PESTS: Ngôn ngữ chéo tiếng Ba Tư_Anh Corpus cho sự tương đồng về văn bản ngữ nghĩa

Mohammad Abdous một , Poorya Piroozfar a và Behrouz Minaei Bidgoli a
Khoa Kỹ thuật Máy tính, Đại học Khoa học và Công nghệ Iran, Iran
Email: mohammadabdous@comp.iust.ac.ir, Poorya_piroozfar@comp.iust.ac.ir, b_minaei@iust.ac.ir

Tóm tắt

Một trong những thành phần của xử lý ngôn ngữ tự nhiên nhận được nhiều nghiên cứu gần đây là sự tương đồng về ngữ nghĩa của văn bản. Trong ngôn ngữ học tính toán và xử lý ngôn ngữ tự nhiên, việc đánh giá sự giống nhau về ngữ nghĩa của các từ, cụm từ, đoạn văn và văn bản là rất quan trọng. Các hệ thống trả lời câu hỏi, tìm kiếm ngữ nghĩa, phát hiện gian lận, dịch máy, truy xuất thông tin và các ứng dụng khác tận dụng sự tương đồng về ngữ nghĩa của văn bản. Việc tính toán mức độ giống nhau về ngữ nghĩa giữa hai đoạn văn bản, đoạn văn hoặc cụm từ được cung cấp ở cả phiên bản đơn ngữ và đa ngôn ngữ được gọi là độ tương tự về ngữ nghĩa. Sự tương tự về ngữ nghĩa giữa các ngôn ngữ đòi hỏi ngữ liệu trong đó có các cặp câu ở cả ngôn ngữ nguồn và ngôn ngữ đích với mức độ tương tự về ngữ nghĩa giữa chúng. Nhiều mô hình tương tự ngữ nghĩa giữa các ngôn ngữ hiện có sử dụng bản dịch máy do không có sẵn bộ dữ liệu về độ tương tự ngữ nghĩa giữa các ngôn ngữ, điều này khiến việc lan truyền lỗi dịch máy làm giảm độ chính xác của mô hình. Mặt khác, khi chúng ta muốn sử dụng các tính năng tương tự về ngữ nghĩa cho dịch máy thì không nên sử dụng các bản dịch máy giống nhau cho tính tương tự về ngữ nghĩa. Đối với tiếng Ba Tư, một trong những ngôn ngữ có nguồn tài nguyên thấp, chưa có nỗ lực nào được thực hiện trong vấn đề này và nhu cầu về một mô hình có thể hiểu ngữ cảnh của hai ngôn ngữ đang trở nên cấp thiết hơn bao giờ hết. Trong bài viết này, kho ngữ liệu về sự tương đồng về mặt ngữ nghĩa giữa các câu trong tiếng Ba Tư và tiếng Anh lần đầu tiên được tạo ra bởi các chuyên gia ngôn ngữ học. Chúng tôi đặt tên cho tập dữ liệu này là PESTS (Tương tự văn bản ngữ nghĩa tiếng Anh Ba Tư). Kho ngữ liệu này chứa 5375 cặp câu. Ngoài ra, các mô hình khác nhau dựa trên máy biến áp đã được tinh chỉnh bằng bộ dữ liệu này. Kết quả cho thấy khi sử dụng bộ dữ liệu PESTS, hệ số tương quan Pearson của mô hình XLM_ROBERTa tăng từ 85,87% lên 95,62%.

Từ khóa: Tương đồng về ngữ nghĩa; Đa ngôn ngữ; Corpus tiếng Anh-Ba Tư

1. Giới thiệu

Đo lường sự giống nhau về ngữ nghĩa giữa các phần văn bản (từ, câu, đoạn văn hay thậm chí là tài liệu) là một lĩnh vực nghiên cứu rất quan trọng trong xử lý ngôn ngữ tự nhiên. Tính toán độ tương tự ngữ nghĩa giữa các câu trong nhiều ứng dụng ngôn ngữ tự nhiên như tìm kiếm ngữ nghĩa (Manjula và Geetha, 2004), Tóm tắt (Aliguliyev, 2009), Hệ thống trả lời câu hỏi (De Boni và Manandhar, 2003), phân loại tài liệu (Al-Anzi và AbuZeina , 2017), Phân tích tình cảm (Žižka và Dařena, 2010) và đạo văn (Alzahrani et al., 2011) được sử dụng.
Tìm mức độ tương đồng về ngữ nghĩa với mục đích hiểu và tạo ra ngôn ngữ tự nhiên là một nghiên cứu hấp dẫn trong lĩnh vực khoa học máy tính, trí tuệ nhân tạo và ngôn ngữ học tính toán. Tương tự ngữ nghĩa được coi là bài toán phân loại hai lớp từ năm 2006 đến năm 2012 (xác định xem hai câu có giống nhau về mặt ngữ nghĩa hay không), nhưng từ năm 2012 đến nay, mức độ tương tự thể hiện bằng số đã được tính toán (Majumder et al., 2016) .).

Các câu trong các ngôn ngữ khác nhau cũng có thể giống nhau về mặt ngữ nghĩa. Ví dụ: câu tiếng Anh "Anh ấy muốn chơi bóng đá" và câu tiếng Ba Tư "

"có thể giống nhau về mặt ngữ nghĩa. Hai câu này thuộc hai ngôn ngữ có cấu trúc khác nhau. Trong nhiều nghiên cứu được thực hiện, do thiếu dữ liệu văn bản và dữ liệu được gắn nhãn để đào tạo ở cả hai ngôn ngữ, họ sử dụng máy dịch để dịch câu ngôn ngữ nguồn sang ngôn ngữ đích và sau đó sử dụng các mô hình tương tự ngữ nghĩa của ngôn ngữ đích (trong ví dụ trên: tiếng Anh) để tính độ tương tự về ngữ nghĩa; Nhưng điểm yếu chính của các hệ thống như vậy là sự lan truyền lỗi dịch máy và việc dịch ngôn ngữ nguồn sang ngôn ngữ đích có thể không được thực hiện tốt. Một trong những mục đích của bài viết là ngăn chặn các lỗi dịch máy khi tìm kiếm sự tương đồng giữa các câu tiếng Ba Tư và tiếng Anh bằng cách tạo ra một tập dữ li
Trong Bảng 1, có thể thấy một mẫu PESTS có độ tương tự về ngữ nghĩa, nằm trong khoảng từ 0 đến 5. Điểm 5 biểu thị mức độ tương đồng cao nhất và 0 biểu thị mức độ tương tự thấp nhất.

Bảng 1. Mẫu câu tiếng Anh-Ba Tư có sự tương đồng về ngữ nghĩa

Câu Tiếng Anh	câu tiếng Ba Tư	Mức độ tương đồng
.meet với yêu cầu khó khăn không phải là cái này	Tôi không thể làm được điều đó.	0
Giai đoạn truyền tải của ngọn đuốc bao gồm đỉnh Everest, nằm ở biên giới Tây Tạng và Nepal.	Bạn có thể làm điều đó bằng cách sử dụng nó để giúp bạn có được một công việc tuyệt vời.	1
Sáu người đã thiệt mạng trong các cuộc tấn công và hơn một trăm người bị thương.	Đây là một trong những điều bạn có thể làm được.	2
Nhiệm vụ của Cơ sở Kính viễn vọng Hồng ngoại Không gian là tìm kiếm sự khởi đầu của vũ trụ.	Bạn có thể sử dụng nó để có được một khoản vay lớn .	2,5
Số người chết chính thức do đợt cấp tính và bệnh mãn tính về đường hô hấp SARS đã tăng ở 26 quốc gia lên 293 người.	Đây là một trong những điều bạn có thể làm để có được một công việc tuyệt vời.	3
Cả hai chất liệu này đều kích thích tuần hoàn và giúp giảm mức cholesterol.	Bạn có thể làm điều đó và bạn có thể làm điều đó và bạn có thể làm điều đó.	4
Chỉ một thập kỷ trước, các công ty chủ yếu quan tâm đến việc nghiên cứu người tiêu dùng.	Bạn có thể tìm thấy nó Bạn có thể làm được điều đó Bạn có thể làm điều đó.	5

Trong bài viết này, kho ngữ liệu Tương tự giữa các ngôn ngữ tiếng Anh Ba Tư được giới thiệu và sau đó bằng cách sử dụng nó, các mô hình khác nhau dựa trên máy biến áp sẽ được tinh chỉnh và cuối cùng, các mô hình này được đánh giá bằng cách sử dụng tập dữ liệu thử nghiệm. Những đổi mới của bài viết này có thể được liệt kê như sau:

1- Sản xuất bộ dữ liệu song ngữ Ba Tư-Anh

2- Tạo mô hình có độ tương quan cao nhất để xác định mức độ tương đồng về ngữ nghĩa giữa câu tiếng Ba Tư và tiếng Anh

3- Loại bỏ giai đoạn dịch máy để đo độ tương tự về ngữ nghĩa ngữ nghĩa giữa hai ngôn ngữ Ba Tư-Anh.

Mục đích của bài viết này là tạo ra một bộ dữ liệu tương tự về ngữ nghĩa giữa tiếng Ba Tư và tiếng Anh. Ở đây, sự giống nhau về ngữ nghĩa đề cập đến khoảng cách ngữ nghĩa giữa hai câu, tức là hai câu giống hay khác nhau như thế nào về nội dung từ vựng và chủ đề chung.
Kết quả đánh giá cho thấy bằng cách sử dụng kho ngữ liệu tương tự về ngữ nghĩa giữa tiếng Ba Tư và tiếng Anh, độ phong phú ngữ nghĩa của các mô hình (hiệu suất của các mô hình để xác định mức độ tương tự) có thể được tăng lên.

Trong phần tiếp theo của bài viết này, trước tiên chúng tôi sẽ mô tả các công trình liên quan trong lĩnh vực tương tự văn bản ngữ nghĩa. Sau đó chúng tôi sẽ giải thích cách chọn các cặp câu và chú thích chúng. Sau đây sẽ mô tả các số liệu thống kê liên quan đến kho ngữ liệu và cuối cùng là các mô hình tương tự về ngữ nghĩa của văn bản được xây dựng, giới thiệu và thử nghiệm. Tập dữ liệu dành cho mục đích sử dụng phi thương mại đã được công bố rộng rãi1 .

2. Công trình liên quan

Sự tương đồng về ngữ nghĩa là một trong những nhiệm vụ quan trọng mà nhiều nhà nghiên cứu trên thế giới đã thực hiện, nhưng trọng tâm chính của họ là tiếng Anh và rất ít văn bản được tạo ra về sự tương đồng về ngữ nghĩa giữa các ngôn ngữ. Để thúc đẩy sự tiến bộ về sự tương đồng về ngữ nghĩa giữa các ngôn ngữ trong các ngôn ngữ khác, năm 2017, người ta đặc biệt nhấn mạnh vào sự tương đồng về ngữ nghĩa của tiếng Anh với tiếng Ả Rập, tiếng Tây Ban Nha và tiếng Thổ Nhĩ Kỳ (Cer et al., 2017). Các kho ngữ liệu tương tự về ngữ nghĩa khác nhau đã được tạo ra từ năm 2005, một số trong đó chúng tôi sẽ đề cập đến.

Lee và cộng sự. (Lee, Pincombe và Welsh, 2005) Bộ dữ liệu của họ bao gồm 65 cặp câu dựa trên từ điển tương tự từ (Mayank, 2020). Các chuyên gia về con người đã được sử dụng để chấm điểm các cặp câu có điểm trong phạm vi từ 0 (độ tương tự tối thiểu) đến 4 (độ tương tự tối đa).
Kho dữ liệu của họ quá nhỏ để đào tạo, phát triển và thử nghiệm các hệ thống dựa trên máy học.
Li và cộng sự. (Li và cộng sự, 2006) Bộ dữ liệu của họ bao gồm 50 tài liệu tin tức với số lượng từ từ 51 đến 126 từ. Để tạo ra bức tượng này, những người gắn thẻ được yêu cầu đánh giá mức độ giống nhau giữa mỗi cặp tài liệu trong phạm vi từ 1 (độ tương tự thấp nhất) đến 5 (độ tương tự cao nhất). Tập dữ liệu này có nhiều cặp tài liệu hơn tập dữ liệu đầu tiên nhưng nó vượt xa sự giống nhau của câu và giống với sự giống nhau của tài liệu hơn. Sau đó, các bộ dữ liệu về độ tương đồng văn bản ngữ nghĩa không được xuất bản cho đến năm 2012.

Tập dữ liệu được xuất bản năm 2012 (Agirre và cộng sự, 2012) sử dụng nhiều nguồn khác nhau như Dữ liệu nghiên cứu của Microsoft (MSR), bao gồm hai bộ dữ liệu. Một trong số chúng (MSRpar) bao gồm 5.801 cặp câu và được thu thập từ hàng nghìn nguồn tin tức trên mạng trong hơn 18 năm.
trong khoảng thời gian tháng mà 67% cặp câu được gắn nhãn là bằng nhau và trong tập dữ liệu này, mức độ đồng ý của người chú thích là từ 82% đến 84%. Một tập dữ liệu khác của bộ dữ liệu Microsoft Research (MSR) là MSR Video Paraphrase Corpus, trong đó tác giả hiển thị các phân đoạn video ngắn gọn cho người chú thích và yêu cầu họ cung cấp mô tả bằng một câu về hành động hoặc sự kiện chính trong video. Dựa trên đó, gần 120.000 câu đã được thu thập cho 2.000 video.

Năm 2013, Agirre và cộng sự. (Agirre et al., 2013b) trình bày một kho ngữ liệu gồm hai nhiệm vụ: nhiệm vụ chính, tương tự như SemEval 2012 nhưng khác ở thể loại cặp câu, và nhiệm vụ còn lại là nhiệm vụ đánh máy giống nhau, cho biết hai câu như thế nào. giống nhau. Trong cả hai nhiệm vụ, mối tương quan giữa người chú thích là 62% đến 87%.

[1] https://github.com/mohammadabdous/PESTS

Trong STS 2014, được trình bày bởi Agirre et al. (Agirre và cộng sự, 2014), có hai nhiệm vụ phụ: tiếng Anh và tiếng Tây Ban Nha. Đối với nhiệm vụ phụ tiếng Anh, điểm giữa các cặp câu nằm trong khoảng từ 0 đến 5 và năm bộ dữ liệu kiểm tra được trình bày: hai bộ dữ liệu mở rộng các thể loại đã xuất bản trong các bộ dữ liệu trước đó. Các bộ sưu tập này bao gồm ánh xạ ontoNotes-WordNet Sense (750 cặp câu) và tiêu đề Tin tức (750 cặp câu). Ngoài ra, ba thể loại mới bao gồm mô tả hình ảnh (750 cặp câu), dữ liệu diễn đàn thảo luận DEFT và dây tin tức (300 cặp câu cho tin tức và 450 cặp câu cho dữ liệu diễn đàn) và ánh xạ tiêu đề tweet-newswire (750 cặp câu); Ngoài ra, tất cả các bộ dữ liệu được xuất bản vào năm 2012 và 2013 đều được sử dụng làm dữ liệu đào tạo.

Đối với nhiệm vụ phụ tiếng Tây Ban Nha, điểm giữa các cặp câu nằm trong khoảng từ 0 đến 4 và hai bộ dữ liệu đa dạng về các thể loại khác nhau được giới thiệu, cụ thể là các mô tả bách khoa được trích từ Wikipedia tiếng Tây Ban Nha (324 cặp câu) cũng như bản tin đơn ng đại của Tây Ban Nha (480 cặp câu).). Đối với nhiệm vụ phụ tiếng Tây Ban Nha, có một lượng dữ liệu được dán nhãn hạn chế, bao gồm 65 cặp câu, để huấn luyện, cho biết rằng tập dữ liệu bằng tiếng Tây Ban Nha có một số lượng nhỏ các cặp câu để thực hiện thao tác huấn luyện.

STS 2015, được trình bày bởi Agirre và cộng sự. (Agirre và cộng sự, 2015), xác định ba nhiệm vụ phụ: tiếng Anh, tiếng Tây Ban Nha và nhiệm vụ thí điểm có thể giải thích được. Đối với tiếng Anh, đã có các cặp câu từ tiêu đề tin tức (750 cặp câu) và mô tả hình ảnh (750 cặp câu), đồng thời các thể loại mới đã được giới thiệu, bao gồm các cặp câu trả lời từ hệ thống đối thoại hướng dẫn (750 cặp câu) và từ các trang web Hỏi đáp (375 cặp câu), các cặp từ tập dữ liệu niềm tin đã cam kết (375 cặp câu). Đối với tiếng Tây Ban Nha, các cặp câu được chọn từ Wikipedia tiếng Tây Ban Nha (251 câu) và các bài báo đơn ng đại được thu thập từ các phương tiện truyền thông bằng tiếng Tây Ban Nha (500 câu). Nhãn của mỗi cặp câu có được bằng cách lấy điểm trung bình của chuyên gia. Cuối cùng, với nhiệm vụ con thí điểm có thể diễn giải, người ta sẽ kiểm tra xem liệu hệ thống có thể giải thích tại sao hai câu có liên quan/không liên quan hay không và trên thực tế, một lớp giải thích đã được thêm vào điểm tương tự.

Trong SemEval 2015, bằng tiếng Anh, độ giống nhau của các câu được đánh giá bằng các số từ 0 đến 5, nhưng trong tiếng Tây Ban Nha, việc đánh giá này được thực hiện với các số từ 0 đến 4 và trên thực tế, nó được đánh giá theo cách sao cho điểm 3 và 4 cho tiếng Anh là 3 cho tiếng Tây Ban Nha.

Agirre và cộng sự. (Agirre và cộng sự, 2016) Trong tập dữ liệu của họ, dữ liệu kiểm tra đa ngôn ngữ được chia thành hai bộ kiểm tra: tin tức và đa nguồn. Các bộ dữ liệu tin tức được thu thập thủ công từ các nguồn tin tức đa ngôn ngữ, trong khi các bộ dữ liệu đa nguồn được trích xuất từ các nguồn khác nhau trong văn bản tiếng Anh, nhưng các câu từ ngôn ngữ khác có được bằng cách dịch các câu tiếng Anh sang ngôn ngữ đó bởi người dịch. Tập dữ liệu này bao gồm các tập dữ liệu cho các cặp ngôn ngữ Tây Ban Nha-Anh và có các tập dữ liệu tiếng Anh-Anh để đánh giá mức độ tương tự về ngữ nghĩa đơn ngữ. Bộ dữ liệu đánh giá tin tức bao gồm 301 cặp câu và bộ dữ liệu đa nguồn chứa 2973 cặp câu được sử dụng để đánh giá các mô hình và phương pháp.

Marelli và cộng sự. (M. Marelli và cộng sự, 2014) tập dữ liệu của họ bao gồm 10.000 cặp câu tiếng Anh và mỗi cặp câu được chú thích cho hai nhiệm vụ ngữ nghĩa quan trọng, một là mối quan hệ ngữ nghĩa của hai câu được đánh dấu từ 1 đến 5 và còn lại là quan hệ kế thừa, được đặc trưng bởi 3 nhãn: kế thừa, mâu thuẫn và trung lập.

Tập dữ liệu này được tạo thành từ hai tập dữ liệu hiện có: tập dữ liệu imageFlickr và tập dữ liệu mô tả video trong SemEval 2012. Để tạo các cặp câu cho tập dữ liệu Sick, một số cặp câu được chọn ngẫu nhiên từ mỗi tập dữ liệu nguồn và được xử lý trước trong 3 bước. Đầu tiên các câu chính được chuẩn hóa để loại bỏ các hiện tượng ngôn ngữ không mong muốn, sau đó các câu chuẩn hóa được mở rộng để thu được tối đa ba câu mới với những tính năng đặc biệt phù hợp cho việc đánh giá hệ thống, và cuối cùng là bước cuối cùng, tất cả các câu được tạo ra trong

bước mở rộng, chúng được ghép nối thành các câu chuẩn hóa để thu được tập dữ liệu cuối cùng. Nhãn vàng của mỗi cặp câu cho nhiệm vụ tương tự về ngữ nghĩa có được bằng cách lấy điểm trung bình của 10 người tham gia, điều này cho thấy rằng, trung bình, các đánh giá của người tham gia thay đổi tới 0,76% số điểm xung quanh điểm cuối cùng được giao cho mỗi người. đôi.

Ferrero và cộng sự. (Ferrero và cộng sự, 2016) đã cung cấp một bộ dữ liệu để đánh giá sự giống nhau của các văn bản song ngữ có thể rất hữu ích trong việc phát hiện gian lận. Bộ dữ liệu này đa ngôn ngữ (tiếng Pháp, tiếng Anh và tiếng Tây Ban Nha) và cung cấp sự liên kết thông tin theo nhiều ngôn ngữ ở các cấp độ khác nhau của tài liệu, câu và đoạn, đồng thời bao gồm văn bản của con người hoặc sử dụng bản dịch máy. Bộ dữ liệu này cũng bao gồm nhiều loại tài liệu được viết bởi nhiều nhà văn cấp trung đến cấp cao. Tập dữ liệu này đã khắc phục được nhiều hạn chế của các tập dữ liệu trước đó, một trong số đó là khả năng chỉ căn chỉnh ở một mức nhất định (ví dụ: cấp độ câu).

Trong độ tương tự ngữ nghĩa giữa các ngôn ngữ, mục tiêu là tính toán mức độ tương tự giữa các câu trong hai ngôn ngữ khác nhau. Không có vấn đề gì trong việc tính toán độ tương tự về mặt ngữ nghĩa trong các ngôn ngữ có tài nguyên cao, nhưng ở một số ngôn ngữ không có nguồn phù hợp, việc tính toán độ tương tự là một thách thức nghiêm trọng. Một trong những giải pháp sẵn có để giải quyết thách thức này là sử dụng các phương pháp tiếp cận dựa trên dịch máy để chuyển đổi các câu từ ngôn ngữ nguồn thấp như tiếng Ba Tư sang ngôn ngữ nguồn cao như tiếng Anh. Vấn đề chính của các phương pháp này là sự tồn tại của lỗi trong dịch máy và nó phụ thuộc nhiều vào chất lượng dịch (Bjerva và Ostling, 2017).

Trong nghiên cứu của họ, Tang et al. (Tang và cộng sự, 2018) đã phát triển mô hình cho các ngôn ngữ có nguồn tài nguyên thấp như tiếng Tây Ban Nha, tiếng Ả Rập, tiếng Indonesia và tiếng Thái. Bằng cách sử dụng khung mô hình tương tự ngữ nghĩa đơn ngữ, họ đã mở rộng nó sang chế độ đa ngôn ngữ và cho thấy rằng bằng cách sử dụng một bộ mã hóa đa ngôn ngữ chung, mỗi câu có thể hiển thị các phần nhúng khác nhau tùy theo ngôn ngữ đích.

Brychcin (Brychcín, 2020) đề xuất ý tưởng rằng các không gian ngữ nghĩa đa ngôn ngữ được đặt trong một không gian chung bằng từ điển song ngữ. Họ chỉ ra rằng không gian ngữ nghĩa chung có thể được cải thiện bằng cách đánh trọng số từ. Kết quả của họ cho thấy tiêu chí tương quan Pearson là 61,8% trong câu tiếng Ả Rập-tiếng Anh.

Tại Hội nghị đánh giá ngữ nghĩa năm 2017 (Cer và cộng sự, 2017), trọng tâm chính là sự tương đồng về ngữ nghĩa giữa các ngôn ngữ và đa ngôn ngữ. Trong hội nghị này, 17 người tham gia thi đấu ở 31 đội và trong hội nghị này, bộ dữ liệu STSBenchMark đã được trình bày. Một số bộ dữ liệu chéo ngôn ngữ bao gồm tiếng Ả Rập-Anh, tiếng Tây Ban Nha-Anh và tiếng Thổ Nhĩ Kỳ-Anh cũng được những người tham gia giới thiệu và đánh giá.

Công việc quan trọng được thực hiện trong lĩnh vực tương đồng về ngữ nghĩa giữa các ngôn ngữ dựa trên việc nhúng các ngôn ngữ chéo (Klementiev, Titov và Bhattarai, 2012)(Zou và cộng sự, 2013)(Mikolov, Le và Sutskever, 2013)(Gouws, Bengio và Corrado, 2015)(Ammar và cộng sự, 2016). Chidambaram và cộng sự. (Chidambaram và cộng sự, 2018) đã tạo ra không gian vectơ xuyên ngôn ngữ bằng mô hình dựa trên bộ mã hóa kép. Mục tiêu của họ là đào tạo một mô hình tạo ra sự giống nhau tối đa giữa các cặp câu trong kho ngữ liệu diễn giải. Việc nhúng kết quả được cải thiện bằng cách sử dụng bộ dữ liệu đơn ngữ và đào tạo đa nhiệm đồng thời. Chúng sử dụng không gian vectơ chung có được trong nhiều nhiệm vụ và có lợi thế so sánh so với các nhiệm vụ khác. Cốt lõi của phương pháp được đề xuất của họ là mô hình hóa các nhiệm vụ dựa trên việc xếp hạng các cặp câu bằng cách sử dụng bộ mã hóa kép và tạo ra các phần nhúng đa ngôn ngữ bằng cách sử dụng bản dịch máy. Trong kiến trúc bộ mã hóa dùng chung, có ba máy biến áp, mỗi máy biến áp có các lớp con chuyển tiếp nguồn cấp dữ liệu và sự chú ý nhiều đầu. Đầu ra của máy biến áp là một chuỗi có độ dài thay đổi mà bằng cách lấy trung bình chúng,

thu được sự nhúng của các câu. Sau đó, các phần nhúng được tạo trong các lớp chuyển tiếp nguồn cấp dữ liệu sẽ được sử dụng để tinh chỉnh từng tác vụ.

Conneau và cộng sự. (Conneau và cộng sự, 2018) đã phát triển một bộ dữ liệu đa ngôn ngữ có tên XNLI.
Bởi vì việc thu thập dữ liệu bằng tất cả các ngôn ngữ là một quá trình tốn kém nên mối quan tâm đến việc hiểu và truyền tải ngôn ngữ đa ngôn ngữ bằng các ngôn ngữ có nguồn lực thấp đã tăng lên. Trong bài báo này, một tập dữ liệu thử nghiệm về Hiểu ngôn ngữ đa ngôn ngữ đã được phát triển và các tập dữ liệu thử nghiệm đã được mở rộng sang 15 ngôn ngữ, bao gồm các ngôn ngữ có nguồn tài nguyên thấp như tiếng Swahili và tiếng Urdu.
Các nhãn trong kho văn bản được tạo ra là tiền đề và giả thuyết. Để chứng minh giá trị của tập dữ liệu, họ đã thử nghiệm nó trên một số tác vụ, chẳng hạn như dịch máy, túi từ đa ngôn ngữ và bộ mã hóa LSTM.

Conneau và cộng sự. (Conneau và Lample, 2019) đã đề xuất mô hình xuyên ngôn ngữ có tên XLM.
Họ đã sử dụng hai phương pháp để học các mô hình đa ngôn ngữ. Phương pháp đầu tiên là không giám sát, chỉ dựa trên dữ liệu đơn ngữ và phương pháp thứ hai được giám sát, sử dụng kho ngữ liệu diễn giải với mục đích mô hình hóa ngôn ngữ đích. Phương pháp được đề xuất của họ hoạt động tốt nhất trên các tác vụ dịch máy được giám sát và không giám sát cũng như XNLI.
Mô hình ngôn ngữ mặt nạ có mục đích tương tự như được Davlin giới thiệu trong bài báo Bert (Devlin và cộng sự, 2018), ngoại trừ việc trong mô hình này, chúng ta phải đối mặt với một chuỗi các cặp câu liên tục. Trong mô hình ngôn ngữ dịch máy, giống như mô hình ngôn ngữ mặt nạ, các cặp câu song song được đưa vào máy. Để dự đoán một từ tiếng Anh, mô hình này có thể xem xét cả bản dịch tiếng Anh và tiếng Pháp, đồng thời tìm cách căn chỉnh các phần nhúng tiếng Anh và tiếng Pháp. XLM sử dụng phương pháp Mã hóa cặp byte (BPE) và cơ chế học ngôn ngữ Bert để tìm hiểu mối quan hệ giữa các từ trong các ngôn ngữ khác nhau. Mã hóa cặp byte là phương pháp nén dữ liệu thay thế nhất quán các cặp ký tự được lặp lại thường xuyên nhất (về cơ bản là byte) trong một tập dữ liệu cụ thể bằng ký hiệu không có sự kiện trong văn bản. Trong mỗi lần lặp, thuật toán sẽ tìm các cặp ký tự thường xuyên nhất và hợp nhất chúng để tạo ra một ký hiệu mới. Trong mô hình XLM, thay vì sử dụng từ hoặc ký tự làm đầu vào cho mô hình, nó sử dụng Mã hóa cặp byte, chia đầu vào thành từ phụ phổ biến nhất trong tất cả các ngôn ngữ, do đó làm tăng vốn từ vựng đa ngôn ngữ phổ biến. Mô hình XLM nâng cao kiến trúc của Bert theo hai cách:

trong mô hình Bert, mỗi phiên bản tàu bao gồm một ngôn ngữ, trong khi ở mô hình XLM, mỗi phiên bản tàu bao gồm một ngôn ngữ. cá thể tàu bao gồm hai ngôn ngữ. Giống như dự đoán các từ bị che trong mô hình của Bert, mô hình này sử dụng ngữ cảnh của câu nguồn để dự đoán các từ bị che của câu đích. Mô hình này cũng nhận mã định danh ngôn ngữ và thứ tự từ trong từng ngôn ngữ riêng biệt dưới dạng bộ mã hóa vị trí. Siêu dữ liệu mới này giúp mô hình tìm hiểu mối quan hệ giữa các từ liên quan trong các ngôn ngữ khác nhau. Bảng 6 trong phần phụ lục tóm tắt các bộ dữ liệu về độ tương đồng văn bản ngữ nghĩa được tạo ra.

Trong bài viết này, sự tương đồng về ngữ nghĩa giữa câu tiếng Ba Tư và tiếng Anh đã được tạo ra. Sau đây, chúng tôi sẽ giải thích cách tạo tập dữ liệu.

3. Quy trình sản xuất Corpus

Như đã đề cập trước đó, để tạo kho ngữ liệu đa ngôn ngữ, trước tiên chúng ta tạo kho ngữ liệu tương tự về ngữ nghĩa tiếng Ba Tư - tiếng Ba Tư. Bởi vì trong tiếng Ba Tư, chúng ta có thể tìm thấy những chuyên gia ngôn ngữ học có thể phân biệt ngữ nghĩa giữa các câu nhưng việc tìm được chuyên gia ngôn ngữ học đồng thời cả tiếng Anh và tiếng Ba Tư thì khó khăn hơn.

Để tạo ra kho ngữ liệu tiếng Ba Tư-Ba Tư, 20.000 cặp câu tiếng Ba Tư phù hợp đầu tiên đã được chọn và điểm ngữ nghĩa giữa chúng được trình bày bởi những người chú thích; sau đó, một phần của toàn bộ kho ngữ liệu được sử dụng để tạo thành kho ngữ liệu tiếng Ba Tư-Anh. Sau đây, chi tiết về sản xuất kho ngữ liệu sẽ được mô tả.

Khi xây dựng bộ dữ liệu của chúng tôi, việc thu thập các cặp câu tự nhiên có mức độ tương đồng về ngữ nghĩa khác nhau đã là một thách thức. Nếu chúng ta xem xét một cặp câu một cách ngẫu nhiên, phần lớn chúng sẽ hoàn toàn không liên quan với nhau và chỉ một đoạn rất nhỏ sẽ hiển thị một số loại tương đương về mặt ngữ nghĩa.

Một trong những thách thức lớn là tìm các cặp câu tương tự nhau. Nếu điểm quan trọng như vậy không được tuân thủ trong quá trình tạo kho ngữ liệu thì hầu hết kho ngữ liệu dựa trên câu đều có điểm 0 hoặc tối đa là 1, và do đó kho ngữ liệu không đủ chất lượng cho các mô hình học máy và không thể được sử dụng trong luyện tập. Để giải quyết thách thức này, chúng tôi đã sử dụng nhiều nguồn văn bản và kho văn bản khác nhau để có thể trích xuất các cặp câu từ kho văn bản này dài từ 7 đến 25 từ.
Tất nhiên, còn có các quy trình tiền xử lý khác để chọn một cặp câu được mô tả dưới đây.

3.1. Tiền xử lý câu

Một trong những vấn đề quan trọng nhất có thể thấy trong kho văn bản lớn như Ferdowsi[2], Persica (Eghbalzadeh và cộng sự, 2012), Wikipedia và những trang tương tự, là sự tồn tại của một số vấn đề xảy ra do thiếu quá trình xử lý trước các câu của nó. Các câu và văn bản khác nhau tồn tại trong các ngữ liệu này cần được xem xét và để sử dụng nó trong thực tế, cần loại bỏ những thiếu sót của nó. Sau đây, chúng tôi sẽ mô tả các quá trình tiền xử lý khác nhau được thực hiện.

1. Từ kho văn bản, chọn ra những câu có số từ từ 7 đến 25.
2. Trong kho văn bản, một số câu, một số câu (ví dụ: một hoặc nhiều từ trong đó) ở các ngôn ngữ khác, đã bị loại bỏ để các mô hình muốn tạo ra trong tương lai sử dụng kho văn bản này có chất lượng mong muốn .
3. Các câu được kiểm tra bằng công cụ sửa lỗi chính tả và sửa nếu sai chính tả.
4. Các cặp câu được chọn từ kho ngữ liệu phải giống nhau về mặt ngữ nghĩa, được thực hiện bởi các chuyên gia con người.
5. Ký tự của các từ trong câu đã được kiểm tra bằng công cụ chuẩn hóa.
6. Nếu câu có trích dẫn thì phần sau “:” (phần trích dẫn) được chọn làm câu chính
7. Sự khác biệt về độ dài câu có thể lên tới 5 từ.

3.2. Lựa chọn các cặp câu Sau khi

xử lý trước các câu và áp dụng các quy tắc liên quan, các cặp câu được trích xuất theo quyết định của chuyên gia con người từ kho ngữ liệu sao cho việc phân phối điểm tương tự trong kho ngữ liệu được tạo là mong muốn và từ tất cả các điểm (0 đến 5) có một cặp câu phù hợp.
Nếu chọn ngẫu nhiên các cặp câu từ một tập hợp các câu đã được xử lý trước thì hơn 99% trong số đó có điểm 0 hoặc cuối cùng là 1, và thực tế hầu hết các cặp câu này ít có sự giống nhau về ngữ nghĩa và sự phân bố mức độ tương đồng của chúng là không đồng nhất.

3.3. Quá trình chú thích

[2] https://github.com/Text-Mining/Ferdowsi-Annotated-Academic-Linguistic-Corpus

Trong quá trình sản xuất văn bản lớn, các hệ thống khác nhau được sử dụng để chú thích. Với việc sử dụng các hệ thống chú thích, chất lượng của kho dữ liệu sản xuất được tăng lên và tốc độ sản xuất của nó cũng tăng lên. Để tạo ra kho ngữ liệu tương tự về mặt ngữ nghĩa, một hệ thống dựa trên web đã được sử dụng và chúng tôi sẽ mô tả sau đây.

Hệ thống chú thích cho các cặp câu gồm có hai phần. Trong phần đầu tiên, mỗi người chú thích có thể xem các cặp câu liên quan đến mình bằng ID người dùng có sẵn và gán điểm từ -1 đến 5 cho mỗi cặp câu. Bởi vì, ngoài việc xem xét và xử lý trước, các cặp câu vẫn có thể mắc lỗi chính tả hoặc các sai sót về ngữ nghĩa khác, người chú thích sẽ ấn định điểm -1 cho cặp câu đó và cuối cùng, một chuyên gia khác sẽ đánh giá điểm -1 và nếu cặp câu không thể sửa được sẽ bị loại khỏi tập dữ liệu; nếu không, cặp câu sẽ được sửa và gửi lại để chú thích.

Trong phần thứ hai của hệ thống, việc sản xuất kho văn bản được quản lý. Bằng cách vào hệ thống này, người dùng quản trị có thể xem các cặp câu đã được tất cả người chú thích cho đến nay theo các bộ lọc khác nhau như giờ, số lượng người dùng chú thích, điểm trung bình của ba người chú thích, v.v. Người dùng cũng có thể sử dụng hệ thống này để sửa đổi hoặc chỉnh sửa các cặp câu đã được chú thích. Để các chuyên gia chú thích từng cặp, Hướng dẫn dưới đây được mô tả trong Bảng 2 sẽ được sử dụng.

Bảng 2. Điểm tương đồng với phần giải thích và ví dụ tiếng Anh từ (Agirre et al., 2015)

điểm	sự định nghĩa
5	Hai câu hoàn toàn tương đương nhau vì chúng có nghĩa giống nhau.
4	Hai câu hầu hết tương đương nhau, nhưng một số chi tiết không quan trọng lại khác nhau.
3	Hai câu gần như tương đương, nhưng một số thông tin quan trọng khác nhau/thiếu.
2	Hai câu không tương đương nhưng có chung một số chi tiết.
1	Hai câu không tương đương nhau nhưng có cùng một chủ đề.
0	Hai câu hoàn toàn khác nhau

3.4. Đặc điểm kỹ thuật của chuyên gia

Để chấm điểm các cặp câu, ba nhà ngôn ngữ học có bằng Tiến sĩ. bằng ngôn ngữ học và có nhiều kinh nghiệm trong lĩnh vực phân tích dữ liệu ngôn ngữ và văn bản đã được sử dụng. Tất cả người chú thích đều tuân theo cùng một cơ sở và hướng dẫn cho điểm. Mối tương quan giữa các bộ chú thích rất quan trọng và làm cho kho văn bản chính xác hơn. Bảng 3 cũng cho thấy mức độ tương quan giữa các người chú thích dựa trên mức độ họ cho điểm đối với các cặp câu. Như bạn có thể thấy trong Bảng 3, tỷ lệ tương quan Pearson giữa các nhãn là trên 90%, đây là một con số rất tốt.

Bảng 3. Mức độ tương quan giữa điểm của người chú thích

Giữa điểm của Người chú thích	tương quan Pearson
Người chú thích1 & người chú thích 2	90,32
Chú thích 1 & điểm trung bình của chú thích 2 và 3	92,66

Người chú thích1 & người chú thích 3	90,80
Chú thích 2 & điểm trung bình của chú thích 1 và 3	92,86
Người chú thích2 & người chú thích 3	92,05
Chú thích 3 & điểm trung bình của chú thích 1 và 2	93,21

3.5. Review các cặp câu của chuyên gia

Mặc dù có cùng cơ sở tính điểm nhưng người chú thích thường chỉ định các điểm khác nhau cho một cặp, điều này có thể có một số lý do: Thứ nhất, một người có thể cho điểm sai do mệt mỏi, thiếu tập trung hoặc tốc độ; Nguyên nhân thứ hai là sự tương đồng về ngữ nghĩa chưa có định nghĩa rõ ràng và thống nhất, trong nhiều trường hợp mức độ tương đồng về ngữ nghĩa còn phụ thuộc vào quan điểm cá nhân của mỗi cá nhân; Và thứ ba, điểm số mà chúng ta đang tìm kiếm (0 đến 5) có ranh giới xác định, trong khi điểm số chính xác có thể nằm giữa hai điểm. Ví dụ: nếu điểm khoảng 2,5, một người có thể đạt điểm 2 và người khác có thể đạt điểm 3. Trong kho văn bản cuối cùng, điểm cuối cùng về mức độ giống nhau về ngữ nghĩa của cặp câu được lấy từ điểm trung bình của ba người chú thích.

Theo đó, nếu chênh lệch điểm số là do sai sót cá nhân thì phải sửa lại. Nếu liên quan đến hai lý do khác (nghĩa là khái niệm về sự tương đồng không rõ ràng hoặc ranh giới điểm số là xác định) thì có thể chấp nhận được miễn là sự khác biệt về điểm số chỉ là một đơn vị. Do đó, nếu chênh lệch điểm lớn (tức là nhiều hơn một đơn vị), điều đó thường cho thấy đã xảy ra lỗi về điểm số và cần được kiểm tra.

3.6. Tạo kho ngữ liệu tiếng Anh-Ba Tư

Để xây dựng một tập dữ liệu phản ánh sự phân bố đồng đều của các phạm vi điểm tương đồng, một tập dữ liệu nhỏ hơn bao gồm 5.374 cặp câu đã được trích xuất và lấy mẫu từ hơn 20.000 cặp câu tiếng Ba Tư đã được ghi điểm. Câu đầu tiên của tuyển tập này đã được các nhà ngôn ngữ học thông thạo dịch sang tiếng Anh và đã thay thế câu tiếng Ba Tư. Như vậy, thu được một kho ngữ liệu trong đó bên thứ nhất của cặp câu là tiếng Anh và bên thứ hai là các câu tiếng Ba Tư, cột thứ ba thể hiện mức độ tương đồng giữa hai câu tiếng Ba Tư và tiếng Anh.

3.7. Thống kê tử thi

Trong kho văn bản được tạo, 5374 cặp câu đã được ba người chú thích chấm điểm và sử dụng tập huấn luyện, chiếm 90% trong số 5374 cặp câu, một mô hình tương tự về ngữ nghĩa đã được tạo. Thống kê về kho văn bản được tạo ra được mô tả trong Bảng 4, có thể được sử dụng cho các nhiệm vụ khác nhau trong lĩnh vực xử lý ngôn ngữ tự nhiên.

Bảng 4. Thống kê ngữ liệu

Số lượng cặp	Tàu hỏa (80%)	Nhà phát triển (10%)	Kiểm tra (10%)	Tất cả (100%)
tất cả đều ghi điểm	4298	538	538	5375
điểm từ 0 đến 1	920	130	131	1181

điểm từ 1 đến 2	488	60	60	608
điểm từ 2 đến 3	1087	136	136	1359
điểm từ 3 đến 4	640	80	80	800
điểm từ 4 đến 5	1163	132	131	1427
Số từ trung bình trong một câu tiếng Ba Tư	14.17	13,88	13,88	14.1
Số từ trung bình trong một câu tiếng anh	14	13,84	13,84	13,97

4. thí nghiệm

Trong bài báo này, chúng tôi đã sử dụng một số mô hình ngôn ngữ dựa trên Transformer để thực hiện các thí nghiệm trên kho ngữ liệu. Hiệu suất của các mô hình này bằng cách sử dụng kho văn bản được tạo sẽ được kiểm tra và so sánh cho các nhiệm vụ tương tự về mặt ngữ nghĩa.

Transformers được thiết kế để giải quyết vấn đề sắp xếp thứ tự trong mạng nơ ron, nghĩa là chúng lấy một chuỗi (như các từ trong câu) và sau khi xử lý thường xuyên, xuất ra theo một chuỗi cụ thể và sẽ không được phép rời đi cho đến khi toàn bộ chuỗi cuối cùng đã được phê duyệt. Máy biến áp được tạo thành từ các bộ mã hóa được sử dụng để xem xét ý nghĩa và khái niệm trong vectơ và để thu được vectơ ngữ nghĩa. Một số máy biến áp này là đa ngôn ngữ (đa ngôn ngữ Bert, XLM, v.v.) và thực sự được đào tạo bằng các ngôn ngữ khác nhau và có thể được sử dụng để biểu diễn các biểu diễn vectơ bằng các ngôn ngữ khác nhau. Tuy nhiên, hiệu suất của chúng cũng được cải thiện khi các máy biến áp này được tinh chỉnh bằng cách sử dụng ngữ liệu đa ngôn ngữ. Cần lưu ý rằng trong quá trình vận hành tinh chỉnh, trọng lượng của lớp máy biến áp cuối cùng được cập nhật bằng dữ liệu huấn luyện.

Trong các thử nghiệm này, chúng tôi đã tinh chỉnh các mô hình được sử dụng bằng cách sử dụng dữ liệu huấn luyện của kho ngữ nghĩa tương tự ngữ nghĩa tiếng Ba Tư-tiếng Anh và chúng tôi cũng đã cho thấy sự cải thiện về hiệu suất của chúng so với chế độ không tinh chỉnh, có thể thấy trong Bảng 5.

Hoạt động tinh chỉnh trong các thử nghiệm của chúng tôi được thực hiện trong 4 kỷ nguyên và ở cỡ lô 32 và hàm mất độ tương tự Cosine đã được sử dụng.

Tiêu chí tương tự cosine đã được sử dụng để đo lường sự tương tự về ngữ nghĩa giữa hai câu. Tiêu chí độ tương tự cosine giữa hai vectơ là một trong những tiêu chí được sử dụng rộng rãi nhất trong việc đo lường độ tương tự ngữ nghĩa giữa các câu.

Hệ số tương quan Pearson (Benesty et al., 2009) và Spearman (SPEARMAN, 1910) được sử dụng để đánh giá đầu ra của các hệ thống tương tự văn bản ngữ nghĩa. Mục đích là để tính toán mối tương quan giữa điểm tương đồng được hệ thống phát hiện và điểm tương tự thực sự của nó.

Cách tính hệ số tương quan Pearson theo Công thức 1:

$$= \frac{\sum_{=1} (\quad \_)(\quad \_)}{\sqrt{\quad =1 \quad (\quad \rightarrow)} \sqrt{\quad =1 \quad (\quad \_)}} \qquad (1)$$

Trong công thức trên cho biết điểm đầu tiên (hoặc dự đoán) và cho biết điểm thứ hai (hoặc vàng). cho biết điểm trung bình của điểm thứ nhất (hoặc dự đoán) và cho biết điểm trung bình của điểm thứ hai (hoặc vàng). Điểm dự đoán hoặc điểm vàng được sử dụng trong giai đoạn thử nghiệm.

Nếu hệ số tương quan Pearson gần bằng 1 thì mô hình thu được sẽ chính xác hơn. Sau đây trình bày các kết quả liên quan đến việc triển khai mô hình trên dữ liệu thử nghiệm độ tương đồng ngữ nghĩa của văn bản giữa các ngôn ngữ Anh-Ba Tư.

Để đo lường hiệu suất của mô hình được tạo ra, chúng tôi đã thử nghiệm và đánh giá nó bằng cách sử dụng dữ liệu thử nghiệm về độ tương tự văn bản ngữ nghĩa giữa các ngôn ngữ. Tiêu chí tương tự cosine được sử dụng để tính điểm tương tự về ngữ nghĩa giữa hai vectơ câu và tiêu chí Pearson và Spearman được sử dụng để đo lường mối tương quan giữa điểm mô hình và điểm vàng. Kết quả của các mô hình như trong Bảng 5 và cho thấy rằng bằng cách sử dụng kho ngữ liệu được tạo ra, tỷ lệ tương quan của các mô hình dựa trên máy biến áp có thể tăng lên để đo lường mức độ tương tự về ngữ nghĩa giữa tiếng Anh và tiếng Ba Tư, ví dụ: paraphrase-xlm- Mô hình r-multilingual-v1 có tương quan Pearson là 85,87 ở chế độ không tinh chỉnh và tương quan Pearson là 95,62 ở chế độ tinh chỉnh, điều này đã làm tăng tỷ lệ tương quan lên khoảng 10% khi sử dụng kho ngữ liệu được tạo. Cũng cần lưu ý rằng trong kết quả thu được, dữ liệu thử nghiệm để đánh giá và dữ liệu huấn luyện để tinh chỉnh đều được cố định trong tất cả các thử nghiệm.

Bảng 5. Kết quả triển khai mô hình dựa trên biến đổi đa ngôn ngữ cho sự tương đồng về ngữ nghĩa giữa các ngôn ngữ

			Mô hình tinh chỉnh	
Người mẫu[3]	Pearson	người cầm thước	Pearson	người cầm thước
Cơ sở Xlm-roberta (Conneau và cộng sự, 2019)	23:80	28,99	89,48	90,40
diễn giải-xlm-r-đa ngôn ngữ-v1	85,87	85,91	95,62	95,17
bert-base-multilingual-cased (Reimers và Gurevych, 2019)	47	45,47	91,88	91,55
Distilbert-base-đa ngôn ngữ	45,56	45,18	89,51	89,08
stsb-xlm-r-đa ngôn ngữ	78,37	76,57	94,43	94.02
xlm-r-100langs-bert-base-nli-stsb-mean- mã thông báo	78,37	76,57	94,4	94.03
Cơ sở Twitter-xlm-roberta	27,94	28.06	90,99	90,28

[3] https://huggingface.co/models

5. Kết luận

Ngày nay, với sự phát triển ngày càng tăng của nguồn tài nguyên văn bản bằng các ngôn ngữ khác nhau, nhu cầu tạo ra các mô hình có khả năng hiểu đồng thời nhiều ngôn ngữ đang trở nên cần thiết hơn bao giờ hết. Một trong những nhiệm vụ quan trọng nhất được sử dụng trong xử lý ngôn ngữ tự nhiên là hiểu ý nghĩa của câu hoặc cụm từ, được gọi là sự tương đồng về ngữ nghĩa của văn bản.

Sự tương đồng về văn bản ngữ nghĩa là một trong những nhiệm vụ quan trọng của xử lý ngôn ngữ tự nhiên đã thu hút nhiều nghiên cứu sâu rộng có thể được sử dụng làm ngôn ngữ chéo. Trong nghiên cứu này, một kho ngữ liệu tương tự về mặt ngữ nghĩa giữa các ngôn ngữ đã được tạo ra. Để tạo ra kho ngữ liệu này, đầu tiên một kho ngữ liệu tiếng Ba Tư-Ba Tư đã được tạo ra và sau đó phần đầu tiên của mỗi cặp câu đã được các nhà ngôn ngữ học thông thạo dịch sang tiếng Anh.

Các thí nghiệm được thực hiện trong nghiên cứu này đã cho thấy tầm quan trọng của việc tạo ra kho ngữ liệu này có thể được sử dụng để tăng hiệu suất của các mô hình cho các nhiệm vụ tương tự về ngữ nghĩa giữa ngôn ngữ Ba Tư-Anh và cũng để đánh giá và so sánh các mô hình sử dụng dữ liệu thử nghiệm của cùng một kho văn bản.

Các mô hình được sản xuất cũng có thể được sử dụng trong các hệ thống trả lời câu hỏi, phát hiện gian lận, dịch máy, truy xuất thông tin và những thứ tương tự bằng tiếng Ba Tư và tiếng Anh.

Tài liệu tham khảo

Agirre, E., Cer, D., Diab, M. và Gonzalez-Agirre, A., 2012. Nhiệm vụ Semeval-2012 6: Thí điểm về sự tương đồng về ngữ nghĩa của văn bản*. Trong SEM 2012: Hội nghị chung đầu tiên về ngữ nghĩa từ vựng và tính toán–Tập 1: Kỷ yếu hội nghị chính và nhiệm vụ chung, và Tập 2: Kỷ yếu Hội thảo quốc tế lần thứ sáu về đánh giá ngữ nghĩa (SemEval 2012) (trang 385-393).

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A. và Guo, W., (2013A) * Nhiệm vụ chung của SEM 2013: Sự tương đồng về ngữ nghĩa của văn bản. Trong Hội nghị chung lần thứ hai về ngữ nghĩa từ vựng và tính toán (* SEM), tập 1: Kỷ yếu của Hội nghị chính và nhiệm vụ chung: sự tương đồng về ngữ nghĩa của văn bản (trang 32-43).

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A. và Guo, W., (2013b) SEM 2013 nhiệm vụ chung: Tương tự văn bản ngữ nghĩa, *SEM 2013 - Hội nghị chung lần thứ 2 về Từ vựng và Ngữ nghĩa tính toán, 1, trang 32–43.

Agirre, E., Banea, C., Cardie, C., Cer, DM, Diab, MT, Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G. và Wiebe, J., 2014, tháng 8. SemEval-2014 Nhiệm vụ 10: Sự tương đồng về văn bản ngữ nghĩa đa ngôn ngữ. Trong SemEval@ COLING (trang 81-91).

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea , R. và Rigau, G., 2015, tháng 6. Nhiệm vụ Semeval-2015 2: Tương tự văn bản ngữ nghĩa, tiếng Anh, tiếng Tây Ban Nha và thí điểm về khả năng diễn giải. Trong Kỷ yếu hội thảo quốc tế lần thứ 9 về đánh giá ngữ nghĩa (SemEval 2015) (trang 252-263).

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez Agirre, A., Mihalcea, R., Rigau Claramunt, G. và Wiebe, J., 2016. Nhiệm vụ 1 của Semeval-2016: Sự tương đồng về ngữ nghĩa của văn bản, đánh giá đơn ngữ và đa ngôn ngữ. Trong SemEval-2016. Hội thảo quốc tế lần thứ 10 về đánh giá ngữ nghĩa; 2016 ngày 16-17 tháng 6; San Diego, CA. Stroudsburg (PA): ACL; 2016. tr. 497-511.. ACL (Hiệp hội Ngôn ngữ học tính toán).

Al-Anzi, FS và AbuZeina, D., 2017. Hướng tới phân loại văn bản tiếng Ả Rập nâng cao bằng cách sử dụng độ tương tự cosine và Lập chỉ mục ngữ nghĩa tiềm ẩn. Tạp chí Khoa học Thông tin và Máy tính của Đại học King Saud, 29(2), tr.189-195.

Aliguliyev, RM, 2009. Một thước đo độ tương tự câu mới và kỹ thuật trích xuất dựa trên câu để tóm tắt văn bản tự động. Hệ thống chuyên gia với các ứng dụng, 36(4), tr.7764-7772.

Alzahrani, SM, Salim, N. và Abraham, A., 2011. Tìm hiểu các mô hình ngôn ngữ, đặc điểm văn bản và phương pháp phát hiện đạo văn. Giao dịch của IEEE về Hệ thống, Con người và Điều khiển học, Phần C (Ứng dụng và Đánh giá), 42(2), tr.133-149.

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C. và Smith, NA, 2016. Khả năng nhúng từ đa ngôn ngữ với số lượng lớn. bản in trước arXiv arXiv:1602.01925.

Benesty, J., Chen, J., Huang, Y. và Cohen, I., 2009. Hệ số tương quan Pearson. Trong Giảm tiếng ồn trong xử lý giọng nói (trang 1-4). Springer, Berlin, Heidelberg.

Bjerva, J. và Östling, R., 2017. Học đa ngôn ngữ về sự tương đồng về ngữ nghĩa của văn bản với cách biểu thị từ đa ngôn ngữ. Trong Hội nghị Bắc Âu lần thứ 21 về Ngôn ngữ học tính toán, NoDaLiDa, Gothenburg, Thụy Điển, ngày 22-24 tháng 5 năm 2017 (trang 211-215). Nhà xuất bản điện tử của Đại học Linköping.

De Boni, M. và Manandhar, S., 2003. Việc sử dụng độ tương tự của câu như một thước đo mức độ liên quan về mặt ngữ nghĩa để trả lời câu hỏi. Trong Hướng đi Mới trong Trả lời Câu hỏi (trang 138-144).

Brychcín, T., 2020. Các phép biến đổi tuyến tính cho sự tương đồng về ngữ nghĩa văn bản giữa các ngôn ngữ. Hệ thống dựa trên tri thức, 187, p.104819.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I. và Specia, L., 2017. Nhiệm vụ Semeval-2017 1: Đánh giá tập trung vào sự tương đồng về ngữ nghĩa của văn bản-đa ngôn ngữ và đa ngôn ngữ. bản in trước arXiv arXiv:1708.00055.

Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, YH, Strope, B. và Kurzweil, R., 2018. Học cách trình bày câu đa ngôn ngữ thông qua bộ mã hóa kép đa tác vụ người mẫu. bản in trước arXiv arXiv:1810.12836.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, SR, Schwenk, H. và Stoyanov, V., 2018. XNLI: Đánh giá cách trình bày câu đa ngôn ngữ. bản in trước arXiv arXiv:1809.05053.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. và Stoyanov, V., 2019. Học tập biểu diễn đa ngôn ngữ không giám sát trên quy mô lớn. bản in trước arXiv arXiv:1911.02116.

Conneau, A. và Lample, G., 2019. Đào tạo trước mô hình ngôn ngữ đa ngôn ngữ. Những tiến bộ trong hệ thống xử lý thông tin thần kinh, 32.

Devlin, J., Chang, MW, Lee, K. và Toutanova, K., 2018. Bert: Đào tạo trước các máy biến áp hai chiều sâu để hiểu ngôn ngữ. bản in trước arXiv arXiv:1810.04805.

Eghbalzadeh, H., Hosseini, B., Khadivi, S. và Khodabakhsh, A., 2012, tháng 11. Persica: Kho ngữ liệu tiếng Ba Tư để khai thác văn bản đa mục đích và xử lý ngôn ngữ tự nhiên. Trong Hội nghị chuyên đề quốc tế lần thứ 6 về Viễn thông (IST) (trang 1207-1214). IEEE.

Ferrero, J., Agnes, F., Besacier, L. và Schwab, D., 2016, tháng 5. Bộ dữ liệu đa ngôn ngữ, đa phong cách và đa chi tiết để phát hiện sự tương đồng về văn bản giữa các ngôn ngữ. Trong phiên bản thứ 10 của Hội nghị đánh giá và tài nguyên ngôn ngữ..

Gouws, S., Bengio, Y. và Corrado, G., 2015, tháng 6. Bilbowa: Biểu diễn phân tán song ngữ nhanh chóng mà không cần căn chỉnh từ. Trong Hội nghị quốc tế về học máy (trang 748-756). PMLR..

Klementiev, A., Titov, I. và Bhattarai, B., 2012, tháng 12. Tạo ra các biểu diễn phân tán đa ngôn ngữ của các từ. Trong Kỷ yếu của COLING 2012 (trang 1459-1474).

Lee, MD, Pincombe, B. và Welsh, M., 2005 'Đánh giá thực nghiệm các mô hình về sự tương đồng của tài liệu văn bản', trong Kỷ yếu cuộc họp thường niên của xã hội khoa học nhận thức.

Li, Y., McLean, D., Bandar, ZA, O'shea, JD và Crockett, K., 2006. Độ tương tự của câu dựa trên mạng ngữ nghĩa và thống kê kho ngữ liệu. Các giao dịch của IEEE về kiến thức và kỹ thuật dữ liệu, 18(8), tr.1138-1150.

Majumder, G., Pakray, P., Gelbukh, A. và Pinto, D., 2016. Các phương pháp, công cụ và ứng dụng tương tự về ngữ nghĩa của văn bản: Một cuộc khảo sát. Computación y Sistemas, 20(4), tr.647-665.

Manjula, D. và Geetha, TV, 2004. Công cụ tìm kiếm ngữ nghĩa. Tạp chí Quản lý thông tin và tri thức, 3(01), tr.107-117.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R. và Zamparelli, R., 2014, Tháng 5. Một phương pháp chữa bệnh SICK để đánh giá các mô hình ngữ nghĩa phân bố tổng hợp. Trong Kỷ yếu của Hội nghị quốc tế lần thứ chín về nguồn lực và đánh giá ngôn ngữ (LREC'14) (trang 216-223).

Mayank, M., 2020. Phân tích nội tại cho các mô hình không gian nhúng từ kép. bản in trước arXiv arXiv:2012.00728.

Mikolov, T., Le, QV và Sutskever, I., 2013. Khai thác sự tương đồng giữa các ngôn ngữ để dịch máy.

Reimers, N. và Gurevych, I., 2019, tháng 11. Câu-BERT: Nhúng câu bằng cách sử dụng Mạng BERT của Xiêm. Trong Kỷ yếu của Hội nghị về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên năm 2019 và Hội nghị chung quốc tế lần thứ 9 về xử lý ngôn ngữ tự nhiên (EMNLP-IJCNLP) (trang 3982-3992).

Spearman, C., 1910. Mối tương quan được tính toán từ dữ liệu bị lỗi. Tạp chí tâm lý học Anh, 3(3), tr.271.

Tang, X., Cheng, S., Do, L., Min, Z., Ji, F., Yu, H., Zhang, J. và Chen, H., 2018. Cải thiện sự tương đồng về ngữ nghĩa của văn bản đa ngôn ngữ với câu chia sẻ bộ mã hóa cho các ngôn ngữ có nguồn tài nguyên thấp. bản in trước arXiv arXiv:1810.08740.

Žiž ka, J. và Dařena, F., 2010, tháng 9. Phân tích cảm xúc tự động bằng cách sử dụng sự tương đồng về nội dung mẫu văn bản trong ngôn ngữ tự nhiên. Trong Hội nghị quốc tế về văn bản, lời nói và đối thoại (trang 224-231). Springer, Berlin, Heidelberg.

Zou, WY, Socher, R., Cer, D. và Manning, CD, 2013, tháng 10. Nhúng từ song ngữ cho dịch máy dựa trên cụm từ. Trong Kỷ yếu hội nghị năm 2013 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên (trang 1393-1398).

Phụ lục

Bảng 6. Bộ dữ liệu về độ tương tự văn bản ngữ nghĩa trong quá khứ

Corpus	Bộ dữ liệu ngôn ngữ		Cặp
		MSRpar	1500

SemEval -2012 Nhiệm vụ 6 (Agirre và cộng sự, 2012)	Tiếng Anh	MSRvid	1500
		OnWN	750
		tin tức SMT	750
		SMTeuroparl	750
		TỔNG	5250
Nhiệm vụ chia sẻ SEM 2013 (Agirre và cộng sự, 2013a)	Tiếng Anh	HDL	750
		FNWN	189
		OnWN	561
		SMT	750
		TỔNG	2250
SemEval -2014 Nhiệm vụ 10 (Agirre và cộng sự, 2014)	Tiếng Anh	HDL	750
		OnWN	750
		diễn đàn Deft	450
		tin tức Deft	300
		Hình ảnh	750
		Tweet-tin tức	750
		TỔNG	3750
	tiếng Tây Ban Nha	Wikipedia	324
		Tin tức	480
		TỔNG	804
SemEval -2015 nhiệm vụ 2 (Agirre và cộng sự, 2015)	Tiếng Anh	HDL	750
		Hình ảnh	750
		sinh viên trả lời	750
		diễn đàn trả lời	375
		Sự tin tưởng	375
		TỔNG	3000
	tiếng Tây Ban Nha	Wikipedia	251
		Tin tức	500
		TỔNG	751

SemEval -2016 nhiệm vụ 1 (Agirre và cộng sự, 2016)	Tiếng Anh	HDL	249
		Đạo văn	230
		Đang đăng bài	244
		Trả lời-Trả lời.	254
		Nhiệm vụ.-Quest.	209
		TỔNG	1186
	Tây Ban Nha-Anh	Sự thử nghiệm	103
		Tin tức	301
		Đa nguồn	294
		TỔNG	698
BỆNH (Marco Marelli và cộng sự, 2014)	Tiếng Anh	Kỳ đánh giá 2012 MSRvid Hình ảnhFlick	9840
SemEval -2017 nhiệm vụ1 (Cer và cộng sự, 2017)	Tiếng Anh	sự đánh giá	250
		Sự thử nghiệm	23
		TỔNG	273
	tiếng Tây Ban Nha	sự đánh giá	250
		Sự thử nghiệm	23
		TỔNG	273
	tiếng Ả Rập	sự đánh giá	250
		Sự thử nghiệm	23
		MSRpar	510
		MSRvid	368
		SMTeuroparl	203
		TỔNG	1354
	Tây Ban Nha-Anh	sự đánh giá	500
		Sự thử nghiệm	23
		MT	1000
		TỔNG	1523

	Tiếng Ả Rập-Anh	sự đánh giá	250
		Sự thử nghiệm	23
		MSRpar	1020
		MSRvid	736
		SMTeuroparl	406
		TỔNG	2435
	Thổ Nhĩ Kỳ-Anh	sự đánh giá	250