

Ứng dụng văn phạm liên kết trong dịch máy Việt – Anh

Application of Link Grammar Formalism in Vietnamese - English Translation

Nguyễn Thị Thu Hương và Lê Ngọc Minh

Abstract: Link grammar is a new model built for Vietnamese. This model is able to specify most of grammatical phenomena and lexical features of Vietnamese in an intuitive mode. In this paper, we discuss the ability to apply the link grammar model to automatic translation from Vietnamese into English. We also present our translation system using Annotated Disjuncts (ADJ).

I. ĐẶT VẤN ĐỀ

Dịch máy là lĩnh vực đang rất được quan tâm do nhu cầu chuyển ngữ một số lượng lớn văn bản trong thời gian nhanh nhất. Lịch sử phát triển dịch máy đã nhắc đến nhiều phương pháp khác nhau: dịch trực tiếp từng từ, dịch dựa trên luật, dịch máy đa ngôn ngữ, dịch máy thông kê, dịch máy dựa trên ví dụ, dịch máy theo ngữ nghĩa đệ quy tối thiểu, dịch dựa trên ràng buộc...

Hiện nay ở Việt Nam đã tồn tại một số hệ thống dịch máy như: EVtran - VEtran của Nacentech [9], Vietgle của Lạc Việt, EVTS của Đại học Công nghệ - ĐHQG Hà Nội [13], Google Translation của Google, hệ thống dịch của Đại học Bách khoa - ĐHQG TP Hồ Chí Minh [21], hệ thống dịch dựa trên việc học luật chuyển đổi của Đại học Khoa học Tự nhiên - ĐHQG TP Hồ Chí Minh [8]. Các hệ thống kể trên chủ yếu là hệ thống dịch Anh - Việt để tận dụng nguồn tài nguyên ngôn ngữ phong phú của Tiếng Anh. Số lượng hệ thống dịch Việt - Anh ít hơn nhiều: chỉ có hai hệ thống nổi bật là VEtran dựa trên luật và Google Translation theo cách tiếp cận thống kê.

Do tài nguyên phục vụ cho dịch máy liên quan đến Tiếng Việt không nhiều, cách tiếp cận dựa trên luật vẫn được phổ biến nhất ở Việt Nam, đặc biệt là trong dịch máy Việt - Anh.

Mô hình văn phạm liên kết do D. Sleator và D. Temperley đưa ra [19] là mô hình hình thức theo kiểu phụ thuộc. Văn phạm không chứa tập ký hiệu không kết thúc nên kết quả phân tích cú pháp gọn nhẹ hơn cây ngữ cấu. Do đặc điểm hoàn toàn từ vựng hóa, văn phạm liên kết không chỉ mô tả cú pháp mà còn mô tả từ pháp của ngôn ngữ. Do vậy mô hình văn phạm liên kết cho phép biểu diễn nhiều hiện tượng đặc biệt chỉ có trong Tiếng Việt. Ví dụ: thì, thể của động từ, quan hệ sở hữu ẩn, quan hệ chỉ chất liệu của hai danh từ đi cạnh nhau, những cụm từ được dùng như đại từ xưng hô trong Tiếng Việt ("cô ấy", "các bạn")... Những thông tin này rất hữu ích nếu được sử dụng trong các hệ thống dịch tự động sang ngôn ngữ khác.

Mô hình văn phạm liên kết đã được ứng dụng để xây dựng các hệ thống dịch từ Tiếng Anh sang các ngôn ngữ châu Âu như Tiếng Đức [24], Nga [23], Thổ Nhĩ Kỳ [22] hay Sanskrit - một ngôn ngữ của Ấn độ [12] dựa trên việc chuyển đổi tương ứng các liên kết giữa hai ngôn ngữ (riêng hệ thống dịch Anh-Nga có kết hợp thống kê). Các hệ thống đó đã đáp ứng khá tốt những biến đổi về hình thái như thì, thể, giống, số, cách. Tuy nhiên, khi ngôn ngữ nguồn và ngôn ngữ đích có sự khác biệt lớn về cú pháp và từ pháp, khó có thể đưa ra những bản dịch có chất lượng. Đó cũng là lý do các hệ thống nói trên chỉ dừng ở mức thử nghiệm cho một tập câu nhỏ. Một hệ thống dịch khác dựa trên văn phạm liên kết là hệ thống dịch của đại

học Petronas, Malaysia [1], [2], [3], dịch từ Tiếng Anh sang Tiếng Indonesia. Hệ thống cho phép tra nghĩa từ, biến đổi cú pháp và từ pháp thông qua dạng tuyển có chú giải (Annotated Disjunct - ADJ). Hệ thống đã cho kết quả dịch khá tốt sang Tiếng Indonesia - một ngôn ngữ Đông Nam Á cũng chưa có được nguồn tài nguyên phong phú để xử lý tự động như Tiếng Việt.

Dựa trên khả năng biểu diễn nhiều hiện tượng cú pháp và từ pháp đặc trưng của Tiếng Việt, chúng tôi đã xây dựng và thử nghiệm hệ thống dịch máy Việt - Anh dựa trên dạng tuyển có chú giải. Đây là một hệ thống dịch máy dựa trên luật với mục đích minh họa các chức năng biểu diễn Tiếng Việt của mô hình văn phạm liên kết.

Hệ thống cho phép dịch các cụm từ Tiếng Việt một cách mềm dẻo hơn các hệ thống dịch khác. Với tập mẫu thử nghiệm, hiệu quả đạt được của hệ thống không thua kém các hệ thống dịch máy khác, trong khi tập luật chuyển đổi của hệ thống lại khá nhỏ gọn.

Trong các mục sau đây, chúng tôi trình bày một số khái niệm về mô hình văn phạm liên kết, dạng tuyển có chú giải và giới thiệu hệ thống dịch do chúng tôi xây dựng.

II. VĂN PHẠM LIÊN KẾT

Văn phạm liên kết bao gồm một tập các từ, mỗi từ có một yêu cầu liên kết. Một câu được định nghĩa bởi văn phạm nếu tồn tại một cách để vẽ các cung (liên kết) phía trên các từ thỏa mãn 5 điều kiện về *tính phẳng, tính liên thông, tính thoả mãn, tính thứ tự, tính loại trừ* [19].

Những yêu cầu liên kết được lưu trữ trong một từ điển. Từ điển chứa các công thức liên kết như trong bảng 1 dưới đây:

Bảng 1. Công thức liên kết của các từ

Từ	Công thức
tôi	SV+
yêu	SV- & {O+}
cô, bà, anh	SV+ or O- & {NtPd+}
ấy, đó, ta	{NtPd-}

Với việc tra cứu trên từ điển và khử nhập nhằng (nếu cần), bộ phân tích cú pháp cho kết quả phân tích câu "tôi yêu cô ấy" như trong hình 1.

Định nghĩa hình thức của văn phạm liên kết được nêu trong [6].



Hình 1. Phân tích câu "tôi yêu cô ấy"

Văn phạm liên kết được xếp vào dòng phụ thuộc vì cũng biểu diễn mối liên hệ trực tiếp giữa các từ trong câu. Tuy nhiên văn phạm liên kết có một số điểm khác biệt:

- *Liên kết không định hướng*, không có khái niệm cai trị - phụ thuộc.
- *Liên kết có nhãn*. Điều này không bắt buộc trong văn phạm phụ thuộc.
- Văn phạm liên kết *cho phép chu trình* nên có thể biểu diễn đồng thời liên kết cú pháp và ngữ nghĩa giữa các từ.
- *Tính phẳng* của liên kết, thực chất là *tính xạ ảnh* trong văn phạm phụ thuộc, đòi hỏi các cung biểu diễn liên kết trong một câu không giao nhau khi vẽ trên các từ. Với yêu cầu tính phẳng, một số quan hệ từ không lân cận có thể không biểu diễn được trong mô hình văn phạm liên kết. Ví dụ, trong câu "Cái áo này, dù rất đắt, tôi vẫn mua", sau khi vẽ liên kết giữa danh từ "áo" với tính từ "đắt" và động từ "mua", sẽ không thể vẽ liên kết giữa từ "dù" với dấu phẩy mà vẫn đảm bảo tính phẳng. Tồn tại những văn phạm phụ thuộc không có tính xạ ảnh, nhưng tính phẳng là bắt buộc trong văn phạm liên kết. Rất may, những câu như trong ví dụ trên không thường gặp trong thực tế.

Dạng tuyển

Để dễ dàng xử lý tự động, công thức trong từ điển được chuyển thành các *dạng tuyển* (disjunct). Một dạng tuyển bao gồm hai danh sách có thứ tự của các tên liên kết: danh sách liên kết bên trái và danh sách

liên kết bên phải, được ký hiệu là $((L_1, L_2, \dots, L_m) (R_n, R_{n-1}, \dots, R_1))$, trong đó $L_1, L_2, \dots, L_m$ là các kết nối về phía trái và $R_n, R_{n-1}, \dots, R_1$ là các kết nối về phía phải.

Ví dụ, $((SV) ())$ và $((O) (NPd))$ là các dạng tuyển của từ "cô" trong từ điển được nêu ở trên.

Khi phân tích cú pháp, thay vì phân tích công thức liên kết, bộ phân tích cú pháp liên kết tìm ra cách tổ hợp các dạng tuyển của mỗi từ trong câu thỏa mãn các yêu cầu liên kết. Kết quả phân tích cú pháp của một câu sẽ chứa những dạng tuyển thỏa mãn các yêu cầu liên kết của câu đó, chẳng hạn, kết quả phân tích câu "tôi yêu cô ấy" sẽ như sau:

(tôi, ((SV)))
(yêu, ((SV)(O)))
(cô, ((O)(NtPd)))
(ấy, ((NtPd)()))

III. DẠNG TUYỂN CÓ CHÚ GIẢI

Dạng tuyển có chú giải (Annotated Disjunct, viết tắt là ADJ) lưu trữ nghĩa của một từ khi đi với dạng tuyển nào đó. Ví dụ, từ "cô" với dạng tuyển $((O) (NtPd))$ sẽ có nghĩa là "her", trong khi đi với dạng tuyển $((O)(NtPd,SV))$ có nghĩa là "she", còn đi với $((O)(SHA))$ từ đó lại có nghĩa là "aunt"

Dạng tuyển có chú giải là một tổ hợp bao gồm ($\langle$ từ nguồn $\rangle$, $\langle$ từ đích $\rangle$, $\langle$ dạng tuyển $\rangle$) trong đó dạng tuyển thuộc về từ nguồn. Từ đích là nghĩa của từ nguồn trong ngôn ngữ đích khi đi với dạng tuyển tương ứng. Trong hệ thống dịch Việt Anh, các dạng tuyển có chú giải của câu "tôi yêu cô ấy" sẽ là

(tôi, I, ((SV)))
(yêu, love, ((SV)(O)))
(cô, her, ((O)(NtPd)))
(ấy, I, ((NtPd)()))

(Dấu ! đại diện cho xâu rỗng, chỉ ra từ sẽ bị xóa khi dịch).

Từ điển ADJ sẽ tập hợp tất cả các ADJ của ngôn ngữ. Về nguyên tắc, từ điển ADJ phải bao gồm các bộ ba: từ, dạng tuyển và nghĩa của từ trong Tiếng Anh khi sử dụng với dạng tuyển tương ứng. Tuy nhiên, số dạng tuyển của mỗi từ là rất lớn. Khi từ điển song ngữ đã gồm gần 100.000 mục từ, mỗi mục từ đó lại được

gắn với mọi dạng tuyển của từ nguồn, kích cỡ của từ điển ADJ sẽ vô cùng lớn.

Để giảm kích cỡ, từ điển ADJ được xây dựng từ từ điển văn phạm liên kết, chú giải thêm nghĩa Tiếng Anh của từ bên cạnh công thức. Để làm được như vậy, từ điển văn phạm liên kết được xây dựng và chỉnh sửa theo hướng mỗi công thức chỉ thể hiện một cách sử dụng của từ. Từ điển ADJ cũng được dùng thay cho từ điển văn phạm liên kết trong giai đoạn phân tích cú pháp. Dưới đây là ví dụ một số mục từ trong từ điển ADJ:

bởi_vì, vì because: (GT_DT+ or CL+) & {PH+} & (EV- or (CO+ or QHT+))
sờ_đi !
là_do, là_vì because
rằng_thất: R- & CL+
/verb.transitive.trans: ((({TĐT1-} & {TĐT2_1-} & {RpVt- or RtVt- or RfVt- or RhVt-} & {TĐT4-}) or TT_ĐT- or ({TT_ĐT-} & TĐT5-)) & {DT_ĐT- or ĐT_ĐT- or THI_ĐT- or LT_ĐT- or BI-} & {ĐT_XONG+} & {ĐT_DT+} & {ĐT_TT+} & {ĐT_GT+} & {ĐT_LT+} & {THT- or THS+} & ({EV+} & {SDT5- or CL-} & {CO-})) or ({DT_ĐT-} & BI-)

Chú ý rằng trong từ điển trên, mục /verb.transitive.trans là mục chứa công thức liên kết của các động từ ngoại động (trừ một số ngoại lệ có công thức liên kết riêng) nên được liên kết với một từ điển song ngữ con chứa nghĩa của từng động từ với công thức liên kết đã nêu. Ví dụ:

a_tòng act_as_an_accomplice_to
a_tòng imitate
am_hiếu know_well
am_hiếu realize
an_bài predestine
an_bài preordain...

IV. SO SÁNH HAI NGÔN NGỮ VIỆT - ANH

Sự khác biệt giữa Tiếng Việt và Tiếng Anh là không nhỏ. Đó là sự khác biệt giữa một ngôn ngữ phương Đông và một ngôn ngữ phương Tây. Theo Đinh Điền [7], do bị ảnh hưởng của văn hoá phương Đông – nền văn hoá thiên về âm tính, nên trong ngôn ngữ, ngữ pháp Tiếng Việt có tính linh động cao, chứ không chặt chẽ (chia thì, thể, sự phù hợp về giống, số, cách ...)

như ngữ pháp phương Tây. Tiếng Anh không quá chặt chẽ về ngữ pháp như Tiếng Nga, nhưng cũng yêu cầu biến đổi nhiều về hình thái.

Theo Nguyễn Phương Thái và Shimazu [20], hai khác biệt quan trọng nhất giữa Tiếng Việt và Tiếng Anh mà hệ thống dịch máy cần giải quyết là: khác biệt về hình thái và khác biệt về trật tự từ.

Khác biệt hình thái

Theo phân loại về ngôn ngữ của Stankevich được nhắc đến trong [18], Tiếng Việt được xếp vào loại ngôn ngữ phân tích (đơn) hay còn gọi là loại hình phi hình thái với đặc điểm: Trong hoạt động ngôn ngữ, từ không biến đổi hình thái, ý nghĩa ngữ pháp nằm ở ngoài từ [7]. Tiếng Anh là ngôn ngữ thuộc loại hình phân tích ở mức độ vừa phải, thiên về loại hình tổng hợp nhiều hơn, biến đổi hình thái theo thì thể, giống, số...

Những khác biệt về hình thái chủ yếu giữa Tiếng Việt và Tiếng Anh thể hiện ở: số của danh từ, chia động từ, quan hệ sở hữu, tính từ so sánh, tính từ làm vị ngữ, danh từ chỉ loại... Ngoài ra, còn một số vấn đề cần xử lý đặc biệt, ví dụ:

Đại từ xưng hô: Có thể nói việc xưng hô trong Tiếng Việt thuộc loại phức tạp nhất thế giới, trong khi số lượng đại từ xưng hô trong Tiếng Anh là khá nhỏ, liên hệ chặt chẽ với cấu trúc ngữ pháp. Do vậy một đại từ xưng hô Tiếng Anh có thể tương ứng với khá nhiều đại từ xưng hô và các cụm từ có vai trò như đại từ xưng hô trong Tiếng Việt. Ví dụ, đại từ “they” Tiếng Anh có thể là bản dịch của “các anh ấy”, “các cô ấy”, “chúng nó”, “bọn chúng”, “chúng”, “bọn nó”, “họ”. Không phải tất cả các từ này đều xuất hiện trong từ điển Tiếng Việt như những mục từ riêng biệt mà chúng được coi là những cụm từ.

Động từ “đi”: Trong Tiếng Việt, động từ đi có thể dùng với động từ khác như “đi học”, “đi chơi”, “đi bơi”... Khi dịch sang Tiếng Anh, để bản dịch sát với thực tế sử dụng, cần có một cách thức chuyển đổi riêng cho trường hợp đặc biệt này.

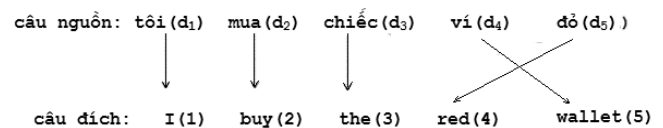
Do đặc điểm từ vựng hóa hoàn toàn của văn phạm liên kết, những dấu hiệu liên quan đến các hiện tượng nói trên đều có thể nhận biết được qua các công thức liên kết. Điều đó cho phép xây dựng các luật chuyển đổi như trong mục V.

Khác biệt về trật tự từ

Theo [7], cả Tiếng Anh và Tiếng Việt đều cùng loại hình S – V – O, cùng giống nhau về trật tự từ ở phần định từ (determiner), mạo từ (article) và giới ngữ (prepositional phrase) sau danh từ. Tuy nhiên, những khác biệt nổi bật về trật tự từ giữa Tiếng Việt và Tiếng Anh có thể liệt kê dưới đây.

Trật tự trong cụm danh từ-tính từ

Trong Tiếng Anh, tính từ luôn đứng trước danh từ trong khi ở Tiếng Việt danh từ lại đứng trước tính từ. Hình 2 dưới đây minh họa việc sắp xếp lại trật tự từ trong trường hợp này:



Hình 2. Sắp xếp lại trật tự từ

Khi câu “tôi mua một chiếc ví đỏ” được áp dụng luật dịch, kết quả trả ra là “I buy a red wallet”. Ở câu đích, tính từ “red” đứng trước danh từ “wallet”. Điều này là kết quả của việc từ “ví” ở vị trí thứ 4 trong câu nguồn đã được ánh xạ thành từ “wallet” ở vị trí thứ 5 trong câu đích, trong khi từ “đỏ” ở vị trí thứ 5 trong câu nguồn thì được ánh xạ thành từ “red” ở vị trí thứ 4. Việc thay đổi trật tự cho cụm hai, ba tính từ đi sau danh từ cũng tương tự như vậy. Chú ý rằng vị trí nói đến ở đây là vị trí tương đối giữa các từ được nhắc đến trong luật. Giữa chúng có thể tồn tại các từ khác. Xét ví dụ “cô gái nhỏ rất xinh”, hệ thống dịch cần thay đổi vị trí các từ “cô gái”, “nhỏ”, “xinh”, từ “rất” luôn phải giữ vị trí đứng ngay trước từ “xinh”. Bản dịch Tiếng Anh của câu này cần phải là “very pretty little girl”. Văn phạm liên kết cho phép kết nối hai từ không liên

kề nên việc thể hiện những luật chuyển đổi trật tự từ khá thuận tiện.

Trật tự câu nghi vấn, câu nghi vấn – phủ định

Dạng câu hỏi thường dùng nhất trong Tiếng Việt liên quan đến từ (cụm từ) để hỏi, ví dụ “*tại sao*”, “*ai*”, “*như thế nào*”. Các từ này thường ở đầu hoặc cuối câu. Trong dạng câu hỏi “*wh*” Tiếng Anh, từ để hỏi luôn luôn đứng trước. Do vậy cần biến đổi trật tự từ cùng với một số biến đổi khác.

Ngoài ra cũng có một số trường hợp cần thay đổi trật tự từ như trong câu chứa phụ từ chỉ tần suất, câu dùng sở hữu cách...

Trật tự từ không lân cận: là sự phụ thuộc được xác lập giữa các từ cách nhau một khoảng nhất định. Phụ thuộc dạng này có thể do “khoảng trống” (gap) tạo nên bởi sự thay đổi vị trí của từ nào đó, ví dụ “*Quà cưới cho cô dâu, tôi đã gửi rồi*”, hay trong các câu ghép, câu phức với nhiều mệnh đề như “*Cái áo tôi mua hôm qua rất đẹp*”. Khi dịch sang Tiếng Anh, để đảm bảo nghĩa của câu, thường phải thay đổi vị trí của từ, nhưng sự thay đổi vị trí cho loại câu này khó có thể biểu diễn một cách đơn giản bằng những luật dịch. Việc phân tích cũng như dịch loại câu này đòi hỏi những phương pháp thống kê, học máy với những bộ ngữ liệu lớn mà chúng tôi chưa có điều kiện xây dựng. Một số dạng câu có thể không biểu diễn được bằng văn phạm liên kết như đã trình bày ở mục 2.

V. HỆ THỐNG DỊCH VIỆT ANH SỬ DỤNG ADJ

Như vậy, để xây dựng bộ dịch dựa trên dạng tuyên có chú giải, ba vấn đề quan trọng nhất phải giải quyết là

1. Tìm nghĩa từ
2. Chuyển đổi cấu trúc câu
3. Hoàn thiện bản dịch

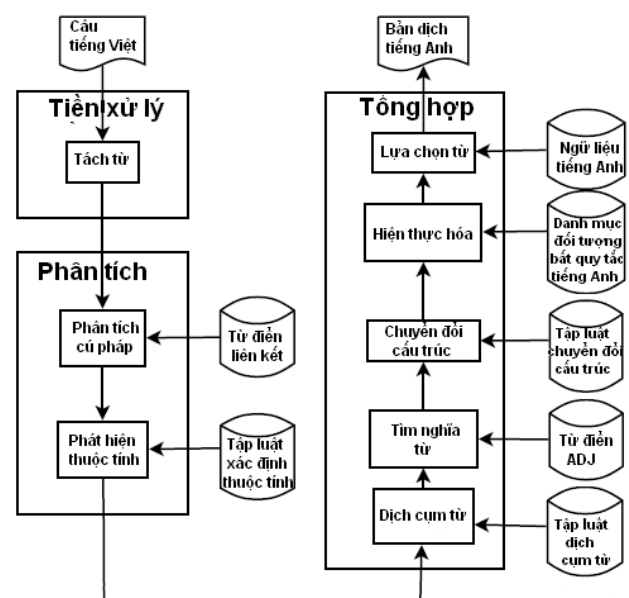
Hình 3 dưới đây mô tả kiến trúc của hệ dịch Việt Anh dựa trên dạng tuyên có chú giải.

Sơ đồ kiến trúc cho thấy hệ thống gồm 3 phần chính:

- Phần tiền xử lý thực hiện tách từ cho câu đưa vào. Giai đoạn này không gán nhãn từ vì bộ phân tích cú

pháp sẽ gán nhãn từ theo liên kết tìm được. Để tách từ, chúng tôi đã sử dụng bộ vnTokenizer của TS Lê Hồng Phương [16].

- Phần phân tích thực hiện phân tích cú pháp bằng bộ phân tích cú pháp liên kết. Đầu ra của quá trình phân tích cú pháp là một phân tích liên kết với cấu trúc đã trình bày ở mục trên. Trong khuôn khổ của bài báo, chúng tôi chưa đề cập vấn đề dịch câu ghép và câu phức nên kết quả nhận được từ bộ phân tích cú pháp là một phân tích liên kết của câu đơn hoặc câu ghép hai mệnh đề có thể phân tích dựa trên liên kết của từ nối. Qua phân tích các liên kết tìm được, hệ thống sẽ xác định các thuộc tính liên quan đến ngôi, số, thì, thể, vv...
- Phần tổng hợp cho phép tạo ra bản dịch bao gồm các công việc sau :
 - Dịch một số cụm từ đặc biệt : “*đi học*”, “*bọn chúng nó*”...
 - Tra nghĩa từ theo dạng tuyên trong từ điển ADJ.
 - Thay đổi hình thái từ dựa trên các thuộc tính tìm được (hiện thực hóa).
 - Tìm phương án dịch tổng thể tốt nhất.



Hình 3. Kiến trúc của hệ thống dịch dựa trên dạng tuyên có chú giải

VI. XÂY DỰNG BỘ LUẬT DỊCH

Như đã mô tả ở trên, hệ thống dịch cần sử dụng ba tập luật liên quan đến các công việc khác nhau: phát hiện thuộc tính, dịch cụm từ, chuyển đổi cấu trúc. Trong phần này chúng tôi mô tả chi tiết các luật điển hình và văn phạm phi ngữ cảnh sinh ra bộ luật.

Trong các luật điển hình dưới đây của hệ thống, ký hiệu W_1, W_2, W_3 đại diện cho các từ, D_1, D_2, D_3 chỉ đây các tên kết nối thuộc danh sách trái hay phải của một dạng tuyến nào đó.

Luật phát hiện thuộc tính

Thuộc tính ở đây là những thông tin cần được lưu trữ lại cho mỗi từ để biến đổi hình thái thích hợp, ví dụ số nhiều của danh từ, thì, thể của động từ, ngôi của đại từ, loại cấu trúc so sánh (so sánh ngang bằng, so sánh hơn kém, so sánh bậc nhất). Dưới đây là một số ví dụ về luật phát hiện thuộc tính:

Luật phát hiện thuộc tính về số nhiều của danh từ:
Căn cứ vào kết nối DpNt của những định từ chỉ số nhiều “*những*”, “*các*”, “*số đông*” với danh từ, có thể đưa giá trị PLURAL vào thuộc tính của danh từ liên kết với chúng:

$W_1(D_1)(DpNt) \quad W_2(DpNt)(D_2)$
 $\rightarrow W_1'W_2'(\text{number} = \text{PLURAL})$

Luật phát hiện thuộc tính thì của động từ: Thì của động từ được thể hiện bằng các liên kết với các phụ từ chỉ thì. Thì quá khứ được thể hiện qua các liên kết RpVt, RpVs, thì tương lai RfVt, RfVs, tiếp diễn RcVt, RcVc, hoàn thành RhVt, RhVc. Các liên kết của thời quá khứ hay tương lai xuất hiện trong phân tích sẽ được xử lý theo các luật nhằm xác định thuộc tính tense cho động từ. Với những thì phức tạp hơn như các thì tiếp diễn hay hoàn thành, không chỉ biến đổi hình thái động từ mà còn thêm từ khác như “*to be*”, “*to have*”, do vậy thuộc tính được xác định giá trị là thuộc tính form. Ví dụ,

$W_1(D_1)(RpVt) \quad W_2(RpVt)(D_2) \rightarrow$
 $W_1'W_2'(\text{tense} = \text{PAST})$
 $W_1(D_1)(RfVt) \quad W_2(RfVt)(D_2) \rightarrow$
 $W_1'W_2'(\text{tense} = \text{FUTURE})$

Luật xác định ngôi của đại từ xưng hô: đối với đại từ xưng hô thì không cần căn cứ vào liên kết vì trong Tiếng Việt số lượng đại từ xưng hô là khá nhỏ, do vậy luật căn cứ vào chính giá trị từ và loại của từ để xác định ngôi:

$tôi[p] \rightarrow I(\text{person} = \text{FIRST})$
 $nó[p] \rightarrow he(\text{person} = \text{THIRD})$

Ký hiệu [p] trong luật thể hiện loại của từ đứng ngay bên trái. Thông tin này có trong từ điển liên kết.

Sau khi xác định được ngôi của đại từ xưng hô, thuộc tính về ngôi phải được lan truyền cho động từ để chia đúng ngôi, đặc biệt là động từ “*là*” vì động từ “*to be*” tương ứng chia khác nhau ở tất cả các ngôi thể hiện qua các luật sau:

$W_1(D_1)(SV) \quad W_2(SV)(D_2)$
 $\rightarrow W_1'W_2'(\text{person} = W_1'.\text{person})$

Động từ “*là*”

$W_1(D_1)(DT_LA) \quad W_2(DT_LA)(D_2)$
 $\rightarrow W_1'W_2'(\text{person} = W_1'.\text{person})$

Luật dịch cụm từ

Dịch cụm từ sang đại từ xưng hô Tiếng Anh

Cách xưng hô trong Tiếng Việt vô cùng phức tạp như đã trình bày trong mục IV. Do vậy cần một tập luật để dịch chính xác nghĩa đại từ xưng hô sang Tiếng Anh. Ví dụ, mối liên hệ của từ “*cô*” thông qua liên kết NtPd+ và SV+ cho phép dịch “*cô ấy*” thành “*she*” còn NtPd+ và O- hay EpNt cho phép dịch “*cô ấy*” thành “*her*” được thể hiện trong các luật:

$cô(D_1)(NtPd, SV) \quad W_2(NtPd)(D_2) \rightarrow she \quad W_2'$
 $W_1(D_1)(EpNt) \quad cô(EpNt)(NtPd) \quad W_2(NtPd)(D_2) \rightarrow W_1' \quad her$

Sau đó, tất nhiên cũng cần những luật để gán thông tin về ngôi cho động từ tương ứng.

Cụm từ là thành ngữ: sử dụng từ điển thành ngữ.

Cụm từ với từ “đi”

Từ “*đi*” đóng một vai trò khá đặc biệt khi dịch Việt Anh. Trong Tiếng Việt, từ “*đi*” có thể kết hợp với một số động từ thường, như “*đi bơi*”, “*đi mua sắm*”, “*đi học*”, “*đi chơi*”. Hiện tượng này đã được thể hiện trong từ điển liên kết với liên kết DI cho động từ “*đi*”

và một số động từ khác. Trong nhiều trường hợp, công thức dịch sẽ là go + verb + ing (go swimming, go fishing...)

$di(D_1)(DI) w_2(DI)(D_2) \rightarrow$

go w_2' (form = PRESENT_PARTICIPLE)

Một số ngoại lệ : “đi học” --> “go to school”, “đi chơi” --> “go out”, “đi ngủ” --> “go to bed” đã được thể hiện qua những luật dịch riêng

Luật dịch quan hệ sở hữu

Quan hệ sở hữu trong Tiếng Việt có thể ở dạng hiện thông qua giới từ “của” (đôi khi là danh từ “nhà”, ví dụ “lợn nhà tôi”). Các luật dịch sẽ chuyển đại từ xưng hô hoặc danh từ đi sau giới từ “của” sang tính từ sở hữu, danh từ sang sở hữu cách.

Ví dụ, luật

của() (EoPp) tôi(EoPp)() $\rightarrow$ my

Trong Tiếng Việt còn có quan hệ sở hữu ẩn (không có từ “của”). Khi phân tích câu có quan hệ sở hữu ẩn, cần những luật dịch thể hiện quan hệ đó, ví dụ

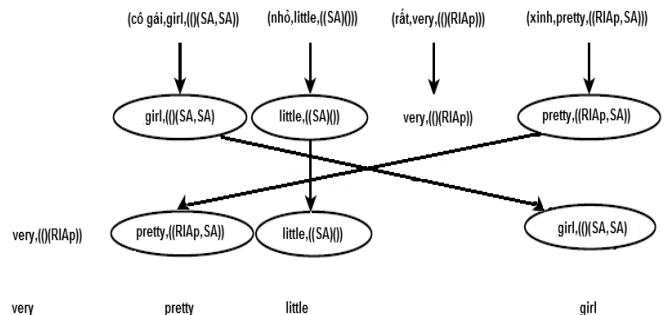
$w_1(D_1)(SHA) tôi(SHA)(D_2) \rightarrow my w_1'$

Luật thay đổi trật tự từ

Việc hoán đổi vị trí các từ và cụm từ cần chú ý tới tính đúng cú pháp của câu đích. Chúng tôi nhận thấy rằng hai từ có liên kết trong câu nguồn thì liên kết đó vẫn cần bảo tồn trong câu đích, trừ trường hợp có từ bị loại bỏ. Sau khi thực hiện hoán đổi vị trí, các từ trong câu đích cũng cần phải tạo thành câu đúng cú pháp liên kết, đặc biệt là tính phẳng. Khi hai từ đổi vị trí cho nhau, liên kết giữa chúng ở câu đích sẽ đổi chiều, kết nối trong dạng tuyển tương ứng sẽ chuyển từ danh sách bên trái của dạng tuyển sang danh sách bên phải hoặc ngược lại. Để các liên kết trong câu đích không cắt nhau, các từ nằm trong khoảng giữa hai từ cũng cần di chuyển thích hợp. Ví dụ, xét các bước dịch trong hình 4 của cụm từ “cô gái nhỏ rất xinh”

Nếu các từ “girl”, “little” và “pretty” đổi vị trí, liên kết giữa chúng cũng sẽ đổi chiều, dạng tuyển thay đổi. Do vậy không chỉ có hai từ đổi vị trí mà phải đổi vị trí hai cụm từ “very pretty” với các từ “little” và “girl” do có liên kết RIAp giữa “very” và “pretty”, liên kết SA

giữa “girl” với “little” và “pretty”. Việc đổi chỗ không ảnh hưởng tới tính phẳng của phân tích nên ta nhận được kết quả cuối cùng là “very pretty little girl”



Hình 4. Thay đổi trật tự từ cho bản dịch cụm từ “cô gái nhỏ rất xinh”

Luật đổi vị trí danh từ - tính từ

Sau đây là một số luật cho phép thay đổi vị trí giữa tính từ bổ nghĩa cho danh từ và tính từ bổ nghĩa cho tính từ khác. Trong Tiếng Việt thường chỉ có đến 2 tính từ bổ nghĩa cho một danh từ (không phân cách bằng dấu phẩy).

$w_1(D_1)(ApAp) w_2(ApAp)(D_2) \rightarrow w_2' w_1'$

$w_1(D_1)(SA SA) w_2(SA)(D_2) w_3(SA)() \rightarrow w_3' w_2' w_1'$

$w_1(D_1)(SA) w_2(SA)(D_2) \rightarrow w_2' w_1'$

Luật dịch cấu trúc phủ định

Các từ chỉ ra cấu trúc phủ định là các phụ từ phủ định “không”, “chẳng”, “chưa”... (loại từ là Rn). Kết nối giữa loại từ này với động từ là RnV. Khi dịch cấu trúc này cần chú ý đến việc chia động từ phù hợp với thì của động từ ở câu nguồn (từ “chưa” tương ứng thì hiện tại hoàn thành, các từ khác dịch sang thời hiện tại):

$w_1(D_1)(SV) chưa(D_2)(RnV) w_3(RnV SV)(D_3) \rightarrow w_1'(D_1')(D_2') have(number = w_3. number, person = w_3. person)[v](D_2')(N_PP) not(N)(D_3') w_3'(negative = FALSE)(_PP)(D_4')$

Luật dịch cấu trúc nghi vấn

Trong Tiếng Việt có một số từ (cụm từ) để hỏi như “tại sao”, “khi nào”... Từ để hỏi có thể đứng đầu hoặc cuối câu. Nhận biết cấu trúc này thông qua kết nối THT (từ hỏi đứng đầu câu) hoặc THS (từ hỏi đứng cuối câu). Khi dịch sang Tiếng Anh với các động từ

thông thường (không phải động từ “to be”) cần thêm trợ động từ vào giữa từ để hỏi và chủ ngữ. Ví dụ luật dưới đây cho phép dịch câu hỏi với từ để hỏi đứng đầu câu và phủ định từ

$W_1() (THT) W_2(D_2) (SV) W_3() (RnV) W_4(RnV SV THT) () \rightarrow W_1'$
 $do(number = W_4.number, person = W_4.person, tense =$
 $W_4.tense, contracted-negation = TRUE)[v]() (I_) W_2$
 $W_4(I_)()$

Cú pháp của các luật được mô tả bằng EBNF như sau:

```
<rule> ::= <name><lhs><arrow><rhs>
<name> ::= <word> " : "
<lhs> ::= <pattern> <pattern>*
<arrow> ::= "-->" | ">"
<rhs> ::= <expression> <expression>*
<pattern> ::=
<string-pattern><disjunct-pattern>
<string-pattern> ::= "*" | <word>
<disjunct-pattern> ::=
"("<link-list>")" "("<link-list> ")"
<link-list> ::= (<connector>)*
<expression> ::=
[<word-ref>][<function-call>][<word-expr>]
<function-call> ::=
<name> "(" (<expression>)* ")"
<word-ref> ::=
"$" <number> [<disjunct-pattern>]
<word-expr> ::= <word> [<disjunct-pattern>]
```

Các đối tượng sau được tiền xử lý và coi như ký hiệu kết thúc: <name> (tên tự đặt cho luật hay hàm), <word> (từ), <number> (số tự nhiên).

Tập luật được lưu dưới dạng XML với cấu trúc sau:

```
<rules>
<rule> <!-- luật doc lap --> </ rule>
<group>
<!-- luật trong nhóm -->
<rule name="past"> ... </ rule>
<rule name="future"> ... </ rule>
<rule name="present"
excludes="past,future"> ...</ rule>
<rule requires="present"> ... </ rule>
</group>
</ rules>
```

Trong đó:

- Ký hiệu ... chỉ nội dung của luật, có thể viết trên nhiều dòng.
- Thuộc tính *requires* chứa tập các luật cần thực hiện trước luật đang xét

- Thuộc tính *excludes* ngăn cản việc thực hiện luật đang xét nếu một trong các luật trong danh sách sau từ khóa *excludes* đã được thực hiện

Luật được đọc từ tệp XML phân tích và xử lý tự động theo văn phạm. Với cú pháp khá đơn giản, người sử dụng có thể dễ dàng thêm luật mới.

VII. HOÀN THIỆN BẢN DỊCH

Sau khi đã phát hiện thuộc tính, di chuyển vị trí, biến đổi hình thái thích hợp, còn một số vấn đề cần thực hiện để tạo ra bản dịch có chất lượng tốt. Đó là:

- Hiện tượng chuyển loại từ:
 - Danh từ có vai trò tính từ. Khi dịch sang Tiếng Anh cần chuyển thành tính từ, ví dụ “*huy chương vàng*”, *vàng* ở đây là từ chỉ chất liệu với liên kết CH, cần chuyển sang tính từ tương ứng.
 - Tính từ hay động từ đi sau “*sử*”, “*việc*” làm chuyển loại cả cụm từ thành danh từ.
- Tính từ bỏ nghĩa cho động từ : Tiếng Việt không có phó từ, nhưng sang Tiếng Anh cần dịch thành phó từ.
- Loại bỏ danh từ chỉ loại.

Chúng tôi đã xây dựng tập luật bao trùm được các hiện tượng này.

Lựa chọn nghĩa từ

Dù đã qua nhiều công đoạn khử nhập nhằng trong quá trình phân tích cú pháp, đến đây ta vẫn có thể gặp hiện tượng nhập nhằng về nghĩa từ, đó là khi một từ Tiếng Việt tương ứng với nhiều nghĩa Tiếng Anh khác nhau. Chẳng hạn từ “*làm*” có thể mang nghĩa “*make*” hay “*manufacture*” (ở đây đã loại trừ trường hợp từ “*làm*” với nghĩa tương ứng “*work*” gắn với một công thức khác so với từ “*làm*” mang hai nghĩa trên).

Như vậy, dù chỉ có một phân tích liên kết duy nhất được chọn, vẫn có thể có nhiều bản dịch tương ứng với nhiều tổ hợp nghĩa khác nhau của các từ. Vấn đề ở đây là chọn ra tổ hợp “trôi chảy” nhất, tức là cách thức hay được người bản ngữ sử dụng nhất. Chúng tôi đã giải quyết bằng cách dùng bộ ngữ liệu Tiếng Anh và tính xác suất để chọn bản dịch tốt nhất. Bộ ngữ liệu

được chọn là COCA rút gọn 0 (chỉ chứa những n gram xuất hiện ba lần trở lên).

Câu được chọn dựa theo tiêu chí sau (đã làm tròn)

$$\operatorname{argmax}_{(w_1, w_2, \dots, w_s)} P(S) = \max_{(w_1, w_2, \dots, w_s)} \left\{ \log(P(w_1)) + \sum_{i=2}^s \log P(w_i | w_{i-1}) \right\}$$

VIII. KẾT QUẢ THỬ NGHIỆM

Như đã giới hạn phạm vi quan tâm ngay từ đầu, việc thử nghiệm hệ thống dịch với ADJ là để minh họa khả năng biểu diễn của văn phạm liên kết Tiếng Việt, còn phải hoàn thiện nhiều để trở thành một công cụ dịch phổ biến. Tuy nhiên, đề xây dựng bộ ngữ liệu thử nghiệm cũng là một khó khăn. Hiện chưa có bộ ngữ liệu chuẩn cho dịch Việt - Anh nên chúng tôi đã tự xây dựng bộ ngữ liệu riêng cho mình.

Bộ ngữ liệu chúng tôi đã thu thập gồm 336 câu được lấy từ sách dạy Tiếng Việt cho người nước ngoài trình độ nâng cao [15]. Ưu điểm của tập mẫu này là chúng được viết bởi các giáo sư về Tiếng Việt và bản dịch Tiếng Anh đã được hiệu đính cẩn thận. Dưới đây là ví dụ minh họa hoạt động của hệ thống dịch xử lý một số mẫu câu:

1. "Mẹ tôi là một bác sĩ giỏi"

Kết quả phân tích cú pháp:

```
# +----DT_LA-----+-----LA_DT-----+
# +-SHA--+ | +----McN-----+DT_TT--+
# | | | | |
# Mẹ.n tôi.p là.v một bác_sĩ.n giỏi.a
```

Các dạng tuyến tìm được cho mỗi từ là:

```
mẹ: ()(SHA DT_LA)
tôi: (SHA)()
là: (DT_LA)(LA_DT)
một: ()(McN)
bác sĩ: (McN LA_DT)(DT_TT)
giỏi: (DT_TT)()
```

Để dịch câu này, các luật sau đã được áp dụng

- Luật xác định ngôi cho danh từ “mẹ”.
- Luật dịch quan hệ sở hữu ẩn.
- Luật chuyển đổi thứ tự giữa nghĩa của danh từ “bác sĩ” và tính từ “giỏi”.

Bản dịch được hệ thống của chúng tôi và hệ thống Google đưa ra:

My mother is a good doctor

Kết quả dịch với VETran:

My mother is a jurisprudent physician.

Kết quả dịch của ba hệ thống không có sự khác biệt lớn. Hệ thống của chúng tôi và Google có sự đối chiếu với ngữ liệu mẫu khi chọn từ nên đưa ra bản dịch “good doctor” trong khi VETran dùng cụm từ “jurisprudent physician” về nghĩa trong từ điển Việt Anh không khác biệt nhưng không sát nghĩa thực tế.

2. “Báo săn là loài động vật nhanh nhất thế giới”

Kết quả phân tích cú pháp

```
# +-----LA_DT-----+
# +-DT_LA--+ | +---DV---+DT_TT--+TT_SS+---NHAT_DTV--+
# | | | | |
# Báo_săn.n là.v loài động_vật.n nhanh.a nhất thế_giới.n
```

Các dạng tuyến nhận được:

```
báo săn: ()(DT_LA)
là: (DT_LA)(LA_DT)
loài: ()(DV)
động vật: (DV LA_DT)(DT_TT)
nhanh: (DT_TT)(TT_SS)
nhất: (TT_SS)(NHAT_DT)
thế giới: (NHAT_DTV)()
```

Quá trình dịch thông qua nhiều luật được mô tả trong Hình 5.



Hình 5. Quá trình dịch câu “Báo săn là loại động vật nhanh nhất thế giới”

Kết quả thực hiện của hệ thống dịch ADJ
Cheetah is the quickest animal world

Kết quả thu được với VETran:
Cheetah is world' s fast animal the kind.

Kết quả do Google đưa ra:
Alert hunt is the world's fastest animal

Cả ba bản dịch đều có những hạn chế. Bản dịch của ADJ đã không phát hiện được quan hệ sở hữu của từ “*thế giới*” trong khi bản dịch của VETrans dịch được quan hệ đó nhưng lại không dịch được tính từ so sánh cấp cao nhất và dịch sai từ “*loài*”. Bản dịch của Google dịch rất tốt cụm từ “*nhANH NHẤT thế giới*” nhưng lại gặp lỗi ở từ “*báo săn*”.

So sánh trên toàn thể bộ ngữ liệu, có thể thấy với những cụm từ thường dùng, Google cho kết quả giống ngôn ngữ thực hơn. Tuy nhiên kết quả của hệ thống dịch của chúng tôi và VETran thường đúng về cú pháp và hình thái, trong đó hệ thống dịch của chúng tôi có phần “trôi chảy” hơn, do mỗi liên kết được xác lập đến từng từ cá biệt. Rõ ràng việc kết hợp các phương pháp khác nhau sẽ nâng chất lượng của bản dịch lên cao hơn.

Hiện nay chúng tôi chưa có một bộ ngữ liệu đủ lớn để giải quyết triệt để vấn đề nhập nhằng trong phân tích cú pháp, do vậy độ chính xác của bộ phân tích cú pháp còn chưa cao. Đối với bộ ngữ liệu 336 câu hội thoại, còn nhiều cụm ở dạng văn nói, chưa có trong ngữ liệu toàn các câu văn viết, nên kết quả trả về của bộ phân tích cú pháp còn thấp (độ chính xác (precision): 22.7%, độ phủ (recall): 28.8%, độ đo F-score: 0.28). Nếu dùng nguyên dạng kết quả của bộ phân tích cú pháp, có thể không đánh giá chính xác chất lượng của hệ thống dịch. Để nghiên cứu một cách tổng thể về ảnh hưởng của các thành phần trong hệ thống đến chất lượng dịch, chúng tôi đã thử nghiệm trên 2 hệ thống sau:

ADJ1: Cho phép loại bỏ bớt những sai sót có thể có trong quá trình phân tích cú pháp bằng cách xác định một số ràng buộc để chọn chính xác phân tích của câu, cụ thể là báo trước một số cặp từ chắc chắn xuất hiện liên kết. Đây cũng là kỹ thuật được dùng trong [14] để

hạn chế số phân tích đưa ra. Câu đưa vào đã chỉnh kết quả tách từ. Độ chính xác của bộ phân tích cú pháp cho ADJ1 là 80.2%, độ phủ 81.4%, F-score 0.81.

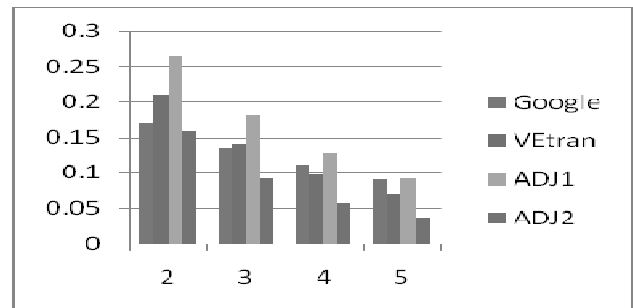
ADJ2: Không cho phép ràng buộc và tách từ.

Dù bộ ngữ liệu còn nhỏ, chúng tôi đã sử dụng phương pháp BLEU [17] với tham số $n = 2, 3, 4, 5$ để so sánh với kết quả đạt được của VETran và Google. Kết quả nhận được thể hiện trong Bảng 2.

Bảng 2. So sánh kết quả các hệ thống dịch

	Google	VETran	ADJ1	ADJ2
2	0.169816	0.209987	0.263627	0.157450
3	0.133085	0.140612	0.181787	0.091807
4	0.109895	0.096798	0.127502	0.056950
5	0.090472	0.069292	0.091302	0.036461

Biểu đồ trong hình 6 cho phép so sánh điểm BLEU của các hệ thống dịch nói trên của chúng tôi với hai hệ thống dịch Việt-Anh phổ biến là Google Translation và VETrans.



Hình 6. So sánh điểm BLEU của các hệ thống

IX. KẾT LUẬN

Hệ thống dịch sử dụng dạng tuyến có chú giải được chúng tôi xây dựng với mục đích minh họa cho khả năng biểu diễn Tiếng Việt của văn phạm liên kết. Tuy vậy, nếu đánh giá như một hệ thống dịch, kết quả nhận được cũng rất đáng chú ý: với kết quả tốt của bộ tách từ và phân tích cú pháp, hệ thống của chúng tôi đạt kết quả nhỉnh hơn một chút so với Google và VETran. Chú ý rằng, để đạt kết quả này, bộ luật dịch của chúng tôi chưa có tới 300 luật, ít hơn rất nhiều so với VETran và

cũng không cần sử dụng bộ ngữ liệu song ngữ. Dù mới là thử nghiệm trên bộ ngữ liệu nhỏ, có thể thấy khả năng sử dụng mô hình văn phạm liên kết cho bài toán dịch máy là rất có triển vọng.

Tập luật dịch của chúng tôi bao gồm ba tập con. Thứ tự áp dụng luật như trên sơ đồ ở hình 4: xác định thuộc tính → dịch cụm từ → chuyển đổi cấu trúc. Sự nhập nhằng khi áp dụng luật (nếu có) chỉ có thể xảy ra trong từng tập con. Tuy nhiên, với văn phạm liên kết, luật chỉ được sử dụng khi thỏa mãn cả hai yếu tố:

- Từ đang xét xuất hiện trong luật
- Tất cả các mối liên kết của từ được nêu trong luật phải thỏa mãn.

Ngoài ra, thuộc tính *exclude* của một số luật (đã mô tả ở trên) cũng góp phần khử nhập nhằng. Do vậy, khi phân tích cú pháp của câu đã xác định, rất khó xảy ra việc nhập nhằng khi áp dụng luật. Trong ba tập luật của chúng tôi, chưa có luật nào có thể gây nhập nhằng trong lúc lựa chọn. Sự nhập nhằng chủ yếu xảy ra khi phân tích, chẳng hạn với hai câu “tôi bán hoa rất nhanh” và “tôi bán hoa rất tươi” có thể dẫn đến nhầm lẫn khi không có dấu hiệu nào cho thấy tính từ chỉ tính chất bổ nghĩa cho từ “hoa” hay từ “bán”. Tuy nhiên khi đã xác định phân tích, nếu là:

```
# +---SV---+ +-----SA-----+
# + +---O---+ +---RLAp---+
# | | | | |
# Tôi.p bán.v hoa.n rất.r tươi.a
```

thì luật về thay đổi trật tự từ được áp dụng. Nếu phân tích được chọn là:

```
# +-----VtAp-----+
# +---SV---+ +---O---+ +---RLAp---+
# | | | | |
# Tôi.p bán.v hoa.n rất.r nhanh.a
```

thì luật về chuyển loại nghĩa của từ “nhanh” sang phó từ Tiếng Anh lại được áp dụng.

Tuy đã đạt được kết quả nhất định trong dịch máy Việt – Anh, hệ thống của chúng tôi vẫn còn những vấn đề chưa giải quyết trọn vẹn:

• *Dịch câu có cấu trúc liên hợp* (coordination) sử dụng kết nối lớn, ví dụ, cụm từ Tiếng Việt trong [7] “một sinh viên khỏe mạnh, cao và tử tế”. Việc dịch loại câu này đòi hỏi một phân tích cú pháp chính xác,

chỉ có được khi khử nhập nhằng liên hợp trong câu chứa từ “và” và dấu phẩy.

• *Dịch câu ghép và câu phức*: Câu ghép và câu phức chứa từ hai nòng cốt trở lên, trong đó câu phức có chứa một nòng cốt bao các nòng cốt còn lại [5]. Trong tập ngữ liệu mẫu, đã có một số câu ghép 2 mệnh đề, tuy nhiên chất lượng dịch các câu này chưa được tốt. Có thể thấy việc xử lý các loại câu ghép là khả thi vì chúng tôi đã xây dựng được bộ phân tích cú pháp xử lý khá tốt trường hợp nhập nhằng liên hợp và phân tích câu ghép với nhiều mệnh đề.

Việc nhận biết giới hạn các mệnh đề trong câu phức, cũng như các thành phần cụm chủ vị đôi khi đòi hỏi thiết lập một liên kết giữa các từ không liền kề. Các xử lý để phân tách mệnh đề hay giải quyết vấn đề nhập nhằng về cụm trạng từ hiện nay đều theo hướng tiếp cận học máy trên tập ngữ liệu lớn. Chúng tôi sẽ tiếp tục phát triển hệ thống theo hướng này khi đã xây dựng được bộ ngữ liệu mẫu đủ lớn.

Một vấn đề khác cũng gây khó khăn cho chúng tôi: dịch cụm từ dạng n - 1 (n từ Tiếng Việt sang 1 từ Tiếng Anh). Ngoài những cụm từ rất phổ biến mà chúng tôi đã xử lý, cần đến sự hỗ trợ của từ điển thành ngữ và bộ ngữ liệu mẫu song ngữ.

Với đặc điểm hoàn toàn từ vựng hóa của văn phạm liên kết, bộ luật dịch của hệ thống thể hiện được những đặc điểm hết sức riêng và cá biệt của ngôn ngữ nguồn và ngôn ngữ đích. Công việc này chắc chắn cần những hiểu biết sâu về cú pháp, từ pháp của cả hai ngôn ngữ. Bộ luật dịch có thể thay đổi hoàn toàn nếu thay đổi cặp ngôn ngữ nguồn – đích, nghĩa là khó có thể sử dụng cho cặp ngôn ngữ khác. Tuy nhiên, để mở rộng hệ thống dịch, có thể quan tâm đến công cụ cho phép các nhà ngôn ngữ định nghĩa các quy tắc cú pháp [4]. Nếu theo hướng tiếp cận này, việc phân tích cú pháp theo biểu đồ (chart parsing) từ văn phạm liên kết cũng dễ hơn so với các mô hình khác vì phân tích liên kết thực chất đã có dạng biểu đồ. Như vậy có thể tính đến khả năng mở rộng hệ thống dịch cho các cặp ngôn ngữ khác.

Như đã trình bày ở đầu bài, do chưa đủ tài nguyên để xây dựng một hệ thống dịch máy thật hoàn thiện, hệ thống dịch máy của chúng tôi nhằm mục đích minh họa khả năng biểu diễn Tiếng Việt của văn phạm liên kết. Tuy nhiên, với chất lượng dịch khá thuyết phục, việc kết hợp mô hình dịch này với hệ thống dịch theo cách tiếp cận thống kê chắc chắn sẽ nâng cao được chất lượng bản dịch do có thể kết hợp sự trôi chảy của phương pháp thống kê với sự chính xác của những biến đổi hình thái và cú pháp. Một trong những minh chứng cho điều đó là sự kết hợp phân tích liên kết để hoàn chỉnh bản dịch ở hệ dịch máy trên nền ví dụ. Tỷ lệ câu dịch hoàn toàn đúng với câu mẫu đã tăng khá nhiều. Kết hợp giữa cách tiếp cận thống kê và văn phạm liên kết là hướng phát triển của hệ thống trong thời gian tới.

PHỤ LỤC. Chú thích ý nghĩa các kết nối

SV: Kết nối danh từ, đại từ xưng hô làm chủ ngữ với động từ.

SA: Kết nối danh từ, đại từ xưng hô với tính từ.

DT_LA: Kết nối danh từ, đại từ xưng hô với động từ quan hệ “là”.

LA_DT: Kết nối động từ “là” với danh từ.

O: Kết nối động từ và bổ ngữ trực tiếp.

DI: Kết nối động từ “đi” với động từ khác.

NtPd: Kết nối danh từ cụ thể với đại từ chỉ định.

DpNt: Kết nối định từ chỉ số nhiều với danh từ cụ thể.

RpVt: Kết nối định từ chỉ thì quá khứ và động từ.

RfVt: Kết nối định từ chỉ thì tương lai và động từ.

EpNt: Kết nối giới từ chỉ vị trí và danh từ cụ thể.

VtEp: Kết nối ngoại động từ và giới từ vị trí.

VtAp: Kết nối ngoại động từ và tính từ chỉ tính chất.

VmVt: Kết nối động từ tình thái và động từ cụ thể.

EoPp: Kết nối giới từ “của” với đại từ xưng hô.

SHA: Kết nối hai danh từ chỉ quan hệ sở hữu ẩn.

RnV: Kết nối phủ định từ và động từ.

THT: kết nối các từ để hỏi và động từ.

SS_NHAT: Kết nối tính từ với từ “nhất”.

NHAT_DT: Kết nối từ “nhất” với danh từ đứng sau.

Các kết nối có tên bắt đầu và kết thúc bằng “_” là kết nối của Tiếng Anh (theo [19]).

TÀI LIỆU THAM KHẢO

- [1] ADJI, T.B. *Applying Link Grammar Formalism in the Development of English-Indonesian Machine Translation System*. Proceedings of the 9th AISC International Conference, Intelligent Computer Mathematics, Springer, 2008, p17-23
- [2] T.B. ADJI, B. BAHARUDIN, N. ZAMIN. *Annotated Disjunct in Link Grammar for Machine Translation*. International Conference on Intelligent and Advanced Systems, Kuala Lumpur, 2007, p. 205-208
- [3] T.B. ADJI, BAHARUDIN B., N. ZAMIN, *Building Transfer Rules using Annotated Disjunct: An Approach for Machine Translation*, The 8 th Student Conference on Research and Development, December 2007, Malaysia.
- [4] AMTRUP. J.W., MERGERDOOMIAN K, ZAJAC R. *Rapid Development of Translation Tool*. Proceedings of Machine Translation Summit, 1999, p 385-389
- [5] DIỆP QUANG BAN. *Ngữ pháp Tiếng Việt*. NXB Giáo dục, 1989
- [6] D. BÉCHET. *k-Valued Link Grammars are Learnable from Strings*. Proceedings of Formal Grammar, Vienna, Austria, 2003, pp. 9-18.
- [7] ĐINH ĐIỀN. *So-sánh trật-tự từ của định-ngữ giữa Tiếng Anh và Tiếng Việt*. Tập san Khoa học xã hội và nhân văn- Đại học KHXHNV Thành phố Hồ Chí Minh, 2011, tr 69-80
- [8] ĐINH ĐIỀN. *Dịch tự động Anh Việt dựa trên việc học luật chuyển đổi từ ngữ liệu song ngữ*. Luận án Tiến sĩ, ĐHKHTN- ĐHQG TP Hồ Chí Minh, 2002.
- [9] LÊ KHÁNH HÙNG. *Nghiên cứu xây dựng thử nghiệm phần mềm dịch tự động Việt-Anh*. Báo cáo tổng kết đề tài nhánh cấp nhà nước KC-01-03, 2004.
- [10] NGUYEN THI THU HUONG, PHAM NGUYEN QUANG ANH *A link Grammar for Vietnamese*. Journal on Information and Communication Technology, 8/2011 p 27-38.
- [11] IYER R., OSTENDORF. *Modelling Long Distance Dependency in Language: topic mixture and Dynamic Case Model*, Speech and Audio Processing, IEEE Transactions on, Jan.1999, p30-39.
- [12] KADAMBINI K, RAMA SREE R.J., RAMA KRISHNAMACHARYULU K.V. *An English-Sanskrit Machine Translation Using Link Parser*. In Proceedings of National Seminar, Tirupati, 2008.

- [13] N. L. MINH, N. P. THAI, L. A. CUONG, N. V. VINH, P.H. NGUYEN, H. S. DAM. *LVT: An English-Vietnamese Machine Translation System*. Hội nghị quốc gia FAIR'03, Hanoi, 10.2003.
- [14] D. MOLLA, B. HUTCHINSON. *Intrinsic versus Extrinsic Evaluations of Parsing Systems*. Proceedings of EACL Workshop on Evaluation Initiatives in Natural Language Processing, 2003, p 15-21.
- [15] NGUYỄN THIỆN NAM. *Tiếng Việt nâng cao cho người nước ngoài*. NXB Giáo dục, 1998
- [16] L. H. PHUONG, N. T. M. HUYEN, A. Roussanaly, H. T. Vinh *A hybrid approach to word segmentation of Vietnamese texts*. Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, LATA 2008, Springer LNCS 5196
- [17] KISHORE PAPINENI, SALIM ROUKOS, TODD WARD, AND WEI-JING ZHU. *BLEU: a Method for Automatic Evaluation of Machine Translation*. Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, July 2002, p. 311-318.
- [18] STANKEVICH N.V. *Các loại hình ngôn ngữ*. NXB Đại học và Trung học chuyên nghiệp, Hà Nội, 1982.
- [19] DANIEL K. SLEATOR, DAVY TEMPERLEY. *Parsing English with Link Grammar*. CMU-CS-91-96, October 1991.
- [20] NGUYEN PHUONG THAI, AKIRA 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, p 138-147, Cambridge, 8/ 2006.
- [21] TRAN NGOC TUAN, PHAN THI TUOI. *Syntax-based SMT Model in Adaption to Vietnamese-English Translation*. Poster of the 4th Conference RIVF, 2006.
- [22] <http://www.fen.bilkent.edu.tr/~aykutlu/ceviri>
- [23] <http://statmt.ru>
- [24] <http://www.link.cs.cmu.edu/link/trans-explanation.html>
<http://corpus.byu.edu/coca/>

SƠ LƯỢC VỀ TÁC GIẢ



NGUYỄN THỊ THU HƯƠNG

Sinh năm 1964.

Tốt nghiệp Đại học Bách khoa Hà Nội năm 1986. Bảo vệ luận án Thạc sỹ ngành CNTT tại Đại học Bách khoa Hà Nội năm 1998.

Hiện là giảng viên bộ môn Khoa học Máy tính, Viện Công nghệ Thông tin và truyền thông, Đại học Bách khoa Hà Nội.

Lĩnh vực nghiên cứu: Xử lý ngôn ngữ tự nhiên, Lý thuyết ngôn ngữ và phương pháp dịch.

Email: huongnt@soict.hut.edu.vn



LÊ NGỌC MINH

Sinh năm 1989.

Tốt nghiệp ngành Khoa học Máy tính, Đại học Bách khoa Hà Nội năm 2012.

Hiện đang học Cao học tại khoa Khoa học Nhận thức, Đại học Trento, Trento, Italia.

Lĩnh vực nghiên cứu: Dịch máy, Ngữ nghĩa phân phối, Ngôn ngữ học nhận thức.

Email: ngocminh.oss@gmail.com

Nhận bài ngày: 31/07/2012