PHÂN LOẠI TÊN THỂ LOẠI Ở WIKIPEDIA TIẾNG VIỆT

Tạ Hoàng Thắng[a*]

[a]*Khoa Công nghệ Thông tin, Trường Đại học Đà Lạt, Lâm Đồng, Việt Nam*

Lịch sử bài báo
Nhận ngày 09 tháng 01 năm 2017 | Chỉnh sửa ngày 17 tháng 04 năm 2017
Chấp nhận đăng ngày 17 tháng 05 năm 2017

Tóm tắt

Wikipedia nổi tiếng là một bách khoa toàn thư mở lớn nhất hiện nay với mục đích phổ cập kiến thức cho tất cả mọi người trên thế giới. Với việc áp dụng robot trong khâu tạo bài tự động, dự án tiếng Việt là một trong 13 dự án ngôn ngữ có hơn một triệu bài viết. Tuy nhiên, điều đó tạo cho Wikipedia tiếng Việt nhiều thách thức trong việc nâng cao chất lượng bài, sắp xếp thể loại, chống phá hoại nội dung và nhiều công tác khác. Trong bài báo này, chúng tôi phân loại thể loại ở Wikipedia tiếng Việt, chi tiết hơn là cấu trúc và các quy ước đặt tên thể loại. Phương pháp chính là áp dụng các tiêu chuẩn và cấu trúc thể loại sẵn có ở tiếng Anh, một dự án Wikipedia lớn nhất về mặt thông tin đóng góp, từ đó áp dụng cho phiên bản tiếng Việt. Tuy nhiên, điều đó không thực hiện dễ dàng, do đó chúng tôi phải kết hợp nhiều phương pháp xã hội cũng như chuyên môn để đạt được sự kỳ vọng. Việc phân tích tên thể loại và dữ liệu từ Wikidata được chúng tôi áp dụng là một tiền đề xây dựng một công cụ chuyển dịch tên thể loại từ tiếng Anh sang tiếng Việt.

Từ khóa: Phân loại thể loại; Quy ước đặt tên; Thể loại Wikipedia.

1. GIỚI THIỆU

Cây thể loại tại dự án Wikipedia tiếng Anh là đối tượng nghiên cứu của nhiều học giả trên thế giới với nhiều bài báo về tái cấu trúc thể loại, loại bỏ thể loại dư thừa và phân tích cấu trúc thể loại, rút trích các quan hệ ngữ nghĩa trên thể loại… Việc quản lý cấu trúc thể loại khó thực hiện một cách hiệu quả ở các dự án ngôn ngữ Wikipedia nhỏ và trung bình vì vấn đề hạn chế về mặt nhân lực. Do đó, nhu cầu quản lý thể loại tại các dự án cũng hết sức cần thiết. Trước hết, chúng tôi phân tích tên thể loại tiếng Anh và tiếng Việt thành các cấu trúc mẫu NLP tương ứng với nhau, từ đó áp dụng việc dịch thuật để tạo mới tên thể loại tiếng Việt từ tiếng Anh thông qua các cấu trúc này. Tiếp đến, dựa vào cấu trúc thể loại tiếng Anh, chúng tôi cũng có thể sắp xếp các bài viết vào các thể loại

* Tác giả liên hệ: Email: thangth@dlu.edu.vn

tiếng Việt vừa mới tạo một cách hợp lý. Điều này giúp tăng độ mịn cấu trúc thể loại tại dự án Tiếng Việt, giảm bớt việc thao tác tạo thể loại mới bằng tay nhằm thúc đẩy chất lượng cộng tác tại Wikipedia tiếng Việt.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Trong bài báo này, chúng tôi dựa chủ yếu vào bài báo của Nastase và Strube (2008) về việc phân loại các dạng tên thể loại trong tiếng Anh. Nghiên cứu này chỉ ra tên và cấu trúc thể loại trong Wikipedia là một nguồn quan hệ giữa các khái niệm. Từ các phân tích và thực nghiệm chỉ ra các thể loại có thể phân tích thành dạng thể loại, mẫu phân tích và các quan hệ kèm theo. Tên thể loại khi được phân tích thành các mẫu với các từ nối có ký hiệu dựa theo tập Penn Treebank (Santorini, 1990).

Việc phân loại quy mô lớn thông tin các quan hệ được xây dựng dựa trên hệ thống thể loại ở Wikipedia, được phân tích từ các mối quan hệ giữa các thể loại sử dụng các phương pháp dựa trên kết nối trong mạng lưới và việc so khớp cú pháp từ vựng. Các mẫu phân tích từ tên thể loại cũng tương tự như cách của Nastase và Strube (2008), nhưng tập trung sâu vào các mẫu *isa* và *nonisa*. Nghiên cứu của Nguyễn, Lê, Tôn, và Nguyễn (2012) cũng chứa các mẫu *isa* và *nonisa* cho thể loại, tuy nhiên nhóm xây dựng cách tiếp cận mô hình Ontology tiếng Việt hơn là tập trung vào thể loại.

Tên thể loại còn được sử dụng là đầu vào của việc phân loại văn bản trong bài báo của Barak, Dagan, và Shnarch (2009). Kết quả nghiên cứu dựa vào tính giống nhau trong không gian LSA, từ đó nhận biết sự tương tự về bối cảnh ở dạng thô. Barak và ctg. (2009) cũng nhận diện các tham chiếu bền vững theo ngữ nghĩa tên thể loại, chứa biến thể đặc biệt để mở rộng từ vựng. Ngoài ra, cũng có nghiên cứu chủ yếu về cấu trúc thể loại Wikipedia từ đó để đo độ tương đồng giữa 2 thể loại cụ thể nào đó dựa theo các mối quan hệ thể loại cha, con (Xu, Takeda, Hamasaki, & Wu, 2010) hay là xây dựng đồ thị thể loại Wikipedia dựa trên các thuật toán xử lý ngôn ngữ tự nhiên (Zesch & Gurevych, 2007).

Các nghiên cứu trên chỉ đơn thuần thực hiện ở tiếng Anh, ý tưởng của chúng tôi là chuyển hóa thành tiếng Việt để cho thấy sự tương đồng về tên thể loại trong tiếng Việt,

từ đó xây dựng một tiêu chuẩn so khớp giữa tiếng Anh và tiếng Việt, đồng thời hướng tới việc tiếp cận dịch mẫu tên thể loại từ tiếng Anh sang tiếng Việt.

3. TIÊU CHUẨN ĐẶT TÊN THỂ LOẠI Ở WIKIPEDIA TIẾNG VIỆT

Để phân loại tên thể loại tiếng Việt, trước hết chúng ta cũng phải xem xét đến một số tiêu chuẩn đặt tên thể loại ở Wikipedia tiếng Việt. Tiêu chuẩn về thể loại và cách đặt tên thể loại chủ yếu kế thừa từ phiên bản tiếng Anh tương đương và kèm theo sự đóng góp của các biên tập tại dự án tiếng Việt. Nhìn chung tên thể loại phải ngắn gọn, súc tích mà vẫn mô tả đầy đủ ý nghĩa của các bài viết, thể loại con chứa bên trong thể loại đó. Nội dung các tiêu chuẩn về thể loại và các thông tin liên quan có thể tìm thấy tại bài viết có tên Wikipedia: Thể loại ở Wikipedia[1]. Theo đó, một số quy ước về tên thể loại phổ biến như sau.

Sử dụng thể loại có "năm" với tất cả các thể loại, ví dụ *Thể loại:Khoa học năm 1990* thay vì *Thể loại:Khoa học 1990*. Sử dụng thể loại chứa tên các quốc gia phổ biến trong tiếng Việt như *Thể loại:Văn hóa Úc* thay vì *Thể loại:Văn hóa Australia.*

Ưu tiên sử dụng số nhiều cho thể loại, ví dụ *Category:Cities of France* được dịch là *Thể loại:Các thành phố ở Pháp* thay vì *Thể loại:Thành phố ở Pháp*. Tuy nhiên, theo dự án về thể loại (Wikimedia, 2015) thì trường hợp này vẫn nên dùng số ít. Vì vậy, kết quả vẫn là *Thể loại:Thành phố ở Pháp* được ưu tiên. Do đó, trong bài viết này chúng tôi khuyến cáo sử dụng số ít trong tiếng Việt khi dịch từ các cụm từ số nhiều tiếng Anh.

4. PHÂN LOẠI TÊN THỂ LOẠI

Để phục vụ cho mục đích chủ yếu là dịch tên thể loại từ tiếng Anh sang tiếng Việt, chúng tôi phân loại tên thể loại theo số lượng biến trong mẫu phân tích được. Phương pháp gần giống cách phân tích về dạng thể loại của Nastase và Strube (2008), chỉ khác ở chỗ chúng tôi chú trọng về số lượng biến hơn là các dạng thể loại mang tính ngữ pháp. Cách tiếp cận này giúp chúng tôi định rõ số lượng từ/cụm từ cần dịch để phục vụ cho mục

[1]https://vi.wikipedia.org/wiki/Wikipedia:Thể_loại

đích dịch tên thể loại từ tiếng Anh sang tiếng Việt trong công cụ dịch thuật và các nghiên cứu tiếp theo.

4.1. Mẫu đơn

Các mẫu đơn (mẫu một biến) sử dụng một biến để định nghĩa tên thể loại. Biến này thường là một danh từ, cụm danh từ hay một số, và không chứa các liên từ và cũng như không thể phân chia thành các thành phần nhỏ hơn. Ký hiệu mẫu đơn được định nghĩa là $p = x_1$. Một số ví dụ về mẫu đơn như trong Bảng 1.

Bảng 1. Phân tích một số trường hợp mẫu đơn

Tên thể loại	Mẫu	Dạng thể loại
Khoa học *Science*	p = x_1, x_1 = Khoa học	mẫu đơn
Động vật đặc hữu *Endemic fauna*	p = x1, x1 = Động vật đặc hữu (endemic fauna = adj + noun)	mẫu đơn
1990	p = x1, x1 = 1990	mẫu đơn

Trong Bảng 1, thể loại *Khoa học* là một mẫu đơn vì nó chỉ chứa duy nhất một từ. Thể loại *Động vật đặc hữu* là mẫu đơn vì đó là một cụm danh từ không thể phân tách thành các cụm nhỏ hơn. Các thể loại về số cũng có thể coi một mẫu đơn, chẳng hạn như *1990*.

4.2. Mẫu hai biến

Mẫu này chứa 2 biến (x_1, x_2) và chứa một liên từ (c_1) hoặc không có liên từ. Chúng tôi mô tả mẫu này bằng ký hiệu $p=x_1c_1x_2$, và một vài ví dụ về mẫu hai biến được liệt kê như Bảng 2.

Trong Bảng 2, các liên từ (*conjunction*) chủ yếu là các giới từ, mệnh đề quan hệ giản lược hay đôi khi là rỗng. Chúng tôi nhận ra rằng không có thể loại nào có tên có liên từ ở vị trí đầu hoặc cuối. Vì vậy, các thể loại tiếng Anh dạng như *By country*, *Cities in* hay *By country in* chắc chắn sẽ không tồn tại vì không đủ nghĩa cấu thành tên thể loại.

Thể loại *Films directed by Charles Frend* có mẫu phân tích theo Nastase và Strube (2008) là X [VBN] Y. Trong đó [VNB] là dạng giản lược của động từ quá khứ 3

trong tiếng Anh. Thể loại này khi được dịch sang tiếng Việt theo dự án Wikimedia (2015) thì có 2 trường hợp đó là:

- *Phim được đạo diễn bởi Charles Frend*: Mang tính bị động, và ít có tính thuần Việt, rất dễ nhận diện đây là cách dịch từng từ từ tiếng Anh.
- *Phim do Charles Frend đạo diễn*: Mang tính thuần Việt hơn.

Bảng 2. Một số mẫu thể loại là dạng 2 biến

Tên thể loại	Mẫu $p=x_1c_1x_2$		
	x_1	c_1	x_2
Nông nghiệp theo quốc gia *Agriculture by country*	Nông nghiệp *Agriculture*	theo *by*	quốc gia *country*
Thành phố ở Pháp *Cities in France*	Thành phố *Cities*	ở *in*	Pháp *France*
Người từ California *People from California*	Người *People*	từ *from*	California *California*
Phim được đạo diễn bởi Charles Frend *Films directed by Charles Frend*	Phim *Films*	được đạo diễn bởi *directed by*	Charles Frend *Charles Frend*
Văn hóa giao thông *Transport culture*	Giao thông *Transport*	Ø	Văn hóa *Culture*
Khoa học năm 2015 *2015 in science*	năm 2015 *2015*	- *in*	Khoa học *Science*
Sách về Việt Nam *Books about Vietnam*	Sách *Books*	về *about*	Việt Nam *Vietnam*

Thể loại *Transport culture* là trường hợp mẫu 2 biến mà không có liên từ ở giữa. Mẫu này được xem là một cụm danh từ. Theo Nastase và Strube (2008) thì mẫu này có dạng XY với *X = transport* và *Y = culture*. Khi dịch về tiếng Việt thì mẫu thành trở thành YX. Thể loại *Transport culture* được dịch trong tiếng Việt là *Văn hóa giao thông*.

Thể loại *2015 in science* là thể loại có chứa năm, theo tiêu chuẩn đặt tên thể loại của Wikipedia tiếng Việt, các mẫu này đều phải có chữ "năm" ở trước số năm để xác định rõ ràng nghĩa. Trường hợp này thể loại *2015 in science* được dịch ngược thành *Khoa học năm 2015* và không có dịch giới từ *in*. Cách dịch giới từ tiếng Anh sang tiếng Việt cũng khá đa dạng và phức tạp và tùy theo nhiều trường hợp vì vậy sẽ không đề cập đến trong bài viết này.

4.3. Mẫu ba biến

Mẫu này bao gồm 3 biến (x_1, x_2, x_3) và có từ 0 đến 2 liên từ, là một mẫu kết hợp giữa mẫu hai biến và mẫu đơn. Chúng ta có thể biểu diễn mẫu 3 thành ký hiệu $p = x_1c_1x_2c_2x_3$. Các danh mục, thuộc về mẫu này, thường ít phổ biến hơn các mẫu đơn và mẫu 2 biến ở Wikipedia. Bảng 3 mô tả một số ví dụ của mẫu ba biến.

Bảng 3. Một số thể loại được liệt kê là dạng mẫu ba biến

Tên thể loại	Mẫu $p=x_1c_1x_2c_2x_3$				
	x_1	c_1	x_2	c_2	x_3
Khoa học và công nghệ theo quốc gia *Science and technology by country*	Khoa học *Science*	và *and*	công nghệ *technology*	theo *by*	quốc gia *country*
Nợ chính phủ theo quốc gia *Government debt by country*	Chính phủ *Goverment*	∅	nợ *debt*	theo *by*	quốc gia *country*
Tiểu thuyết khoa học giả tưởng *Science fiction novels*	Khoa học *science*	∅	giả tưởng *fiction*	∅	tiểu thuyết *novels*

Thể loại *Government debt by country* là một trường hợp phổ biến của mẫu 3 biến khi mà thường chỉ chứa một liên từ (trong trường hợp này là giới từ *by*). Thể loại được phân tích thành 2 cụm dựa theo giới từ *by* đó là: *Government debt* và *country*, trong đó cụm *Government debt* chính là một mẫu 2 biến kiểu XY. Thể loại được dịch trong tiếng Việt là *Nợ chính phủ theo quốc gia.*

Thể loại *Science fiction novels* là một trường hợp đặc biệt của mẫu 3 biến khi không hề có bất kỳ liên từ nào, trường hợp này được xem là một cụm danh từ với 3 danh từ kết hợp liên tiếp nhau. Việc dịch cụm này ra tiếng Việt cũng là một vấn đề, trường hợp có 2 cách dịch như sau:

- *Tiểu thuyết khoa học giả tưởng*: Đây là cách dịch thông thường, tức là đi từ bên phải sang, lấy từng từ dịch, phần còn lại xem có thể diễn dịch được hay không, nếu không lại tiếp tách cụm như trên cho đến khi tất cả mọi cụm đều được dịch, ngược lại thì sẽ không dịch được cụm này. Chẳng hạn, lấy *novels* dịch thành *tiểu thuyết*, tiếp đến cụm từ *Science fiction* dịch thành *Khoa học giả tưởng*. Đảo ngược các cụm dịch được thì có kết quả là *Tiểu thuyết khoa học giả tưởng.* Các cụm từ dịch được hoàn toàn lấy giá trị từ Wikidata với

các cụm từ được liên kết với nhau ở các phiên bản ngôn ngữ (Vrandečić & Krötzsch, 2014).

- *Tiểu thuyết giả tưởng khoa học*: Kiểu dịch này ngược với cách trên là đi từ bên trái sang, lấy từng từ, và dịch các từ này và cụm còn lại cho đến khi dịch hết toàn bộ từ. Trong trường hợp này, chúng ta có thể tách làm 2 cụm: *science* và *fiction novels*. Sau đó tiến hành dịch thì được kết quả là: *Khoa học và tiểu thuyết giả tưởng*, đảo ngược thì có kết quả như trên. Tuy nhiên, khi lấy giá trị từ Wikidata, chúng ta không thể dịch được cụm từ *fiction novels*, vì vậy cụm từ này nếu dùng từ điển dịch hay các công cụ khác thì đây có thể là cách dịch sai hoặc không phổ biến.

4.4. Các mẫu nhiều hơn ba biến

Ngoài ra, tên thể loại ở Wikipedia còn được phân tích thành các mẫu với số biến lớn hơn 3. Các mẫu này đều có thể phân tách thành các mẫu con, từ đó có thể hiểu được cấu trúc các mẫu. Chẳng hạn, thể loại tiếng Anh *Science fiction novels by nationality* có thể phân tách thành mẫu 3 biến *Science fiction novels* và mẫu đơn *nationality* với liên từ là giới từ *by*. Thể loại này được dịch ra tiếng Việt là *Tiểu thuyết khoa học giả tưởng theo quốc tịch*.

5. MỘT SỐ TRƯỜNG HỢP DỊCH THUẬT

Từ việc phân tích tên thể loại thành các mẫu sử dụng số lượng biến trong bài, chúng tôi thử áp dụng các mẫu này trong việc dịch thuật thông qua công cụ chúng tôi tự tạo. Trong bài báo này, chúng tôi sẽ không nêu chi tiết cách thức thực hiện và tập dữ liệu đầu vào mà cũng như phương pháp thực hiện mà chỉ nêu một số ví dụ về dịch thuật được sử dụng thông qua công cụ dịch thuật. Các phần trên được chúng tôi tiếp tục nghiên cứu và xuất bản ở các nghiên cứu khác.

5.1. Quy trình dịch thuật

Trong phần này, chúng tôi trực tiếp đề cập các quy trình chính về cách dịch mẫu thay vì trình bày toàn bộ các xử lý chi tiết mà công cụ dịch thuật thực hiện. Vì việc dịch

không đơn giản với một số mẫu phức tạp, chúng tôi chọn lựa các mẫu có giới từ, mẫu một biến và mẫu không chứa liên từ làm các mẫu thử nghiệm trong công cụ dịch. Các quy trình dịch trong công cụ chia làm các bước chính sau đây:

- **Bước 1** *(Dịch mẫu một biến*): Đầu tiên, đầu vào là chúng tôi có các tên thể loại tiếng Anh cần dịch, chúng tôi xem các đầu vào này mặc định là mẫu một biến, chúng tôi kiểm tra xem các tên này có ở Wikidata hay không, nếu có thì dừng việc dịch và cho ra kết quả. Nếu không thì cho ra kết quả là không dịch được.

- **Bước 2** *(Dịch mẫu chứa giới từ)*: Chúng tôi dò tìm xem trong tên thể loại có chứa giới từ hay không, nếu không có chúng tôi chuyển sang Bước 3. Trường hợp tên thể loại chứa giới từ thì tách tên thể loại thành 3 phần, phần trước giới từ, giới từ, phần sau giới từ. Tiếp tục lặp lại Bước 1 với ba phần này, sau đó gom các kết quả có được lại. Chỉ cần một trong các bước cho ra kết quả không tìm thấy kết quả dịch được từ Wikidata thì dừng ngay quá trình dịch và cho kết quả không dịch được.

- **Bước 3** *(Dịch mẫu là cụm danh tính từ*): Tên thể loại được tách làm 2 phần gồm từ cuối cùng của thể loại và phần còn lại. Tiếp tục lặp Bước 1 cho hai phần này. Nếu một trong 2 phần không dịch được, thì chúng tôi lại tách cụm thành 2 từ cuối cùng của thể loại và phần còn lại. Tiếp tục lặp Bước 1 cho hai phần này cho đến khi thể loại được tách thành 2 phần: Từ đầu tiên và phần còn lại mà vẫn không cho ra kết quả dịch thì dừng việc dịch. Kết quả dịch được sẽ được đảo ngược vị trí.

- **Bước 4** *(So khớp và giám sát tay):* Chúng tôi sử dụng một module của tác giả Dao và Simpson (2015) để so khớp cụm từ kết quả với các cụm từ dịch được trước đó. Nếu kết quả cho ra kết quả tổng điểm trung bình >0.5 (tổng trung bình của phần so khớp với cụm tiếng Anh và cụm tiếng Việt với một thể loại tương tự) thì chúng tôi giữ kết quả này. Tiếp tục, chúng tôi kiểm tra sự tương tự giữa cách dịch của thể loại cần dịch và thể loại tương tự thông qua cấu trúc

tên phân tích được (*NameAnalysis*) để đồng bộ về cách dịch cho các thể loại cùng một cụm đặc trưng. Cuối cùng, chúng tôi sử dụng phương pháp giám sát bằng tay để kiểm tra kết quả dịch có hợp lý trước khi đưa ra kết quả chính xác cuối cùng.

5.2. Các ví dụ dịch thuật

- **Dịch trường hợp một biến**: Đầu tiên, xét đến trường hợp dịch một biến. Đầu vào như sau:

Category:Honiara---Q7403236---Real Kakamora FC

Trong đó *Category:Honiara* có chỉ số Q-id là Q7403236 có nội dung đặc tả các liên kết ngôn ngữ, mệnh đề RDF, nguồn và nhiều phần khác ở Wikidata. Nếu dịch thành công tên thể loại này thì tên bài viết *Real Kakamora FC* sẽ được xếp vào thể loại mới này. Tiếp đến lấy từ *Honiara* (tên thủ đô của quần đảo Solomon) tìm kiếm ở Wikidata thì nhận thấy ở Q40921 có liên kết với tên giống với một bài viết trong tiếng Việt. Do đó, *Category:Honiara* dịch thành *Thể loại:Honiara* trong tiếng Việt. Trường hợp chúng tôi đặt điểm chính xác (điểm so khớp) là một vì hiển nhiên là lấy trực tiếp từ Wikidata với tên bài tương ứng.

Trường hợp khác của mẫu một biến có thể là một thể loại có tên được dịch trực tiếp từ Wikidata mặc dù nó có thể chứa cụm danh tính từ được phân tích thành các cụm nhỏ hơn và chứa giới từ. Chúng tôi xét đến trường hợp sau:

Category:French Republican Calendar---Q8472929---Lịch

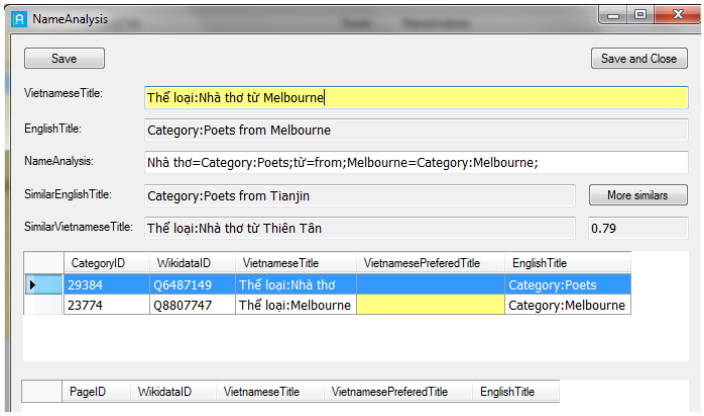
Thể loại trên có thể phân tích thành 2 cụm *Calendar* và *French Republican*, tuy nhiên do khi dịch để nguyên cụm *Frech Republican Calendar* chúng ta có kết quả tương ứng ở Wikidata là *Lịch cộng hòa*. Trong công cụ dịch, chúng tôi vẫn xem là đây là mẫu một biến vì nó được dịch trực tiếp nguyên cụm từ Wikidata.

- **Dịch mẫu chứa giới từ**: Mẫu chứa giới từ (liên từ) có thể chứa nhiều hơn một giới từ. Xét đến trường hợp mẫu như sau:

Category:Poets from Melbourne---Q8767587---Diane Fahey

Sau khi kiểm tra ở Bước 1 ở trên thì tên thể loại trên không phải là mẫu một biến cũng không thể dịch nguyên một cụm từ được từ Wikidata. Chúng ta xét đến xem mẫu này có giới từ hay không, vì có giới từ *from* trong tên thể loại nên chúng ta tách tên này thành 3 phần: *Poets*, *from* và *Melbourne*. Khi dò tìm ở Wikidata, chúng ta được các kết quả tương ứng với các phần: *Nhà thơ, từ* (dịch mặc định trong chương trình) và *Melbourne*. Ghép các cụm kết quả chúng ta có thể loại dịch được ra tiếng Việt là *Thể loại:Nhà thơ từ Melbourne*. Về vấn đề giới từ *from* có thể dịch thành *từ, rỗng* hoặc *đến từ, ...* Chúng tôi mặc định dịch *from* thành *từ*. Sau đó, chúng tôi tìm mẫu tương tự với tên thể loại tiếng Anh để xem cách dịch trước đó như thế nào để cho ra tên giới từ phù hợp như trong Hình 1.

Trong Hình 1, *Thể loại:Nhà thơ từ Melbourne* có thể loại tương tự là *Thể loại:Nhà thơ từ Thiên Tân* với điểm so khớp là 0.79, dựa vào việc so khớp cấu trúc phân tích tên (*NameAnalysis*) chúng tôi xác định được vì thể loại tương ứng dịch giới từ *from* thành *từ* vì vậy chúng tôi giữ cách dịch này. Hơn nữa, chức năng *More similars* cũng cho phép xem nhiều hơn các thể loại tương tự.



Hình 1. Ví dụ về phân tích tên thể loại của *Thể loại:Nhà thơ từ Melbourne*

- **Dịch mẫu không chứa liên từ**: Để tăng độ chính xác của các mẫu này, chúng tôi sử dụng đa số 2 bước cuối cùng của quy trình dịch (được nêu trong Mục 5.1): So khớp độ tương đương với các kết quả trước đó và giám sát bằng tay các kết quả dịch trên các mẫu này. Rõ ràng, chúng tôi cũng nhận thấy một vài cách dịch sai trong các mẫu này, tuy nhiên thông qua 2 bước dịch trên đã giảm thiểu phần lớn các kết quả bị dịch sai. Chúng tôi xét đến trường hợp

đầu tiên như sau:

Category:Water technology---Q6968052---Thẩm thấu

Trong đó *Category:Water technology* được phân tích là một mẫu 2 biến và không chứa liên từ, chúng ta dựa vào Bước 3 để tách thành 2 cụm: *Water* và *technology*. Dựa vào Wikidata, chúng ta cũng dịch được sang tiếng Việt là: *Nước và công nghệ*. Chúng ta đảo ngược các thành phần kết quả để có kết quả cuối cùng là *Thể loại:Công nghệ nước*. Tuy nhiên, đây là trường hợp đơn giản và cũng khá dễ dịch. Chúng tôi xét đến trường hợp khác như sau:

Category:Satellite navigation systems---Q6392458--- Dẫn đường chi tiết

Dựa theo Bước 3, thể loại trên cũng được tách thành 2 phần: *Satellite navigation và Systems*. Dựa vào Wikidata, chúng ta thu được các tên dịch được là: *GNSS* và *hệ thống*. Đảo ngược 2 phần này thì có kết quả *Thể loại:Hệ thống GNSS*. Tiếp đến chúng tôi so khớp với thể loại tương tự và dùng phương pháp giám sát bằng tay thì được kết quả vẫn là *Thể loại:Hệ thống GNSS*. Ở phần này, chúng tôi đề xuất một phương pháp bổ sung có thể là xem xét độ phổ biến của cụm từ ở Google và đánh giá bằng một thang điểm nào đó để đưa ra kết luận cuối cùng về kết quả dịch. Chẳng hạn, nếu dùng Google chúng ta có cụm từ *Hệ thống GNSS* với 977 kết quả tìm được.

Cuối cùng, với các trường hợp phức tạp và có thể dẫn đến trường hợp dịch sai ở mẫu không chứa liên từ. Giải pháp cuối cùng là có thể đưa vào danh sách đen để tránh dịch các từ, cụm từ này trong tương lai và tìm kiếm các phương pháp giải quyết tốt hơn.

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài viết đã phân tích một số dạng tên thể loại ở Wikipedia tiếng Việt, chủ yếu dựa trên số lượng biến được sử dụng trong mẫu phân tích. Việc phân tích này có ý nghĩa quan trọng trong việc tìm hiểu về tên thể loại trong tiếng Việt, đặc biệt hơn nữa có vai trò quan trọng trong việc chuyển dịch tên thể loại từ tiếng Anh sang tiếng Việt dựa trên bảng so khớp các mẫu thể loại trong cả 2 thứ tiếng này. Điều đó giúp biên tập viên giảm thiểu thời

gian dịch tên thể loại có thể lặp lại nhiều lần và mang tính thủ công, từ đó tập trung vào việc nâng cao chất lượng nội dung bài viết ở Wikipedia.

Hiện tại, chúng tôi đang xây dựng công cụ dịch từ tên thể loại từ tiếng Anh sang tiếng Việt với một dự án của Wikimedia (2015). Phương pháp chuyển dịch chủ yếu là lấy một thể loại tiếng Anh, phân tách thể loại này thành các thành phần con, dựa vào Wikidata để dịch các thành phần con này sang tiếng Việt theo các liên kết ngoại ngữ, so khớp các cách dịch trước đó để đạt được sự đồng nhất về cách dịch thuật ngữ, cho phép con người giám sát quá trình dịch và lặp lại thao tác dịch thuật này cho đến khi đạt kết quả như yêu cầu. Công cụ này đã dịch bán tự động hơn 5000 thể loại mới cũng như các tập triple kèm theo để sắp xếp bài viết vào từng thể loại. Chúng tôi cũng sử dụng bộ nhớ đệm gồm hơn 56000 thể loại, 10000 trang bài viết để tăng tốc độ lấy dữ liệu dịch thuật. Đánh giá một cách chủ quan, kết quả mang lại là khả quan và có những đóng góp quan trọng trong việc xây dựng và phát triển chất lượng bài viết tại Wikipedia tiếng Việt. Công cụ dịch thuật không được chúng tôi kỳ vọng để dịch tất cả các thể loại từ Wikipedia tiếng Anh sang Wikipedia tiếng Việt mà là một công cụ giúp ích cho quá trình dịch thuật và có đóng góp nhất định ở Wikipedia tiếng Việt.

Trong tương lai, chúng tôi tiếp tục xây dựng các dự án nhỏ để dịch các mẫu thể loại mới đa dạng hơn. Chúng tôi cũng xây dựng mô hình Ontology để phân loại các cụm từ trong tên thể loại, dựa vào công cụ Google Translate và đo độ phổ biến cụm từ trên Google để từ đó đưa ra kết quả dịch chính xác và hoàn thiện hơn. Chúng tôi hi vọng mang lại một cách nhìn khác về cách dịch cụm từ và là nguồn tham khảo hữu ích cho các nghiên cứu tiếp theo.

TÀI LIỆU THAM KHẢO

Barak, L., Dagan, I., & Shnarch, E. (2009). *Text categorization from category name via lexical reference*. Paper presented at The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, USA.

Dao, T. N., & Simpson, T. (2005). *Measuring similarity between sentences*. Retrieved from http://trac.research.cc.gatech.edu/ccl/export/184/SecondMindProject/SM/SM.WordNet/Paper/WordNetDotNet_Semantic_Similarity.pdf

Nastase, V., & Strube, M. (2008). *Decoding Wikipedia categories for knowledge acquisition*. Paper presented at The Twenty-third AAAI Conference on Artificial Intelligence, USA.

Nguyễn, Q. C., Lê, T. N., Tôn, L. P., & Nguyễn, V. T. (2012). Một hướng tiếp cận xây dựng Ontology tiếng Việt. *Tạp chí Đại học Công nghiệp, 14*(6), 23-31.

Ponzetto, S. P., & Strube, M. (2007). Deriving a large-scale taxonomy from Wikipedia. Paper presented at The AAAI Conference on Artificial Intelligence, USA.

Santorini, B. (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision)*. Philadelphia, USA: University of Pennsylvania.

Tuc, H. D. (2003). *Vietnamese-English bilingualism: Patterns of code-switching*. London, UK: Routledge Curzon Press.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM, 57*(10), 78-85.

Wikimedia (2015). *Project: Semi-automatically generated categories for Vietnamese Wikipedia*. Retrieved from https://meta.wikimedia.org/wiki/Grants:IEG/Semi-automatically_generate_Categories_for_Vietnamese_Wikipedia

Xu, L., Takeda, H., Hamasaki, M., & Wu, H. (2010). *Typing software articles with Wikipedia category structure*. Retrieved from http://www.nii.ac.jp/TechReports/public_html/10-002E.pdf

Zesch, T., & Gurevych, I. (2007). *Analysis of the Wikipedia category graph for NLP applications*. Paper presented at The TextGraphs-2 Workshop, USA.

CLASSIFYING CATEGORY NAMES IN VIETNAMESE WIKIPEDIA

Ta Hoang Thang[a*]

[a]The Faculty of Information Technology, Dalat University, Lamdong, Vietnam
[*]Corresponding author: thangth@dlu.edu.vn

Article history
Received: January 09th, 2017 | Received in revised form: April 17th, 2017
Accepted: May 17th, 2017

Abstract

Wikipedia is famous to be the biggest encyclopedia currently, the purpose of which is to spread knowledge for everyone in the world. By using robots in the process of article generation, Vietnamese Wikipedia is one of 13 language projects which has more than 1 million articles. However, this raises a lot of challenges for Vietnamese Wikipedia in article quality improvement, category classification, anti-vandalism and other tasks. In this paper, we classify categories in Vietnamese Wikipedia, particularly in category taxonomy and naming conventions. The crucial method is to adopt standards and category taxonomy in the English project, the biggest Wikipedia project in term of the amount of contributed information. Then we apply these to Vietnamese Wikipedia. To do this, we have to combine many social methods as well as techniques to gain expected results. The evaluation of category names and data results from Wikidata which we obtained is a first step to build a tool to translate English categories into Vietnamese categories.

Keywords: Naming convention; Name taxonomy; Wikipedia category.