CHUYỂN NGỮ TRUY VẤN TRONG TRUY XUẤT THÔNG TIN XUYÊN NGÔN NGỮ

Nguyễn Chánh Thành, Nguyễn Văn Hiếu, Phan Thị Tươi[9]

Trường Đại học Công nghệ Tp. Hồ Chí Minh

Trong truy xuất thông tin xuyên ngôn ngữ, việc chuyển ngữ truy vấn từ dạng từ khóa đơn giản đến dạng cụm danh từ phức tạp là bài toán chính yếu. Nhiều giải pháp chuyển ngữ cụm danh từ đã được đề xuất trong các nghiên cứu về xử lý ngôn ngữ tự nhiên đều dựa trên cơ sở chuyển ngữ các từ thành phần thông qua từ điển và tính độ kết hợp. Nhờ sử dụng kết hợp lý thuyết tập hợp với mẫu luật sinh trong văn phạm của ngôn ngữ tự nhiên, bài báo đề xuất một cách tiếp cận khác để chuyển ngữ cụm danh từ từ ngôn ngữ nguồn sang ngôn ngữ đích. Thực nghiệm trong cặp ngôn ngữ Việt-Anh cho kết quả với độ chính xác tương đối khả quan cho thấy tính khả thi của phương pháp đề xuất.

1. Giới thiệu

Bài toán chuyển ngữ cụm danh từ, trong xử lý ngôn ngữ tự nhiên, khi áp dụng vào trường hợp tổng quát cho cặp ngôn ngữ L_1 và L_2 bất kỳ thường gặp khó khăn vì phụ thuộc vào độ phức tạp văn phạm của ngôn ngữ, tính chính xác trong kết quả thu được do đó bị ảnh hưởng nhiều. Phương pháp chung, dựa trên một từ điển, được áp dụng ngày càng nhiều do tính đơn giản và sẵn có của các từ điển song ngữ máy khả dọc L_1 - L_2. Tuy nhiên, hai vấn đề chính cần quan tâm là tính đầy đủ của từ điển và sự nhập nhằng do có nhiều phương án chuyển từ một từ trong ngôn ngữ nguồn sang ngôn ngữ đích.

Trong việc nghiên cứu tìm cách khắc phục những vấn đề này, chúng tôi đã kết hợp lý thuyết tập hợp với tập mẫu luật sinh để tìm giải pháp chuyển ngữ cụm danh từ cho độ chính xác cao hơn, đồng thời tránh được các nhược điểm nêu trên.

2. Các nghiên cứu liên quan

Ngày nay, việc cụm danh từ thường được sử dụng như một khối thống nhất khi chuyển từ ngôn ngữ nguồn L_1 sang ngôn ngữ đích L_2 làm cho tất cả các cụm danh từ sẵn có trong L_2 phải được lưu trữ trong CSDL tương ứng. Với một cụm danh từ trong L_1, cần phải tìm trong CSDL những cụm danh từ dự tuyển tương ứng. Để thực hiện việc này, một số giải pháp như sử dụng *lý thuyết tập hợp* và *luật sinh trong văn phạm của ngôn ngữ tự nhiên đã tạo được sự* quan tâm.

Giải pháp dùng luật sinh trong chuyển ngữ là một hướng nghiên cứu kinh điển được áp dụng rất nhiều trong dịch máy. Một trong số đó đã được nhóm Jianfeng Gao [1] trình bày với những kết quả tương đối khả quan cho cặp ngôn ngữ Hoa-Anh. Nhóm tác giả này đã đưa ra khái niệm cụm danh từ cơ sở và khuôn mẫu của cụm danh từ phức tạp dựa trên cụm cơ sở. Nhờ đó, thay vì thực hiện việc xử lý chuyển ngữ cho cụm danh từ phức tạp họ đã xử lý nhiều cụm từ cơ sở. Tuy nhiên, nghiên cứu này chỉ tập trung xây dựng các khuôn mẫu cho việc chuyển ngữ Anh - Hoa. Một nhóm khác, nhóm Rossitsa Petcova [2], đã nghiên cứu việc chuyển ngữ từ tiếng Bulgary sang tiếng Anh, phân tích chuyên sâu từ loại của hai ngôn ngữ này và các luật sinh tương ứng liên quan đến cụm danh từ, trên cơ sở đó đề xuất bổ sung các tiền tố và hậu tố vào các luật chuyển đổi. Tuy nhiên, điểm hạn chế của hướng tiếp cận này là

[9] Tại Hội nghị Quốc tế Thư viện số châu Á lần thứ X (ICADL), tháng 12 năm 2007, bài viết "A Hybrid Approach of Noun Phrase Translation in Cross-Language Information Retrieval" của các tác giả cũng đã được chọn đăng trong Kỷ yếu "Asian Digital Libraries: Looking Back 10 years and Forging New Frontiers".

khả năng ứng dụng trong các ngôn ngữ có độ phức tạp và tính bất quy tắc cao của cấu trúc cú pháp của văn phạm.

Theo hướng dùng lý thuyết tập hợp, nhóm Fernando López - Ostenero đã đề xuất phương pháp chuyển ngữ cho cặp ngôn ngữ Anh-Tây ban nha [3] mà không đề cập sâu đến việc xử lý nhập nhằng, nhóm Juan M. Cigarr'an [4] đã tiến hành nghiên cứu và vận dụng việc chọn lựa cụm danh từ, nhóm Ryan Richardson [5] đã phát triển phương pháp của Fernando López - Ostenero [3] để xây dựng bản đồ ngữ cảnh có hỗ trợ chuyển ngữ Anh - Tây Ban Nha, với việc mở rộng xử lý cho cụm danh từ.

Trong cả hai hướng trên, khó khăn ảnh hưởng đến tính hiệu quả là không thể xây dựng một CSDL với đầy đủ các cụm danh từ đích, nghĩa là nhiều cụm danh từ sẽ không được chuyển vào $D(L_2)$. Ngoài ra, độ phức tạp trong văn phạm của mỗi ngôn ngữ, số lượng trường hợp bất quy tắc của cấu trúc cú pháp (theo hướng dùng luật sinh) và số lượng tính toán liên quan đến chọn lựa phần tử trong tập hợp khá lớn (theo hướng dùng tập hợp) làm cho kết quả chưa đạt độ chính xác cao. Hiện nay, trong dịch máy và truy xuất thông tin xuyên ngôn ngữ, hướng tiếp cận dùng xác suất thống kê được quan tâm ngày càng nhiều, nhờ dễ áp dụng trong nhiều ngôn ngữ khác nhau, lại không quá bị phụ thuộc vào tập luật sinh của văn phạm. Một số kết quả theo hướng nghiên cứu này đã được công bố bởi nhóm Philipp Koehn [6], Ashish Venugopal [7], Juan Miguel Vilar [8], và Stephan Kanthak [9]. Tuy nhiên, kết quả thu được thường bị hạn chế bởi độ lớn và độ chính xác của tập dữ liệu huấn luyện.

Từ thực tế nói trên, việc xây dựng một phương pháp chuyển ngữ cụm danh từ hàm chứa các ưu điểm của các hướng tiếp cận trên là thật sự có ý nghĩa.

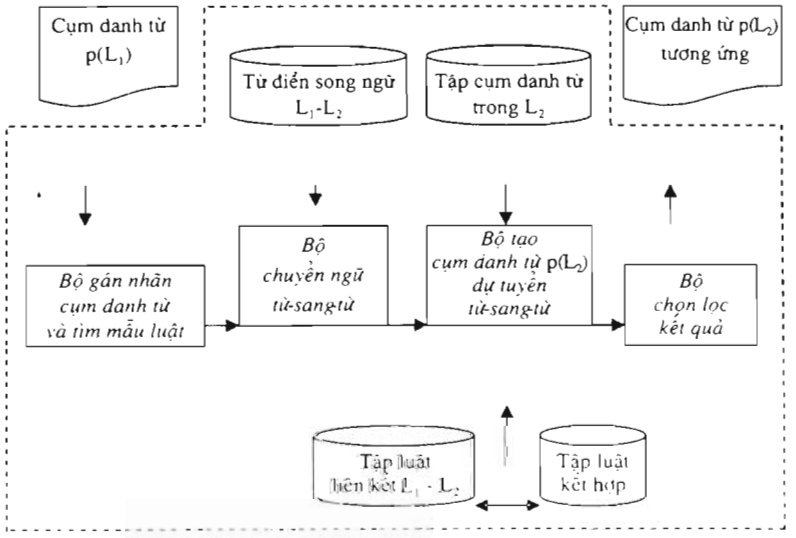
3. Phương pháp tiếp cận chuyển ngữ cụm danh từ

3.1. Mô hình đề xuất

Phương pháp "lai" cho việc xử lý chuyển ngữ cụm danh từ do chúng tôi đề xuất là sự kết hợp giải pháp tập hợp ở trên và luật chuyển đổi để chuyển ngữ những cụm danh từ mà các phương pháp trước không thực hiện được. Đây cũng chính là sự nghiên cứu tìm tòi của riêng chúng tôi.

Để thực hiện phương pháp đề xuất, đã sử dụng một từ điển song ngữ L_1 - L_2, một kho ngữ liệu đơn ngữ chứa các cụm danh từ trong L_1 đã được gán nhãn từ loại, một tập các luật thực hiện việc kết hợp các cụm danh từ cơ sở để tạo thành các cụm danh từ phức tạp hơn, và tập các luật dùng để chuyển đổi thứ tự từ trong các cụm danh từ trong L_1 và L_2.

Việc tổ chức mô hình chuyển ngữ được trình bày trong Hình 1.



Hình 1. *Tổ chức mô hình chuyển ngữ*

Hoạt động của mô hình này được diễn ra tuần tự theo các khối chức năng. Ví dụ: cụm danh từ "đơn vị xử lý" trong L_1 (tiếng Việt), sau khi gán nhãn từ loại có dạng "*đơn vị*/N_1 *xử lý*/N_2", sẽ được chuyển ngữ dạng từ-sang-từ và tạo thành 2 cụm danh từ dự tuyển trong L_2 (tiếng Anh) "*procesing*/N_2 *unit*/N_1" và "*unit*/N_1 *processing*/N_2". Trong bộ chọn lọc, nhờ áp dụng luật , có thể loại đi cụm danh từ "unit processing" vì không có luật phù hợp. Kết quả cuối cùng "*processing unit*" là kết quả đúng.

Cách tiếp cận do chúng tôi đề xuất là nhằm chuyển ngữ các cụm danh từ như một đơn vị thống nhất. Trong đó, ngoài những cụm danh từ $p(L_1)$ đã có cụm tương ứng $p(L_2)$ được lưu ở CSDL trong L_2, các cụm danh từ $p(L_1)$ khác đều sẽ được xử lý bằng các mẫu luật chuyển ngữ và phương pháp thống kê sự đồng xuất hiện. Phương pháp sẽ chọn sự chuyển ngữ tốt nhất dựa trên sự đồng xuất hiện của các từ được chuyển ngữ từ các từ gốc trong cụm từ tiếng Việt.

Cơ sở lý thuyết của cách tiếp cận này là việc kết hợp giữa lý thuyết tập hợp và mô hình Markov ẩn (HMM) trong việc tạo cụm danh từ, đồng thời tính toán các độ đo liên kết tương quan, từ đó, hạn chế sự nhập nhằng và nâng cao độ chính xác của kết quả.

3.2. Một số vấn đề lý thuyết

Độ chính xác của việc chuyển ngữ từ ngôn ngữ L_1 sang L_2 phụ thuộc vào tính đầy đủ của từ điển và việc lựa chọn một từ chuyển ngữ chính xác trong số các khả năng do từ điển đưa ra (điều này không đơn giản).

Về tính đầy đủ của từ điển: nhiều giải pháp nhằm thu thập và tạo ra các tài nguyên lớn, cũng như cập nhật từ điển từ các nguồn tài nguyên này bằng tay hay tự động đã được đề xuất. Nhờ đó tính đầy đủ của từ điển đã tăng lên. Vấn đề chủ yếu, tức là việc chọn lựa sự tương ứng giữa cặp mục từ trong từ điển L_1-L_2, tuy thế vẫn phụ thuộc rất nhiều vào con người. Do vậy, đây không phải là hướng quan tâm của chúng tôi.

Về việc chuyển ngữ một từ chính xác: chúng tôi cố gắng chuyển ngữ cụm danh từ như là một đơn vị thống nhất bằng cách sử dụng một từ điển đơn ngữ chứa cụm danh từ chuẩn trong L_1 hay L_2, được lấy từ các từ điển cũng như rút trích tự động từ các tạp chí uy tín. Tuy nhiên, cần quan tâm đến trường hợp khi có một cụm danh từ trong L_1 nhưng cụm danh từ tương ứng trong L_2 không có sẵn trong từ điển cụm danh từ L_1-L_2. Ví dụ: chuyển cụm danh từ "*bài ca chiến thắng*" trong L_1 (tiếng Việt) sang cụm danh từ ở L_2 (tiếng Anh), được viết là "*song of victory*". Nếu "song of victory" được lưu trong từ điển L_2, thì phép giao (sẽ được trình bày ở phần sau) sẽ đưa ra chính xác cụm từ này. Nếu "song of victory" chưa có trong từ điển, thì bằng các mẫu luật và phương pháp thống kê sự đồng xuất hiện, sẽ đưa ra được sự chuyển ngữ là "victory song" - thông qua việc dùng luật $N_1^{L1} N_2^{L1} ::= N_2^{L2} N_1^{L2}$ - mặc dù "victory song" không phải là sự chuyển ngữ hoàn hảo song nó là kết quả chấp nhận được.

Trong thực tế, chúng ta khó có thể tạo ra được một từ điển chứa tất cả các cụm danh từ đã được dùng trong cuộc sống, bởi các cụm từ mới liên tục được tạo ra trong đời sống xã hội. Chính vì thế, việc bắt gặp những cụm từ chưa được lưu trong từ điển, bất kể lớn đến đâu, là điều tất yếu. Một cơ chế tự động cập nhật dữ liệu cho từ điển vì vậy rất cần thiết. Đây chính là công việc chúng tôi dự định làm trong tương lai.

3.3. Phương pháp tiếp cận

Phương pháp chính mà chúng tôi hướng đến là *thống kê tần suất đồng xuất hiện các từ được chuyển ngữ và xây dựng một tập các luật chuyển ngữ từ cụm danh từ trong L_1 sang L_2, từ đó thực hiện việc chọn lựa kết quả dựa vào các phép toán của lý thuyết tập hợp.*

Định nghĩa cụm danh từ cơ sở của J.Gao [1]:

Cụm danh từ cơ sở là cụm danh từ không chứa những cụm danh từ khác trong nó.

đã được chúng tôi sử dụng trong phương pháp của mình.

Theo đó, khi chuyển ngữ cụm danh từ cơ sở (trong L_1), hai trường hợp có thể xảy ra như sau:

- Nếu cụm danh từ tương ứng, hoặc cụm danh từ tương ứng (trong L_2) đã có trong từ điển, thì danh từ sẽ được xác định bằng phương pháp giao các kết quả chuyển ngữ;

- Nếu cụm danh từ tương ứng (trong L_2) chưa có trong từ điển, thì kết quả chuyển ngữ sẽ được xác định bởi phương pháp dùng tập các luật chuyển đổi và thống kê tần suất đồng xuất hiện của từ được chuyển ngữ.

Việc chuyển ngữ cụm danh từ (phức tạp) trong L_1 sang L_2 sẽ được tiến hành theo hai bước như sau:

- Đầu tiên, các cụm danh từ phức tạp sẽ được phân tích thành các cụm danh từ cơ sở và chuyển ngữ;

- Sau đó, dùng phương pháp kết hợp để tạo ra sự chuyển ngữ cho cụm danh từ phức hợp.

Hai bước xử lý trên giúp chương trình giải quyết các bài toán chuyển ngữ đơn giản cho các cụm danh từ cơ sở thay cho việc xử lý chuyển ngữ một cụm từ phức tạp. Điều này góp phần làm giảm độ phức tạp của việc tính toán cũng như tận dụng được năng lực của từ điển.

Ví dụ về một số mẫu luật cho cặp ngôn ngữ Việt-Anh như sau:

- $NP_1^{L1} <$ của $> NP_2^{L1} ::= NP_1^{L2} <$ of $> NP_2^{L2}$
- $NP_1^{L1} <$ và $> NP_2^{L1} ::= NP_1^{L2} <$ and $> NP_2^{L2}$
- $NP_1^{L1} <$ trong $> NP_2^{L1} ::= NP_1^{L2} <$ in $> NP_2^{L2}$

Theo một số bài báo, trong CSDL của Penn Tree Bank, có khoảng 6.000 mẫu luật tạo NP, trong số đó khoảng 1.100 mẫu luật hay được sử dụng thường xuyên. Hiện tại, chúng tôi đang xây dựng các mẫu luật chuyển đổi tương ứng cho tiếng Việt từ 1.100 mẫu luật này, chẳng hạn $N_1^{L1}\ N_2^{L1} ::= N_2^{L2}\ N_1^{L2}$, $N^{L1}\ ADJ^{L1} ::= ADJ^{L2}\ N^{L2}$,... Công việc tuy không khó về mặt học thuật nhưng đòi hỏi nhiều thời gian và công sức, vì vậy chúng tôi quyết định kế thừa kết quả nghiên cứu trước đó của các thành viên trong nhóm.

Ngoài ra, một cách tổng quát, một mẫu luật trong L_1 có thể có một số mẫu luật tương ứng trong L_2, vì thế các mẫu luật này sẽ mang xác suất cho biết mức độ ưu tiên được dùng khi chuyển đổi.

Dạng tổng quát của mẫu luật như sau:

$PAT=[\ LHS^{L1} ::= RHS^{L2}$, xác_suất$]$

Xác suất này được xác định từ một kho ngữ liệu song ngữ theo công thức:

$$\wp(RHS^{L2} \mid LHS^{L1}) = C(RHS^{L2}, LHS^{L1}) / C(LHS^{L1})$$

Trong đó:

- $C(RSH^{L2}, LHS^{L1})$ là số lần mẫu luật RSH^{L2} tương ứng với mẫu luật LHS^{L1}

- $C(LHS^{L1})$ là số lần mẫu luật xuất hiện trong kho ngữ liệu.

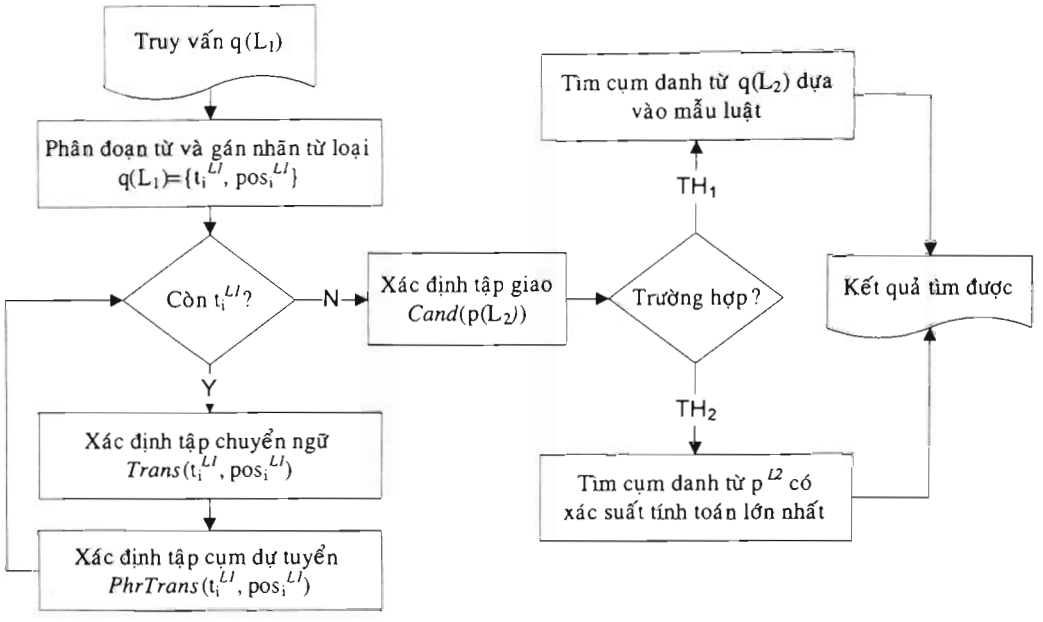
Ví dụ về một số mẫu luật trong cặp ngôn ngữ Việt (L_1) và Anh (L_2) như sau:

- $[N_1^{L1}\ N_2^{L1} ::= N_2^{L2}\ N_1^{L2}, 0.89]$
- $[NP_1^{L1} <$ tại $> NP_2^{L1} ::= NP_1^{L2} <$ at $> NP_2^{L2}, 0.57]$
- $[NP_1^{L1} <$ tại $> NP_2^{L1} ::= NP_1^{L2} <$ in $> NP_2^{L2}, 0.32]$

3.4. Xử lý chuyển ngữ cụm danh từ

Phương pháp xử lý này dựa trên lý thuyết tập hợp để xác định cụm danh từ trong L_2 dựa vào các từ thành phần trong cụm danh từ tương ứng ở L_1, từ đó, tính độ phụ thuộc bigram của mô hình Markov ẩn. Sơ đồ các thuật toán này được thể hiện trong Hình 2.

Từ sơ đồ trong Hình 2, có thể xem xét các thuật toán tương ứng với hai trường hợp sau đây:



Hình 2. *Sơ đồ thuật toán xử lý chuyển ngữ*

3.4.1. Thuật toán 1

Thuận toán này được áp dụng khi cụm danh từ tương ứng trong L_2 đã có trong từ điển $D(L_2)$ trong L_2.

Bước 1. Xét cụm danh từ $p(L_1) = t_1^{L1} t_2^{L1} ... t_n^{L1}$, với t_i^{L1} không phải là các stopword, và pos_i^{L1} là nhãn từ loại tương ứng của t_i^{L1}.

Bước 2. Với mỗi t_i^{L1}, xây dựng tập *Trans* với: $Trans(t_i^{L1}, pos_i^{L1}) = \{k_i^{L2} / k_i^{L2} = m_{L1L2}(t_i^{L1})$ trong từ điển L_1-L_2, và có từ loại là $pos_i^{L1}\}$

Bước 3. Xây dựng tập *PhrTrans* với:

$PhrTrans(t_i^{L1}, pos_i^{L1}) = \{ p(L_2) / p(L_2)$ chứa một từ trong $Trans(t_i^{L1}, pos_i^{L1})$ ở vị trí bất kỳ$\}$

Bước 4. Tập tất cả các cụm danh từ dự tuyển được xác định bởi công thức:

$$Cand(p(L_2)) = \bigcap_{i=1..n} PhrTrans(t_i^{L1}, pos_i^{L1})$$

Trong đó phép giao $A \cap B = \{x / x \in A, x \in B\}$, với x ở đây là cụm danh từ.

Bước 5. Nếu $|Cand(p)| \geq 1$ thì áp dụng các mẫu luật phù hợp để tìm kết quả.

Tuy nhiên, nếu cụm danh từ $p(L_2)$ tương ứng cho cụm danh từ $p(L_1)$ chưa được lưu trong từ điển cụm danh từ, phép giao cho kết quả là một tập rỗng hoặc là tập chứa một số cụm danh từ nhưng nó không thỏa mãn các mẫu luật chuyển đổi, thì cần sử dụng thuật toán dưới đây.

3.4.2. Thuật toán 2

Thuận toán này được áp dụng khi cụm danh từ trong L_2 tương ứng không có trong tập cụm danh từ trong L_2, khi đó mẫu luật và thống kê sự đồng xuất hiện sẽ được dùng để chuyển ngữ cho những cụm danh từ đó.

Bước 1. Cho cụm từ $p(L_1) = t_1^{L1} t_2^{L1} ... t_n^{L1}$, và vế trái LHS^{L1} trong mẫu luật tương ứng của nó.

Bước 2. Với mỗi t_i^{L1}, xây dựng tập *Trans* với: $Trans(t_i^{L1}, pos_i^{L1}) = \{k_i^{L2} / k_i^{L2} = m_{L1L2}(t_i^{L1})$ trong từ điển L_1-L_2, và có từ loại là $pos_i^{L1}\}$

Bước 3. Xây dựng tập *PhrTrans* với:

$PhrTrans(t_i^{L1}, pos_i^{L1}) = \{ p(L_2) / p(L_2)$ chứa một từ trong $Trans(t_i^{L1}, pos_i^{L1})$ ở vị trí bất kỳ$\}$

Bước 4. Tập tất cả các cụm danh từ dự tuyển được xác định bởi công thức:

$Cand(p^{L_2}) = \bigcap_{i=1,n} PhrTrans(\ t_i^{L1}\ ,\ pos_i^{L1})$

Bước 5. Nếu $|Cand(p^{L_2})|\geq 1$ thì cụm $p(L_2)^*$ phù hợp nhất sẽ được xác định bởi:

$p(L2)^* = \underset{Cand(p(L_1))}{\arg\max}(\wp(RHS^{L2} \mid LHS^{L2})\wp(p^{L2}))$

trong đó giả sử $p(L_2) = t'^{L2}_1 t'^{L2}_2 ... t'^{L2}_n$

3.4.3. Trường hợp ngoại lệ

Do đặc thù về yếu tố Hán-Việt của tiếng Việt nên trong một số trường hợp cần phân biệt chính xác giữa từ ghép Hán-Việt với cụm danh từ. Ví dụ: "*ảnh ảo*" là một cụm danh từ hai thành phần, còn "*ảo ảnh*" lại là một từ ghép với chức năng danh từ, có nghĩa là "delusion", "halucination",... Đây là công việc của giai đoạn phân đoạn từ và gán nhãn từ loại, nếu việc làm này không chính xác, sẽ rất dễ gây nhầm lẫn cho giai đoạn chuyển ngữ sau này.

4. Thực nghiệm và đánh giá

4.1. Nguồn dữ liệu

Những khó khăn lớn nhất mà chúng tôi gặp phải là (a) việc xây dựng tập cụm danh từ tiếng Anh và tiếng Việt, (b) năng lực của từ điển song ngữ máy khả đọc và (c) tập mẫu luật chuyển đổi Việt-Anh cũng như tập luật kết hợp tương ứng (Hình 1). Tuy vậy chúng tôi đã đạt được một số kết quả như sau:

Vấn đề (a): một kho ngữ liệu gồm 785 tài liệu tiếng Anh (từ nguồn TREC [11]) với tổng dung lượng 702MB đã được xây dựng , 5.210 tài liệu tiếng Việt (từ nguồn PCWorld Vietnam [15]) với tổng dung lượng 16.4MB và một số tài liệu tiếng Anh khác (nguồn PCWorld [12]) đã được chuyển đổi. Từ kho ngữ liệu này, sau khi tiến hành rút trích cụm danh từ tương ứng, một lượng đáng kể các cụm danh từ cần thiết đã được lưu vào từ điển các cụm danh từ.

Vấn đề (b): nguồn từ điển mà nguồn mở của Hồ Ngọc Đức [14] với khoảng 103.000 mục từ tiếng Anh và 23.000 mục từ tiếng Việt đã được sử dụng để xây dựng một từ điển song ngữ MRD có khả năng mở rộng sau này.

Vấn đề (c): tập luật từ đề tài nghiên cứu của một thành viên trong nhóm với khoảng 1.100 luật sinh tiếng Anh và số lượng tương ứng luật tiếng Việt liên quan đến cụm danh từ đã được sử dụng.

4.2. Một số ví dụ

Ví dụ 1, thuật toán 1:

Xét cụm $p(L_1)$="*thể thao*/N_1 *mùa đông*/N_2".

Giả sử trong từ điển $D(L_2)$ đã có "*winter sports*".

Ta có:

PhrTrans("thể thao", N_1) = {"sports club", "water sports", "winter sports", "outdoor sports" }

với *PhrTrans*("thể thao", N_1) là một tập hợp mà mỗi phần tử là một cụm danh từ, và "winter sports" là một thành phần trong tập này vì nó chứa từ "sports".

Tương tự:

PhrTrans("mùa đông", N_2) = {"winter clothes", "winter day", "winter sports", ... }

Phép giao hai tập hợp *PhrTrans*("thể thao", N_1) và *PhrTrans*("mùa đông", N_2) cho ta kết quả:

Cand($p(L_2)$) = *PhrTrans*("thể thao", N_1)$\cap$ *PhrTrans*("mùa đông", N_2)

= {"*winter sports*"}

vì phần tử này thuộc hai tập hợp, do đó $p(L_2)$ = "*winter sports*".

Ví dụ 2, thuật toán 1:

Xét trường hợp phức tạp hơn với cụm $p(L_1)$="*đơn vị*/N_1 *xử lý*/N_2". Giả sử trong từ điển $D(L_2)$ ngoài cụm từ "*processing unit*" còn có "*unit processing*", "*central processing unit*", ...

Ta có:

PhrTrans("đơn vị", N_1) = {"central processing unit", "unit procesing", "processing unit",

"unit of length", "sample unit", ... }

PhrTrans("xử lý", N_2) = {"central processing unit", "unit processing",

"processing unit", "image processing", ...}

Thực hiện phép giao hai tập hợp này, kết quả thu được là:

Cand($p(L_2)$)= {"central processing unit", "unit procesing", "processing unit"}

Khi áp dụng mẫu luật $N_1^{L1} N_2^{L1} ::= N_2^{L2} N_1^{L2}$, kết quả thu được chính là "processing unit".

Ví dụ 3, thuật toán 2:

Xét mẫu luật $N_1^{L1} ADJ_2^{L1} ::= ADJ_2^{L2} N_1^{L2}$ có xác suất $\wp(RHS^{L2} | LHS^{L2}) = 1.0$,

và $\wp(p(L_2)) = \wp(t_1^{L2} t_2^{L2} ... t_n^{L2}) = \prod_{i=2..n} \wp(t_i^{L2} | t_{i-1}^{L2})$

Điểm cần chú ý ở đây là thứ tự các t_i^{L2} được sắp theo thứ tự của mẫu luật RHS^{L2} đã áp dụng, và với t_1^{L2} thì có thể mặc định là $\wp(t_1^{L2} | the) = 1$

Từ đó, để chuyển ngữ "*ảnh*/N *ảo*/ADJ", áp dụng luật $N_1^{L1} ADJ_2^{L1} ::= ADJ_2^{L2} N_1^{L2}$ với = 1.0, sẽ có kết quả như sau:

"ảo"/ADJ_1	"ảnh"/N_1
Virtual	Picture
Imaginary	Image
	Portrait
	Photo

Trong đó:

$\wp$ ("virtual picture") = $\wp$ ("virtual" | "the")* $\wp$ ("picture" | "virtual")

$\wp$ ("virtual image") = $\wp$ ("virtual" | "the")* $\wp$ ("image" | "virtual")

.....

$\wp$ ("imaginary photo") = $\wp$ ("imaginary" | "the")* $\wp$ ("photo" "imaginary")

Trong các giá trị này thì $\wp$ ("virtual image") = $\wp$ ("virtual" | "the")* $\wp$ ("image" | "virtual") là lớn nhất, vậy "*virtual image*" là kết quả chuyển ngữ cho "*ảnh ảo*".

4.3. Thực nghiệm

Chúng tôi chuẩn bị kho tài liệu tiếng Anh gồm 51.802 câu (mỗi câu được lưu trên một dòng), được lấy từ các văn bản tiếng Anh thuộc nhiều lĩnh vực khác nhau (kinh tế, lịch sử, thể thao, kiến trúc) với số lượng lượt từ phân biệt là 30.660 từ (chọn lựa từ là 930.874 có sự trùng lặp), và chiều dài trung bình của một câu là 17,96 từ/câu; một số cặp cụm từ Anh - Việt từ văn bản song ngữ "Cam kết gia nhập WTO" do ban công tác của Chính phủ Việt Nam biên soạn [10].

Ngoài ra, chúng tôi còn xây dựng: từ điển cụm danh từ tiếng Anh gồm các cụm danh từ được rút trích từ các câu tiếng Anh trong từ điển LONGMAN; một file gồm 15.000 cặp cụm danh từ Việt - Anh là bản dịch của nhau được lấy từ các cặp câu mẫu trong từ điển Lạc Việt, được dùng để thống kê và rút ra các quy luật chuyển đổi giữa các cụm từ Anh - Việt; tập các mẫu luật nhận dạng và rút trích cụm danh từ được thực hiện bởi công cụ GATE và đặc biệt là phần tiện ích JAPE của công cụ này [13].

Dù đây chưa phải là những kho ngữ liệu thực sự lớn, kích thước dữ liệu tuy vậy tương đối đủ, số từ phân biệt tương đối đa dạng để có thể tin cậy vào kết quả thực nghiệm.

Độ chính xác khi đánh giá được xác định bằng công thức sau:

Độ chính xác = số cụm từ chuyển ngữ đúng / số cụm từ được chuyển ngữ

Dưới đây là một số kết quả thực nghiệm:

(a) Kết quả thực nghiệm bị tác động bởi việc so sánh trên chiều dài các cụm danh từ của L_2:

Cụm từ tiếng Việt	Kết quả chuyển ngữ	Xử lý bằng cách so sánh chiều dài
Truyền thống anh hùng	Heroic tradition	Heroic tradition
	Revolutionary Heroic tradition	
Cuộc chiến tranh xâm lược	A dirty war of aggression	A war of aggression
	A war of aggression	
Hội đồng bầu cử	Central electoral council	Electoral council
	Electoral council	
Bệnh viện trung ương	Central hospital	Central hospital
	Hue central hospital	
Quy luật kinh tế	The economic rule	The economic rule
	The economic rule of socialism	

Đối với cụm danh từ tiếng Việt, những danh từ chỉ loại (sự, việc, cuộc, niềm,...), những hư từ và liên từ (và, cho, của, các, những,...) và các giới từ (về, tại,...) sẽ không được tính vào chiều dài cụm danh từ, mà được lưu trong một danh sách ***vn_stoplist***.

Tương tự như vậy đối với cụm danh từ tiếng Anh, những từ như a, an, the, of, and, about, for, ... sẽ được lưu trong danh sách ***eng_stoplist***, và không được tính vào chiều dài cụm danh từ.

(b) Kết quả thực nghiệm bị tác động bởi việc kiểm tra độ dài cụm danh từ của L_2:

Độ dài	Cụm từ tiếng Việt	Kết quả chuyển ngữ
2	bài ca chiến thắng	Song ***of*** victory
2	giá trị đồng tiền	***the*** value ***of the*** currency
2	yêu cầu ***về*** tài chính	***the*** financing requirement
3	Luật Đầu tư nước ngoài	***the*** law ***on*** foreign investment
3	môi trường thuận lợi ***cho*** đầu tư	**an** enabling environment **for** investment
3	mục tiêu ***của*** chính sách tiền tệ	**the** monetary policy objective

Ngoài ra, với việc sắp xếp kết quả giảm dần theo xác suất của cụm danh từ, nếu phép giao cho nhiều hơn một cụm danh từ, thì tất cả các cụm danh từ sẽ được hiển thị ra màn hình với thứ tự giảm dần của xác suất. Dưới đây là kết quả minh hoạ khi thực hiện chương trình.

(c) Kết quả thực nghiệm bị tác động bởi việc kiểm tra xác suất chuyển ngữ:

Cụm từ tiếng Việt	Kết quả chuyển ngữ	Xác suất chuyển ngữ
Con ngựa bất kham	an unruly horse	1E-08
	a restive horse	1E-08
Lợi nhuận béo bở	fat profits	0.333333333333333
	big profits	0.0067567567567567

Cụm từ tiếng Việt	Kết quả chuyển ngữ	Xác suất chuyển ngữ
Hiệp ước bất tương xâm	non-aggression pact	0.0001
	non-aggression treaty	0.0001
Cuộc chiến đấu bền bỉ	an enduring struggle	5.30503978779841E-08
	a presevering struggle	1E-08
Bàn bida	billiard table	0.0001
	pool table	0.0001

Ngoài ra, vấn đề nhập nhằng cũng được xử lý khi chuyển ngữ bằng các luật. Sau đây là một số minh hoạ kết quả xử lý:

Cụm từ tiếng Việt	Cụm từ tiếng Anh	Kết quả dịch	Kết quả chọn
Động lực chủ yếu	Prime movers	Main movers	Prime movers
		Main engines	
		Prime movers	
		Prime engines	
		Chief movers	
		Chief engines	
Chính sách kinh tế	Economic policies	*Economic policies*	*Economic policies*
Tuyên bố mở đầu	Introductory statements	Introductory statements Introductory declaration Introductory proclamation	Do dữ liệu thừa, cả ba cụm từ có cùng xác suất E–12
Quyền kinh doanh	Trading right	Business right Business authority Business power Trading right Trading authority Trading power	Do dữ liệu thừa, cả ba cụm từ có cùng xác suất E–12
Hạn chế xuất khẩu	Export restriction	Export limitation *Export restriction*	Export restriction
Chỉ dẫn địa lý	Geographical indications	*Geographical indications* Geographical directions Geographical instruction Geographic indications Geographic directions Geographic instruction	*Geographical indications*
Kiểu dáng công nghiệp	Industrial designs	*Industrial designs*	*Industrial designs*

Cụm từ tiếng Việt	Cụm từ tiếng Anh	Kết quả dịch	Kết quả chọn
Bí mật thương mại	Trade secrets	*Trade secrets* Trade secrecies Commerce secrets Commerce secrecies	Trade secrets
Kinh tế quốc gia	National economy	*National economy*	*National economy*
Đầu tư trực tiếp nước ngoài	Foreign direct investment	Foreign live investment Foreign direct investment Overseas direct investment Overseas live investment	Đúng (phần in nghiêng)
Quỹ tiền tệ quốc tế	International monetary fund	*International monetary fund* International monetary budget	*International monetary fund*

350 cụm danh từ tiếng Việt, gồm 171 cụm có độ dài 2 từ, 126 cụm - 3 từ, 53 cụm - 4 từ đã được chuyển ngữ, kết quả thực hiện chương trình như sau:

Chiều dài cụm danh từ	Số cụm danh từ chuyển ngữ	Số cụm danh từ chuyển ngữ đúng	Độ chính xác
2	171	147	86%
3	126	81	64%
4	53	23	43%
Tổng cộng	**350**	**251**	**72%**

5. Kết luận

Trong phạm vi bài báo này, chúng tôi đã giới thiệu cách tiếp cận chuyển ngữ truy vấn dạng cụm danh từ trên cơ sở xác định tập các chuyển ngữ riêng cho từng từ thuộc cụm danh từ, trước khi tìm tập hợp giao của các phần tử chung, có tham khảo xác suất sử dụng. Phương pháp này cho phép tận dụng năng lực của kho ngữ liệu trong tìm kiếm tập các cụm từ dự tuyển, đồng thời hạn chế sự phụ thuộc của giải thuật vào từ điển song ngữ. Tuy nhiên việc tính toán độ liên kết có bị giảm thiểu đôi phần so với các phương pháp đã có, bởi việc xử lý phép giao của các tập *PhrTrans* để tìm phần tử chung ít nhiều cũng ảnh hưởng đến chi phí bộ nhớ và tài nguyên của hệ thống trong thời gian thực thi. Đây chính là nội dung chúng tôi sẽ quan tâm cải tiến trong quá trình nghiên cứu tiếp theo.

Tài liệu tham khảo

1. Jianleng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, Changning Huang, Improving Query Translation for Cross - Language Information Retrieval using Statistical Model. ACM 1-58113-331-6/01/0009. 2001.

2. Rossitsa Petcova, Katya Alahverdzhieva. Transfer rules in Bulgarian-English Machine Translation - The POS (Noun) and the Noun Phrase. RoCoLi 2005 Summer School (Romania Computational Linguistics Summer School).

3. Fernando López - Ostenero, Julio Gonzalo, Felisa Verdejo. Noun phrases as building blocks for cross - language Search Assistance, Information Processing and Management: an International Journal archive, Volume 41, Issue 3, pp549-568, ISSN:0306-4573. May 2005.

4. Juan M. Cigarr˜an, Anselmo Pe˜nas, Julio Gonzalo, and Felisa Verdejo. Automatic Selection of Noun Phrases as Document Descriptors in an FCA-Based Information Retrieval System. Spnnger LNCS 3403; ICFCA 2005.

5. Ryan Richardson, Ben Goertzel, Edward A. Fox, Hugo Pinto, Automatic creation and translation of concept maps for Computer science-related theses and dissertations. Second International Conference on Concept Mapping/Segundo Congreso Internacional Sobre Mapas Conceptuales San José, Costa Rica Sept. 5-8, 2006.

6 Philipp Koehn, Christof Monz, Shared Task: Statistical Machine Translation between European Languages. Building and Using Parallel Texts: Data-Driven. Machine Translation and Beyond, Proceedings of the Workshop, ACL-05, 29-30 June 2005, University of Michigan - Ann Arbor, Michigan, USA.

7. Ashish Venugopal, Andreas Zollmann, Alex Waibel, Training and Evaluating Error Minimization Rules for Statistical Machine Translation. Building and Using Parallel Texts: Data-Driven, Machine Translation and Beyond. Proceedings of the Workshop, ACL-05, 29-30 June 2005. University of Michigan - Ann Arbor, Michigan, USA.

8. Juan Miguel Vilar, Enrique Vidal, A Recursive Statistical Translation Model. Building and Using Parallel Texts: Data-Driven, Machine Translation and Beyond, Proceedings of the Workshop, ACL-05, 29-30 June 2005, University of Michigan - Ann Arbor, Michigan, USA.

9. Slephan Kanthak, David Vilar, Evgeny Matusov, Richard Zens, and Hermann Ney. Novel Reordering Approaches in Phrase-Based Statistical Machine Translation, Building and Using Parallel Texts: Dala-Driven, Machine Translation and Beyond, Proceedings of the Workshop, ACL-05, 29-30 June 2005, University of Michigan - Ann Arbor, Michigan, USA.

10. Vietnamese Ministry of Trade, http://www.mot.gov.vn 11. Text REtrieval Conference (TREC), http://trec.nist.gov 12. PC World, http://www.pcworld.com 13. GATE, A General Architecture for Text Engineering, http://gate.ac.uk 14. Ho Ngoc Duc, Open-source dictionary, http://www.informalik.uni-leipzig.de/~duc/Dict 15. PC World Vietnam, http://www.pcworld.com.vn