Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00077

NÂNG CAO HIỆU QUẢ ĐO ĐỘ TƯƠNG TỰ NGỮ NGHĨA DỰA TRÊN MẠNG TỪ

Bùi Văn Tân[1], Nguyễn Phương Thái[2], Nguyễn Minh Thuận[2]

[1]Trường Đại học Kinh tế Kỹ thuật Công nghiệp

[2]Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

bvtan@uneti.edu.vn, thainp@vnu.edu.vn, athuan255@gmail.com

TÓM TẮT: *Đo lường độ tương tự ngữ nghĩa giữa các từ là một bài toán nghiên cứu cốt lõi và có nhiều ứng dụng trong xử lý ngôn ngữ tự nhiên. Hiện nay, kết quả của các kỹ thuật word similarity theo hướng tiếp cận Mạng từ được đánh giá trên bộ dữ liệu chuẩn là khá thấp. Trong bài viết này, chúng tôi trình bày một số kỹ thuật đo độ tương tự của từ theo tiếp cận Mạng từ, qua đó đề xuất một lược đồ làm tăng hiệu quả của các kỹ thuật này dựa trên việc biểu diễn các mối quan hệ của từ bằng cấu trúc đồ thị. Cuối cùng, chúng tôi trình bày kết quả thực nghiệm và đánh giá hiệu quả của lược đồ cải tiến.*

Từ khóa*: Xử lý ngôn ngữ tự nhiên, độ tương tự ngữ nghĩa, Mạng từ, đồ thị.*

I. GIỚI THIỆU

Sự tương đồng về ngữ nghĩa (semantic similarity) đóng vai trò trung tâm trong cách thức con người xử lý tri thức và là tiêu chí để phân loại các đối tượng, xây dựng các khái niệm, biểu diễn sự tổng quát và trừu tượng. Do đó, semantic similarity đóng vai trò then chốt trong nhiều tác vụ xử lý ngôn ngữ tự nhiên như truy vấn thông tin (infomation retrieval); mô hình ngôn ngữ (language modeling); phân cụm văn bản (document clustering); phát hiện kế thừa văn bản (recognizing textual entailment)… Đo lường độ tương tự ngữ nghĩa một cách hiệu quả là một thách thức cốt lõi trong xử lý các tài liệu văn bản phi cấu trúc của lĩnh vực xử lý dữ liệu lớn (Big Data).

Các kỹ thuật tương tự ngữ nghĩa lượng giá mức độ giống nhau của hai từ (word similarity), hay định lượng khoảng cách nhận thức giữa hai khái niệm với sự quan tâm về loại của chúng (ví dụ, từ '*trâu*' sẽ rất tương tự với từ '*bò*' bởi vì cả hai đều là gia súc ăn cỏ được con người nuôi dưỡng) hoặc chức năng của chúng (ví dụ, từ '*xe máy*' sẽ có độ tương tự lớn với từ '*xe đạp*' vì cả hai đều là phương tiện mà con người dùng để di chuyển). Ngược lại, các kỹ thuật đo mức độ liên quan của các từ thường tập trung vào mối quan hệ về loại của chúng, ví dụ từ "*ô tô"* có liên quan đến từ "*xăng"* nhưng chúng không tương tự với nhau về nghĩa, bởi vì giữa "*ô tô*" và "*xăng*" không chia sẻ một kiểu hay chức năng chung, tuy nhiên giữa chúng có mối quan hệ chung, "*xăng*" là nhiên liệu được dùng cho "*ô tô*". Khái niệm tương tự (similarity) và liên quan (relatedness) không loại trừ, độc lập với nhau cũng như không được phân biệt rõ ràng bởi con người và thuật toán.

Một số kỹ thuật được đề xuất cho bài toán word similarity với các hướng tiếp cận khác nhau: thứ nhất, dựa trên Cơ sở tri thức (Knowledge-based). Hướng tiếp cận này phát triển vào đầu thập niên 80, khai thác tri thức tự động từ các từ điển điện tử (Machine – Readable Dictionaries) như các từ điển đồng nghĩa. Kết quả của hướng tiếp cận này là sự ra đời của Mạng từ (WordNet) – một cơ sở tri thức lớn về ngữ nghĩa theo hướng liệt kê nét nghĩa; thứ hai, dựa trên kho ngữ liệu (Corpus-based). Hướng tiếp cận này sẽ rút ra các qui luật xử lý ngữ nghĩa (bằng thống kê, bằng học máy,…) từ những kho dữ liệu lớn đã có sẵn và áp dụng các luật này cho trường hợp mới. Hiện nay, cách tiếp cận dựa trên ngữ liệu kết hợp với tri thức có sẵn đang được nhiều nhà nghiên cứu quan tâm.

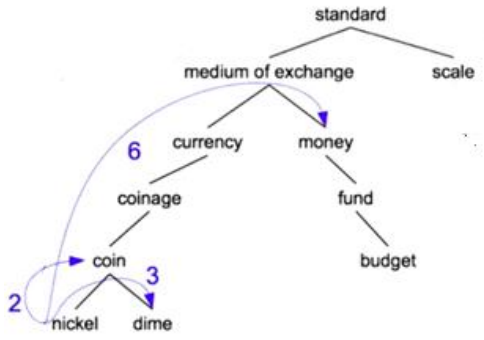
Nội dung tiếp theo của bài viết này, chúng tôi trình bày: phần II, thực nghiệm một số kỹ thuật word similarity với bộ dữ liệu chuẩn SimLex-999; phần III, đề xuất lược đồ tăng hiệu quả một số kỹ thuật word similarity dựa trên Mạng từ và cuối cùng là phần phân tích, kết luận.

II. MỘT SỐ KỸ THUẬT WORD SIMILARITY

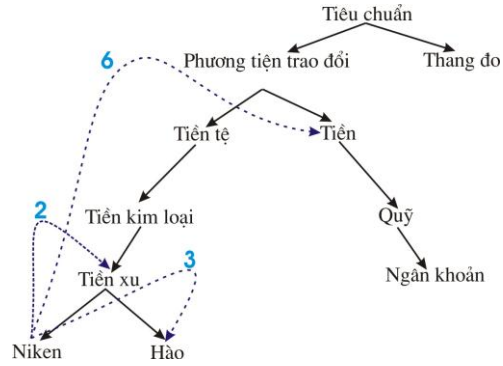
A. Hướng tiếp cận dựa trên Mạng từ

Đối với tiếng Anh, từ năm 1978, George Miller (Fellbaum, 1998) bắt đầu nghiên cứu phát triển một cơ sở dữ liệu về từ và quan hệ ngữ nghĩa giữa chúng. Cơ sở dữ liệu này được gọi là Mạng từ (WordNet) và có thể được coi là một mô hình của từ vựng tinh thần (mental lexicon). Có thể hình dung Mạng từ là một đồ thị rời rạc khổng lồ trong đó mỗi nút là một loạt đồng nghĩa (synset) và mỗi cạnh là một quan hệ ngữ nghĩa giữa các loạt đồng nghĩa [16]. Hiện nay, Mạng từ tiếng Việt (Hình 2) đã được xây dựng và ứng dụng trong một số nghiên cứu về xử lý ngôn ngữ tự nhiên. Các kỹ thuật đo độ tương tự dựa trên Mạng từ thường dựa vào cấu trúc cây phân loại của các khái niệm (Hình 1, 2).

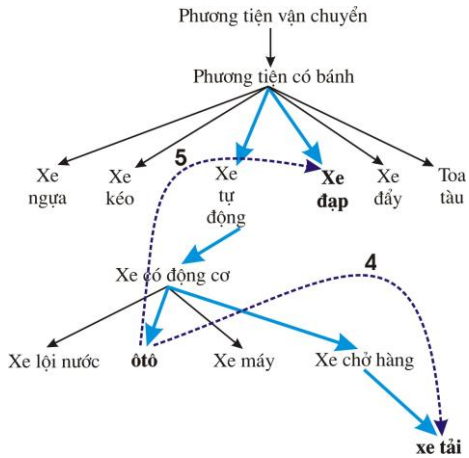
Bằng trực giác có thể thấy hai khái niệm tương tự nhau hơn nếu chúng gần nhau hơn trên cây phân loại so với các khái niệm khác (Hình 3). Do đó có thể đo độ tương tự bằng cách đếm số nút hoặc số cạnh (shortest path approach) giữa các khái niệm trên cây phân loại (công thức 1). Quan sát từ "*ôtô*", "*xe tải*" và "*xe đạp*" trên Hình 3, có thể thấy rõ khoảng cách từ "*ôtô*" đến "*xe tải*", gần hơn so với từ "*ôtô*" đến "*xe đạp*".



Hình 1. Một phần cây phân loại trong WordNet tiếng Anh



Hình 2. Một phần cây phân loại trong WordNet tiếng Việt



Hình 3. Một phần cây phân loại trong WordNet tiếng Việt

$$sim_{Path}(w_1, w_2) = \frac{1}{pathlen(w_1, w_2)} \quad (1)$$

Theo công thức 1, Sim_{Path}('*ôtô*', '*xe đạp*') = $\frac{1}{5} = 0.2$, Sim_{Path}('*ôtô*', '*xe tải*') = $\frac{1}{4} = 0.25$.

Tuy nhiên, hướng tiếp cận này có nhược điểm là đã bỏ qua tính không đồng đều giữa các quan hệ trên cây phân loại (ví dụ, khoảng cách ngữ nghĩa giữa từ "*động vật*" đến từ '*sinh vật*' lớn hơn đáng kể so với khoảng cách từ "*chó sói*" đến từ "*chó*"). Leacock và Chodorow [17] đã khắc phục nhược điểm này bằng cách thiết lập độ dài đường dẫn theo độ sâu tối đa trên cây loại (công thức 2).

$$Sim_{LCh}(w_1, w_2) = -\log \frac{\min\_length(w_1, w_2)}{2 * Depth_{\max}} \quad (2)$$

Trong đó *min_length*(w_1,w_2) là độ dài của đường đi ngắn nhất giữa hai khái niệm; *Depth* là độ sâu lớn nhất của hệ thống cây phân loại.

Theo cách tiếp cận tương tự, Wu và Palmer [8]: độ tương tự được đo bởi độ sâu của hai khái niệm trong Mạng từ và độ sâu của nút cha chung gần nhất (Least common subsumer - LCS) của cả hai khái niệm đó (công thức 3).

$$Sim_{Wup}(w_1, w_2) = 2 * \frac{depth(LCS(w_1, w_2))}{depth(w_1) + depth(w_2)} \quad (3)$$

Hướng tiếp cận thuần túy dựa trên cấu trúc cây phân loại sẽ gặp khó khăn khi khoảng cách ngữ nghĩa của các quan hệ không đồng đều. Do đó, một số kỹ thuật sử dụng thêm khái niệm độ đo nội dung thông tin (information content – IC). Nội dung thông tin của một từ là xác suất gặp phải từ này trong một kho ngữ liệu lớn, những từ thường xuyên xuất hiện sẽ có lượng thông tin nhỏ hơn những từ ít xuất hiện (công thức 4).

$$IC(c) = -\log P(c) \quad (4)$$

Theo Resnik [9], mức tương tự nhau của hai từ có thể được đánh giá bằng mức độ chia sẻ thông tin giữa chúng. Resnik định nghĩa độ tương tự giữa hai từ là "*hàm lượng thông tin*" của cha chung gần nhất của chúng.

$$Sim_{\text{Res}}(w_1, w_2) = -\log P(LCS(w_1, w_2)) = IC(LCS(w_1, w_2)) \quad (5)$$

Các nghiên cứu sau đó của Lin[10], Jiang và Conrath [11] đã mở rộng ý tưởng này và chuẩn hóa độ đo tương tự.

$$Sim_{Lin}(w_1, w_2) = 2 * \frac{IC(LCS(w_1, w_2))}{IC(w_1 + w_2)} \quad (6)$$

$$Sim_{JC}(w_1, w_2) = \frac{1}{IC(w_1) + IC(w_2) - 2 * IC(LCS(w_1, w_2))} \quad (7)$$

B. Hướng tiếp cận dựa trên nhúng từ (Word embeddings)

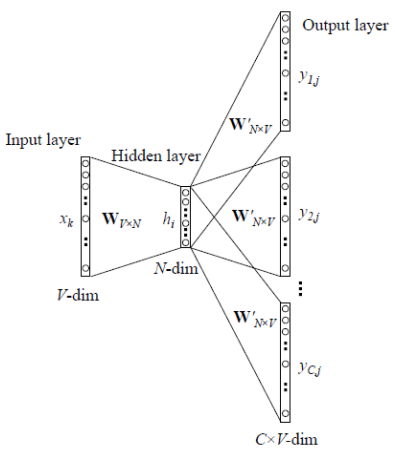
Gần đây một số kỹ thuật nhúng từ lấy cảm hứng từ mô hình ngôn ngữ dựa trên mạng nơron nhân tạo (Neural Network Language Models). Các mô hình ngôn ngữ mạng nơron sẽ chuẩn đoán các từ ngữ cảnh dựa trên từ được cung cấp. Về trực giác, những từ có nghĩa tương tự nhau thường xuất hiện gần nhau trong văn bản. Các mô hình mạng nơron học các nhúng từ bắt đầu bằng việc khởi tạo các vector biểu diễn các từ một cách ngẫu nhiên, sau đó lặp đi lặp lại việc luyện mạng, tạo cho vector của từ nhúng gần với vector biểu diễn các từ lân cận và khác các vector biểu diễn các từ mà không xuất hiện ở lân cận. Tiêu biểu nhất trong số các kỹ thuật này được cho là word2vec do T. Mikolov và các cộng sự đề xuất [12]. Cũng giống như các mô hình ngôn ngữ mạng nơron, mô hình word2vec học các nhúng từ bằng cách huấn luyện mạng nơron để dự đoán các từ lân cận, với hai kiến trúc Skip-gram và Continuous bag of words (CBOW). Trong đó, kiến trúc Skip-gram (Hình 4) dự đoán (predict) các từ lân cận trong một cửa sổ ngữ cảnh (context window) bằng cách cực đại hóa trung bình logarit của các xác suất có điều kiện (công thức 9).

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=-c}^{c}\log p(w_{t+i} \mid w_t) \quad (9)$$

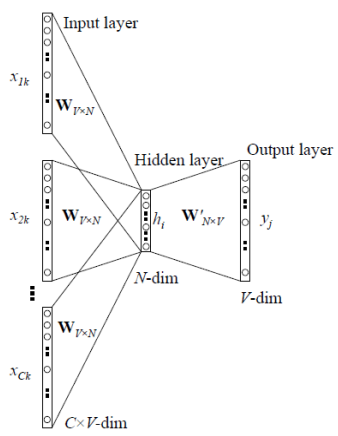
Trong đó $\{w_i : i \in T\}$ là toàn bộ tập huấn luyện, w_t là từ trung tâm và w_{t+j} là các từ trong cửa sổ ngữ cảnh. Xác suất có điều kiện được định nghĩa bằng hàm softmax (công thức 10).

$$p(w_j \mid w_I) = \frac{\exp({v'_{w_O}}^T v_{wI})}{\sum \exp({v'_{w'_j}}^T v_{w_I})} \quad (10)$$

Trong đó, v_w và v'_w là vector biểu diễn của từ w, v_w là một hàng của ma trận trọng số W giữa lớp đầu vào (input) và lớp ẩn (hidden), v'_w là một cột của ma trận trọng số W' giữa lớp ẩn và lớp ra (output) của mạng. Ta gọi v_w là vector đầu vào (input vector) và v'_w là vector đầu ra (output vector) của từ w.



Hình 4. Kiến trúc Skip-gram



Hình 5. Kiến trúc Continuous bag of words

Một trong những ưu điểm lớn nhất của kỹ thuật word2vec là chỉ cần huấn luyện với ngữ liệu thô. Khi sử dụng kho ngữ liệu lớn, tập từ vựng khá đầy đủ, có thể tính được độ tương tự của một cặp từ bất kỳ. Bên cạnh đó, các vector

biểu diễn từ được tạo ra sau khi huấn luyện, ngoài khả năng đo được độ tương tự ngữ nghĩa còn có thể được sử dụng trong nhiều tác vụ xử lý ngôn ngữ khác. Nhược điểm của kỹ thuật này là không phân biệt rõ tính tương tự và tính liên quan của cặp từ.

C. Thực nghiệm

Trong phần này, chúng tôi trình bày kết quả thực nghiệm kỹ thuật Wu & Palmer (WuP), Shortest Path (Path), Leacock-Chodorow (LCh), Resnik (Res), Jiang-Conrath (JC), Lin với bộ dữ liệu SimLex-999[1] (SimLex) và trích dẫn kết quả thực nghiệm word2vec của Christoph Lofi [7]. Bộ dữ liệu SimLex gồm 999 cặp từ, trong đó 111 cặp tính từ, 666 cặp danh từ và 222 cặp động từ. Kết quả thực hiện với các kỹ thuật khác nhau được chuẩn hóa về thang đo [0→10]. Kết quả đo độ tương tự được trình bày trong bảng 1 là tương quan Pearson giữa độ tương tự được đánh giá bởi con người và kết quả đo được bằng một kỹ thuật nhất định. Kết quả thực nghiệm của chúng tôi[2] hoàn toàn trùng khớp với kết quả được báo cáo bởi Christoph Lofi năm 2016 [7]. Để cài đặt chương trình thực nghiệm, chúng tôi sử dụng NLTK toolkit [5].

Bảng 1. Kết quả thực nghiệm một số kỹ thuật word similarity

Method	*SimLex*
Word2vec	0.44
Wu & Palmer	0.32
Shortest Path	0.45
Leacock-Chodorow	0.29
Resnik	0.35
Jiang-Conrath	0.20
Lin	0.39

Mạng từ chứa cấu trúc quan hệ ở mức nghĩa của từ (sense) đã được các chuyên gia ngôn ngữ xây dựng một cách công phu. Một trong những ưu điểm các kỹ thuật đo độ tương tự ngữ nghĩa của từ dựa trên Mạng từ (WordNet-based) là có thể đo độ tương tự ở mức sense. Để cải thiện kết quả một số ứng dụng, một số nghiên cứu thực hiện đo độ tương tự ở mức sense (sense similarity). Qua thực nghiệm, nhóm kỹ thuật này cho kết quả chính xác nhất đối với các cặp từ có mối quan hệ ngữ nghĩa rõ ràng như quan hệ đồng nghĩa (synonymy) trái nghĩa (antonymy), bao thuộc (hyponymy), tổng phân (meronymy), kéo theo (entailment), hơn nữa các kết quả đo được là khá tường minh và có thể lý giải với cấu trúc cây phân loại của Mạng từ. Chúng tôi cho rằng nếu sử dụng các kỹ thuật WordNet-based để đo độ tương tự giữa các sense sẽ cho kết quả rất sát với đánh giá của con người. Bên cạnh đó, nhóm kỹ thuật WordNet-based có ba hạn chế chính: thứ nhất, không thực hiện được phép đo với các cặp tính từ; thứ hai, số lượng từ của Mạng từ hạn chế dẫn đến không thể đo được độ tương tự của một số cặp từ; thứ ba, dữ liệu Mạng từ không được cập nhật thường xuyên, do đó kết quả đo được là "*tĩnh*" so với tính "*động*" của ngôn ngữ (sự thay đổi ngữ nghĩa của ngôn ngữ theo thời gian). Trong nhóm kỹ thuật này, kỹ thuật Shortest Path đạt kết quả tốt nhất đối với bộ dữ liệu SimLex.

III. LƯỢC ĐỒ WORD SIMILARITY CẢI TIẾN

A. Ý tưởng

Qua thực nghiệm các kỹ thuật đo độ tương tự ngữ nghĩa của từ dựa trên Mạng từ với bộ dữ liệu SimLex, chúng tôi thấy rằng, các kỹ thuật này cho kết quả đo độ tương tự bằng 0 hoặc kết quả thấp hơn so với đánh của con người ở nhiều cặp từ. Đặt Δ là hiệu của giá trị tương tự được đo bởi một kỹ thuật với giá trị tương tự được đánh giá bởi con người trong bộ dữ liệu SimLex. Thống kê trên kết quả thực nghiệm cho thấy, số cặp cho giá trị bằng 0 là hoặc có độ lệch (Δ) lớn chiếm một tỷ lệ khá cao (bảng 2).

Bảng 2. Thống kê số cặp từ có độ tương tự bằng 0 hoặc Δ lớn trong 999 cặp từ thuộc bộ SimLex

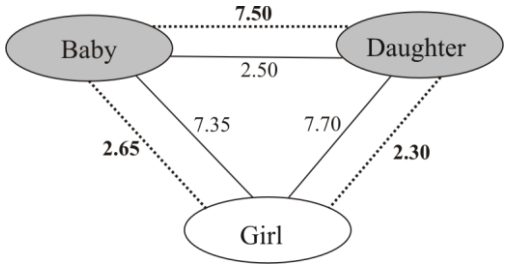
Tiêu chí thống kê	Path	WuP	LCh	Res	JC	Lin
Số cặp có độ tương tự bằng 0	82	82	100	190	102	190
Số cặp có Δ>3	281	76	75	195	653	139
Số cặp có Δ>4	171	61	61	130	527	106
Số cặp có Δ>5	93	44	55	91	405	82

Các kỹ thuật đo độ tương tự dựa trên Mạng từ thường dựa vào cấu trúc cây phân loại của các khái niệm. Trong Mạng từ, tính từ không được cấu trúc theo cây phân loại (taxonomy), do đó không thể đo được độ tương tự ngữ nghĩa. Bên cạnh đó, động từ cũng không được tổ chức tốt theo taxonomy, hiệu quả đo trên động từ là thấp. Nhìn chung, các kỹ thuật word similarity dựa trên Mạng từ thực hiện tồi với tính từ và động từ [15].

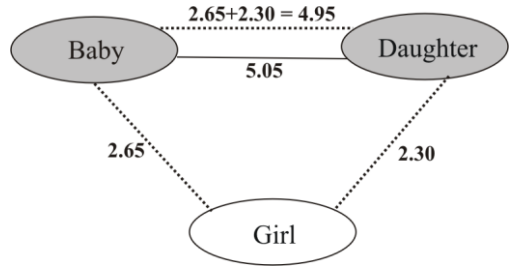
Để nâng cao hiệu quả đo độ tương tự ngữ nghĩa của các kỹ thuật word similarity dựa trên Mạng từ, chúng tôi đề xuất hai cải tiến: thứ nhất, lược đồ được áp dụng cho các kỹ thuật nhằm giảm số cặp từ cho kết quả đo có Δ lớn; thứ

[1]Available from *http://www.cl.cam.ac.uk/~fh295/simlex.html*
[2]Available from https://github.com/BuiVanTan2017/Word-Similarity

hai, quy tắc tính giá trị ngưỡng α, giá trị này được thay cho các kết quả đo bằng 0 của các cặp từ.



Hình 6. Một phần đồ thị tương tự của các từ



Hình 7. Tính độ tương tự của cặp từ

Xuất phát từ ý tưởng có thể đo độ tương tự của một cặp từ dựa vào độ tương tự của chúng với những từ khác, Hình 6, 7 trực quan hóa ý tưởng này, xác định độ tương tự của cặp từ *baby* và *daughter* thông qua từ trung gian *girl*, đường nối nét liền có trọng số trong hình vẽ biểu thị cho độ tương tự, đường nối nét đứt có trọng số biểu thị cho khoảng cách giữa hai từ. Do tính đối ngẫu giữa độ đo tương tự và độ đo khoảng cách, chúng tôi đề xuất một lược đồ áp dụng cho các kỹ thuật word similarity gồm năm bước.

Bước 1. Xác định tập các từ phổ biến *V*.

Bước 2. Từ tập *V*, xây dựng tập các cặp từ phổ biến *E*.

Bước 3. Xây dựng đồ thị vô hướng có trọng số *G(V,E)*. Trong đó mỗi đỉnh của đồ thị là một từ thuộc *V*, mỗi cạnh của đồ thị tương ứng với một cặp từ thuộc *E*. Đồ thị *G* có ma trận trọng số *W* (Hình 6), trọng số của cạnh nối hai từ *u*, *v* được tính như sau:

$$W(u,v) = 10 - simlilarity(u,v) \tag{11}$$

Bước 4. Sử dụng thuật toán Floyd-Warshall để tìm đường đi ngắn nhất giữa mọi cặp đỉnh của đồ thị *G*, tạo ra ma trận trọng số *W*’ và ma trận lưu vết đường đi *P* (Hình 7).

Bước 5. Sử dụng ma trận trong số *W*’ để tính *α*:

$$\alpha = \frac{\sum_{u,v \in V} W'(u,v)}{|V|} \tag{12}$$

Bước 6. Tính độ tương tự của các cặp từ theo kỹ thuật *m* (công thức 14). Độ tương tự của từ *u* và *v* bằng hệ số *α* nếu chúng không liên thông trên đồ thị tương tự *G*. Ngược lại, nếu *u* liên thông với *v*, độ tương tự của hai từ được đo bằng tổng của độ tương tự tính theo kỹ thuật *m* với một lượng tỷ lệ thuận với độ tương tự cực đại và tỷ lệ nghịch với số đỉnh trung gian trên đường đi tối ưu giữa *u* và *v*. Trong đó *Length(u,v)* là số đỉnh trung gian trên đường đi ngắn nhất từ *u* đến *v*.

$$Similarity(u,v) = \begin{cases} \alpha & if \;\; W'(u,v) \geq 10 \\ Similarity_m(u,v) + \dfrac{10 - W'(u,v)}{Length(u,v)} & if \;\; W'(u,v) < 10 \end{cases} \tag{14}$$

Trong đó:

$$Length(u,v) = \begin{cases} 1 & if \;\; p[u,v] = 0 \\ Length(u, p[u,v]) + Length(p[u,v], v) & if \;\; p[u,v] \neq 0 \end{cases} \tag{15}$$

B. Thực nghiệm

Chúng tôi lựa chọn hơn hai triệu cặp từ[3] được sử dụng phổ biến trong tiếng Anh để đo độ tương tự ngữ nghĩa, từ đó xây dựng đồ thị ngữ nghĩa. Để tìm đường đi tối ưu trên đồ thị, chúng tôi sử dụng thuật toán Floyd-Warshall. Thuật toán Floyd-Warshall được Robert Floyd đề xuất năm 1962 cho bài toán xác định đường đi ngắn nhất giữa các cặp đỉnh của đồ thị, thuật toán được trình bày dưới dạng giả mã như sau:

Floyd-Warshall Algorithm
Input: ma trận trọng số *W*.
Output: ma trận trọng số *W*’, ma trận lưu vết đường đi *P*.
for (k = 0; k<n; k++) {
* for (i = 0; i<n; i++){*

```
for (j = 0; j<n; j++) {
 if (W[i, j] > W[i, k] + W[k, j])
   {
    W[i, j] = W[i, k] + W[k, j];
    P[i,j] = k;
   }
```

Kết quả thực nghiệm đối với các kỹ thuật word similarity khi áp dụng lược đồ cải tiến (kỹ thuật cải tiến) không còn cặp từ có độ tương tự bằng 0, số cặp có Δ lớn giảm đi rõ rệt (bảng 4). Ví dụ, cặp từ *pretend* và *imagine* có độ tương tự là 3.3 khi đo với kỹ thuật WuP, khi áp dụng lược đồ cải tiến, chúng ta thu được độ tương tự của cặp từ này là 6.1, kết quả này gần hơn với giá trị được đánh giá bởi con người là 8.5 (bảng 3).

Bảng 3. Độ tương tự một số cặp từ được đo bởi kỹ thuật gốc so với kỹ thuật cải tiến

Cặp từ	SimLex	WuP	WuP cải tiến
dictionary - definition	6.3	1.0	7.2
communication - television	5.6	1.5	5.3
necessary - important	7.4	0.0	3.9
hard - simple	1.4	0.0	0.5
save - protect	6.6	3.3	6.3
pretend , imagine	8.5	3.3	6.1

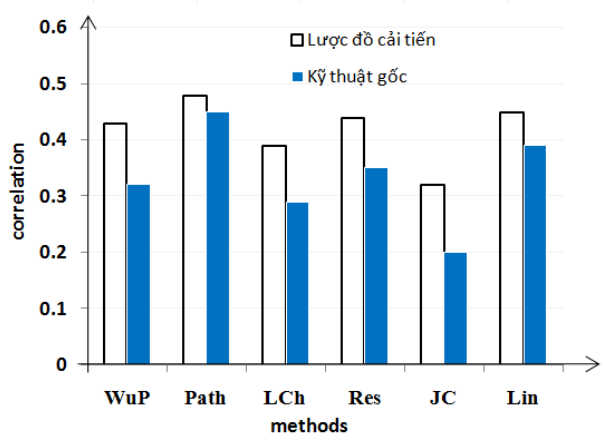
Bảng 4. Thống kê số cặp từ có độ tương tự bằng 0 hoặc Δ lớn trong kết quả thực hiện với lược đồ cải tiến

Tiêu chí thống kê	Path	Wu & Palmer	Leacock-Chodorow	Resnik	Jiang-Conrath	Lin
Số cặp có độ tương tự bằng 0	0	0	0	0	0	0
Số cặp có Δ>3	20	53	45	150	0	36
Số cặp có Δ>4	11	31	33	97	0	19
Số cặp có Δ>5	10	16	15	41	0	7

Để đánh giá hiệu quả của lược đồ cải tiến, chúng tôi thực nghiệm các kỹ thuật gốc và kỹ thuật cải tiến. Kết quả thực nghiệm các kỹ thuật gốc thu được hoàn toàn trùng khớp với kết quả do Christoph Lofi thực hiện, công bố trong [7] năm 2016. Chi tiết dữ liệu cũng như kết quả thực nghiệm được chúng tôi công bố để cộng đồng tham khảo[3].

Bảng 5. Kết quả thực nghiệm lược đồ cải tiến

Kỹ thuật	Áp dụng kỹ thuật gốc	Áp dụng lược đồ cải tiến
Wu & Palmer	0.32	0.43
Path	0.45	0.48
Leacock-Chodorow	0.29	0.39
Resnik	0.35	0.44
Jiang-Conrath	0.20	0.32
Lin	0.39	0.45



Hình 8. Kết quả thực nghiệm với lược đồ cải tiến

Bảng 5 trình bày kết quả thực nghiệm, biểu đồ trong hình 8 biểu diễn trực quan hiệu quả của lược đồ cải tiến đã đề xuất. Quan sát kết quả thực nghiệm trong bảng 5, có thể thấy các kỹ thuật đo độ tương tự sử dụng lược đồ cải tiến cho kết quả tốt hơn hẳn so với kỹ thuật gốc. Nhược điểm của kỹ thuật cải tiến là có chi phí thời gian lớn để thực hiện: đo độ tương tự của danh sách các cặp từ phổ biến với độ phức tạp thời gian $O(n^2)$; tìm đường đi ngắn nhất giữa mọi cặp đỉnh của đồ thị với độ phức tạp thời gian $O(n^3)$, với n là số từ vựng phổ biến của ngôn ngữ.

Chúng tôi thực nghiệm lược đồ cải tiến với 2609470 cặp từ được tổ hợp từ 2285 từ phổ biến trong tiếng Anh, chương trình được thực hiện trên máy tính cá nhân có cấu hình: hệ điều hành Windows 7 Untimate 32 bit; CPU Intel core 2 Duo T6600; bộ nhớ Ram 2 GigaByte. Có thể thấy rằng, chi phí thời gian để tính độ tương tự của hơn hai triệu cặp từ trong bước 3 là lớn, qua thực nghiệm chúng tôi đo được thời gian thực hiện bước này từ 195 đến 287 phút tùy theo từng kỹ thuật. Tuy nhiên bước này chỉ cần thực hiện một lần duy nhất, kết quả thực hiện được lưu lại để sử dụng

[3]Available from https://github.com/BuiVanTan2017/Word-Similarity

cho các lần tính toán sau. Chúng tôi cũng đo được thời gian thực hiện các bước 4 là 15 phút, các bước 5 và 6 có thời gian thực hiện nhỏ hơn 1 phút.

IV. KẾT LUẬN

Trong bài viết này, chúng tôi đã thực nghiệm một số kỹ thuật đo độ tương tự dựa trên Mạng từ. Đặc biệt, chúng tôi đề xuất một lược đồ nâng cao hiệu quả cho kỹ thuật word similarity dựa trên Mạng từ. Trong nghiên cứu này, do hạn chế về thời gian và phương tiện tính toán, chúng tôi mới chỉ thực nghiệm với hơn hai triệu cặp từ, chúng tôi hy vọng thực nghiệm mà chúng tôi đang tiến hành với số lượng cặp từ lớn hơn sẽ cho kết quả tốt hơn nữa. Trên cơ sở những nghiên cứu và thực nghiệm đã tiến hành, chúng tôi tiếp tục nghiên cứu sử dụng Mạng từ để đo lường độ tương tự ngữ nghĩa cho tiếng Việt.

LỜI CẢM ƠN

Bài viết này nhận được hỗ trợ bởi đề tài nghiên cứu khoa học "Xây dựng hệ thống dịch tự động hỗ trợ việc dịch các tài liệu giữa tiếng Việt và tiếng Nhật nhằm giúp các nhà quản lý và các doanh nghiệp Hà Nội tiếp cận và làm việc hiệu quả với thị trường Nhật Bản", chúng tôi biết ơn sự hỗ trợ phương tiện, tài liệu và kinh phí trong khuôn khổ đề tài nghiên cứu này. Chúng tôi cũng rất biết ơn cán bộ phản biện kín về những nhận xét hữu ích của họ, giúp chúng tôi hoàn thiện bài viết của mình.

TÀI LIỆU THAM KHẢO

[1] A. Tversky, Features of similarity, Psychol. Rev, vol. 84, no. 4, pp. 327–352, 1977.

[2] J. Camacho-Collados, M.T. Pilehvar, R. Navigli, A framework for the construction of monolingual and cross-lingual word similarity datasets, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing – Short Papers, Beijing, China, pp.1–7, 2015.

[3] Ira Leviant, Roi Reichart, Separated by an Un-common Language: Towards Judgment Language Informed Vector Space Modeling, CoRR, abs/1508.00106, 2015.

[4] Felix Hill, Roi Reichart and Anna Korhonen, SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation, Computational Linguistics, 41:665–695, 2015.

[5] Steven Bird, Edward Loper NLTK: the natural language toolkit, in Int. Conf. on Computation Linguistics (COLING), 2006.

[6] M. Lesk, Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In Proceedings of SIGDOC '86, 1986.

[7] Christoph Lofi, Measuring semantic similarity and relatedness with distributional and knowledge-based approaches, Database Society of Japan (DBSJ) Journal, vol. 14, 2016.

[8] Z. Wu and M. Palmer, Verbs semantics and lexical selection, in Annual Meeting of the Association for Computational Linguistics (ACL), 1994.

[9] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in Int. Joint Conference on AI (IJCAI), 1995.

[10] D. Lin, An information-theoretic definition of similarity, in Int. Conf. on Machine Learning (ICML), 1998.

[11] J. J. Jiang and D. W. Conrath, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, in Conference on Linguistics and Speec Processing (ROCLING), 1997.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, Distributed Representations of Words and Phrases and their Compositionality, In Advances in neural information processing systems, pages 3111–3119, 2013.

[13] Rajendra Banjade, Nabin Maharjan, Nobal B. Niraula, Vasile Rus, and Dipesh Gautam, Lemon and Tea Are Not Similar: Measuring Word-to-Word Similarity by Combining Different Methods. Computational Linguistics and Intelligent Text Processing, 2015.

[14] Nicolai Erbs, Iryna Gurevych, Torsten Zesch, Sense and Similarity: A Study of Sense-level Similarity Measures, proceedings in: Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics (*SEM 2014), pp. 30-39, 2014.

[15] Daniel Jurafsky & James H. Martin, Vector Semantics, Speech and Language Processing, 2015.

[16] Phuong-Thai Nguyen, Van-Lam Pham, Hoang-Anh Nguyen, Huy-Hien Vu, Ngoc-Anh Tran, Thi-Thu Ha Truong. A Two-Phase Approach for Building Vietnamese WordNet, the 8th Global Wordnet Conference (GWC), 2015.

[17] C. Leacock and M. Chodorow, Combining Local Context and Wordnet Similarity for Word Sense Identification, in WordNet: An Electronic Lexical Database, MIT Press, pp. 265–283, 1998.

ENHANCEMENT OF MEASUREMENT EFFICIENCY FOR SEMANTIC SIMILARITY BASED ON WORDNET

Bui Van Tan, Nguyen Phuong Thai, Nguyen Minh Thuan

ABSTRACT: *Evaluation of word similarity is a core issue because it has many applications in natural language processing. At present, efficiency of word similarity techniques based on Wordnet with benchmark datasets is low. In this paper, we present some word similarity techniques based on wordnet, thereby proposed an improved scheme that enhance performance of these techniques by representing relationships between words by graph structure. Finally, we present experimental results and evaluate effectiveness of the improvement scheme.*