Tạp chí Tin học và Điều khiển học, T.29, S.4 (2013), 358–368

DỊCH MÁY THỐNG KÊ PHÁP-VIỆT KẾT HỢP THÔNG TIN GIÓNG HÀNG PHÂN ĐOẠN NGỮ

LÊ NGỌC TẤN[1], LÊ NGỌC TIẾN[1], ĐINH ĐIỀN[2]

[1]*Khoa Công nghệ Thông tin, Trường Đại học Công nghiệp Tp. HCM;*
letan, letien.dhcn@gmail.com

[2]*Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên Tp.HCM;*
ddien@fit.hcmus.edu.vn

Tóm tắt. Hiện nay, trong các mô hình dịch máy thống kê, mô hình dịch dựa trên ngữ được đánh giá cao nhất. Tuy nhiên, mô hình này vẫn còn thiếu sự tích hợp các tri thức ngôn ngữ ở mức cao hơn, như thông tin từ pháp, thông tin cú pháp và ngữ nghĩa. Điều này dẫn đến kết quả của phương pháp này vẫn còn bị hạn chế đối với bài toán câu dài. Chính vì vậy, việc sử dụng các thông tin hình thái như phân đoạn ngữ với mục đích giảm độ dài câu để cải tiến chất lượng dịch là một trong những hướng tiếp cận đầy tiềm năng trong những năm gần đây và qua đó, góp phần khử nhập nhằng trong gióng hàng từ trong bài toán câu dài. Bài báo đề xuất hướng tiếp cận dịch máy thống kê Pháp-Việt kết hợp thông tin phân đoạn ngữ cho cặp ngôn ngữ Pháp-Việt nhằm khắc phục hạn chế đối của hệ dịch với những câu dài. Tiến hành thử nghiệm mô hình hệ thống với kho ngữ liệu song ngữ Pháp-Việt gồm 10.000 cặp câu và kết quả độ đo BLEU tăng gần 2% so với mô hình cơ sở.

Từ khóa. Ngữ liệu song ngữ, dịch máy thống kê, gióng hàng phân đoạn ngữ.

Abstract. Nowadays, among Statistical Machine Translation (SMT) models, the phrase-based SMT is highly appreciated, however, this model is still lacked of linguistics knowledge at a higher level such as morphological, syntactic and semantic information. Consequently, the results of this approach are still limited by the issue of long sentences. So, using morphological information from such as phrase chunking on the purpose of reducing the length of sentences to improve the translation quality is a promising approach. And thus, it contributes to disambiguate the chunk alignment in the long sentences. In this paper, we present an approach of a chunk alignment applied to French-Vietnamese SMT. We tested the model system with a French-Vietnamese bilingual corpus which consists of 10,000 pairs and assessed the metrics measures. The result of the model of French-Vietnamese SMT based on chunk alignment is considerable with the BLEU metric measure which increases almost 2% compared to the baseline model.

Key words. Bilingual corpus, statistical machine translation, chunk alignment.

1. GIỚI THIỆU

Hiện nay, phương pháp dịch máy thống kê càng ngày càng có nhiều bước tiến đáng kể. Đến nay, đã có nhiều cách tiếp cận nhằm cải tiến kết quả dịch máy thống kê như dịch máy

thống kê dựa trên ngữ (phrase-based SMT) [1], dịch máy thống kê dựa trên phân tích cây cú pháp (syntaxbased SMT) [2]. Nhưng hướng tiếp cận dựa trên cây cú pháp không khả thi đối với các ngôn ngữ chưa có công cụ phân tích cú pháp hiệu quả (tiếng Việt của chúng ta là một ví dụ). Còn hướng tiếp cận dựa trên ngữ thì đến nay vẫn còn bị hạn chế khi gặp các câu dài. Vì vậy việc dùng phân đoạn ngữ (phrasechunking) với mục đích giảm độ dài câu để cải tiến độ chính xác của gióng hàng từ và góp phần làm tăng chất lượng dịch là một trong những hướng tiếp cận đầy tiềm năng trong những năm gần đây.

Trong bài báo này, chúng tôi sẽ đề xuất hướng tiếp cận dùng thông tin phân đoạn ngữ cho câu tiếng Pháp kết hợp với từ điển song ngữ và tính chất biên trong các đoạn ngữ để tạo nên các cặp phân đoạn ngữ cho cặp ngôn ngữ Pháp-Việt nhằm khắc phục hạn chế của hệ thống cho những câu dài trong dịch máy thống kê.

Dựa trên ý tưởng từ hai bài báo của nhóm Isabelle Tellier [3] và nhóm Sun Le [5], đã đề xuất một mô hình dịch máy thống kê dựa trên thông tin gióng hàng phân đoạn ngữ nhằm mục đích giải quyết bài toán câu dài dựa trên phân đoạn ngữ cho cặp ngôn ngữ Pháp-Việt. Ở đây, ta sử dụng mô hình gióng hàng phân đoạn ngữ cho cặp ngôn ngữ Pháp-Việt kết hợp với phương pháp gióng hàng từ bằng Giza++ và phương pháp dịch máy thống kê dựa trên ngữ (phrase based SMT) bằng Moses.

Phần còn lại của bài báo sẽ được trình bày như sau: phần 2 trình bày về những hướng tiếp cận liên quan hiện nay, phần 3 mô tả về phương pháp gióng hàng phân đoạn ngữ cho cặp ngôn ngữ Pháp-Việt và mô hình kết hợp gióng hàng phân đoạn ngữ vào dịch máy thống kê. Phần 4 trình bày về các thực nghiệm–đánh giá. Và phần cuối cùng kết luận và hướng phát triển.

2. CÁC CÔNG TRÌNH LIÊN QUAN

Ở Việt Nam, với sự phát triển của ngành công nghệ thông tin, các nghiên cứu về dịch máy từ Anh-Việt đã bắt đầu từ những năm 1980. Cho đến nay, đã có nhiều công trình nghiên cứu liên quan đến dịch máy Anh-Việt, Việt-Anh, tuy nhiên, vẫn có rất ít các nghiên cứu về dịch máy Pháp-Việt hay Việt–Pháp và những kết quả đạt được vẫn còn khá khiêm tốn. Sự phát triển đối với cặp ngôn ngữ Pháp-Việt chỉ được thực hiện đơn lẻ ở các giai đoạn phân tích về ngữ pháp như phân tích cây cú pháp của tác giả Lê Hồng Phương [10] hay phân tích về thời gian trong tiếng Việt của tác giả Nicolas Boffo [11]. Ngoài ra còn có công trình về gióng hàng văn bản đa ngữ Pháp–Việt của tác giả Nguyễn Thị Minh Huyền [12] được phát triển trong luận án tiến sĩ năm 2006.

Năm 2010, đối với cặp ngôn ngữ Pháp-Việt có công trình khai thác kho ngữ liệu không thực sự song song cho hệ thống dịch máy thống kê Pháp-Việt của nhóm tác giả Đỗ Thị Ngọc Diệp cùng cộng sự [9]. Hệ thống này sử dụng các đặc trưng như các từ viết hoa làm tên riêng để tìm các cặp văn bản song song nhau.

Trong các hướng nghiên cứu áp dụng phân đoạn ngữ trong dịch máy, có hai hướng tiêu biểu. Hướng tiếp cận thứ nhất, xác định các gióng hàng phân đoạn ngữ trong câu. Người ta sử dụng ngữ liệu này làm ngữ liệu chính cho mô hình dịch thống kê (như công trình của nhóm Sun Le [5]). Trong khi đó hướng tiếp cận thứ hai, sử dụng các phân đoạn ngữ để xây dựng các tập luật chuyển đổi trật tự (chunk reordering), giúp cho kết quả dịch chính xác hơn (như công trình của nhóm Vinh Van Nguyen [6]). Bên cạnh đó, cũng còn có hướng nghiên cứu lai kết hợp cả thống kê và dùng mẫu dịch, tận dụng các từ khóa để phân chia câu thành các phân đoạn ngữ cho từng ngôn ngữ. Sau đó, sử dụng quy hoạch động và thống kê để tìm các mẫu

gióng hàng phân đoạn ngữ (như công trình của nhóm Francisco Nevado [7]).

Đối với hướng nghiên cứu thứ hai, nhóm [6] đã thực hiện thử nghiệm trên ngữ liệu song ngữ Anh-Việt, tuy nhiên kết quả còn một số hạn chế, đặc biệt là việc xác định phân đoạn ngữ cho tiếng Việt hiện này kết quả chưa cao. Còn với hướng nghiên cứu thứ nhất hiện nay vẫn chưa có bài báo nào được công bố cho tiếng Việt. Đây chính là một lý do khiến nhóm tác giả thực hiện nghiên cứu theo hướng này.

Ngoài ra, cũng có nhóm tác giả dựa trên gióng hàng từ và xác suất gióng hàng để tìm các điểm cắt phân chia câu thành các phân đoạn có xác suất tốt nhất, từ đó xây dựng được các gióng hàng phân đoạn ngữ. Đối với nhóm [7], tác giả lại sử dụng các từ đánh dấu (marker-word) để chia cắt câu thành từng phân đoạn cho từng ngôn ngữ riêng lẻ. Sau đó, sử dụng quy hoạch động với xác suất gióng hàng từ để xác định gióng hàng phân đoạn ngữ. Dựa trên ý tưởng của nhóm Sun Le [5] , nhóm thực hiện phân đoạn ngữ của tiếng Pháp (là ngôn ngữ đã được nghiên cứu khá sâu và đã đạt được các kết quả phân đoạn ngữ khá chính xác), dùng từ điển song ngữ Pháp-Việt và tính chất biên của các từ được dịch để khử nhập nhằng về vị trí của các từ trong phân đoạn ngữ và xác định vùng biên của từng phân đoạn ngữ tiếng Pháp tương ứng với từng phân đoạn ngữ trong câu tiếng Việt. Từ đó, xây dựng được các phân đoạn ngữ gióng hàng Pháp-Việt đồng thời đưa ra một mô hình dịch máy thống kê Pháp-Việt kết hợp với kết quả gióng hàng phân đoạn ngữ này nhằm mục đích cải thiện chất lượng dịch cơ sở.

3. HƯỚNG TIẾP CẬN CỦA BÀI BÁO

3.1. Phương pháp gióng hàng phân đoạn ngữ

Trong bài báo này, sẽ sử dụng một bộ ngữ liệu song ngữ Pháp-Việt đã được gióng hàng ở cấp độ câu. Các câu tiếng Pháp được gán nhãn từ loại (POS tagging) và nhãn ranh giới ngữ (chunk tagging) bằng công cụ SEM đã được huấn luyện với ngữ liệu French Tree Bank [4] dùng giải thuật Conditional Random Fields (CRF) và từ điển song ngữ Pháp-Việt được sử dụng để tham chiếu kết quả dịch của một từ tiếng Pháp sang tiếng Việt.

Ta xác định các phân đoạn ngữ trong ngữ liệu tiếng Pháp. Sau đó dự đoán, thông qua kết quả dịch từ của từ điển song ngữ, những ranh giới phân đoạn ngữ trong câu tiếng Việt tương ứng với từng phân đoạn ngữ trong câu tiếng Pháp. Sự khử nhập nhằng về ranh giới ngữ trong tiếng Việt sẽ được giải quyết dựa trên vị trí dịch của các từ liền kề trong cùng phân đoạn ngữ tiếng Pháp.

Quy trình thực hiện gồm 5 bước sau đây:

Bước 1: Ngữ liệu song ngữ Pháp-Việt sẽ được tách ra thành hai tập ngữ liệu đơn ngữ. Tiếp đến tiến hành phân đoạn ngữ cho ngữ liệu tiếng Pháp. Các phân đoạn có số lượng từ nhỏ hơn ngưỡng θ thì sẽ được nhóm lại với phân đoạn liền kề. Các nhóm này sẽ được đánh nhãn thứ tự tăng dần.

Bước 2: Tham chiếu trong từ điển song ngữ Pháp-Việt tìm các từ tiếng Việt là từ dịch của tập hợp từ trong các phân đoạn ngữ tiếng Pháp. Nếu tìm thấy trong từ điển này thì từ tiếng Việt sẽ được gán cùng nhãn phân đoạn ngữ với từ tiếng Pháp đang xét.

Bước 3: Khử nhập nhằng cho các từ tiếng Việt có nhiều nhãn thứ tự dựa trên vị trí dịch của các từ liền kề trong cùng phân đoạn ngữ tiếng Pháp.

Bước 4: Thiết lập các phân đoạn ngữ trong ngữ liệu tiếng Việt.

Bước 5: Tạo các cặp câu được gióng hàng phân đoạn ngữ Pháp-Việt dựa trên các nhãn thứ tự trùng nhau trong cùng một cặp câu song ngữ Pháp-Việt.

3.1.1. Gán nhãn phân đoạn ngữ (chunk tagging) cho tiếng Pháp

Ngữ được phân đoạn ngữ là những thành phần liên tục và tạo nên cấu trúc cú pháp của một câu. Để thực hiện việc gán nhãn phân đoạn ngữ tiếng Pháp, chúng tôi áp dụng mô hình học Conditional Random Fields (CRF) được giới thiệu bởi [6]. Mô hình học này là một mô hình xác suất cho phép gán nhãn các chuỗi dữ liệu tuyến tính. Hơn nữa, nó cho phép kết hợp cho một quan sát x một nhãn y dựa trên một tập hợp ví dụ đã được gán nhãn (x, y). Trong trường hợp này:

Cho $x = (x_1, x_2, x_3, ..., x_k)$ là tập hợp những dữ liệu đầu vào được quan sát hay nói cách khác x là một chuỗi đơn vị từ vựng tương ứng với nhãn từ loại (POS).

Và $y = (y_1, y_2, y_3, ..., y_k)$ là tập hợp những trạng thái hay nói cách khác y là một chuỗi các nhãn BIO tương ứng kết hợp từng loại phân đoạn ngữ (chunk). Mô hình CRF định nghĩa xác suất có điều kiện của một chuỗi trạng thái, biết rằng với một chuỗi đầu vào cho trước, bằng công thức như sau

$$p(y/x) = \frac{1}{Z(x)} \prod \exp(\sum_k \lambda_k f_k(y, x, c)) \quad (1)$$

trong đó:

$Z(x)$ là hệ số chuẩn hoá, được định nghĩa rằng tổng trên y của tất cả các xác suất $p(y/x)$ đối với một giá trị x nhất định được gán giá trị bằng 1 trong trường hợp này.

ζ là tập hợp các phần tử con trên y. Các phần tử này bao gồm hoặc duy nhất một nút đơn lẻ hoặc một cặp các nút liền kề.

f_k là hàm đặc trưng (features) được định nghĩa trong mỗi phần tử con c, và thường được chọn để trả về giá trị nhị phân 0 hoặc 1. Theo định nghĩa, giá trị của các hàm đặc trưng này có thể phụ thuộc vào các nhãn y tồn tại trong một phần tử con c bất kỳ cũng như giá trị của các nhãn từ loại x trong dữ liệu đầu vào.

λ_k là trọng số ở vị trí k, điều chỉnh giá trị tối ưu nhất ứng với mỗi hàm đặc trưng f_k.

Trong ngữ liệu tiếng Pháp, mỗi từ sau khi được gán nhãn từ loại, sẽ được gán nhãn ranh giới ngữ kết hợp với một nhãn theo mô hình BIO (Begin, In, Out). Mô hình BIO cho phép đánh dấu giới hạn ranh giới ngữ. Một ngữ bao gồm nhiều từ. Từ đầu tiên sẽ được đánh nhãn B, tiếp theo là I. Nhãn O được gán cho các từ không thuộc bất kỳ ngữ nào hoặc nằm riêng lẻ trong một câu. Ví dụ ta có hai trường hợp gán nhãn phân đoạn ngữ như sau:

(a) (La commercialisation efficace)NP est plus exigeante.

(b) (La commercialisation efficace)NP (est)VN (plus exigeante)AP.

Kết hợp mô hình BIO, ta sẽ có kết quả:

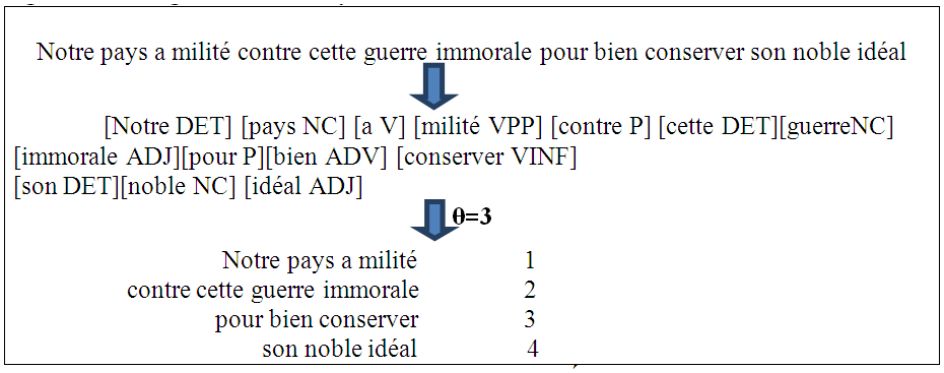
(a') La/B-NP commercialisation/I-NP efficace/I-NP est/O plus/O exigeante/O.

(b') La/B-NP commercialisation/I-NP efficace/I-NP est/B-VN plus/B-AP exigeante/I-AP.

Công cụ gán nhãn từ loại và gán nhãn phân đoạn ngữ cho tiếng Pháp là công cụ SEM. SEM (viết tắt của Segmenteur-Étiqueteur Markovien) là một bộ đánh nhãn được huấn luyện trên tập French Tree Bank (Abeillé, 2003) [4] của Đại Học Paris 7, Pháp.

3.1.2. Rút trích các phân đoạn ngữ tiếng Pháp

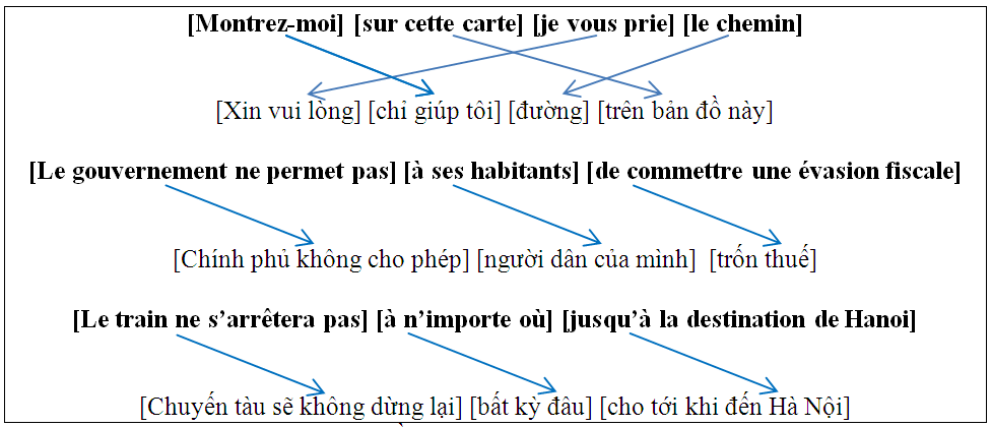
Một ngữ được xác định là đúng khi và chỉ khi những ranh giới của nó và loại của nó đúng. Kết hợp mô hình BIO và thuật toán Conditional Random Fields trên tập huấn luyện French Tree Bank của Đại Học Paris 7, Pháp, ta có mô hình gán nhãn phân đoạn ngữ cơ sở (basechunking) cho tiếng Pháp. Ta thấy nếu chỉ sử dụng mô hình gán nhãn phân đoạn ngữ cơ sở này thì kết quả các ngữ tìm thấy rất thấp. Điều đó kéo theo hiệu quả thấp của việc khử nhập nhằng. Do đó ta thực hiện thao tác đánh thứ tự các nhãn từ trái sang phải cho các ngữ. Tiếp đến, dựa trên giới hạn biên của các phân đoạn ngữ, chúng được ghép lại chung với nhau theo tiêu chí số lượng từ tối thiểu trong ngữ, với ngưỡng $\theta = 3$ trong ví dụ sau đây:



Hình 1. Ví dụ rút trích phân đoạn ngữ trong tiếng Pháp với ngưỡng $\theta = 3$

3.1.3. Xác định vùng biên cho các phân đoạn ngữ tiếng Việt

Theo nhận xét của [5], khi dịch một câu từ tiếng Anh sang tiếng Hoa thì các từ trong một phân đoạn ngữ tiếng Anh có xu hướng được dịch thành một cụm các từ tiếng Hoa liền kề nhau và điều này cũng đúng trong cặp ngôn ngữ Pháp-Việt. Có nghĩa là một phân đoạn ngữ tiếng Pháp sẽ được gióng hàng với một phân đoạn ngữ trong tiếng Việt dựa trên các mục từ được xem là dịch của nhau tham chiếu qua từ điển song ngữ Pháp-Việt.



Hình 2. Một vài ví dụ về gióng hàng phân đoạn ngữ Pháp-Việt

Như vậy, để xác định các phân đoạn ngữ tiếng Việt tương ứng với các phân đoạn ngữ tiếng Pháp, đầu tiên ta cần gán thứ tự nhãn phân đoạn ngữ tiếng Pháp dựa theo mô hình gán nhãn CRF và mô hình BIO. Trong phân đoạn ngữ tiếng Việt, khi ta xét một từ được xem là từ dịch của một từ trong tiếng Pháp thì từ đó sẽ được gán cùng nhãn với từ tiếng Pháp

đang xét. Trong ví dụ minh họa hình 3 dưới đây, khi tra từ điển song ngữ Pháp-Việt, ta sẽ tìm ra "ma" = "của tôi" (tính từ sở hữu), "ne voient jamais" = "không hề thấy" (động từ voir và ne...jamais). Và trong câu tiếng Pháp, "ma" có số thứ tự 2 nên "của tôi" sẽ có số thứ tự là 2. Tương tự "ne voient jamais" có số thứ tự 5 nên cụm từ "không hề thấy" sẽ được gán thứ tự là 5. Tuy nhiên, trong trường hợp một từ tiếng Việt có nhiều nhãn trong nhiều phân đoạn ngữ tiếng Pháp, ta cần khử nhập nhằng chúng bằng cách xét vùng biên là các nhãn thứ tự của các từ liền kề. Cuối cùng, các từ tiếng Việt có cùng nhãn thứ tự trong câu sẽ được nhóm lại thành các phân đoạn ngữ.

Bước 1: Xác định các phân đoạn ngữ của câu tiếng Pháp và gán số thứ tự tương ứng
[Toutes les personnes 1] [dans ma famille 2] [ont affirmé 3] [à un policer 4]
[qu' ils ne voient jamais un voleur 5]

Bước 2: Gán nhãn thứ tự các từ trong câu tiếng Việt được xem là dịch của từ trong câu tiếng Pháp
tất_cả (1) mọi (1) người (1) trong (2) gia_đình (2) của_tôi (2) đã quả_quyết (3) với (4) một (4) người (1,4) công_an (4) rằng (5) họ (5) không hề thấy (5) một (4,5) kẻ_trộm (5)

Bước 3: Khử nhập nhằng nhiều nhãn thứ tự của các từ trong câu tiếng Việt
tất_cả (1) mọi (1) người (1) trong (2) gia_đình (2) của_tôi (2) đã quả_quyết (3) với (4) một (4) người (4) công_an (4) rằng (5) họ (5) không hề thấy (5) một (5) kẻ_trộm (5)

Bước 4: Thiết lập các phân đoạn ngữ trong câu tiếng Việt
[tất_cả mọi người 1] [trong gia_đình của tôi 2] [đã quả_quyết 3] [với một người công_an 4] [rằng họ không hề thấy một kẻ_trộm 5]

Bước 5: Tạo các cặp câu được gióng hàng phân đoạn ngữ Pháp-Việt

tất_cả mọi người	Toutes les personnes
trong gia đình của tôi	dans ma famille
đã quả quyết	ont affirmé
với một người công an	à un policier
rằng họ không hề thấy một kẻ trộm	qu' ils ne voient jamais un voleur

Hình 3. Ví dụ minh họa kết quả mong muốn rút trích các cặp phân đoạn ngữ song ngữ Pháp-Việt

3.2. Mô hình hệ thống dịch máy thống kê kết hợp thông tin phân đoạn ngữ

Ta có mô hình hệ thống dịch máy thống kê kết hợp thông tin phân đoạn ngữ tổng hợp như sau.

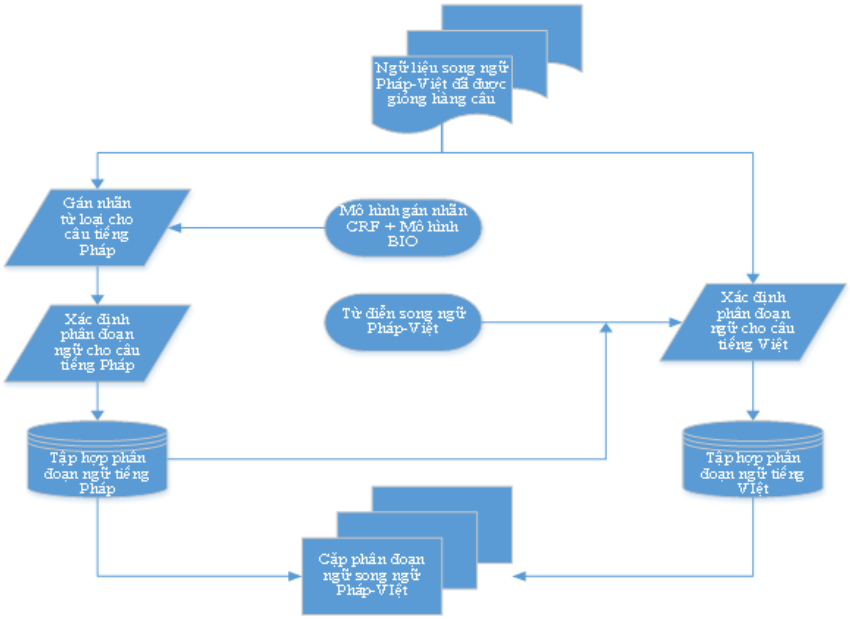
Sau giai đoạn xây dựng tập hợp các cặp phân đoạn ngữ song ngữ Pháp-Việt, ta thực hiện việc gióng hàng từ cho các cặp phân đoạn ngữ này rồi tiến hành huấn luyện, xây dựng mô hình dịch (translationmodel) và mô hình ngôn ngữ (languagemodel) dựa trên tập ngữ liệu song ngữ trên.

Trong hệ thống dịch máy thống kê, mô hình dịch máy thống kê dựa trên ngữ được xây dựng trên việc huấn luyện các cặp câu song ngữ để tạo nên mô hình dịch và dữ liệu đơn ngữ để tạo nên mô hình ngôn ngữ. Trong quá trình huấn luyện, các cặp câu song ngữ sẽ được gióng hàng ở cấp độ từ trước tiên và trong bộ giải mã (decoder), kết quả dịch sẽ được kết hợp từ hai mô hình dịch và mô hình ngôn ngữ. Ở đây, ta tích hợp các thông tin phân đoạn ngữ Pháp-Việt vào quá trình huấn luyện mô hình dịch.

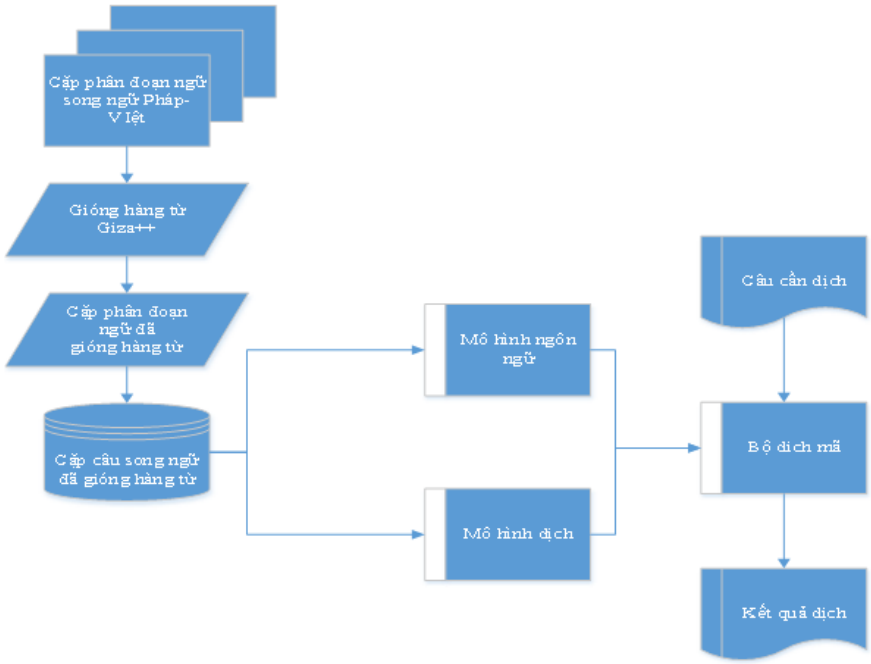
4. KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1. Ngữ liệu và công cụ

Kho ngữ liệu song ngữ thử nghiệm bao gồm 10.000 cặp câu song ngữ Pháp-Việt được tổng hợp từ các sách giáo khoa đàm thoại tiếng Pháp và từ điển song ngữ Pháp-Việt với hơn 70.000



Hình 4a. Mô hình hệ thống dịch máy thống kê Pháp-Việt kết hợp gióng hàng phân đoạn ngữ



Hình 4b. Mô hình hệ thống dịch máy thống kê Pháp-Việt kết hợp gióng hàng phân đoạn ngữ

mục từ.

Ngữ liệu song ngữ gồm 10.000 cặp câu được chuẩn hóa theo các tiêu chí sau đây:

- Đồng nhất về mặt nội dung và về mặt hình thức: mỗi câu trên một dòng duy nhất và kết thúc bằng dấu câu rõ ràng.
- Đều được kiểm lỗi chỉnh tả, được loại bỏ các câu trùng nhau.
- Các câu và các phân đoạn ngữ có độ dài từ 1-20 từ.

Công cụ gán nhãn từ loại và gán nhãn phân đoạn ngữ cho tiếng Pháp là công cụ SEM. Các môđun dịch trong bài báo này được xây dựng bằng cách áp dụng bộ dịch máy thống kê Moses (Koehn and al., 2007) [15] với các tham số cài đặt mặc định. Công cụ Giza++ được sử dụng để gióng hàng từ, và thuật toán "grow-diag-final-and" được chọn. Công cụ huấn luyện mô hình ngôn ngữ là SRILM và công cụ huấn luyện mô hình dịch là hệ thống dịch dựa trên ngữ.

Chia tập ngữ liệu thành ba tập: tập huấn luyện (training set), tập phát triển (developing set) và tập đánh giá(testing set). Trong đó, sử dụng 90% các cặp câu song ngữ làm tập huấn luyện, 5% cho tập phát triển và 5% còn lại cho tập đánh giá.

4.2. Kết quả của hệ thống

Các thử nghiệm sau để đánh giá một cách thủ công chất lượng của công cụ gióng hàng phân đoạn ngữ cho cặp câu Pháp-Việt của bài báo đề xuất bằng cách tính toán độ chính xác, độ bao phủ, hệ số cân bằng và nhờ chuyên gia ngôn ngữ.

Mô hình 1. (Baseline) Hệ thống dịch máy thống kê cơ sở không dùng thêm tri thức ngôn ngữ, chỉ được tách tokens, tách khoảng trắng.

Mô hình 2. Hệ thống dịch máy thống kê chỉ chứa duy nhất các cặp phân đoạn ngữ, mỗi cặp phân đoạn ngữ Pháp-Việt được coi như một cặp câu song ngữ. Ngữ liệu đầu vào cho mô hình dịch là các cặp phân đoạn ngữ này. Tiếp đến chúng được gióng hàng từ bằng công cụ Giza++. Sau đó chúng được huấn luyện và rút trích thành bảng chuyển ngữ (phrase table).

Mô hình 3. Hệ thống dịch máy thống kê tích hợp thêm các cặp phân đoạn ngữ trong mô hình cơ sở, có nghĩa là gộp lại tất cả cặp phân đoạn ngữ Pháp-Việt và các câu song ngữ cơ sở lại thành một tập đồng nhất. Như vậy, tập ngữ liệu đầu vào bao gồm hai phần. Phần thứ nhất là các câu song ngữ cơ sở và phần thứ hai là các cặp phân đoạn ngữ đã được gióng hàng. Tiếp đến kho ngữ liệu này sẽ được gióng hàng ở cấp độ từ với công cụ Giza++. Tương tự mô hình 2, sau đó chúng được huấn luyện và rút trích thành bảng chuyển ngữ.

*B*ảng 1. Kết quả thực nghiệm dịch máy Pháp-Việt kết hợp các mô hình

MÔ HÌNH	BLEU	NIST	TER
Mô hình 1	24.39%	3.224	69.29%
Mô hình 2	23.57%	2.689	74.19%
Mô hình 3	25.76%	3.188	68.48%

Nhận xét: Theo thống kê, tỉ lệ về số lượng từ chính tả (tương đương với tiếng của tiếng Việt hay từ của tiếng Anh) trong văn bản giữa ngôn ngữ Anh-Việt là 1:1,55. Có nghĩa là mỗi từ tiếng Anh thường được dịch thành 1,55 tiếng của tiếng Việt. Và đối với cặp ngôn ngữ Pháp-Việt, [9] đã thống kê tỉ lệ của số từ chính tả Pháp-Việt là 0,8:1,3 . Do đó, trong mô hình

2, ta chỉ giữ lại các cặp câu song ngữ mà tỉ lệ độ dài nằm trong ngưỡng $\alpha = [0,8 - 1,3]$ và nhận thấy rằng khi loại bỏ các cặp câu Pháp-Việt không thỏa điều kiện trên, chất lượng dịch bị giảm xuống. Với kết quả thực nghiệm trên cho thấy, hướng tiếp cận kết hợp thông tin phân đoạn ngữ trong dịch máy thống kê Pháp-Việt đã đạt hiệu quả cao, đặc biệt với mô hình 3 thì điểm BLEU tăng hơn gần 2 điểm so với mô hình cơ sở. Điểm BLEU của mô hình 3 đạt kết quả cao nhất với giá trị 25,76%.

Phân tích ảnh hưởng của bộ phân đoạn đối với độ chính xác của dịch máy:

Ưu điểm: nếu bộ phân đoạn ngữ được xử lý tốt sẽ làm giảm tỷ lệ sai trong gióng hàng từ hay nói một cách khác, điều đó sẽ làm tăng độ chính xác của dịch máy.

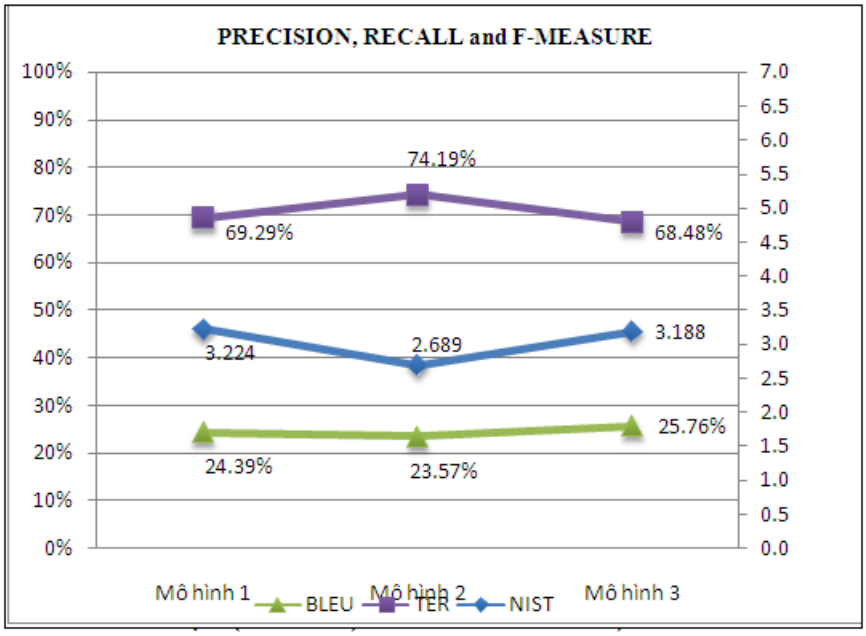
Nhược điểm: nếu bộ phân đoạn ngữ chứa quá nhiều phân đoạn ngữ nhỏ, độ bao phủ cao, thì điều đó sẽ làm tăng sự nhập nhằng trong việc chọn lựa các từ hay ngữ chính xác trong một số trường hợp trong hệ thống dịch máy.

Phân tích lỗi:

1) Lỗi xác định chunk trong câu tiếng Pháp:

Các ngữ trong câu tiếng Pháp	Các ngữ trong câu tiếng Việt
6 | le fils	6 | cậu con_ trai đã
6 | a hérité	6 | thừa_ hưởng
6 | de toute la	**6 | tất_ cả**
6 | fortune	**6 | tài_ sản**
6 | à la mort de	6 | sau cái chết của
6 | ses parents	6 | bố_ mẹ mình

Đây là lỗi của tiếng Pháp khi xác định ngữ (chunk). Điều đó kéo theo lỗi xác định ngữ sai trong câu tiếng Việt. Kết quả mong muốn sau khi xác định các ngữ trong câu tiếng Pháp như



Hình 5. Biểu đồ so sánh kết quả các mô hình dịch thống kê Pháp-Việt

sau:

Le fils | a hérité | de toute la fortune | à la mort de | ses parents.

2) Lỗi trong dịch máy:

Câu cần dịch: Nguyen Xuan Ly, chef du département de police de l'environnement, a déclaré qu'il s'agissait d'un nouveau type de crime dont les importateurs ont déclaré à la douane que ce sont des matériaux pour la production, mais ils sont en fait des déchets toxiques.

Cơ sở: Nguyễn Xuân Lý, người đứng đầu của Cục Cảnh sát môi trường, nói rằng điều này là một loại mới của crime, trong đó các nhà nhập khẩu tuyên bố tại hải quan mà là những tài liệu cho sản xuất, nhưng họ đang ở trong thực tế chất thải độc hại.

Mô hình: Nguyễn Xuân Lý, người đứng đầu của Cục Cảnh sát môi trường, nói rằng điều này là một loại tội phạm mới, trong đó các nhà nhập khẩu tuyên bố tại hải quan rằng đây là những tài liệu cho sản xuất, nhưng họ đang ở trong thực tế chất thải độc hại.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đề xuất một hướng tiếp cận mới bằng cách kết hợp phương pháp gióng hàng phân đoạn ngữ trong dịch máy thống kê cho cặp ngôn ngữ Pháp-Việt để giải quyết bài toán câu dài một cách hiệu quả. Kết quả thực nghiệm của mô hình hệ thống cho thấy độ đo điểm BLEU đã tăng hơn gần 2 điểm so với mô hình cơ sở.

Ta nhận thấy rằng việc tích hợp thêm các tri thức ngôn ngữ đã có những đóng góp đáng kể trong việc nâng cao chất lượng dịch Pháp-Việt. Đây là những kết quả đáng quan tâm trong quá trình phát triển nghiên cứu đối với cặp ngôn ngữ Pháp-Việt ở Việt Nam nói riêng và trên thế giới nói chung. Hơn nữa, việc tăng độ chính xác của các gióng hàng phân đoạn ngữ sẽ giúp ích trong việc cải thiện đáng kể chất lượng dịch. Trong tương lai, nhóm tác giả sẽ tập trung vào giai đoạn cải tiến trật tự từ trong các phân đoạn ngữ để tăng độ chính xác của gióng hàng từ trong các cặp phân đoạn ngữ được rút trích bởi hệ thống.

TÀI LIỆU THAM KHẢO

[1] Philipp Koehn, Franz Josef Och and Daniel Marcu, Statistical phrase-based translation, *Proceedings of the HLT-NAACL 2003 Conference*, Edmonton, Alberta, Canada, 2003 (127–133).

[2] Eugene Charniak, Kevin Knight, and Kenji Yamada, Syntax-based language models for statistical machine translation, *Proceedings of the Ninth Machine Translation Summit of the International Association for Machine Translation*, New Orleans, Louisiana, September, 2003 (id=#21).

[3] Isabelle Tellier, Denys Duchier, Iris Eshkol, Arnaud Courmet, Mathieu Martinet: Apprentissage automatique d'un chunker pour le français, Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, Grenoble, 4 au 8 Juin 2012 Copyright 2012 ATALA & AFCP (pp.431–438).

[4] Anne Abeillé, Lionel Clément et François Toussenel: Building a treebank for french, *A. ABEILLE,* éditeur: Treebanks. Kluwer, Dordrecht, 2003 (165–187).

[5] Sun Le, Jin Youbing, Du Lin, Sun Yufang, Word alignment of english-chinese bilingual corpus based on chunks, *Proc. 2000 EMNLP and VLC*, Hong Kong, 2000 (110–116).

[6] Vinh Van Nguyen, Thai Phuong Nguyen, Akira Shimazu, and Minh Le Nguyen, Reordering phrase-based machine translation over chunks, *IEEE International Conference on Research, Innovation and Vision for the Future - RIVF,* Ho Chi Minh city, Vietnam, 2008 (114–119).

[7] Francisco Nevado, Francisco Casacuberta, Enrique Vidal: Parallel corpora segmentation using anchor words, *Proc. of the EAMT/EACL Workshop on MT and Other Language Technology Tools,* Budapest, Hungary, April, 2003 (12–17).

[8] John Lafferty, Andrew McCallumand Fernando Pereira: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *The Proceedings of International Conference on Machine Learning (ICML),* Pittsburgh, Pennsylvania, USA, 2001 (282–289).

[9] Thi-Ngoc-Diep Do ,Viet-Bac Le, Brigitte Bigi, Laurent Besacier, Eric Castelli: Mining a comparable text corpus for a Vietnamese - French statistical machine translation system *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March, 30-31, 2009 (165–172).

[10] LE Hong Phuong, "TAG – Tree Adjoining Grammar" PhD thesis, Loria, Nancy, France, 2009.

[11] Nicolas BOFFO, "Formation de la temporalité en Vietnamien pour la traduction automatique", PhD thesis in process, France-Vietnam, 2012.

[12] NGUYEN Thi Minh Huyen, "Outils et ressources linguistiques pour l'alignement de textes multilingues français-vietnamien" PhD thesis, Nancy, France, 2006.

[13] TRAN Tuan Duc, DYALANG, Université de Rouen, France: Système de recherche d'information medical par croisement de langue: Vietnamien–Français–Anglais, *GLOTTOPOL* (8) (2006) 45–52.

[14] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: open source toolkit for statistical machine translation, *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, international conference*, Prague, Czech Republic, 2007 (177–180).

[15] Ngoc Tan LE, Ngoc Tien LE, Dien Dinh, An approach of chunk alignment for French-Vietnamese Bilingual corpora, *The Proceedings of International Journal of Computer Science Issues (IJCSI), Vol. 10, Issue 2*, Republic of Mauritius, 2013 (111–117).

Ngày nhận bài 04 - 5 - 2013
Nhận lại sau sửa ngày 22 - 11 - 2013