Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.000112

ỨNG DỤNG HỖ TRỢ TRA CỨU CỤM TỪ DÙNG TRONG BÀI BÁO KHOA HỌC BẰNG TIẾNG ANH

Đặng Văn Thìn, Nguyễn Văn Kiệt, Nguyễn Lưu Thùy Ngân

Trường Đại học Công nghệ thông tin, ĐHQG Tp. Hồ Chí Minh

dangvanthin.uit@gmail.com, kietnv@uit.edu.vn, ngannlt@uit.edu.vn

TÓM TẮT: *Tiếng Anh là một ngôn ngữ quốc tế được sử dụng để trình bày các công trình nghiên cứu khoa học tại các hội nghị trên thế giới. Tuy nhiên, việc sử dụng tiếng Anh một cách thành thạo là một khó khăn đối với người của những nước không sử dụng tiếng Anh, bao gồm Việt Nam và các nước khác trên thế giới. Để viết một bài báo khoa học bằng tiếng Anh, chúng ta phải dành nhiều thời gian để tìm hiểu và tra cứu cách dùng các cụm từ chưa biết. Trong bài báo này, chúng tôi trình bày kết quả thử nghiệm phương pháp rút trích các cụm từ từ kho bài báo khoa học có sẵn và xây dựng một ứng dụng hỗ trợ tra cứu cụm từ này để giúp người dùng tham khảo và học các cấu trúc, cụm từ chuyên dụng trong viết bài báo khoa học bằng tiếng Anh. Ứng dụng được mong đợi sẽ giúp người dùng diễn tả chính xác nội dung trong bài báo khoa học tiếng Anh.*

Từ khóa: *hỗ trợ viết tài liệu học thuật, tìm kiếm cụm từ, tìm kiếm ví dụ.*

I. GIỚI THIỆU

Ngày nay, nghiên cứu khoa học ngày càng được đầu tư và phát triển ở các quốc gia và các trường đại học. Trong hoạt động nghiên cứu khoa học nói chung thì bài báo khoa học đóng một vai trò cực kỳ quan trọng. Bài báo khoa học không chỉ là một sản phẩm tri thức mà còn là minh chứng cho khả năng của những người làm khoa học đặc biệt là các bài báo khoa học quốc tế. Từ đó, chúng ta nhận thấy rằng tầm quan trọng của một bài báo khoa học quốc tế và tiếng Anh luôn là ngôn ngữ chính được sử dụng để trình bày nội dung. Tuy nhiên, đối với những người không sử dụng tiếng Anh là ngôn ngữ chính thì khi viết bài báo khoa học bằng tiếng Anh sẽ gặp rất nhiều khó khăn. Họ phải tra cứu các thuật ngữ cụm từ trong các từ điển hay các sách hướng dẫn viết bài báo để chắc chắn rằng từ ngữ, cụm từ diễn tả chính xác nội dung của họ. Điều đó cho thấy rằng, họ sẽ tốn rất nhiều thời gian để hoàn thành chính xác nội dung bài báo của mình. Thêm vào đó, việc tìm kiếm các thông tin không phải tiếng bản ngữ là một khó khăn và các công cụ dịch thuật cũng chưa hiệu quả và đạt kết quả như mong muốn.

Nhận thấy những khó khăn đó nên năm 2016, các nhà nghiên cứu ở Ý đã ra mắt công cụ tìm kiếm ngôn ngữ Lugdwig [1] để giúp mọi người viết các câu tiếng Anh chuẩn xác qua việc tìm ra các mẫu câu tương tự trong các bài báo đăng tải trên các trang tin nổi tiếng như New York Time, BBC,… Tuy nhiên, công cụ này chỉ hỗ trợ viết tiếng Anh thông thường, còn hạn chế cho các lĩnh vực nghiên cứu khoa học và học thuật. Bên cạnh đó, tác giả Yu-Chih Sun [2] tại trường đại học National Chiao Tung University, Taiwan đã xây dựng một ứng dụng website hỗ trợ viết bài báo khoa học và tra cứu các cụm từ hữu ích. Tuy nhiên, phương pháp của tác giả sử dụng các cụm từ rút trích bằng tay, cụ thể là các sinh viên phân công để đọc và lựa chọn các cụm từ hữu ích trong bài báo, sau đó các giáo viên có kinh nghiệm sẽ kiểm tra xem cụm từ đó có phải là cụm từ hữu ích trong viết báo hay không. FLOW (Chen et al, 2012) [3] là một hệ thống hỗ trợ viết tương tác trực tiếp với người dùng nhằm đề xuất ra các cụm từ được sử dụng trong viết bài báo khoa học dành cho người Trung Quốc. Điều đặc biệt của hệ thống là giúp người dùng hoàn thành được bài báo mà không bị gián đoạn bởi các từ vựng mà họ không biết trong tiếng Anh. Đối với tiếng Việt hiện nay, có trang web "www.hellochao.com" có chức năng hỗ trợ người học tiếng Anh, khi người dùng truy vấn từ khóa tiếng Việt thì ứng dụng sẽ tìm kiếm trong cơ sở dữ liệu được xây dựng sẵn để đưa ra các câu nói tiếng Anh hoàn chỉnh thường được dùng để giao tiếp hàng ngày. Hệ thống cơ sở dữ liệu của website có khoảng 300.000 cặp câu song ngữ Việt-Anh, các câu tiếng Anh được chuyển ngữ thành các câu tiếng Việt phù hợp với ngữ cảnh và văn hóa.

Tận dụng nguồn ngữ liệu là các bài báo đã được công bố trên các hội nghị quốc tế uy tín, chúng tôi tiến hành thử nghiệm phương pháp rút trích tự động các cụm từ hữu ích và xây dựng ứng dụng hỗ trợ tra cứu cụm từ và câu ví dụ để người dùng có thể tham khảo và học được các cụm từ thường được sử dụng trong bài báo khoa học bằng tiếng Anh. Chúng tôi trình bày mô hình xây dựng ứng dụng tra cứu câu ví dụ dựa trên truy vấn với tra cứu cụm từ được sử dụng trong viết bài báo khoa học ở mục 2.1 và phương pháp rút trích tự động các cụm từ hữu ích ở mục 2.2. Còn đối với tập ngữ liệu được chúng tôi trình bày chi tiết ở mục 2.3.

II. CÀI ĐẶT CHƯƠNG TRÌNH

2.1. Ứng dụng tra cứu cụm từ

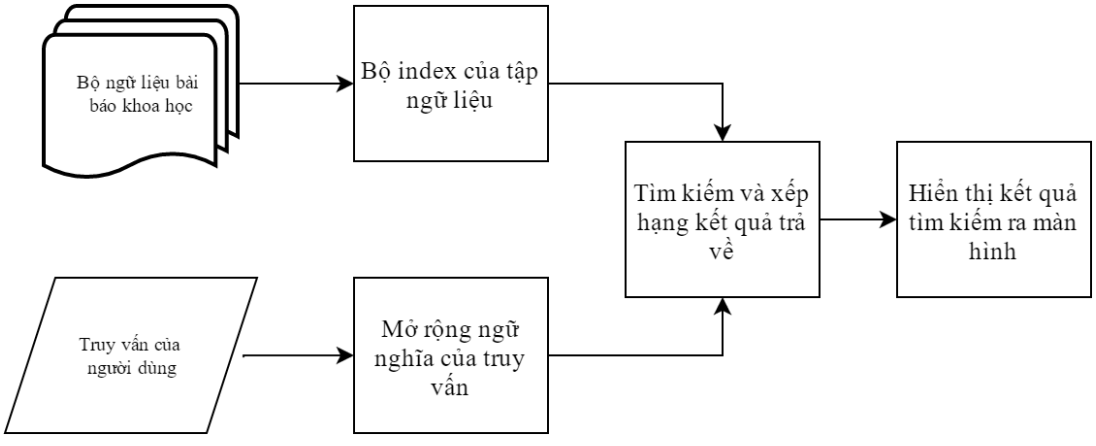
Chúng tôi phát triển ứng dụng hỗ trợ tra cứu cụm từ để hỗ trợ người dùng trong việc viết các bài báo khoa học bằng tiếng Anh. Ứng dụng gồm hai chức năng chính: 1) Tìm kiếm câu ví dụ - Trả về các câu ví dụ dựa trên truy vấn các cụm từ nhập vào từ người dùng; 2) Tra cứu thông tin cụm từ - Tra cứu các cụm từ, mẫu câu được sử dụng trong các bài báo khoa học.

2.1.1. Mô hình tìm kiếm ví dụ

Mục tiêu chính của ứng dụng này là giúp người dùng hình dung và kiểm tra cách sử dụng của các cụm từ trong các bài báo khoa học thông qua các câu ví dụ như tìm hiểu giới từ đi kèm, vị trí cụm từ trong câu, hoặc người dùng có thể biết cụm từ đó có được sử dụng trong viết bài báo khoa học hay không (ví dụ như người dùng không biết sử dụng giới từ đi kèm với cụm danh từ training data là "in" hay "on".). Chúng tôi cũng tiến hành khảo sát mười bạn sinh viên chưa có kinh nghiệm viết bài báo khoa học thì đánh giá chức năng này mang lại hiệu quả khi tra cứu cụm từ. Ví dụ như các sinh viên muốn dịch câu "*Chúng tôi thực hiện các tiến trình một cách đồng thời*" trong tiếng Việt sang tiếng Anh, họ sử dụng từ điển trực tuyến để tìm kiếm các từ có nghĩa "*đồng thời*" như "*at once, at the same time, simultaneously, ...*". Tuy nhiên, theo ngữ cảnh trong các bài báo khoa học thì từ "at once" có nghĩa tương đương "cùng một lúc" thông qua các câu ví dụ khi sử dụng phương pháp như "*Furthermore, a hashtag can encode multiple topics* ***at once***." hay "*They process several examples* ***at once*** *and use a short-list vocabulary v with only the most frequent words.*" Khi tìm kiếm các câu của vị của cụm từ "*as the same time*", kết quả trả về các câu ví dụ như "*To the best of our knowledge, this is the first model which trains two tasks* ***at the same time***." hoặc "***At the same time****, with the help of the yago knowledge, we borrow the distant supervision technique to mine the implicit facts from the text*.". Từ những câu ví dụ, sinh viên sẽ chọn cụm từ "at the same time" thay vì cụm từ "at once". Do đó, dựa vào các câu ví dụ được đưa ra bởi ứng dụng, người dùng sẽ tham khảo được ngữ cảnh sử dụng của cụm từ trong các bài báo khoa học một cách chính xác, trách các trường hợp sử dụng sai từ để làm sai ngữ nghĩa của câu văn. Dựa vào các ví dụ trên, người sử dụng cũng có thể tham khảo được vị trí của từ khi được dùng trong câu tiếng Anh như thế nào (ví dụ như cụm từ "at once" thường được sử dụng ở cuối câu trong khi cụm từ "at the same time" có thể sử dụng ở các vị trí khác nhau trong câu).

Ngoài ra, để thuận tiện cho người sử dụng, chúng tôi cũng phân lớp các câu ví dụ vào bảy lớp ("abstract", "introduction", "related work", "methods", "results and discussions", "conclusions", "acknowledgements") theo cấu trúc chuẩn của một bài báo khoa học quốc tế dựa vào tiêu đề của câu ví dụ khi được rút trích từ báo khoa học. Điều này giúp người dùng hình dung câu ví dụ nằm trong phần nào của bài báo.

Mô hình thực hiện chức năng tìm kiếm câu ví dụ được trình bày ở Hình 1 và mô tả chi tiết từng thành phần của mô hình được chúng tôi trình bày ngay sau đó.



Hình 1. Sơ đồ chức năng tìm kiếm câu ví dụ

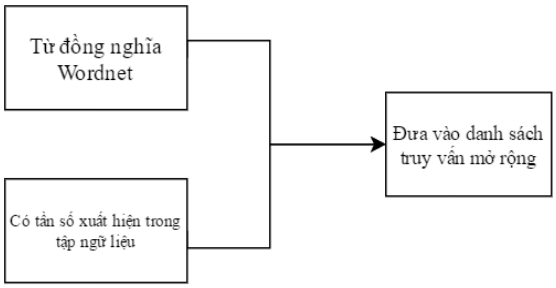
Truy vấn của người dùng: Truy vấn có thể là tiếng Anh hoặc tiếng Việt. Đối với tiếng Việt, hệ thống sẽ kiểm tra trong tập từ điển đã được chúng tôi thu thập từ các nguồn khác nhau [1], [2]. Sau đó các cụm từ này được dịch bởi các tác giả có kinh nghiệm trong lĩnh vực học thuật (ví dụ - query = "bài báo trình bày" thì ứng dụng sẽ trả ra các kết quả như "the paper presents", "the paper shows", …). Điều này sẽ giúp người dùng chưa có kinh nghiệm viết bài báo khoa học có thể tra cứu bằng tiếng Việt. Tuy nhiên, tập từ điển này rất hạn chế, vì thế chúng tôi tập trung vào việc truy vấn bằng tiếng Anh. Bên cạnh đó, để đa dạng truy vấn, ứng dụng cũng chấp nhận truy vấn thiếu như "we develop a * method to …" – điều này giúp cho người dùng đa dạng lựa chọn các tính từ phù hợp với nội dung của câu.

Bộ index: Tập ngữ liệu sẽ được chúng tôi đánh chỉ mục. Các đề mục trong bài báo được phân loại thành bảy lớp theo bộ cục của một bài báo khoa học. Nhằm biết vị trí tương đối của câu ví dụ trong bài báo khoa học, chúng tôi thực hiện một phương pháp phân lớp đơn giản dựa trên các từ khóa của đề mục. Đề mục X sẽ được phân vào lớp Y nếu X chứa các từ dấu hiệu trong lớp Y. Chúng tôi nhận thấy rằng, mỗi bài báo khoa học đều có đề mục "Introduction", thế nên chúng tôi sử dụng luật heuristic là sau khi đề mục "Introduction" mà đề mục X không có các từ dấu hiệu trong lớp Y thì sẽ được phân lớp vào lớp "Methods". Bảng 1 liệt kê ra các cụm từ dấu hiệu cho mỗi lớp. Chúng tôi lựa chọn khoảng 150 bài báo trong tập ngữ liệu để đánh giá kết quả. Kết quả độ chính của phân lớp này là 95,6% (1097/1147).

Bảng 1. Các cụm từ dấu hiệu của mỗi đề mục trong bố cục bài báo khoa học tiếng Anh

Đề mục	**Cụm từ dấu hiệu**
Abstract	Abstract
Introduction	Introduction
Related Word	Past work, Related work, Previous work, Recent work, Overview
Methods	Data, Model, Framework, Approach, Corpora, Method, Background, System
Results and Discussions	Result, Evaluation, Experiment, Analysis, Discussion
Conclusion	Conclusion, Future Work
Acknowledgements	Acknowledgements

Mở rộng truy vấn ngữ nghĩa: Mở rộng query là một bước rất cần thiết dựa trên tập ngữ liệu nhỏ. Thành phần này của mô hình sẽ làm đa dạng và phong phú các câu ví dụ. Hiện tại, chúng tôi chỉ tập trung giải quyết mở rộng truy vấn ngữ nghĩa như sau: Nếu truy vấn là một từ đơn thì sẽ mở rộng các từ đồng nghĩa, còn đối với truy vấn dài chúng tôi mở rộng dựa trên động từ của câu truy vấn. Giả sử như câu truy vấn của người dùng là "*In this paper, we present ...*" thì việc ứng dụng mở rộng truy vấn dựa trên ngữ nghĩa của động từ sẽ là "*present*" được mở rộng với các động tự khác như "*show, demo, introduce, represent, ...*" sẽ trả về kết quả thêm của các truy vấn "In this paper, we show …" hoặc vấn "*In this paper, we introduce ...*". Từ đó người dùng có nhiều lựa chọn cho bài báo của mình cũng như học thêm được các từ ngữ tương đương. Chúng tôi sử dụng từ đồng nghĩa Wordnet[1] để mở rộng ngữ nghĩa dựa trên động từ của truy vấn.



Hình 2. Mô hình mở rộng truy vấn ngữ nghĩa

Tìm kiếm và xếp hạng: Các câu ví dụ của cụm từ truy vấn trong tập ngữ liệu đã đánh index sẽ được tìm kiếm và xếp hạng để người dùng có thể tham khảo một cách tốt nhất. Để xếp hạng các kết quả của truy vấn, chúng tôi sử dụng độ đo TF-IDF để xếp hạng kết hợp với trên độ dài của câu ví dụ. Việc xếp hạng sẽ đưa ra các câu ví dụ càng ngắn thì người dùng sẽ có khả năng hiểu được nội dung ngữ cách và cách sử dụng cụm từ hiệu quả. Hình 3 là một kết quả trả về cho người dùng của truy vấn "*we present a * method*". Trong một vài trường hợp người dùng sẽ không hiểu rõ được nội dung của câu ví dụ, vì thế chúng tôi đưa ra thêm một đoạn văn của câu ví dụ bao gồm câu trước, câu ví dụ và câu sau để người dùng tham khảo và hiểu được nghĩa của cụm từ truy vấn theo ngữ cảnh. Nếu người dùng muốn đọc toàn bộ bài báo thì nhấn vào link của tên bài báo khoa học ở bên dưới mỗi câu ví dụ thì báo báo sẽ được trực tiếp tìm kiếm trên google.

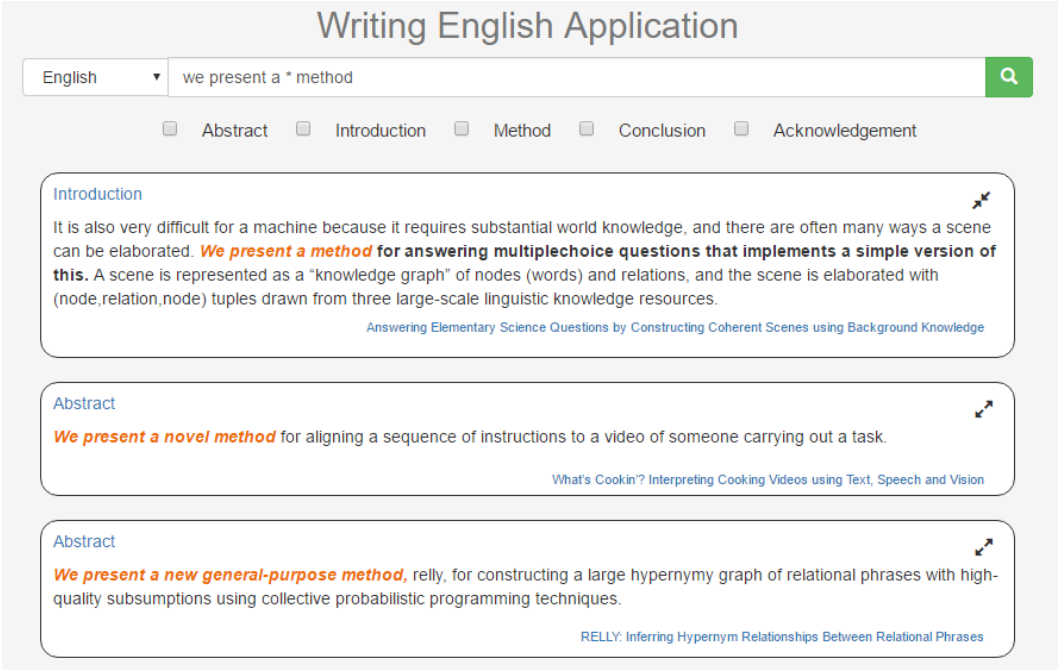
2.1.2. Mô hình tham khảo các cụm từ

Mục tiêu của mô hình này giúp người dùng tham khảo được các cụm từ, mẫu câu được gợi ý để sử dụng khi viết bài báo khoa học. Hiện tại, chúng tôi chỉ thu thập bằng tay và đưa vào cơ sở dữ liệu của chúng tôi các cụm từ, mẫu câu và nội dung mà cụm từ diễn tả đến trong bài báo khoa học (ví dụ như là cụm từ *"... is a classic problem in ...*" được dùng để diễn tả cho nội dung như "*Establishing the importance of the topic for the discipline*" trong bài báo khoa học). Bên cạnh đó, chúng tôi cũng phân loại nội dung này vào các đề mục chính của bài báo khoa học (ví dụ như nội dung "*Establishing the importance of the topic for the discipline*" thuộc đề mục trong bài báo là "*Introducing work*") và tìm kiếm các câu ví dụ chứa cụm từ trong tập ngữ liệu của chúng tôi để người dùng hình dung được cách sử dụng và ngữ nghĩa của câu. Ngoài ra, các cụm từ sẽ được dịch sang tiếng Việt để những người chưa có khả năng tiếng Anh tốt có thể hiểu và hình dung được ngữ nghĩa cụm từ. Tất cả các cụm từ chúng tôi thu thập đều được lấy từ các cuốn sách nổi tiếng về hướng dẫn viết bài báo khoa học bằng tiếng Anh [4], [5] để đưa vào cơ sở dữ liệu.

Ngoài ra, chúng tôi còn cho người dùng tham khảo các cụm từ hữu ích được rút trích tự động dựa trên tập ngữ liệu có sẵn. Bên cạnh việc tra cứu được các cụm từ hữu ích, người dùng còn xem thêm được nhiều ví dụ của cụm từ đó

[1] https://wordnet.princeton.edu/

trong tập ngữ liệu của chúng tôi. Với chức năng này, chúng tôi hy vọng ứng dụng sẽ giúp người dùng có thêm đa dạng nhiều lựa chọn để diễn tả nội dung, kết quả nghiên cứu trong bài báo khoa học.



Hình 3. Màn hình chụp của truy vấn "we present a * method" trên ứng dụng

2.2. Rút trích tự động cụm từ hữu ích

Các cụm từ hữu ích bao gồm các mẫu cụm từ, thuật ngữ, thành ngữ, cụm từ đi chung với nhau được sử dụng trong viết các bài báo khoa học (ví dụ *như "... is aligned with ...", "... is independent of ..." hay "Table 1 gives the performance of ..."*). Lưu ý rằng ở đây, chúng tôi không rút trích các cụm từ thuật ngữ chuyên môn trong lĩnh vực đặc biệt. Dựa vào kết quả phân tích của tác giả Kozawa et al. 2010, thì một cụm từ hữu ích phải có có các đặc điểm như sau:

- Cụm từ được sử dụng thường xuyên trong các bài báo khoa học.
- Độ dài các cụm từ không quá ngắn.
- Cụm từ có các từ đằng trước và các từ đằng sau khác nhau trong mỗi bài báo.

Dựa theo những phân tích ở trên, chúng tôi thực hiện phương pháp rút trích các cụm từ hữu ích theo như Hình 4 được đề xuất bởi Kozawa, nhưng chúng tôi bổ sung một bước để xác định xem cụm danh từ nào sẽ được thay thế bằng nhãn <NP> đối với mô hình của tác giả Kozawa et al 2010 [6]. Chúng tôi mô tả chi tiết các bước thực hiện ở sau đây.

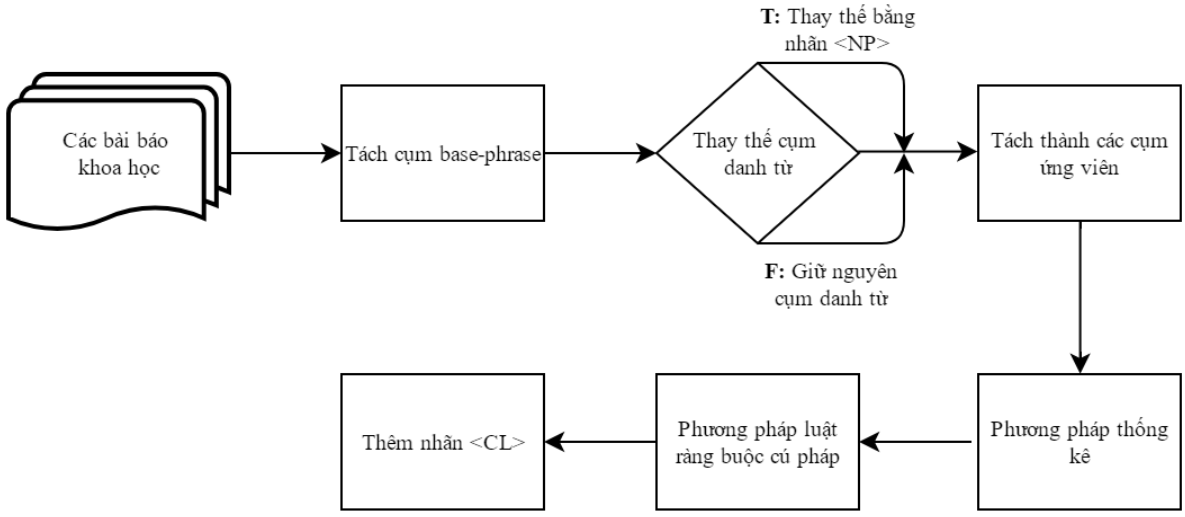
Đầu tiên chúng tôi phân tích các câu trong tập ngữ liệu thành các cụm base-phrase. Sau đó chúng tôi sẽ xác định xem cụm danh từ nào sẽ được thay thế thành nhãn <NP>. Bởi vì các cụm danh từ khi được phân tích có thể là các cụm danh từ chuyên ngành, các cụm danh từ riêng. Vì thế, để xác định xem cụm danh từ nào sẽ được thay thế thành nhãn <NP>, chúng tôi tiến hành rút trích các cụm danh từ trong các danh sách cụm từ hữu ích [4], [5] mà chúng tôi đã thu thập để tạo thành bộ từ điển các danh từ. Sau đó, chúng tôi sẽ xét trường hợp nếu cụm danh từ xuất hiện trong bộ từ điển này sẽ được giữ nguyên, trong trường hợp ngược lại cụm danh từ sẽ được thay thế bằng nhãn <NP>. Ngoài ra chúng tôi sẽ thay thế các trường hợp cụm danh từ kép như là <NP> <NP> hoặc <NP> of <NP> thành một nhãn <NP> theo những kết luận của Kozawa et al. 2010. Tiếp theo, chúng tôi tách các câu thành các chuỗi base-phrase – lưu ý là chuỗi base-phrase không quá bốn cụm danh từ <NP>, vì các cụm từ hữu ích thường không quá bốn cụm <NP>. Cuối cùng, chúng tôi sử dụng phương pháp thống kê và luật ràng buộc cú pháp để loại bỏ những cụm từ không hữu ích. Phương pháp thống kê và luật ràng buộc cú pháp sẽ được trình bày ở các mục 2.2.1 và mục 2.2.2 của bài báo. Kết quả đánh giá của phương pháp được chúng tôi trình bày ở mục 2.2.3.

2.3. Phương pháp thông kê

Chúng tôi sử dụng hàm tính toán của Ikeno et al để xác định các chuỗi cụm từ có phải là các cụm từ hữu ích hay không. Ý tưởng chính của bước này là xác định những cụm từ nào có các từ ngữ bên trái và các từ ngữ bên phải đa dạng trong tập ngữ liệu, đây là các cụm từ được sử dụng trong viết bài báo khoa học. Hàm tính toán được mô tả như sau:

$$\text{Lscore} = \log(\text{tf}(E)) \times \text{length}(E) \times \text{Hl}(E) \quad (1)$$

$$\text{Rscore} = \log(\text{tf}(E)) \times \text{length}(E) \times \text{Hr}(E) \quad (2)$$



Hình 4. Sơ đồ rút trích tự động các cụm từ hữu ích từ các bài báo khoa học tiếng Anh

Trong đó:

- E là chuỗi cụm từ đang được xét.
- Tf(E) là tần số xuất hiện của cụm từ E trong tập ngữ liệu bài báo khoa học.
- Length(E) là độ dài của chuỗi E.
- Hl(E), Hr(E) là phân bố xác suất của cụm base-phrase đằng trước và đằng sau của cụm từ E.

Phân bố xác suất Hl(E) và Hr(E) được tính toán theo công thức như sau:

$$\mathrm{Hl(E)} = -\sum_{i} Pl_i(E) log Pl_i(E) \tag{3}$$

$$\mathrm{Hr(E)} = -\sum_{i} Pr_i(E) log Pr_i(E) \tag{4}$$

Với Pl_i là xác suất của cụm từ đằng trước của E và Pr_i là xác suất của cụm từ đằng sau của E. Pl_i/Pr_i được tính toán bằng công thức xác suất như sau:

$$Pl_i(E) = P(X_i E|E) = \frac{P(X_i E)}{P(E)} \approx \frac{tf(X_i E)}{tf(E)} \tag{5}$$

$$Pr_i(E) = P(EX_i|E) = \frac{P(EX_i)}{P(E)} \approx \frac{tf(EX_i)}{tf(E)} \tag{6}$$

Cụm từ ứng viên E được rút trích ra nếu hàm Lscore(E) và hàm Rcore(E) thõa một trong các bất đẳng thức sau đây với XE/EX, trong đó X với các từ đằng trước và các từ đằng sau của cụm từ E.

$$\mathrm{Lscore(E)} > \mathrm{Lscore(XE)}$$

$$\mathrm{Rscore(E)} > \mathrm{Rscore(EX)}$$

2.4. Luật ràng buộc cú pháp

Khi phân tích kết quả kết quả sử dụng phương pháp thống kê ở mục 2.2.1, chúng tôi nhận thấy vẫn còn nhiều cụm từ được rút trích tự động không có hữu ích người dùng tham khảo (*ví dụ như là cụm từ "<NP> to improve <NP>"*, có tần số xuất hiện là 301, nhưng cụm từ lại không có ý nghĩa tham khảo đối với người dùng). Vì thế để loại bỏ các cụm từ như trên, chúng tôi lựa chọn ra khoảng 1000 cụm từ được rút trích thành công từ bước thống kê để xây dựng tập luật ràng buộc. Nếu một cụm từ thỏa một trong các luật ràng buộc cú pháp thì sẽ được loại bỏ khỏi kết quả cuối cùng. Các luật được chúng tôi liệt kê ở bảng 2 dưới đây.

Bảng 2. Luật ràng buộc cú pháp[2]

Nếu cụm từ có chứa các động từ thì đi theo sau không có các giới từ (in, on, to,…) hoặc cụm từ có chứa chủ từ như "we" nhưng không có động từ hay cụm từ có chứa "<NP> (or) ".
Bắt đầu hoặc kết thúc của cụm từ là từ "and" hoặc từ "if"
Các cụm từ bắt đầu bằng: <NP> such as <NP>; <NP> to; <NP> of; <NP> for; <NP> and; <NP> but; <NP> from; <NP> if; <NP> are <NP>; <NP> is <NP>; <NP> than; <NP> because.
Nếu chuỗi từ loại (POS) của cụm từ thuộc một trong các trường hợp sau đây: "NP TO WDT NP"; "NP TO VB NP"; "NP RB IN NP"; "NP JJ IN NP", "NP IN WDT NP"; "NP VBZ IN NP"; "NP VBZ RB NP".
Nếu chuỗi từ loại của cụm từ bắt đầu bằng: "NP WDT"; "NP WP".
Nếu kết quả chunking của cụm từ là một trong các trường hợp sau: "NP VP NP"; "NP ADVP"; "NP PP (NP PP) *".
Các chuỗi chunking của cụm từ kết thúc như sau: "NP PP"; "NP (of|and|or|in) NP".

III. KẾT QUẢ THỬ NGHIỆM

Sau khi thực hiện phương pháp rút trích tự động các cụm từ hữu ích, chúng tôi đã rút ra tổng cộng được 9536 cụm từ hữu ích từ tập ngữ liệu bài báo khoa học có sẵn. Để đánh giá kết quả thử nghiệm của phương pháp rút trích tự động các cụm từ hữu ích trong bài báo khoa học, chúng tôi tiến hành lựa chọn ngẫu nhiên 10 bài báo khoa học và tiến hành rút trích các cụm từ hữu ích. Kết quả chúng tôi rút trích ra được 498 cụm từ hữu ích và sau đó chúng tôi tiến hành tính độ chính xác và độ phủ trên kết quả của mười bài báo được chọn. Nếu các cụm từ được xuất hiện trong các từ điển chúng tôi thu thập từ các nguồn [4][5] thì cụm từ đó được xem là cụm từ hữu ích, các cụm từ còn lại sẽ được hai nhà nghiên cứu có nhiều kinh nghiệm viết các bài báo khoa học quốc tế bằng tiếng Anh đánh giá độc lập xem cụm từ có phải là cụm từ hữu ích hay không. Số lượng cụm từ được đánh giá đúng là tổng số cụm từ xuất hiện trong tài liệu tham khảo [4], [5] và cụm từ được đánh giá là đúng bởi cả hai nhà nghiên cứu - ở đây chúng tôi lấy phần giao chung các cụm từ mà hai nhà nghiên cứu đều gán nhãn giá trị đúng giống nhau thì được xem là cụm từ hữu ích. Bên cạnh đó, chúng tôi kiểm tra loại kết quả loại bỏ của phương pháp thống kê và phương pháp ràng buộc cú pháp để xác định các cụm từ đúng nhưng bị loại bỏ. Dựa vào đó, chúng tôi đánh giá phương pháp dựa trên độ đo của 498 cụm từ được rút trích tự động trên mười bài báo khoa học bằng tiếng Anh được rút ngẫu nhiên bằng độ chính xác và độ phủ. Công thức tính độ chính xác và độ phủ của phương pháp rút trích tự động các cụm từ hữu được tính toán như sau:

Độ chính xác = Số cụm từ rút trích chính xác / Tổng số cụm từ được rút trích

Độ phủ = Số cụm từ rút trích chính xác / Tổng số cụm từ chính xác

Sau khi tổng hợp kết quả đánh giá của nhà nghiên cứu, chúng tôi đạt kết quả độ chính xác trung bình gần 66% (330/498). Thống kê kết quả trên của phương pháp thống kê và luật ràng buộc thì có 72 cụm từ bị loại bỏ bước thống kê do không thỏa một trong hai điều kiện bất đẳng thức và 19 cụm từ hữu ích bị loại bỏ bởi các luật ràng buộc. Do đó, độ phủ của phương pháp là 78,4% (330/(330+72+19)) và độ đo F1 là 71,66%. Từ đó cho thấy phương pháp cho kết quả khá tốt và có thể áp dụng rút trích tự động cụm từ hữu ích cho các ngôn ngữ khác như các bài báo khoa học bằng tiếng Việt.

Một vài kết quả thí nghiệm của phương pháp rút trích cụm từ hữu ích tự động được chúng tôi đã liệt kê ở bảng 4 và bảng 3 là kết quả thống kê chi tiết của tập ngữ liệu và số cụm hữu ích mà chúng tôi thực nghiệm được trên nguồn ngữ liệu hiện tại của chúng tôi. Ngoài một vài cụm xuất hiện trong tập từ điển như là *"The amount of <NP>"* hay *"To the best of our knowledge <NP>"* còn có các cụm như *"In this paper, we consider <NP>"* hoặc *"We conduct experiments on <NP>"* cũng được rút trích. Điều này giúp cho thấy rằng phương pháp có thể rút trích được các cụm từ mà tập từ điển không có, từ đó có thể làm đa dạng các cụm từ, các khuôn mẫu (template) cho người dùng tham khảo và sử dụng trong các bài báo khoa học bằng tiếng Anh.

Bảng 3. Thống kê kết quả của phương pháp rút trích tự động các cụm từ hữu ích

Số bài báo	Số câu	Số cụm rút trích tự động
1565	218847	9536

Sau khi quan sát các cụm từ bị loại bỏ ở phương pháp luật ràng buộc ngữ nghĩa thì chúng tôi nhận thấy rằng nhiều cụm từ hữu ích như là "*We observe that <CL>*", "*This indicates that <CL>*" hay "*Table 1 summarizes <NP>*" không được rút trích từ phương pháp thống kê là do hàm tính toán Rscore(E) < Lscore(EX). Do cụm từ EX có tần số xuất hiện và hàm phân bố xác suất Hr gần bằng với cụm từ hữu ích E. Vì thế đối với các dạng trường hợp này, chúng tôi sẽ tìm hiểu và tính toán lại độ đo trên phương pháp thống kê để giải quyết các trường hợp trên. Bên cạnh đó, chúng tôi cũng nhận thấy rằng nhiều động từ được rút trích chưa có giới từ chính xác trong câu cũng như là cụm danh từ. Hiện tại phương pháp trên chỉ xử lý trên bề mặt của câu là thay thế các cụm danh từ bằng nhãn <NP> và chưa có phân tích sâu vào câu trúc ngữ

[2] Trong đó: NP là cụm danh từ khác với nhãn <NP>, VP là cụm động từ, PP là cụm giới từ, ADVP là cụm các trạng từ.
Các ký tự nhãn từ loại (POS) tham khảo ở đây: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

nghĩa của câu. Thêm vào đó, chúng ta sẽ phân tích cụm danh từ thành các tính từ, trạng từ và danh từ và thay thế các tính từ trong cụm danh từ thành nhãn <JJ> và dựa vào danh từ để xác định xem cụm danh từ có được thay thế thành nhãn <NP>. Vì thế, chúng tôi nghĩ sử dụng phương pháp bổ sung như phân tích cấu trúc phụ thuộc dựa trên ngữ nghĩa của câu, xử lý cụm danh từ sẽ đưa ra độ chính xác và tin cậy hơn cho các cụm từ mà hệ thống rút trích.

Bảng 4. Kết quả rút trích thành công từ tập ngữ liệu của phương pháp rút trích tự động

The amount of <NP>
In this paper, we consider <NP>
We evaluate our approach on <NP>
<NP> is to convert NP to <NP>
<NP> aims to maximize <NP>
<NP> is learned from <NP>
<NP> learn <NP> using <NP>
<NP> can be formulated as <NP>
In our experiments, we consider <NP>
We conduct experiments on <NP>
Figure <digital> demonstrates that <CL>
It is not surprising that <CL>
To the best of our knowledge <NP>
<NP> provides a list of <NP>
It can be seen that <CL>

Xây dựng dữ liệu

Association for Computational Linguistics (ACL) là một tổ chức khoa học quốc tế và là nơi nghiên cứu hàng đầu về các vấn đề liên quan Ngôn ngữ học tính toán hay Xử lý Ngôn ngữ Tự nhiên. Do đó, các bài báo tại các hội nghị thuộc tổ chức ACL như là hội nghị ACL, CoNLL, EACL, NAACL, EMNLP từ năm 2014 đến năm 2016 được sử dụng để làm tập ngữ liệu sẽ đảm bảo được chất lượng và độ tin cậy cho người dùng khi tham khảo câu ví dụ hay các cụm từ được rút trích tự động. Tuy nhiên, các bài báo được định dạng theo một bài báo khoa học với nhiều thông tin không cần thiết như là công thức, bảng số liệu, chú thích và tài liệu tham khảo. Vì thế, chúng tôi phải loại bỏ các thông tin bằng thủ công sau đó sẽ sử dụng thư viện PDFxStream[3] để đọc nội dung của các bài báo khoa học. Tiếp theo, chúng tôi sẽ tổ chức ngữ liệu thành định dạng .xml[4] như Hình 5 để dễ dàng lưu trữ và quản lý. Để đảm bảo cho tập ngữ liệu được chính xác, chúng tôi phải kiểm tra thủ công từng bài báo để tránh trường hợp sai sót trong quá trình xử lý ngữ liệu.

```
<paper>
<title>paper's title</title>
<body>
<sec>
        <title sec>Section</title sec>
        <p>
                <sub sec>
                <title subsec>Sub-section</title subsec>
                <p>
                        <s>sentence 1</s>
                        <s>sentence 2</s>
                </p>
                </sub sec>
        </p>
</sec>
</body>
</paper>
```

Hình 5. Định dạng .xml của một bài báo khoa học sau khi đã xử lý

[3] https://www.snowtide.com/

[4] Số lượng bài báo khoa học có các đề mục con như 2.1.1, 2.1.2 cũng như là 2.1.1.1, … chiếm số lượng không đáng kể trên tổng số các bài báo. Vì thế chúng tôi quyết định không biểu diễn cho các trường hợp đó mà sẽ gộp chung với đề mục lớn hơn.

IV. KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày kết quả thử nghiệm của phương pháp rút trích cụm từ một cách tự động dựa trên tập ngữ liệu có sẵn. Phương pháp rút trích dựa trên các phương pháp thống kê và luật cú pháp để rút trích tự động ra các cụm từ, mẫu câu (template) để làm đa dạng và phong phú cụm từ cho người sử dụng. Kết quả rút trích tự động các cụm từ của chúng tôi đạt độ đo F1 là 71,66%. Bên cạnh đó, chúng tôi cũng đã xây dựng một ứng dụng hỗ trợ người dùng tra cứu cụm từ được sử dụng trong các bài báo khoa học bằng tiếng Anh trên nền tảng web. Ứng dụng có hai chức năng là tìm kiếm các câu ví dụ dựa trên truy vấn và tra cứu cụm từ được sử dụng trong các bài báo khoa học. Ứng dụng hy vọng sẽ giúp người dùng chưa có kinh nghiệm viết bài báo khoa học bằng tiếng Anh cải thiện được nội dung và chất lượng của bài báo khoa học quốc tế.

Trong tương lai, chúng tôi sẽ nghiên cứu công thức tính độ đo của cụm từ trên phương pháp thống kê để giải quyết các trường hợp của cụm từ hữu ích khác và tiến hành phân tích các câu để lựa chọn ra các cụm ứng viên dựa trên việc phân tích ngữ nghĩa của câu. Sau đó, chúng tôi sẽ cài đặt thử nghiệm phương pháp rút trích tự động các cụm từ hữu ích dành cho các bài báo khoa học tiếng Việt. Trong phần phát triển tiếp theo của ứng dụng, chúng tôi sẽ mở rộng tập ngữ liệu của ứng dụng về các lĩnh vực chuyên môn khác nhau để đa dạng hóa người dùng, đồng thời áp dụng các phương pháp gợi ý tự động dựa trên đầu vào là các truy vấn của người dùng.

LỜI CẢM ƠN

Chúng tôi cảm ơn các thành viên trong nhóm nghiên cứu Xử lý Ngôn ngữ Tự nhiên – Phòng Thí nghiệm Truyền Tthông Đa phương tiện (NLP-MMLab) thuộc trường Đại học Công nghệ thông tin đã góp ý và xây dựng tập ngữ liệu cho nghiên cứu này. Chúng tôi cũng cảm ơn phòng thí nghiệm MMLab đã hỗ trợ các cơ sở thí nghiệm để chúng tôi hoàn thành nghiên cứu.

TÀI LIỆU THAM KHẢO

[1] Ludwig s.r.l.s, Via Fiume 6,90133 Palermo, Italy. Find your sentence. Website "https://ludwig.guru/".

[2] Yu-Chih Sun. "Learner Perceptions of a Concordance Tool for Academic Writing". Computer Assisted Language Learning. Vol.20, No. 4, October 2007, pp. 323 – 343.

[3] Mei-Hua Chen, Shih-Ting Huang, Hung-Ting Hsieh, Ting-Hui Kao, Jason S. Chang. FLOW: A First-Language-Oriented Writing Assistant System. ACL 2012, pages 157–162.

[4] John Morley. "Academic Phrasebank". The University of Manchester, 2014.

[5] "English for Writing Research Papers Useful Phrases". Springer. Website: "http://www.springer.com/cda/content/document/cda_downloaddocument/Free%2BDownload%2B-%2BUseful%2BPhrases.pdf%3FSGWID%3D0-0-45-1543172-p177775190+&cd=1&hl=en&ct=clnk&gl=vn".

[6] Shunsuke Kozawa, Yuta Sakai, Kenji Sugiki, and Shigeki Matsubara. "Automatic Collection of Useful Phrases for English Academic Writing". Innovations in Intell. Machines- 2, SCI 376, pp. 45–59.

[7] Yuanchao Liu, Xin Wang, Ming Liu, Xiaolong Wang. Write-righter: "An Academic Writing Assistant System". Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16).

[8] Sayako Maswana, Toshiyuki Kanamaru, Akira Tajino. "Analyzing the Journal Corpus Data on English Expressions Across Disciplines". The Journal of ASIA TEFL, vol.10, No. 4, pp. 71-96, Winter 2013.

[9] ZHAO Lili. "The Application of Corpus in English Writing and Its Influences". Studies in Sociology of Science, Vol. 6, No. 6, 2015, pp. 78-82.

[10] Kato, Y., Egawa, S., Matsubara, S., Inagaki. "English sentence retrieval system based on dependency structure and its evaluation". In Proceedings of 3rd International Conference on Information Digital Management, pp. 279–285.

THE PHRASE SEARCHING APPLICATION FOR ENGLISH SCIENTIFIC PAPERS

Thin Van Dang, Kiet Nguyen Van, Ngan Nguyen Luu Thuy

ABSTRACT: *English is an international language used to present scientific research at conferences around the world. However, using English in a proficient way is difficult for non-native speakers including Vietnamese and other countries' people. To write a scientific paper in English, non-native novice writers need to spend time to learn unknown phrases or look up in dictionaries. In this paper, we present a method for extracting useful expressions automatically from available scientific papers that were written by native speakers, and using the extracted expressions in an application for phrase-searching. Our application is expected to help the user accurately express the content of the scientific English papers.*

Keyword: *writing assistance, academic writing, phrase search, example search.*