Các phư ơ ng pháp đo lư ờng dựa trên việc nhúng từ Sự giống nhau về ngữ nghĩa của câu tiếng Ả Rập-tiếng Anh

El Moatez Billah Nagoudi, Jérémy Ferrero, Didier Schwab, Hadda Cherroun

Các phương pháp đo lường dựa trên việc nhúng từ Sự giống nhau về ngữ nghĩa của câu tiếng Ả Rập-tiếng Anh

El Moatez Billah Nagoudi1, Jer' emy Ferrero 2,3, Didier Schwab3 và Hadda Cherroun1

{e.nagoudi,h.cherroun}@lagh-univ.dz
{jeremy.ferrero,didier.schwab}@imag.fr
(1) Phòng thí nghiệm d'Informatique et de Mathematique LIM,
Đại học Amar Telidji, Laghouat, Algeria.
(2) Compilatio, 276 rue du Mont Blanc, 74540 Saint-Flix, Pháp.
(3) LIG-GETALP, Đại học. Grenoble Alpes, Grenoble, Pháp.

Trừu tượng. Tương tự văn bản ngữ nghĩa (STS) là một thành phần quan trọng trong nhiều ứng dụng xử lý ngôn ngữ tự nhiên (NLP) và đóng vai trò quan trọng trong lĩnh vực khác nhau như truy xuất thông tin, dịch máy, trích xuất thông tin và phát hiện đạo văn. Trong bài viết này, chúng tôi đề xuất hai phương pháp tiếp cận dựa trên việc nhúng từ nhằm đo lường sự giống nhau về ngữ nghĩa giữa các câu xuyên ngôn ngữ Ả Rập-Anh. Ý tưởng chính là khai thác Dịch máy (MT) và cách trình bày nhúng từ được cải tiến để nắm bắt
thuộc tính cú pháp và ngữ nghĩa của từ. MT được dùng để dịch tiếng Anh sang tiếng Ả Rập để áp dụng cách so sánh đơn ngữ cổ điển. Sau đó, hai phương pháp dựa trên việc nhúng từ được phát triển để xếp hạng sự tương đồng về mặt ngữ nghĩa. Ngoài ra, Căn chỉnh từ (WA), Tài liệu nghịch đảo Trọng số Tần suất (IDF) và Phần lời nói (POS) được áp dụng trên các câu được kiểm tra để hỗ trợ việc xác định các từ mang tính mô tả nhất
trong mỗi câu. Hiệu suất của các phương pháp tiếp cận của chúng tôi được đánh giá trên tập dữ liệu đa ngôn ngữ chứa hơn 2400 cặp câu tiếng Ả Rập-tiếng Anh. Hơn nữa, các phương pháp đề xuất còn được khẳng định thông qua mối tương quan Pearson giữa điểm tương đồng của chúng tôi và xếp hạng của con người.

Từ khóa: Tương đồng câu ngữ nghĩa, Đa ngôn ngữ, Ả Rập-Anh, Dịch máy, Nhúng từ.

1 Giới thiệu

Tương tự văn bản ngữ nghĩa (STS) là nhiệm vụ đo lường mức độ ngữ nghĩa sự tương đương giữa hai đơn vị văn bản (văn bản, đoạn văn hoặc câu) [1]. STS là một lĩnh vực cốt lõi của Xử lý ngôn ngữ tự nhiên (NLP) và đóng vai trò quan trọng trong một số lĩnh vực ứng dụng, chẳng hạn như Truy xuất thông tin (IR), Phân biệt nghĩa của từ (WSD), Trả lời câu hỏi (QA) và Tóm tắt văn bản (TS) cùng những thứ khác. Ở đó có hai loại STS được biết đến: đơn ngữ và đa ngôn ngữ [3]. Phương pháp đầu tiên đánh giá mức độ ngữ nghĩa cơ bản của hai đơn vị văn bản được viết bằng cùng một ngôn ngữ, tương đương với nhau, trong khi ngôn ngữ chéo STS nhằm mục đích định lượng mức độ mà hai đơn vị văn bản có liên quan về mặt ngữ nghĩa, độc lập với ngôn ngữ chúng được viết [15].

Việc xác định sự giống nhau giữa các câu đã được xem xét rộng rãi trong miền đơn ngữ [20], [4], [37] và [43]. Trong khi sự tương đồng về ngữ nghĩa giữa các ngôn ngữ tương đối khó xác định hơn do mối quan hệ liên quan của các từ được nghiên cứu giữa hai ngôn ngữ khác nhau [15]. Vì vậy, cần phải giải quyết nhiệm vụ này để nâng cao hiệu suất trong một số ứng dụng, chẳng hạn như Dịch máy (MT), Phát hiện đạo văn đa ngôn ngữ (CLPD) và Truy xuất lại thông tin đa ngôn ngữ (CLIR).

Trong bài viết này, chúng tôi tập trung điều tra vào việc đo lường sự tương đồng về ngữ nghĩa giữa các câu chéo ngôn ngữ Ả Rập-Anh bằng cách sử dụng dịch máy và biểu diễn nhúng từ. Chúng tôi cũng xem xét việc căn chỉnh các từ, tính trọng số tần suất của thuật ngữ và gắn thẻ Phần lời nói để cải thiện việc xác định các từ có tính mô tả cao trong mỗi câu.

Phần còn lại của bài viết này được tổ chức như sau, phần tiếp theo mô tả công việc liên quan đến mô hình nhúng từ và phát hiện ngôn ngữ chéo STS. Trong Phần 3, chúng tôi trình bày các phương pháp dựa trên việc nhúng từ đa ngôn ngữ được đề xuất của chúng tôi. Phần 4 mô tả kết quả thử nghiệm của các hệ thống này. Cuối cùng, kết luận của chúng tôi và một số hướng nghiên cứu tiếp theo được rút ra ở Phần 5.

2 công việc liên quan

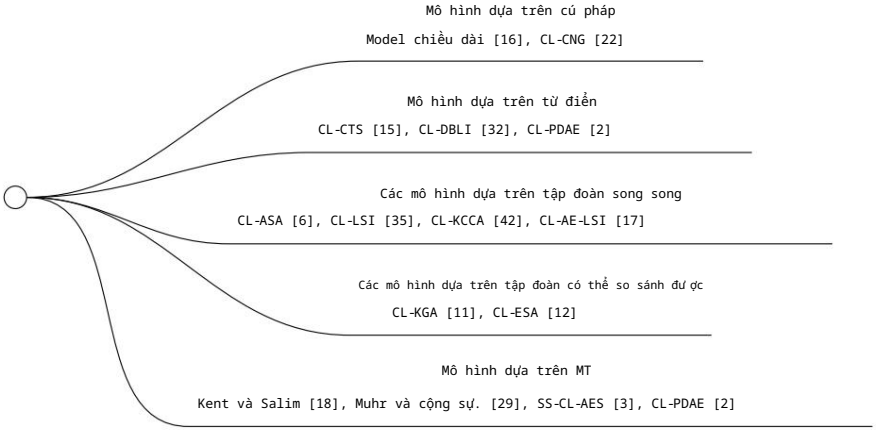
Trong phần này, chúng tôi xem xét các phương pháp phù hợp nhất để đo lường văn bản ngữ nghĩa đa ngôn ngữ. Sau đó, chúng tôi nghiên cứu những thứ dành riêng cho sự tương đồng về ngữ nghĩa giữa tiếng Ả Rập và tiếng Anh. Cuối cùng, chúng ta nhắc lại một số khái niệm liên quan đến việc nhúng từ.

2.1 Phát hiện sự tương đồng về văn bản ngữ nghĩa giữa các ngôn ngữ

Trong tài liệu, nhiều phương pháp được đề xuất để phát hiện sự giống nhau về văn bản giữa các ngôn ngữ. Chúng ta có thể phân loại chúng theo chiến lược mà chúng đã sử dụng để phát hiện sự tương đồng đó thành năm lớp: Mô hình dựa trên cú pháp, dựa trên từ điển, song song và dựa trên tập đoàn có thể so sánh và dựa trên MT [10]. Hình 1 cho thấy sự phân loại của các phương pháp khác nhau để phát hiện sự giống nhau giữa các ngôn ngữ. Sau đây chúng ta sẽ xem xét các phương pháp được sử dụng phổ biến nhất.

Liên quan đến các mô hình dựa trên cú pháp, ý tưởng chính nằm ở việc so sánh các văn bản đa ngôn ngữ không có bản dịch. Ví dụ, Pouliquen và cộng sự. [16] đã đề xuất một "Mô hình độ dài" để ước tính độ tương tự của văn bản giữa các ngôn ngữ. Nó chủ yếu dựa trên việc so sánh kích thước văn bản. Họ quan sát thấy thực tế là độ dài của văn bản trong các ngôn ngữ khác nhau được liên kết chặt chẽ bởi một yếu tố và có một yếu tố khác nhau đối với mỗi cặp ngôn ngữ.
McNamee và Mayfield [22] đã giới thiệu mô hình N-Gram ký tự đa ngôn ngữ (CL-CNG) để so sánh hai đơn vị văn bản bằng cách sử dụng biểu diễn vectơ n-gram của chúng.
Kỹ thuật này đạt được hiệu quả tốt với các ngôn ngữ gần nhau do có các từ gốc chung.

Trong các mô hình dựa trên từ điển, độ tương tự về ngữ nghĩa được đo bằng cách xây dựng mô hình không gian vectơ của các đơn vị văn bản. Vì vậy, một vectơ khái niệm được xây dựng cho từng đơn vị văn bản bằng cách sử dụng từ điển hoặc từ điển đồng nghĩa. Sự giống nhau giữa các vectơ của các khái niệm có thể được đo bằng độ tương tự Cosine, khoảng cách Euclide hoặc bất kỳ



Hình 1: Phân loại các phư ơ ng pháp tiếp cận khác nhau để phát hiện sự giống nhau giữa các ngôn ngữ [10].

thư ớc đo tư ơ ng tự khác. Trong [15], mô hình Sim-ilarity dựa trên khái niệm đa ngôn ngữ (CL-CTS) đư ợc đề xuất để đo lư ờng sự giống nhau giữa các đơ n vị văn bản đư ợc viết bằng các ngôn ngữ khác nhau (tiếng Tây Ban Nha, tiếng Anh và tiếng Đức). CL-CTS dựa trên các vectơ khái niệm từ điển đồng nghĩa đư ợc trình bày trong Eurovoc[1] trong đó độ tư ơ ng tự Cosine đư ợc tính toán giữa các vectơ này. Trong bối cảnh tư ơ ng tự, Pataki [32] đã đề xuất mô hình Độc lập ngôn ngữ dựa trên từ điển đa ngôn ngữ (CL-DBLI). CL-DBLI xem xét một từ điển dịch thuật đồng nghĩa để trích xuất các khái niệm trừu tư ợng từ các từ trong các đơ n vị văn bản

Đối với các mô hình dựa trên tập hợp có thể so sánh đư ợc, Gabrilovich và Markovitch [12] đã gửi trư ớc mô hình Phân tích ngữ nghĩa rõ ràng xuyên ngôn ngữ (CL-ESA). CL-ESA là dựa trên Phân tích ngữ nghĩa rõ ràng (ESA), thể hiện ý nghĩa của văn bản bởi một vectơ các khái niệm bắt nguồn từ Wikipedia. Trong bối cảnh đa ngôn ngữ, Potthast và cộng sự. [36] sử dụng Wikipedia làm kho ngữ liệu có thể so sánh để ư ớc tính độ giống nhau của hai tài liệu bằng cách tính toán độ giống nhau của hai cách trình bày ESA của chúng. Một mô hình khác đư ợc gọi là Phân tích sơ đồ tri thức đa ngôn ngữ (CL-KGA), đư ợc giới thiệu cho lần đầu tiên bởi Franco-Salvador et al. [11]. CL-KGA sử dụng đồ thị tri thức đư ợc xây dựng từ mạng ngữ nghĩa đa ngôn ngữ (tác giả sử dụng BabelNet [31]) để thể hiện văn bản và sau đó so sánh chúng trong một không gian đồ thị ngữ nghĩa ngôn ngữ chung.

Về các mô hình dựa trên tập đoàn song song, một số cách tiếp cận đư ợc đề xuất. Ví dụ, Barron-Cede no et al. [6] đã giới thiệu phư ơ ng pháp Phân tích tư ơ ng tự liên kết ngôn ngữ chéo (CL-ASA). CL-ASA ư ớc tính sự giống nhau giữa hai đơ n vị văn bản sử dụng từ điển thống kê song ngữ đư ợc trích từ kho ngữ liệu song song. các ý tư ởng tư ơ ng tự đã đư ợc Pinto et al sử dụng một cách độc lập. [34]. Mô hình lập chỉ mục ngữ nghĩa tiềm ẩn đa ngôn ngữ (CL-LSI) đư ợc phát triển bởi Potthast et al. [35]. CL-LSI sử dụng một ngữ liệu song song với chiến lư ợc Ngữ nghĩa tiềm ẩn phổ biến đư ợc áp dụng trong các hệ thống IR cho sự kết hợp đơ n vị thuật ngữ-văn bản. Một mô hình khác có tên Cross-Language Kernel Canonical

[1] http://eurovoc.europa.eu/

Mô hình Phân tích tư ơ ng quan (CL-KCCA) do Vinokourov et al. [42], nó phân tích
sự tư ơ ng ứng giữa hai không gian LSI để đo lư ờng mối tư ơ ng quan của các không gian tư ơ ng ứng
các giá trị chiếu.

Ý tư ởng chính của các mô hình dựa trên dịch máy bao gồm việc sử dụng các công cụ MT
dịch các đơ n vị văn bản sang cùng một ngôn ngữ (ngôn ngữ xoay) để áp dụng một
so sánh đơ n ngữ giữa chúng [5]. Với mục đích này, Kent và Salim [18] có
đã sử dụng API Google Translate để dịch văn bản, trong khi Muhr et al. [29] thay từng chữ
của văn bản gốc bằng các bản dịch có khả năng nhất của nó sang ngôn ngữ đích.

2.2 Sự tư ơ ng đồng về ngữ nghĩa giữa các ngôn ngữ Ả Rập-Anh

Trong bối cảnh có sự tư ơ ng đồng về ngữ nghĩa giữa các ngôn ngữ Ả Rập và Anh, Hattab [17] có
đã sử dụng Lập chỉ mục ngữ nghĩa tiềm ẩn (LSI) để xây dựng ngữ nghĩa tiếng Ả Rập-tiếng Anh đa ngôn ngữ
dấu cách (CL-AE-LSI), từ đó nó kiểm tra sự giống nhau về ngữ cảnh của hai văn bản nhất định,
một bằng tiếng Ả Rập và một bằng tiếng Anh.

Gần đây, Alzahrani [3] đã trình bày hai mô hình Tư ơ ng tự ngữ nghĩa cho các câu đa ngôn ngữ Ả
Rập-Anh (SS-CL-AES). Cái đầu tiên sử dụng một từ điển
bản dịch, trong đó một câu tiếng Ả Rập đư ợc dịch sang thuật ngữ tiếng Anh, thì độ tư ơ ng tự về mặt
ngữ nghĩa đư ợc tính toán bằng cách sử dụng kỹ thuật tư ơ ng tự bản dịch tối đa. TRONG
mô hình thứ hai, MT đư ợc áp dụng cho câu tiếng Ả Rập. Sau đó, các thuật toán
đư ợc đề xuất bởi Lee [19], và Liu et al. [21] đư ợc sử dụng để tính toán độ tư ơ ng tự về ngữ nghĩa.

Alaa và cộng sự. [2] quan tâm đến Phát hiện đạo văn đa ngôn ngữ của các tài liệu tiếng Ả Rập-Anh
(CL-PDAE). Trên thực tế, sau khi truy xuất tài liệu ứng viên từng bư ớc
trích xuất cụm từ khóa, họ dịch văn bản nguồn bằng cách lấy cho một từ tất cả các bản dịch có sẵn
của tất cả các từ đồng nghĩa có sẵn của nó từ WordNet [27], sau đó họ sử dụng một
sự kết hợp của các biện pháp đơ n ngữ (Dãy chung dài nhất (LCS), Cosine
tư ơ ng tự và N-Gram) để phát hiện các cụm từ tư ơ ng tự.

2.3 Mô hình dựa trên nhúng từ

Gần đây kỹ thuật Word Embedding (WE) nhận đư ợc rất nhiều sự quan tâm trong NLP
cộng đồng và đã trở thành tòa nhà cốt lõi cho nhiều ứng dụng NLP. CHÚNG TÔI đại diện
các từ dư ới dạng vectơ trong một không gian nhiều chiều liên tục. Những cách biểu diễn này cho phép
nắm bắt các thuộc tính ngữ nghĩa và cú pháp của ngôn ngữ [23]. Trong tài liệu, một số kỹ thuật đư ợc
đề xuất để xây dựng mô hình nhúng từ.

Ví dụ, Collobert và Weston [9] đã trình bày một hệ thống thống nhất dựa trên
một mạng lư ới thần kinh sâu và đư ợc đào tạo chung với nhiều nhiệm vụ NLP, chẳng hạn như : gắn thẻ
POS, Gắn nhãn vai trò ngữ nghĩa và Nhận dạng thực thể đư ợc đặt tên. Mô hình của họ đư ợc lư u trữ trong
ma trận M 2 Rd |D| , trong đó D đại diện cho từ điển của tất cả các từ duy nhất trong
dữ liệu huấn luyện và mỗi từ trong D đư ợc nhúng vào một vectơ d chiều. Các câu câu đư ợc thể hiện
bằng cách nhúng các từ tạo thành chúng. Một ý tư ởng tư ơ ng tự đã đư ợc
đư ợc đề xuất và sử dụng độc lập bởi Turian et al. [41].

Mnih và Hinton [28] đã giới thiệu một dạng khác để biểu diễn các từ trong vector
không gian, đư ợc đặt tên là Mô hình log-song tuyến phân cấp (HLBL). Giống như hầu hết các mô hình
ngôn ngữ thần kinh, mô hình HLBL đư ợc sử dụng để biểu diễn mỗi từ bằng một đặc điểm có giá trị thực.

Machine Translated by Google

vectơ . HLBL nối (n 1) các từ nhúng đầu tiên (w1..wn1) và học mô hình tuyến tính thần kinh để xác định từ cuối cùng wn.

Trong Mikolov và cộng sự. [26] mạng thần kinh tái phát (RNN) [24] được sử dụng để xây dựng mô hình ngôn ngữ thần kinh. Mô hình RNN mã hóa từng từ ngữ cảnh và dự đoán từ tiếp theo. Sau đó, trọng số của mạng được huấn luyện được coi là vectơ nhúng từ.

Dựa trên mô hình ngôn ngữ thần kinh đơn giản hóa của Bengio et al. [7], Mikolov và cộng sự. [23] [25] trình bày hai kỹ thuật khác để xây dựng mô hình biểu diễn từ. Trong nghiên cứu của họ, hai mô hình được đề xuất: mô hình túi từ liên tục (CBOW) [23] và mô hình Skip-gram (SKIP-G) [25]. Mô hình CBOW dự đoán một từ xoay vòng theo ngữ cảnh bằng cách sử dụng một cửa sổ gồm các từ theo ngữ cảnh xung quanh nó. Cho một chuỗi các từ S = w1, w2, ..., wi, mô hình CBOW học cách dự đoán tất cả các từ wk từ các từ xung quanh chúng (wkl, ..., wk1, wk+1, ..., wk+l). Mô hình thứ hai, SKIP-G, dự đoán các từ xung quanh của từ xoay hiện tại wk [25].

Pennington và cộng sự. [33] đã đề xuất mô hình Global Vectors (GloVe) để biểu diễn các từ trong không gian vectơ . Mô hình GloVe xây dựng ma trận đồng xuất hiện M bằng cách sử dụng số liệu thống kê toàn cầu về sự xuất hiện của từng từ. Sau đó, ma trận M được sử dụng để ước tính xác suất từ wi xuất hiện trong ngữ cảnh của một từ khác wj , xác suất P(i/j) này thể hiện mối quan hệ giữa các từ.

Trong một nghiên cứu so sánh được thực hiện bởi Mikolov et al. [23] tất cả các phương pháp [9], [41], [28], [26], [23] và [25] đã được đánh giá và so sánh và cho thấy CBOW [23] và SKIP-G [25] các mô hình được đào tạo nhanh hơn đáng kể với độ chính xác tốt hơn. Vì lý do này, chúng tôi đã sử dụng cách biểu diễn từ CBOW cho mô hình tiếng Ả Rập, do Zahran et al đề xuất. [45]. Để đào tạo mô hình này, họ đã sử dụng một bộ sưu tập lớn từ nhiều nguồn khác nhau với hơn 5,8 tỷ từ[2].

Trong mô hình CBOW tiếng Ả Rập [45] mỗi từ w được biểu thị bằng một vectơ v có chiều d. Sự giống nhau giữa hai từ wi và wj (ví dụ: từ đồng nghĩa, số ít, số nhiều, nữ hóa hoặc có liên quan chặt chẽ về mặt ngữ nghĩa) có được bằng cách so sánh cách biểu diễn vectơ vi và vj của chúng tương ứng [23]. Sự tương tự này có thể được đánh giá bằng cách sử dụng độ tương tự Cosine, khoảng cách Euclide, khoảng cách Manhattan hoặc bất kỳ độ tương tự nào khác È™" Am. Ã'@" (đại lường. Ví dụ: đặt " ÈJ æÀ @" (khoa) là học) từ."Z AÇ œ@"ng (buổi tối) và " được đánh bằng cách tính độ tương tự cosine giữa vectơ của chúng như sau: Sim(Z AÇ œ@, È™" Am. Ã'@) = Cos(v(Z AÇ œ@), v(È™" Am. Ã'@)) = 0,13

Sim(ÈJ æÀ @, È™" Am. Ã'@) = Cos(v(ÈJ æÀ @), v(È™" Am. Ã'@)) = 0,72

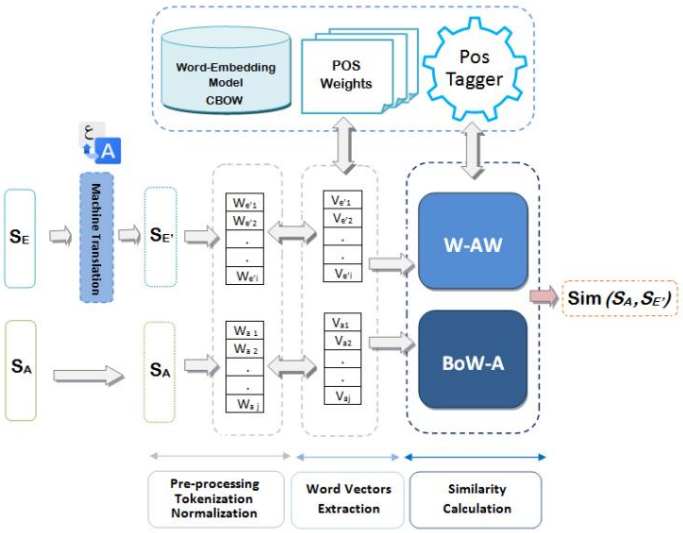
nghĩa. Điều đó có nghĩa là, các từ ÈJ æÀ @" (khoa) và " È™" Am. Ã'@" (đại học) có tính chất ngữ " gần hơn "Z AÇ œ@ " (buổi tối) và " có È™" Am. Ã'@" (đại học). Sau đây chúng tôi khai thác điều này tính chất đo lường sự tương đồng về ngữ nghĩa ở cấp độ câu.

[2] https://sites.google.com/site/mohazahran/data

3 phư ơ ng pháp đề xuất

Trong phần này, chúng tôi trình bày hai phư ơ ng pháp đư ợc đề xuất cho ngôn ngữ chéo tiếng Ả Rập-Anh sự giống nhau của câu. Các phư ơ ng pháp này sử dụng Mô hình dựa trên dịch máy, theo sau bằng cách phân tích sự tư ơ ng đồng ngữ nghĩa đơ n ngữ dựa trên việc nhúng từ. Chúng bao gồm ba bư ớc, bao gồm dịch thuật, tiền xử lý và ghi điểm tư ơ ng đồng.

Đầu tiên, MT đư ợc sử dụng để dịch các câu tiếng Anh sang tiếng Ả Rập. Sau đó, hai từ của chúng tôi phư ơ ng pháp dựa trên nhúng đư ợc sử dụng để đo lư ờng sự tư ơ ng tự về ngữ nghĩa của tiếng Ả Rập câu. Trong phần đầu tiên, chúng tôi đề xuất sử dụng kỹ thuật căn chỉnh từ đư ợc đề xuất bởi Sulatan và cộng sự. [39] với các từ phư ơ ng pháp trọng số của Nagoudi và Schwab [30], chúng tôi gọi phư ơ ng pháp này là Trọng số các từ đư ợc căn chỉnh (W-AW). Cái thứ hai tạo ra một Túi từ cho các từ đư ợc căn chỉnh để tạo thành một biểu diễn vectơ cho mỗi câu. Sau đó sự giống nhau có đư ợc bằng cách so sánh hai vectơ câu, chúng tôi đặt tên cho phư ơ ng pháp này Căn chỉnh túi từ (BoW-A). Hình 2 đư a ra cái nhìn tổng quan về các phư ơ ng pháp đư ợc đề xuất.



Hình 2: Tổng quan về các phư ơ ng pháp đề xuất

Đặt SE = we1 , we2 , ..., wei và SA = wa1 , wa2 , ..., waj là tiếng Anh và tiếng Ả Rập câu, và các vectơ từ của chúng lần lư ợt là (ve1 , ve2 , ..., vei) và (va1 , va2 , ..., vaj). Sự giống nhau về ngữ nghĩa giữa SE và SA đư ợc tính toán theo ba bư ớc: dịch thuật, tiền xử lý và phân bổ điểm tư ơ ng tự đơ n ngữ.

1) Dịch thuật: ở bư ớc này, chúng tôi đã sử dụng API Google Translate [3] để dịch các câu tiếng Anh sang tiếng Ả Rập, chúng tôi ký hiệu là câu đã dịch SE0 . Bằng bản dịch này

[3] https://cloud.google.com/translate/

vấn đề đư ợc giảm xuống thành vấn đề tư ơ ng tự về ngữ nghĩa đơ n ngôn ngữ.

2) Tiền xử lý : nhằm chuẩn hóa các câu để đánh giá độ tư ơ ng tự
bư ớc này, một tập hợp tiền xử lý đư ợc thực
hiện: – Tokenization: các câu đầu vào đư ợc chia thành các
từ; – Loại bỏ dấu chấm câu, dấu phụ và ký tự không chữ và số; – Chuẩn hóa @,
@ , như trong mô hình CBOW tiếng Ả Rập [45]; – Thay thế cuối cùng ¯ theo sau
@ đến @ và Ả Rập
là Z bằng ¯ .

Tại thời điểm này, chúng tôi nên đề cập rằng chúng tôi sẽ không loại bỏ các từ dừng vì chúng có thể
ảnh hư ởng đến điểm tư ơ ng đồng, Ví dụ: È™" Am. Ã'@ ˙Õ @ SE
đi học đại học" và SA = ˝ ="Joseph đã TÔI Î YK 'À ÉÒK " (Joseph
không học đại học). Nếu chúng ta loại bỏ các từ 'À , ˙Õ @ và to làm từ dừng, cả hai đều sen-
các thì trở nên giống nhau, trong khi chúng có ý nghĩa trái ngư ợc nhau.

3) Độ tư ơ ng tự của câu, chúng tôi đề xuất hai phư ơ ng pháp đo độ tư ơ ng tự về ngữ
nghĩa giữa SE0 và SA: Phư ơ ng pháp căn chỉnh trọng số các từ (W-AW) và Phư ơ ng
pháp căn chỉnh túi từ (BoW-A). Sau đây, chúng tôi phát triển các phư ơ ng pháp đư ợc
đề xuất và cung cấp cho mỗi phư ơ ng pháp cách đo lư ờng sự tư ơ ng đồng về ngữ nghĩa.

3.1 Phư ơ ng pháp căn chỉnh trọng số từ (W-AW)

Một cách đơ n giản để so sánh câu dịch SE0 và câu SA tiếng Ả Rập là tính tổng các
vectơ từ của chúng [30]. Sau đó, độ tư ơ ng tự thu đư ợc bằng cách tính độ tư ơ ng
tự Cosine Cos(VE0 ,VA), trong đó:

$$VE0 = \sum_{k=1} ve0_k \qquad VA = \sum_{k=1} vak$$

Ví dụ: đặt SE và SA là hai câu: SE ="Joseph đi
học đại học".
SA = ˝ È™" Aj. À A'QÂÑ" ˙Ê î fl ÉÒK " (Joseph nhanh chóng vào đại học).

Bư ớc 1: Dịch thuật Ở
bư ớc này Google Translate API đư ợc sử dụng để dịch câu tiếng Anh SE sang tiếng Ả Rập SE0 =
" ÈJ æÀ @ ˙Õ @ ÉÒK I. Î X".

Bư ớc 2: Tổng các vectơ từ VE0 = V (
ÈJ æÀ @) + V (˙Õ @) + V (ÉÒK) + V (I. Î X)

VA = V (È™" Aj. À) + V (A`QÂÑ") + V (˙Ê î fl) + V (ÉÒK)

Bư ớc 3: Điểm tư ơ ng đồng
Độ tư ơ ng tự giữa SE0 và SA có đư ợc bằng cách tính độ tư ơ ng tự cosin giữa các
vectơ câu VE0 và VA như sau:

$$Sim(SE, SA) = Sim(SE0 , SA) = Cos(VE0 , VA)=0,71$$

Để cải thiện kết quả tương tự, chúng tôi đã sử dụng phương pháp căn chỉnh các từ được trình bày bởi Sultan et al. [39], với sự khác biệt là chúng tôi căn chỉnh các từ dựa trên sự giống nhau về ngữ nghĩa của chúng trong mô hình nhúng từ chứ không phải trong từ điển. Chúng tôi cũng giả định rằng tất cả các từ không có tầm quan trọng như nhau đối với ý nghĩa của câu. Để làm được điều đó, chúng tôi sử dụng ba hàm trọng số (idf, pos và idf-pos) do Nagoudi và Schwab đề xuất trong [30] để tính trọng số cho các từ được căn chỉnh. Cuối cùng, độ tương tự giữa SE0 và SA được tính như sau:

$$Sim(SE0, SA) = \frac{1}{2}\left(\frac{\sum_{w2SE0} WT(w) \quad BM(w, SA)}{\sum_{w2SE0} WT(w)} + \frac{\sum_{w2SA} WT(w) \quad BM(w, SE0)}{\sum_{w2SA} WT(w)}\right) \quad (1)$$

trong đó WT(w) là hàm trả về trọng số của từ w. WT sử dụng ba phương pháp tính trọng số: idf, pos và kết hợp cả hai. Hàm BM(w, Sk) biểu thị điểm Phù hợp nhất giữa w và tất cả các từ trong câu Sk. Do đó, hàm BM căn chỉnh các từ dựa trên sự tương đồng về ngữ nghĩa của chúng có trong mô hình nhúng từ.
Hàm BM được định nghĩa là:

$$BM(w, Sk) = Max\{Cos(v, vk), tuần 2 Sk\} \quad (2)$$

Ví dụ, chúng ta tiếp tục với ví dụ tương tự ở trên, sự giống nhau giữa SE0 và SA được lấy theo 4 bước như sau:

Bước 1: Gắn thẻ POS
Đầu tiên, trình gắn thẻ POS của Gahbiche-Braham et al. [13] được sử dụng để dự đoán thẻ từ loại của mỗi từ trong Sk .

$$Pos\_tag(SA) = \text{danh từ hỗ trợ động từ tính từ}$$
$$Pos\_tag(SE0) = \text{danh từ động từ trợ động từ tính từ}$$

Bước 2: Trọng số IDF & POS Để tính
trọng số cho các từ được căn chỉnh mang tính mô tả, chúng tôi truy xuất cho mỗi từ wk trong Sk trọng số IDF của nó idf(wk), chúng tôi cũng sử dụng trọng số POS được đề xuất trong [30].

Bước 3: Căn chỉnh từ Ở bước này,
chúng ta căn chỉnh các từ có nghĩa giống nhau trong cả hai câu. Để làm được điều đó, chúng tôi tính toán độ tương tự giữa mỗi từ trong SE0 và từ gần nhất về mặt ngữ nghĩa trong SA bằng cách sử dụng hàm BM, ví dụ BM(˙ Ê î fl , SE0)
= Max{Cos(˙ Ê î fl ,vk), tuần 2 SA}
= Cos(v(˙ Ê î fl), v(I. Î X)) = 0,85

Bước 4: Tính độ tương tự Độ tương tự
giữa SE0 và SA có được bằng cách sử dụng công thức (1), cho ta: Sim(SE0 , SA)=0,82.

3.2 Phương pháp căn chỉnh túi từ (BoW-A)

Một trong những ưu điểm của việc nhúng từ là nó cho phép truy xuất danh sách các từ được sử dụng trong cùng ngữ cảnh đối với một từ nhất định w [14]. Chúng tôi đặt tên này

danh sách các từ k-gần nhất với w. Ví dụ: Bảng 1 hiển thị 10 từ gần nhất của È™" Am. Ã'@ và ÈJ æÀ @ trong mô hình CBOW tiếng Ả Rập.

Cây cung(È™" Am. Ã'@)	BoW(cÀ @ ÈJ æÀ @)
È™" Am. Ã'AK. , ÈJ æÀ @, È™" Aj. À, HA'Km Ã'@Aø	J ø, TÔI. ÈJ ø, È™"Am. Ã'@, ÈJ fl X Aø B @,
AJJ™" Ag. , ÈJ fl X Aø B @, È™" Ag. , Ë XA"™À @,	È™" Aj. À, È™"Ag. , Ë HA™" Am Ã'@, È™" Aj. MỘT,
H A™" Am Ã'@, È™" Am. Ã'@ – Qk	XA"™À @, È™" Am. MỘT'@ HAJ ø

Bảng 1: 10 từ gần nhất của È™" Am. Ã'@ và ÈJ æÀ @.

Chúng tôi đã sử dụng thuộc tính này để đánh giá mức độ giống nhau về ngữ nghĩa giữa SE0 và SA, trư ớc tiên chúng tôi tiến hành xây dựng vectơ biểu diễn RV cho mỗi câu. Gọi RVE0 và RVA lần lư ợt là các vectơ biểu diễn của SE0 và SA , kích thư ớc của mỗi vectơ là số từ trong câu tư ơ ng ứng của nó. Giá trị của một mục trong vectơ biểu diễn đư ợc xác định như sau:

1. Với mỗi từ w chúng ta truy xuất từ w0 đã đư ợc căn chỉnh của nó trong câu còn lại bằng cách sử dụng BM hàm xác định theo công thức (2).
2. Chúng tôi sử dụng mô hình nhúng để xây dựng cho cả w và w0 Túi từ của chúng. BoWw (BoWw0) chứa các Boww và Boww0 . trong mô từ k gần nhất với w (w0) hình nhúng.
3. Chúng tôi tính toán độ tư ơ ng tự Jaccard giữa BoWw và BoWw0 :

$$\text{Jacc(BoWw, BoWw0)} = \frac{\text{BoWw} \setminus \text{BoWw0}}{[\ \text{BoWw0}}$$

4. Giá trị của mục nhập RV [w] đư ợc đặt thành Jacc(BoWw, BoWw0).
5. Quy trình này đư ợc áp dụng cho tất cả các từ trong cả hai câu để xây dựng RVE0 và RVA.
6. Cuối cùng, sự tư ơ ng đồng giữa SE0 và SA có đư ợc bằng cách:

$$\text{Sim(SE0 , SA)} = \frac{1}{2}\left(\frac{\sum_{w2SE0} WT(w)\ RV[w]}{\sum_{w2SE0} WT(w)} + \frac{\sum_{w2SA} WT(w)\ RV[w0]}{\sum_{w2SA} WT(w)}\right)_{1} \quad (3)$$

4 thí nghiệm và kết quả

Để đánh giá hệ thống của chúng tôi và theo dõi hiệu suất của chúng, chúng tôi đã sử dụng bốn bộ dữ liệu đư ợc lấy từ nhiệm vụ chung STS SemEval-2017 (Nhiệm vụ 1: STS Đa ngôn ngữ Ả Rập-Anh)4 [8], với tổng số 2412 cặp câu. Các cặp câu đã đư ợc năm ngư ời chú thích dán nhãn thủ công và điểm tư ơ ng đồng là giá trị trung bình của các đánh giá của năm ngư ời chú thích. Điểm này là một số thực giữa “0” (biểu thị ý nghĩa của các câu hoàn toàn độc lập) đến “5” (biểu thị ý nghĩa tư ơ ng đư ơ ng). Thông tin thêm về các bộ dữ liệu đư ợc sử dụng đư ợc liệt kê trong Bảng 2.

4 http://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools

Tập dữ liệu	Nguồn	Cặp
MSRvid	Tập tài liệu mô tả video nghiên cứu của Microsoft	736
MSRpar	Tập hợp cụm từ nghiên cứu của Microsoft	1020
SMTeuroparl	Bộ dữ liệu phát triển WMT2008	406
Dữ liệu đánh giá STS SNLI Corpus		250

Bảng 2: Bộ đánh giá tiếng Ả Rập-Anh.

4.1 Kết quả thực nghiệm

Chúng tôi đã nghiên cứu hiệu suất của cả hệ thống Trọng số các từ được căn chỉnh (W-AW) và Hệ thống túi từ căn chỉnh (A-BoW) với ba chức năng trọng số: IDF, POS và sự kết hợp của cả hai. Ngoài ra, đối với phương pháp A-BoW, chúng tôi đã sử dụng bốn giá trị khác nhau của k để tạo ra 5 từ gần nhất, 10 từ gần nhất, 15 từ gần nhất và 20 từ gần nhất. Sau đó, Để đánh giá độ chính xác của từng phương pháp, chúng tôi tính toán hệ số tương quan Pearson giữa điểm tương đồng về ngữ nghĩa được chỉ định của chúng tôi và đánh giá của con người trên SemEval Bộ dữ liệu nhiệm vụ STS. Bảng 3 trình bày kết quả của các phương pháp đề xuất.

Phương pháp	MSRvid	MSRpar	SMTeuro.	Đánh giá STS	Nghĩa là
W-AW-IDF	0,6895	0,7019	0,7274	0,6951	0,7034
W-AW-POS 0,6924	W-AW-IDF-POS	0,7402	0,7478	0,7205	0,7252
0,7015		0,7385	0,7512	0,7375	0,7321
k = 5					
A-BoW-IDF	0,6863	0,7119	0,7174	0,6881	0,7009
A-BoW-POS	0,6933	0,7349	0,7364	0,7187	0,7218
A-BoW-IDF-POS 0,7074		0,7365	0,7482	0,7362	0,7320
k = 10					
A-BoW-IDF	0,6879	0,7131	0,7291	0,7114	0,7103
A-BoW-POS	0,7084	0,7437	0,7514	0,7305	0,7335
A-BoW-IDF-POS 0.7216		0,7418	0,7603	0,7565	0,7450
k = 15					
A-BoW-IDF A-	0,6954	0,7089	0,7284	0,7254	0,7145
BoW-POS 0,7124	A-BoW-IDF-POS	0,7402	0,7578	0,7391	0,7398
0,7575		0,7485	0,7672	0,7739	0,7603
k = 20					
A-BoW-IDF	0,6912	0,7055	0,7283	0,7254	0,7244
A-BoW-POS	0,7254	0,7382	0,7514	0,7351	0,7351
A-BoW-IDF-POS 0,7525		0,7477	0,7689	0,7613	0,7576

Bảng 3: Phương pháp của chúng tôi so với đánh giá của con người

Những kết quả này chỉ ra rằng khi sử dụng phương pháp tính trọng số IDF thì tỷ lệ tương quan trung bình không giảm xuống dưới 70% trong tất cả các phương pháp được thử nghiệm. Khi áp dụng POS và

trọng số hỗn hợp, tỷ lệ tư ơ ng quan của trọng số IDF tốt hơ n ở cả hai phư ơ ng pháp A-AW và A-BoW với giá trị trung bình lần lư ợt là +2,35% và +3,91%. Điều thú vị là, việc tăng tham số k để tạo ra các từ gần k nhất trong phư ơ ng pháp A-BoW, mỗi lần đều dẫn đến sự nâng cao về tỷ lệ tư ơ ng quan. Ví dụ: việc sử dụng 15 từ gần nhất sẽ tốt hơ n hệ thống 5 từ gần nhất với mức tư ơ ng quan trung bình là +2,01%. Tuy nhiên, khi k đư ợc nâng lên 20, tỷ lệ tư ơ ng quan trung bình sẽ thấp hơ n một chút. Điều này là do số lư ợng từ có ý nghĩa khác nhau trong BoW ngày càng tăng.

Từ các kết quả trên, chúng ta có thể thấy rằng độ tư ơ ng tự ư ớc tính đư ợc cung cấp bởi các phư ơ ng pháp của chúng tôi khá phù hợp với các đánh giá của con ngư ời. Tuy nhiên, sự tư ơ ng quan chư a đủ tốt khi hai câu có từ gần giống nhau như ng nghĩa hoàn toàn khác nhau, ví dụ: "H. Aċ m Ã'@ ·K . Q'' ·` AK A Jª Y™É @Q ÆK " và (Saad đọc một cuốn sách về Omar Ibn Al-Khattab) "H. Aċ m Ã'@ ·K . Q"™À AK. A Jª @Q ÆK Y™É" (Saad đọc sách cho Omar Ibn Al-Khattab). Trong ví dụ này, các câu có chung vectơ , trọng số POS và IDF. Thực tế này dẫn đến điểm tư ơ ng quan cao, như ng thực tế không phải vậy. Vấn đề này đư ợc để lại cho công việc trong tư ơ ng lai.

4.2 So sánh với ngư ời chiến thắng SemEval-2017

Chúng tôi đã so sánh kết quả tối ư u của mình với ba hệ thống tốt nhất đư ợc đề xuất trong nhiệm vụ đánh giá song ngữ tiếng Ả Rập-tiếng Anh SemEval-2017 [8] (ECNU [40], BIT [44] và HCTI [38]) và hệ thống cơ sở [8]. Trong đánh giá này, ECNU đạt hiệu suất tốt nhất với điểm tư ơ ng quan là 74,93%, tiếp theo là BIT và HCTI với lần lư ợt là 70,07% và 68,36%. Bảng 4 cho thấy sự so sánh giữa kết quả tốt nhất của chúng tôi với kết quả thu đư ợc từ ba hệ thống đã đư ợc thử nghiệm trên Dữ liệu đánh giá STS5.

phư ơ ng pháp	Đánh giá STS
W-BoW-IDF-POS (k = 15) 77,39% ECNU 74,93 % 73,75%	
W-AW-IDF-POS	
CHÚT	70,07 %
HCTI	68,36%
Đư ờng cơ sở cosin	51,55 %

Bảng 4: So sánh kết quả tư ơ ng quan với 3 hệ thống tốt nhất trong SemEval-2017.

Kết quả quan sát cho thấy phư ơ ng pháp trọng số hỗn hợp của chúng tôi với k = 15 là phư ơ ng pháp hoạt động tốt nhất với tỷ lệ tư ơ ng quan là 77,39%. Phư ơ ng pháp W-BoW-IDF-POS (k = 15) mang lại mức tăng tư ơ ng quan +9,03%, +7,32% và +2,46% về tỷ lệ tư ơ ng quan so với ECNU, BIT và HCTI.

5 http://alt.qcri.org/semeval2017/task1/data/uploads/sts2017.eval.v1.1.zip

5 Kết luận và công việc trong tư ơ ng lai

Trong bài báo này, chúng tôi trình bày hai phư ơ ng pháp đo lư ờng mối quan hệ ngữ nghĩa giữa các câu xuyên ngôn ngữ Ả Rập-Anh bằng cách sử dụng Dịch máy (MT) và cách thể hiện nhúng từ. Ý tư ởng chính dựa trên việc sử dụng ngữ nghĩa thuộc tính của các từ có trong mô hình nhúng từ. Để làm thêm tiến bộ trong việc phân tích sự giống nhau về ngữ nghĩa của câu, chúng tôi đã sử dụng sự kết hợp giữa căn chỉnh từ, trọng số IDF và POS để hỗ trợ việc xác định những từ miêu tả rõ ràng nhất trong mỗi câu. Ngoài ra, chúng tôi đã đánh giá các đề xuất của mình trên bốn bộ dữ liệu của nhiệm vụ chung STS SemEval-2017. Trong các thí nghiệm chúng tôi đã chỉ ra cách phư ơ ng pháp Bag-of-words nâng cao rõ ràng các kết quả tư ơ ng quan. Hiệu suất của các phư ơ ng pháp đề xuất của chúng tôi đã đư ợc xác nhận thông qua mối tư ơ ng quan Pearson giữa điểm tư ơ ng đồng về ngữ nghĩa đư ợc chỉ định của chúng tôi và đánh giá của con ngư ời. Trong thực tế, chúng tôi đã đạt đư ợc tỷ lệ tư ơ ng quan tốt nhất so với tất cả các hệ thống tham gia STS Nhiệm vụ phụ liên ngôn ngữ Ả Rập-Anh của SemEval-2017. Với công việc trong tư ơ ng lai, chúng tôi sẽ để kết hợp các phư ơ ng pháp này với các phư ơ ng pháp của các kỹ thuật cổ điển khác trong lĩnh vực NLP, bao gồm định hư ớng nghĩa của từ, tài nguyên ngôn ngữ và dấu vân tay tài liệu theo thứ tự để cải thiện hơ n nữa việc phát hiện đạo văn đa ngôn ngữ.

Tài liệu tham khảo

1. E. AGIRRE, C. BANEA, D. CER, M. DIAB, A. GONZALEZ-AGIRRE, R. MIHALCEA, G. RIGAU, AND J. WIEBE, Nhiệm vụ 1 Semeval-2016: Sự tư ơ ng đồng về ngữ nghĩa của văn bản, đơ n ngữ và đánh giá đa ngôn ngữ, Kỷ yếu của SemEval, (2016), trang 497–511.
2. Z. ALAA, S. TIUN, AND M. ABDULAMEER, Đạo văn xuyên ngôn ngữ của các tài liệu tiếng Ả Rập-tiếng Anh sử dụng hồi quy logistic tuyến tính, Tạp chí Công nghệ thông tin lý thuyết và ứng dụng, 83 (2016).
3. S. ALZAHRANI, Sự tư ơ ng đồng về ngữ nghĩa giữa các ngôn ngữ của các cụm từ ngắn và tiếng Ả Rập-tiếng Anh câu, Tạp chí Khoa học Máy tính, 12 (2016), trang 1–18.
4. D. BAR¨ , C. BIEMANN, I. GUREVYCH, AND T. ZESCH, Ukp: Tính toán văn bản ngữ nghĩa sự tư ơ ng đồng bằng cách kết hợp nhiều biện pháp đo lư ờng sự tư ơ ng đồng về nội dung, trong Kỷ yếu đầu tiên Hội nghị chung về ngữ nghĩa từ vựng và tính toán, Hiệp hội tính toán Ngôn ngữ học, 2012, trang 435–440.
5. A. Phát hiện BARRON ' , P. GUPTA, AND P. ROSSO, Các phư ơ ng pháp đạo văn đa ngôn ngữ -CEDENO˜ , Hệ thống dựa trên tri thức, 50 (2013), trang 211–217.
6. A. BARRON ' -CEDENO, P. ROSSO, D. PINTO, AND A. JUAN, Về đạo văn xuyên ngôn ngữ phân tích sử dụng mô hình thống kê., trong PAN, 2008, trang 1–10.
7. Y. BENGIO, R. DUCHARME, P. VINCENT, VÀ C. JAUVIN, Một ngôn ngữ xác suất thần kinh model, Tạp chí nghiên cứu máy học, 3 (2003), trang 1137–1155.
8. D. CER, M. DIAB, E. AGIRRE, I. LOPEZ-GAZPIO, VÀ L. SPECIA, nhiệm vụ 1 Semeval-2017: Đánh giá sự tư ơ ng đồng về ngữ nghĩa của văn bản đa ngôn ngữ và đa ngôn ngữ, trong Kỷ yếu của Hội thảo quốc tế lần thứ 11 về đánh giá ngữ nghĩa (SemEval-2017), Vancouver, Canada, tháng 8 năm 2017, Hiệp hội Ngôn ngữ học tính toán, trang 1–14.
9. R. COLLOBERT VÀ J. WESTON, Kiến trúc hợp nhất để xử lý ngôn ngữ tự nhiên: Mạng lư ới thần kinh sâu với học tập đa nhiệm, trong Kỷ yếu của hội nghị quốc tế lần thứ 25 về Học máy, ACM, 2008, trang 160–167.

10. J. FERRERO, F. AGNES, L. BESACIER, VÀ D. SCHWAB, Một ngư ời đa ngôn ngữ, đa phong cách và
tập dữ liệu đa chi tiết để phát hiện sự giống nhau về văn bản giữa các ngôn ngữ, trong phiên bản thứ 10 của
Hội nghị đánh giá và tài nguyên ngôn ngữ, 2016.
11. M. FRANCO-SALVADOR, P. GUPTA, AND P. ROSSO, Phát hiện đạo văn đa ngôn ngữ
sử dụng mạng ngữ nghĩa đa ngôn ngữ, trong Hội nghị truy xuất thông tin châu Âu lần thứ 35 (ECIR'13),
LNCS 7814, Springer Berlin Heidelberg, 2013, trang 710–713.
12. E. GABRILOVICH VÀ S. MARKOVITCH, Tính toán mối quan hệ ngữ nghĩa bằng cách sử dụng
Phân tích ngữ nghĩa rõ ràng dựa trên Wikipedia, trong Kỷ yếu của Liên hợp quốc tế lần thứ 20
Hội nghị về Trí tuệ nhân tạo (IJCAI'07), Hyderabad, Ấn Độ, tháng 1 năm 2007, Morgan
Nhà xuất bản Kaufmann Inc., trang 1606–1611.
13. S. GAHBICHE-BRAHAM, H. BONNEAU-MAYNARD, T. LAVERGNE VÀ F. YVON, Chung
phân đoạn và gắn thẻ pos cho tiếng Ả Rập bằng cách sử dụng bộ phân loại dựa trên crf., trong LREC, 2012,
trang 2107–2113.
14. D. GANGULY, D. ROY, M. MITRA, AND GJ JONES, Khái quát hóa dựa trên việc nhúng từ
mô hình ngôn ngữ để truy xuất thông tin, trong Kỷ yếu của ACM quốc tế lần thứ 38
Hội nghị SIGIR về Nghiên cứu và Phát triển trong Truy xuất Thông tin, ACM, 2015,
trang 795–798.
15. P. GUPTA, A. BARRON' -CEDENO, AND P. ROSSO, Tìm kiếm tư ơ ng tự cao đa ngôn ngữ
sử dụng từ điển đồng nghĩa về khái niệm, trong Hội nghị quốc tế về đánh giá đa ngôn ngữ
Diễn đàn Ngôn ngữ Châu Âu, Springer, 2012, trang 67–75.
16. A. HAPPE, B. POULIQUEN, A. BURGUN, M. CUGGIA, AND P. LE BEUX, Trích xuất khái niệm tự động
từ các báo cáo y khoa nói, Tạp chí Quốc tế về Tin học Y tế,
70 (2003), trang 255–263.
17. E. HATTAB, Phư ơ ng pháp phát hiện đạo văn đa ngôn ngữ: tiếng Ả Rập và tiếng Anh, trong Phát triển Kỹ
thuật Hệ thống Điện tử (DeSE), Hội nghị Quốc tế về 2015, IEEE, 2015,
trang 141–144.
18. CK KENT AND N. SALIM, Phát hiện đạo văn xuyên ngôn ngữ dựa trên web, trong Trí tuệ tính toán, Mô hình
hóa và Mô phỏng (CIMSiM), Hội nghị quốc tế lần thứ hai về 2010, IEEE, 2010, trang 199–204.

19. MC LEE, Một thư ớc đo độ tư ơ ng tự câu mới cho các hệ chuyên gia dựa trên ngữ nghĩa, Expert
Hệ thống có ứng dụng, 38 (2011), trang 6392–6399.
20. Y. LI, D. MCLEAN, ZA BANDAR, JD O'SHEA, AND K. CROCKETT, Tính tư ơ ng tự của câu dựa trên mạng ngữ nghĩa
và thống kê kho ngữ liệu, các giao dịch IEEE về kiến thức và
kỹ thuật dữ liệu, 18 (2006), trang 1138–1150.
21. C. LIU, C. CHEN, J. HAN, AND PS YU, Gplag: phát hiện đạo văn phần mềm bằng
phân tích biểu đồ phụ thuộc chư ơ ng trình, trong Kỷ yếu của Hội nghị quốc tế ACM SIGKDD lần thứ 12
hội nghị về Khám phá tri thức và khai thác dữ liệu, ACM, 2006, trang 872–881.
22. P. MCNAMEE VÀ J. MAYFIELD, Mã thông báo n-gram ký tự cho ngôn ngữ châu Âu
truy xuất văn bản, Truy xuất thông tin, 7 (2004), trang 73–97.
23. T. MIKOLOV, K. CHEN, G. CORRADO, AND J. DEAN, Ư ớc tính hiệu quả các lần lặp lại từ trong không gian
vectơ , trong: ICLR: Kỷ yếu của Hội nghị quốc tế về
Theo dõi Hội thảo về Trình bày Học tập, 2013, trang 1301–3781.
24. T. MIKOLOV, M. KARAFIAT' , L. BURGET, J. CERNOCKY` thuê mô , VÀ S. KHUDANPUR, Tái diễn-
hình ngôn ngữ dựa trên mạng thần kinh., trong Interspeech, tập. 2, 2010, tr. 3.
25. T. MIKOLOV, I. SUTSKEVER, K. CHEN, GS CORRADO, VÀ J. DEAN, Đã phân phối
cách thể hiện các từ và cụm từ cũng như thành phần của chúng, trong Những tiến bộ trong hệ thống xử
lý thông tin thần kinh, 2013, trang 3111–3119.
26. T. MIKOLOV, W.-T. YIH, AND G. ZWEIG, Quy tắc ngôn ngữ trong từ không gian liên tục
đại diện., trong Hlt-naacl, tập. ngày 13 tháng 1 năm 2013, trang 746–751.
27. GA MILLER, Wordnet: cơ sở dữ liệu từ vựng cho tiếng Anh, Communications of the ACM, 38
(1995), trang 39–41.

28. A. MNIH VÀ GE HINTON, Mô hình ngôn ngữ phân tán có cấp bậc có thể mở rộng, trong Những cải tiến trong Hệ thống xử lý thông tin thần kinh 21, D. Koller, D. Schuurmans, Y. Bengio, và L. Bottou, biên tập, Curran Associates, Inc., 2009, trang 1081–1088.
29. M. MUHR, R. KERN, M. ZECHNER, AND M. GRANitzER, Phát hiện đạo văn bên ngoài và bên trong bằng cách sử dụng hệ thống phân đoạn và truy xuất đa ngôn ngữ, trong Notebook Papers của Phòng thí nghiệm và Hội thảo CLEF 2010, 2010.
30. EM B. NAGOUDI VÀ D. SCHWAB, Kỷ yếu của Ngôn ngữ tự nhiên Ả Rập thứ ba Hội thảo xử lý, Hiệp hội Ngôn ngữ học tính toán, 2017, ch. Sự giống nhau về mặt ngữ nghĩa của các câu tiếng Ả Rập với cách nhúng từ, trang 18–24.
31. R. NAVIGLI VÀ SP PONZETTO, BabelNet: Tự động xây dựng, đánh giá và ứng dụng mạng ngữ nghĩa đa ngôn ngữ có phạm vi bao phủ rộng, trong Trí tuệ nhân tạo Pro-ceedings, tập. 193, 2012, trang 217–250.
32. M. PATAKI, Một cách tiếp cận mới để tìm kiếm đạo văn đã dịch, (2012).
33. J. PENNINGTON, R. SOCHER, AND CD MANNING, Glove: Các vectơ toàn cầu cho từ đại diện- sự phẫn uất., trong EMNLP, tập. ngày 14 tháng 1 năm 2014, trang 1532–1543.
34. D. PINTO, J. CIVERA, A. BARRON ' -CEDENO, A. JUAN, AND P. ROSSO, Một cách tiếp cận thống kê đối với các nhiệm vụ ngôn ngữ tự nhiên xuyên ngôn ngữ, Tạp chí Thuật toán, 64 (2009), trang 51– 60.
35. M. POTTHAST, A. BARRON' -CEDENO˜ phát hiện , B. STEIN, AND P. ROSSO, Đạo văn đa ngôn ngữ- rism, Tài nguyên và Đánh giá Ngôn ngữ, 45 (2011), trang 45–62.
36. M. POTTHAST, B. STEIN, VÀ M. ANDERKA, Truy xuất đa ngôn ngữ dựa trên Wikipedia Model, tại Hội nghị Châu Âu lần thứ 30 về Nghiên cứu IR (ECIR'08), tập. 4956 của LNCS của Ghi chú bài giảng về Khoa học Máy tính, Glasgow, Scotland, tháng 3 năm 2008, Springer, trang 522–530.
37. M. RIOS VÀ L. SPECIA, Uow: Quá trình gaussian học đa tác vụ cho văn bản ngữ nghĩa sự tư ơ ng đồng, Kỷ yếu của SemEval, (2014), trang 779–784.
38. Y. SHAO, Trong kỷ yếu hội thảo quốc tế lần thứ 11 về đánh giá ngữ nghĩa, (Se- mEval 2017).
39. MA SULTAN, S. BETHARD, AND T. SUMNER, Dls@ cu: Câu giống với từ sự liên kết và thành phần vectơ ngữ nghĩa, trong Kỷ yếu của Hội thảo quốc tế lần thứ 9 về Đánh giá ngữ nghĩa, 2015, trang 148–153.
40. J. TIAN, Z. ZHOU, M. LAN, AND Y. WU, Trong kỷ yếu hội thảo quốc tế lần thứ 11 về đánh giá ngữ nghĩa, (SemEval 2017).
41. J. TURIAN, L. RATINOV, VÀ Y. BENGIO, Cách biểu diễn từ: một cách đơ n giản và tổng quát phư ơ ng pháp học bán giám sát, trong Kỷ yếu cuộc họp thư ờng niên lần thứ 48 của hiệp hội ngôn ngữ học tính toán, Hiệp hội Ngôn ngữ học tính toán, 2010, trang 384–394.
42. A. VINOKOUROV, J. SHAWE-TAYLOR, VÀ N. CRISTIANINI, Suy ra sự tái hiện ngữ nghĩa của văn bản thông qua phân tích tư ơ ng quan đa ngôn ngữ, NIPS-02: Những tiến bộ trong thần kinh Hệ thống xử lý thông tin, (2003), trang 1473–1480.
43. W. WALI, B. GARGOURI, AND AB HAMADOU, Tăng cư ờng độ đo độ tư ơ ng tự của câu bằng kiến thức ngữ nghĩa và cú pháp-ngữ nghĩa, Tạp chí Khoa học Máy tính Việt Nam, (2016), trang 1–10.
44. H. WU, H. HUANG, P. JIAN, Y. GUO, VÀ C. SU, Trong tố tụng Quốc tế lần thứ 11 workshop về đánh giá ngữ nghĩa (semeval 2017), (SemEval 2017), 2017.
45. MA ZAHRAN, A. MAGOODA, AY MAHGOUB, H. RAAFAT, M. RASHWAN, VÀ A. ATYIA, Biểu diễn từ trong không gian vectơ và ứng dụng của chúng cho tiếng Ả Rập, trong Hội nghị quốc tế về xử lý văn bản thông minh và ngôn ngữ học tính toán, Springer, 2015, trang 430–443.