con gì, cái gì, ở đâu, v.v

4. **Tạo câu hỏi**

Nhiệm vụ cuối cùng sau khi xác định được ba bước trên, thì chúng ta sẽ thực hiện bước chuyển đổi câu thành câu hỏi, sắp xếp, thêm, xóa các thành phần cú pháp trong câu, định dạng chúng thành một câu hỏi đúng nghĩa

Đây là danh sách thứ tự các bước trong bài luận văn thạc sĩ của David Lindberg [3], nên những bước xử lý trên không phải lúc nào cũng rời rạc và có thể không nhất thiết xảy ra theo thứ tự này. Bước 2 không phải lúc nào cũng đi trước bước 3. Việc xác định mục tiêu có thể thúc đẩy việc xác định loại câu hỏi và ngược lại. Một hệ thống bị ràng buộc trong việc tạo ra các loại câu hỏi cụ thể sẽ chỉ chọn các mục tiêu thích hợp cho các loại câu hỏi đó. Ngược lại, một hệ thống có khả năng tạo rộng hơn có thể chọn mục tiêu tự do hơn và (lý tưởng là) chỉ tạo ra các câu hỏi thích hợp cho các mục tiêu đó. Ví dụ về cả hai trường hợp này sẽ được thảo luận trong các phần sau của bài báo cáo.

Sau đây, chúng tôi sẽ thảo luận về các phương pháp phát sinh câu hỏi của các nhà nghiên cứu trong *Hội nghị Quốc tế (ITS2010*), và xác định phương pháp tốt để làm bước đệm phát triển cho bài toán phát sinh câu hỏi trong tiếng Việt.

3.1Các phương pháp dựa trên phân tích cú pháp (Syntax-based)

3.1.1 Đơn giản hóa câu (Sentece Simplification)

Đơn giản hóa câu thường được xem như một bước tiền xử lý cần thiết. Ngoại trừ phương pháp của Varga và Hà, mỗi công trình nghiên cứu được trích dẫn ở trên sử dụng một hoặc nhiều bước đơn giản hóa, bao gồm tách các câu có chứa các mệnh đề độc lập, loại bỏ phụ tố, loại bỏ cụm từ tiền tố, loại bỏ dấu hiệu diễn ngôn và loại bỏ mệnh đề tương đối. Varga và Hà thích sử dụng cơ chế lọc để ngăn chặn việc tạo ra câu hỏi từ các câu phức tạp hơn là cố gắng đơn giản hóa chúng [17]. Trong khi việc đơn giản hóa làm cho một số khía cạnh của việc hình thành câu hỏi dễ dàng hơn, nó cũng đưa ra những vấn đề cần phải xử lý. Đó chính là vấn đề nhập nhằng về đại từ về sau như đã nói ở Chương 2, nếu ta không giải quyết vấn đề này, thì câu hỏi được tạo ra sẽ trở thành câu hỏi mơ hồ như kiểu "*Anh ấy đi đâu*?".

3.1.2 Nhận dạng cum từ chứa thông tin (Key phrase Identification)

Cụm từ khóa là những cụm từ (đôi khi là một từ) là mục tiêu tốt nhất để tạo câu hỏi. Nhiều cách tiếp cận khác nhau đã được áp dụng để xác định các cụm từ chính này. Một cách tiếp cận vay mượn các kỹ thuật từ tóm tắt tự động để xác định tất cả các cụm từ khóa trong toàn bộ tài liệu văn bản trước khi đi sâu vào tạo câu hỏi ở cấp độ câu [14]. Một cách tiếp cận khác là coi chủ ngữ, tân ngữ và các cụm giới từ có chứa một thực thể được đặt tên là các cụm từ khóa [17]. Cách tiếp cận thứ hai cũng đã được mở rộng để bao gồm các cụm từ khác, chẳng hạn như trạng ngữ [15].

3.1.3. Biến đổi cú pháp và thay thế từ/cụm từ chính bằng từ nghi vấn

Các bước cuối cùng trong phương pháp dựa trên phân tích cú pháp là chuyển đổi ngữ pháp câu văn ban đầu thành câu hỏi bằng các thao tác biến đổi trên cây cú pháp và chèn các từ câu hỏi. Heilman và Smith [18] sử dụng các quy tắc tương tác được xác định trước để xác định trước các loại câu hỏi sẽ được tạo ra từ các mẫu giới từ chủ ngữ-động từ-tân ngữ có các loại thực thể được đặt tên cụ thể ở mỗi vị trí đó. Kalady và cộng sự [14], những nhà nghiên cứu sử dụng phương pháp nhận dạng cụm từ khóa cấp độ tài liệu đã được nêu trước đây, đã đưa ra các phương pháp phân tích riêng biệt để tạo định dạng câu hỏi tùy thuộc vào các cụm từ khóa có nhãn từ loại là gì, cụm danh từ, cụm danh từ tân ngữ, phụ ngữ, cụm tiền đề , hoặc trạng ngữ.

3.1.3 Biến đổi cú pháp và thay thế từ/cụm từ chính bằng từ nghi vấn

Các bước cuối cùng trong phương pháp dựa trên phân tích cú pháp là chuyển đổi ngữ pháp câu văn ban đầu thành câu hỏi bằng các thao tác biến đổi trên cây cú pháp và chèn các từ câu hỏi. Heilman và Smith [18] sử dụng các quy tắc tương tác được xác định trước để xác định trước các loại câu hỏi sẽ được tạo ra từ các mẫu giới từ chủ ngữ-động từ-tân ngữ có các loại thực thể được đặt tên cụ thể ở mỗi vị trí đó. Kalady và cộng sự [14], những nhà nghiên cứu sử dụng phương pháp nhận dạng cụm từ khóa cấp độ tài liệu đã được nêu trước đây, đã đưa ra các phương pháp phân tích riêng biệt để tạo định dạng

câu hỏi tùy thuộc vào các cụm từ khóa có nhãn từ loại là gì, cụm danh từ, cụm danh từ tân ngữ, phụ ngữ, cụm tiền đề , hoặc trạng ngữ.

3.1.4 Ví dụ

Một số ví dụ minh hoạ kết quả của phương pháp phân tích dựa trên cú pháp. Xét ví dụ sau đây:

Barack Obama is the president of America.

Câu này không cần phải đơn giản hóa, cụm danh từ chủ ngữ của nó được tách ra và xác định như một thực thể loại người (*person*). Câu hỏi trong được hình thành bằng cách thay thế chủ ngữ bằng từ nghi vấn thích hợp với loại người là *who*.

Who is the president of America?

Kalady và cộng sự cung cấp một ví dụ khác, trong đó việc đơn giản hóa câu sẽ được thực hiện như sau:

Mexico City, the biggest city in the world, has many interesting archaeological sites.

Câu này chứa một phụ tố (appositive), "*the biggest city in the world*" (là phụ tố của Mexico City). Trích xuất phụ tố và biến nó thành câu hỏi có kết quả như bên dưới:

Which/Where is the biggest city in the world?

Một mẫu câu ngắn hơn thể hiện hành vi có thể dự đoán được của QG dựa trên cú pháp. Hãy xem xét ví dụ sau đây, một ví dụ được đưa ra bởi Ali và cộng sự [17].

Tom ate an orange at 7 pm.

Các câu hỏi đến được tạo ra bởi Ali và cộng sự.

Who ate an orange?

Who ate an orange at 7 pm?

What did Tom eat?

When did Tom eat an orange?

Tất cả các ví dụ này đều dùng để minh họa hành vi cơ bản của các phương thức dựa trên cú pháp. QG từ văn bản về cơ bản là rút gọn các thành phần cú pháp của một câu và thay thế các từ hoặc cụm từ bằng các từ nghi vấn.

Điểm mạnh quan trọng nhất của các phương thức dựa trên cú pháp này là tính di động của chúng. Các phương pháp này dựa trên các phép biến đổi có mục đích chung có thể áp dụng cho bất kỳ miền nào. Tuy nhiên, có thể cần phải đào tạo lại mô hình nhận dạng thực thể được đặt tên để xác định tốt hơn các thực thể được đặt tên (NER) khi áp dụng các phương pháp này cho các lĩnh vực bí truyền hơn. Thật không may, tính khả chuyển của các phương pháp này hạn chế các loại câu hỏi mà chúng có thể tạo ra. Như các ví dụ đã cho thấy, chỉ bằng cách sắp xếp lại các yếu tố cú pháp của một câu thì các phương pháp này sẽ tạo ra rất nhiều câu hỏi factoid.

3.2Các phương pháp dựa trên phân tích ngữ nghĩa (Semantic-based)

Giống như các phương pháp dựa trên cú pháp, phương pháp dựa trên ngữ nghĩa đã dựa vào những phép biến đổi để biến câu khai văn ban đầu thành câu hỏi. Tuy nhiên, sự khác biệt là các phương pháp dựa trên ngữ nghĩa sử dụng nghĩa nhiều hơn thay vì phân tích cú pháp để chuyển đổi câu thành câu hỏi. Mannem và cộng sự [19] cho chúng ta thấy một hệ thống kết hợp SRL (Semantic Role Labeling) với các phép biến đổi cú pháp. Trong bước lựa chọn nội dung, một câu đơn trước tiên sẽ được phân tích cú pháp với một trình gắn nhãn vai trò ngữ nghĩa (SRL) để xác định các mục tiêu tiềm năng. Các mục tiêu được chọn bằng cách sử dụng các tiêu chí lựa chọn đơn giản. Bất kỳ đối số ngữ nghĩa cụ thể của vị từ (Arg0-Arg5), nếu có, đều được coi là mục tiêu hợp lệ. Cần lưu ý rằng bất kỳ vị từ nào có ít hơn hai đối số này đều không được coi là khả thi và bị bỏ qua, lý do là cần có ít nhất hai đối số để hình thành câu hỏi. Mannem và các cộng sự xác định thêm ArgM-MNR, ArgM-PNC, ArgM-CAU, ArgM-TMP, ArgM-LOC và ArgM-DIS là

các mục tiêu tiềm năng. Các vai trò này được sử dụng để tạo ra các câu hỏi bổ sung mà không thể đạt được nếu chỉ sử dụng các vai trò Arg0-Arg5. Ví dụ: ArgM-LOC có thể được sử dụng để tạo câu hỏi where và ArgM-TMP có thể được sử dụng để tạo câu hỏi khi nào. Xem bảng dưới đây để biết danh sách đầy đủ các loại câu hỏi.

Semantic Role	**Question Type**
ArgM-MNR	How
ArgM-CAU	Why
ArgM-PNC	Why
ArgM-TMP	When
ArgM-LOC	Where
ArgM-DIS	How

Bảng 3.1 Các loại câu hỏi

Sau khi các mục tiêu đã được xác định, những mục tiêu này cùng với phân tích cú pháp SRL hoàn chỉnh của câu được chuyển sang giai đoạn xây dựng câu hỏi. Bước đầu tiên trong giai đoạn này là xác định động ngữ cho mỗi mục tiêu đã xác định trong câu. Điều này bao gồm động từ chính và bất kỳ trợ động từ hoặc động từ khiếm khuyết nào được xác định từ phân tích cú pháp phụ thuộc của câu. Đối với mỗi mục tiêu, các câu hỏi được tạo ra bằng cách sử dụng một loạt các phép biến đổi đơn giản. Đầu tiên, bản thân mục tiêu sẽ bị xóa, nếu mục tiêu chứa bất kỳ giới từ nào, mục tiêu sẽ được thay thế bằng giới từ đầu tiên của nó. Thứ hai, một từ trong câu hỏi được chọn, từ trong câu hỏi chính xác được xác định dựa trên vai trò ngữ nghĩa của mục tiêu và thực thể được đặt tên (nếu có) mà nó chứa. Thứ ba, bất kỳ tính từ nào xuất hiện bên trái vị ngữ sẽ được chuyển xuống cuối câu. Ở bước chuyển đổi cuối cùng, câu được chuyển thành câu nghi vấn. Các trợ từ hoặc phương thức, nếu có, được chuyển lên đầu câu trước khi từ câu hỏi được chèn vào vị trí ban đầu. Nếu không có trợ giúp hoặc phương thức nào có mặt, một trong số do, does, hoặc did sẽ được thêm vào tùy thuộc vào thẻ POS của vị từ.

Mannem và cộng sự đã tham gia dự án Question Generation Shared Task Evaluation Challenge (QGSTEC) năm 2010 [20], kêu gọi tạo ra sáu câu hỏi từ một đoạn văn nhất định. Họ xếp hạng các câu hỏi của mình bằng cách sử dụng hai phương pháp phỏng đoán và sau đó chọn ra sáu câu hỏi hàng đầu. Trong sơ đồ xếp hạng của họ, các câu hỏi được xếp hạng đầu tiên theo độ sâu của vị ngữ của chúng trong phân tích phụ thuộc của câu gốc. Điều này dựa trên giả định rằng các câu hỏi phát sinh từ các mệnh đề chính được mong muốn hơn các câu hỏi được tạo ra từ các vị từ sâu hơn. Trong giai đoạn thứ hai, các câu hỏi có cùng thứ hạng được xếp lại theo số lượng đại từ mà chúng chứa, với những câu hỏi có ít đại từ hơn sẽ có thứ hạng cao hơn.

Là một phần của việc tham gia vào nhiệm vụ được chia sẻ, họ cũng được yêu cầu đưa ra các câu hỏi ở ba phạm vi: tổng quát, trung bình và cụ thể. Các câu hỏi tổng quát nhằm mục đích là phạm vi đoạn văn, các câu hỏi trung bình nhằm về các ý chính được diễn đạt trong một đoạn văn và các câu hỏi cụ thể được cho là có thể trả lời được từ một câu hoặc cụm từ. Mannem và cộng sự đã không thực hiện bất kỳ phân tích cấp độ đoạn văn nào để tạo ra các câu hỏi chung của họ. Thay vào đó, họ tạo ra một câu hỏi từ câu đầu tiên của mỗi đoạn văn bằng cách sử dụng một bộ quy tắc chuyển đổi khác nhau. Nếu động từ chính là copula (động từ to be, từ "*là*" trong tiếng Việt), họ đặt ra câu hỏi về lập luận đúng của nó. Hai ví dụ dưới đây là kết quả tương ứng trong phương pháp của Mannem.

Intelligent tutoring systems consist of four different subsystems or modules.

Vị từ câu trên là "*consist*", là một copula, vì vậy câu hỏi trong được hình thành bằng cách sử dụng đối số đúng, trong trường hợp này là Arg2.

What are the four different subsystems or modules that intelligent tutoring systems consist of?

Vị từ có trong ví dụ dưới đây sẽ không phải là một copula, vì vậy quy tắc chuyển đổi thứ hai được sử dụng để tạo ra.

But one-handed backhands have some advantages over two-handed players.

What are some advantages over two-handed players that one-handed backhands have?

Mặc dù kỹ thuật tạo câu hỏi dựa trên ngữ nghĩa khác nhiều so với các phương pháp phân tích dựa trên cú pháp, nhưng phương pháp dựa trên ngữ nghĩa này vẫn đạt kết quả rất giống với phân tích cú pháp, vì vậy điểm mạnh và điểm yếu của chúng rất giống nhau. Tuy chúng thực sự có khả năng linh hoạt, nhưng chúng cung cấp quá nhiều các câu hỏi factoid.

3.3Các phương pháp dựa trên khuôn mẫu (Template-based)

Các khuôn mẫu của câu hỏi cung cấp khả năng đặt câu hỏi không cần sự kết hợp chặt chẽ với từ ngữ chính xác của văn bản nguồn như các phương pháp dựa trên cú pháp và ngữ nghĩa. Mẫu câu hỏi được áp dụng cho bất kỳ văn bản nào được xác định trước với các biến phù hợp thì sẽ được thay thế bằng nội dung từ văn bản ban đầu.

Cai và cộng sự [21] trình bày NLGML (Natural Language Generation Markup Language), một loại công cụ có thể được sử dụng để tạo ra các câu hỏi thuộc bất kỳ lĩnh vực nào. NLGML kết hợp việc phân tích cú pháp và các thuộc tính ngữ nghĩa để hình thành một khuôn mẫu phù hợp để hỏi. Khi viết một tập lệnh NLGML, tác giả xác định các khối khuôn-mẫu-thể loại (category-pattern-template). Category ràng buộc pattern và template với nhau. Các pattern kết hợp cấu trúc cụm từ cú pháp và các đặc điểm ngữ nghĩa, được sử dụng để xác định các câu thích hợp và các template xác định các quy tắc được sử dụng để chuyển câu nguồn thành câu hỏi mong muốn. Ví dụ minh hoạ bên dưới ta có thể thấy rõ cấu trúc của NLGML của Cai.

Như ví dụ minh họa này, một mẫu không cần là một cây cú pháp hoàn chỉnh. Thẻ <star /> là một ký tự biểu diễn cho bất kỳ cây con nào. Các biến _person_ và _place_ được sử dụng bởi các mẫu để trích xuất văn bản của các cụm danh từ tương ứng, nói

chung, các biến có thể được gán cho bất kỳ phần nào của cây con. Các biến cung cấp cơ chế trích xuất chuỗi con bất kể cú pháp của cây con bên dưới.

Các pattern cũng có thể áp đặt ràng buộc ngữ nghĩa. Trong ví dụ này, thuộc tính person = "true" ở chủ ngữ cho thấy chỉ những chủ ngữ là cụm danh từ có chứa một thực thể được đặt tên (NER) của loại người (person) mới được coi là hợp lệ. Tương tự, cụm danh từ thứ hai yêu cầu một thực thể được đặt tên của loại vị trí (loc). Việc kết hợp các đặc điểm ngữ nghĩa này cho phép hệ thống dựa trên NLGML xác định thời điểm thích hợp để đặt câu hỏi cho ai và ở đâu. Tương tự, đặc điểm ngữ nghĩa thời gian (time) cũng có thể được sử dụng để tạo khi câu hỏi.

```
<category>
  <pattern>
    <S>
      <NP person="true">_person_</NP>
      <VP> <VBD>went</VBD>
        <PP> <TO>to</TO> <NP location="true">_place_</NP> </PP>
      </VP>
      <star />
    </S>
  </pattern>
  <template> Where did _person_ go? </template>
  <template> Who went to _place_? </template>
  <template> Why did _person_ go to _place_? </template>
</category>
```

Hình 3.1 Natural Language Generation Markup Language

Mô hình NLGML trên có thể được coi là mã hóa ý tưởng "*somebody went somewhere*". Khi có một câu trong văn bản nguồn khớp với pattern này, ba câu hỏi sẽ được tạo ra bằng cách sử dụng các template đã xác định. Cai và cộng sự phát sinh ra ba câu hỏi phù hợp với câu "*The boy went to school*." như sau:

Where did The boy go?

Who went to school?

Why did The boy go to school?

Như chúng ta thấy thì bước thứ tư đơn giản chỉ cần "copy và paste" các thuộc tính vào trong template, một nhiệm vụ vô cùng dễ dàng, ít tốn sức và làm cho cấu trúc câu hỏi chặt chẽ hơn nhiều.

Bài nghiên cứu phương pháp QG của Ceist [22] cũng là một ví dụ khác về cách tiếp cận dựa trên khuôn mẫu. Giống như NLGML, Ceist sử dụng đối sánh các pattern, biến và template để chuyển đổi câu trong văn bản nguồn thành câu hỏi. Một điểm khác biệt chính giữa Ceist và NLGML là cách các pattern được chỉ định. Trong bài nghiên cứu của Ceist, các pattern được chỉ định bằng cú pháp của Stanford Tregex [23]. Điểm khác biệt thứ hai là Ceist không dựa vào nhận dạng thực thể được đặt tên rõ ràng. Thay vào đó, Ceist coi nhận dạng thực thể được đặt tên như một bài tập so khớp mẫu bổ sung.

Chúng ta cũng dễ hình dung được một điểm mạnh rõ ràng của các phương pháp dựa trên khuôn mẫu này là khả năng tạo ra các câu hỏi với cách diễn đạt phong phú, đa dạng. Do đó, về nguyên tắc chúng không bị ràng buộc chỉ giới hạn trong một số loại câu hỏi nhất định, chẳng hạn như *who, what, when, where, why, v.v.* Điều này dường như có khả năng phá vỡ sự cứng nhắc trong việc tạo câu hỏi, giúp câu hỏi của chúng ta ít nhiều trên mức factoid, và cho phép các câu hỏi diễn giải ý nghĩa văn bản gốc dễ hiểu hơn. Phương pháp tiếp cận dựa trên mẫu chắc chắn là lựa chọn lý tưởng cho hệ thống đọc hiểu của Mostow và Chin [24], vì các template rất phù hợp để tạo ra các câu hỏi siêu nhận thức.

Sự linh hoạt này đi kèm với cái giá là cần thêm nỗ lực của con người để làm cho các hệ thống này hoạt động tốt. Các hệ thống được thiết kế để khả chuyển theo miền dựa trên đối sánh pattern dựa trên cú pháp. Điều này cần các ràng buộc đối sánh rất sát nghĩa, nên nó có thể làm cho các pattern trở nên khó viết. Tuy nhiên, quan trọng hơn, trong khi điều này có thể ngăn chặn việc tạo quá nhiều, nó cũng có thể khiến các câu hỏi tiềm năng bị bỏ sót. Một sự khác biệt nhỏ về cú pháp sẽ khiến các câu giống nhau về ngữ nghĩa bị xử lý khác nhau. Sử dụng các phương pháp dựa trên mẫu hiện có, chúng ta phải giải quyết vấn đề này bằng cách có nhiều mẫu hoặc nhiều mẫu phức tạp hơn. Trong chương

tiếp theo, chúng tôi sẽ kết hợp các phương pháp cú pháp, ngữ nghĩa và khuôn mẫu lại với nhau, áp dụng cho phù hợp cấu trúc của tiếng Việt.

CHƯƠNG 4 – PHƯƠNG PHÁP CỦA CHÚNG TÔI

4.1Mô hình phát sinh câu hỏi dựa trên phương pháp phân tích cú pháp và khuôn mẫu cho từng kiểu câu (Syntax-based và Template-based)

4.1.1 Câu hai thành phần có vị ngữ danh từ

Nhóm này có hệ từ cùng với danh từ làm vị ngữ. Mô hình tổng quát của nhóm này là:

<Câu> = <Chủ ngữ> <Vị ngữ> (là <danh từ>)

Ví dụ:

Tôi là sinh viên.

Máy bay là phát minh vĩ đại của con người.

Thầy Phước là tiến sĩ trường Đại học Tôn Đức Thắng.

Hà Nội là thủ đô Việt Nam.

Cây phân tích cú pháp cho thấy trong câu loại này, *root* là danh từ đứng sau từ "*là*", và từ "*là*" có quan hệ *copular* với *root* (do từ "*là*" trong tiếng Việt cùng nghĩa và chức năng với động từ to be trong tiếng Anh). Chủ ngữ là danh từ sẽ có quan hệ *nsubj* với *root* trong câu, trường hợp chủ ngữ là một cụm danh từ thì từ có mối quan hệ *nsubj* với *root* sẽ là danh từ trung tâm, trong Chương 2 khi nhắc đến khái niệm danh ngữ, các nhà nghiên cứu VLSP chia danh ngữ ra làm bốn thành phần đối với mô hình tổng quát, nếu danh ngữ có danh từ làm trung tâm khi không có danh từ chỉ loại, và ngược lại thì danh từ chỉ loại sẽ là thành phần trung tâm. Còn đối với cây phân tích phụ thuộc của Underthesea [14], vậy nên mọi chuyện ngược lại, danh từ sẽ luôn là thành phần trung tâm bất kể trong danh ngữ có danh từ chỉ loại hay không. Nếu trường hợp chủ ngữ là một mệnh đề đầy đủ chủ-vị thì vị ngữ của mệnh đề chủ ngữ chính sẽ có quan hệ *csubj* với *root* trong nòng cốt của câu. Chủ ngữ là một cụm động từ thì động từ chính sẽ có quan hệ *vsubj* với *root*, tương tự với tính từ là *asubj*.

Đối với chủ ngữ nếu là quan hệ *nsubj* với *root* của câu, thì ta sẽ chia ra làm ba loại câu hỏi đó là "*ai?", "con gì?", "cái gì?*". Để phân loại được ba trường hợp này thì chúng tôi sẽ kiểm tra theo thứ tự như sau, đầu tiên là kiểm tra chủ ngữ có phải chỉ người không, sau đó kiểm tra chủ ngữ phải là con vật không, nếu chủ ngữ không thuộc các loại trên thì nó sẽ thuộc loại còn lại đó là "*cái gì*?". Vấn đề cần giải quyết bây giờ là làm sao chúng ta có thể biết chủ ngữ là con người, không thể dựa trên cây phân tích cú pháp vì nó đơn giản cho chúng ta biết các phụ thuộc cú pháp của câu, công cụ xác định thực thể có tên cũng là một cách giải quyết, nếu trong chủ ngữ tồn tại tên riêng và được mô hình NER xác định là person thì chủ ngữ chắc chắn sẽ là người, nếu không tính sai số của mô hình NER. Nhưng như thế vẫn là chưa đủ để xem chủ ngữ có phải là người hay chưa, chúng tôi sẽ thêm bước xác định từng từ trong cụm chủ ngữ, nếu phát hiện được đại từ nhân xưng "*anh", "chị", "tôi", "nó*", v.v. Vì công cụ gán nhãn từ loại (POS) trong Underthesea không có nhãn cho đại từ nhân xưng, nên chúng tôi quyết định sẽ tự tìm một danh sách bao gồm các đại từ nhân xưng và kiểm tra từng từ trong chủ ngữ, nếu từ nào nằm trong danh sách này thì đều chỉ người. Ngoài ra chúng ta còn có các từ chỉ người khác như là "*bộ đội", "cảnh sát", "công nhân", "bác sĩ", "giám đốc", "y tá", "diễn viên", "nghệ sĩ*", v.v. Trong trường hợp này thì chúng tôi cũng sẽ liệt kê trong một danh sách và cố gắng tìm nhiều trường hợp nhất có thể để phân loại được người. Như vậy là chúng ta đã phân loại được trường hợp thứ nhất, và trước khi đi qua trường hợp con gì thì NER của Underthesea cũng hỗ trợ thêm hai loại nhãn phổ biến đó là *location* và *organization*, nếu là *location* thì chủ ngữ của ta sẽ là nơi chốn, chủ ngữ trả lời cho câu hỏi "*đâu*?", tương tự với *organization*, chủ ngữ sẽ trả lời cho câu hỏi "*tổ chức nào*?". Để xác định được chủ ngữ có phải con vật hay không thì đây là nhiệm vụ vẫn còn tồn tại nhiều thách thức và công cụ xử lý ngôn ngữ tự nhiên của chúng ta, ít nhất là chúng tôi chưa tìm được công cụ để xử lý loại này. Phương pháp của chúng tôi rất đơn giản là xét nếu chủ ngữ có danh từ chỉ loại là "*con*" thì ta sẽ coi chủ ngữ đó trả lời cho câu hỏi "*con gì*?", nhưng không phải trường hợp chủ ngữ có danh từ chỉ loại là "*con*" thì chúng ta đều

chấp nhận, bỏ các trường hợp là "*con trai", "con gái", "con nuôi", "con tàu", "con thuyền", "con song", "con nước", "con kênh", "con ngươi", "con mắt", v.v.* Tất cả những trường hợp còn lại chủ ngữ trả lời cho câu hỏi *"cái gì?"*.

Trong trường hợp chủ ngữ là *vsubj* hoặc *asubj* thì câu hỏi được sinh ra sẽ tuỳ kiểu câu chúng tôi quyết định khai thác thông tin này, trong trường hợp câu có vị ngữ là "*là danh từ*" thì chúng ta có thể hỏi là "*Hành động nào là + danh từ?"* đối với *vsubj* hoặc "*Danh từ + là gì?*" đối với *asubj*. Riêng với chủ ngữ là một mệnh đề thì chúng tôi sẽ tạo câu hỏi là "*Sự việc nào là + danh từ*?".

Bây giờ chúng tôi sẽ đưa ra phương pháp xử lý câu hỏi ở thành phần vị ngữ. Cấu trúc của vị ngữ đa dạng, phức tạp, khó xử lý hơn nhiều so với cấu trúc của chủ ngữ. Và để giảm bớt sự phức tạp trong cấu trúc của vị ngữ, chúng tôi sẽ không xử lý những câu có thành phần vị ngữ bao gồm những quan hệ *parataxis* (câu ghép đẳng lập), *advcl* (trạng ngữ), *mark* (từ nối với mệnh đề quan hệ khác), *ccomp* (mệnh đề bổ ngữ), với riêng quan hệ ccomp thì chúng ta chỉ chấp nhận khi câu có vị ngữ là động từ. Một lưu ý nữa là nếu vị ngữ có quan hệ *conj* (conjunction) thì chúng ta sẽ xét cụm conj này nếu chúng là một cụm từ cùng từ loại với *root* thì sẽ được nhận, lý do chúng tôi có hệ lọc như vậy vì trong quan hệ *conj* đôi khi là một mệnh đề, cấu trúc câu sẽ trở nên phức tạp và khó để câu hỏi diễn tả được ý nghĩa muốn hỏi. Nói chung, trong bước này chúng tôi cố gắng tạo một bộ lọc cho vị ngữ, để đơn giản hóa cấu trúc phức tạp của chúng, phù hợp với các phương pháp hiện tại của chúng tôi, tuy nhiên trong tương lai, có thể bộ lọc này sẽ được thay đổi tùy thuộc vào quá trình thêm bớt các phương pháp trong mô hình.

Khi câu đơn thuộc trường hợp này, *root* của câu sẽ là danh từ, như vậy để khai thác thông tin của cụm danh từ này thì chúng ta có thể dựa vào phương pháp tạo câu hỏi trên chủ ngữ, khi danh từ trung tâm nằm trong chủ ngữ có mối quan hệ nsubj với *root*. Trên thực tế thì hai trường hợp này khá giống nhau, và trong một số tình huống thì có thể đổi chỗ với nhau (Danh từ là danh từ). Vì lý do này trong mô hình câu đơn có vị ngữ là cụm "*là + danh từ*" thì phần vị ngữ chúng ta xử lý giống với chủ ngữ khi nó là

danh từ. Khi nó được gán nhãn thực thể là *person* ở ngay *root*, thì câu hỏi có tên là gì, sẽ được chọn. Và đây là một số câu hỏi mô hình chúng tôi phát sinh:

Tôi là sinh viên.

=> *Sinh viên là ai?;*

Tôi là gì?

Máy bay là phát minh vĩ đại của con người.

=> *Phát minh vĩ đại của con người là gì?;*

Máy bay là gì?

Thầy Phước là tiến sĩ trường Đại học Tôn Đức Thắng.

=> *Tiến sĩ trường Đại học Tôn Đức Thắng có tên là gì?;*

Thầy Phước là gì?

Hà Nội là thủ đô Việt Nam.

=> *Đâu là thủ đô Việt Nam?;*

Hà Nội là gì?

4.1.2 Câu hai thành phần có vị ngữ tính từ

Cấu trúc của nó gồm: chủ ngữ là danh từ, đại từ, còn vị ngữ là tính từ có hoặc không có hệ từ.

Mô hình tổng quát của nhóm này là:

<Câu> = <Chủ ngữ> <Vị ngữ> (là <tính từ>)

Ví dụ:

Cô ta thông minh.

Chỉ cô ta là thông minh thôi.

Nếu chủ ngữ trong câu là cụm tính từ thì chúng tôi sẽ không phát sinh câu hỏi trong dạng này. Ngoài ra chủ ngữ là cụm danh từ thì khi tạo câu hỏi về chủ ngữ, chúng ta cũng có một bộ lọc để biết được chủ ngữ sẽ trả lời cho câu hỏi loại nào, cấu trúc cũng là hỏi về ai, con gì, cái gì, đâu, v.v. Điểm khác biệt so với cấu trúc câu có vị ngữ là cụm "*là tính từ*", khi chủ ngữ là một cụm động từ thì template câu hỏi của chúng ta sẽ là "*Hành động nào có đặc điểm + tính từ*?".

Đối với vị ngữ, trước khi khởi tạo câu hỏi, chúng ta cần phải lọc thành phần này giống như những gì chúng ta đã làm với cấu trúc câu "là danh từ". Các thành phần quan hệ như *parataxis* (câu ghép đẳng lập), *advcl* (trạng ngữ), *mark* (từ nối với mệnh đề quan hệ khác), *ccomp* (mệnh đề bổ ngữ), Nếu trong vị ngữ có quan hệ *conj* thì chúng ta cần thực hiện bước kiểm tra xem cây con của *conj* có là tính từ hay không, nếu thoả điều kiện này, bước đặt câu hỏi mới được thực thi. Và *root* phải là tính từ, thành phần trung tâm của tính ngữ sẽ là *root* của cả câu. Chúng ta sẽ có template câu hỏi cho thành phần vị ngữ là "*Chủ ngữ + có đặc điểm gì*?". Đây là một số ví dụ và kết quả của mô hình chúng tôi chạy được:

Cô ta thông minh.

=> *Ai thông minh?;*

Cô ta có đặc điểm gì?

Con mèo nhà em rất giỏi bắt chuột.

=> *Con gì rất giỏi bắt chuột?;*

Con mèo nhà em có đặc điểm gì?

Trường đại học Tôn Đức Thắng giàu thành tích nhất trong các trường đại học ở Việt Nam.

=> *Đâu là nơi giàu thành tích nhất trong các trường đại học ở Việt Nam?;*

Trường đại học Tôn Đức Thắng có đặc điểm gì?

4.1.3 Câu hai thành phần có vị ngữ danh từ và không có hệ từ

Câu hai thành phần với vị ngữ là danh từ hoặc tổ hợp danh từ không có hệ từ. Loại câu này thường biểu thị ý nghĩa địa điểm, sự kiện, hiện tượng, bản chất. Vị ngữ của nó là tổ hợp danh từ với một số từ loại khác. Về nguồn gốc và chức năng, loại này có thể là biến thể của vị ngữ động từ.

Mô hình tổng quát của nhóm này là:

<Câu> = <Chủ ngữ> <Vị ngữ> (<danh từ>)

Ví dụ:

Đồng hồ này ba kim.

Cả nước một lòng.

Các câu trên đây có khả năng ứng với câu mà bộ phận vị ngữ thêm yếu tố có.

Dạng này để xử lý một cách gọn gàng thì chúng ta sẽ thêm động từ "*có*" vào giữa chủ ngữ và vị ngữ. Sau đó xử lý câu với này dưới dạng cấu trúc chủ ngữ và vị ngữ là động từ, chúng ta có thể xem tiếp mục 4.1.4 để hiểu thêm cách hoạt động của câu đơn loại này. Và đây là một số câu hỏi mô hình chúng tôi phát sinh ứng với việc đã thêm từ có:

Đồng hồ này ba kim.

=> *Cái gì có ba kim?;*

Đồng hồ này có cái gì?

Cả nước một lòng.

=> *Cái gì có một lòng?;*

Cả nước có cái gì?

4.1.4 Câu hai thành phần có vị ngữ động từ

Mô hình tổng quát của nhóm này là:

<Câu> = <Chủ ngữ> <Vị ngữ> (<Động từ>)

Đây là câu trúc chắc chắn chiếm phần lớn trong câu đơn, nếu một câu có đầy đủ chủ ngữ và vị ngữ thì việc vị ngữ là động từ sẽ chiếm hơn 70% tổng số câu. Như vậy chúng tôi muốn khai thác triệt để cấu trúc câu này để mô hình đưa ra kết quả khả quan nhất. Đầu tiên chúng ta xét trường hợp đơn giản nhất trong cấu trúc này đó là vị ngữ chỉ là động từ độc lập, như câu dưới đây:

An chạy bộ.

Chú công nhân đi làm.

Với các loại câu như trên thì chúng ta chỉ có thể phát sinh hai dạng câu hỏi tương ứng về chủ ngữ và vị ngữ. Tương tự như những cấu trúc trước đó, chủ ngữ trả lời câu hỏi ai, con gì, cái gì. Còn vị ngữ sẽ trả lời câu hỏi làm gì, các dạng câu hỏi sau:

An chạy bộ.

=> *An làm gì?*

Ai chạy bộ?

Ngoài cấu trúc tổng quát trên thì câu còn có các cấu trúc đặc thù như bên dưới.

Nếu động từ vị ngữ là ngoại động thì mô hình có thêm bổ ngữ:

<Câu> = <Chủ ngữ> <Vị ngữ> <Bổ ngữ>

Ví dụ:

Chúng tôi thường viết thư bằng bút chì.

Phân tích một ít về ví dụ trên, như đã nhắc đến trong Chương 2, bổ ngữ bao gồm ba loại là bổ ngữ hình thái, bổ ngữ đối tượng và bổ ngữ miêu tả. Trong ba loại này bổ ngữ hình thái không mang lại nhiều giá trị để hỏi được, và hai loại còn lại giàu khả năng

chứa thông tin cao. Giống như trong ví dụ trên "*Chúng tôi viết thư bằng bút chì*", trong câu này gồm chủ ngữ là "*Chúng tôi*", vị ngữ động từ là "*viết*", bổ ngữ đối tượng trực tiếp là "*thư*" và bổ ngữ miêu tả là "*bằng bút chì*". Bổ ngữ đối tượng trực tiếp trả lời câu hỏi ai, cái gì, nếu không tính chúng là mệnh đề, và bổ ngữ gián tiếp sẽ trả lời câu hỏi cho ai, cho cái gì. Còn bổ ngữ miêu tả biểu thị cách thức, trạng thái, tính chất, mục đích, nơi chốn. Từ đó, nếu câu có tồn tại bổ ngữ đối tượng trực tiếp chúng ta sẽ có template như "*Chủ ngữ + vị ngữ + ai/cái gì + (cho + bổ ngữ đối tượng gián tiếp) + (bổ ngữ miêu tả)?*", cũng tuân theo thứ tự này khi chúng ta đặt câu hỏi cho bổ ngữ đối tượng gián tiếp và bổ ngữ miêu tả. Như vậy một lưu ý rằng nếu một câu tồn tại một lúc nhiều bổ ngữ đối tượng trực tiếp, nhiều bổ ngữ miêu tả thì khi phát sinh câu hỏi trên một thành phần bổ ngữ đối tượng trực tiếp nào đó chúng ta không gắn toàn bộ các thành phần cùng loại còn lại vào trong câu hỏi, làm vậy câu hỏi sẽ rất dài dòng và lan man, thay vào đó thành phần bổ ngữ miêu tả tuy nhiều nhưng khác loại, nếu chúng ta lược bỏ phần nào trong số chúng thì câu hỏi có thể sẽ bị thay đổi ý nghĩa không ít thì nhiều. Giả sử như câu sau:

Tôi mua hoa, làm bánh tặng cho mẹ vào dịp sinh nhật.

Trong câu này, ta thấy tồn tại hai thành phần bổ ngữ đối tượng trực tiếp là "hoa" và "bánh". Khi phát sinh câu hỏi về thành phần này, cụ thể là bổ ngữ đối tượng "hoa", chúng tôi sẽ đặt câu hỏi như "Tôi mua cái gì tặng cho mẹ vào dịp sinh nhật?", không phải là "Tôi mua cái gì, làm bánh tặng cho mẹ vào dịp sinh nhật?". Như vậy với loại câu này thì chúng ta có thể hỏi về bổ ngữ đối tượng, bổ ngữ miêu tả nếu có. Mô hình của chúng tôi có thể sinh các câu hỏi cho loại này như:

Chúng tôi thường viết thư bằng bút chì.

=> *Chúng tôi thường viết thư bằng gì?;*

Chúng tôi thường viết cái gì bằng bút chì?;

Chúng tôi thường làm gì?;

Ai thường viết thư bằng bút chì?

Câu với động từ sai khiến có thêm mô hình:

<Câu> = <Chủ ngữ 1> <Vị ngữ 1> <Chủ ngữ 2> <Vị ngữ 2>

Ví dụ:

Tôi khuyên nó học tập.

Trong cấu trúc này, chủ ngữ 2 và vị ngữ 2 thật ra là một mệnh đề bổ ngữ cho câu, vị ngữ 2 có quan hệ ccomp với root, tượng tự "*học tập*" có quan hệ *ccomp* với "*khuyên*". Như vậy để tạo câu hỏi trong trường hợp vị ngữ có bổ ngữ là một mệnh đề, phương pháp của chúng tôi là tách mệnh đề này thành một câu riêng biệt "*Nó học tập*" sau đó sẽ đưa mệnh đề này vào mô hình phát sinh câu hỏi một lần nữa và kết quả tương ứng sẽ được đem gắn vào phần trước khi tách. Quan sát ví dụ trên sẽ rõ, ta có câu "*Nó học tập*" là mệnh đề được tách ra và chuyển thành hai câu hỏi như "*Nó làm gì?"* và "*Ai học tập?*". Gắn thành phần này vào nòng cốt chính của câu hỏi ban đầu sẽ được câu hỏi như sau:

Tôi khuyên nó học tập.

=> *Tôi khuyên nó làm gì?*

Tôi khuyên ai học tập?

Tôi làm gì?

Ai khuyên nó học tập?

Câu có động từ không độc lập làm thành tố chính ngữ. Mô hình:

<Câu> = <Chủ ngữ 1> <Vị ngữ 1> <Vị ngữ 2> <Bổ ngữ>

Ví dụ:

Tôi hứa giúp đỡ anh.

Động từ không độc lập là động từ không biểu thị một nội dung ý nghĩa hoàn chỉnh (ý nghĩa hành động, hoạt động hay trạng thái) do đó, về nguyên tắc, không thể đứng một mình để đảm đương một chức năng ngữ pháp mà đòi hỏi phải có một từ khác (ví dụ: danh từ, động từ …) đi theo sau để bổ sung ý nghĩa. Vậy nếu vị ngữ thuộc động từ không

độc lập thì chúng ta sẽ giữ nguyên động từ này khi đặt câu hỏi, không thay đổi hoặc khai thác thông tin ở chúng, trong ví dụ trên ta thấy "*anh*" là bổ ngữ đối tượng trực tiếp nên câu hỏi có thể đặt ra là "*Tôi hứa giúp ai*?" hoặc là "*Tôi hứa làm gì*?". Vậy trong trường hợp vị ngữ chính trong câu là động từ không độc lập thì chúng tôi sẽ giữ nguyên động từ này lúc tạo câu hỏi. Và trong phương pháp của chúng tôi, không có cách nào có thể xác định giữa đâu là động từ không độc lập và động từ độc lập nên chúng tôi sẽ khai báo một danh sách bao gồm các động từ không độc lập, và đó trở thành một trong các bộ lọc trước khi được đưa vào cấu trúc câu có *root* là động từ. Không chỉ riêng ở dạng câu này, mà tất cả các dạng trong cấu trúc này, nếu vị ngữ là động từ không độc lập, chúng ta sẽ không phát sinh câu hỏi trên chúng, cụ thể là câu hỏi "*làm gì?*".

Về phần chủ ngữ và vị ngữ của câu, chúng tôi đảm bảo rằng thông tin để hỏi trong chủ ngữ không khác nhiều so với những cấu trúc câu trước đây, nhưng chúng tôi quyết định nếu chủ ngữ là cụm động từ và tính từ thì không đặt câu hỏi về chủ ngữ. Với trường hợp vị ngữ, bộ lọc của chúng ta sẽ có sự khác biệt một chút về các mối quan hệ trong đây, ngoài việc *root* phải là động từ, thì chúng ta sẽ chấp nhận phụ thuộc *ccomp*, như đã nêu ở dạng câu có mệnh đề bổ ngữ cho vị ngữ **<Chủ ngữ 1> <Vị ngữ 1> <Chủ ngữ 2> <Vị ngữ 2>.** Ngoài ra chúng tôi cần phải xét xem vị ngữ chính, root, có phải là động từ không độc lập hay không. Nếu phải thì chúng ta sẽ không phát sinh câu hỏi "làm gì", cụ thể trong ví dụ dưới đây:

Tôi muốn một chiếc xe hơi.

=> *Tôi muốn cái gì?*

Ai muốn một chiếc xe hơi?

Bên cạnh những loại câu nêu trên, chúng tôi còn phát hiện những trường hợp đặc biệt sau của cấu trúc vị ngữ là động từ là: động từ mang ý nghĩa kết nối *("pha", "trộn", v.v),* động ngữ mang thành tố chính là ngữ khứ hồi ("*đi...về*"), *root* là động từ sở hữu có, và cuối cùng là sau vị ngữ động từ là một tính từ bổ nghĩa cho động từ. Với trường hợp

vị ngữ là động từ sở hữu có, chúng tôi quyết định cho nó tham gia vào danh sách các từ không độc lập, và không phát sinh câu hỏi "*làm gì*?", ví dụ như:

Bình có một cây đàn guitar cũ.

=> Bình có cái gì?

Và một trường hợp nữa đó chính là tính từ bổ ngữ cho động từ làm vị ngữ, quan hệ *acomp*. Sau khi xác định được quan hệ acomp bằng cây phân tích cú pháp, chúng ta sẽ thay thế nhánh con của acomp bằng cụm từ "*như thế nào?*". Xét ví dụ sau đây:

Bình chơi đàn rất hay.

=> Bình chơi đàn như thế nào?

4.1.5 Câu bị động

Đây là loại câu có động từ với ý nghĩa may rủi, nguyên nhân. Thường được dùng các từ bị, được, do, bởi, phải, mắc, nhưng thông dụng nhất là các từ bị, được, phải.

Ví dụ:

Tôi bị phạt

Tôi được khen.

Tôi được thầy khen.

Để đặt câu hỏi trong loại câu này cũng rất đơn giản, về phần chủ ngữ, chúng tôi vẫn sẽ đặt câu hỏi "*ai?", "con gì?", "cái gì?*" nếu là danh ngữ. Và câu hỏi cho vị ngữ cũng không khó khăn khi chúng ta chỉ cần thay thế vị ngữ, *root* bằng từ để hỏi "*gì*" và bỏ *agent* nếu có. Một câu hỏi nữa để chúng ta khai thác ở dạng câu bị động đó chính là hỏi về người chủ động (*agent*), nếu trong câu có tồn tại một *agent* (được xác định bởi cây phân tích phụ thuộc) thì sẽ có hai cách để sinh câu hỏi. Một là thay thế cụm từ agent bằng từ để hỏi "*ai*", cách hai là chúng ta khai báo một template có kiểu "*Ai + vị ngữ (bỏ động từ bị động) + chủ ngữ*". Mô hình của chúng tôi cho ra kết quả câu hỏi loại này là:

Minh được thầy khen.

=> *Ai được thầy khen?*

Minh được ai khen? / Ai khen Minh?

Minh được gì?

4.2Mô hình tách các vế trong câu ghép đẳng lập thành các câu đơn

Mục này, chúng tôi sẽ dựa vào quan hệ *parataxis* hoặc quan hệ *conjunction* để phân tách các vế trong câu ghép đẳng lập thành câu đơn, hỗ trợ việc xử lý, phát sinh câu hỏi tự động trên câu đơn trở nên dễ dàng hơn. Sau đây là một số trường hợp câu ghép đẳng lập (các vế không liên quan đến nhau):

Anh ấy đi du lịch; anh ấy là cổ đông của công ty.

Thầy giáo bước vào, cả lớp đứng lên.

Anh ấy đi du lịch và cô ấy đi làm.

Hai ví dụ đầu tiên, đó là quan hệ *parataxis*, ví dụ cuối cùng là quan hệ *conj*. Trong câu có tồn tại quan hệ *parataxis* hoặc *conj* giữa *root* và một từ bất kỳ, thì chúng ta sẽ xét hai vế này có chủ ngữ hay không, trong điều kiện hai vế này tồn tại đầy đủ chủ ngữ và vị ngữ, chúng ta sẽ tiến hành tách các vế câu, trong quá trình này, chúng ta sẽ lược bỏ những từ nối như "*và*" và các dấu câu. Sau đây là kết quả của mô hình:

Anh ấy đi du lịch; anh ấy là cổ đông của công ty.

=> *Anh ấy đi du lịch.*

Anh ấy là cổ đông của công ty.

Thầy giáo bước vào, cả lớp đứng lên.

=> *Thầy giáo bước vào.*

Cả lớp đứng lên.

Anh ấy đi du lịch và cô ấy đi làm.

=> *Anh ấy đi du lịch.*

Cô ấy đi làm.

CHƯƠNG 5 – THỬ NGHIỆM VÀ ĐÁNH GIÁ MÔ HÌNH

5.1Kho ngữ liệu tiếng Việt (Corpus)

Chúng tôi đã đánh giá hệ thống của mình bằng cách sử dụng một kho tài liệu gồm 1000 câu tiêu đề của các bài báo [26], phù hợp với khả năng phát sinh câu đơn của mô hình.

5.2Kết quả mô hình

Sau đây là kết quả của mô hình phát sinh đối với những câu đơn có cấu trúc đơn giản:

```
Lan là học sinh giỏi nhất lớp. -------- [['Học sinh giỏi nhất lớp có tên là gì?', 'Lan'], ['Lan là gì?', 'Học sinh gi
ỏi nhất lớp']]
Bác nông dân dắt con trâu đi cày ruộng. -------- [['Ai dắt con trâu đi cày ruộng?', 'Bác nông dân'], ['Bác nông dân d
ắt con trâu làm gì?', 'Đi cày ruộng'], ['Bác nông dân dắt con gì đi cày ruộng?', 'Con trâu']]
Những người thanh niên kéo nhau đi tình nguyện. -------- [['Ai kéo nhau đi tình nguyện?', 'Những người thanh niên'],
['Những người thanh niên làm gì?', 'Kéo nhau đi tình nguyện']]
Bác Hồ là tấm gương sáng của dân tộc Việt Nam. -------- [['Tấm gương sáng của dân tộc Việt Nam có tên là gì?', 'Bác H
ồ'], ['Bác Hồ là gì?', 'Tấm gương sáng của dân tộc Việt Nam']]
Thầy Trần Thanh Phước là giáo sư trường Đại học Tôn Đức Thắng. -------- [['Giáo sư trường Đại học Tôn Đức Thắng có tê
n là gì?', 'Thầy Trần Thanh Phước'], ['Thầy Trần Thanh Phước là gì?', 'Giáo sư trường Đại học Tôn Đức Thắng']]
Tí học rất chăm chỉ. -------- [['Tí học có đặc điểm gì?', 'Rất chăm chỉ']]
Con trâu gặm cỏ non. -------- [['Con gì gặm cỏ non?', 'Con trâu'], ['Con trâu làm gì?', 'Gặm cỏ non']]
Con trai ông hàng xóm đỗ vào trường Y Dược. -------- [['Ai đỗ vào trường Y Dược?', 'Con trai ông hàng xóm'], ['Con tr
ai ông hàng xóm làm gì?', 'Đỗ vào trường Y Dược']]
An đi Hà Nội. -------- [['Ai đi Hà Nội?', 'An'], ['An làm gì?', 'Đi Hà Nội']]
Anh bác sĩ chăm sóc bệnh nhân chu đáo. -------- [['Ai chăm sóc bệnh nhân chu đáo?', 'Anh bác sĩ'], ['Anh bác sĩ làm g
ì?', 'Chăm sóc bệnh nhân chu đáo']]
Con trâu có làn da xám xịt. -------- [['Con gì có làn da xám xịt?', 'Con trâu'], ['Con trâu có cái gì?', 'Có làn da x
ám xịt']]
Nguyễn Ngọc Ngạn rất đẹp trai. -------- [['Ai rất đẹp trai?', 'Nguyễn Ngọc Ngạn'], ['Nguyễn Ngọc Ngạn có đặc điểm g
ì?', 'Rất đẹp trai']]
Elon Musk là tỷ phú. -------- [['Tỷ phú có tên là gì?', 'Elon Musk'], ['Elon Musk là gì?', 'Tỷ phú']]
Hoa hậu hoàng vũ năm 2019 thật xinh xắn. -------- [['Hoa hậu hoàng vũ năm 2019 có đặc điểm gì?', 'Thật xinh xắn']]
Bà Nguyễn Thị Thu Hà là mẹ của cô dâu Thu Trinh. -------- [['Mẹ của cô dâu Thu Trinh có tên là gì?', 'Bà Nguyễn Thị T
hu Hà'], ['Bà Nguyễn Thị Thu Hà là ai?', 'Mẹ của cô dâu Thu Trinh']]
Bác hàng xóm đi du lịch. -------- [['Ai đi du lịch?', 'Bác hàng xóm'], ['Bác hàng xóm làm gì?', 'Đi du lịch']]
Cô ấy mua bánh mì. -------- [['Ai mua bánh mì?', 'Cô ấy'], ['Cô ấy làm gì?', 'Mua bánh mì']]
Con ngựa chạy đua với con lừa. -------- [['Con gì chạy đua với con lừa?', 'Con ngựa'], ['Con ngựa làm gì?', 'Chạy đua
với con lừa']]
Con ngựa chạy rất nhanh. -------- [['Con gì chạy rất nhanh?', 'Con ngựa'], ['Con ngựa chạy như thế nào?', 'Rất nhan
h']]
Tấm là nàng dâu hiếu thảo. -------- [['Nàng dâu hiếu thảo có tên là gì?', 'Tấm'], ['Tấm là gì?', 'Nàng dâu hiếu thả
o']]
Cám rất nhạm hiểm. -------- [['Ai rất nhạm hiểm?', 'Cám'], ['Cám có đặc điểm gì?', 'Rất nhạm hiểm']]
```

Hình 5.1 Kết quả mô hình phát sinh đối với những câu đơn

```
Tấm là nàng dâu hiếu thảo. -------- [['Nàng dâu hiếu thảo có tên là gì?', 'Tấm'], ['Tấm là gì?', 'Nàng dâu hiếu thả
o']]
Cám rất nham hiểm. -------- [['Ai rất nham hiểm?', 'Cám'], ['Cám có đặc điểm gì?', 'Rất nham hiểm']]
Mẹ con Cám rất độc ác. -------- [['Ai rất độc ác?', 'Mẹ con Cám'], ['Mẹ con Cám có đặc điểm gì?', 'Rất độc ác']]
Hoàng tử vô cùng mê gái đẹp. -------- [['Cái gì vô cùng mê gái đẹp?', 'Hoàng tử'], ['Hoàng tử có đặc điểm gì?', 'Vô c
ùng mê gái đẹp']]
Thầy Trần Thanh Phước khoa Công Nghệ Thông Tin dạy rất hay. -------- [['Ai dạy rất hay?', 'Thầy Trần Thanh Phước khoa
Công Nghệ Thông Tin'], ['Thầy Trần Thanh Phước khoa Công Nghệ Thông Tin dạy như thế nào?', 'Rất hay']]
Con ếch tìm thức ăn. -------- [['Con ếch làm gì?', 'Tìm thức ăn']]
Con người làm việc cực khổ. -------- [['Ai làm việc cực khổ?', 'Con người'], ['Con người làm việc như thế nào?', 'Cực
khổ']]
Bà già Noel tặng quà cho trẻ em. -------- [['Ai tặng quà cho trẻ em?', 'Bà già Noel'], ['Bà già Noel tặng quà cho a
i?', 'Cho trẻ em'], ['Bà già Noel tặng cho trẻ em cái gì?', 'Quà']]
```

Hình 5.2 Kết quả mô hình phát sinh đối với những câu đơn

Còn dưới đây là kết quả của corpus trên, những câu có cấu trúc phức tạp hơn hẳn các các câu văn trên.

```
Chây ì nộp phạt nguội. -------- [['Cái gì nộp phạt nguội?', 'Chây ì'], ['Chây ì làm gì?', 'Nộp phạt nguội']]
Phổ biến nhất là lỗi đỗ không đúng nơi quy định. -------- [['Phổ biến nhất là gì?', 'Lỗi đỗ không đúng nơi quy đị
n'], ['Lỗi đỗ không đúng nơi quy định là gì?', 'Phổ biến nhất']]
Điều này có thể thấy,. -------- [['Cái gì có thể thấy,?', 'Điều này'], ['Điều này làm gì?', 'Có thể thấy,']]
Họ sử dụng công nghệ hiện đại từ lâu để giảm lực lượng xử phạt tại chỗ. -------- [['Ai sử dụng công nghệ hiện đại từ
lâu để giảm lực lượng xử phạt tại chỗ?', 'Họ'], ['Họ sử dụng công nghệ hiện đại làm gì?', 'Để giảm lực lượng xử phạt
tại chỗ'], ['Họ sử dụng ai từ lâu để giảm lực lượng xử phạt tại chỗ?', 'Công nghệ hiện đại']]
PC67 đưa ra các nguyên nhân dẫn đến người nộp phạt đạt tỷ lệ thấp nói trên là cách trả lời chưa thuyết phục. --------
[['Ai đưa ra các nguyên nhân dẫn đến người nộp phạt đạt tỷ lệ thấp nói trên là cách trả lời chưa thuyết phục?', 'PC6
7'], ['PC67 đưa đi đâu?', 'Đưa ra các nguyên nhân dẫn đến người nộp phạt đạt tỷ lệ thấp nói trên là cách trả lời chưa
thuyết phục']]
Ông Sanh nói. -------- [['Ai nói?', 'Ông Sanh'], ['Ông Sanh làm gì?', 'Nói']]
Đây là cội nguồn. -------- [['Cội nguồn là gì?', 'Đây'], ['Đây là gì?', 'Cội nguồn']]
Nguyên nhân dẫn đến việc cưỡng chế người vi phạm nộp phạt không hiệu quả. -------- [['Ai dẫn đến việc cưỡng chế người
vi phạm nộp phạt không hiệu quả?', 'Nguyên nhân'], ['Nguyên nhân làm gì?', 'Dẫn đến việc cưỡng chế người vi phạm nộp
phạt không hiệu quả']]
Ông Sanh nói. -------- [['Ai nói?', 'Ông Sanh'], ['Ông Sanh làm gì?', 'Nói']]
Cháu đòi tiền cơm. -------- [['Cái gì đòi tiền cơm?', 'Cháu'], ['Cháu làm gì?', 'Đòi tiền cơm']]
Dì đòi tiền nhà. -------- [['Ai đòi tiền nhà?', 'Dì'], ['Dì làm gì?', 'Đòi tiền nhà']]
Hai bên đã hoàn tất thủ tục đăng bộ sang tên. -------- [['Cái gì đã hoàn tất thủ tục đăng bộ sang tên?', 'Hai bên'],
['Hai bên làm gì?', 'Đã hoàn tất thủ tục đăng bộ sang tên']]
Bà S có hứa khi nào chết sẽ để lại cho ông căn nhà này. -------- [['Ai có hứa khi nào chết sẽ để lại cho ông căn nhà
này?', 'Bà S'], ['Bà S có cái gì?', 'Có hứa khi nào chết sẽ để lại cho ông căn nhà này']]
Ông T từng làm nghề sản xuất vỏ xe ). -------- [['Ai từng làm nghề sản xuất vỏ xe )?', 'Ông T'], ['Ông T làm gì?', 'T
ừng làm nghề sản xuất vỏ xe )']]
Ông T kháng cáo. -------- [['Ai kháng cáo?', 'Ông T'], ['Ông T làm gì?', 'Kháng cáo']]
Một thẩm phán khuyên ông T. -------- [['Cái gì khuyên ông T?', 'Một thẩm phán'], ['Một thẩm phán làm gì?', 'Khuyên ôn
g T']]
Ông làm ăn trên đất của bà ấy mấy chục năm nay là đã mang một ân tình lớn. -------- [['Ông làm ăn trên đất của bà ấy
làm gì?', 'Mấy chục năm nay là đã mang một ân tình lớn']]
```

Hình 5.3 Kết quả mô hình phát sinh đối với những câu có cấu trúc phức tạp hơn

5.3Đánh giá mô hình

Độ chính xác, tính hợp lý của các câu hỏi trong mô hình của chúng tôi tỉ lệ thuận với độ chính xác trong mô hình cây cú pháp phụ thuộc, nếu một câu bị gán các nhãn quan hệ phụ thuộc sai thì câu hỏi cũng trở nên không đúng hoặc mô hình không phát sinh được câu hỏi. Ngược lại nếu cấu trúc phụ thuộc chính xác thì câu hỏi của chúng tôi sẽ đảm bảo được như kỳ vọng.

Không có cách tiêu chuẩn nào để đánh giá đầu ra của hệ thống QG. Chúng tôi sử dụng *recall* và *precision*, và áp dụng chúng như sau:

Precision = correct /(correct + spurious)

Recall = correct / (correct + missed)

*F-score = 2*Precision*Recall / (Precision + Recall)*

Trong đó correct là số lượng câu hỏi được tạo bởi hệ thống QG và cũng có trong các câu hỏi được tạo theo cách thủ công từ một tài liệu, spurious là số lượng câu hỏi

được tạo bởi hệ thống QG nhưng không có trong các câu hỏi được tạo thủ công và missed là số lượng các câu hỏi có trong các câu hỏi được tạo thủ công nhưng không có trong các câu hỏi do hệ thống QG tạo ra. Chúng tôi sẽ đánh giá mô hình tạo câu hỏi theo phương pháp này.

Dựa vào những câu chúng tôi đã tạo được ở phần trên thì kết quả đánh giá các giá trị *Recall*, *Precision* và *F-Score* như sau:

Số lượng từ trong câu văn	Precision	Recall	F-Score
0-5	0.57	0.80	0.66
5-10	0.48	0.71	0.57
10-15	0.42	0.66	0.51
Hơn 15	0.38	0.58	0.46
Giá trị trung bình	0.46	0.68	0.55

Bảng 5.1 Kết quả đánh giá mô hình

CHƯƠNG 6 – KẾT LUẬN

6.1Những vấn đề đã đạt được trong bài báo cáo

Tìm hiểu về phương pháp phát sinh câu hỏi tự động, là một trong những bài toán có ý nghĩa thức tiễn cao của lĩnh vực xử lý ngôn ngữ tự nhiên.

Xây dựng mô hình phát sinh câu hỏi tự động trên câu đơn dựa vào phương pháp phân tích cú pháp (syntax-based) và khuôn mẫu (template-based) cùng với các công cụ xử lý ngôn ngữ tự nhiên có sẵn của mã nguồn mở Underthesea [14].

Xây dựng mô hình tách các vế câu trong câu ghép đẳng lập thành các câu đơn, hỗ trợ mạnh mẽ cho mô hình phát sinh câu hỏi tự động trên câu đơn.

6.2Những vấn đề cần phát triển

Cải tiến hiệu quả của mô hình phát sinh câu hỏi động dựa trên câu đơn và nghiên cứu, tìm hiểu thêm một số template và dạng câu hỏi tốt hơn, tránh các câu hỏi factoid, mang tính giáo dục trong tiếng Việt.

Tìm hiểu phát triển mô hình phát sinh câu hỏi tự động trên câu ghép và các kiểu câu có cấu trúc phức tạp.

Xây dựng một công cụ phát sinh câu hỏi hoàn thiện cho tiếng Việt không chỉ dừng lại ở mức câu, mà có thể phát sinh câu hỏi mức đoạn văn hoặc cả văn bản.

TÀI LIỆU THAM KHẢO

[1] Đào Minh Thu, Đào Thị Minh Ngọc, Nguyễn Mai Vân, Lê Kim Ngân, Lê Thanh Hương, Nguyễn Phương Thái, Đỗ Bá Lâm (2009), *Tập quy tắc cú pháp tiếng Việt, Báo cáo kỹ thuật VLSP.*

[2] Đinh Điền, nhóm biên soạn VCL (2008), *Hướng dẫn tách câu Tiếng Việt, Báo cáo kỹ thuật VLSP.*

[3] David Lindberg (2013), *Automatic Question Generation From Text For Self-Directed Learning, Master degree of Simon Fraser University*

[4] Robert Dale, Hermann Moisl, Harold Somers (2000), *Handbook of Natural Language Processing*

[5] Phạm Nguyên Khang, Trần Nguyễn Minh Thư, Phạm Thế Phi, Đỗ Thanh Nghị (2016), *Sự ảnh hưởng của phương pháp tách từ trong bài toán phân lớp văn bản tiếng Việt*

[6] Nghia Ticy, Hung Le, *https://streetcodevn.com/blog/vntok*

[7] Trần Thu Trang (2012), *Nghiên cứu gán nhãn từ loại cho văn bản tiếng Việt bằng phương pháp học máy không có hướng dẫn*

[8] Deepika Kumawat, Vinesh Jain (2015), *POS Tagging Approaches: A comparison*

[9] Phạm Thị Thu Trang (2018), *Nhận dạng thực thể định danh từ văn bản ngắn tiếng Việt và đánh giá thực nghiệm*

[10] Lhioui Chahira, Zouaghi Anis, Zrigui Mounir (2017), *A rule-based Named Entity Method and Syntactico-Semantic Annotation for Arabic Language*

[11] Thinh Hung Truong, Mai Hoang Dao, Dat Quoc Nguyen, *COVID-19 Named Entity Recognition for Vietnamese*

[12] Luong Nguyen Thi, Linh Ha My, Hung Nguyen Viet, Huyen Nguyen Thi Minh, Phuong Le Hong (2013), *Building a Treebank for Vietnamese Dependency Parsing*

[13] Hà Mỹ Linh (2015), *Phân tích cú pháp phụ thuộc tiếng Việt*

[14] Saidalavi Kalady, Ajeesh Elikkottil, and Rajarshi Das. Natural language question generation using syntax and keywords. In Proceedings of QG2010: The Third Workshop on Question Generation, pages 1–10, 2010.

[15] Andrea Varga and Le An Ha. Wlv: A question generation system for the qgstec 2010 task b. In Proceedings of QG2010: The Third Workshop on Question Generation, pages 80–83, 2010.

[16] John H Wolfe. Automatic question generation from text-an aid to independent study. In ACM SIGCUE Outlook, volume 10, pages 104–112. ACM, 1976.

[17] Husam Ali, Yllias Chali, and Sadid A Hasan. Automation of question generation from sentences. In Proceedings of QG2010: The Third Workshop on Question Generation, pages 58–67, 2010.

[18] Michael Heilman and Noah A Smith. Extracting simplified statements for factual question generation. In Proceedings of QG2010: The Third Workshop on Question Generation, pages 11–20, 2010.

[19] Prashanth Mannem, Rashmi Prasad, and Aravind Joshi. Question generation from paragraphs at upenn: Qgstec system description. In Proceedings of QG2010: The Third Workshop on Question Generation, pages 84–91, 2010.

[20] Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cris- tian Moldovan. Overview of the first question generation shared task evaluation chal- lenge. In Proceedings of the Third Workshop on Question Generation, pages 45–57, 2010.

[21] Zhiqiang Cai, Vasile Rus, Hyun-Jeong Joyce Kim, Suresh C. Susarla, Pavan Karnam, and Arthur C. Graesser. Nlgml: A markup language for question generation. In

Thomas Reeves and Shirley Yamashita, editors, Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2006, pages 2747–2752, Honolulu, Hawaii, USA, October 2006. AACE.

[22] Brendan Wyse and Paul Piwek. Generating questions from openlearn study units. In AIED 2009: 14 th International Conference on Artificial Intelligence in Education Workshops Proceedings, 2009.

[23] Roger Levy and Galen Andrew. Tregex and tsurgeon: tools for querying and manipu- lating tree data structures. In LREC 2006, 2006.

[24] Jack Mostow and Wei Chen. Generating instruction automatically for the reading strat- egy of self-questioning. In Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling, pages 465–472. IOS Press, 2009.

[25] Underthesea, Vietnamese python open source NLP, *https://pypi.org/project/underthesea/*

[26] Title Corpus Vietnamese (2018), news corpus, *https://github.com/binhvq/news-corpus/blob/master/sample/demo-title.txt*