

Tìm kiếm Thông tin Tiếng Việt theo Khái niệm

Đỗ Thị Thanh Tuyền

Trường Đại học Công nghệ Thông tin, ĐHQG-HCM, TP. HCM, Việt Nam
tuyendtt@uit.edu.vn

Tóm tắt. Bài báo này trước tiên giới thiệu chung về mô hình tìm kiếm thông tin tiếng Việt theo khái niệm trong các văn bản tiếng Việt, kể đến trình bày kết quả nghiên cứu của một phần nội dung trong nghiên cứu trên. Phần nội dung này tập trung vào việc xác định các đối tượng hoặc các khái niệm xuất hiện trong văn bản và câu truy vấn, làm cơ sở cho việc tìm kiếm kết quả phù hợp với mong muốn của người sử dụng. Để xác định các đối tượng hoặc các khái niệm này, cần giải quyết một số vấn đề xác định ranh giới từ tổ, từ đồng nghĩa và từ đồng âm. Trong đó, vấn đề từ đồng nghĩa và từ đồng âm được tập trung nghiên cứu bằng cách xây dựng một phương pháp dựa trên khái niệm *semantic memory* và *Head-driven Phrase Structure Grammar (HPSG)*.

Từ khóa. Phân tích văn bản tiếng Việt, tìm kiếm thông tin theo khái niệm, từ đồng nghĩa, từ đồng âm.

1 Giới thiệu

Tìm kiếm thông tin tiếng Việt theo khái niệm (Concept-based Information Retrieval for Vietnamese) là tìm kiếm thông tin dựa trên sự so sánh các đối tượng hoặc các khái niệm được nói đến trong văn bản và truy vấn tiếng Việt. Một hệ thống tìm kiếm thông tin tiếng Việt theo khái niệm cần phân tích văn bản và câu truy vấn thành các khái niệm thay vì phân tích thành các tiếng¹ dựa vào khoảng trắng. Việc tìm kiếm sẽ được thực hiện dựa trên sự so khớp các khái niệm có trong câu truy vấn và văn bản thay vì so khớp các tiếng của chúng.

Khác với phương pháp tìm kiếm này, tìm kiếm thông tin theo mô hình Extended Boolean², được áp dụng rộng rãi cho các tài liệu viết theo ngôn ngữ thuộc ngữ hệ Ấn-Âu vốn là ngôn ngữ biến hình, dựa trên việc so sánh các term³. Khi áp dụng phương pháp tìm kiếm thông tin này cho các văn bản tiếng Việt sẽ gặp khó khăn trong việc tìm những văn bản chứa nội dung mong đợi. Nguyên nhân là mỗi tiếng trong tiếng Việt được xử lý tương tự như một từ trong ngôn ngữ thuộc ngữ hệ Ấn-Âu; trong khi theo Cao Xuân Hạo [1], tiếng Việt mang tính phân tích cao nên đa phần phải dùng nhiều tiếng để định danh một đối tượng hoặc một khái niệm. Như vậy trong tiếng Việt, ngữ đoạn có nhiều tiếng giống nhau có thể hoàn toàn khác nội dung. Ví dụ có

¹ Theo Cao Xuân Hạo [1], trong tiếng Việt mỗi “tiếng” là một từ và không có từ ghép

² Là sự kết hợp của hai mô hình luận lý và không gian véc-tơ được trình bày chi tiết trong [2]

³ Là dạng tương đương với nguyên mẫu của một từ sử dụng để lập chỉ mục, tìm kiếm theo [2]

hai ngữ đoạn: 1) máy tính khoa học (scientific calculator), và 2) khoa học máy tính (computer science). Hai ngữ đoạn này đều chứa các tiếng: “khoa”, “học”, “máy” và “tính” nhưng nội dung của chúng hoàn toàn khác nhau. Khi dùng Google để tìm kiếm tài liệu tiếng Việt với ngữ đoạn 1, kết quả trả về đại đa số có nội dung là ngữ đoạn 2.

Như vậy, để việc tìm kiếm thông tin văn bản tiếng Việt trở nên chính xác hơn cần phải tiến hành việc tìm kiếm dựa trên sự so khớp các khái niệm hoặc các đối tượng được mô tả trong văn bản thay cho sự so khớp các tiếng có trong văn bản như phương pháp tìm kiếm theo từ khóa đã được xây dựng cho ngôn ngữ Âu châu. Việc xác định các đối tượng hoặc khái niệm được đề cập trong một văn bản tiếng Việt được thực hiện theo trình tự: đầu tiên là xác định ranh giới các từ tố⁴ (hay còn gọi là ngữ đoạn), kế đến phải xác định các từ đồng âm và các từ đồng nghĩa để đảm bảo đối tượng hay khái niệm được xác định không phụ thuộc vào biểu hiện của nó bằng ngôn ngữ, sau cùng là dùng một cấu trúc biểu diễn các khái niệm và các đối tượng này để phục vụ cho việc lập chỉ mục và tìm kiếm.

Theo sự phân tích ở trên, nội dung chính của việc nghiên cứu đề xuất mô hình tìm kiếm thông tin tiếng Việt theo khái niệm gồm có: 1) nghiên cứu phương pháp xác định ranh giới các từ tố; 2) nghiên cứu giải quyết các vấn đề từ đồng âm, từ đồng nghĩa; 3) xây dựng cấu trúc để biểu diễn các khái niệm và các đối tượng trong một văn bản tiếng Việt; và cuối cùng, 4) nghiên cứu phương pháp tìm kiếm dựa trên cấu trúc đã xây dựng.

Trong phạm vi bài báo này, phương pháp xác định ranh giới các từ tố tạm thời sử dụng phương pháp tách từ dựa trên từ điển có quá trình tiền xử lý tên riêng. Nội dung được tập trung giới thiệu là mô hình tìm kiếm thông tin tiếng Việt theo khái niệm và phương pháp xác định đối tượng hoặc khái niệm không bị chi phối bởi từ đồng âm và từ đồng nghĩa dựa trên ý tưởng về semantic memory⁵ và head-driven phrase structure grammar⁶. Đồng thời kết quả thử nghiệm độ chính xác tìm kiếm theo các khái niệm hoặc đối tượng đã được xác định cũng được trình bày.

2 Những nghiên cứu liên quan

Ở phạm vi ngoài nước, các phương pháp giải quyết vấn đề tìm thông tin theo ngữ nghĩa, vốn là vấn đề bao quát của vấn đề tìm kiếm thông tin theo khái niệm, đã được đề xuất trong các công trình nghiên cứu của một số tác giả như Thomas C. Rindfleisch [5], Julio Gonzalo [6], Atanas Kiryakov [7], Fausto Giunchiglia [8], Miriam Fernández Sánchez [9], Stein L. Tomassen [10], Ofer Egozi [11] và Julian Szymanski [12]. Các công trình nghiên cứu này giải quyết vấn đề theo hai hướng chính. Hai hướng này là query enrichment (còn gọi là query expansion) như trong [9], [10] và [12] và semantic annotation như trong [5], [6], [8] và [11]. Các hướng này có đặc điểm sau:

⁴ Theo Cao Xuân Hạo [1], một nhóm các tiếng chỉ một khái niệm được gọi là từ tố hay tổ hợp

⁵ Khái niệm ngữ nghĩa trí nhớ được trình bày trong [3]

⁶ Khái niệm ngữ pháp cấu trúc hướng tâm được trình bày trong [4]

2.1 Query enrichment

Các giải pháp theo hướng này tập trung phân tích câu truy vấn thành các từ khóa, sau đó sản sinh tập từ khóa mới như trong [10]. Tập từ khóa mới này gồm các từ khóa đã phân tích được và các từ khóa đồng nghĩa với chúng để hình thành các câu truy vấn mở rộng. Quá trình phân tích từ khóa và sản sinh tập từ khóa có thể dùng từ điển đồng nghĩa hoặc dùng ontology thuộc miền tri thức mà hệ thống tìm kiếm sẽ được áp dụng. Việc mở rộng câu truy vấn như trong [9] và [12] có thể dùng cả từ bao hàm hoặc từ bộ phận của từ cần mở rộng. Các câu truy vấn mở rộng và câu truy vấn gốc sau đó được dùng để tìm kiếm trong tập tài liệu theo mô hình tìm kiếm Extended Boolean.

Theo hướng nghiên cứu này, Julian Szymanski [12] sử dụng khái niệm semantic memory, trong đó quan niệm các khái niệm là biểu diễn của trí nhớ về sự vật, hiện tượng trong thế giới thực và từ ngữ chỉ là những nhãn của các khái niệm này. Những khái niệm này được biểu diễn bằng các bộ ba *object – relation – feature*. Khi tìm kiếm thông tin, các đặc điểm trong truy vấn sẽ được rút trích. Sau đó, dựa vào các đặc điểm này sẽ xác định sự vật chứa các đặc điểm đó và tìm kiếm các tài liệu chứa các sự vật đã được xác định. Công trình này của Julian Szymanski giải quyết được vấn đề định danh⁷ trong ngôn ngữ. Cụ thể là trường hợp một khái niệm được biểu diễn bằng tên riêng, nhưng cũng được biểu diễn bằng một ngữ đoạn mang tính phân tích⁸. Ví dụ: “laptop” với “personal computer for mobile use” trong tiếng Anh và “tivi” với “máy truyền hình” trong tiếng Việt.

2.2 Semantic annotation

Các giải pháp theo hướng semantic annotation tập trung vào việc phân tích tài liệu và câu truy vấn để xác định các ngữ đoạn có ý nghĩa trong tài liệu. Việc phân tích được thực hiện qua ba bước chính:

- (1) Phân tích ngữ pháp các câu trong tài liệu để xác định các ngữ đoạn.
- (2) Xác định các ngữ đoạn có cấu trúc phù hợp với yêu cầu tìm kiếm của miền tri thức mà hệ thống cần phục vụ.
- (3) Dùng ontology thuộc miền tri thức mà hệ thống sẽ phục vụ để sinh các ngữ đoạn có nội dung tương đương.

Việc tìm kiếm tài liệu được thực hiện theo mô hình Extended Boolean nhưng đối tượng được dùng để so sánh không phải là từ khóa mà là các ngữ đoạn đã được xác định trong quá trình chú giải ngữ nghĩa.

Trong hướng nghiên cứu này, chú giải ngữ nghĩa có thể là kết quả phân lớp tài liệu dựa vào tập các lớp ngữ nghĩa đã được xác định trước như trong [6] và [11]. Các lớp ngữ nghĩa này có thể là các bài viết trên một nguồn đáng tin cậy nào đó như wikipedia⁹ theo [11]. Các lớp ngữ nghĩa này sẽ là các chiều trong không gian ngữ nghĩa và mỗi tài liệu sẽ được biểu diễn bằng một vector ngữ nghĩa xác định.

⁷ Vấn đề định danh được Cao Xuân Hạo trình bày trong [1]

⁸ Tính phân tích trong tiếng Việt được Cao Xuân Hạo khẳng định trong [1]

⁹ <http://www.wikipedia.org>

Ở phạm vi trong nước, hiện chưa có kết quả công bố chính thức về một mô hình tìm kiếm thông tin theo khái niệm cho tiếng Việt. Tuy nhiên, có nhiều công trình nghiên cứu liên quan đến xử lý tiếng Việt như phân tích từ loại trong câu tiếng Việt trong [13] theo phương pháp Maximum Entropy dùng trong bộ phân tích từ loại Stanford-Tagger trong [14]. Vấn đề xử lý ngữ nghĩa tiếng Việt áp dụng cho phạm vi ứng dụng cụ thể có các công bố về xử lý ngữ nghĩa câu hỏi tiếng Việt cho hệ thống trả lời câu hỏi (QA) tiếng Việt trong [15], hệ thống tìm kiếm trong thư viện điện tử trong [16]. Các kết quả nghiên cứu này cung cấp nhiều thông tin có giá trị cho việc phân tích và xác định ngữ nghĩa văn bản tiếng Việt.

3 Tìm kiếm thông tin tiếng Việt theo khái niệm

Như đã đề cập đến trong phần giới thiệu, vấn đề tìm kiếm thông tin tiếng Việt theo khái niệm cần giải quyết bốn nội dung chính: 1) nghiên cứu phương pháp xác định ranh giới các từ tố; 2) nghiên cứu giải quyết các vấn đề từ đồng âm, từ đồng nghĩa; 3) xây dựng cấu trúc để biểu diễn các khái niệm và các đối tượng trong một văn bản tiếng Việt; và cuối cùng, 4) nghiên cứu phương pháp tìm kiếm dựa trên cấu trúc đã xây dựng. Trong bài báo này, nội dung tập trung trình bày mô hình tìm kiếm thông tin tiếng Việt theo khái niệm và cách giải quyết các vấn đề từ đồng âm và từ đồng nghĩa trong tiếng Việt. Các nội dung còn lại sẽ được trình bày trong các nghiên cứu tiếp theo.

3.1 Mô hình tìm kiếm thông tin tiếng Việt theo khái niệm

Để thực hiện việc tìm kiếm thông tin theo khái niệm, qua nghiên cứu các giải pháp trong các công trình [5-12], mô hình tìm kiếm thông tin tiếng Việt theo khái niệm được xác định như Hình 1. Mô hình này chứa các thành phần chính như sau:

Thành phần phân tích từ tố

Thành phần này có nhiệm vụ phân tích tài liệu văn bản hoặc truy vấn thành các từ tố theo các quy tắc ngữ pháp tiếng Việt. Các loại từ tố được tập trung xử lý là các tổ hợp danh từ và các tổ hợp vị từ¹⁰.

Thành phần xác định khái niệm

Thành phần này dựa vào các từ điển tương ứng với các lĩnh vực thông tin mà hệ thống tìm kiếm phục vụ để xác định các khái niệm từ các từ tố được đề cập đến trong tài liệu văn bản hoặc truy vấn. Các khái niệm ở đây bao gồm các khái niệm về đối tượng, tính chất và hành vi tương ứng với các danh ngữ và vị ngữ. Độ chính xác của kết quả xác định khái niệm phụ thuộc vào cấu trúc biểu diễn của một mục và tổ chức của từ điển. Một mục của từ điển có thể là một từ hay một ngữ đoạn.

Thành phần xác định quan hệ

¹⁰ Theo Cao Xuân Hạo [1], vị từ trong tiếng Việt bao gồm động từ và “tính từ”.

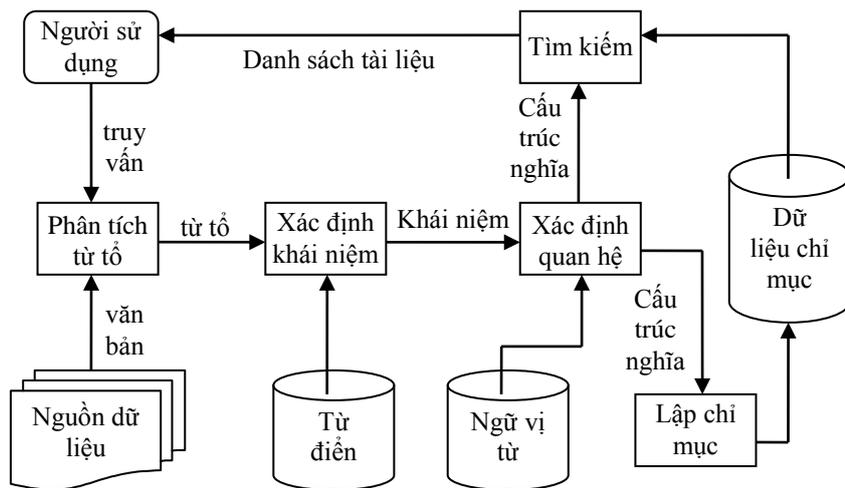
Dựa trên các khái niệm được đề cập đến cùng với biểu diễn của nó là các ngữ đoạn, xác định quan hệ giữa các khái niệm theo ngữ nghĩa của các vị ngữ tiếng Việt. Ngữ nghĩa của các vị ngữ tiếng Việt được xây dựng sẵn và lưu trữ trong một tập dữ liệu ngữ vị từ.

Thành phần lập chỉ mục

Lập chỉ mục các cấu trúc nghĩa tương tự như chỉ mục trong mô hình tìm kiếm thông tin Extended Boolean, trong đó đối tượng lập chỉ mục là từng cấu trúc nghĩa thay vì là term.

Thành phần tìm kiếm

Phương pháp tìm kiếm sẽ được thực hiện theo mô hình Extended Boolean, trong đó mỗi cấu trúc nghĩa được xem như một chiều trong không gian vector. Việc so khớp các cấu trúc nghĩa được thực hiện bằng những phép toán được định nghĩa trên cấu trúc nghĩa. Nội dung về việc so khớp các cấu trúc nghĩa sẽ được trình bày trong các nghiên cứu sau.



Hình 1. Mô hình tìm kiếm thông tin tiếng Việt theo khái niệm

3.2 Xác định các khái niệm trong văn bản tiếng Việt

Cấu trúc mô tả khái niệm

Trong tiếng Việt, có nhiều từ ngữ cùng là biểu diễn về ngôn ngữ của một khái niệm. Hiện tượng này xuất hiện do các địa phương khác nhau trong nước có sử dụng phương ngữ khác nhau cũng như do các từ ngữ gốc Hán, vốn chiếm đến hơn 70% khối lượng từ ngữ trong tiếng Việt theo [1], được sử dụng chung với các từ ngữ thuần Việt. Hiện tượng này khác với từ ngữ đồng nghĩa ở chỗ các từ ngữ đồng nghĩa là biểu diễn về ngôn ngữ của các khái niệm tương đương nhau trong một ngữ cảnh xác định.

Ví dụ “đen” và “mực” là hai từ đồng nghĩa khi mô tả đặc điểm màu sắc của loài chó, theo đó, “chó đen” đồng nghĩa với “chó mực”. Ở đây “chó mực” không phải là một giống chó như “cà chua” là một giống cà. Mặc dù “đen” và “mực” có nghĩa tương đương nhau, nhưng không thể dùng “màu đen” như “màu mực” vì “mực” có nghĩa là một vật chất có màu đen được dùng để ghi lại các ký hiệu, và khi dùng tổ hợp “chó mực”, màu đen được hiểu từ nghĩa của “mực”. Còn đối với trường hợp “heo” và “lợn” là các từ địa phương, chúng có thể thay thế cho nhau ở tất cả các trường hợp như “con heo/lợn”, “thịt heo/lợn”, “nuôi heo/lợn”. Như vậy, trong trường hợp này, “heo” và “lợn” chỉ là hai biểu diễn về ngôn ngữ của một khái niệm.

Bên cạnh vấn đề từ đồng nghĩa, khi xác định khái niệm cần phải chú ý đến vấn đề từ đồng âm. Hai từ đồng âm là hai từ có biểu hiện về ngôn ngữ giống nhau nhưng là biểu diễn về ngôn ngữ của những khái niệm khác nhau. Trong quá trình nghiên cứu, có hai mức độ khác nhau được chú ý là khác nhau hoàn toàn và khác nhau theo ngữ cảnh. Ở mức độ khác nhau hoàn toàn, hai khái niệm được biểu diễn bằng hai từ đồng âm hoàn toàn tách rời nhau. Ví dụ từ “mực” chỉ một loại vật chất màu đen dùng để viết và từ “mực” chỉ một loài động vật sống trong môi trường nước mặn. Ở mức độ khác nhau theo ngữ cảnh, hai từ cùng chỉ một khái niệm nhưng khái niệm đó được quan tâm ở những khía cạnh khác nhau. Ví dụ từ “gan” trong một văn bản thuộc lĩnh vực sinh học sẽ quan tâm đến khía cạnh chức năng của nó trong cơ thể động vật; trong khi cũng từ “gan” trong văn bản thuộc lĩnh vực chế biến thực phẩm sẽ quan tâm đến khía cạnh dinh dưỡng của nó. Như vậy, đối với vấn đề đồng âm, ngữ cảnh là một yếu tố gần như duy nhất để xác định khái niệm mà từ đồng âm muốn biểu diễn cũng như khía cạnh được quan tâm của khái niệm mà nó chỉ đến.

Dựa trên các đặc điểm nêu trên, kết hợp với khái niệm semantic memory và head-driven phrase structure grammar, mỗi khái niệm đều được biểu diễn theo cấu trúc C như sau:

$$C = (L, s, p, cat)$$

Trong đó:

- (i) L là tập các nhãn biểu diễn bằng ngôn ngữ;
- (ii) s là một phù hiệu được đặt cho nghĩa hay khái niệm được biểu diễn bằng tập các nhãn L .
- (iii) p là chức năng ngữ pháp của nhãn trong câu. Trong phạm vi nghiên cứu, các chức năng ngữ pháp tập trung xử lý gồm danh ngữ, ngữ vị từ động, ngữ vị từ tĩnh được ký hiệu tương ứng là DN, VNĐ, VNT.
- (iv) cat là phân loại của khái niệm trong hệ thống phân loại dạng cây của các khái niệm trong một lĩnh vực. Yếu tố phân loại này được đưa vào dựa trên ý tưởng của *head-driven phrase structure grammar* nhằm phục vụ cho việc xác định khái niệm cũng như cấu trúc nghĩa ở thành phần xác định quan hệ.

Ví dụ 1.

- Khái niệm “heo” biểu diễn theo cấu trúc trên là
({heo, lợn}, con_heo, động_vật_có_vú_heo, DN)
- khái niệm “mực” biểu diễn theo cấu trúc trên là
({đen, mực}, màu_đen, màu_sắc_đen, VNT);

Việc sử dụng cấu trúc này để xây dựng từ điển phục vụ cho xác định khái niệm sẽ cho kết quả đơn giản và khắc phục được vấn đề đồng nghĩa. Tuy nhiên, việc xây dựng từ điển theo cấu trúc này phải được thực hiện thủ công thông qua việc xét ý nghĩa từng từ tố trong những ngữ cảnh xác định. Tức là, mỗi lĩnh vực chuyên môn cần có từ điển riêng. Mục đích là để xác định trước ngữ cảnh mà các từ tố được dùng để biểu diễn các khái niệm. Việc xác định phạm vi của từng lĩnh vực chuyên môn phụ thuộc vào đối tượng cần phục vụ của việc tìm kiếm, trong đó phạm vi của lĩnh vực chuyên môn càng hẹp, từ điển càng chi tiết thì kết quả càng chính xác nhưng chi phí xây dựng từ điển sẽ càng lớn.

Phương pháp xác định khái niệm

Việc xác định khái niệm được thực hiện dựa trên hai giả thuyết sau:

Giả thuyết 1. Một từ hoặc từ tố có thể được dùng để mô tả nhiều khái niệm nhưng trong một lĩnh vực cụ thể, nó chỉ có thể dùng để mô tả một khái niệm. Để xác định một khái niệm mà từ hoặc từ tố đó mô tả, phải xác định lĩnh vực mà nó đang được dùng.

Giả thuyết 2. Một câu được xem là dùng trong một lĩnh vực cụ thể nếu nó diễn tả một quá trình, một tác động hoặc một sự biến đổi các khái niệm được biểu diễn bằng các thuật ngữ trong lĩnh vực đang xét của khái niệm đó. Nghĩa là câu đó phải chứa các khái niệm được biểu diễn bằng các danh ngữ có quan hệ cú pháp với vị ngữ trong lĩnh vực đó. Nếu một câu được xem là dùng trong một lĩnh vực nào đó, các khái niệm có trong câu đó được xem là thuộc lĩnh vực đó.

Ví dụ 2. Trong chế biến thực phẩm có các từ như “gan” chỉ một loại nguyên liệu, “xào” chỉ một cách chế biến. Xét hai câu sau: 1) “Gan là cơ quan quan trọng nhất trong cơ thể” và 2) “Gan xào rất tốt cho cơ thể”. Có thể thấy trong câu 1, từ “gan” không được quan tâm ở khía cạnh là thực phẩm của nó; nhưng trong câu 2, từ “gan” cho thấy sự quan tâm ở khía cạnh là thực phẩm. Nguyên nhân là từ “gan” trong câu 1 chỉ được nhắc đến mà không nằm trong một quá trình, một tác động hay một biến đổi nào trong chế biến thực phẩm; trong khi từ “gan” trong câu 2 được kết hợp với vị từ “xào” cũng là một khái niệm trong chế biến thực phẩm để tạo thành một từ tố “gan xào” cho thấy một biến đổi trong chế biến thực phẩm. Như vậy, câu 2 được xem là thuộc về lĩnh vực chế biến thực phẩm theo giả thuyết 2. Từ đó, các từ “gan” và từ “xào” cũng thuộc lĩnh vực chế biến thực phẩm theo giả thuyết 1. Như vậy, từ “gan” trong câu 1 có ý nghĩa khác với từ “gan” trong câu 2.

Phương pháp xác định khái niệm, được trình bày trong thuật toán xác định khái niệm, sẽ thực hiện cho từng câu trên mỗi từ điển của từng lĩnh vực chuyên môn, khi một từ điển của một lĩnh vực chuyên môn nào có thể dùng để xác định được các khái niệm từ những từ tố trong câu đó theo hai giả thuyết trên thì xem như các khái niệm đó đã được xác định. Nếu không thể xác định khái niệm với tất cả từ điển thì xem như không xác định được khái niệm, khi đó các khái niệm chỉ chứa nhãn trong cấu trúc C.

Thuật toán xác định khái niệm.

Đầu vào: Tập các từ tố $W = \{w_1, \dots, w_n\}$,
Tập các Từ điển $D = \{D_1, \dots, D_m\}$,

```

Đầu ra:      Tập các khái niệm  $T = \{t_1, \dots, t_n\}$ .

i=1
T=∅
while (T≠∅) or (i<=m)
begin
  T1=∅
  for j=1 to n
    c={({wj}, null, null, null)
    if (t ∈ Di: c.L ∩ t.L ≠ ∅)
      c=t
    endif
    T1=T1 ∪ {c}
  endfor
  if (t ∈ T1: t.p=DN) and (t ∈ T1: t.p=VND or t.p=VNT)
    T = T1
  endif
end
return T

```

3.3 Thử nghiệm

Mô hình được thử nghiệm với thành phần phân tích từ tổ là một công cụ phân đoạn từ dựa trên từ điển. Từ điển là tập các từ tổ được xây dựng thủ công dựa trên 30 văn bản hướng dẫn chế biến thực phẩm theo cấu trúc C đã giới thiệu. Thành phần xác định khái niệm của mô hình được xây dựng theo thuật toán xác định khái niệm. Thành phần xác định quan hệ hiện tại sử dụng kết quả của thành phần xác định khái niệm làm đầu ra của nó. Thành phần lập chỉ mục và thành phần tìm kiếm sử dụng thư viện Lucene¹¹ để thực hiện theo mô hình Extended Boolean, trong đó việc tìm kiếm ưu tiên so khớp giá trị s trong cấu trúc C , trường hợp s không xác định sẽ thực hiện trên giá trị L . Công cụ được xây dựng để thử nghiệm mô hình được gọi tên là tìm kiếm khái niệm.

Để đánh giá kết quả thử nghiệm, một công cụ tìm kiếm dựa trên tiếng trong tiếng Việt (tương ứng với từ khóa trong tiếng Anh) và một công cụ tìm kiếm dựa trên từ tổ tiếng Việt theo mô hình Extended Boolean được xây dựng trên Lucene cũng được thử nghiệm trên cùng một mẫu kiểm thử. Các công cụ tìm kiếm này được gọi tên lần lượt là tìm kiếm từ khóa và tìm kiếm ngữ đoạn. Công cụ tìm kiếm ngữ đoạn tách các từ tổ bằng các sử dụng thành phần xác định từ tổ trong mô hình.

Mẫu kiểm thử là một tập gồm 41 văn bản và 10 câu truy vấn được lựa chọn thủ công. Tập văn bản có nội dung trong các lĩnh vực chế biến món ăn, y tế, ca nhạc và tiếp thị. Một số tài liệu được lựa chọn theo các tiêu chí sau:

¹¹ <http://lucene.apache.org/>

- Chứa các tiếng của các khái niệm trong câu truy vấn nhưng không chứa các khái niệm đó. Ví dụ “món ăn”, “bỏ” và “xào gan” với truy vấn “món ăn bỏ gan”.
- Chứa các từ đồng âm ở mức độ khác nhau theo ngữ cảnh với từ trong truy vấn. Ví dụ “xào gan” với truy vấn “bệnh viêm gan”.
- Chứa từ đồng nghĩa với từ của câu truy vấn nhưng không chứa các tiếng của từ đồng nghĩa đó. Ví dụ “xào gan” với truy vấn “chế biến gan”.

Độ phủ và độ chính xác theo [2] của các công cụ tìm kiếm nêu trên qua thử nghiệm 10 truy vấn với tập 41 văn bản tiếng Việt có kết quả như các bảng 1 và 2.

Bảng 1. Độ phủ của các công cụ tìm kiếm từ khóa, tìm kiếm ngữ đoạn và tìm kiếm khái niệm

Truy vấn	Tìm kiếm từ khóa	Tìm kiếm ngữ đoạn	Tìm kiếm khái niệm
bánh bông lan	100,00%	100,00%	100,00%
gan	100,00%	100,00%	100,00%
chế biến gan	29,63%	25,93%	66,67%
chế biến dứa	25,00%	21,43%	64,29%
bệnh viêm gan	100,00%	100,00%	100,00%
chế biến lòng trắng trứng gà	44,83%	27,59%	68,97%
mua bánh bông lan	100,00%	100,00%	100,00%
chế biến cá bống đục	46,15%	15,38%	65,38%
món ăn bỏ gan	100,00%	85,71%	92,86%
món ăn bỏ làm từ gan	93,10%	79,31%	68,97%
Trung bình	73,87%	65,53%	82,71%

Bảng 2. Độ chính xác của công cụ tìm kiếm từ khóa, tìm kiếm ngữ đoạn và tìm kiếm khái niệm

Truy vấn	Tìm kiếm từ khóa	Tìm kiếm ngữ đoạn	Tìm kiếm khái niệm
bánh bông lan	50,00%	100,00%	100,00%
gan	84,62%	84,62%	84,62%
chế biến gan	50,00%	53,85%	81,82%
chế biến dứa	46,67%	66,67%	90,00%
bệnh viêm gan	15,38%	25,00%	25,00%
chế biến lòng trắng trứng gà	52,00%	80,00%	90,91%
mua bánh bông lan	35,00%	58,33%	58,33%
chế biến cá bống đục	63,16%	66,67%	89,47%
món ăn bỏ gan	43,75%	75,00%	76,47%
món ăn bỏ làm từ gan	72,97%	79,31%	86,96%
Trung bình	51,35%	68,94%	78,36%

Trong Bảng 1, độ phủ của công cụ tìm kiếm ngữ đoạn là thấp nhất vì việc tìm kiếm thực hiện trên một từ tổ biểu diễn một khái niệm thay vì tìm kiếm tất cả từ tổ biểu diễn khái niệm đó, hay đúng hơn là tìm kiếm chính khái niệm đó. Độ phủ của công cụ tìm kiếm từ khóa xếp thứ hai vì tiếng Việt dùng nhiều tiếng để biểu diễn một khái niệm và những khái niệm khác nhau có thể dùng chung một số tiếng như nhau, ví dụ “con gà” và “xe con” dùng chung tiếng “con”; vì vậy số lượng tài liệu thỏa một truy vấn nhiều hơn. Độ phủ của công cụ tìm kiếm khái niệm cao nhất 82,71% vì đã được xử lý hiện tượng đồng âm và đồng nghĩa.

Trong Bảng 2, độ chính xác của công cụ tìm kiếm từ khóa là thấp nhất vì việc tìm kiếm thực hiện bằng cách so khớp các tiếng xuất hiện trong truy vấn với các tiếng trong tài liệu. Độ chính xác của công cụ tìm kiếm ngữ đoạn xếp thứ hai vì các khái niệm đã được xác định bằng các từ tổ, tuy nhiên việc tìm kiếm chỉ thực hiện trên từ tổ mà chưa tìm chính khái niệm mà từ tổ biểu diễn. Độ chính xác của công cụ tìm kiếm khái niệm là cao nhất, với 78,36% vì việc tìm kiếm thực sự dựa trên khái niệm.

4 Kết luận

Bài báo này trình bày các vấn đề trong tìm kiếm thông tin tiếng Việt theo khái niệm và đề xuất một mô hình nhằm giải quyết các vấn đề trên, trong đó tập trung trình bày mô hình và phương pháp giải quyết các hiện tượng từ đồng âm và từ đồng nghĩa xác định được trong quá trình nghiên cứu. Phương pháp giải quyết này dùng cấu trúc khái niệm *C* được xây dựng dựa trên khái niệm semantic memory và head-driven phrase structure grammar nhằm xác định chính xác khái niệm được mô tả bằng từ tổ trong văn bản cũng như trong truy vấn.

Mặc dù công cụ tìm kiếm khái niệm chưa hiện thực được thành phần xác định quan hệ trong mô hình và hiện thực chưa đầy đủ thành phần xác định từ tổ, nhưng với kết quả thử nghiệm với độ phủ 82,71% và độ chính xác 78,36% cho thấy khả năng ứng dụng của mô hình này trong tìm kiếm thông tin tiếng Việt theo khái niệm.

5 Hướng phát triển

Các nghiên cứu sắp tới trước tiên tập trung vào việc xác định từ tổ theo các nguyên tắc xác lập có được từ các nghiên cứu về ngôn ngữ học tiếng Việt. Kế đến vấn đề xác định cấu trúc nghĩa và quan hệ giữa các khái niệm cần được nghiên cứu, từ đó xây dựng các phép toán dựa trên cấu trúc này để phục vụ cho việc lập chỉ mục và tìm kiếm.

Cùng với việc nghiên cứu các phương pháp xử lý ngữ nghĩa, vấn đề tổ chức các từ điển theo lĩnh vực chuyên môn đảm bảo sự đúng đắn về ngữ cảnh và đơn giản trong việc xây dựng rất cần được nghiên cứu. Các từ điển này là nền tảng cho các phương pháp xử lý nêu trên.

Vấn đề tạo một tập dữ liệu kiểm thử có tầm quan trọng không kém. Việc xây dựng một tập dữ liệu kiểm thử chặt chẽ, đủ lớn để đánh giá cũng được thực hiện nhằm đánh giá mô hình một cách toàn diện và chính xác hơn.

Tài liệu tham khảo

1. Cao Xuân Hạo. Tiếng Việt: mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa. NXB Giáo Dục, 2007, Mã số: 7X290t7-DAI.
2. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008, ISBN: 978-0-521-86571-5.
3. Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., and Patterson, K. (2004). The structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, 111(1), 205-235.
4. Pollard, C. and Sag, I.A. Head-Driven Phrase Structure Grammar. University of Chicago Press, 1994.
5. T. Rindflesch and A. Aronson. "Semantic processing in information retrieval". In Seventeenth Annual Symposium on Computer Applications in Medical Care (SCAMC 93), pages 611-615. McGraw-Hill Inc., New York, 1993.
6. Julio Gonzalo, Irina Chugur, Felisa Verdejo, "Sense clusters for information retrieval: evidence from Semcor and the EuroWordNet InterLingual Index", Proceedings of the ACL-2000 workshop on Word senses and multi-linguality, p.10-18, October 07-07, 2000, Hong Kong [doi>10.3115/1117724.1117726].
7. Atanas Kiryakov, Borislav Popov, Ivan Terziev, DimitarManov, and Damyan Ognyanoff. "Semantic annotation, indexing, and retrieval". *J.Web Sem.*, 2(1):49-79, 2004.
8. Fausto Giunchiglia, Uladzimir Kharkevich, Ilya Zaihrayeu. "Concept Search: Semantics Enabled Syntactic Search". In CEUR Workshop Proceedings of SemSearch 2008, Volume 334, pages 109-123, ISSN 1613-0073, 2008.
9. Miriam Fernández Sánchez. Semantically enhanced Information Retrieval: An ontology-based approach. Doctoral dissertation, Universidad Autonoma de Madrid, 2009.
10. Tomassen, S.L. and Strasunskas, D. "Measuring intrinsic quality of semantic search based on feature vectors", *Int. J. Metadata, Semantics and Ontologies*, 2010, Vol. 5, No. 2, pp.120-133.
11. Ofer Egozi, Shaul Markovitch, Evgeniy Gabrilovich. "Concept-Based Information Retrieval Using Explicit Semantic Analysis". *ACM Trans. Inf. Syst.* 29(2): 8, 2011.
12. Julian Szymanski, Wlodzislaw Duch. "Information retrieval with semantic memory model". *Cognitive Systems Research*, 2011, doi:10.1016/j.cogsys.2011.02.002.
13. Le-Hong P., Roussanaly A., Nguyen T. M. H., Rossignol M., "An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts", In Proceedings of TALN 2010.
14. Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network", In Proceedings of HLT-NAACL 2003, pp. 252-259.
15. Tuoi T. Phan, Thanh C. Nguyen and Thuy N. T. Huynh, "Question Semantic Analysis in Vietnamese QA system", *Studies in Computational Intelligence*, Volume 283/2010, 2010, pp. 29-40.
16. Dang Tuan Nguyen, Ha Quy-Tinh Luong, "Document searching system based on natural language query processing for Vietnam Open Courseware library", *Int' J. of Computer Science Issues (IJCSI)*, vol. 6, no. 2, November 2009, pp. 7-13. ISSN (online): 1694-0784, ISSN (print): 1694-0814.