CẢI TIẾN MÔ HÌNH GIÓNG HÀNG TRONG DỊCH MÁY THỐNG KÊ CẶP NGÔN NGỮ VIỆT-ANH VỚI KỸ THUẬT CHIA NHỎ TỪ

Đặng Thanh Quyền[1*], Nguyễn Chí Thành[1], Nguyễn Phương Thái[2]

Tóm tắt: *Trong hệ thống dịch máy thống kê (Statistical Machine Translation - SMT), gióng hàng từ là một nhiệm vụ quan trọng và có ảnh hưởng lớn đến chất lượng hệ dịch. Hiện nay, chưa có nghiên cứu nào sử dụng các kỹ thuật chia nhỏ từ cho hệ thống dịch máy thống kê cặp ngôn ngữ Việt-Anh. Trong bài báo này, chúng tôi đề xuất một hướng tiếp cận sử dụng các kỹ thuật chia nhỏ từ vào hệ thống dịch máy thống kê nhằm nâng cao chất lượng gióng hàng từ, từ đó nâng cao chất lượng hệ dịch cho cặp ngôn ngữ Việt-Anh. Ngoài việc áp dụng kỹ thuật chia nhỏ từ như một bước tiền xử lý, chúng tôi còn đề xuất cải tiến mô hình gióng hàng từ để nâng cao chất lượng hệ dịch. Phương pháp đề xuất đã được cài đặt, thử nghiệm với các kỹ thuật chia nhỏ từ khác nhau như BPE, Wordpiece, unigram và Morfessor, kết quả thử nghiệm cho thấy, việc áp dụng phương pháp đề xuất đều giúp tăng điểm BLEU so với kết quả baseline, với kết quả cao nhất sử dụng kỹ thuật BPE giúp tăng 0.81 điểm BLEU.*

Từ khóa: Subword; Gióng hàng từ; Dịch máy thống kê.

1. ĐẶT VẤN ĐỀ

Trong hệ thống dịch máy thống kê (SMT), việc gióng hàng từ trên một kho ngữ liệu song ngữ đã gióng hàng mức câu là một bước quan trọng và có ảnh hưởng lớn đến chất lượng hệ dịch [1]. Hiện nay, các mô hình gióng hàng từ phổ biến nhất là các mô hình gióng hàng IBM [2]. Các mô hình này được áp dụng rộng rãi trong các hệ thống dịch máy thống kê. Các tham số của các mô hình IBM được ước tính bằng cách sử dụng nguyên lý hợp lý cực đại (Maximum Likelihood), tức là bằng cách đếm sự đồng xuất hiện của các từ trong văn bản song song. Các mô hình gióng hàng IBM đòi hỏi một lượng lớn dữ liệu song ngữ được gióng hàng mức câu và thường gặp vấn đề khi gióng hàng với các từ có tần suất xuất hiện ít (từ hiếm - rare words). Đã có nhiều nghiên cứu nhằm tăng chất lượng gióng hàng từ cho dịch máy thống kê cho các cặp ngôn ngữ tài nguyên hạn chế, trong đó tập trung vào vấn đề xử lý từ hiếm [4], [3],...

Trong dịch máy Việt-Anh, bên cạnh vấn đề từ hiếm, ta gặp các vấn đề về sự không tương đồng về cấu trúc giữa hai ngôn ngữ, trong đó có sự khác biệt về hình thái. Tiếng Việt là ngôn ngữ đơn hình, trong đó, tiếng Anh là ngôn ngữ đa hình (một từ tiếng Anh có nhiều hình thái khác nhau tùy thuộc vào ngữ cảnh sử dụng, các hình thái từ này có chung một từ gốc và được bổ sung thêm các tiền tố, hậu tố tùy theo ngữ cảnh sử dụng). Hiện tượng tương tự đối với các tiếng Anh dạng từ kết hợp (một từ được tạo ra kết hợp bởi hai hoặc nhiều thành phần có nghĩa, khi kết hợp lại được một từ mới có nghĩa mới tương ứng với một hoặc nhiều từ phía tiếng Việt, ví dụ supermarket: siêu thị, wonderland: xứ sở thần tiên,...).

Trong bài báo này, chúng tôi đề xuất một phương pháp cải tiến mô hình gióng hàng từ sử dụng các kỹ thuật chia nhỏ từ trong hệ thống dịch máy thống kê cho cặp ngôn ngữ Việt-Anh nhằm giải quyết vấn đề từ hiếm và khác biệt về hình thái giữa hai ngôn ngữ. Đầu tiên, các kỹ thuật chia nhỏ từ (ví dụ như BPE [4], unigram [5],…) được sử dụng để chia nhỏ từ trong các câu phía tiếng Anh của kho ngữ liệu song ngữ, sau đó thực hiện gióng hàng từ và xây dựng bảng gióng hàng từ Việt-Anh. Chúng tôi đề xuất một thuật toán cải tiến bảng gióng hàng từ để sử dụng huấn luyện mô hình dịch máy Việt-Anh. Kết quả đạt được, hệ thống dịch máy sau khi cải tiến tăng 0.81 điểm BLEU so với hệ thống trước khi cải tiến.

Các đóng góp mới của nghiên cứu này bao gồm:

1. Đề xuất việc áp dụng kỹ thuật chia nhỏ từ đối với các câu tiếng Anh trước khi đưa vào gióng hàng trong hệ thống dịch máy Việt-Anh.

2. Đề xuất thuật toán tạo bảng gióng hàng từ A* từ bảng A trước khi xây dựng mô hình dịch, giúp giữ nguyên chất lượng mô hình ngôn ngữ trong hệ thống dịch máy.

Bài báo được trình bày theo thứ tự sau: Phần 2 trình bày các nghiên cứu liên quan; Phần 3 trình bày phương pháp cải tiến mô hình gióng hàng từ sử dụng các kỹ thuật chia nhỏ từ; Phần 4 trình bày các kết quả thử nghiệm, đánh giá; Cuối cùng, kết luận được trình bày trong phần 5.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Dịch máy thống kê được quan tâm và nghiên cứu cách đây hơn 20 năm. Chất lượng của một hệ dịch máy thống kê phụ thuộc vào hai yếu tố chính là ngữ liệu huấn luyện và mô hình dịch. Đối với các cặp ngôn ngữ tài nguyên hạn chế (như cặp ngôn ngữ Việt-Anh), việc cải tiến mô hình dịch được ưu tiên vì khó khăn trong bổ sung ngữ liệu huấn luyện. Trong mô hình dịch máy thống kê, gióng hàng từ là một bước quan trọng ảnh hưởng lớn đến chất lượng hệ dịch, xây dựng nên mô hình dịch cho hệ thống. Có nhiều nghiên cứu nhằm nâng cao chất lượng gióng hàng từ cho các cặp ngôn ngữ trên thế giới, tuy nhiên, với cặp ngôn ngữ Việt-Anh chưa có nhiều nghiên cứu về gióng hàng từ. Ngoài ra, các kỹ thuật chia nhỏ từ được sử dụng rộng rãi trong dịch máy trên mạng nơ-ron nhưng chưa có nhiều nghiên cứu áp dụng cho dịch máy thống kê. Trong phần này, chúng tôi sẽ giới thiệu một số nghiên cứu về gióng hàng từ và chia nhỏ từ trong dịch máy.

2.1. Gióng hàng từ trong dịch máy thống kê

Trong mô hình dịch máy thống kê, gióng hàng từ (word alignment) có nhiệm vụ xác định sự tương ứng giữa các từ trong một văn bản song ngữ [6]. Đây là bước đầu tiên trong hầu hết các cách tiếp cận hiện tại của SMT và cũng là bước đóng vai trò rất quan trọng cho sự thành công của một hệ thống SMT. Trong các mô hình gióng hàng từ, các mô hình IBM của Brown và cộng sự [2] được sử dụng rộng rãi nhất.

Các phương pháp nâng cao chất lượng gióng hàng từ có thể được chia thành 2 hướng: cải tiến mô hình gióng hàng và tiền/hậu xử lý dữ liệu trước/sau khi gióng hàng. Các nghiên cứu theo hướng cải tiến mô hình gióng hàng phần lớn là các nghiên cứu nhằm cải tiến mô hình IBM. Một số nghiên cứu khác đã đề xuất các phương pháp đưa thêm các ràng buộc về ngôn ngữ vào mô hình gióng hàng để cải tiến chất lượng gióng hàng [7]. Trong hướng nghiên cứu thứ hai, nhiều nghiên cứu tập trung vào việc sử dụng các thông tin về từ loại để nâng cao độ chính xác của gióng hàng, áp dụng trong giai đoạn tiền xử lý [8] và hậu xử lý [9].

Đối với dịch máy thống kê giữa hai ngôn ngữ tiếng Việt và tiếng Anh cũng đã có một số nghiên cứu nâng cao chất lượng gióng hàng từ cho dịch máy từ tiếng Anh-Việt và ngược lại (Việt-Anh). Đối với bài toán dịch máy Anh-Việt, Lê Quang Hùng và cộng sự đã có một số công trình theo hướng cải tiến mô hình gióng hàng bằng cách đưa thêm một số ràng buộc như ràng buộc neo, ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ [10]. Nhóm nghiên cứu đưa ra phương pháp để tích hợp các ràng buộc vào thuật toán EM trong quá trình ước lượng tham số của mô hình và đưa ra một phương pháp để kết hợp các ràng buộc. Vuong Van Bui và cộng sự đã đề xuất một phương pháp tiền xử lý bằng cách phân tích hình thái các từ tiếng Anh trước khi đưa vào mô hình IBM [11]. Kết quả thực nghiệm trên bài toán dịch máy Anh-Việt cho thấy đề xuất này giúp nâng cao chất lượng của dịch máy, tuy nhiên, các kết quả chỉ cải thiện đối với các trường hợp dữ liệu huấn luyện có kích thước từ 35.000 cặp câu trở xuống. Tại hội nghị IWSLT 2015, Takahiro Nomura và cộng sự đã đề xuất hai phương pháp tiền xử lý cho dịch máy thống kê cặp ngôn ngữ Việt-Anh, tuy nhiên, kết quả thực nghiệm cho thấy, các cải tiến này không cải thiện chất lượng của hệ thống dịch máy. Trần Hồng Việt và cộng sự đã đề xuất một số phương pháp đảo trật tự từ của các câu đầu vào trước khi đưa vào hệ thống dịch máy thống kê cho cả dịch máy Anh-Việt và Việt-Anh [12].

2.2. Các kỹ thuật chia nhỏ từ

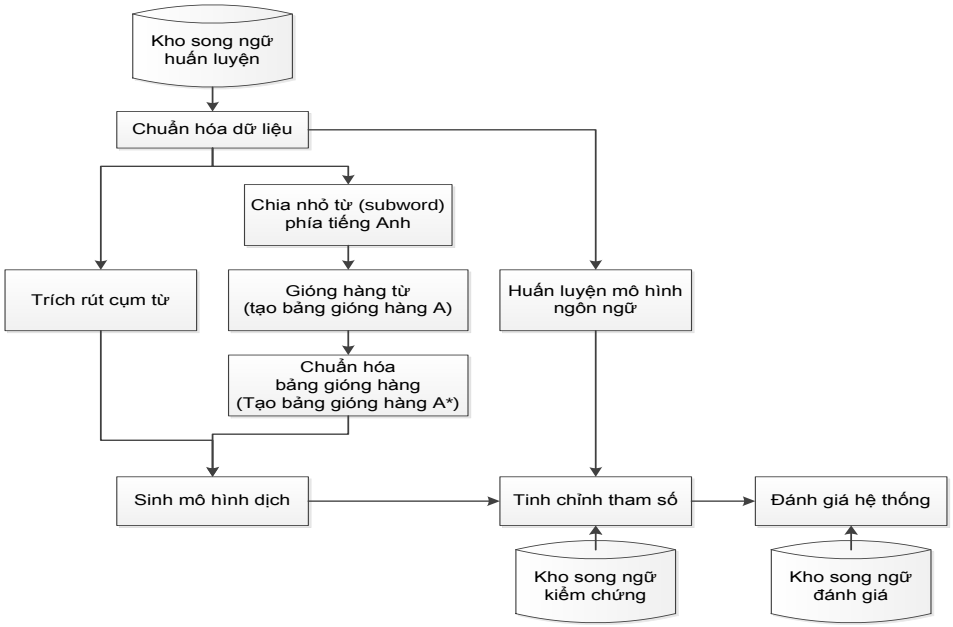
Trong dịch máy trên mạng nơ-ron, kỹ thuật chia nhỏ từ thường được sử dụng như một

phương pháp biểu diễn từ nhằm mục đích giảm kích thước bảng từ vựng, từ đó hạn chế hiện tượng OOV (Out of Vocabulary – từ nằm ngoài bảng từ vựng). Các từ hiếm và từ chưa biết được mã hóa dưới dạng chuỗi các từ con. Các kỹ thuật chia nhỏ từ hiện nay đang được sử dụng phổ biến và đem lại hiệu quả là BPE [4], Wordpiece [17], unigram [13].

Các kỹ thuật chia nhỏ từ trên thường được sử dụng cho các bài toán xử lý ngôn ngữ tự nhiên nói chung và bài toán dịch máy nói riêng trên mạng nơ-ron để giải quyết vấn đề từ hiếm, từ chưa biết. Hiện nay có rất ít công bố sử dụng các kỹ thuật này cho dịch máy thống kê nói chung và chưa có nghiên cứu nào cho dịch máy thống kê cặp ngôn ngữ Việt-Anh. Trong bài báo này, chúng tôi đề xuất một hướng tiếp cận áp dụng các kỹ thuật chia nhỏ từ để huấn luyện hệ thống dịch máy thống kê. Trong hướng tiếp cận này, bên cạnh áp dụng kỹ thuật chia nhỏ từ, chúng tôi còn cải tiến mô hình gióng hàng từ để nâng cao chất lượng hệ dịch. Trong các phần tiếp theo, chúng tôi trình bày chi tiết về hướng tiếp cận này và thực hiện một số thực nghiệm để chứng minh hiệu quả của phương pháp.

3. CẢI TIẾN MÔ HÌNH GIÓNG HÀNG VỚI KỸ THUẬT CHIA NHỎ TỪ

Chúng tôi đề xuất một phương pháp cải tiến mô hình gióng hàng nhằm nâng cao chất lượng hệ dịch cho dịch máy thống kê cặp ngôn ngữ Việt-Anh sử dụng các kỹ thuật chia nhỏ từ. Ý tưởng chính của đề xuất là trước khi thực hiện gióng hàng từ, các câu phía tiếng Anh được chia nhỏ từ (bước này được coi là encode ngữ liệu phía tiếng Anh), sau đó thực hiện gióng hàng từ giữa các cặp song ngữ tiếng Việt và tiếng Anh (đã encode), thu được bảng gióng hàng từ A. Ở bước tiếp theo, bảng gióng hàng từ A được chuẩn hóa để sinh ra bảng gióng hàng từ A* giữa các cặp câu tiếng Việt và tiếng Anh ban đầu (bước này được coi là decode bảng gióng hàng từ). Sau đó, bảng gióng hàng từ A* được sử dụng để huấn luyện mô hình dịch máy. Phương pháp đề xuất được mô tả trong hình 1. Việc áp dụng chia nhỏ từ sẽ giúp giải quyết được hai vấn đề ảnh hưởng đến chất lượng của dịch máy thống kê: (i) vấn đề từ hiếm, (ii) sự khác biệt về hình thái từ giữa hai ngôn ngữ tiếng Anh (đa hình) và tiếng Việt (đơn hình).



Hình 1. *Mô hình đề xuất áp dụng chia nhỏ từ vào dịch máy thống kê.*

Phương pháp đề xuất bao gồm 2 cải tiến trong quá trình tạo bảng gióng hàng: (i) Chia nhỏ từ phía tiếng Anh trước khi đưa vào gióng hàng; (ii) Đề xuất thuật toán tạo bảng gióng hàng từ mới A*($V \rightarrow E$) từ bảng gióng hàng A($V \rightarrow E$').

- Chia nhỏ từ phía tiếng Anh trước khi đưa vào gióng hàng: Việc chia nhỏ từ nhằm mục đích giảm kích thước bảng từ vựng, từ đó tăng tần suất xuất hiện của từ trong ngữ liệu huấn luyện, giảm số lượng các từ có tần suất xuất hiện thấp. Ngoài ra, do tiếng Anh là ngôn ngữ đa hình, trong đó, mỗi từ có nhiều hình thái khác nhau bằng cách thêm vào các tiền tố, hậu tố khi có thay đổi về từ loại, thì của động từ,... Khi dịch một câu tiếng Việt sang tiếng Anh, một từ tiếng Việt có thể chỉ tương ứng với một phần của một từ tiếng Anh. Nếu kỹ thuật chia nhỏ từ tách được từ gốc và các tiền tố, hậu tố thì chất lượng của bảng gióng hàng từ sẽ được nâng lên [11].

Kỹ thuật chia nhỏ từ được áp dụng trên kho ngữ liệu huấn luyện C(*V,E*), trong đó, *V* là tập các câu tiếng Việt, *E* là tập các câu tiếng Anh tương ứng. Chia nhỏ từ chỉ thực hiện đối với các câu phía tiếng Anh, tập *E* sau khi thực hiện chia nhỏ từ được tập *E'*.

$$E' = subword(E)$$

Trong đó, *subword(E)* là kỹ thuật chia nhỏ từ các câu trong tập E bằng các thuật toán chia nhỏ từ (*BPE, Wordpiece, Unigram, Morfessor*), sau bước này thu được kho ngữ liệu C'(*V,E'*). Bước gióng hàng từ trong huấn luyện mô hình dịch máy được thực hiện trên kho ngữ liệu C'(*V,E'*) và thu được bảng gióng hàng từ A($V \rightarrow E'$).

Thuật toán DecodeAlignmentTable

Input: Bảng gióng hàng từ *A*, Tập các câu tiếng Anh đã được chia nhỏ *E'*
Output: Bảng gióng hàng từ *A**

```
For each a in A
    s ← GetEnglishSentenceOf(a)
    Loop
        For each s[i] in s
            If s[i] is subword    // Từ s[i] là từ đã được chia nhỏ
                s[i] ← s[i] + s[i+1]
                For all a[j] in a
                    If a[j] include alignment k>i
                        Update_alignment a[j]: k ← k – 1
    Until Number of subwords in s = 0
    a* ← RemoveDuplicateAlignment(a)
    A* ← A* + a*
Return(A*)
```

Hình 2. *Thuật toán tạo bảng gióng hàng từ A*(V → E) từ bảng gióng hàng từ A(V → E').*

- Thuật toán tạo bảng gióng hàng từ mới A*($V \rightarrow E$) từ bảng gióng hàng từ A($V \rightarrow E'$): Cải tiến thứ hai là thay vì sử dụng bảng gióng hàng A($V \rightarrow E'$) được tạo ra từ bước gióng hàng từ, chúng tôi đề xuất một thuật toán để sinh bảng gióng hàng từ mới A*($V \rightarrow E$) từ bảng gióng hàng A($V \rightarrow E'$). Nếu sử dụng bảng A($V \rightarrow E'$) để huấn luyện hệ thống dịch máy sẽ có hai vấn đề cần giải quyết: (i) câu dịch nhận được sau khi dịch là câu tiếng Anh đã chia nhỏ do vậy cần giải mã lại câu này để nhận được câu dịch đúng, (ii) mô hình ngôn ngữ huấn luyện trên tập *E'* đã bị chia nhỏ nên các thống kê *n-gram* sẽ không đúng với định nghĩa thống kê. Để xây dựng bảng gióng hàng từ A* từ bảng gióng hàng từ A, chúng tôi đề xuất phương pháp tạo gióng hàng như sau: (i) trong bảng A nếu một từ tiếng Việt được gióng với mới một từ con của một từ tiếng Anh thì ta thêm một gióng hàng giữa từ tiếng Việt với từ tiếng Anh đó vào A*, (ii) trong trường hợp còn lại thì ta giữ nguyên gióng hàng đó để thêm vào A*. Phương pháp này được mô hình hóa như sau:

Cho tập ngữ liệu song ngữ C(V,E) và bảng gióng hàng từ A(V,E'). Trong bảng A(V,E'), mỗi cặp câu (v,e'), với v∈V và e'∈E', có nhiều gióng hàng từ ($v_j \rightarrow e'_i$), trong đó, $v_j \in v$ ($j \in [1..n]$) và $e'_i \in e'$ ($i \in [1..m]$). Với mỗi cặp câu (v,e') trong A(V,E'), xét tất cả các gióng hàng từ ($v_j \rightarrow e'_i$):

- Nếu e'_i là từ con và e'_i được chia nhỏ từ e_k thì thêm gióng hàng ($v_j \rightarrow e_k$) vào bảng A.*

- Nếu e'_i không phải là từ con thì thêm gióng hàng ($v_j \rightarrow e'_i$) vào A.*

Sau đó thực hiện xóa bỏ các gióng hàng giống nhau trong A để loại bỏ trùng lặp.*

Thuật toán DecodeAlignmentTable tạo bảng bảng gióng hàng từ A*($V \rightarrow E$) theo phương pháp trên được trình bày ở hình 2.

4. THỬ NGHIỆM, ĐÁNH GIÁ

4.1. Dữ liệu và môi trường thử nghiệm

Cặp ngôn ngữ Việt-Anh là cặp ngôn ngữ có nguồn ngữ liệu song ngữ hạn chế, không có nhiều bộ dữ liệu công khai (các bộ dữ liệu được sử dụng trong các công bố gần đây không được công khai). Trong nghiên cứu của chúng tôi, chúng tôi sử dụng bộ dữ liệu của nhóm Stanford NLP (https://nlp.stanford.edu/projects/nmt/): IWSLT'15 English-Vietnamese data [Small]. Thống kê về bộ dữ liệu này được trình bày trong bảng 1.

***Bảng 1.** Kho ngữ liệu IWSLT15.*

Tên file	train.en	train.vi	tst2012.en	tst2012.vi	tst2013.en	tst2013.vi
Sử dụng	Huấn luyện (train)		Tinh chỉnh (tuning)		Đánh giá (evaluation)	
Số lượng câu	133.317	133.317	1.553	1.553	1.268	1.268
Số lượng từ	2.706.404	3.311.620	27.983	34.297	26.728	33.682

Để đánh giá phương pháp đề xuất, chúng tôi tiến hành các thử nghiệm như sau:

- Thử nghiệm thứ nhất (Baseline): tính điểm baseline.

- Thử nghiệm thứ hai (SMT-BPE-A): huấn luyện và đánh giá hệ thống chỉ chia nhỏ từ, không tạo bảng gióng hàng A*.

- Thử nghiệm thứ ba (SMT-BPE-A*), thứ tư (SMT-Wordpiece-A*), thứ năm (SMT-Unigram-A*) và thứ sáu (SMT-morfessor-A*): thực hiện chia nhỏ từ bằng kỹ thuật BPE, Wordpiece, unigram, Morfessor, sử dụng bảng gióng hàng A* được chuẩn hóa từ bảng gióng hàng A bằng phương pháp được đề xuất ở Phần 3.

Các thử nghiệm được cài đặt trên hệ thống dịch máy thống kê Moses ((http://www.statmt.org/moses/) với bộ số liệu IWSLT2015 được thống kê trong bảng 1. Gióng hàng từ sử dụng công cụ GIZA++ Toolkit. Mô hình ngôn ngữ trong các thử nghiệm 1, 3, 4, 5 và 6 được huấn luyện bằng bằng công cụ Kenlm [15] trên dữ liệu huấn luyện phía tiếng Anh. Đối với thử nghiệm 2, mô hình ngôn ngữ huấn luyện trên dữ liệu tiếng Anh đã được chia nhỏ. Độ đo BLEU [14] được sử dụng để đánh giá chất lượng hệ dịch máy. Chi tiết các thử nghiệm mô tả trong phần sau.

4.2. Kết quả thử nghiệm

4.2.1. Thử nghiệm baseline trên hệ thống dịch máy thống kê MOSES với dữ liệu huấn luyện IWSLT15 gốc

Kết quả baseline được trình bày trong bảng 2.

***Bảng 2.** Kết quả thử nghiệm Baseline.*

Hệ thống	**BLEU**	**BLEU-c**
Baseline	24,45	24,05

4.2.2. Các thử nghiệm sử dụng các kỹ thuật chia nhỏ từ

Thử nghiệm thứ hai (hệ thống SMT-BPE-A): sử dụng kỹ thuật chia nhỏ từ BPE để chia nhỏ (encode) các từ thuộc phía tiếng Anh với kích thước bảng từ vựng là 5.000 (5K), để tính điểm BLEU cần kết hợp các từ đã được phân đoạn (decode) các câu thu được để so sánh với các câu tham chiếu (reference) trong tập *tst2013*.

Thử nghiệm thứ ba (hệ thống SMT-BPE-A*), chúng tôi sử dụng kỹ thuật chia nhỏ từ BPE để chia nhỏ các từ thuộc phía tiếng Anh với kích thước bảng từ vựng lần lượt được sử dụng là 4K (4.000), 5K (5.000), 6K (6000), 8K (8.000).

Thử nghiệm thứ tư (hệ thống SMT-Wordpiece-A*), chúng tôi sử dụng kỹ thuật chia nhỏ từ wordpiece để chia nhỏ các từ thuộc phía tiếng Anh.

Thử nghiệm thứ năm (hệ thống SMT-Unigram-A*), chúng tôi sử dụng kỹ thuật chia nhỏ từ unigram để chia nhỏ các từ thuộc phía tiếng Anh, kích thước bảng từ vựng lần lượt được sử dụng là 4K (4.000), 5K (5.000), 6K (6000).

Thử nghiệm thứ sáu, chúng tôi sử dụng công cụ Morfessor 2.0 [16] để chia nhỏ các từ thuộc phía tiếng Anh. Các kỹ thuật BPE, Wordpiece và unigram là các kỹ thuật chia nhỏ được sử dụng cho dịch máy trên mạng nơ-ron. Khác với các kỹ thuật này, Morfessor là một công cụ phân tích hình thái từ tiếng Anh dựa trên học máy thống kê. Mặc dù Morfessor không được sử dụng cho dịch máy trên mạng nơ-ron, chúng tôi tiến hành thử nghiệm chia nhỏ từ bằng Morfessor để đánh giá hiệu quả của chia nhỏ từ bằng phân tích hình thái đối với dịch máy thống kê cặp ngôn ngữ Việt-Anh. Phương pháp này đòi hỏi mô hình dùng để chia nhỏ từ phải được huấn luyện từ dữ liệu đơn ngữ. Trong thử nghiệm, này, chúng tôi huấn luyện hai mô hình chia nhỏ từ các kho ngữ liệu đơn ngữ khác nhau: (i) đối với hệ thống **SMT-Morfessor1-A*** thì sử dụng các câu phía tiếng Anh của kho ngữ liệu huấn luyện IWSLT2015 và (ii) đối với hệ thống **SMT-Morfessor2-A*** thì sử dụng ngữ liệu đơn ngữ từ kho europarl-v7.en, sau đó sử dụng các mô hình này để thực hiện việc chia nhỏ từ. Các bước còn lại thực hiện như thử nghiệm thứ ba.

Bảng 3. *Kết quả thử nghiệm với dữ liệu huấn luyện đã được chia nhỏ từ các câu tiếng Anh bằng các kỹ thuật BPE, Wordpiece, Unigram, Morfessor.*

Hệ thống	Kích thước bảng từ vựng	BLEU	ΔBLEU	BLEU-c	ΔBLEU-c
Baseline		24,45		24,05	
SMT-BPE-A		23,40	-1.05	22,16	-1.59
SMT-BPE-A*	4.000	24,86	0,41	24,42	0,37
	5.000	**25,26**	**0,81**	**24,85**	**0,80**
	6.000	24,65	0,20	24,28	0,23
	8.000	24,82	0,27	24,43	0,38
SMT-Wordpiece-A*		24,87	0,42	24,49	0,44
SMT-Unigram-A*	4.000	24,73	0,28	24,30	0,25
	5.000	24,80	0,35	24,39	0,34
	6.000	24,69	0,24	24,30	0,25
SMT-Morfessor1-A*		24,46	0,01	24,06	0,01
SMT-Morfessor2-A*		24,95	0,50	24,55	0,50

Các kết quả thử nghiệm được trình bày trong bảng 3. Trong đó, ΔBLEU và ΔBLEU-c là kết quả so sánh với điểm baseline.

Bảng kết quả thử nghiệm trên hệ thống dịch máy thống kê MOSES ở trên cho thấy, hệ thống SMT-BPE-A chỉ áp dụng chia nhỏ từ và thực hiện huấn luyện hệ thống dịch máy thống kê, khiến cho hệ thống này trở nên kém hơn so với hệ thống gốc. Điều này có thể lý giải do việc chia nhỏ

từ đã làm thay đổi các câu dùng để huấn luyện mô hình ngôn ngữ, mô hình ngôn ngữ không còn hiệu quả trong việc lựa chọn các câu dịch tốt nhất. Việc chia nhỏ từ chỉ giúp ích cho công đoạn huấn luyện mô hình dịch, trực tiếp ở đây là bảng gióng hàng từ.

Các kết quả thử nghiệm đều cho thấy: áp dụng chia nhỏ từ và sử dụng thuật toán tạo bảng gióng từ A* đều khiến cho hệ thống tốt hơn theo đánh giá bằng điểm BLEU. Từ đó cho thấy việc sử dụng kỹ thuật chia nhỏ từ cho phía tiếng Anh trong dịch máy Việt-Anh và áp dụng thuật toán tạo bảng gióng hàng từ đã đề xuất trong bài báo giúp nâng cao chất lượng cho mô hình dịch máy thống kê Việt-Anh. Ngoài các phương pháp chia nhỏ từ thông dụng cho dịch máy trên mạng nơ-ron, việc áp dụng phương pháp do bài báo đề xuất với kỹ thuật chia nhỏ từ dựa trên phân tích hình thái (Morfessor) cũng giúp nâng cao chất lượng hệ dịch, tăng thêm 0,5 điểm BLEU.

5. KẾT LUẬN

Trong bài báo, chúng tôi đề xuất một phương pháp cải tiến mô hình gióng hàng từ sử dụng các kỹ thuật chia nhỏ từ trong hệ thống dịch máy thống kê cho cặp ngôn ngữ Việt-Anh để giải quyết vấn đề từ hiếm và tăng chất lượng gióng hàng từ. Phương pháp đề xuất bao gồm 2 cải tiến đối với mô hình gióng hàng: (i) Áp dụng kỹ thuật chia nhỏ từ đối với các câu tiếng Anh trước khi đưa vào gióng hàng; (ii) Đề xuất thuật toán tạo bảng gióng hàng từ A* từ bảng A. Kết quả thử nghiệm cho thấy, đối với cặp ngôn ngữ Việt-Anh có tài nguyên hạn chế, khi sử dụng các kỹ thuật BPE, Wordpiece, unigram và Morfessor để chia nhỏ từ trong các câu phía tiếng Anh, sau đó thực hiện gióng hàng từ và xây dựng bảng gióng hàng từ Việt-Anh bằng thuật toán đề xuất thì hệ thống dịch máy sau khi cải tiến tăng 0,81 điểm BLEU so với hệ thống trước khi cải tiến.

TÀI LIỆU THAM KHẢO

[1]. Brown, Peter F., et al. *"A statistical approach to machine translation."* Computational linguistics 16.2 (1990): 79-85.

[2]. Brown, Peter F., et al. *"The mathematics of statistical machine translation: Parameter estimation."* Computational linguistics 19.2 (1993): 263-311.

[3]. Poerner, Nina, et al. *"Aligning Very Small Parallel Corpora Using Cross-Lingual Word Embeddings and a Monogamy Objective."* arXiv preprint arXiv:1811.00066 (2018).

[4]. Sennrich, Rico, Barry Haddow, and Alexandra Birch. *"Neural machine translation of rare words with subword units."* arXiv preprint arXiv:1508.07909 (2015).

[5]. Kudo, Taku. *"Subword regularization: Improving neural network translation models with multiple subword candidates."* arXiv preprint arXiv:1804.10959 (2018).

[6]. Liu, Yang, Qun Liu, and Shouxun Lin. *"Discriminative word alignment by linear modeling."* Computational Linguistics 36.3 (2010): 303-339.

[7]. Kamigaito, Hidetaka, et al. *"Unsupervised Word Alignment Using Frequency Constraint in Posterior Regularized EM."* Journal of Natural Language Processing 23.4 (2016): 327-351.

[8]. Ghaffar, Shady Abdel, Mohamed Waleed Fakhr, and Cairo Sheraton. *"English to arabic statistical machine translation system improvements using preprocessing and arabic morphology analysis."* Recent Researches in Mathematical Methods in Electrical Engineering and Computer Science (2011): 50-54.

[9]. Clifton, Ann, and Anoop Sarkar. *"Combining morpheme-based machine translation with post-processing morpheme prediction."* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011.

[10]. Quang-Hung, L. E., and L. E. Anh-Cuong. *"Syntactic pattern based Word Alignment for Statistical Machine Translation."* International Journal of Knowledge and Systems Science (IJKSS) 5.3 (2014): 36-45.

[11]. Van Bui, Vuong, et al. *"Improving Word Alignment Through Morphological Analysis."* International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making. Springer, Cham, 2015.

[12]. Viet, Tran Hong, et al. *"Dependency-based pre-ordering for English-Vietnamese statistical machine translation."* VNU Journal of Science: Computer Science and Communication Engineering 33.2 (2017).

[13]. Kudo, Taku. *"Subword regularization: Improving neural network translation models with multiple subword candidates."* arXiv preprint arXiv:1804.10959 (2018).
[14]. Papineni, Kishore, et al. *"Bleu: a method for automatic evaluation of machine translation."* Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
[15]. Heafield, Kenneth. *"KenLM: Faster and smaller language model queries."* Proceedings of the sixth workshop on statistical machine translation. 2011.
[16]. Smit, Peter, et al. *"Morfessor 2.0: Toolkit for statistical morphological segmentation."* The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014. Aalto University, 2014.
[17]. Wu, Yonghui, et al. *"Google's neural machine translation system: Bridging the gap between human and machine translation."* arXiv preprint arXiv:1609.08144 (2016).

ABSTRACT

SUBWORD FOR VIETNAMESE-ENGLISH STATISTICAL MACHINE TRANSLATION

In this paper, we propose an approach for applying subword methods in SMT to improve word alignment in Vietnamese-English SMT systems. In addition to applying subword methods as a preprocessing step, we propose a new algorithm for decoding alignment table of translation model. The proposed method has been implemented and evaluated with various subword methods: BPE, Wordpiece, unigram, and Morfessor. Experimental results show that the proposed method produces better results with every subword method, and the highest improvement is 0.81 BLEU from the model with the BPE subword method.

Keywords: Subword; Word alignment; Statistical machine translation.

Nhận bài ngày 26 tháng 4 năm 2021
Hoàn thiện ngày 17 tháng 5 năm 2021
Chấp nhận đăng ngày 30 tháng 7 năm 2021

Địa chỉ: [1] Viện Công nghệ thông tin, Viện KH-CN quân sự;
[2] Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.
[*]Email: dangthanhquyen@gmail.com.