KỸ THUẬT VÀ CÔNG NGHỆ

MỘT HƯỚNG TIẾP CẬN XÂY DỰNG ONTOLOGY TIẾNG VIỆT

Nguyễn Quang Châu[*]
Lê Trọng Ngọc[**], *Tôn Long Phước*[**], *Nguyễn Văn Tân*[**]

TÓM TẮT

Trong bài báo này, chúng tôi xây dựng một Ontology tiếng Việt phục vụ bài toán rút trích cụm danh từ. Ontology bao gồm một thành phần mô tả các tri thức từ vựng tương tự tổ chức danh từ trong Wordnet, một thành phần mô tả các tri thức về thế giới thực. Các tri thức về từ vựng sẽ được ánh xạ đến các khái niệm tương ứng mô tả tri thức về thế giới thực. Ontology về các khái niệm hình thức được xây dựng bằng cách khai thác từ Wikipedia. Các tri thức về từ vựng được rút trích từ các tài liệu và được ánh xạ một cách bán tự động tới các khái niệm tương ứng.

VIETNAMESE ONTOLOGY BUILDING APPROACH

SUMMARY

In the paper, we present an approach for building Vietnamese ontology for noun phrase extraction. The Ontology includes two parts: a) the first is a lexical part, it looks like the noun organization part in the Wordnet; b) the rest is a formal part, it constructs knowledge of the real world. The lexical knowledge is mapping to the concept in the real world. Formal concepts of the formal part in Ontology are built by mined from Wikipedia. Lexical knowledge is extracted from documents and mapped to concepts in the formal part in Ontology.

1. TỔNG QUAN

Cụm từ đặc trưng ngữ nghĩa là các cụm từ mô tả tóm tắt nội dung của tài liệu. Chúng có thể được ứng dụng trong các hệ thống truy hồi thông tin, như mô tả ngữ nghĩa của các tài liệu kết quả của truy vấn, các chỉ mục tìm kiếm, hoặc cho phép xây dựng các phương pháp đo độ tương tự giữa các tài liệu, … [15]. Do đó, việc rút trích chính xác các cụm từ đặc trưng có ý nghĩa rất lớn và là mối quan tâm của các nhà ngôn ngữ học cũng như các nhà khoa học trong lĩnh vực xử lý ngôn ngữ tự nhiên bằng máy tính.

Hướng tiếp cận máy học được sử dụng phổ biến để giải quyết bài toán rút trích cụm danh từ đặc trưng. Phương pháp này không đòi hỏi nhiều công sức xây dựng cơ sở tri thức hay từ điển nhưng lại có độ chính xác cao. Tuy nhiên, một khó khăn lớn của phương pháp này là nó không thể rút trích các cụm từ hợp lý nhưng có tần suất thấp. Trong khi đó, Ontology và cơ sở tri thức đã được sử dụng rộng rãi trong các hệ thống chú thích ngữ nghĩa [4, 9, 13]. Ngữ nghĩa của các khái niệm và thực thể được đề cập đến trong tài liệu có thể được nắm bắt một các chính xác nếu chúng được liên kết với các khái niệm và thực thể trong Ontology và cơ sở tri thức. Do đó, sử dụng Ontology và cơ sở tri thức là một hướng tiếp cận hoàn toàn hợp lý để giải quyết bài toán rút trích cụm từ đặc trưng. Một vấn đề lớn trong hướng tiếp cận sử dụng Ontology và cơ sở tri thức là làm cách nào xác định được khái niệm, thực thể trong Ontology tương ứng với từ, cụm từ được đề cập trong tài liệu, hay còn gọi là chú thích ngữ nghĩa cho các từ, cụm từ đó. Phần lớn các hệ thống chú thích ngữ nghĩa tự động hiện nay đều tập trung giải quyết bài toán chú thích ngữ nghĩa cho các *thực thể có*

[*] TS, Khoa Sau Đại học, Trường Đại học Công nghiệp TPHCM
[**] ThS, Khoa Công nghệ thông tin, Trường Đại học Công nghiệp TPHCM

tên (Named Entity-NE). Việc chú thích ngữ nghĩa cho các khái niệm là tương đối khó khăn.

Trong bài báo này, chúng tôi xây dựng một Ontology Tiếng Việt phục vụ cho việc rút trích cụm danh từ đặc trưng ngữ nghĩa trong văn bản tiếng Việt. Từ những khảo sát về các nghiên cứu liên quan, chúng tôi tiến tới ý tưởng xây dựng bán tự động một Ontology phục vụ bài toán rút trích cụm danh từ đặc trưng. Ontology bao gồm một thành phần mô tả các tri thức từ vựng tương tự như tổ chức danh từ trong Wordnet, một thành phần mô tả các tri thức về các khái niệm, thực thể trong thế giới thực. Ontology chứa các tri thức về thế giới thực được xây dựng bằng tay nhằm đảm bảo độ chính xác. Các tri thức từ vựng (cụm danh từ) sẽ được rút trích bán tự động từ các tài liệu và được ánh xạ đến các khái niệm, thực thể tương ứng.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Các nghiên cứu xây dựng Ontology được khảo sát trong [2]. Các phương pháp có thể chia thành hai hướng chính:

Xây dựng Ontology mới, hướng tiếp cận này chủ yếu sử dụng các phương pháp *gom cụm (clustering)* để xây dựng cây phân cấp ngữ nghĩa. Một độ đo khoảng cách giữa các thuật ngữ phải được định nghĩa làm tiêu chuẩn cho việc gom cụm. Yếu tố quyết định của các phương pháp này là phải chọn một độ đo khoảng cách tốt và một giải thuật gom cụm phù hợp. Khoảng cách ngữ nghĩa giữa các thuật ngữ trong trường hợp này phụ thuộc rất nhiều vào ngữ cảnh của các thuật ngữ đó, vì theo Harris [5], các từ tương tự nhau xuất hiện trong những ngữ cảnh giống nhau. Nghĩa là, nếu các từ càng chia sẻ những thông tin ngữ cảnh giống nhau, thì chúng càng tương tự nhau. Thông tin ngữ cảnh có thể được tính dựa trên các từ xuất hiện xung quanh từ trung tâm. Một trong những nghiên cứu đầu tiên về gom cụm ngữ nghĩa là của Hindle[7], tác giả tính toán sự tương tự ngữ nghĩa giữa các danh từ dựa trên các động từ chúng chia sẻ. Các cặp động từ-chủ từ (verb-subject) và động từ-túc từ (verb-subject) sẽ được tính trọng số đồng xuất hiện trong kho ngữ liệu. Sự tương tự ngữ nghĩa của hai danh từ trên một động từ được tính là giá trị nhỏ nhất trọng số của hai danh từ với động từ đó. Và độ tương tự ngữ nghĩa của hai danh từ là tổng độ tương tự của hai danh từ trên tất cả các động từ chung. Trong [3], tác giả sử dụng các mẫu nhận dạng của Hearst [6] giúp nhận dạng các quan hệ hypernymy (quan hệ cha trong cây phân cấp ngữ nghĩa). Với mỗi danh từ trong tập huấn luyện, một vector chứa tần suất xuất hiện của các danh từ khác ở dạng đồng vị ngữ với chúng được tạo ra. Độ tương tự giữa hai danh từ được tính toán dựa trên công thức cosine giữa hai vector biểu diễn ngữ cảnh của chúng. Để có được các quan hệ hypernymy, tác giả sử dụng các mẫu nhận dạng. Ví dụ : với mẫu câu "B is a (kind of) A", dễ dàng nhận thấy A là hypernym của B. Hoặc với mẫu "X, Y and other Zs", có thể suy ra rằng Z là hypernym của X và Y.

Mở rộng Ontology đã tồn tại, việc mở rộng Ontology được xem như là công việc phân loại các khái niệm mới vào Ontology. Các thông tin của Ontology sẽ được sử dụng như tập huấn luyện để tạo ra bộ phân loại cho các đối tượng chưa biết. Trong [14], tác giả thêm các từ mới vào Wordnet với ý tưởng từ mới sẽ được thêm vào vị trí các từ gần nghĩa nhất với nó tập trung. Nghiên cứu của Witschel [16] sử dụng hệ phân cấp như là một cây quyết định. Khi một khái niệm mới được thêm vào, cây quyết định sẽ được duyệt từ gốc để chọn vị trí thích hợp nhất. Một vấn đề của phương pháp này là, sự trừu tượng cao của các khái niệm ở mức trên dẫn đến việc chọn sai nút trên đường đi. Vấn đề này được tác giả giải quyết bằng cách lan truyền các mô tả ngữ nghĩa của các nút con lên đến nút gốc.

3. XÂY DỰNG ONTOLOGY

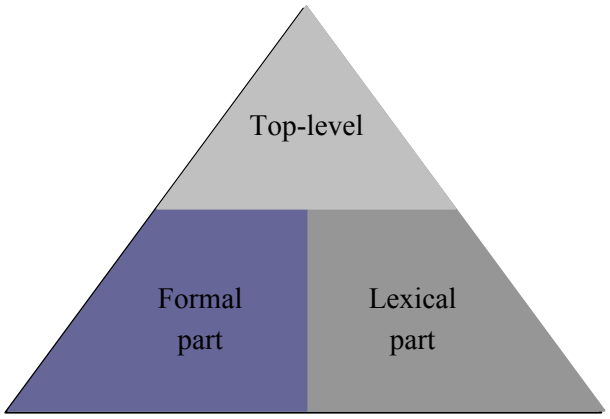
3.1. Phương pháp luận

Như đã đề cập ở trên, Ontology được xây dựng gồm ba phần: một Ontology mức trên mô tả các khái niệm và các quan hệ trừu tượng như

thực thể, đối tượng, tiến trình…; Một Ontology từ vựng tương tự như Wordnet mô tả các từ vựng trong lĩnh vực máy tính; Một Ontology hình thức mô tả các khái niệm tương ứng. Tuy nhiên, việc xây dựng bằng tay các Ontology và ánh xạ các khái niệm của chúng sẽ mất khá nhiều công sức. Do vậy, ý tưởng xây dựng tự động một số bước của Ontology là hoàn toàn tự nhiên. Thông qua khảo sát các nghiên cứu liên quan trong việc xây dựng Ontology, chúng tôi nhận thấy, việc xây dựng tự động hệ phân cấp ngữ nghĩa là hết sức khó khăn. Các nghiên cứu trong hướng tiếp cận này có độ chính xác khá thấp trong khi yêu cầu đặt ra là Ontology phải có độ chính xác cao. Từ những nhận xét trên, hiện tại chúng tôi chọn giải pháp: Ontology mức trên và Ontology hình thức được xây dựng bằng tay. Các từ vựng trong Ontology từ vựng sẽ được xây dựng và ánh xạ một các bán tự động đến các khái niệm tương ứng trong Ontology hình thức.

Ontology được xây dựng phải thoả mãn yêu cầu nắm bắt được ngữ nghĩa của tài liệu. Ngữ nghĩa của tài liệu bao gồm các thực thể được đề cập trong tài liệu và các từ, ngữ mô tả các khái niệm, thuộc tính, quan hệ,… Do đó, Ontology cũng phải mô tả được các thực thể cũng như các khái niệm chung, các thuộc tính, các mối quan hệ,… Sau khi xem xét các Ontology mức trên như OpenCyc, Proton, SUMO,… chúng tôi quyết định xây dựng Ontology dựa trên các khái niệm đã được xây dựng trong SUMO. Ontology của SUMO được xây dựng với mục tiêu cung cấp một nền tảng chung duy nhất, các Ontology mức ứng dụng có thể được kế thừa và xây dựng từ nó, tạo nên sự nhất quán và khả năng tích hợp dễ dàng giữa các Ontology khác nhau. Các khái niệm trong SUMO là khá logic và toàn diện, đáp ứng được yêu cầu đặt ra. SUMO cũng không quá phức tạp và nặng nề, dễ dàng cho việc xây dựng một phiên bản cho tiếng Việt.

Để xây dựng Ontology hình thức trong tiếng Việt, miền tri thức chúng tôi chọn là các khái niệm liên quan đến lĩnh vực máy tính. Như chúng ta đã biết, OpenCyc được xây dựng như một bách khoa toàn thư chứa đựng tri thức của nhân loại. Các khái niệm trong Ontology OpenCyc được xây dựng khá logic và toàn diện. Do đó, ngữ nghĩa của chúng khá phức tạp và nặng nề và việc thao tác đòi hỏi đến những chuyên gia. Tuy nhiên, chỉ xét trong miền tri thức máy tính, các khái niệm được mô tả trong OpenCyc rất đầy đủ và chi tiết, có thể dễ dàng Việt hoá. Do đó, chúng tôi quyết định chọn OpenCyc làm cơ sở để xây dựng Ontology cho tiếng Việt. Cấu trúc Ontology được thể hiện như trong hình 1 sau đây.



Hình 1. *Cấu trúc Ontology*

3.2. Khai thác Wikipedia

Wikipedia[3] là một bách khoa toàn thư trực tuyến với nội dung mở bằng nhiều ngôn ngữ, được viết và xây dựng do nhiều người dùng cùng cộng tác với nhau. Wikipedia và Wiktionary được xem như là một tài nguyên mới về ngữ nghĩa từ vựng do tính năng được cập nhật liên tục nên nó trở thành nguồn tham khảo hữu ích với hàng triệu người. Đặc biệt, tiềm năng của Wikipedia được khai thác gần đây như một cơ sở tri thức ngữ nghĩa từ vựng. Nó được ứng dụng trong các công việc xử lý ngôn ngữ tự nhiên như phân loại văn bản [10], truy hồi thông tin [11], hệ thống hỏi đáp, tính toán quan hệ ngữ nghĩa [9]. Một lý do quan trọng là Wikipedia có phiên bản tiếng Việt, đó là Vi.Wikipedia và Vi.Wiktionary, gọi là

Vi.wiki[1]. Theo thống kê xếp hạng của Zesch [9] thì Vi.Wiktionary đứng thứ 3 trong danh sách xếp hạng 10 ngôn ngữ có số đầu mục cao nhất với 225.000 đầu mục (như trong Bảng 1). Đây thực sự là một kho cơ sở tri thức tiếng Việt rất hữu ích cho cộng đồng nghiên cứu xử lý ngôn ngữ tiếng Việt bằng máy tính (như Bảng 2).

Bảng 1: *các phiên bản Wiktionary (29/1/2008)*

Ngôn ngữ	Xếp hạng	Số đầu mục
French	1	730.193
English	2	682.982
Vietnamese	3	225.380
Turkish	4	185.603
Russian	5	132.386
Ido	6	128.366
Chinese	7	115.318
Greek	8	102.198
Arabic	9	95.020
Polish	10	85.494

Bảng 2: *Số trang thông tin (# danh hiệu), thể loại, và trang đổi hướng của phiên bản Vi.Wikipedia(1/3/2009)*

Ngôn ngữ	# trang	# thể loại	# trang đổi hướng
Tiếng Việt	157.994	322.631	36.301

Với hướng tiếp cận trên, bài báo đi nghiên cứu khai thác Vi.Wiki như một ontology tiếng Việt để phục vụ cho việc rút trích cụm danh từ đặc trưng ngữ nghĩa cho câu tiếng Việt.

Trong Vi.wiki[2], đầu vào cơ bản là các **trang** thông tin. Một trang thông tin có thể là: một bài viết bình thường nói về một khái niệm hay thực thể; một **trang đổi hướng,** là trang dẫn bạn đến trang có tên khác (có thể thông dụng hơn) nói về cùng một đề tài; hay một **trang định hướng**, là một bài viết giải thích về ý nghĩa phổ biến nhất của thuật ngữ, bên dưới liệt kê các liên kết đến các bài viết có tựa đề (tên bài viết) tương tự hoặc có khái niệm tương tự, giúp định hướng cho người đọc đến đúng bài viết mà họ đang tìm.

Mỗi **trang** thông tin được định danh bằng danh hiệu duy nhất, danh hiệu được đặt phù hợp với nội dung mô tả đối tượng được đề cập trong trang này. Trong mỗi trang, ngoài thông tin mô tả về đối tượng, nó còn chứa nhiều **liên kết** đến các trang liên quan khác, các trang liên quan có thể mô tả về đối tượng có quan hệ thành phần, đồng nghĩa, hay phản nghĩa với đối tượng mà trang chứa liên kết đề cập. Hệ thống **trang đổi hướng** có thể được xem như là một từ điển về cụm từ đồng nghĩa, cụm từ biến thể, hay cụm từ viết tắt.

Ngoài ra, Vi.wiki có một hệ thống phân chia **thể loại** các đối tượng, đây là một nguồn thông tin ngữ nghĩa rất hữu ích, nó được dùng để phân loại các chủ đề của các trang thông tin. Hệ thống phân loại của Vi.wiki không chỉ cung cấp hệ thống phân cấp các đối tượng trong thế giới thật, mà còn có thể biểu diễn được các quan hệ giữa các thể loại của các đối tượng như các quan hệ thành phần (thuộc quan hệ isa), và các quan hệ đồng nghĩa (thuộc quan hệ non-isa),... Như vậy, mỗi trang thông tin được liên kết với một hoặc nhiều thể loại, các thể loại này có thể có các tiểu thể loại của chúng với các quan hệ thành phần và quan hệ đồng nghĩa.

Dựa trên sự nghiên cứu về nguồn tài nguyên của Vi.wiki, hướng tiếp cận của bài báo là: 1) rút trích cấu trúc phân cấp của Vi.wiki cùng các quan hệ của chúng (như các quan hệ thành phần, và các quan hệ không thành phần,..); 2) rút trích các danh hiệu của các trang thông tin cùng với các danh hiệu của các trang liên kết với chúng để tạo một Ontology VnO phục vụ bài toán rút trích cụm từ tiếng Việt.

[1] www.vi.wikipedia.org/

[2] Tất cả các dữ liệu sử dụng trong phần này được lấy từ nguồn Vietnamese Wikipedia database dump 1/3/2009.

3.3. Xây dựng Ontology tiếng Việt VnO và từ điển VnDic

Nghiên cứu về nguồn tài nguyên của Vi.wiki, hướng tiếp cận của bài báo bao gồm hai bước sau:

Bước một: Rút trích cây phân cấp của Vi.wiki cùng các quan hệ của chúng như các quan hệ thành phần và các quan hệ không thành phần,… để tạo một Ontology VnO phục vụ bài toán rút trích cụm từ đặc trưng ngữ nghĩa (CTĐT) trong tiếng Việt.

Bài báo sử dụng Java-based Wikipedia Library (JWPL)[16] để rút trích các tài nguyên từ Wikipedia như các trang thông tin, các liên kết, các thể loại và các trang đổi hướng. Kết quả đạt được Ontology VnO có 157.994 khái niệm (danh hiệu) và 322.631 thể loại.

Bước hai: Rút trích các danh hiệu của các trang thông tin cùng với các danh hiệu của các trang đổi hướng với chúng để tạo một từ điển VnDic. Vì mục tiêu là xác định cụm từ đặc trưng nên bài báo xem mỗi trang thông tin trong Wikipedia là một định nghĩa cho đối tượng mà trang mô tả và danh hiệu tương ứng của nó có cụm từ đặc trưng cho đối tượng. Danh hiệu là cụm từ đặc trưng của một đối tượng được định nghĩa trong mỗi trang nếu thỏa mãn một trong các tiêu chí sau:

- Nếu danh hiệu của một trang thông tin là một câu thì trong trường hợp này CTĐT tương ứng sẽ là CTĐT cho câu.

- Nếu danh hiệu là một cụm từ thì CTĐT tương ứng là chính cụm từ đó.

Theo phương pháp như trên, cấu trúc của từ điển VnDic là một tập các đầu mục, mỗi đầu mục bao gồm: CTĐT, danh hiệu, cụm từ đồng nghĩa có được là CTĐT của trang đổi hướng. Mỗi đầu mục trong từ điển được ánh xạ tới thể loại trong VnO. Kết quả đạt được từ điển VnDic có tổng cộng 152.450 đầu mục, mỗi đầu mục có cấu trúc được minh họa như sau:

< CTĐT >< danh hiệu của trang thông tin>< CTĐT của trang đổi hướng>.

Trong trường hợp có nhiều trang mà kết quả của quá trình rút trích các danh hiệu cho cùng một CTĐT thì mỗi đầu mục trong từ điển ViDic có dạng:

< CTĐT >< danh hiệu của trang thông tin 1>< CTĐT của trang đổi hướng 1>,…,< danh hiệu của trang thông tin n>< CTĐT của trang đổi hướng n>.

Trường hợp một CTĐT có nhiều danh hiệu chỉ chiếm tỉ lệ 52 trong tổng 152450 đầu mục trong từ điển VnDic.

Trong phương pháp tiếp cận này, mặc dù bài báo sử dụng các thông tin từ Wikipedia để tạo ra một từ điển VnDic. Tuy nhiên phương pháp này còn có thể áp dụng cho các Ontology hay các cơ sở tri thức khác.

3.4. Huấn luyện Ontology

Các từ vựng trong Ontology từ vựng được rút trích tự động từ tập các tài liệu được chọn lựa trong tạp chí Thế Giới Vi Tính. Tiếp đó, các từ vựng sẽ được ánh xạ một cách bán tự động sang các khái niệm trong Ontology hình thức. Việc ánh xạ các từ vựng có thể được xem như việc chú thích ngữ nghĩa cho các từ vựng đó. Hình 2 mô tả các công đoạn rút trích và ánh xạ các khái niệm, trong đó:

** Nhận diện các khái niệm, thực thể:*

Mục đích của bước này là rút trích các khái niệm cũng như các thực thể được mô tả trong tài liệu. Thông thường, các khái niệm, các thực thể xuất hiện trong tài liệu dưới dạng danh từ và cụm danh từ. Do vậy, nghiên cứu của chúng tôi sử dụng cấu trúc cụm danh từ được trình bày trong [11]. Theo đó, cụm danh từ được cấu tạo bởi phần phụ trước, phần trung tâm và phần phụ sau, trong đó:

Phần phụ trước và phần phụ sau thể hiện thuộc tính, tính chất của đối tượng.

Phần trung tâm thể hiện đối tượng được đề cập đến, gồm hai bộ phận T_1 và T_2. T_1 mô tả đơn vị tính toán, chủng loại khái quát. Các từ loại có thể thực hiện chức năng T_1 là: danh từ chỉ đơn vị (Nu), danh từ chỉ số lượng (Nn), danh từ tổng thể (Ng), danh từ loại thể (Nt), danh từ đơn thể (Nc). T_2 mô tả đối tượng được đem ra tính toán, đối tượng cụ thể. Các từ loại thực hiện chức năng T_2: danh từ riêng (Np), danh từ đơn thể (Nc), danh từ trừu tượng (Na). Trong [11], tác giả đã chỉ ra rằng, T1 là trung tâm về mặt ngữ pháp, T2 là trung tâm về mặt ý nghĩa từ vựng. Trong thực tế, phần trung tâm có thể có đầy đủ hoặc cũng có thể thiếu T1 hoặc T2 hoặc là thiếu cả hai. Nếu phần trung tâm chỉ chứa T1, nghĩa là phần T2 bị thiếu đã được đề cập trước đó. Nghiên cứu của chúng tôi chỉ xét những trường hợp phần trung tâm có chứa ít nhất là T2.

** Chú thích ngữ nghĩa cho các khái niệm:*

Quá trình ánh xạ các từ, cụm danh từ vào các khái niệm trong Ontology hình thức có thể xem như quá trình chú thích ngữ nghĩa cho các từ, cụm danh từ đó. Tuy nhiên, chú thích ngữ nghĩa cho các từ nói chung là một vấn đề khó. Một từ có thể mang nhiều ý nghĩa khác nhau trong các ngữ cảnh khác nhau, ở các vùng miền khác nhau, thậm chí ngay cả con người cũng không thể nhận biết được. Phần nhiều các hệ thống chú thích ngữ nghĩa như [13], [8], [10] đều tập trung vào vấn đề chú thích ngữ nghĩa cho các thực thể có tên. Các hệ thống chú thích ngữ nghĩa cho các từ như [4], [9] đều dựa trên phương pháp học máy, thống kê. [9] tập trung vào miền dữ liệu cụ thể (về trang phục), trong đó nhập nhằng ngữ nghĩa rất ít xảy ra. Nghiên cứu của chúng tôi tập trung giải quyết vấn đề chú thích ngữ nghĩa cho các từ trong miền dữ liệu liên quan đến máy tính. Các tài liệu được sử dụng là các bài báo trong tạp chí Thế Giới Vi Tính. Với miền dữ liệu được thu hẹp như trên, ta có được những thuận lợi: i) các từ thường chỉ mô tả một khái niệm, vấn đề nhập nhằng ngữ nghĩa là rất ít, ii) Các từ ngữ, mẫu câu thường được lặp lại theo mẫu. Quá trình chú thích ngữ nghĩa gồm các bước:

Chú thích ngữ nghĩa dựa trên từ điển: từ nhận xét là nhập nhằng ngữ nghĩa xảy ra là rất ít trong miền dữ liệu quan tâm, ví dụ như các cụm từ hệ điều hành, RAM, ổ đĩa cứng, …, chúng tôi xây dựng một tập các từ, cụm danh từ tương ứng với các khái niệm đã tồn tại trong Ontology. Các từ và cụm danh từ này không nhập nhằng và biểu diễn các khái niệm thông dụng. Thông qua đó, các khái niệm biểu diễn bởi các từ, cụm danh từ này sẽ được nhận diện một cách nhanh chóng.

Chú thích ngữ nghĩa dựa trên khoảng cách ngữ nghĩa: để tính toán khoảng cách ngữ nghĩa giữa từ, cụm danh từ với một nút trong Ontology, chúng tôi khảo sát công thức độ tương tự được trình bày trong [1]. Tương tự trong [16], Ontology sẽ được duyệt từ gốc để tìm khái niệm phù hợp nhất. Ở mỗi nút hiện hành, độ tương tự của từ hay cụm danh từ với các nút con của nó sẽ được xét.

+ Nếu giá trị độ tương tự lớn nhất nhỏ hơn giá trị của nút hiện hành, thì nút hiện hành sẽ là nút phù hợp nhất với từ hay cụm từ đó.

+ Ngược lại, tiếp tục duyệt cây với nút con có giá trị độ tương tự lớn nhất.

Thủ tục dừng khi điều kiện dừng được thoả mãn hoặc gặp nút lá.

** Chú thích ngữ nghĩa cho các thực thể:*

Chúng tôi không cố gắng xây dựng khối chú thích ngữ nghĩa cho thực thể từ ban đầu, mà sử dụng kết quả của khối chú thích ngữ nghĩa cho các từ, cụm danh từ. Các tác vụ thực hiện bao gồm:

Chú thích ngữ nghĩa dựa trên từ điển: trong các bài báo trên tạp chí Thế Giới Vi Tính, các thực thể có tên thông dụng như Microsoft, IBM, Windows,… được sử dụng thường xuyên mà không chú thích gì thêm vì người đọc đã quá quen thuộc với chúng. Do đó, để giúp cho quá trình chú thích được nhanh chóng và chính xác hơn, một tập các tên riêng ứng với các thực thể có tên thông dụng đã được xây dựng.

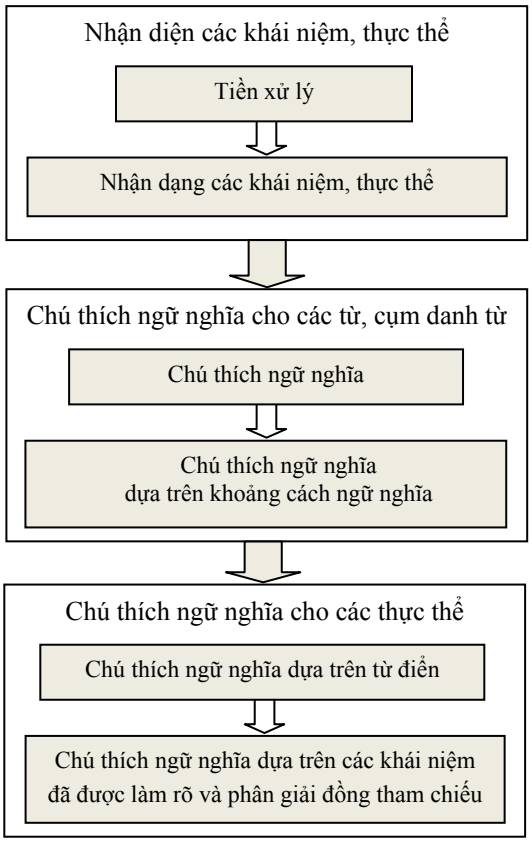
Chú thích ngữ nghĩa dựa trên các khái niệm đã được làm rõ: dựa trên [12], các đồng tham chiếu được khai thác xử lý chuyên biệt ở mức đồng tham chiếu tên riêng dựa trên các luật đề xuất như:

Luật 1: N1 và N2 giống nhau hoàn toàn (kế thừa từ [12])

Luật 2: N2 là viết tắt các ký tự đầu trong N1 (kế thừa từ [12])

Luật 3: N2 là phần cuối của N1 (áp dụng cho tên sản phẩm)

Luật 4: N2 là phần đầu của N1 (áp dụng cho tên công ty, tổ chức kinh doanh)



Hình 2. *Mô hình rút trích và chú thích các khái niệm và thực thể cho Ontology*

4. KẾT QUẢ THỰC NGHIỆM

Sự chính xác của việc xây dựng Ontology thể hiện ở sự chính xác của phương pháp nhận diện và chú thích ngữ nghĩa cho các cụm từ. Hai tham số là độ chính xác (Precision) và độ đầy đủ (Recall) được sử dụng [17] để đánh giá hiệu quả của phương pháp.

Việc thực nghiệm được tiến hành đánh giá trên 5 tài liệu chọn từ tạp chí Thế Giới Vi Tính. Đầu vào của chương trình là các từ và cụm danh từ đã được chú thích ngữ nghĩa và được hiệu chỉnh bằng tay. Tổng số cụm danh từ mô tả các khái niệm và thực thể là 210. Kết quả thể hiện trong bảng 3 và 4, trong đó công thức GO là công thức tính độ tương tự dựa trên các từ chung giữa các định nghĩa, và công thức cosine là công thức độ tương tự dựa trên cosine giữa hai vector ngữ cảnh:

Bảng 3: *Kết quả rút trích và ánh xạ các từ, cụm danh từ chỉ khái niệm*

R	A	Ra	Precision	Recall
105	76	39	**51,3%**	**37,1%**

Bảng 4. *Kết quả rút trích và ánh xạ các từ, cụm danh từ chỉ thực thể*

R	A	Ra	Precision	Recall
21	41	16	**39%**	**76,2%**

Theo Bảng 3 và Bảng 4, có thể thấy kết quả đạt được của hai công thức là gần như nhau và đều khá thấp. Đối với các khái niệm trong Bảng 3, quá trình nhận dạng sai do hai nguyên nhân: (a) khái niệm không nằm trong lĩnh vực quan tâm nhưng hệ thống không thể phân biệt và (b) nhận dạng sai do khối chú thích ngữ nghĩa dựa trên độ tương tự. Do đó, cần khảo sát các công thức độ tương tự khác để tăng độ chính xác cho kết quả.

Đối với các thực thể ở Bảng 4 cũng tương tự, các thực thể được nhận diện chủ yếu do khối nhận diện dựa trên từ điển. Các trường hợp bị sai phần lớn do ảnh hưởng của khối chú thích ngữ nghĩa cho các khái niệm. Tuy nhiên, độ đầy đủ trong cả hai công thức đều khá cao. Lý do là một phần rất lớn các thực thể không được nhận diện bởi khối nhận diện thực thể và sẽ được nhận diện lại bởi khối nhận diện khái niệm.

Từ các kết quả trên cho thấy, quá trình chú thích ngữ nghĩa cho các khái niệm là tương đối

khó khăn và cần có những nghiên cứu tiếp theo để cải thiện độ chính xác. Khối chú thích ngữ nghĩa dựa trên độ tương tự được xây dựng trong nghiên cứu chủ yếu dựa trên các thông tin về ngữ cảnh để xác định ngữ nghĩa của từ, cụm danh từ. Tuy nhiên, để xác định được chính xác ngữ nghĩa của chúng, các thông tin mang tính chất nền tảng đóng vai trò rất quan trọng.

5. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã xây dựng một Ontology kết hợp giữa tri thức từ vựng và tri thức về các khái niệm. Ontology gồm có một thành phần chứa các tri thức từ vựng tương tự như Wordnet và một thành phần chứa các tri thức ngữ nghĩa về thế giới. Các từ vựng (từ, cụm danh từ) được rút trích và ánh xạ đến các lớp mô tả ngữ nghĩa tương ứng một cách bán tự động. Hơn nữa, các từ vựng sẽ được ánh xạ đến các lớp tương ứng tùy theo miền dữ liệu cụ thể. Ngoài ra, các khái niệm trong Ontology có thể được truy vấn gần đúng thông qua công thức độ tương tự được đề nghị trong nghiên cứu. Nghiên cứu cũng đề xuất một phương pháp chú thích ngữ nghĩa cho từ, cụm từ kết hợp giữa Ontology và thống kê. Phương pháp đã được áp dụng trong quá trình xây dựng Ontology.

Từ đó, việc nghiên cứu cải tiến hệ thống trong tương lai được chúng tôi dự trù thực hiện dựa trên một số điểm giới hạn trong kết quả nghiên cứu hiện tại như: (a) Ontology hiện tại chỉ lưu trữ danh từ và cụm danh từ và các khái niệm, thực thể tương ứng. Trong khi đó, động từ cũng đóng vai trò hết sức quan trọng đối với ngữ nghĩa của tài liệu. Việc nắm bắt và lưu trữ ngữ nghĩa của các động từ là vấn đề cần được khảo sát và nghiên cứu. (b) Việc mở rộng miền dữ liệu tri thức là cần thiết để có thể giải quyết các bài toán trong các lĩnh vực rộng hơn. Tuy nhiên, việc mở rộng Ontology sẽ dẫn đến những nhập nhằng ngữ nghĩa vì các từ có thể mang những nghĩa khác nhau trong những lĩnh vực khác nhau. Việc tổ chức lưu trữ Ontology hữu ích cho việc giải quyết nhập nhằng ngữ nghĩa là một hướng mở rộng cần quan tâm nghiên cứu. (c) Trong công thức độ tương tự, các từ có xác suất đồng xuất hiện cao đối với một khái niệm sẽ được thống kê như là mô tả ngữ cảnh của khái niệm đó. Tuy nhiên trong các tài liệu, các ngữ cảnh xuất hiện với một khái niệm tuỳ thuộc vào chủ đề của tài liệu. Đây là vấn đề dữ liệu thưa, có ảnh hưởng lớn đến kết quả tính toán. Việc giải quyết vấn đề này là một trong những hướng nghiên cứu đáng quan tâm. (d) Nghiên cứu cũng chưa giải quyết vấn đề khi ánh xạ một từ, cụm danh từ đến một khái niệm nhưng khái niệm đó chưa có trong Ontology. Vấn đề này theo ý kiến của chúng tôi là thường xuyên xảy ra trong thực tế và có ý nghĩa trong việc mở rộng một Ontology. Do đó, vấn đề này rất cần được quan tâm nghiên cứu trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Banerjee S. and Pederson T., 2003, Extended Gloss Overlaps as a Measure of Semantic Relatedness, In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, IJCAI-03, pp. 805–810.

[2] Biemann C., 2005, Ontology learning from text: A survey of methods, *LDV-Forum*, Vol. 20, Issue 2, pp. 75-93.

[3] Caraballo S. A., 1999, Automatic Construction of a Hypernym-Labeled Noun Hierrachy from Text, In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL-99*, pp. 120–126.

[4] Dill S., Gibson N., Gruhl D., Guha R., Jhingran A., Kanungo T., Rajagopalan S., Tomkins A., Tomlin J.A. and Zien J.Y., 2003, SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation, In *Proceedings of Twelfth International World Wide Web Conference*, *Budapest, Hungary*, pp. 178-186.

[5] Harris Z. S., 1968, *Mathematical Structures of Language*, Interscience Publishers John Wiley & Sons, New York.

[6] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora, In *Proceedings of the Fourteenth International Conference on Computational Linguistics, COLING 1992*, Nantes, France, pp. 539–545.

[7] Hindle D., 1990, Noun Classification from Predicate-Argument Structures, In *Meeting of the Association for Computational Linguistics*, pp. 268–275.

[8] Kogut P. and Holmes W., 2001, AeroDAML : Applying Information Extraction to Generate DAML Annotations from Web Pages, In *Proceedings of First International Conference on Knowledge Capture*.

[9] Li J., Zhang L. and Yu Y., 2001, Learning to Generate Semantic Annotation for Domain Specific Sentences, In *the Workshop on Knowledge Markup and Semantic Annotation, the First International Conference on Knowledge Capture, K-CAP 2001*, Victoria B.C., Canada.

[10] Marynard D., Tablan V., Cunningham K. and Wilks Y., 2003, MUSE : a multisource entity recognition system, *Computers and the Humanities*, Website Reference : http://gate. ac. uk/sale/muse/muse.pdf.

[11] Nguyễn Tài Cẩn, "*Ngữ pháp tiếng Việt*", Nhà xuất bản Đại học và Trung học chuyên nghiệp, 1981.

[12] Nguyễn Thanh Hiên, 2005, *Phân giải sự đồng tham chiếu các thực thể có tên tiếng Việt*, Luận văn thạc sĩ, Trường Đại học Bách khoa Thành phố Hồ Chí Minh.

[13] Popov B., Kiryakov A., Kirilov A., Manov D., Ognyanoff D. and Goranov, M., 2003, KIM – Semantic Annotation Platform, In *Proceedings of 2nd International Semantic Web Conference (ISWC2003)*, Florida, USA, pp. 834-849.

[14] Widdows D., 2003, Unsupervised methods for developing taxonomies by combining syntactic and statistical information, In *Proceedings of HLT/NAACL 2003*, Edmonton, Canada, pp. 276–283.

[15] Witten I., Paynter G., Frank E., Gutwin C. and Neville-Manning C., 1999, KEA : Practical Automatic Keyphrase Extraction, *In Proceedings of ACM DL '99*, pp. 254-255.

[16] Witschel H. F., 2005, Using decision trees and text mining techniques for extending taxonomies. In *Proceedings of Learning and Extending Lexical Ontologies by using Machine Learning Methods*, Workshop at ICML-05.

[17] Nguyễn Quang Châu, 2011, *Mô hình rút trích Cụm từ đặc trưng ngữ nghĩa trong tiếng Việt*, Luận văn tiến sĩ , Trường Đại học Bách khoa – Đại học Quốc gia Thành phố Hồ Chí Minh.