Chuyển đổi cây cú pháp cụm danh từ tiếng Anh theo tiếng Việt

Transferring Syntax Trees of English Noun Phrases via Vietnamese Syntax

Nguyễn Chí Hiếu

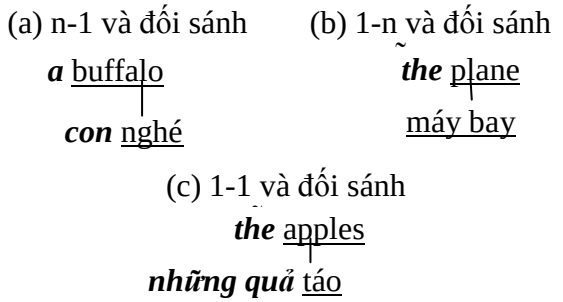
***Abstract**: Transferring syntax trees is one of the key tasks of machine translation systems. To transfer syntax trees, they can be performed by different models of the statistical translation method, rule-based translation method, or a hybrid of statistics and rule-based translation method [9]. In this paper we present a combination of bilingual corpus and knowledge-based method, which transfers syntax trees of English Base Noun Phrases via the Vietnamese syntax from a pair of bilingual sentences to identify anchor points. Our combination technique can help to invert word order in noun phrases of the source language to suit those of target language and improve the performance of miss-alignment, null-alignment, overlap and conflict projection of the existing methods. The proposed technique can be easily applied to other language pairs. Experiment on pairs of sentences in the English-Vietnamese bilingual corpus showed that our proposed method is satisfactory.*

***Keywords**: Npbase, đối sánh từ, NLP*

I. GIỚI THIỆU

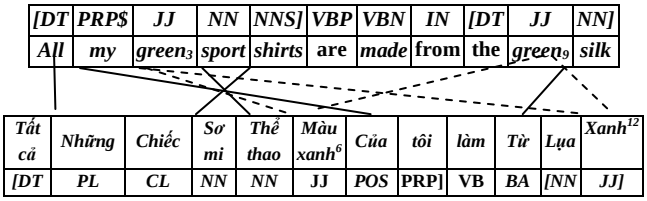
Rút trích cụm danh từ (*NPbase*) từ ngữ liệu song ngữ được nhiều nhà nghiên cứu về xử lý ngôn ngữ tự nhiên đề xuất bằng các phương pháp khác nhau trong nhiều thập kỷ qua như J.Kupiec [7], H.Wantanabe [14], Yarowsky [15], W.Wang [13], Y.S.Hwang [6], Đ.Điền [2] và N.P.Thai [12]. Tuy nhiên việc nghiên cứu ngôn ngữ tự nhiên để giúp máy tính hiểu được ngôn ngữ con người vẫn còn là một thách thức và bài toán nhận biết cụm danh từ tiếng Việt trong cặp câu song ngữ vẫn còn là bài toán mở [4]. Bài báo này đề xuất một mô hình chuyển đổi cây phân tích cụm danh từ cơ sở tiếng Anh theo tiếng Việt nhằm giải quyết một số hạn chế của mô hình chiếu trên cơ sở đối sánh từ của Yarowsky [15] (sau này gọi là "*phương pháp Yarowsky*") để nhận biết cụm danh từ tiếng Việt [5].

Phương pháp Yarowsky cho phép các nước còn thiếu nguồn tài nguyên về xử lý ngôn ngữ tự nhiên (NLP) có thể thu được nhanh kết quả nghiên cứu thông qua một ngôn ngữ giàu tài nguyên nghiên cứu khác như tiếng Anh. Tuy nhiên hạn chế của phương pháp Yarowsky là chưa giải quyết vấn đề đối sánh rỗng, vấn đề chồng chéo và xung đột như ví dụ ở Hình 1 và 2.



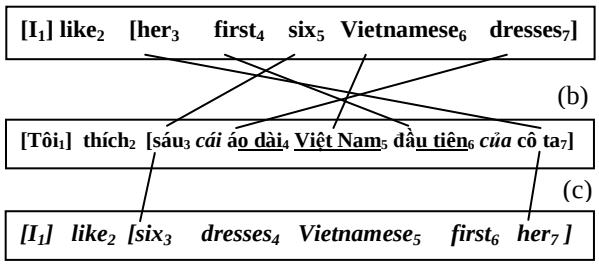
Hình 1. Ví dụ về một số loại đối sánh

Hình 1(a) cho chúng ta thấy, khi hai từ tiếng Anh đối sánh với một từ tiếng Việt người ta gọi là đối sánh *nhiều-một* (n-1). Hình 1(b) là ví dụ về một từ tiếng Anh đối sánh với hai từ tiếng Việt người ta gọi là đối sánh *một-nhiều* (1-n). Hình 1(c) là đối sánh *một-một* (1-1).



Hình 2. Ví dụ về chồng chéo và xung đột

Hình 2 là một ví dụ cho cặp câu Anh–Việt có sự chồng chéo. Trong ví dụ này, từ "green" xuất hiện nhiều lần trong câu ở vị trí thứ 3 và vị trí thứ 9. Do đó từ "green" có thể đối sánh với hai vị trí tương ứng trong câu tiếng Việt là "màu xanh" ở vị trí thứ 6 và "xanh" ở vị trí thứ 12, nên nó gây lên vấn đề chồng chéo và xung đột. (a)



Hình 3. Ví dụ về chuyển đổi trật tự từ để khắc phục vấn đề chồng chéo khi đối sánh

Giải pháp giải quyết cho vấn đề chồng chéo và xung đột là chuyển đổi trật tự từ tiếng Anh theo tiếng Việt trước khi đối sánh như ví dụ mô phỏng trên hình 3. Trên Hình 3(c) là câu tiếng Anh của Hình 3(a) đã được chuyển đổi trật tự từ trong cụm danh từ cơ sở tiếng Anh theo tiếng Việt. Bằng trực quan chúng ta dễ dàng nhận thấy rằng: Nếu chiếu theo câu trên Hình 3(c) với câu trên Hình 3(b), thì chúng ta sẽ khắc phục được vấn đề chồng chéo và xung đột. Vậy làm thế nào có thể giúp máy tính tự động thực hiện chuyển đổi sẽ được chúng tôi trình bày trong mục III. Trước tiên là mục phân rã cấu trúc cụm danh từ.

II. CẤU TRÚC CỤM DANH TỪ

Bảng 1. Cấu trúc cụm danh từ tiếng Anh [10]

định tố trước	định tố	định tố sau	bổ ngữ	từ trung tâm	bổ ngữ sau
all	the	three	good	books	on sale
Y1	Y2	Y3	Y4	Y5	Y6

Bảng 2. Cấu trúc cụm danh từ tiếng Việt [11]

Thành tố phụ trước			Trung tâm		Thành tố phụ sau	
tổng lượng	số lượng	tiểu từ cái	từ phân lớp	danh từ	cụm tính từ	chỉ định / sở hữu
Cả	sáu	cái	con	gà	đen	ấy
X1	X2	X3	X4	X5	X6	X7

Theo V.N.Tú [10], cụm danh từ tiếng Anh được phân tích thành 6 thành phần như luật (a). Theo T.H.Nguyen [11], cụm danh từ tiếng Việt được phân tích thành 7 thành phần như luật (b).

(a) NPe →Y1 Y2 Y3 Y4 Y5 Y6

(b) NPv →X1 X2 X3 X4 X5 X6 X7

Trong đó:

NPe: ký hiệu cụm danh từ tiếng Anh

NPv: ký hiệu cụm danh từ tiếng Việt

Yj: là từ loại của từ ở vị trí j , j = 1 đến 6

Xi: là loại ở vị trí thứ i, i = 1 đến 7

Cụm danh từ tiếng Anh bao gồm sáu thành phần (bốn thành phần phụ trước, một thành phần *trung tâm* và một thành phần phụ sau) và cụm danh từ tiếng Việt bao gồm bảy thành phần (bốn thành phần phụ trước, một thành phần *trung tâm* và hai thành phần phụ sau), đó là xét khả năng các thành phần đều xuất hiện. Trong thực tế có thể không xuất hiện đầy đủ các thành phần như vậy, ngoại trừ thành phần thứ 5 là bắt buộc (*Y5 trong tiếng Anh và X5 trong tiếng Việt)*.

Bảng 3. Cấu trúc của cụm danh từ Anh - Việt

Ngôn ngữ	Cấu trúc chung		
	Thành tố phụ trước	Danh từ trung tâm	Thành tố phụ sau
Anh	Y1 Y2 Y3 Y4	Y5	Y6
Việt	X1 X2 X3 X4	X5	X6 X7

Cấu trúc cụm danh từ tiếng Anh và tiếng Việt đều giống nhau ở chỗ có ba thành phần lớn là thành tố phụ trước, thành phần trung tâm, thành tố phụ sau (Bảng 3). Tuy nhiên, các yếu tố tạo nên thành tố giữa cụm danh từ tiếng Anh và tiếng Việt lại rất khác biệt và khá phức tạp, nhất là trong cụm danh từ tiếng Việt. Vì thế, chúng ta cần có các bước hợp lý để phân tích các thành tố của cụm danh từ vào các thành phần tương

ứng. Các bước này được chúng tôi đề xuất để phân tích cụm danh từ cơ sở tiếng Anh thành các thành phần từ Y1 đến Y6 thông qua các giải thuật 1, 2, 3 và 4 ở mục kế tiếp.

Để phân rã cụm danh từ cơ sở tiếng Anh thành các thành phần, chúng tôi thực hiện theo trình tự như sau. Trước tiên, chúng tôi phải phân tích cụm danh từ cơ sở tiếng Anh thành ba thành phần là phụ trước (Y1 Y2 Y3 Y4), trung tâm (Y5), phụ sau (Y6). Trong đó thành phần Y5 được ưu tiên xác định đầu tiên, Nếu xác định được thành phần Y5, thì thành phần phía bên phải của Y5 sẽ là thành phần Y6 và các thành phần phía bên trái của Y5 sẽ là các thành phần Y1, Y2, Y3, Y4. Cách làm này sẽ loại bỏ được sự xung đột về từ loại giữa Y4, Y5 và Y6. Tiếp theo, chúng tôi phân tích tiếp các thành tố phụ truớc thành các thành phần từ Y1 đến Y4. Sự xung đột giữa Y1 và Y2 về từ loại DT (*Phụ lục A*) được xét thêm bởi từ đứng đầu cụm danh từ cơ sở (Y1 là vị trí của từ chỉ tổng lượng). Thành phần Y2 lại được phân tích thành 2 tiểu thành phần là Y2a và Y2b. Y2a là vị trí của mạo từ như "a", "an", "the". Y2b là vị trí của tính từ sở hữu *"my", "his", "our",…*, từ chỉ thị *"this", "these", "that", "those"*, các từ *"some", "any", "many", "much", "enough"*. Thành phần Y3 cũng được phân tích thành 2 tiểu thành phần là Y3a và Y3b. Y3a là vị trí của từ chỉ số thứ tự. Y3b là vị trí của từ chỉ số đếm. Thành phần Y6 cũng được phân tích thành hai tiểu thành phần là Y6a và Y6b. Y6a là vị trí của từ có từ loại CD mà liền kề trước nó là các từ loại tiền tệ ví dụ như $ CD, # CD. Y6b là các từ loại còn lại.

Ở vị trí của X2, X3, X4 có thể là những từ đặc trưng trong tiếng Việt (X2: *"những", "các"*, X3: "***cái***", X4: từ phân lớp). X2 vẫn có thể tương đương chức năng với tiểu thành phần Y2a trong trường hợp các từ là mạo từ như *"a", "an" "the"*. Ngoài ra, X2 còn tương đương chức năng với tiểu thành phần của Y3 là Y3b, khi ở vị trí Y3 từ chỉ thứ tự và từ chỉ số đếm cùng xuất hiện, X2 cũng tương đương chức năng với tiểu thành phần Y6a. X1 tương đương chức năng với Y1. X6 tương đương chức năng với Y4. Ngoài ra, X6 còn tương đương chức năng với tiểu thành phần của Y3 là Y3a, và X6 cũng tương đương chức năng với Y6b. Cuối cùng X7 tương đương chức năng với tiểu thành phần của Y2 là Y2b như Bảng 4.

Bảng 4. So sánh các thành phần trong cụm danh từ

Sự tương đương chức năng giữa các thành phần trong cụm danh từ tiếng Việt với cụm danh từ tiếng Anh	
Việt	*Anh*
X1	Y1
X2 (một|ϕ)	Y2a (a, an, the), Y3b, Y6a
X3 (cái)	ϕ
X4 (từ phân lớp)	ϕ
X5	Y5
X6	Y3a, Y4, Y6b
X7	Y2b

Cấu trúc chi tiết cụm danh từ tiếng Anh và tiếng Việt có nhiều điểm rất khác nhau như đã trình bày ở trên. Phần tiếp theo sẽ trình bày chi tiết mô hình đề xuất.

III. MÔ HÌNH CHUYỂN ĐỔI TRẬT TỰ TỪ

Trên cở sở nghiên cứu về đối sánh ngôn ngữ của T.H.Nguyen [11] và V.N.Tú [10] về cụm danh từ song ngữ Anh-Việt, chúng tôi đã phân tích cụm danh từ tiếng Anh, cụm danh từ tiếng Việt thành các thành phần theo vị trí chức năng của chúng như bảng 4. Với vị trí và chức năng của mỗi thành phần đã phân tích ở trên cho phép chúng tôi thiết kế một giải thuật hiệu quả để chuyển đổi trật tự từ của cụm danh từ cơ sở tiếng Anh theo tiếng Việt.

Bài toán chuyển đổi trật tự từ trong cụm danh từ cơ sở tiếng Anh theo tiếng Việt bao gồm các bài toán sau:

- Xác định từ trung tâm (Y5) của cụm danh từ cơ sở tiếng Anh;
- Phân tích các thành phần (Y3 đến Y6) khi có liên từ xuất hiện trong chúng;
- Chuyển đổi trật tự từ của cụm danh từ cơ sở tiếng Anh theo tiếng Việt.

(i) Bài toán xác định từ trung tâm

Từ trung tâm là từ quan trọng nhất trong các cụm từ nói chung và cụm danh từ nói riêng. Xét về lý

thuyết, từ trung tâm của cụm danh từ là danh từ, nhưng trong thực tế ngôn ngữ thì từ trung tâm còn bao hàm các loại từ khác, ví dụ câu: *this book costs $ 100*.

(S (NP this/DT book/NN) (VP costs/VBP (NP $/$ 100/CD)))

(a) $/$ 100/CD	(b) #/# 13/CD
Y5	Y5
một trăm/CD đô la Mỹ/$	mười ba/CD bảng Anh/#

Qua khảo sát trên ngữ liệu từ chương 15 đến 18 trong Penn Treebank [16] của luận án [4], thì từ trung tâm trong cụm danh từ cơ sở bao gồm các từ có từ loại là {*NN, NNP, NNPS, NNS, JJR, POS, $, #, CD, JJ, JJS, RB*}(ký hiệu các từ loại - Phụ lục A).

Trên cơ sở khảo sát, thống kê ngữ liệu của Penn Treebank của tài liệu [4] và tham khảo ý tưởng trong giải thuật xác định từ trung tâm của Collins [1], chúng tôi xây dựng giải thuật để tìm từ trung tâm của cụm danh từ cơ sở tiếng Anh như sau.

Giải thuật 1: **Tìm từ trung tâm trong cụm danh từ cơ sở tiếng Anh (Y5)**

Nhập: Cụm danh từ tiếng Anh đã gán nhãn từ loại
Xuất: Từ trung tâm (Y5)

Phương pháp:
Nếu cụm danh từ cơ sở có một phần tử thì thực hiện bước 6; Ngược lại thì thực hiện bước 1;
Bước 1: *Tìm từ phải qua trái, nếu gặp từ có từ loại trong tập {NN, NNP, NNPS, NNS} thì Y5 là từ vừa tìm thấy và thực hiện bước 6; Ngược lại thì thực hiện bước 2;*
Bước 2: *Tìm từ phải qua trái, nếu gặp từ có từ loại trong tập {$, #} thì Y5 là từ vừa tìm thấy và thực hiện bước 6; Ngược lại thì thực hiện bước 3;*
Bước 3: *Tìm từ phải qua trái, nếu gặp từ có từ loại là CD thì Y5 là từ vừa tìm thấy và thực hiện bước 6, ngược lại thì thực hiện bước 4;*
Bước 4: *Tìm từ phải qua trái, nếu gặp từ có từ loại trong tập {JJ, JJS, RB, JJR} thì Y5 là từ vừa tìm thấy và thực hiện bước 6, ngược lại thì thực hiện bước 5;*
Bước 5: *Y5 là từ tận cùng bên phải của cụm danh từ cơ sở và trả về cho chương trình thành phần Y5;*
Bước 6: *trả về thành phần Y5;*

(ii) Bài toán phân tích các thành phần khi có liên từ

Sự xuất hiện của liên từ trong cụm danh từ cơ sở làm tăng thêm tính nhập nhằng khi phân tích các thành phần trong cụm danh từ cơ sở. Ngoại trừ hai vị trí Y1 (Y1 có từ loại thuộc tập {PDT, DT}) và Y2 (Y2 có từ loại thuộc tập {DT, PRP$}) là không có sự xuất hiện của liên từ, bốn vị trí còn lại đều có thể có sự xuất hiện của liên từ. Chi tiết giải quyết bài toán này trình bày dưới đây.

Giải thuật 2: **Xác định từ trung tâm có liên từ và tách cụm danh từ cơ sở**

Nhập: Cụm danh từ cơ sở tiếng Anh đã gán nhãn từ loại
Xuất: phần phụ trước (Y1Y2Y3Y4), trung tâm (Y5) và phần phụ sau (Y6a, Y6b)

Phương pháp:
Bước 1: *Chạy giải thuật 1 để xác định từ trung tâm (Y5) ở vị trí h trong cụm danh từ cơ sở gồm n từ (h ≤ n)*
Bước 2: *Mở rộng từ trung tâm và tách cụm danh từ cơ sở như sau:*

```
begin
    foundCC = true; Y5= từ ở vị trí h; h2 = h;
    while ((h>2) and (foundCC))
   begin  /*nhiều liên từ trong thành phần Y5*/
    if (h-1 là vị trí của từ có từ loại là CC) then
    begin
      Y5 = bộ hai các từ ở vị trí ( h-1, h) + Y5;
      h = h-2;
    end else    foundCC = false;
end
if (h>1) then tách các vị trí từ 1 đến h-1 vào thành tố phụ trước(Y1Y2Y3Y4);
if (h2<n) then tách các vị trí từ h2+1 đến n vào thành tố phụ sau (Y6);
if (trong Y6 có từ loại là CD ở vị trí h+1 và vị trí h có từ loại là $ hoặc #)  then  Y6a = từ ở vị trí h+1 ;
else Y6b = Y6 ;
end;
```

Như vậy, với giải thuật 2 chúng ta đã phân tích được thành phần trung tâm (Y5) và thành phần phụ sau (Y6a và Y6b). Các thành phần phụ trước (Y1 Y2 Y3 Y4) sẽ được phân tích như sau.

Giải thuật 3: **Phân tích các thành phần phụ trước (Y1, Y2, Y3, Y4)**

Nhập: Thành phần phụ trước (Y1Y3Y3Y4)
Xuất: Y1, Y2a, Y2b, Y3a, Y3b, Y4

Phương pháp: Gọi A là chuỗi nhập có dạng $A \rightarrow C_1 \ldots C_n$, i = 1…n; hàm *GetPos(A,i)* trả về từ loại của chuỗi A ở vị trí i;

```
i = 1;
repeat
  c = GetPos(A,i);
  case c of
```

```
    {PDT}: Y1 = c;
    {DT}: if (từ là "all" hay "both") then Y1 = c;
          else if (từ là "the" hay "a" hoặc "an")
          then Y2a = c;  else Y2b = c;
    {PRP$}: Y2b = Y2b + c;
    {CD}:    Y3b = Y3b + c;
    c = GetPos(A,i+1); /* tìm liên từ trong Y3 */
  if ((c={CC})
and ((c1 = GetPos(A,i+2)) = {CD})) then
begin
          Y3b = Y3b + c + c1; i = i + 2; end
      {JJ}:  if ((c1 = GetPos(A,i+1)) = {CD}) then
             begin
               Y3a = c; Y3b = Y3b + c1; i = i + 1;
             end
    else Y4 = tất cả các từ còn lại; //thoát vòng lặp
   end {case}
until (hết chuỗi nhập);
```

Trong giải thuật 3, chúng tôi giả định đã có hàm GetPos(A,i). Hàm này nhận vào hai thông số là chuỗi từ loại A, vị trí từ loại cần lấy i và trả về từ loại ở vị trí i trong chuỗi A. Các biến c và c1 được dùng để lưu các từ loại. Bốn biến Y1, Y2, Y3 và Y4 được dùng để lưu các thành phần được phân rã từ phần phụ trước (Y1 Y2 Y3 Y4). Ví dụ về sự xuất hiện của liên từ trong các thành phần của cụm danh từ tiếng Anh.

Ví dụ 1:

(a) [your/PRP$ old/JJ *apples/NNS* and/CC *bananas/NNS*]
[Y2 Y4 Y5[3, 4, 5]]

(b) [my/PRP$ *dog/NN*] and/CC [your/PRP$ *cat/NN*]
[Y2 Y5] CC [Y2 Y5]

(iii) Bài toán chuyển đổi trật tự từ

Sau khi có kết quả phân tích cấu trúc cụm danh từ cơ sở tiếng Anh thành các thành phần từ Y1 đến Y6 bằng các giải thuật 1, 2 và 3. Giải thuật chuyển đổi trật tự từ chỉ đơn giản là sắp đặt lại các thành phần này vào các vị trí của các thành phần tương ứng trong cụm danh từ cơ sở tiếng Việt như mô tả trong Bảng 4. Mô hình chuyển đổi trật tự từ được mô tả ở Hình 4. Chi tiết giải thuật chuyển đổi trật tự từ trong cụm danh từ cơ sở tiếng Anh theo tiếng Viêt được trình bày sau đây.

Giải thuật 4: **Chuyển đổi trật tự từ của cụm danh từ cơ sở tiếng Anh theo cấu trúc cụm danh từ cơ sở tiếng Việt.**

Nhập: Cụm danh từ cơ sở tiếng Anh đã được gán nhãn từ loại
Xuất: Cụm danh từ cơ sở tiếng Anh đã chuyển đổi trật tự từ theo cấu trúc cụm danh từ cơ sở tiếng Việt

Phương pháp:
Bước 1: *Thực hiện các giải thuật 1, 2 và 3 để phân tích chuỗi nhập thành các thành phần chức năng ở các vị trí tương ứng.*
Bước 2: *Sắp xếp lại các thành phần này vào các vị trí tương ứng trong cấu trúc cụm danh từ cơ sở tiếng Việt như sau: cất Y1 vào X1, cất Y2a hoặc Y3b, hoặc Y6a vào vị trí X2, cất Y5 vào vị trí X5, cất Y4, Y6b, Y3a vào vị trí X6 theo hai cấu trúc "Y3a Y4 Y6b" và "Y4 Y3a Y6b", cất Y2b vào vị trí X7; Xếp các vị trí này theo thứ tự tăng dần của X; Các thành tố trong chuỗi mới X1 X2 X5 X6 X7 là cấu trúc cụm danh từ cơ sở tiếng Anh đã chuyển đổi trật tự theo cấu trúc cụm danh từ cơ sở tiếng Việt;*

Ví dụ 2 mô tả quy trình chuyển đổi trật tự từ tiếng Anh theo tiếng Việt.

Ví dụ 2:

a) [her$_1$/PRP$ first$_2$/JJ six$_3$/CD Vietnamese$_4$/JJ dresses$_5$/NNS]
Sáu/CD *cái/CL* áo dài/NN Việt Nam/JJ đầu tiên/JJ *của/PRO* cô/NN ta/NN

Sau khi phân tích ví dụ 2(a) chúng ta được Y2= her/PRP$, Y3a = first/JJ, Y3b = six/CD,

Y4 = Vietnamese/JJ, Y5 = dresses/NNS và sắp xếp lại theo trật tự từ tiếng Việt là "Y3b Y5 Y4 Y3a Y2" chúng ta có trật tự như ví dụ 2(b).

(b) [six$_1$/CD dresses$_2$/NNS Vietnamese$_3$/JJ first$_4$/JJ her$_5$/PRP$]
Sáu/CD (*cái/CL)* áo dài/NN Việt Nam/JJ đầu tiên/JJ (*của/PRO)* cô/NN ta/NN

CL là từ loại của từ "cái" và POS là từ loại của từ "của".

Quan sát bảng cấu trúc cụm danh từ cơ sở tiếng Việt trên hình 4, chúng ta thấy rằng: ngoại trừ các thành phần X2, X3, X4, X6, các thành phần khác đều có sự tương đương chức năng là 1-1 (X1 ↔ Y1, X5 ↔ Y5, X7 ↔ Y2b). Thành phần X3 và X4 là hai thành phần đặc trưng của cụm danh từ cơ sở tiếng Việt. X3, X4 không có thành phần tương đương chức năng bên cụm danh từ cơ sở tiếng Anh (X3 → ϕ, X4 → ϕ). Ở vị trí của thành phần X2 xuất hiện thành phần

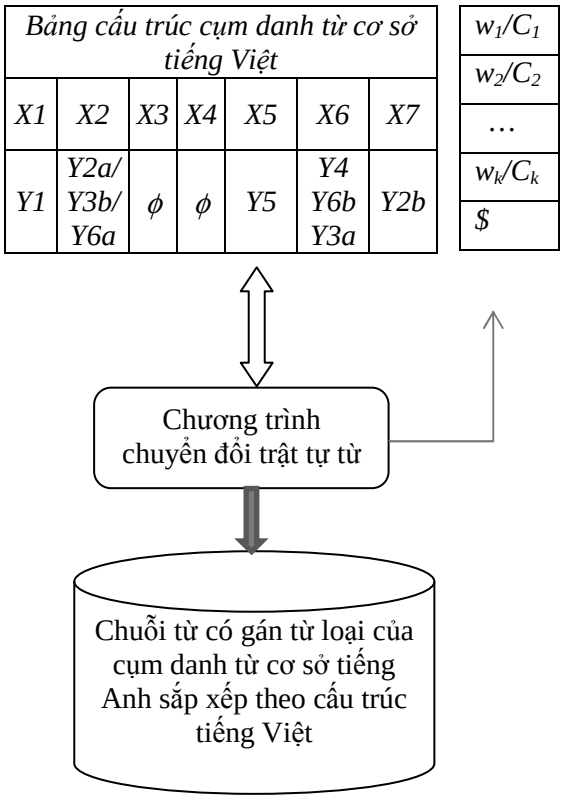
Y2a hoặc thành phần Y3b hoặc thành phần Y6a là những thành phần tương đương chức năng với X2. Ở vị trí của thành phần X6, nếu có nhiều thành phần của cụm danh từ cơ sở tiếng Anh cùng xuất hiện ở vị trí này, thì thứ tự xuất hiện của chúng là X6 = {Y4 Y6b Y3a} như ví dụ 3 (b), hoặc X6 = {Y4 Y3a Y6b} như ví dụ 3 (c), phụ thuộc vào ngữ cảnh của câu tiếng Việt.

Ví dụ 3:

(a) [the/DT first/JJ three/CD young/JJ man/NN
Y2 Y3a Y3b Y4 Y5
absent/JJ] will/MD be/VB punished/VBN
Y6b

(b) ba/CD thanh niên/NN trẻ/JJ vắng mặt/JJ đầu
X2 X5
tiên/JJ sẽ/MD bị phạt/VB.
X6

(c) ba/CD thanh niên/NN trẻ/JJ đầu tiên/JJ vắng
X2 X5
mặt/JJ sẽ/MD bị phạt/VB.
X6



Hình 4. Mô hình chuyển đổi trật tự từ tiếng Anh theo tiếng Việt

Trong đó:

- w_k: Từ thứ *k* trong cụm danh từ tiếng Anh
- C_k: Mã từ loại của từ thứ *k*
- X1 đến X7: là các thành tố trong cụm danh từ tiếng Việt
- Y1 đến Y6: là các thành tố của cụm danh từ tiếng Anh

Phân tích các thành phần của ví dụ 3 bằng các giải thuật 1, 2 và 3. Chúng ta có Y2 = the/DT, Y3a = first/JJ, Y3b = three/CD, Y4 = young/JJ, Y5 = man/NN, Y6 = absent/JJ. Sắp xếp lại chúng theo trật tự của tiếng Việt theo giải thuật 4 cho câu trong ví dụ 3(b) là "Y3b Y5 Y4 Y3a Y6b" và cho câu trong ví dụ 3(c) là "Y3b Y5 Y3a Y4 Y6b".

Như vậy, trật tự của Y3a và Y4 có thể thay đổi cho nhau phụ thuộc vào câu dịch tiếng Việt. Tuy khác nhau về trật tự khi được sắp xếp theo thứ tự của cụm danh từ cơ sở tiếng Việt tương ứng, nhưng cả Y3a và Y4 trong cụm danh từ cơ sở tiếng Anh đều có chức năng tương đương với thành phần X6 trong cụm danh từ tiếng Việt.

IV. KẾT QUẢ THỰC NGHIỆM

✓ **Dữ liệu đánh giá**

Để đảm bảo tính bao phủ của mô hình chuyển đổi trật tự từ trong cụm danh từ cơ sở tiếng Anh theo tiếng Việt, chúng tôi đã chuẩn bị bộ ngữ liệu mẫu. Bộ ngữ liệu này bao gồm 174 mẫu cụm danh từ cơ sở tiếng Anh. Những mẫu này có số từ lớn hơn 2 từ trở lên và có tần suất xuất hiện hơn 10 lần trong ngữ liệu của Penn Treebank [16]. Phần dịch ra tiếng Việt của 174 mẫu cụm danh từ cơ sở tiếng Anh cũng như việc tách các cụm danh từ cơ sở tiếng Anh thành các thành phần từ Y1 đến Y6, các cụm danh từ cơ sở tiếng Việt từ X1 đến X7 do Trung tâm nghiên cứu ngôn ngữ - Viện KHXH tại thành phố Hồ Chí Minh thực hiện.

✓ **Penn Treebank**

Dự án Penn Treebank được thực hiện trong 8 năm (1989-1996), đã gán nhãn bằng tay từ loại cho khoảng 7 triệu lượt từ, 3 triệu từ được phân tính cú pháp và hơn 2 triệu từ trong câu được phân tích cấu trúc với

các chức năng ngữ pháp cho văn bản và 1,6 triệu từ phiên âm cho tiếng nói. Tài liệu dùng để chú giải là các tài liệu về máy tính của hãng IBM, ghi chú của y tá, các bài báo trong Tạp chí Phố Uôn (Wall Street Journal), các cuộc hội thoại qua điện thoại[17].

✓ **Tiêu chuẩn đánh giá**

$$Pre(A,B) = \frac{A}{B} \quad (1)$$

$$Rec(A,C) = \frac{A}{C} \quad (2)$$

$$F_\beta(Pre,Rec) = \frac{(\beta^2+1)*Pre*Rec}{\beta^2*(Pre+Rec)} \quad (3)$$

$$AER(A,B,C) = 1 - \frac{2*A}{B + C} \quad (4)$$

Trong đó:

- A: Số NPbase do con người đánh giá lại từ kết quả đưa ra bởi máy tính
- B: Số NPbase đưa ra bởi máy tính
- C: Số NPbase do con người xác định từ dữ liệu đánh giá
- Pre: Độ chính xác (precision)
- Rec: Độ thu hồi (Recall)
- AER: là tỷ số lỗi
- F_β: là trọng số điều hòa
- β: Thông số cho trước, chúng tôi chọn β = 1

Chúng tôi sử dụng tiêu chuẩn về đánh giá cụm từ (chunker) của Jurafsky và Matin [3] bao gồm các phép đo về độ chính xác Pre, độ thu hồi Rec, trọng số điều hòa F_β theo các công thức (1), (2), (3). Chúng tôi cũng sử dụng công thức đánh giá tỷ suất lỗi đối sánh AER của Och [8] (công thức (4) để đánh giá cho các đối tượng trong chương này.

(i) Phân tích cụm danh từ cơ sở tiếng Anh thành các thành phần từ Y1 đến Y6

Để thực nghiệm cho mục này, chúng tôi sử dụng các giải thuật 1, 2 và 3 để phân tích 174 mẫu đã chọn vào các thành phần từ Y1 đến Y6. Sau đó lấy kết quả thu được từ máy tính để so trùng với các thành phần đã được thực hiện bởi các chuyên gia ngôn ngữ, cụ thể với 174 mẫu thu được như Bảng 5.

Bảng 5. Kết quả phân tích cụm danh từ tiếng Anh

Thành tố	*A*	*B*	*C*	*Pre (A,B)*	*Rec(A,C)*	*AER*	F_β
Y1	13	13	13	100%	100%	0%	100%
Y2a	73	75	75	97,3%	97,3%	2,7%	97,3%
Y2b	16	16	16	100%	100%	0%	100%
Y3a	2	2	2	100%	100%	0%	100%
Y3b	25	26	26	96,2%	96,2%	3,8%	96,2%
Y4	145	147	147	98,6%	98,6%	1,4%	98,6%
Y5	173	174	174	99,4%	99,4%	0,6%	99,4%
Y6a	3	3	3	100%	100%	0%	100%
Y6b	3	3	3	100%	100%	0%	100%

Trong Bảng 5:

- Cột A là thống kê kết quả đánh giá lại của con người về các thành phần từ Y1 đến Y6 (của cụm danh từ cơ sở tiếng Anh) do máy tính đưa ra.
- Cột B là thống kê ghi lại tổng số các thành phần từ Y1 đến Y6 do các giải thuật 1, 2 và 3 mà máy tính xác định được.
- Cột C là thống kê số thành phần được phân tích trong bộ mẫu chuẩn do con người xác định trước.
- Pre là độ chính xác đạt được của các giải thuật phân tích cụm danh từ cơ sở tiếng Anh
- Rec là độ thu hồi đạt được của các giải thuật phân tích cụm danh từ cơ sở tiếng Anh F_β là độ đo trọng số điều hòa
- AER là tỷ suất lỗi

Khi phân tích 174 mẫu thành các thành phần từ Y1 đến Y6, chúng tôi thấy thành phần Y5 có 10 mẫu có liên từ, chiếm tỷ lệ là 10/174, các thành phần khác như Y6 có tỷ lệ là 1/9, Y4 có tỷ lệ là 6/147 và Y3 có tỷ lệ là 1/26.

(ii) Chuyển đổi trật tự từ và xác định điểm neo

Bảng 6 thống kê các điểm neo xác định được nhờ quá trình phân tích cụm danh từ cơ sở tiếng Anh và sắp xếp lại các thành phần này theo tiếng Việt bằng giải thuật 4. Sau đó, chúng tôi xác định các điểm neo trái và neo phải của cụm danh từ cơ sở tiếng Anh đã được sắp xếp để phục vụ cho bài toán chiếu tìm cụm danh từ tiếng Việt tương ứng. Điểm neo được xác định là từ nằm ở tận cùng bên trái làm điểm neo trái, từ nằm ở tận cùng bên phải làm điểm neo phải. Biên trái

và biên phải này là của cụm danh từ cơ sở tiếng Anh đã chuyển đổi trật tự từ theo tiếng Việt (đã loại bỏ stopword trong cụm danh từ cơ sở tiếng Anh).

Bảng 6. Kết quả chuyển đổi trật tự từ và xác định điểm neo

Điểm neo	*D*	*E*	*F*	*Pre(D,F)*	*Rec(D,F)*	*AER*	F_β
L (trái)	173	174	174	99,4%	99,4%	0,6%	99,4%
R (phải)	172	174	174	98,8%	98,8%	1,2%	98,8%
NPed	171	174	174	98,3%	98,3%	1,7%	98,3%

Trong Bảng 6:

- Cột D là số điểm neo hoặc luật được thẩm định lại bởi con người từ kết quả đưa ra bởi máy tính
- Cột E là số điểm neo hoặc luật do máy tính đưa ra.
- Cột F là số điểm neo hoặc luật được định trước bởi con người trong mẫu thẩm định.
- Cột Pre là độ chính xác đạt được (công thức (1))
- Cột Rec là độ thu hồi tính theo công thức (2).
- AER là tỷ suất lỗi tính theo công thức (4).
- Cột Fβ là độ đo trọng số điều hòa tính theo công thức (3) với β=1.
- Hàng L là tổng số điểm neo trái của các mẫu đánh giá.
- Hàng R là tổng điểm neo phải của các mẫu đánh giá.
- Nped là luật sinh cụm danh từ cơ sở tiếng Anh đã chuyển đổi theo trật tự từ của cụm danh từ cơ sở tiếng Việt.

V. KẾT LUẬN

Bài báo đã trình bày mô hình chuyển đổi trật tự từ cụm danh từ cơ sở tiếng Anh theo tiếng Việt, để xác định điểm neo trong cụm danh từ tiếng Anh. Điểm neo trong bài toán này được định nghĩa là từ tận cùng bên trái và từ tận cùng bên phải của cụm danh từ tiếng Anh sau khi đã chuyển đổi trật tự từ theo tiếng Việt.

Kết quả của nghiên cứu này phục vụ cho bài toán nhận biết cụm danh từ cơ sở tiếng Việt thông qua đối sánh điểm neo trong cặp câu song ngữ [5]. Mặc dù kết quả khá khả quan, nhưng đây mới chỉ là kết quả thực nghiệm trên 174 mẫu nghiên cứu do Trung tâm nghiên cứu ngôn ngữ - Viện KHXH tại thành phố Hồ Chí Minh thực hiện biên dịch và đánh giá so sánh. Cho dù số mẫu này thống kê có tần suất xuất hiện thường xuyên trong Penn Treebank và chúng chiếm tỷ lệ đến 92% số luật sinh về cụm danh từ cơ sở trong các chương từ 15 đến 18 của Penn Treebank, chúng tôi vẫn cần khảo sát, nghiên cứu thêm nhiều mẫu cụm danh từ cơ sở khác nữa.

TÀI LIỆU THAM KHẢO

[1] M.COLLINS *(1999), Head driven statistical models for natural language parsing,* PhD dissertation, University of Pennsylvania.

[2] DINH DIEN, THUY NGAN, XUAN QUANG, CHI NAM, *"The Parallel Corpus Approach to Building the Syntactic Tree Transfer Set in the English-to-Vietnamese Machine Translation",* International Conference on Electronics, Informations and Commumications. Hanoi, 16-18/08/2004.

[3] D.JURAFSKY and J.MATIN (2006), *Speech and Language Processing,* http://www.cs.colorado.edu/~martin/slp2.html

[4] NGUYỄN CHÍ HIẾU, *Mô hình khai thác đặc tính ngôn ngữ đích nhằm xác định các cụm danh từ cơ sở tương ứng Anh-Việt,* Luận án Tiến sĩ, Đại học Bách khoa – Thành phố Hồ Chí Minh, 2008.

[5] *NGUYỄN CHÍ HIẾU, "A Combination System for Identifying Base Noun Phrase*", Advanced Methods for Computational Collective Intelligence, SCI 457, pp. 13-23, ©Springer-Verlag Berlin Heidelberg, 2012.

[6] *Y.S.*HWANG, *K.*PAIK, *Y.*SASAKI, *"Bilingual Knowledge Extraction Using Chunk Alignment",* PACLIC 18, December 8th-10th, 2004, Waseda University, Tokyo.

[7] *J.*KUPIEC, *"An Algorithm for finding Noun phrase Correspondences in Bilingual Corpora",* Proceedings of the 31st annual meeting on Association for Computational Linguistics, Columbus, Ohio, USA, Pages: 17 – 22, 1993.

[8] F.J.OCH, H.NEY, *"A Systematic Comparision of Various Statistical Alignment Models",* Association for Computational Linguistics, 2003.

[9] M. POPEL, *"English-Czech Machine Translation Using TectoMT"*, WDS'10 Proceedings of Contributed Papers, Part I, pages: 88–93, 2010.

[10] VŨ NGỌC TÚ, *Nghiên cứu đối chiếu trật tự từ Anh-Việt trên một số cấu trúc cú pháp cơ bản*, Luận án PTS Ngữ văn, ĐHQG Hà nội, 1996.

[11] N.H.TUONG, *The structure of the Vietnamese Noun Phrase*, Ph.D. dissertation, Boston University Graduate School of Arts and Sciences, 2004.

[12] N.P.THAI and A.SHIMAZU, *"Improving Phrase-Based SMT with Morpho-Syntactic Analysis and Transformation"*, Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 138-147, Cambridge, August-2006.

[13] W.WANG and M.ZHOU, *"Structure Alignment Using Bilingual Chunking"*, The 19th International Conference on Computational Linguistics (Coling 2002).

[14] H. WANTANABE, S. KUROHASHI and E. ARAMAKI, Finding Structural Correspondences from Bilingual Parsed Corpus, IBM Research, Tokyo Research Laboratory, 1999.

[15] D.YAROWSKY, G.NGAI and R.WICENTOWSKI, *"Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora"*, Proceedings of NAACL-2001.

[16] http://lcg-www.uia.ac.be/conll2000/chunking

[17] A.TAYLOR, M.MARCUS, B.SANTORINI, *The Penn Treebank: An Overview*, http://www.ldc.upenn.edu

PHỤ LỤC A

BẢNG KÝ HIỆU MÃ TỪ LOẠI

Nhãn	Mô tả
CC	Coordinating conjunction (liên từ)
CD	Cardinal number (số từ)
CD	Determiner (định từ)
EX	Existential "there" ("có")
FW	Foreign word (từ nước ngoài)
IN	Preposition (giới từ)
JJ	Adjective (tính từ)
JJR	Adjective, comparative
JJS	Adjective, superlative
NN	Noun, singular / mass
NNS	Noun, plural (danh từ số nhiều)
NP	Proper noun, singular (danh từ riêng số ít)
NPS	Proper noun, plural (danh từ riêng số nhiều)
PDT	Pre-determiner (tiền chỉ định từ)
POS	Possessive ending
PRO	Personal pronoun (đại từ nhân xưng)
PRP$	Possessive pronoun (đại từ sở hữu)
RB	Adverb (trạng từ)
RB	Adverb, comparative (trạng từ so sánh hơn)
RBS	Adverb, superlative (trạng từ so sánh nhất)
VB	Verb, base form (động từ nguyên thể)
VBD	Verb, past tense (động từ quá khứ)
VBG	Verb, gerund or present participle
VBN	Verb, past participle (động từ quá khứ)
VBP	Verb, non 3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner (định từ bắt đầu bằng Wh)
WP	Wh-pronoun (đại từ bắt đầu bằng Wh)
WP$	Possessive Wh-pronoun
CL	từ phân lớp
CA	tiểu từ "cái"
PL	"những", "các"
BA	"bằng", "từ"
$	"đô la Mỹ"
#	"bảng Anh"

Nhận bài ngày: 14/03/2013

SƠ LƯỢC VỀ CÁC TÁC GIẢ

NGUYỄN CHÍ HIẾU

Sinh ngày 27/12/1959.

Tốt nghiệp đại học ngành Chế tạo máy- Đại học Bách Khoa Hà Nội, năm 1981; đại học ngành CNTT - Đại học Bách Khoa TP. Hồ Chí Minh, năm 1998; Tốt nghiệp Thạc sỹ Khoa học máy tính- Đại học Bách Khoa TP. Hồ Chí Minh, năm 2002; Nhận bằng Tiến sĩ Kỹ thuật ngành Khoa học máy tính - Đại học Bách Khoa TP. Hồ Chí Minh, năm 2008.

Hiện đang công tác tại Đại học Công nghiệp Thành phố Hồ Chí Minh.

Hướng nghiên cứu chính : Xử lý ngôn ngữ tự nhiên

E-mail: nchieu@hui.edu.vn