TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 20, Số 2 (2023): 232-243
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 20, No. 2 (2023): 232-243
https://doi.org/10.54607/hcmue.js.20.2.3618(2023)

Bài báo nghiên cứu

MÔ HÌNH DÓNG HÀNG MỨC TỪ DỰA TRÊN BERT CHO CẶP CÂU VIỆT – NHẬT

Lê Thanh Tùng*, Nguyễn Hồng Bửu Long, Hoàng Khuê
Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam
**Tác giả liên hệ: Lê Thanh Tùng – Email: tungleqb@gmail.com*
Ngày nhận bài: 11-10-2022; ngày nhận bài sửa: 08-11-2022; ngày duyệt đăng: 21-02-2023

TÓM TẮT

Dóng hàng mức từ giữ vai trò quan trọng trong nhiều công đoạn của xử lí ngôn ngữ tự nhiên. Có nhiều công trình nghiên cứu trên nhiều cặp ngôn ngữ khác nhau, tuy nhiên trên cặp câu song ngữ Nhật-Việt vẫn còn hạn chế. Hầu hết các dóng hàng mức từ Nhật-Việt được tạo từ các công cụ dóng hàng dựa trên phương pháp thống kê, hoặc dựa trên phương pháp học không giám sát, cho kết quả có độ chính xác không cao. Trong nghiên cứu này, chúng tôi xây dựng bộ ngữ liệu dóng hàng mức từ Nhật-Việt bằng tay và sau đó cài đặt và huấn luyện mô hình dóng hàng mức từ tự động cho cặp câu song ngữ Nhật-Việt. Mô hình dóng từ của chúng tôi đạt độ chính xác vượt trổi hơn 20.06 điểm F_1 so với công cụ GIZA++. Chúng tôi tạo được mô hình dóng hàng mức từ Nhật-Việt tân tiến ở thời điểm hiện tại.

Từ khóa: BERT; Nhật - Việt; Bộ ngữ liệu; SQuAD; dóng từ

1. Giới thiệu

Dóng hàng mức từ có vai trò quan trọng trong nhiều tác vụ xử lí ngôn ngữ tự nhiên như dịch, phân tích ngôn ngữ, đồng tham chiếu... Việc xây dựng bộ ngữ liệu song ngữ đã được dóng hàng mức từ mất rất nhiều thời gian và công sức nếu thực hiện một cách thủ công. Chính vì vậy, mô hình dóng hàng mức từ đã được quan tâm nghiên cứu trong một thời gian dài, trên nhiều cặp ngôn ngữ khác nhau. Trước đây, phương pháp tiếp cận thống kê đã được áp dụng trong nhiều mô hình như mô hình IBM 1-5, mô hình HMM… Các công cụ giống hàng GIZA (Franz Josef Och, 2003; Chris Dyer, 2013) được xây dựng dựa các mô hình trên.

Trong những năm gần đây, các mô hình dóng từ dựa trên mảng nơron đã được nghiên cứu và cho thấy cải thiện kết quả dóng từ so với phương pháp thống kê. (Nan Yang, 2013) đã đề xuất mô hình dóng từ dựa trên mảng neuron FFNN trên cặp câu Anh-Trung, (Akihiro Tamura, 2014) đề xuất mô hình dóng từ dựa trên mạng hồi quy trên cặp câu Anh-Nhật, (Joel Legrand, 2016) đã đề xuất mô hình dựa nơron trên các cặp Anh-Đức, Anh-Romania, Anh-

Cite this article as: Le Thanh Tung, Nguyen Hong Buu Long, & Hoang Khue (2023). BERT based word alignment model for Japanese –Vietnamese. *Ho Chi Minh City University of Education Journal of Science, 20*(2), 232-243.

Pháp. Các công trình trên đã được so sánh có độ chính xác cao hơn so với GIZA++. Gần đây, các mô hình dóng từ dựa trên Transformer (Ashish Vaswani, 2017) đã vượt trội hơn so với phương pháp học nông, như các nghiên cứu (Thomas Zenkel, 2019; Sarthak Garg, 2019; Elias Stengel-Eskin, 2019; Thomas Zenkel, 2020). Dựa trên các mô hình Transformer đã được huấn luyện nhằm dịch máy cho cặp ngôn ngữ nhất định, họ đã tìm cách khai thác mô hình Transformer cho việc dóng từ. Điều này dẫn đến yêu cầu phải có bộ ngữ liệu song ngữ lớn để huấn luyện mô hình Transformer trước khi khai thác để dóng từ. Cách tiếp cận này không phù hợp với các cặp ngôn ngữ còn hạn chế về ngữ liệu song ngữ như Nhật - Việt.

Đáng chú ý trong những công trình nghiên cứu về dóng từ là phương pháp của (Masaaki Nagata, 2020). Khác với các phương pháp dóng hàng trước đây đó, tác giả đã chuyển bài toán dóng hàng mức từ sang bài toán trả lời câu hỏi (Question Answering) với dạng dữ liệu SQuAD v2.0. Trong hệ thống trả lời câu hỏi, cho một đoạn nội dung và một câu hỏi, hệ thống sẽ dự đoán một phần câu (span) trong đoạn nội dung làm câu trả lời. Tương tự, trong bài toán dóng hàng mức từ khi cho một câu đích như là một đoạn nội dung và cho một từ trong câu nguồn như là câu hỏi, mô hình dự đoán bản dịch của từ nguồn trong câu đích như là câu trả lời. Tác giả đã thực nghiệm trên 5 cặp ngôn ngữ (Zh-En, Ja-En, De-En, Ro-En, En-Fr) và cho thấy rằng chỉ cần một số lượng nhỏ dữ liệu dóng từ bằng tay, mô hình cho kết quả vượt trội hơn so với các phương pháp từng công bố trước đây. Đây là có thể coi là phương pháp tân tiến nhất ở thời điểm hiện tại.

Đối với cặp song ngữ Nhật - Việt, ngữ liệu song ngữ sẵn có còn hạn chế, chưa có sẵn bộ ngữ liệu song ngữ dóng hàng mức từ, các mô hình dóng hàng mức từ theo cách tiếp cận học sâu chưa được thực hiện. Hầu như để thực hiện nhiệm vụ dóng hàng cho cặp song ngữ Nhật - Việt, các nhà nghiên cứu thường sử dụng Giza++ để giải quyết vấn đề trên. Tuy nhiên, Giza++ không cho kết quả cao với bộ ngữ liệu hạn chế như cặp song ngữ Nhật - Việt.

Nhận thấy được sự thiếu sót và tầm quan trọng của bài toán nêu trên, chúng tôi đã xây dựng bộ ngữ liệu dóng hàng mức từ bằng tay gồm 2646 cặp câu và cài đặt, huấn luyện mô hình dóng hàng tự động mức từ tân tiến cho cặp song ngữ Nhật - Việt.

Chúng tôi cài đặt theo (Masaaki Nagata, 2020) để áp dụng cho mô hình dóng hàng mức từ cặp câu Nhật - Việt. Mô hình dóng hàng mức từ được huấn luyện trên bộ ngữ liệu dóng từ bằng tay cặp câu Nhật - Việt mà chúng tôi xây dựng.

2. Nội dung

2.1. Cơ sở lí thuyết

2.1.1. Bài toán dóng hàng mức từ

Chúng tôi dựa theo (Franz Josef Och, 2003) để định nghĩa bài toán dóng hàng mức từ như là việc tìm ra các cặp từ tương đồng trong cặp câu đã cho. Bài toán dóng hàng mức từ (sau đây gọi là dóng từ) được định nghĩa như sau:

Cho vectơ $s_1^N = [s_1, s_2, \ldots, s_j, \ldots, s_N]$ chứa các từ theo thứ tự trong câu nguồn (tiếng Nhật) và vectơ $t_1^M = [t_1, t_2, \ldots, t_i, \ldots, t_M]$ chứa các từ theo thứ tự trong câu đích (tiếng Việt),

hai câu này hoàn toàn là bản dịch của nhau. Chúng ta cần tìm ra tập các cặp từ giữ câu nguồn và câu đích tương đồng với nhau. Như vậy, dóng từ là một tập các cặp từ:

$$a \subseteq \{(i,j): j = 1,2,\dots,N; i = 1,2,\dots,M\} \quad (1)$$

Để đơn giản, chúng tôi biểu diễn ánh xạ dóng hàng một cặp từ (i,j) thành $j = a_i$. Hơn nữa, việc tìm ra đầy đủ một tập các cặp từ như thế là rất khó, do đó việc biểu diễn dóng từ có thêm các ràng buộc sau:

Ràng buộc 1: Mọi từ trong câu đích đều phải được dóng với một từ trong câu nguồn, a_i có thể nhận giá trị 0 để biểu diễn từ thứ i trong câu đích được dóng với từ "null" trong câu nguồn;

Ràng buộc 2: Một từ trong câu đích chỉ được dóng với duy nhất một từ trong câu nguồn.

Như vậy, công thức (1) trở thành:

$$a_1^M = [a_1,\dots,a_M] \quad (2)$$

với $a_i \in \{0,1,2,\dots,N\}, i \in \{1,\dots,M\}$.

Dóng từ là bài toán có hướng, hướng từ câu nguồn sang câu đích và ngược lại. Trong bài toán dóng từ quan hệ giữa từ trong câu đích và từ trong câu nguồn là quan hệ một - nhiều.

	nen	ni	hitori	no	amerika	jin	ga	kono	konkuuru	ni	nyuushou	shi	te	iru
t_sent: 1980	年	に	一人	の	アメリカ	人	が	この	コンクール	に	入賞	し	て	いる

s_sent: Vào năm 1980 , một người Mĩ đã giành giải quán_quân trong cuộc thi này .

a: [3, 2, 1, 0, 4, 7, 6, 0, 12, 12, 12, 11, 10, 10, 9, 16]

Hình 1. *Ví dụ cặp câu Nhật-Việt được dóng từ theo chiều tiếng Việt đến tiếng Nhật*

Hình 1 là một ví dụ trong ngữ liệu dóng hàng của chúng tôi xây dựng (trình bày ở phần sau). Trong đó, câu nguồn tiếng Việt và câu đích tiếng Nhật được tách thành các từ và vectơ dóng từ theo chiều từ tiếng Việt đến tiếng Nhật. Từ tiếng Việt đầu tiên trong câu nguồn được đánh chỉ số từ số 0 tương ứng với vị trí 0 trong vectơ dóng từ, còn từ tiếng Nhật đầu tiên được đánh chỉ số từ số 1 (chỉ số 0 dành cho từ "null"). Trong ví dụ này, vectơ dóng từ $a[1]$ có giá trị là 2 có nghĩa là từ "năm" trong câu tiếng Việt được nối với từ "年" (nen) trong câu tiếng Nhật.

2.1.2. Chuyển bài toán dóng từ sang bài toán trả lời câu hỏi

(Masaaki Nagata, 2020) đã xây dựng mô hình dóng từ dựa trên dữ liệu hệ thống trả lời câu hỏi dạng SQuAD (Pranav Rajpurkar, 2018). Trong hệ thống trả lời câu hỏi, cho một đoạn văn và một câu hỏi, mô hình dự đoán một đoạn trong đoạn văn làm câu trả lời. Tương tự trong mô hình dóng từ, cho một câu đích như là đoạn văn và một từ trong câu nguồn cùng với câu nguồn đó như là câu hỏi, mô hình dự đoán từ trong câu đích có nghĩa giống từ nguồn như là câu trả lời.

Hình 2 cho thấy sự chuyển đổi dữ liệu dóng hàng sang mô hình dạng SQuAD v2.0. Trong đó, "t_sent" là câu đích như là đoạn văn, "s_sent" là câu nguồn như là câu hỏi. Từ

"Pháp" trong câu nguồn nằm giữa 2 dấu ¶ (đánh dấu đoạn)[2] để đánh dấu từ cần tìm bản dịch. Câu trả lời "answer" là vị trí của từ trong câu đích "フランス" (furansu). Chúng tôi sử dụng toàn bộ như là câu hỏi nhằm cung cấp ngữ cảnh của từ nguồn để mô hình dự đoán tốt hơn. Các trường thông tin "t_pos", "t_ner", "s_pos", "s_ner", "id" lần lượt là từ loại, thực thể có tên của câu hỏi, từ loại, thực thể có tên của từ nguồn và mã định danh cho dữ liệu nhằm phục vụ trong quá trình huấn luyện mô hình.

Vì có nhiều dóng từ với "null" nên phiên bản SQuAD v2.0 là phù hợp. Để chuyển đổi sang dữ liệu dạng SQuAD v2.0 (Pranav Rajpurkar, 2018) với mỗi cặp câu Việt-Nhật chúng tôi lần lượt cho một câu làm ngữ cảnh đoạn văn, câu còn lại sẽ chọn từng từ đánh dấu với ¶ để làm câu hỏi.

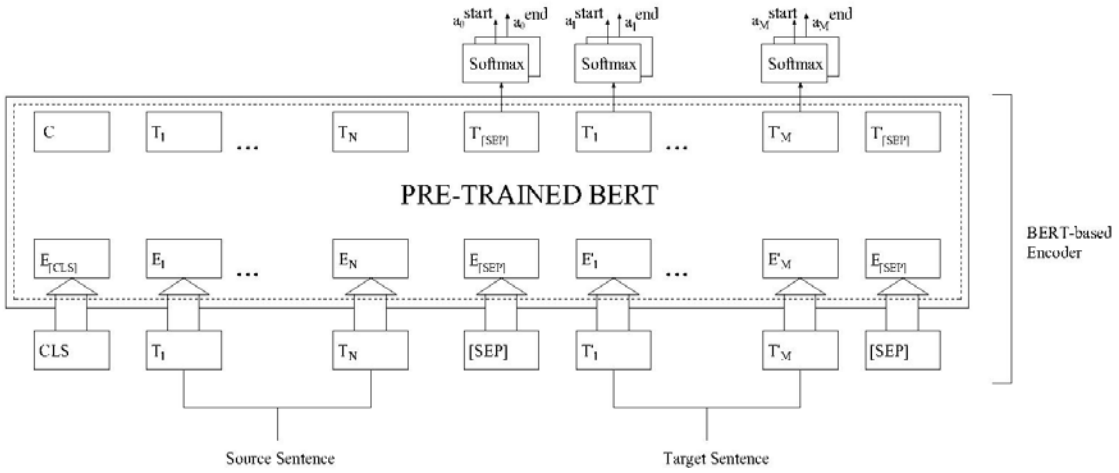

```
{                    nen   de furansu   go  ni fukutatsu su   ru   no wa  kiwamete
musukashii
   "t_sent": "2, 3 年 で フランス 語 に 熟達 す る の は きわめて 難しい
。",
   "s_sent": "Rất là khó để thành_thạo tiếng ¶ Pháp ¶ trong hai hay ba năm .",
   "s_word": "Pháp",
   "answer": 4
   "t_pos": "名詞 名詞 助詞 名詞 名詞 助詞 名詞 動詞 助詞 助詞 副詞 形容詞 補助記号",
   "t_ner": "DAT DAT O LOC O O O O O O O O O",
   "s_pos": "Np",
   "s_ner": "I-MISC",
   "id": "GoldData2646.2296trainP1_2245_ja_6_8",
 }
```


Hình 2. *Dữ liệu dóng từ chuyển sang dữ liệu dạng SQuAD v2.0*

2.1.3. Mã hoá và tích hợp ngữ cảnh

Chúng tôi sử dụng mô hình BERT đa ngôn ngữ (Jacob Devlin, 2019) trong kiến trúc dóng từ bởi nó được tiền huấn luyện trên nhiều ngôn ngữ bao gồm cả tiếng Việt và tiếng Nhật. Cụ thể, lớp mã hoá đầu vào sử dụng phiên bản $\text{BERT}_{\text{BASE}}$ gồm 12 lớp Transformer, kích thước các lớp ẩn là 768, gồm 12 lớp attention. Bộ tách từ của BERT tách các từ trong câu gốc thành các từ con (subword unit).

[2] Chúng tôi sử dụng ¶ để đánh dấu vì nó thuộc bảng mã Unicode và nó có trong từ điển BERT đa ngôn ngữ và hiếm xuất hiện trong các văn bản thông thường.



Hình 3. *Kiến trúc mô hình dóng từ với dữ liệu dạng SQuAD v2.0*

Giả sử, chúng ta có câu nguồn gồm N từ gốc $s_1^N = [s_1, s_2, \ldots, s_j, \ldots, s_N]$ được tách thành $s_1^X = [bs_1, bs_2, \ldots, bs_j, \ldots, bs_X]$ chứa X từ con và câu đích gồm M từ gốc $t_1^M = [t_1, t_2, \ldots, t_i, \ldots, t_M]$ được tách thành $t_1^Y = [bt_1, bt_2, \ldots, bt_i, \ldots, bt_Y]$ chứa Y từ con với $X \geq N$, $Y \geq M$. Với một từ trong câu nguồn đã được tách thành các từ con liên tục $s_{i:j} = bs_i \ldots bs_j$ trong câu nguồn, mô hình sẽ tìm ra từ $t_{l:k} = bt_l \ldots bt_k$ trong câu đích.

Để mã hoá đoạn văn và câu hỏi trước khi đưa vào BERT, chúng được nối với nhau như sau "[CLS] câu nguồn [SEP] câu đích [SEP]". Sau khi nối với nhau chuỗi đầu vào trở thành:

$$w_1^z = [[CLS], bs_1, \ldots, bs_i, \ldots, bs_X, [SEP], bt_1, \ldots, bt_j, \ldots, bt_Y, [SEP]] \quad (3)$$

Lớp mã hoá BERT sẽ sinh ra đặc trưng tích hợp e_r cho mỗi từ con thứ r như sau:

$$e_r = BERT_{base}(bw_1^z, r) \quad (4)$$

2.1.4. Dòng hàng mức từ

Chúng tôi theo đề xuất của (Masaaki Nagata, 2020) sử dụng kiến trúc mô hình trả lời câu hỏi dạng dữ liệu SQuAD v2.0 cho nhiệm vụ dóng từ. Để dự đoán vị trí bắt đầu và kết thúc cho câu trả lời là khác với (Masaaki Nagata, 2020), bổ sung phía sau BERT hai lớp Softmax.

Điểm số xác suất của vị trí bắt đầu và vị trí kết thúc của câu trả lời được tính như sau:

$$a_r^{start} = softmax(e_r) \quad (5)$$

$$a_r^{end} = softmax(e_r) \quad (6)$$

2.2. Xây dựng bộ ngữ liệu

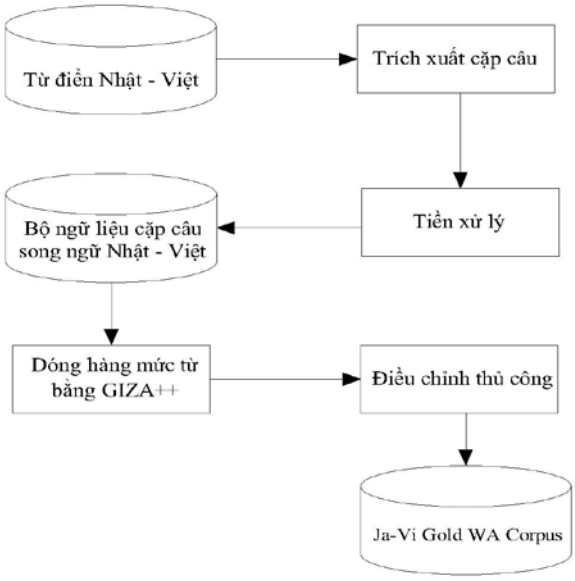
Trong nghiên cứu này, chúng tôi xây dựng bộ ngữ liệu dóng từ bằng tay 2646 cặp câu song ngữ Nhật-Việt. Chúng tôi đặt tên bộ ngữ liệu của chúng tôi là Ja-Vi Gold WA Corpus. Hình 4 mô tả quy trình tạo bộ ngữ liệu gồm 4 bước, cụ thể các bước của quy trình trên như sau:

2.2.1. Trích xuất cặp câu song ngữ Nhật-Việt

Để có ngữ liệu song ngữ cặp câu Nhật - Việt chúng tôi trích xuất từ các cặp mẫu câu trong các bộ từ điển Nhật - Việt. Trong các bộ từ điển, các mẫu câu chủ yếu là các câu thường dùng trong đời sống hằng ngày và mỗi cặp câu đều là bản dịch của nhau. Chúng tôi đã trích xuất được hai bộ ngữ liệu song ngữ khác nhau được đặt tên là Ja-Vi Giza Corpus và Ja-Vi Gold Sent Corpus. Các bộ ngữ liệu được sử dụng trong các mục đích khác nhau trong bài báo này. Các bộ ngữ liệu với kích thước được mô tả theo bảng sau:

Bảng 1. *Thông tin về các bộ ngữ liệu song ngữ*

Thông tin \ Bộ ngữ liệu	Giza Corpus	Ja-Vi Gold WA Corpus
Số lượng cặp câu	**1562527 cặp** (gồm cặp câu và cặp từ)	**2646 cặp câu**



Hình 4. *Quy trình xây dựng bộ ngữ liệu dóng từ*

Giza Corpus là bộ ngữ liệu nhằm mục đích để đưa vào công cụ dóng từ GIZA++ ở bước thứ 3 trong quy trình xây dựng bộ ngữ liệu dóng từ. Giza Corpus chứa các cặp câu và cặp từ song ngữ. GIZA++ là mô hình dóng từ dựa trên thống kê, do đó khi đưa các cặp từ song ngữ vào dữ liệu đầu vào, mô hình thống kê sẽ thống kê chính xác hơn xác suất dóng từ của mỗi cặp từ trong các cặp câu.

Ja-Vi Gold WA Corpus là bộ ngữ liệu vàng được dóng từ mà chúng tôi xây dựng. Được sử dụng để huấn luyện mô hình dóng từ Nhật - Việt. Bộ ngữ liệu chứa 2646 cặp câu được chọn từ bộ ngữ liệu Giza Corpus.

2.2.2. Tiền xử lí văn bản

Bước tiền xử lí thực hiện tách từ, gán nhãn từ loại, nhận dạng thực thể có tên (NER) cho mỗi câu trong các cặp từ. Chúng tôi sử dụng công cụ VnCoreNLP (Vu Thanh, 2018) để xử lí cho tiếng Việt. Đối với câu tiếng Nhật, chúng tôi sử dụng công cụ Mecab (Toshinori Sato & Okumura, 2017) để tách từ, gán nhãn từ loại và sử dụng dịch vụ của trang web dịch vụ goo labs[3] để nhận dạng thực thể có tên. Đối với số đếm, để đồng nhất, chúng tôi sử dụng kí tự unicode cho số đếm cả tiếng Nhật và tiếng Việt.

2.2.3. Dóng hàng bằng Giza++

Sau khi tiền xử lí cho bộ ngữ liệu Giza Corpus, chúng tôi sử dụng công cụ GIZA++ để dóng từ cho bộ ngữ liệu. Các thông số của GIZA++ được đặt ở giá trị mặc định và thực hiện dóng từ độc lập 02 chiều từ tiếng Nhật sang tiếng Việt và từ tiếng Việt sang tiếng Nhật. Kết quả dóng từ bằng GIZA++ được sử dụng để điều chỉnh ở bước tiếp theo giúp giảm chi phí và tăng độ chính xác khi điều chỉnh thủ công.

2.2.4. Điều chỉnh thủ công

Chúng tôi điều chỉnh thủ công những dóng hàng sai trong kết quả dóng hàng bằng GIZA++ để tạo ra bộ ngữ liệu dóng từ vàng. Đây là bước tốn nhiều chi phí nhân công, chúng tôi được các công tác viên là những người có chứng chỉ năng lực tiếng Nhật JLPT N2 thực hiện điều chỉnh thủ công trên từng cặp từ.

Chúng tôi dựa theo (João Graça, 2008) để xác định liên kết từ giữa các cặp từ. Một số ví dụ như sau:

- Đối với những đoạn câu mà có trong câu này những không bản dịch trong câu kia thì đoạn câu đó được liên kết với "null".
- Đối với dấu câu: Trong tiếng Nhật kết thúc câu hỏi thường sử dụng dấu chấm hỏi "？" hoặc dấu chấm câu "。" do đó chúng tôi xác định liên kết giữa dấu chấm hỏi và chấm câu trong tiếng Nhật với dấu chấm câu trong tiếng Việt là liên kết đúng.
- Đối với thời gian: Trong câu tiếng Nhật nếu sử dụng chữ cái romaji để biểu diễn con số thì sẽ được chuyển sang chữ số mã unicode đồng nhất với tiếng Việt.
- Đối với các đoạn dịch các cụm từ: chúng tôi xác định liên kết các từ đến từ có nghĩa gần nhất trong bản dịch là hợp lệ.

2.3. Thực nghiệm

2.3.1. Cài đặt mô hình

Chúng tôi sử dụng BERT-base[4], Multilingual Cased (104 ngôn ngữ, 12 lớp, 768 trạng thái ẩn, 12 lớp đầu ra, 110 triệu tham số huấn luyện) trong thực nghiệm của chúng tôi. Hai lớp Softmax bổ sung phía sau BERT có kích thước 768×1. Các tham số được cài đặt để cho phép tinh chỉnh.

Trong quá trình huấn luyện, thuật toán tối ưu AMSGrad (Sashank, 2018) với các tham số theo mặc định. Chúng tôi sử dụng phương pháp làn truyền ngược (*back propagation)* để

[3] https://labs.goo.ne.jp
[4] https://huggingface.co/bert-base-cased

điều chỉnh các tham số của mô hình. Hàm mất mát (loss function) là hàm cross entropy được tính như sau:

$$loss = -log(\frac{1}{2}(a_{+}^{start} + a_{+}^{end})) \quad (7)$$

Trong đó, $a_{+}^{start}, a_{+}^{end}$ lần lượt là xác suất đầu ra của lớp Softmax của vị trí bắt đầu và kết thúc của subtoken của nhãn dóng hàng.

Các tham số huấn luyện như sau: train_batch_size = 32, learning_rate = 1e-5, num_train_epochs = 2, max_seq_length = 256. Chúng tôi huấn luyện mô hình trên công cụ nghiên cứu máy học của Google Colab[5] với GPU 12GB. Thời gian huấn luyện khoảng 10 tiếng đồng hồ. Chỉ huấn luyện một mô hình để dóng hàng theo hai chiều.

2.3.2. Dữ liệu huấn luyện

Bộ ngữ liệu dóng hàng mức từ Nhật-Việt (Ja-Vi Gold WA Corpus) mà chúng tôi đã giới thiệu ở trên được sử dụng để huấn luyện mô hình dóng từ. Bộ ngữ liệu được chia làm hai tập ngẫu nhiên gồm 2303 cặp câu cho tập huấn luyện và 343 cặp câu cho tập kiểm tra. Chúng tôi thực hiện việc chia bộ ngữ liệu 5 lần riêng biệt và huấn luyện mô hình dóng hàng riêng biệt, sau đó lấy kết quả trung bình để đánh giá.

Ngoài ra, bộ ngữ liệu song ngữ Nhật-Việt (Giza Corpus) được sử dụng cho công cụ GIZA++.

2.3.3. Mô hình cơ sở

Mô hình đề xuất được so sánh với công cụ GIZA++ và của mô hình dóng từ dựa trên BERT theo (Le, 2021).

Đối với công cụ GIZA++, các tham số được cài đặt ở giá trị mặc định. Thực hiện chạy trên bộ ngữ liệu Giza Corpus, theo hai chiều độc lập từ tiếng Nhật sang tiếng Việt và từ tiếng Việt sang tiếng Nhật. Kết quả GIZA++ sau đó được trích xuất để lấy kết quả dóng từ của 343 cặp thuộc tập kiểm tra.

Thực hiện cài đặt và huấn luyện mô hình dóng từ dựa theo (Le, 2021) cho cặp câu song ngữ Nhật - Việt, ngoại trừ việc đưa thêm tri thức ngữ liệu tiếng Hán. Sử dụng cùng bộ ngữ liệu huấn luyện với mô hình đề xuất. Huấn luyện 02 mô hình riêng biệt cho 02 chiều dóng từ.

2.3.4. Độ đo

Theo (Och & Ney, 2003) để tính $Precision$ và $Recall$ yêu cầu hai tập dóng từ S và P, trong đó tập S là các dóng từ chắc chắn giữa cặp từ và tập P là các dóng từ có thể chấp nhận được giữa cặp từ. Chú ý rằng $S \subseteq P$. Để tính $Precision$ và $Recall$ của một tập A các dóng từ dự đoán của mô hình, sử dụng công thức như sau:

$$Precision(A, P) = \frac{|A \cap P|}{|A|} \quad (8)$$

$$Recall(A,\ S) = \frac{|A \cap S|}{|S|} \quad (9)$$

[5] https://colab.research.google.com

Trong bộ ngữ liệu của chúng tôi xây dựng, các dóng từ có thể chấp nhận không được chấp nhận. Điều này có nghĩa là tập $P = \emptyset$, do đó cho phép chúng tôi đánh giá độ chính xác của mô hình dựa trên độ đo F_1 được tính theo công thức sau:

$$F_1 = 2 \times Precision \times Recall / (Precision + Recall) \quad (10)$$

2.4. Kết quả

Bảng 2 so sánh kết quả của mô hình đề xuất với các mô hình cơ sở. Trong cả hai chiều dóng từ, mô hình của chúng tôi vượt trội hơn trên cả bả độ đo so với các mô hình được sử dụng gần đây. Mô hình dựa theo theo (Le, 2021) yêu cầu phải huấn luyện các mô hình khác nhau cho các chiều dóng hàng khác nhau, trong khi mô hình của chúng tôi chỉ cần huấn luyện một lần cho cả hai chiều dóng hàng.

Mô hình của chúng tôi hơn 20.06 điểm số F_1 so với GIZA++, tuy nhiên thật không công bằng nếu so sánh mô hình học có giám sát với mô hình học theo không có giám sát. Mục tiêu của nghiên cứu này là thực nghiệm phương pháp học có giám sát.

Bảng 2. *So sánh kết quả giữa các mô hình dóng từ*

Chiều dóng hàng	Mô hình	***Precision***	***Recall***	F_1
Ja-Vi	GIZA++	64.89	64.08	64.48
	BERT for Word Alignmnet (Le, 2021)	77.45	78.18	77.81
	Mô hình chúng tôi đề xuất	**84.32**	**84.77**	**84.54**
Vi-Ja	GIZA++	63.66	62.53	63.09
	BERT for Word Alignmnet (Le, 2021)	76.20	75.83	76.01
	Mô hình chúng tôi đề xuất	**82.91**	**84.39**	**83.64**

Chúng ta thấy kết quả chiều dóng hàng từ tiếng Nhật sang tiếng Việt cho kết quả tốt hơn so với chiều ngược lại trên cả ba độ đo. Điều đó là do trong cặp câu song ngữ, thường câu tiếng Việt có số lượng từ ngắn hơn so với câu tiếng Nhật.

Đối với dóng hàng đối xứng, chúng tôi sử dụng ba giải thuật heuristic gồm phép hợp, phép lấy giao và phương pháp trung bình được đề xuất bởi (Masaaki Nagata, 2020). Trong đó, phương pháp trung bình sẽ dựa vào trung bình cộng của dóng từ giữa hai chiều của một cặp từ bất kì nếu vượt ngưỡng 0.4 thì xác định từ đó được dóng từ với nhau. Bảng 3 là kết quả dóng từ đối xứng trên mô hình của chúng tôi.

Để thấy được quy mô bộ ngữ liệu dóng từ mà chúng tôi xây dựng so với các bộ ngữ liệu dóng từ trên các cặp ngôn ngữ khác. Bảng 4 cho thấy kích thước các bộ ngữ liệu dóng từ của các cặp ngôn ngữ Anh-Trung (Xuansong Li, 2015), các cặp ngôn ngữ Anh - Nhật (Neubig, 2015), Anh - Đức (David Vilar, 2016), Anh-Romania (Och & Ney, 2003) và Anh-Pháp (Pedersen, 2003). Tất cả 05 cặp ngôn ngữ này đã được (Masaaki Nagata, 2020) sử dụng để huấn luyện mô hình dóng từ cho từng cặp ngôn ngữ. Bộ ngữ liệu dóng từ mà chúng tôi xây dựng có kích thước lớn thứ 2 chỉ sau cặp tiếng Anh - tiếng Trung, và có kính thước lớn hơn nhiều so với 04 cặp ngữ liệu còn lại.

Bảng 3. *Kết quả dóng hàng đối xứng trên mô hình của chúng tôi*

Giải thuật	Recall	Precisison	F1
Bidi_int (2 chiều với phép hợp)	82.31	86.54	84.37
Bidi_uni (2 chiều với phép giao)	84.22	85.13	84.67
Bidi_avg (2 chiều theo Masaaki Nagata, 2020)	88.72	68.39	77.23

Bảng 4. *Kích thước các bộ ngữ liệu dóng từ*

Cặp ngôn ngữ	Kích thước (Cặp câu)
tiếng Anh - tiếng Trung	6.099
tiếng Nhật - tiếng Việt	2.646
tiếng Anh - tiếng Nhật	1.235
tiếng Anh - tiếng Đức	508
tiếng Anh - tiếng Pháp	447
tiếng Anh - tiếng Romania	247

3. Kết luận

Chúng tôi đã xây dựng được bộ ngữ liệu dóng từ bằng tay với 2646 cặp câu song ngữ Nhật - Việt. Có kích thước lớn thứ 2 so với các bộ ngữ liệu dóng từ nổi tiếng. Chúng tôi đã cài đặt và huấn luyện mô hình dóng từ Nhật - Việt trên bộ ngữ liệu của chúng tôi. Mô hình của chúng tôi cho kết quả vượt trội so với các mô hình dóng từ cặp câu Nhật - Việt trước đây.

Trong thời gian tiếp theo, chúng tôi sẽ nghiên cứu cải thiện mô hình bằng cách sử dụng các tri thức ngôn ngữ như nhãn từ loại (POS), thực thể có tên (NER) để đưa vào mô hình. Chúng tôi cũng sẽ nghiên cứu sử dụng các mô hình ngôn ngữ khác thay vì dùng $\text{BERT}_{\text{base}}$.

Chúng tôi cũng sẽ nghiên cứu mở rộng dóng từ trên các cặp ngôn ngữ giữa tiếng Việt và các thứ tiếng khác.

❖ ***Tuyên bố về quyền lợi:*** *Các tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

TÀI LIỆU THAM KHẢO

Akihiro Tamura, T. W. (2014). Recurrent Neural Networks for Word Alignment Model. *In Proceedings of the ACL-2014*, (pp. 1470-1480).

Ashish Vaswani, N. S. (2017). Attention Is All You Need. *In Proceedings of the NIPS 2017*, (pp. 5998-6008).

Chris Dyer, V. C. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. *In Proceedings of the NAACL-HLT-2013*, (pp. 644-648).

David Vilar, M. P. (2016). AER: Do we need to “improve” our alignments? *In Proceedings of IWSLT-2006*, (pp. 2005-212).

Elias Stengel-Eskin, T. R. (2019). A Discriminative Neural Model for Cross-Lingual Word Alignment. *In Proceedings of the EMNLP-IJCNLP-2019*, (pp. 910-920).

Franz Josef Och, a. H. (2003, 3). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics, 29*.

Jacob Devlin, M.-W. C. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the NAACL-2019*, (pp. 4171-4186).

João Graça, J. P. (2008). Building a Golden Collection of Parallel Multi-Language Word Alignment. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).* Marrakech, Morocco: European Language Resources Association (ELRA).

Joel Legrand, M. A. (2016). Neural Network-based Word Alignment through Score Aggregation. *In Proceedings of the WMT-2016*, (pp. 66-73).

Josef, F., & Ney, H. (2003, 3). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics, 29.*

Le H. B., T. V. (2021). Automatic Word Alignment For English-Vietnamese Bilinguals Corpus Using A Deep Learning Approach. *FAIR2021: Fundamental and Applied Information Technology*, (pp. 491-498). Ho Chi Minh.

Masaaki Nagata, K. C. (2020). A Supervised Word Alignment Method based on Cross-Language Span Prediction using Multilingual BERT. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 555-565). Association for Computational Linguistics.

Nan Yang, S. L. (2013). Word Alignment Modeling with Context Dependent Deep Neural Network. *In Proceedings of the ACL-2013*, (pp. 166-175).

Neubig, G. (2015). *Kyoto Free Translation Task alignment data package.* http://www.phontron.com/kftt/.

Och, F. J., & Ney, H. (2003, 3). A Systematic Comparison of Various Statistical Alignment Models. *Comput. Linguist., 29*, 19-51.

Pedersen, R. M. (2003). An Evaluation Exercise for Word Alignment. *In Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, (pp. 1--10).

Pranav Rajpurkar, R. J. (2018). Know What You Don’t Know: Unanswerable Questions for SQuAD. *In Proceedings of the ACL-2018*, (pp. 784-789).

Sarthak Garg, S. P. (2019). Jointly Learning to Align and Translate with Transformer Models. *In Proceedings of the EMNLP-IJCNLP-2019*, (pp. 4452-4461).

Sashank J. Reddi, S. K. (2018). On the Convergence of Adam and Beyond. *International Conference on Learning Representations (ICLR) 2018.* Vancouver Canada.

Thomas Zenkel, J. W. (2019). Adding Interpretable Attention to Neural Translation Models Improves Word Alignment. *ArXiv:1901.11359.*

Thomas Zenkel, J. W. (2020). End-to-End Neural Word Alignment Outperforms GIZA++. *In Proceeding of the ACL-2020*, (pp. 1605-1607).

Toshinori Sato, T. H., & Okumura, M. (2017). Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval (in Japanese). *Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing* (pp. NLP2017-B6-1). The Association for Natural Language Processing.

Vu Thanh, N. D. (2018, 6). VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 56-60). New: Association for Computational Linguistics.

Vu, T., Nguyen, D. Q., Nguyen, D. Q., Dras, M., & Johnson, M. (2018, 6). VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 56–60). New: Association for Computational Linguistics.

Xuansong Li, S. G. (2015). GALE Chinese-English Parallel Aligned Treebank -- Training. *Linguistic Data Consortium.* Linguistic Data Consortium.

BERT BASED WORD ALIGNMENT MODEL FOR JAPANESE-VIETNAMESE

Le Thanh Tung*[*]*, Nguyen Hong Buu Long, Hoang Khue

University of Science, Ho Chi Minh City, Vietnam National University Ho Chi Minh City, Vietnam

[*]Corresponding author: Le Thanh Tung – Email: tungleqb@gmail.com

Received: October 11, 2022; Revised: November 08, 2022; Accepted: February 21, 2023

ABSTRACT

Word alignment plays an important role in many subtasks of natural language processing. Therefore, a wide range of studies have been conducted on different language pairs. However, those on word alignment for Vietnamese - Japanese pair are still limited. Most Japanese-Vietnamese word alignments are created from word alignment tools based on statistical methods or unsupervised learning methods, giving results with low accuracy. In this study, we build a Japanese-Vietnamese word-level alignment corpus manually and then implement and train an automatic word alignment model for Japanese-Vietnamese bilingual sentence pairs. Our word alignment model achieves an outstanding accuracy of 20.06 F_1 scores compared to the GIZA++ tool. We have created a word alignment model for Japanese-Vietnamese, which is advanced at present.

Keywords: BERT; Japanese-Vietnamese; Parallel corpus; SQuAD; Word alignment model