ĐO ĐỘ TƯƠNG TỰ NGỮA NGHĨA CỦA CẶP NGÔN NGỮ ANH-VIỆT THEO MÔ HÌNH PHÂN PhỐI NGỮ NGHĨA SONG NGỮ

Bùi Văn Tân[1], Nguyễn Phương Thái[2], Đinh Khắc Quý[2]

[1]Trường Đại học Kinh tế Kỹ thuật Công nghiệp
[2]Trường Đại học Công nghệ, Đại học Quốc Gia Hà Nội

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. Những nghiên cứu được công bố gần đây thường giải quyết bài toán này cho đơn ngữ. Gần đây, chứng kiến sự gia tăng không ngừng số lượng những ứng dụng xử lý tự nhiên đa ngôn ngữ, đặt ra yêu cầu cần có các kỹ thuật đo lường độ tương tự ngữ nghĩa song ngữ một cách hiệu quả. Trong bài viết này, chúng tôi trình bày một số kỹ thuật đo độ tương tự ngữ nghĩa song ngữ theo tiếp cận nhúng từ song ngữ; đề xuất một mô hình mạng nơron xây dựng không gian vector biểu diễn ngữ nghĩa song ngữ; xây dựng bộ dữ liệu chuẩn cho bài toán đo độ tương tự ngữ nghĩa song ngữ Việt-Anh. Cuối cùng, chúng tôi thực nghiệm và đánh giá các kỹ thuật trên bộ dữ liệu đã xây dựng.*

Từ khóa— *xử lý ngôn ngữ tự nhiên, độ tương tự ngữ nghĩa song ngữ, nhúng từ song ngữ.*

I. GIỚI THIỆU

Sự tương đồng về ngữ nghĩa giữa các từ (word 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, word similarity đóng vai trò then chốt trong nhiều tác vụ xử lý ngôn ngữ tự nhiên (NLP) 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).

Phần lớn các kỹ thuật được đề xuất cho bài toán word similarity là cho đơn ngữ, chúng thực hiện đo độ tương tự ngữ nghĩa của các cặp từ trong cùng một ngôn ngữ. Sự phát triển của những ứng dụng xử lý đa ngôn ngữ (multilingual) đặt ra yêu cầu đo lường độ tương tự ngữ nghĩa của các cặp từ song ngữ (Cross-Lingual Words Similarity- CLWS). Hiện nay, CLWS là một bài toán quan trọng có ứng dụng trong một số tác vụ xử lý ngôn ngữ tự nhiên như dịch máy (machine translation), tìm kiếm thông tin (information retrival) cũng như trong khai phá dữ liệu (data minning) [6].

Các kỹ thuật word similarity lượng giá mức độ giống nhau của hai từ, 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 ngữ nghĩa (word relatedness) quan tâm đến nhiều loại quan hệ khác nhau giữa các từ, ví dụ từ "*ô tô"* có liên quan ngữ nghĩa với 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. word similarity là trường hợp đặc biệt của word relatedness.

Nội dung tiếp theo của bài viết này được cấu trúc như sau: phần II trình bày một số kỹ thuật CLWS dựa trên kỹ thuật nhúng từ song ngữ (cross-lingual word embeddings); phần III, đề xuất mô hình mạng nơron xây dựng không gian vector biểu diễn ngữ nghĩa song ngữ; phần IV, đề xuất bộ dữ liệu đánh giá kỹ thuật CLWS cho cặp ngôn ngữ Việt-Anh; phần V, trình bày thực nghiệm trên cặp ngôn ngữ Việt-Anh; cuối cùng là phần phân tích, kết luận.

II. MỘT SỐ KỸ THUẬT CLWS DỰA TRÊN NHÚNG TỪ SONG NGỮ

Những kỹ thuật được đề xuất cho bài toán CLWS có thể được chia thành ba nhóm chính: thứ nhất, dựa trên Cơ sở tri thức (Knowledge-based), khai thác tri thức tự động từ các từ điển điện tử (Machine – Readable Dictionaries) như các từ điển đồng nghĩa, mạng từ (WordNet); thứ hai, dựa trên kho ngữ liệu (Corpus-based). Thứ ba, dựa trên nhúng từ song ngữ (cross-lingual word embeddings), những nghiên cứu được công bố gần đây cho thấy, đây là hướng tiếp cận đặc biệt hiệu quả cho bài toán CLWS. Trong bài viết này, chúng tôi trình bày một số kỹ thuật CLWS cho cặp ngôn ngữ Anh-Việt theo hướng cross-lingual word embeddings.

A. Monolingual Embedding Models

Những năm gần đây, phương pháp nhúng từ đơn ngữ hay word embeddings (Mikolov et al., 2013a; Pennington et al., 2014) nhận được sự quan tâm đặc biệt trong lĩnh vực NLP. 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 (Mikolov et al., 2013a). 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

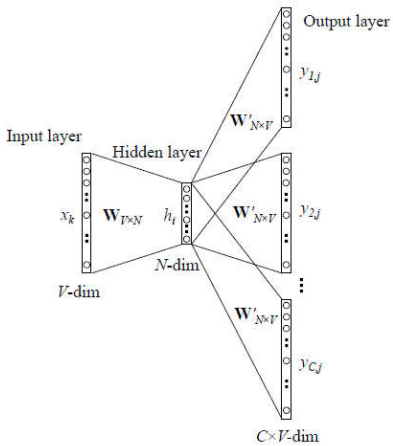
lân cận, với hai kiến trúc Skip-gram và Continuous bag of words (CBOW). Trong đó, kiến trúc Skip-gram (Hình 1) dự đoán (predict) các từ lân cận trong một cửa sổ ngữ cảnh (context window) bằng cách cực đại hóa trung bình logarit của các xác suất có điều kiện (công thức 1).

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

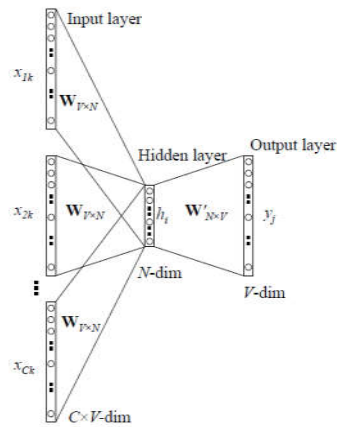
Trong đó $\{w_i : i \in T\}$ là toàn bộ tập huấn luyện, w_t là từ trung tâm và w_{t+j} là các từ trong cửa sổ ngữ cảnh. Xác suất có điều kiện được định nghĩa bằng hàm softmax (công thức 2).

$$p(w_j \mid w_I) = \frac{\exp({v'_{w_O}}^{T} v_{w_I})}{\sum \exp({v'_{w'_j}}^{T} v_{w_I})} \quad (2)$$

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 1. Kiến trúc Skip-gram



Hình 2. 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.

B. Cross-Lingual Word Embedding Models

Cross-lingual word embeddings (CLWE) là mô hình biểu diễn từ cho phép chúng ta biểu diễn ngữ nghĩa của từ trong ngữ cảnh đa ngôn ngữ, nó đóng vai trò chính trong tác vụ chuyển đổi tài nguyên giữa các ngôn ngữ (cross-lingual transfer knowledge) khi phát triển các ứng dụng NLP cho những ngôn ngữ có ít tài nguyên (low-resource languages). Gần đây, chứng kiến sự gia tăng không ngừng về số lượng những ứng dụng NLP trên dữ liệu đa ngôn ngữ, các ứng dụng này đòi hỏi cần có các mô hình CLWE hiệu quả. Các mô hình CLWE tạo ra không gian vector biểu diễn từ đa ngôn ngữ bằng kết nối các không gian vector biểu diễn từ đơn ngữ.

Mô hình Translation Matrix: do Mikolov và các cộng sự đề xuất năm 2013 (Mikolov et al., 2013b) dựa trên tiếp cận xây dựng ánh xạ tuyến tính (Mapping-based approaches). Nghiên cứu này đã cho thấy, có sự tương đồng về quan hệ hình học trong không gian vector biểu diễn từ của các từ giữa những ngôn ngữ khác nhau. Ví dụ, một số từ thuộc chủ đề động vật trong tiếng Anh được biểu diễn bởi tập hợp điểm như trong tiếng Tây Ban Nha (Hình 3). Điều này cho thấy rằng, chúng ta có thể chuyển đổi không gian vector biểu diễn từ của ngôn ngữ nguồn *s* tới không gian vectors biểu diễn từ của ngôn ngữ đích *t* thông qua việc học một ánh xạ tuyến tính là một ma trận chuyển (transformation matrix) *W*. Họ sử dụng 5000 cặp từ song ngữ phổ biến trong hai ngôn ngữ nguồn và đích. Sau đó học ma trận *W* sử dụng thuật toán giảm gradient để cực tiểu hóa hàm lỗi bình phương nhỏ nhất (mean squared error, MSE).

$$\Omega_{MSE} = \sum_{i=1}^{n} \| Wx_i^s - x_i^t \|^2 \quad (3)$$

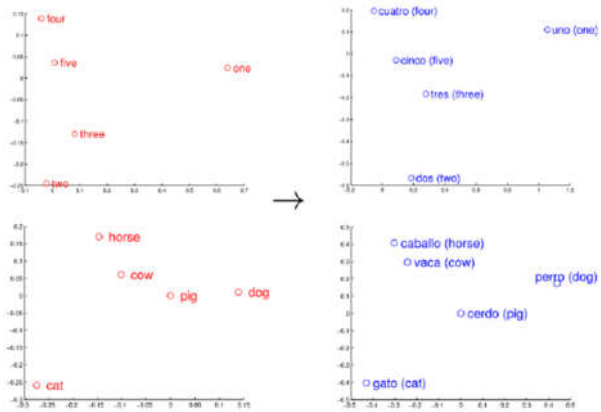
Trong đó X_s và X_t là hai không gian vector biểu diễn các từ mồi trong ngôn ngữ nguồn và ngôn ngữ đích. Trong nghiên cứu của (Xing et al., 2015) đã chỉ ra rằng, kết quả học ma trận tối ưu W^* được cải thiện đáng kể nếu bổ xung

ràng buộc trực giao cho ma trận W ($W.W^T$ = I). Trong trường hợp này, việc tìm W^* quy về giải bài toán trực giao Procrustes. Lời giải tối ưu có thể đạt được thông qua phép phân tích ma trận singular value decomposition (SVD) (công thức 4).

$$W^* = \arg\min_{W \in O_d(R)} \| WX_S - X_t \|_F = UV^T \quad (4)$$

Với $U \sum V^T = SVD(X_S X_T)$

Mô hình Bilingual Bag-of-Words (BilBOWA): do Gouws và cộng sự đề xuất năm 2015 (Gouws et al., 2015), mô hình BilBOWA không dùng dữ liệu từ gióng hàng từ (word alignments), nó là một mở rộng của skip-gram negative sampling (SGNS) để học CLWE. Thay vì dùng dữ liệu cặp từ song ngữ đã được gióng hàng, mô hình này giải thiết mỗi từ trong một câu nguồn sẽ được gióng với mọi từ trong câu đích dưới một mô hình gióng hàng thống nhất (uniform alignment model). Do đó, mô hình này thuộc nhóm dựa trên dữ liệu gióng hàng ở mức câu (Sentence-Level Alignment Methods).



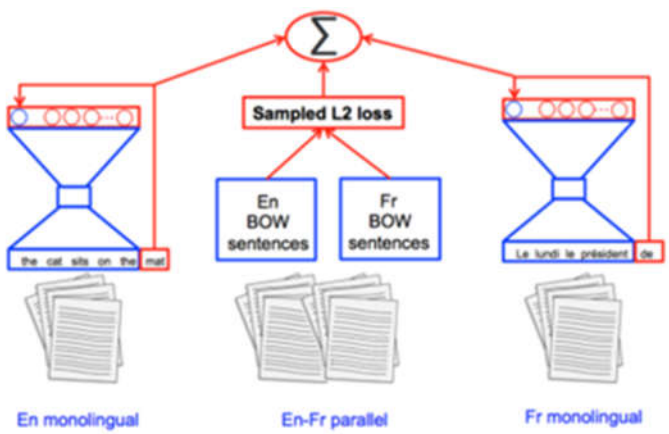
Hình 3. Mô phỏng sự giống nhau về tương quan hình học giữa các từ thuộc chủ đề động vật trong tiếng Anh và Tây Ban Nha [3]).

Thay vì cực tiểu hóa khoảng cách giữa từ đã được gióng hàng, mô hình này cực tiểu hóa khoảng cách trung bình giữa các biểu diễn từ các trong câu đã được gióng hàng. Hàm mục tiêu của BiBOWA được xác định như công thức 5.

$$\Omega_{BILBOWA} = \| \frac{1}{m} \sum_{w_i^s \in sent^s}^{m} x_i^s - \frac{1}{n} \sum_{w_j^t \in sent^t}^{n} x_j^t \|^2 \quad (5)$$

Trong đó x_i^s và x_j^t là các vector embeddings của từ w_i^s và w_j^t trong mỗi câu $sent^s$ và $sent^t$ trong ngôn ngữ s và t. Sử dụng SGNS như hàm mục tiêu cho nhúng từ đơn ngữ, BilBOWA cực tiểu hàm lỗi như trong công thức 6.

$$J = L_{SGNS}^{s \to t} + L_{SGNS}^{t \to s} + \Omega_{BILBOWA} \quad (6)$$



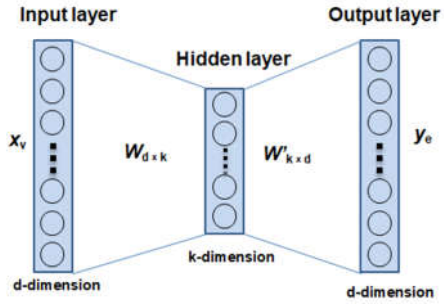
Hình 4. Mô hình BilBOWA [2]

Mô hình BiSkip: được đề xuất bởi Luong và các cộng sự (Luong et al., 2015), mô hình này sử dụng SGNS để dự đoán ngữ cảnh (contexts) của từ ở cả ngôn ngữ nguồn và đích. Khác với tiếp cận của BilBOWA, BiSkip sử dụng SGNS để dự đoán như mục tiêu song ngữ. Mô hình này được tối ưu theo hàm mất mát như sau:

$$J = L_{SGNS}^{s} + L_{SGNS}^{t} + L_{SGNS}^{s \to t} + L_{SGNS}^{t \to s} \quad (7)$$

III. MÔ HÌNH MẠNG NƠRON

Trong nghiên cứu này, chúng tôi đề xuất một mô hình mạng nơron gồm ba lớp để học một ánh xạ tuyến tính từ không gian vector nhúng từ tiếng Việt vào không gian vector những từ tiếng Anh. Kiến trúc của mạng nơron đề xuất trong nghiên cứu này được minh họa như hình 5, gồm ba lớp: lớp đầu vào (input layer) và lớp ẩn (hidden layer) có kích thước là *d*, lớp đầu ra (output layer) có kích thước *k*. Đầu vào nhận x_v là vector embedding của từ tiếng Việt, đầu ra là vector y_e biểu diễn cho từ trong tiếng Anh tương ứng với từ tiếng Việt đã được gióng hàng. Các trọng số giữa lớp input và hidden được biểu diễn bằng ma trận *d* hàng *k* cột (W_{dxk}), các trọng số giữa lớp hidden và lớp output được biểu diễn bằng ma trận *k* hàng *d* cột (W_{kxd}).



Hình 5. Kiến trúc mạng nơron được đề xuất

Cho cặp từ <*v*,*e*> trong *t* cặp từ Việt-Anh thuộc tập huấn luyện, x_v là vector biểu diễn từ *v* trong tiếng Việt, lớp hidden và lớp output được tính như sau:

$$h = \mathrm{Re}\,LU(x_t.W) \quad (8)$$

$$y_e = h.W' \quad (9)$$

Kiến trúc mạng được định nghĩa và các tham số được mô tả bằng giả mã như trong thuật toán 1. Chúng tôi sử dụng hàm lỗi Mean Squared Error (MSE) và thuật toán tối ưu Adam.

Thuật toán 1: thuật toán huấn luyện mạng

```
x= WE_V      # word embedding Vietnamese
y= WE_E      # word embedding English
N # number of loops
model = torch.nn.Sequential(
  torch.nn.Linear(D_in, H),
torch.nn.ReLU(),
torch.nn.Linear(H, D_out))
loss_fn = torch.nn.MSELoss(size_average=False)
learning_rate = 1e-5
optimizer = torch.optim.Adam(model.parameters(),
lr=learning_rate)
for t in range(N):
    y_pred = model(x)
    loss = loss_fn(y_pred, y)
    model.zero_grad()
    loss.backward()
    optimizer.step()
```

IV. XÂY DỰNG BỘ DỮ LIỆU TƯƠNG TỰ NGỮ NGHĨA SONG NGỮ

Bộ dữ liệu kiểm tra độ tương tự ngữ nghĩa song ngữ của từ (cross-lingual semantic word similarity dataset) đóng vai trò là công cụ để đánh giá các kỹ thuật CLWS. Mặc dù vậy, có ít nghiên cứu về CLWS cho tiếng Việt được công bố. Theo sự tra cứu của chúng tôi đối với các nghiên cứu về xử lý ngôn ngữ tự nhiên tính đến thời điểm hiện tại, chưa có nghiên cứu nào công bố các bộ dữ liệu đánh giá cho bài toán này. Do đó, chúng tôi thực hiện nghiên cứu và xây dựng bộ dữ liệu đánh giá các kỹ thuật CLWS cho cặp ngôn ngữ Việt-Anh (English-Vietnamese Words Smilarity Dataset - EVWSD).

Word similarity được thừa nhận rộng rãi trong việc lượng giá các mô hình không gian vector ngữ nghĩa (semantic vector space models) nói riêng và trong các kỹ thuật biểu diễn ngữ nghĩa nói chung (semantic representation techniques). Một trong những vấn đề cốt lõi khi đánh giá các kỹ thuật word similarity là không có một phép đo chính xác tuyệt đối cho các kỹ thuật này. Tính tương tự được đánh giá bằng thang đo sự đồng thuận của con người. Do đó, sự tương tự về ngữ nghĩa có thể thay đổi theo ngữ cảnh, nền tảng văn hóa, nhận thức chủ quan của con người hoặc theo thời gian.

A. Lựa chọn các cặp từ song ngữ

Tham khảo bộ dữ liệu tương tự ngữ nghĩa song ngữ cho cặp ngôn ngữ Anh-Pháp được công bố trong SemEval-2017 về *Multilingual and Cross-lingual Semantic Word Similarity* (Camacho-Collados et al., 2017) và bộ dữ liệu Vsim400 do Kim Anh Nguyen và công sự công bố (Nguyen et al., 2018). Chúng tôi tiến hành xây dựng bộ dữ liệu VEsim400 với 400 cặp từ Việt-Anh để đánh giá kỹ thuật CLWS cho cặp ngôn ngữ này. Các cặp từ Anh-Việt được chọn lựa dựa trên nguyên tắc:

- Là các từ được sử dụng phổ biến, có tần số xuất hiện cao trong các kho ngữ liệu đơn ngữ.

- Hạn chế dùng các từ đa nghĩa.

- Các từ trong cùng một cặp cùng từ loại và thuộc một trong ba từ loại danh từ, tính từ hoặc động từ.

- Đối với từ tiếng Việt, chúng tôi ưu tiên chọn từ thuần Việt, từ đơn âm tiết (so với đa âm tiết).

-Bộ dữ liệu gồm 400 cặp từ, trong đó 200 cặp danh từ, 100 cặp động từ và 100 cặp tính từ.

Từ 1	Từ 2	Độ tương tự	Từ 1	Từ 2	Độ tương tự
dog	*chó*	9.00	*fly*	*bay*	9.10
dog	*dê*	4.50	*fly*	*bầu_trời*	6.87
cat	*mèo*	9.00	*hear*	*nghe*	9.10
language	*ngôn_ngữ*	9.70	*locate*	*định_vị*	8.20
language	*sách*	7.52	*reply*	*trả_lời*	9.00
language	*điện_thoại*	2.45	*smile*	*cười*	8.80
bird	*gà_trống*	6.36	*search*	*tìm_kiếm*	9.40
bird	*chim*	9.20	*sing*	*hát*	9.20
signature	*chữ_ký*	9.45	*happy*	*hạnh_phúc*	9.35
pillow	*gối*	8.55	*happy*	*buồn*	1.25
pillow	*giường*	7.20	*intelligent*	*giỏi*	9.10
fill	*lấp_đầy*	7.90	*intelligent*	*ngu_dốt*	1.75

Bảng 1. Một số cặp từ Việt-Anh trong bộ dữ liệu

B. Đánh giá độ tương đồng ngữ nghĩa các cặp từ

Bộ dữ liệu được chia thành bốn tập con rời nhau, mỗi tập gồm 100 cặp từ Việt-Anh. Mỗi tập con được 15 sinh viên chuyên ngành công nghệ thông tin đánh giá độ tương tự, đây là những người có kiến thức về ngôn ngữ, có trình độ tiếng Anh ở mức cơ bản. Việc đánh giá của mỗi cá nhân được diễn ra độc lập. Để thuận lợi cho người đánh giá, chúng tôi cung cấp cho họ bản dịch sang tiếng Việt của các từ tiếng Anh trong bộ dữ liệu. Thang đo độ đánh giá là từ 0 tới 10. Sau khi nhận được kết quả đánh giá từ 15 người, chúng tôi tổng hợp kết quả đánh giá. Cuối cùng, độ tương đồng ngữ nghĩa của mỗi cặp từ Việt-Anh sẽ là giá trị trung bình do 15 người đánh giá độc lập.

V. THỰC NGHIỆM

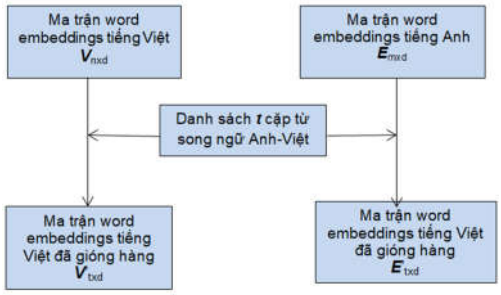
Trong nghiên cứu này, chúng tôi tiến hành hai thực nghiệm: (1-NN) mô hình mạng nơron đã đề xuất để học một ánh xạ tuyến tính từ không gian vector nhúng từ tiếng Việt vào không gian vector những từ tiếng Anh; (2-SVD) sử dụng phân tích ma trận SVD[1] để tính ma trận chuyển W^*. Để tạo ra mô hình nhúng từ đơn ngữ cho tiếng Việt với mô hình skip gram negative sampling, chúng tôi sử dụng một corpus gồm 21 triệu câu với khoảng 560 triệu token, sử dụng công cụ vnTokenizer để tách từ. Đối với nhúng từ tiếng Anh, chúng tôi sử dụng corpus BWLMB[2]. Các vector nhúng có số chiều là 300, thuật toán huấn luyện loại bỏ các từ xuất hiện ít hơn 50 lần trong corpus, kích thước cửa sổ context là 5, số mẫu negative (negative samples) là 10. Chúng tôi sử dụng 1000 cặp từ Anh-Việt phổ biến được lựa chọn từ điển Anh-Việt, Việt Anh[3], từ đó trích ra từ hai không gian nhúng từ đơn ngữ hai ma trận được gióng hàng như hình 6.

Mạng nơron trình bày trong phần III cài đặt sử dụng PyTorch, mạng này được huấn luyện để cực tiểu hóa hàm lỗi MSE sử dụng thuật toán tối ưu Adam. Tốc độ học $\alpha=10^{-5}$, số chiều vector d=300, số nút ẩn k=150.

[1] https://docs.scipy.org/doc/numpy-1.14.0

[2] https://code.google.com/archive/p/1-billion-word-language-modeling-benchmark/

3 https://github.com/



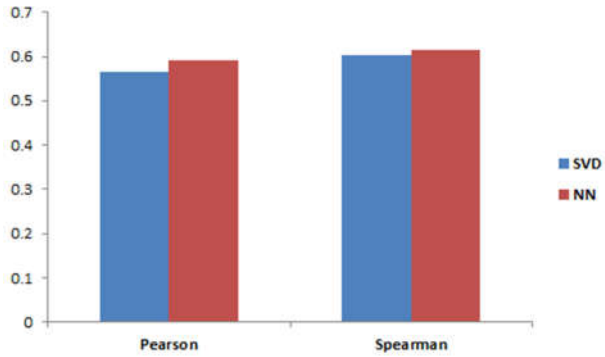
Hình 6. Sơ đồ tạo ma trận word embedding giống hàng

Để tính độ tương tự giữa các cặp từ, chúng tôi sử dụng độ đo khoảng cách cosine.

$$\cos ine(\vec{u},\vec{v}) = \frac{\vec{u}\cdot\vec{v}}{|\vec{u}||\vec{v}|} \quad (10)$$

Bảng 2. Độ tương tự ngữ nghĩa một số cặp từ được đo bởi kỹ thuật nhúng từ song ngữ

Từ 1	Từ 2	VEsim400	SVD	NN
dog	chó	9.00	9.33	8.56
dog	dê	4.50	3.40	3.55
cat	mèo	9.00	8.22	8.43
language	ngôn_ngữ	9.70	9.85	8.86
language	sách	7.52	3.20	5.75
language	điện_thoại	2.45	2.10	1.87
bird	gà_trống	6.36	2.80	4.73
bird	chim	9.20	5.60	6.40
signature	chữ_ký	9.45	4.90	5.80
pillow	gối	8.55	8.89	7.60
pillow	giường	7.20	2.10	5.50
fill	lấp_đầy	7.90	3.20	6.45
…				
Độ tương quan Pearson			0.564	0.592
Độ tương quan Spearman			0.603	0.614



Hình 7. Kết quả thực nghiệm với bộ dữ liệu VEsim400

Bảng 2 trình bày kết quả đo độ tương tự ngữ nghĩa trên một số cặp từ của bộ dữ liệu VEsim400, biểu đồ trong hình 7 biểu diễn trực quan hiệu quả của lược đồ cải tiến đã đề xuất. Kết quả thực nghiệm cho thấy rằng mạng nơron do chúng tôi đề xuất có khả năng sinh ra không gian vector biểu diễn từ song ngữ tốt hơn cho tác vụ đo lường độ tương tự ngữ nghĩa, so với hướng tiếp cận sử dụng phân tích ma trận SVD.

VI. KẾT LUẬN

Trong bài viết này, chúng tôi đã thực trình bày một số hướng tiếp cận cho bài toán CLWS, đề xuất một mô hình mạng nơron nhân tạo xây dựng không gian vector biểu diễn ngữ nghĩa song ngữ. Đặc biệt, chúng tôi đề xuất bộ dữ liệu VEsim400 để đánh giá các kỹ thuật CLWS cho cặp ngôn ngữ Việt-Anh. Trên cơ sở những nghiên cứu và thực nghiệm đã tiến hành, chúng tôi tiếp tục nghiên cứu bài toán đo lường độ tương tự ngữ nghĩa song ngữ dựa trên cross-lingual embeddings.

VII. 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] José Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. InSteven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel M.Cer, and David Jurgens, editors, SemEval@ACL, pages 15–26. Association for Computational Linguistics, 2017.

[2] Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In Francis R. Bach and David M. Blei, editors, ICML, volume 37 of JMLR Workshop and Conference Proceedings, pages 748–756. JMLR.org, 2015.

[3] Thang Luong, Hieu Pham, and Christopher D Manning. Bilingual word representations with monolingual quality in mind. In VS@ HLT-NAACL, pages 151–159, 2015.

[4] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. CoRR, abs/1309.4168, 2013a.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, NIPS, pages 3111–3119, 2013b.

[6] Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. Introducing two vietnamese datasets for evaluating semantic models of (dis-)similarity and relatedness. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, NAACL-HLT (2), pages 199– 205. Association for Computational Linguistics, 2018. ISBN 978-1-948087-29-2.

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In EMNLP, volume 14, pages 1532–1543, 2014.

[8] Yangyang Wu, Siying Wu, and Duansheng Chen. Chinese-english bilingual word semantic similarity based on chinese wordnet. JSW, 10(1):20–31, 2015.

[9] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In Rada Mihalcea, Joyce Yue 2 Chai, and Anoop Sarkar, editors, HLT-NAACL, pages 1006–1011. The Association for Computational Linguistics, 2015. ISBN 978-1-941643-49-5.

Cross-lingual Semantic Similarity via Cross-Lingual Embeddings

Bui Van Tan, Nguyen Phuong Thai, Dinh Khac Quy

ABSTRACT- *measuring semantic similarity between words is a core issue because important applications in natural language processing. Former study on this problem almost to solve on monolingual. Recently, there has been an increase in multilingual natural language processing applications that require there are powerful cross-lingual word semantic similarity methods. In this paper, we present cross-lingual semantic word similarity methods based on cross-lingual word embedding. We proposed a neural network model for constructing a cross-lingual word embeddings space. Construct a benchmark dataset for evaluating these methods on Vietnamese-English; the last, which is experimental on the proposed dataset.*