Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Huế, ngày 07-08/6/2019
DOI: 10.15625/vap.2019.00071

XÁC ĐỊNH TƯƠNG ĐỒNG XUYÊN NGỮ ANH - VIỆT SỬ DỤNG MÔ HÌNH ĐỒ THỊ

Lê Thành Nguyên, Trần Gia Trọng Nhân, Trần Công Hậu, Đinh Điền

Trường Đại học Khoa học Tự Nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh

lethanhnguyen.vn@gmail.com, 1553023@student.hcmus.edu.vn, 1553010@student.hcmus.edu.vn, ddien@fit.hcmus.edu.vn

TÓM TẮT: *Bài toán xác định tương đồng ngữ nghĩa văn bản là một trong những bài toán đóng vai trò rất quan trọng, ảnh hưởng đến chất lượng của nhiều bài toán xử lý ngôn ngữ tự nhiên như truy vấn thông tin, tóm tắt văn bản, phát hiện đạo văn,... Đặc biệt trong thời đại hiện nay, với sự phát triển của các công cụ dịch tự động, thì bài toán xác định tương đồng ngữ nghĩa văn bản còn phải xem xét đến cả các trường hợp các cặp văn bản thuộc các ngôn ngữ khác nhau. Trong bài báo này, chúng tôi đề xuất sử dụng mô hình đồ thị để xác định tương đồng ngữ nghĩa xuyên ngữ Anh- Việt. Bên cạnh đó, chúng tôi cũng áp dụng bổ sung các phương pháp như điều chỉnh gán nhãn từ loại giữa văn bản tiếng Việt và văn bản tiếng Anh, bổ sung danh sách từ tiếng Việt đồng nghĩa, kết hợp các lớp đồ thị khác nhau. Kết quả thực nghiệm cho thấy việc sử dụng các phương pháp trên giúp nâng độ chính xác của mô hình từ 71,9% lên 76,3%.*

Từ khóa: *tương đồng, xuyên ngữ, đồ thị, Tiếng Việt.*

I. GIỚI THIỆU

Hiện nay, bài toán tìm kiếm và phát hiện tương đồng ngữ nghĩa văn bản đóng vai trò rất quan trọng trong nhiều bài toán xử lý ngôn ngữ tự nhiên như đánh giá chất lượng dịch máy, phát hiện đạo văn, tóm tắt văn bản, tìm kiếm văn bản xuyên ngữ,...

Ví dụ như hai câu sau được xem là tương đồng với nhau:

- Câu tiếng Việt: *Nếu tôi đặt hàng bây giờ, không biết khi nào tôi có thể nhận được sản phẩm đó.*

- Câu tiếng Anh: *If I order now, I wonder when I can receive the product.*

Trong khi đó, hai câu sau được xem là không tương đồng do khác nhau về mặt ý nghĩa:

- Câu tiếng Việt: *Nếu tôi đặt hàng bây giờ, không biết khi nào tôi có thể nhận được sản phẩm đó.*

- Câu tiếng Anh: *If I go now, I wonder when I can see the doctor.*

Việc tìm kiếm và phát hiện tương đồng có thể được thực hiện bằng cách thủ công, tuy nhiên cách này mất rất nhiều thời gian và công sức, đặc biệt là việc phát hiện tương đồng đối với hai văn bản sử dụng ngôn ngữ khác nhau. Do đó, việc áp dụng máy học là một cách thức phù hợp giúp giải quyết bài toán so sánh tương đồng ngữ nghĩa xuyên ngữ văn bản Anh - Việt.

Mặc dù đã có nhiều nghiên cứu về bài toán phát hiện tương đồng ngữ nghĩa văn bản xuyên ngữ, tuy nhiên, theo hiểu biết của cá nhân, hiện nay chưa có nhiều nghiên cứu trên cặp ngôn ngữ Anh - Việt. Trong bài báo này, chúng tôi sử dụng hướng tiếp cận đồ thị tri thức để tìm kiếm và phát hiện tương đồng giữa văn bản tiếng Anh và văn bản tiếng Việt. Ưu điểm của phương pháp đồ thị tri thức là việc biểu diễn ngữ cảnh, liên hệ các khái niệm có trong văn bản được xét để có thể so sánh hai văn bản một cách tường tận.

Phần còn lại trong bài báo này được trình bày như sau. Mục II sẽ giới thiệu các nghiên cứu liên quan đối với bài toán phát hiện tương đồng ngữ nghĩa văn bản xuyên ngữ. Chúng tôi sẽ giới thiệu phương pháp đề xuất của chúng tôi ở mục III, cũng như trình bày về kết quả đánh giá ở mục IV. Và cuối cùng, trong mục V, chúng tôi sẽ trình bày phần kết luận và hướng phát triển trong tương lai.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Nghiên cứu của Potthast [1] đã phân loại các phương so sánh độ tương đồng ngữ nghĩa xuyên ngữ theo năm mô hình như trong Bảng 1.

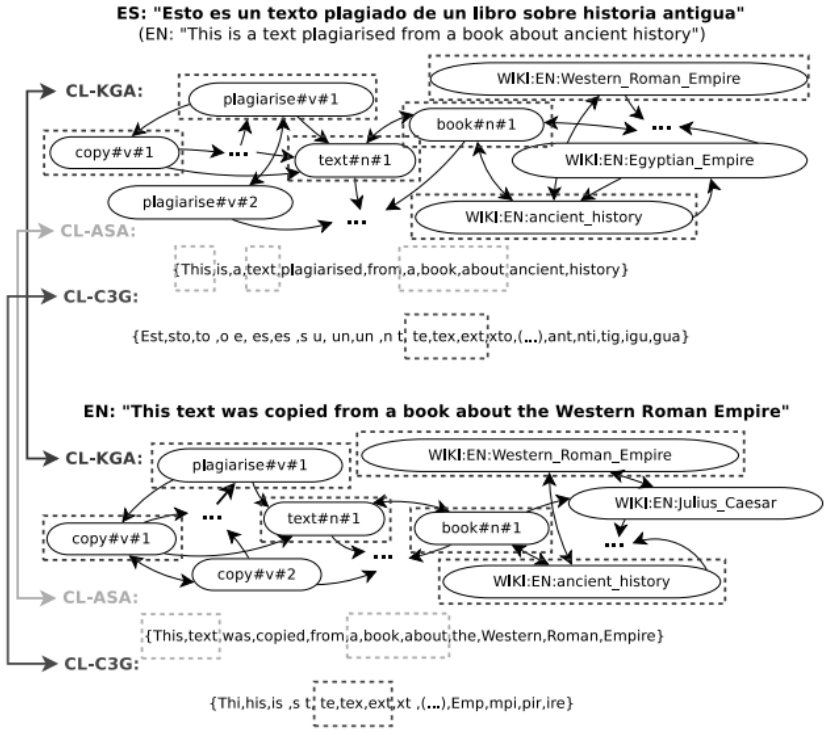
Nhóm mô hình dựa trên cấu trúc với phương pháp có phương pháp CL-CNG [2] làm đại diện. Ý tưởng chính của phương pháp này là so sánh các cặp câu sử dụng các n-gram được trích xuất từ các từ liên tiếp nhau trong câu. Phương pháp này không đạt hiệu quả cao trên các cặp ngôn ngữ khác cấu trúc cú pháp hoặc không cùng nhóm ngôn ngữ, nên không áp dụng được hiệu quả cho cặp ngôn ngữ Anh - Việt. Bên cạnh đó, phương pháp CL-CTS [3] đại diện cho nhóm mô hình dựa trên tự điển, có ý tưởng chính là biểu diễn văn bản dưới dạng véctơ khái niệm và tiến hành so sánh độ tương đồng hai văn bản dựa trên hai véctơ của chúng.

Với nhóm mô hình dựa trên kho ngữ liệu song song, phương pháp CL-ASA [4] được phát triển dựa trên công nghệ dịch máy thống kê. Với hai văn bản d và d' thuộc hai ngôn ngữ khác nhau L và L', phương pháp tính toán xác suất mỗi từ ở d là bản dịch của mỗi từ ở d' dựa trên cặp kho ngữ liệu song song thuộc hai ngôn ngữ L và L'. Từ xác suất các cặp từ là bản dịch của nhau, tính toán xác suất hai văn bản d và d' là bản dịch của nhau. Mô hình này phụ thuộc nhiều vào chất lượng kho ngữ liệu và mô hình Length và chỉ hiệu quả cao với cặp câu được dịch bởi các chuyên gia hay dịch tự động.

Bảng 1. Các mô hình so sánh độ tương đồng ngữ nghĩa xuyên ngữ

Tên nhóm mô hình	Phương pháp đại diện
Mô hình dựa trên cấu trúc (Syntax-based model)	Phương pháp CL-CNG (McNamee và Mayfield, 2004)
Mô hình dựa trên tự điển (Dictionary-based model)	Phương pháp CL-CTS (Gupta, 2012)
Mô hình dựa trên kho ngữ liệu song song (Parallel corpus- based model)	Phương pháp CL-ASA (Pinto, 2009)
Mô hình dựa trên kho ngữ liệu có thể so sánh (Comparable corpus- based model)	Phương pháp CL-KGA (M. Franco-Salvador, 2015)
Mô hình dựa trên dịch tự động (Machine translation- based model)	Phương pháp dịch và phân tích đơn ngữ (Barrón-Cedeno, 2012)

Phương pháp CL-KGA [5] đại diện cho nhóm mô hình dựa trên kho ngữ liệu có thể so sánh, được thực hiện dựa trên việc xây dựng đồ thị tri thức cho từng văn bản trên nền tảng mạng BabelNet [6] (một từ điển bách khoa toàn thư đa ngôn ngữ, được tài trợ bởi Hội đồng Nghiên cứu Châu Âu (ERC)) và so sánh các đồ thị tri thức này với nhau.



Hình 1. Ví dụ về khả năng phát hiện tương đồng của phương pháp CL-KGA trong tương quan với các phương pháp CL-ASA và CL-CNG [7]

Phương pháp này có độ chính xác cao hơn các phương pháp khác như CL-CNG, CL-ASA, CL-ESA [8] (so sánh trên kho ngữ liệu PAN-11, cặp ngôn ngữ Tây Ban Nha - Anh), tuy nhiên việc xây dựng đồ thị tốn nhiều thời gian. Do phương pháp này tiến hành chuyển các văn bản về dạng đồ thị tri thức để so sánh, nên mô hình này hoàn toàn có thể được áp dụng trên cặp ngôn ngữ Anh - Việt.

Cuối cùng là nhóm mô hình dựa trên dịch tự động với Phương pháp dịch và phân tích đơn ngữ (T+MA) [9] làm đại diện. Ý tưởng của phương pháp này là dịch các văn bản trên các ngôn ngữ khác nhau về môt ngôn ngữ chung sử dụng Google Translate [10] hoặc thay thế từng từ bằng những từ gần như là bản dịch [11], sau đó tiến hành so sánh các bản dịch của các văn bản trên ngôn ngữ chung đó. Nghiên cứu của Barron-Cedeno [12] và Muhr [11] cũng khuyến

nghị nên sử dụng phương pháp túi của từ (bag of words) trong giai đoạn so sánh. Độ chính xác của phương pháp này phụ thuộc nhiều vào độ chính xác của công cụ dịch tự động được sử dụng.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Ý tưởng chính của phương pháp đề xuất là sử dụng mô hình so sánh tương đồng ngữ nghĩa xuyên ngữ văn bản dựa trên đồ thị tri thức CL-KGA, đồng thời áp dụng các giải thuật cải tiến như điều chỉnh gán nhãn từ loại giữa văn bản tiếng Việt và văn bản tiếng Anh, bổ sung danh sách từ tiếng Việt đồng nghĩa, cũng như kết hợp các lớp đồ thị khác nhau.

Đồ thị tri thức được áp dụng cho văn bản là mô hình biểu diễn tri thức của văn bản dưới dạng đồ thị. Trong đó các đỉnh là các khái niệm tương ứng với các từ trong văn bản, các cạnh là mối quan hệ giữa các đỉnh trong đồ thị. Dựa vào đó, đồ thị tri thức trình bày một cách trực quan và dễ hiểu về các khái niệm và mối liên hệ giữa chúng.

A. So sánh tương đồng ngữ nghĩa xuyên ngữ văn bản dựa trên đồ thị tri thức (CL-KGA)

Phương pháp so sánh tương đồng ngữ nghĩa xuyên ngữ văn bản dựa trên đồ thị tri thức bao gồm 02 giai đoạn:

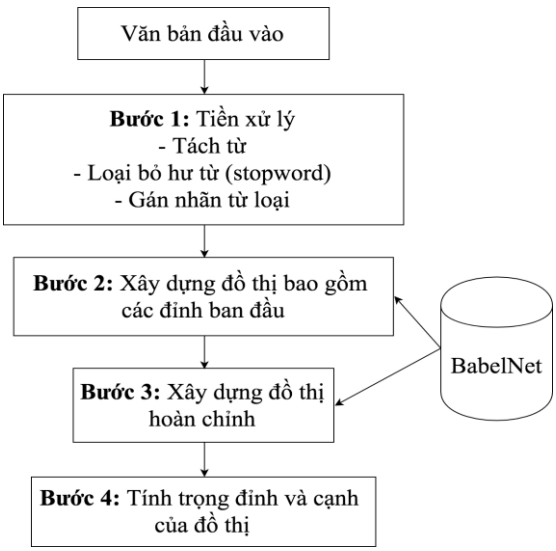
- Giai đoạn 1: Xây dựng đồ thị tri thức cho từng văn bản

- Giai đoạn 2: So sánh hai đồ thị tri thức đã xây dựng

Cụ thể cách thức thực hiện của hai giai đoạn như sau:

1. Xây dựng đồ thị tri thức cho từng văn bản:

Mỗi văn bản được xây dựng thành đồ thị tri thức bằng cách thực hiện theo bốn bước sau:



Hình 2. Các bước xây dựng đồ thị tri thức cho từng văn bản

a) Bước 1: Tiền xử lý

Trong bước tiền xử lý này, các văn bản được tách từ (đối với văn bản tiếng Việt), loại bỏ các hư từ (stopword), gán nhãn từ loại và loại bỏ các từ không được gán nhãn là Danh từ (N), Tính từ (Adj), Động từ (V) và Trạng từ (Adv).

Ví dụ như cặp câu Anh - Việt sau đây:

- Câu tiếng Anh: This is the text with plagiarism.

- Câu tiếng Việt: Đây là văn bản đạo văn.

Sau khi tiền xử lý, chúng ta sẽ thu được tập hợp các từ kèm từ loại của từng câu như sau:

- Câu tiếng Anh: text\N plagiarism\N.

- Câu tiếng Việt: văn_bản\N đạo_văn\Adj.

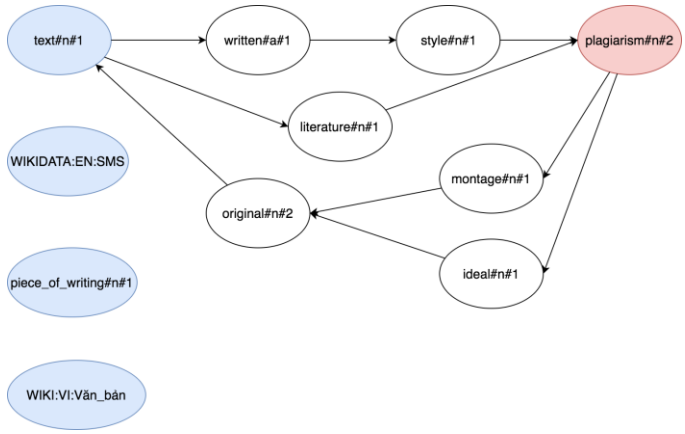
b) Bước 2: Xây dựng đồ thị bao gồm các đỉnh ban đầu

Sau khi có được danh sách các từ kèm từ loại trong văn bản đầu vào, chúng ta sẽ sử dụng BabelNet để lấy các tập các từ đồng nghĩa (synset) chứa các từ với từ loại tương ứng. Những synset ban đầu này sẽ đóng vai trò như các đỉnh ban đầu của đồ thị.

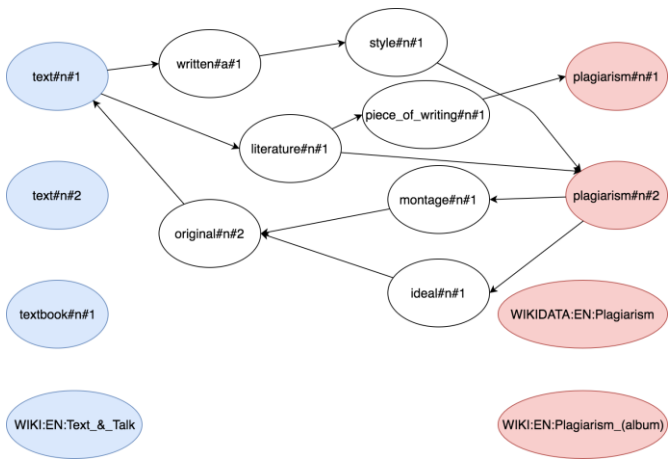
c) Bước 3: Xây dựng đồ thị hoàn chỉnh

Với các đỉnh ban đầu có được trong bước 2, chúng ta tiến hành lần lượt tìm đường nối giữa các cặp đỉnh. Trong BabelNet, hai synset có mối quan hệ ngữ nghĩa với nhau thì có cạnh nối với nhau. Sử dụng tính chất này, các cặp đỉnh được xem là có thể nối với nhau khi có thể tìm thấy đường nối giữa hai đỉnh (đường nối giữa hai synset trong BabelNet) và khoảng cách giữa hai đỉnh tối đa là 3 (tối đa có hai synset trung gian giữa hai synset được so sánh). Sau khi tìm được tất cả các đường nối giữa các đỉnh trong văn bản, chúng ta tiến hành thêm các đỉnh và cạnh trung gian vừa tìm được vào đồ thị ban đầu, thu được đồ thị hoàn chỉnh.

Ví dụ đồ thị hoàn chỉnh thu được trên câu "văn bản đạo văn" và "text plagiarism" như sau:



Hình 3. Đồ thị hoàn chỉnh được tạo ra từ câu "văn bản đạo văn"



Hình 4. Đồ thị hoàn chỉnh được tạo ra từ câu "text plagiarism"

d) Bước 4: Tính trọng đỉnh và cạnh của đồ thị

Tại bước này, chúng ta tiến hành tính trọng tất cả các đỉnh và cạnh có trong đồ thị. Trọng của đỉnh được tính bằng số lượng các cạnh nối đi ra ngoài (outdegree) từ đỉnh đó. Còn để tính trọng của cạnh có trong đồ thị, chúng ta sử dụng phương pháp biểu diễn phân tán của các khái niệm theo 4 bước như sau:

- Xây dựng các véctơ từ bằng cách sử dụng mô hình skip-gram [13].

- Tạo véctơ của các chú thích của synset, gọi là gloss véctơ. Do tính đa ngôn ngữ của các synset trong BabelNet nên chúng ta chỉ cần lấy các chú thích tiếng Anh của các synset để tạo ra các gloss véctơ. Để tạo ra các gloss véctơ, chúng ta áp dụng mô hình SenVec (Doc2Vec) [14] với đầu vào là các word véctơ đã xây dựng từ bước trước.

- Tạo véctơ của synset: do một synset có thể có nhiều chú thích nên để có thể có véctơ của một synset thì chỉ cần tính trung bình cộng của tất cả gloss véctơ mà một synset có. Sau khi áp dụng cách tạo véctơ của synset thì chúng ta có được véctơ biểu diễn cho một synset hay là một đỉnh trong đồ thị.

- Tính trọng của cạnh nối hai đỉnh trong đồ thị: áp dụng phương pháp tính so sánh cosine giữa hai véctơ đỉnh:

$$\text{Độ tương đồng (v, v')} = \frac{\vec{v}_v . \vec{v}_{v'}}{\|\vec{v}_v\| \, \|\vec{v}_{v'}\|}$$ với $\vec{v}_v$ là véctơ của đỉnh v, và $\vec{v}_{v'}$ là véctơ của đỉnh v'.

2. So sánh hai đồ thị tri thức đã xây dựng:

Sau khi đã tiến hành xây dựng đồ thị tri thức G cho văn bản tiếng Anh và G' cho văn bản tiếng Việt, tiến hành so sánh hai đồ thị tri thức đã xây dựng như sau:

Đầu tiên, tính toán độ tương đồng đỉnh giữa hai đồ thị G và G' bằng cách sử dụng phương pháp so sánh Dice coefficient:

$$S_c(G, G') = \frac{2.\sum_{c \in V(G) \cap V(G')} w(c)}{\sum_{c \in V(G)} w(c) \ + \ \sum_{c \in V(G')} w(c)}$$

Trong đó: w(c) là trọng của đỉnh c;

V(G) là tập đỉnh của đồ thị G, V(G') là tập đỉnh của đồ thị G';

V(G) ∩ V(G') là tập đỉnh chung của hai đồ thị G và G'.

Sau đó, tính toán độ tương đồng cạnh giữa hai đồ thị G và G' bằng cách sử dụng phương pháp so sánh Dice coefficient:

$$S_r(G, G') = \frac{2.\sum_{r \in E(G) \cap E(G')} w(r)}{\sum_{r \in E(G)} w(r) \ + \ \sum_{r \in E(G')} w(r)}$$

Trong đó: - w(r) là trọng của cạnh r

- E(G) là tập cạnh của đồ thị G, E(G') là tập cạnh của đồ thị G'

- E(G) ∩ E(G') là tập cạnh chung có ở cả hai đồ thị G và G'

Cuối cùng, tính toán độ tương đồng giữa hai đồ thị G và G' dựa trên độ tương đồng đỉnh S_c(G, G') và độ tương đồng cạnh S_r(G,G'): $S_g(G, G') = a.S_c(G, G') + b.S_r(G, G')$. Trong đó, a và b là các hệ số tương quan giữa các đỉnh và các cạnh, với a+b=1.

Để xác định xem văn bản tiếng Anh và văn bản tiếng Việt có tương đồng ngữ nghĩa với nhau hay không, chúng ta sử dụng ngưỡng T ∈ [0,1]. Khi đó, văn bản tiếng Anh và văn bản tiếng Việt tương đồng với nhau khi $S_g(G, G')$ > T và ngược lại thì văn bản tiếng Anh và văn bản tiếng Việt không tương đồng với nhau.

B. Phương pháp cải tiến

1. Điều chỉnh gán nhãn từ loại giữa văn bản tiếng Việt và văn bản tiếng Anh

Thực tế trong quá trình xử lý các cặp văn bản Anh - Việt cho thấy rằng, có các trường hợp từ trong câu tiếng Việt và từ trong câu tiếng Anh cùng diễn tả một ý nghĩa như nhau, tuy nhiên trong quá trình tiền xử lí sẽ có trường hợp hai từ này lại được gán nhãn từ loại khác nhau, điều này ảnh hưởng đến kết quả khi so sánh tương đồng, do tuy cùng một từ nhưng với các từ loại khác nhau, BabelNet sẽ trả ra các tập synset khác nhau ứng với mỗi từ loại. Ý tưởng chính của phương pháp là trong quá trình truy vấn synset, nếu hai từ này có thể cho ra cùng một danh sách các synset trong BabelNet thì sẽ giúp nâng cao độ chính xác của bài toán xác định tương đồng văn bản Anh - Việt. Ví dụ như từ ***khỏe_mạnh*** trong câu tiếng Việt và từ ***health*** trong câu tiếng Anh được gán nhãn từ loại khác nhau, tuy nhiên có thể dễ dàng nhận thấy hai từ này biểu đạt ý nghĩa giống nhau.

Câu tiếng Việt: *Tôi/Pp cũng/R đã/R cố_gắng/Vv để/Cm được/Vv an_toàn/Aa nhất/R có_thể/Aa ,/PU vì/Cp nó/Pp là/Vc một/Nq phần/Nn của/Cm sự/Nc* ***khỏe_mạnh/Aa*** *./PU*

Câu tiếng Anh: *I/PRP also/RB tried/VBD to/TO be/VB the/DT safest/JJS person/NN I/PRP could/MD be/VB ,/, because/IN that/DT 's/VBZ a/DT part/NN of/IN* ***health/NN*** *./.*

Để thực hiện được điều này, chúng tôi áp dụng phương pháp liên kết các từ tương ứng với nhau giữa văn bản tiếng Anh và văn bản tiếng Việt, sau đó cập nhật từ loại của các từ trong văn bản Việt theo từ loại của các từ tương ứng với chúng trong văn bản tiếng Anh. Phương pháp bao gồm các bước sau đây:

Bước 1: Tiền xử lý hai văn bản đầu vào tiếng Anh E và tiếng Việt V, được loại bỏ các từ dừng (stopword).

Bước 2: Với mỗi từ trong hai văn bản Anh và Việt, sử dụng BabelNet để truy vấn tất cả các synset có chứa từ.

Bước 3: Với mỗi cặp từ trong văn bản tiếng Anh và từ trong văn bản tiếng Việt, sử dụng độ đo Dice coefficient để tính độ tương đồng giữa 2 từ.

Bước 4: Với mỗi từ tiếng Việt, chọn từ tiếng Anh có độ tương đồng cao nhất và lớn hơn 0 để liên kết lại với nhau và nếu hai từ khác từ loại thì chúng ta sẽ thực hiện cập nhật lại từ loại của từ tiếng Việt theo từ loại của từ tiếng Anh.

Cụ thể thuật toán như sau:

SW_E = Danh sách từ trong E \ stopword

SW_V = Danh sách từ trong V \ stopword

Với mỗi từ W_E trong SW_E {

Độ tương đồng lớn nhất M = 0

Với mỗi từ W_V trong SW_V {

S_E = { các synset trong BabelNet có chứa từ W_E }

S_V = { các synset trong BabelNet có chứa từ W_V }

Độ tương đồng(W_V, W_E) = $\frac{2 * |S_V \cap S_E|}{|S_V| + |S_E|}$

Nếu Độ tương đồng(W_V, W_E) > Độ tương đồng lớn nhất M {

Độ tương đồng lớn nhất M = Độ tương đồng(W_V, W_E)

}

}

Nếu Độ tương đồng lớn nhất M > 0 {

Cập nhật lại từ loại của từ tiếng Việt theo từ loại của từ tiếng Anh

}

}

2. Bổ sung danh sách từ tiếng Việt đồng nghĩa

Số liệu tại Bảng 2 cho thấy rằng trong BabelNet, số lượng các từ loại của tiếng Việt rất ít so với số lượng từ loại của tiếng Anh. Điều này dẫn đến việc trong quá trình xử lý văn bản, có rất nhiều từ tiếng Việt không có trong BabelNet, điều này làm giảm chất lượng của thuật toán được áp dụng.

Bảng 2. Số lượng các từ loại giữa tiếng Anh và tiếng Việt trong BabelNet 3.7

Từ loại	**Tiếng Anh**	**Tiếng Việt**
Danh từ	11.638.669	3.038.992
Động từ	29.738	931
Tính từ	85.912	839
Trạng từ	14.886	371

Để nâng cao số lượng các từ tiếng Việt trong BabelNet, chúng tôi sử dụng danh sách các từ đồng nghĩa của vi.wiktionary.org và viet.wordnet.vn [15]. Để truy vấn BabelNet synset của các từ tiếng Việt không có trong BabelNet, chúng tôi thực hiện theo các bước như sau:

- Bước 1: Tìm kiếm tất cả các từ đồng nghĩa với từ tiếng Việt trong danh sách các từ đồng nghĩa.

- Bước 2: Truy vấn trong BabelNet các synset chứa các từ có được ở bước 1.

- Bước 3: Các synset thu được chính là synset của từ tiếng Việt không có trong BabelNet.

Cụ thể thuật toán cho hàm truy vấn synset từ trong BabelNet của một từ tiếng Việt như sau:

Đầu vào: từ tiếng Việt W_V

SW_V = { danh sách các từ đồng nghĩa với W_V trong danh sách các từ đồng nghĩa }

SB = {}

Với mỗi từ W'_V trong SW_V {

BabelSynset = { các synset trong BabelNet có chứa từ W'_V }

SB = SB ∩ BabelSynset

}

Đầu ra: SB là danh sách các synset trong BabelNet có chứa từ W_V

3. Kết hợp các lớp đồ thị khác nhau

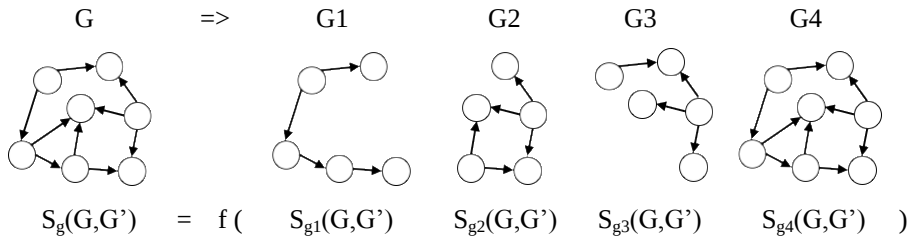
Để cải tiến mô hình phát hiện tương đồng ngữ nghĩa văn bản Anh - Việt, chúng tôi tách đồ thị tri thức thành nhiều đồ thị con, sau đó kết hợp kết quả tính toán tương đồng của các đồ thị con theo hàm số Linear Regression để có kết quả tương đồng cuối cùng. Cụ thể các bước thực hiện như sau:

- Bước 1: Đồ thị ban đầu G được tách thành bốn đồ thị khác nhau: (1) đồ thị G1 chỉ chứa danh từ và động từ, (2) đồ thị G2 chỉ chứa danh từ và tính từ, (3) đồ thị G3 chỉ chứa động từ và trạng từ, và (4) đồ thị G4 chứa tất cả các từ loại.

- Bước 2: Tính toán $S_{gi}(G,G')$ của từng đồ thị i.

- Bước 3: Tính toán $S_g(G,G') = f(S_{g1}(G,G'), S_{g2}(G,G'), S_{g3}(G,G'), S_{g4}(G,G'))$, trong đó f là hàm số Linear Regression.

- Bước 4: Để xác định xem văn bản tiếng Anh và văn bản tiếng Việt có tương đồng ngữ nghĩa với nhau hay không, chúng ta sử dụng ngưỡng $T \in [0,1]$. Khi đó, văn bản tiếng Anh và văn bản tiếng Việt tương đồng với nhau khi $S_g(G,G') > T$ và ngược lại thì văn bản tiếng Anh và văn bản tiếng Việt không tương đồng với nhau.



Hình 5. Mô hình Kết hợp các lớp đồ thị khác nhau

IV. ĐÁNH GIÁ KẾT QUẢ

A. Dữ liệu huấn luyện

Để đánh giá chất lượng của phương pháp đề xuất, chúng tôi xây dựng kho ngữ liệu gồm 1000 cặp câu Anh - Việt, trong đó 500 cặp câu tương đồng và 500 cặp câu không tương đồng. Để thực hiện được điều này, chúng tôi sử dụng kho ngữ liệu các bản dịch Anh - Việt đã được kiểm tra bằng tay, rút trích ngẫu nhiên 500 cặp câu tương đồng. Sau đó, chúng tôi ghép cặp ngẫu nhiên một câu tiếng Anh và một câu tiếng Việt, có tiến hành kiểm tra lại bằng tay để xây dựng 500 cặp câu không tương đồng. Trong 1000 cặp câu Anh - Việt đã xây dựng, chúng tôi sử dụng 900 cặp câu để huấn luyện và 100 cặp câu để đánh giá mô hình.

B. Đánh giá kết quả

Áp dụng phương pháp CL-KGA chưa cải tiến trên dữ liệu huấn luyện cho thấy, trong bảng 3, độ chính xác đạt được là 71,9%. Sau khi áp dụng các cải tiến như gán nhãn từ loại và bổ sung từ đồng nghĩa, độ chính xác đã tăng lên ở mức 76,3%.

Bảng 3. Kết quả độ chính xác các phương pháp

Phương pháp	**Độ chính xác**
Phương pháp CL-KGA chưa cải tiến	71,9%
Phương pháp CL-KGA cải tiến gán nhãn từ loại và bổ sung từ đồng nghĩa	76,2%
Phương pháp CL-KGA cải tiến gán nhãn từ loại, bổ sung từ đồng nghĩa và kết hợp các lớp đồ thị khác nhau	76,3%

Chúng tôi tiếp tục tiến hành phân lớp đồ thị thành bốn đồ thị khác nhau, đồng thời tính toán $S_{gi}(G,G')$ cho từng đồ thị i, trong đó:

- $S_{g1}(G,G')$ tương ứng với độ tương đồng trong đồ thị chỉ chứa danh từ và động từ;

- $S_{g2}(G,G')$ tương ứng với đồ thị chỉ chứa danh từ và tính từ;

- $S_{g3}(G,G')$ tương ứng với đồ thị chỉ chứa động từ và trạng từ;

- $S_{g4}(G,G')$ tương ứng với đồ thị chứa tất cả các từ loại.

Sau đó chúng tôi sử dụng phần mềm Weka [16] để tính toán các trọng số cho hàm số Linear Regression, kết quả thu được hàm số như sau:

$$S_g(G,G') = -0,9675 \times S_{g2}(G,G') + 2,4289 * S_{g4}(G,G') + 0,4033$$

Điều này cho thấy, đồ thị chỉ chứa động từ và trạng từ không có ý nghĩa với việc tính toán độ tương đồng chung giữa hai văn bản tiếng Anh và tiếng Việt. Kết quả thu được khi áp dụng phương pháp CL-KGA cải tiến gán nhãn từ loại, bổ sung từ đồng nghĩa và kết hợp các lớp đồ thị khác nhau cho thấy độ chính xác đạt được là 76,3%.

Những kết quả trên cho thấy rằng, việc áp dụng các phương pháp cải tiến như cập nhật gán nhãn từ loại, bổ sung từ đồng nghĩa và kết hợp các lớp đồ thị khác nhau đã giúp nâng cao độ chính xác của phương pháp CL-KGA. Phương pháp này rất tiềm năng để có thể kết hợp với các phương pháp học sâu trên đồ thị để tạo ra các phương pháp lai. Tuy nhiên, nghiên cứu này cũng còn hạn chế trong việc đánh giá tính chính xác của phương pháp cập nhật nhãn từ loại, cũng như tìm ra phương pháp hiệu quả để bổ sung từ tiếng Việt trong BabelNet.

V. KẾT LUẬN

Bài toán xác định tương đồng ngữ nghĩa xuyên ngữ là một trong những bài toán có vai trò rất quan trọng trong các bài toán xử lý ngôn ngữ tự nhiên khác như tìm kiếm văn bản xuyên ngữ, kiểm tra chất lượng của các mô hình dịch tự động, tóm tắt văn bản, phát hiện đạo văn,... Tuy nhiên, theo hiểu biết của cá nhân, hiện nay vẫn chưa có nhiều nghiên cứu về mô hình xác định tương đồng ngữ nghĩa xuyên ngữ Anh - Việt, đặc biệt là việc áp dụng mô hình đồ thị tri thức cho bài toán này. Trong nghiên cứu này, chúng tôi đã áp dụng phương pháp so sánh tương đồng xuyên ngữ dựa trên đồ thị tri thức, đồng thời áp dụng các phương pháp cải tiến như cập nhật gán nhãn từ loại, bổ sung từ đồng nghĩa và kết hợp các lớp đồ thị khác nhau đã giúp nâng cao độ chính xác của phương pháp so sánh tương đồng xuyên ngữ dựa trên đồ thị tri thức. Kết quả cho thấy rằng, việc áp dụng các phương pháp cải tiến đã giúp nâng cao độ chính xác của phương pháp từ 74% lên 75,9%. Việc nghiên cứu các phương pháp so sánh tương đồng dựa trên đồ thị tri thức có nhiều tiềm năng để phát triển, có thể kết hợp với các mô hình học sâu trên đồ thị để tạo ra các mô hình lai, giúp nâng cao hơn nữa độ chính xác của bài toán phát hiện tương đồng ngữ nghĩa xuyên ngữ Anh - Việt.

TÀI LIỆU THAM KHẢO

[1] Potthast, M., Barron-Cedeno, A., Stein, B., and Rosso, P. (2011). Cross-Language Plagiarism Detection. In Language Ressources and Evaluation, volume 45, pages 45–62.

[2] Mcnamee, P. and Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. In Information Retrieval Proceedings, volume 7, pages 73–97. Kluwer Academic Publishers.

[3] Gupta, P., Barron-Cedeno, A., and Rosso, P. (2012). Cross-language High Similarity Search using a Conceptual Thesaurus. In Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics, pages 67–75. Springer Berlin Heidelberg.

[4] Pinto, D., Civera, J., Juan, A., Rosso, P., and Barron-Cedéno, A. (2009). A Statistical Approach to Crosslingual Natural Language Tasks. In CEUR Workshop Proceedings, volume 64 of Journal of Algorithms, pages 51–60.

[5] M. Franco-Salvador, P. Rosso, and M. Montes-y-Gómez (2015). A Systematic Study of Knowledge Graph Analysis for Cross-language Plagiarism Detection. In: Information Processing & Management, vol. 52(4), pp. 550-570.

[6] R. Navigli and S. Ponzetto (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250.

[7] Mcnamee, P. and Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. In Information Retrieval Proceedings, volume 7, pages 73–97. Kluwer Academic Publishers.

[8] Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07), pages 1606–1611.

[9] Barron-Cedéno, A. (2012). On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism. In PhD thesis, Valencia, Spain.

[10] Kent, C. K. and Salim, N., "Web Based Cross Language Plagiarism Detection," Second International Conference on Computational Intelligence, Modelling and Simulation (CIMSiM), 2010, pp. 199–204.

[11] Muhr, M., Kern, R., Zechner, M., and Granitzer, M., "External and Intrinsic Plagiarism Detection Using a Cross-Lingual Retrieval and Segmentation System," Lab Report for PAN at CLEF 2010, 2010.

[12] Barron-Cedeno, A., Rosso, P., Agirre, E., and Labaka, G., "Plagiarism Detection across Distant Language Pairs," Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10), 2010, pp. 37–45.

[13] Bojanowski, Piotr and Grave, Edouard and Joulin, Armand and Mikolov, Tomas, "Enriching Word Vectors with Subword Information," Journal of Transactions of the Association for Computational Linguistics, Vol. 5, 2017, pp. 135-146.

[14] Quoc Le, Tomas Mikolov, "Distributed Representations of Sentences and Documents," Proceedings of the 31 st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

[15] https://github.com/zeloru/vietnamese-wordnet

[16] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

CROSS-LINGUAL SEMANTIC SIMILARITY DETECTION BETWEEN VIETNAMESE AND ENGLISH TEXTS USING THE KNOWLEDGE GRAPH

Le Thanh Nguyen, Tran Gia Trong Nhan, Tran Cong Hau, Dinh Dien

SUMMARY: *The textual semantic similarity detection task is one of the problems which play a very important role, affects the quality of many Natural Language Processing problems such as information query, text summary, plagiarism detection, etc. Especially in nowaday world, with the development of machine translation tools, the task of detecting textual semantic similarity need to consider the cross-lingual case also. In this paper, we will propose a method that uses the knowledge graph model to detect cross-lingual semantic similarity between English-Vietnamese texts. Besides, we also propose additional methods such as adjusting part of speech tag between Vietnamese text and English text, adding list of Vietnamese synonyms, combining different classes of graphs. The result shows that using above mentioned methods help to increase the accuracy of the model from 71.9% to 76.3%.*

Keywords: *similar, cross-language, graph, Vietnamese.*