Xử lý câu hỏi chính phản trong dịch tự động Hoa-Việt

Processing Yes - No Questions in the Chinese - Vietnamese Machine Translation

Trần Thanh Phước, Đinh Điền

Abstract: *Machine translation is a tradtional and difficult problem, therefore, improving the quality of machine translation is a necessary issue. In this paper, we present a method to improve the quality of Chinese – Vietnamese machine translation. When translating from Chinese into Vietnamese, one of the most frequently encountered errors is the wrong translation "yes - no question", so, we will apply some rules to convert "yes-no question" in the Chinese to common question. We have experimented this problem on the 10,000 pairs of Chinese-Vietnamese parallel corpus and 77,000-entry Chinese Vietnamese Dictionary and we reach the result of 47.23 BLEU score in comparison of 14.15 BLEU of Google Translator.*

I. GIỚI THIỆU

Dịch máy là bài toán lâu đời và khó nhưng đã thu hút được sự quan tâm của rất nhiều nhóm nghiên cứu trong nước và trên thế giới do tiềm năng ứng dụng của nó. Tuy nhiên, do bản chất nhập nhằng của ngôn ngữ ở mọi cấp độ và mọi khía cạnh cộng với sự khác biệt giữa các ngôn ngữ khiến bài toán chuyển ngữ lại càng phức tạp hơn. Đặc biệt đối với ngôn ngữ có sự khác biệt về trật tự từ như giữa tiếng Hoa và tiếng Việt.

Đối với dịch tự động có liên quan tiếng Việt, chủ yếu là dịch tự động Anh - Việt, hiện nay đã có một số nhóm nghiên cứu khác nhau với các cách tiếp cận khác nhau, cụ thể là:

- Nhóm Lê Khánh Hùng đã phát triển hệ dịch Anh-Việt từ đầu những năm 1990 và đã có bản thương mại tên là EVTran. Nhóm này chủ yếu sử dụng phương pháp dịch dựa trên luật (RBMT: Rule-based Machine Translation).
- Nhóm nghiên cứu của Đinh Điền (Trường Đại Học Khoa Học Tự Nhiên TP.HCM): khởi đầu từ năm 1995 là theo cách tiếp cận dựa trên luật nhưng từ năm 2005 họ đã phát triển thêm cách tiếp cận dựa trên thống kê (SMT: Statistical Machine Translation) và có sự kết hợp tri thức ngôn ngữ để nâng cao chất lượng.
- Nhóm nghiên cứu của Phan Thị Tươi (Trường ĐạiHọc Bách Khoa TP.HCM): sử dụng phương pháp phân tích cú pháp có xác suất để dịch văn bản Anh-Việt và Việt-Anh[6].
- Cty Lạc Việt đã phát triển công cụ dịch từ Anh sang Việt có thêm phần dịch chuyên ngành: tin học, toán học, y học và kế toán, kỹ thuật[7].
- Gần đây nhất là Google Translator (www. translate. google.com): sử dụng phương pháp dịch máy thống kê dựa trên kho ngữ liệu song ngữ khổng lồ khai thác từ Internet. Google Translator đã hỗ trợ hơn 50 ngôn ngữ bao gồm cả tiếng Việt. Tốc độ dịch nhanh và có tính năng tương tác với người dùng nhằm tăng chất lượng dịch cho các lần sau. Đối với các bản tin thông thường, chất lượng dịch Anh-Việt, Việt-Anh của Google Translator được xem là tốt nhất hiện nay.

Về dịch tự động, Hoa-Việt hay Việt-Hoa đến nay chưa có nhóm nào công bố chính thức ngoại trừ

Google Translator đã có thực hiện việc dịch Hoa-Việt, Việt-Hoa thông qua ngôn ngữ trung gian là tiếng Anh. Ví dụ: để dịch một văn bản từ tiếng Hoa sang tiếng Việt, trước tiên hệ thống sẽ dịch từ tiếng Hoa sang tiếng Anh, rồi sau đó sẽ dịch tiếp văn bản kết quả tiếng Anh đó sang tiếng Việt. Chính vì phải dịch theo kiểu bắc cầu như vậy đã khiến chất lượng dịch chung cuộc bị giảm đáng kể.

Nhìn chung, do tiếng Hoa và tiếng Việt cùng loại hình đơn lập (isolated) nên giữa hai ngôn ngữ có nhiều điểm tương đồng về văn phạm, nhưng cũng vẫn có một số điểm dị biệt về trật tự từ, hư từ và ở một số cấu trúc đặc biệt. Một trong những cấu trúc đặc biệt này là "câu hỏi chính phản". Kết quả dịch tự động cho các cấu trúc đặc biệt này thường bị sai lệch về ngữ nghĩa, về trật tự từ và tính tự nhiên của bản dịch trong ngôn ngữ đích. Trong nghiên cứu này, chúng tôi sẽ dịch tự động trực tiếp từ tiếng Hoa sang tiếng Việt theo phương pháp dịch máy thống kê (SMT) kết hợp với một số luật chuyển đổi trên các cấu trúc của câu hỏi chính phản. Cụ thể, trong bài báo này chúng tôi sẽ trình bày các nội dung như sau:

- Tổng quan về dịch tự động Hoa - Việt
- Cấu trúc câu hỏi chính phản trong tiếng Hoa
- Chuyển đổi câu hỏi chính phản trong tiếng Hoa
- Thực nghiệm, kết quả và kết luận.

II. TỔNG QUAN VỀ DỊCH TỰ ĐỘNG HOA – VIỆT

Đến nay, chúng ta thấy có hai hướng tiếp cận chính trong dịch tự động, đó là dịch dựa trên luật (RBMT) và dịch dựa trên ngữ liệu (CBMT: Corpus-Based Machine Translation). Theo hướng RBMT đòi hỏi hệ thống phải có được tập luật phân tích, chuyển đổi của ngôn ngữ nguồn, tập luật tạo sinh ngôn ngữ đích cùng với kho tài nguyên ngôn ngữ cần thiết. Để xây dựng được tập luật hay kho tài nguyên này đòi hỏi thời gian và công sức của nhiều chuyên gia ngôn ngữ và tin học cũng như các kết quả hình thức hóa của ngôn ngữ tự nhiên ở mức độ cao. Vì vậy, đối với việc dịch tự động Hoa-Việt hay Việt-Hoa theo hướng tiếp cận này trong giai đoạn hiện nay là chưa khả thi. May mắn là cách tiếp cận CBMT mà cụ thể ở đây là SMT cho phép chúng ta khắc phục được khó khăn nói trên. Theo cách tiếp cận SMT, chúng ta chủ yếu cần xây dựng kho ngữ liệu song ngữ Hoa-Việt với số lượng lớn và chất lượng tốt. Khác với các cặp ngôn ngữ phổ biến khác (như: Anh-Pháp, Anh-Hoa, ,...), một kho ngữ liệu song ngữ Hoa-Việt lớn và tốt như thế hiện chưa có sẵn, do đó, trong nghiên cứu này, chúng tôi cũng phải tự xây dựng kho ngữ liệu Hoa-Việt. Bên cạnh việc phụ thuộc vào ngữ liệu song ngữ dùng để huấn luyện, chất lượng dịch của SMT cũng phụ thuộc nhiều vào đặc điểm của cặp ngôn ngữ cần dịch. Nếu giữa hai ngôn ngữ càng có nhiều điểm tương đồng về từ vựng (về hình thái, về sự từ vựng hóa), về trật tự từ,... thì chất lượng dịch sẽ càng cao. Tiếng Hoa và tiếng Việt cùng thuộc loại hình ngôn ngữ đơn lập, nên giữa hai ngôn ngữ có nhiều điểm tương đồng về mặt hình thái của từ, sự từ vựng hóa (65% vốn từ trong tiếng Việt vay mượn từ tiếng Hoa). Tuy nhiên, trong một số cấu trúc đặc biệt, giữa hai ngôn ngữ Hoa và Việt có sự khác biệt, như: trật tự của giới ngữ, trạng ngữ chỉ thời gian, câu hỏi chính phản (yes/no question),... Trong các trường hợp đặc biệt này, cách tiếp cận SMT thường cho kết quả dịch còn rất thấp do kết quả dóng hàng từ (word alignment) bị sai lệch khá nhiều. Chính vì vậy, với những trường hợp này, chúng tôi đã đưa vào một số luật chuyển đổi của cách tiếp cận RMBT để chuyển đổi những cấu trúc đặc biệt đó sang dạng thông thường nhằm nâng cao chất lượng dóng hàng từ trong SMT và từ đó kết quả dịch chung cuộc của cả hệ thống đã được cải thiện đáng kể.

II.1. Đặc điểm về ngôn ngữ trong dịch Hoa-Việt

Trong phần này, chúng tôi sẽ khảo sát một số đặc điểm trong quá trình dịch từ tiếng Hoa sang tiếng Việt:

Về mặt loại hình ngôn ngữ (language typology), tiếng Hoa và tiếng Việt cùng thuộc loại hình đơn lập, nên có những điểm tương đồng như:

+ Trong hoạt động ngôn ngữ, từ không biến hình:

Tiếng Việt	Tôi hỏi anh ấy	Anh ấy hỏi tôi
Tiếng Hoa	我□ 他	他□ 我
Phiên âm Pinyin	wǒ wèn tā	tā wèn wǒ
Tiếng Anh	I ask him	He asks me

+ Phương thức ngữ pháp chủ yếu là trật tự từ và từ hư:

Tiếng Việt	cửa trước	trước cửa
Tiếng Hoa	前□	□ 前
Phiên âm Pinyin	qián mēn	mēn qián
Tiếng Anh	front door	at somebody's door

+ Ranh giới từ không được xác định bằng khoảng trắng:

Tiếng Việt	Học sinh học sinh học
Tiếng Hoa	学生学生物
Phiên âm Pinyin	xué shēng xué shēng wù
Tiếng Anh	pupils learn biology

+ Có từ chỉ loại trước danh từ:

Tiếng Việt	Tôi mua một cuốn sách
Tiếng Hoa	我□ 一本□
Phiên âm Pinyin	wǒ mǎi yī běn shū
Tiếng Anh	I buy a book

+ Dùng hư từ để thể hiện thể phủ định, nghi vấn:

Thể loại / **Ngôn ngữ**	**Khẳng định**	**Phủ định**	**Nghi vấn**
Tiếng Việt	Anh ấy ăn cơm	Anh ấy không ăn cơm	Anh ấy ăn cơm không ?
Tiếng Hoa	他吃□	他不吃□	他吃□ □ ?
Phiên âm pinyin	tā chī fàn	Tā bú chī fàn	Tā chī fàn mā ?
Tiếng Anh	He eats rice	He doesn't eat rice	Does he eat rice ?

Tuy nhiên, trong một số cấu trúc, giữa, tiếng Hoa và tiếng Việt có một số khác biệt như:

+ Trật tự giới ngữ:

Tiếng Việt	Chúng tôi học tiếng Hoa **tại trường đại học**
Tiếng Hoa	我□ 在大学学□ □ □
Phiên âm pinyin	wō mēn zāi dà xué xué xí hàn yǔ
Tiếng Anh	We learn Chinese at the university

+ Trạng ngữ chỉ thời gian:

Tiếng Việt	Tối mai chúng ta đi xem phim
Tiếng Hoa	明天 □上 我□ 去 看 □ 影
Phiên âm pinyin	míng tiān wǎn shàng wǒ mēn qù kàn diàn yǐng
Tiếng Anh	We'll go to the cinema on tomorrow evening

+ Cấu trúc cố định 之前，之后，以前，以后:

Tiếng Việt	phía sau thư viện là nhà ăn
Tiếng Hoa	□ □ □ 之后 是 食堂
Phiên âm pinyin	tú shū guàn zhī hòu shi shí tang
Tiếng Anh	behind the library is the dining-room

+ Câu hỏi chính phản:

Tiếng Việt	Anh có đi không ?
Tiếng Hoa	你去不去？ Hư từ
Phiên âm pinyin	nǐ qù bú qù?
Tiếng Anh	Do you go ?

Đây là cấu trúc chính mà chúng tôi muốn xử lý trong bài báo này. Chi tiết về cấu trúc này sẽ được trình bày trong phần III.

II.2. Đặc điểm về dịch thống kê Hoa-Việt

Hệ thống dịch máy thống kê cần kho ngữ liệu gồm hai tập tin song ngữ Hoa - Việt như sau:

- Tập tin tiếng Hoa: Tách các từ chính tả tiếng Hoa (kể cả các dấu) với nhau bởi một khoảng trắng.
- Tập tin tiếng Việt: Tách các từ chính tả tiếng Việt với các dấu (bao gồm các dấu , . ! ? “” :)
- Chia mỗi tập tin Hoa – Việt thành ba tập tin riêng lẻ theo tỉ lệ như sau: 90% dành cho huấn luyện (train), 5% dành cho điều chỉnh tham số (dev) và 5% dành cho thử nghiệm dịch (test). Lưu ý: Các câu ở mỗi tập tin Hoa – Việt phải tương ứng 1-1 với nhau.
- Sử dụng công cụ Moses [5] để huấn luyện và dịch với các tham số mặc định. Đây là công cụ dịch thống kê mã nguồn mở phổ biến nhất hiện nay.

Dóng hàng từ Hoa-Việt thường cho kết quả là một từ tiếng Hoa có nghĩa tương ứng với một từ tiếng Việt. Ví dụ trong kho ngữ liệu huấn luyện có các kết quả dóng hàng từ như sau:

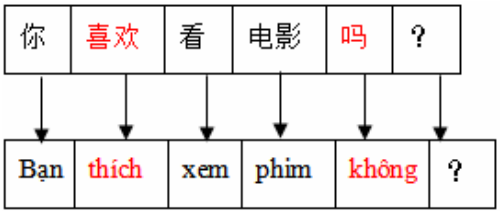
你↔ bạn; 喜□ ↔ thích; 不↔ không; 看↔ xem; □ 影↔ phim; ？↔ ?

Do đó, nếu câu tiếng Hoa không chứa các cấu trúc ngữ pháp đặc biệt thì khi dịch thống kê Hoa-Việt sẽ cho kết quả dịch với chất lượng rất tốt. Tuy nhiên, với các cấu trúc đặc biệt của tiếng Hoa (bao gồm các từ kết hợp lại) thì một số từ trong các cấu trúc đặc biệt này không còn giữ nguyên các ý nghĩa ban đầu của chúng. Trong câu hỏi chính phản, một số từ sẽ không được dịch sang tiếng Việt và câu tiếng Việt thường thêm từ để hỏi “không” ở cuối câu. Ví dụ, ta có câu hỏi chính phản như sau:

你	喜□	不	喜□	看	□ 影	？

Từ “喜□ ” có nghĩa “thích” và từ “不” có nghĩa “không”. Kết quả dịch thống kê thường cho kết quả là “Bạn thích không thích xem phim?”. Trong trường hợp này, từ “不喜□ ” (“không thích”) sẽ không được dịch sang tiếng Việt, nó đóng vai trò là hư từ để tạo nên cấu trúc câu hỏi chính phản, cuối câu hỏi chính phản khi dịch sang tiếng Việt phải có từ “không”. Dịch đúng thì câu tiếng Việt sẽ là “Bạn thích xem phim không?”.

Để có kết quả dịch đúng đối với câu hỏi chính phản này, điều chúng ta cần là xóa từ thừa (là các từ không được dịch sang tiếng Việt) của cấu trúc và thêm từ để hỏi ở cuối câu. Câu hỏi sau khi chuyển đổi đối với ví dụ này sẽ là: 你喜□ 看□ 影□ ？ Với sự thay đổi này, kết quả dịch tương ứng sẽ chính xác hơn.



II.3. Xây dựng kho ngữ liệu

Để phục vụ nghiên cứu này, chúng tôi đã xây dựng 2 kho ngữ liệu: một kho ngữ liệu đơn ngữ gồm 77.000 mục từ (từ điển tiếng Hoa); một kho ngữ liệu song ngữ Hoa – Việt gồm 10.000 cặp câu.

Từ điển đơn ngữ tiếng Hoa 77.000 mục từ được lấy từ trang web [10] gồm 64.000 mục từ và phần còn lại được lấy từ sách [3]. Từ điển này được sắp thứ tự tăng dần, làm ngữ liệu nguồn cho thuật toán Maximum Matching [4] để tách từ tiếng Hoa trong xử lý câu hỏi chính phản.

Kho ngữ liệu song ngữ 10.000 cặp câu được chúng tôi tổng hợp bằng tay từ các sách đàm thoại tiếng Hoa [3], [8] và các diễn đàn tiếng Hoa online, chủ yếu từ trang web [9]. Do ngữ liệu được nhập từ sách, diễn đàn và được định dạng bằng tay nên chất lượng của kho ngữ liệu này rất tốt.

Văn bản trong kho ngữ liệu chủ yếu là văn bản giao tiếp phổ thông, rất thường gặp trong giao tiếp tiếng Hoa, chiều dài của các câu tương đối ngắn, bình quân khoảng 10 từ trong một câu. Vì kho ngữ liệu tập trung phổ biến các câu giao tiếp nên số lượng câu hỏi xuất hiện khá nhiều. Chi tiết về kho ngữ liệu song ngữ được trình bày ở Bảng 1.

Bảng 1. Đặc điểm các câu trong kho ngữ liệu song ngữ

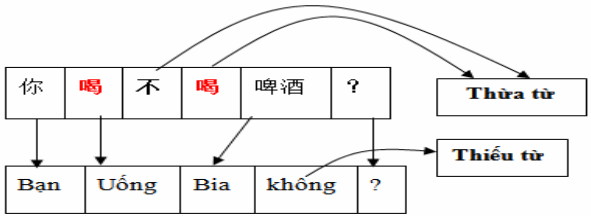
Tiêu đề	Số lượng	%
Sách [3]	4.000	40% tổng số câu
Sách [8]	3.000	30% tổng số câu
Trang web [9]	3.000	30% tổng số câu
Số từ trung bình mỗi câu	10 từ	
Câu hỏi	2.929	29% tổng số câu
Câu hỏi chính phản	212	2,2% tổng số câu 7,3% tổng số câu hỏi

III. CẤU TRÚC CÂU HỎI CHÍNH PHẢN TRONG TIẾNG HOA

Câu hỏi chính phản: Là dạng câu hỏi sử dụng thể khẳng định và phủ định của động từ hoặc hình dung từ (tính từ) để hỏi. Thể phủ định của câu hỏi chính phản là: **不/没 + động từ / hình dung từ**. Có các hình thức như sau:

- **Hình thứ 1:** Liên kết thể khẳng định và phủ định của động từ hoặc cụm tính từ, có 2 dạng:
  + **Dạng 1:** Thể khẳng định của động từ/ tính từ + thể phủ định của động từ / tính từ + tân ngữ (hình 1)

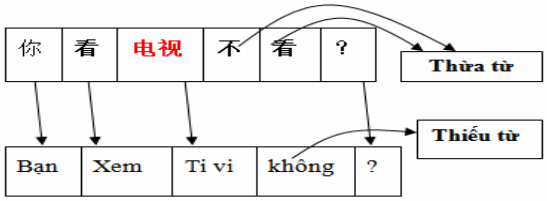
    Tân ngữ là từ đứng sau động từ và bổ nghĩa cho động từ, tân ngữ thường là danh từ hoặc cụm danh từ. Ví dụ trong câu sau thì từ "喝" là động từ ("uống") và từ "啤酒" là tân ngữ ("bia"):



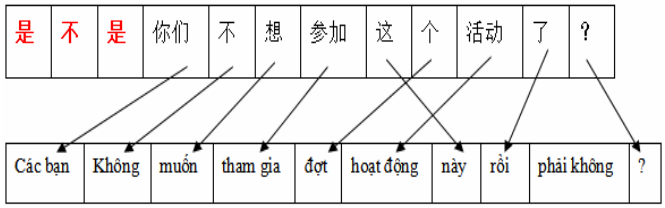
Hình 1. Câu hỏi chính phản hình thức 1-dạng 1

  + **Dạng 2:** Thể khẳng định của động từ/ tính từ + tân ngữ + thể phủ định của động từ / tính từ + tân ngữ (Hình 2)

Trong ví dụ này, từ "看" có nghĩa là "xem" (động từ), từ "□ 影" có nghĩa "ti vi" (tân ngữ).



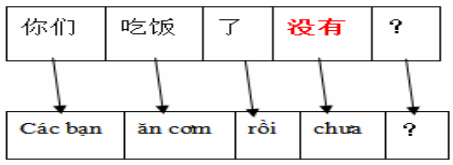
Hình 2. Câu hỏi chính phản hình thức 2-dạng 2



Hình 3. Câu hỏi chính phản hình thức 2

- **Hình thức 2:** Nếu người nói đã có một sự đánh giá hay nhận xét về một sự việc, hiện tượng nào đó; muốn xác minh rõ ràng hơn nữa thì dùng "是不是" để nêu câu hỏi. Trong tiếng việt, "是不是" có nghĩa là "phải không"; nhưng nghĩa của từng từ sẽ là: "是" nghĩa là "là"; "不" nghĩa là "không" (Hình 3)
- **Hình thức 3:** Dùng thể khẳng định của động từ / tính từ + **没有**

Trong câu hỏi, từ "没有" dịch sang tiếng việt có nghĩa là "chưa". Ở thể khẳng định, "没有" có nghĩa là "không có" (Hình 4)



Hình 4. Câu hỏi chính phản hình thức 3

Trong ba hình thức của câu hỏi chính phản, hình thức 3 đơn giản và cố định, chỉ có từ "没有?". Do đó, trong bài báo này chúng tôi sẽ không xử lý hình thức 3 vì nó đơn giản và hệ thống dịch SMT[5] sẽ tự học được. Chúng tôi sẽ chỉ tập trung vào 2 dạng của hình thức 1 và hình thức 2.

IV. CHUYỂN ĐỔI CÂU HỎI CHÍNH PHẢN SANG CÂU HỎI THÔNG THƯỜNG

Như ở phần II.2 (Đặc điểm dịch thống kê Hoa-Việt) chúng tôi đã trình bày về lỗi sai khi dịch câu hỏi chính phản trong dịch thống kê Hoa-Việt. Do đó, đối với câu hỏi chính phản, trước khi cho qua hệ thống dịch, chúng tôi tiến hành chuyển đổi chúng sang câu hỏi thông thường. Câu hỏi thông thường là câu hỏi sau khi đã xóa các từ thừa và bổ sung từ thiếu của câu hỏi chính phản. Tương ứng với câu hỏi thông thường, dựa vào kết quả dóng hàng từ trong kho ngữ liệu huấn luyện, câu dịch đích sẽ cho kết quả chính xác hơn.

Ví dụ: Câu hỏi chính phản: 你喜□ 不喜□ 看□ 影？Sau khi xóa từ thừa "不喜□ " và bổ sung từ để hỏi "□ " ở cuối câu, ta được câu hỏi thông thường "你喜□ 看□ 影□ ？". Kết quả dịch thống kê đối với ví dụ này cho kết quả chính xác như người dịch.

Câu hỏi chính phản và kết quả dịch:

你	喜□	不	喜□	看	□ 影	?
Bạn	thích	Không	thích	Xem	phim	?

Sau khi chuyển đổi sang câu hỏi thông thường và kết quả dịch tương ứng:

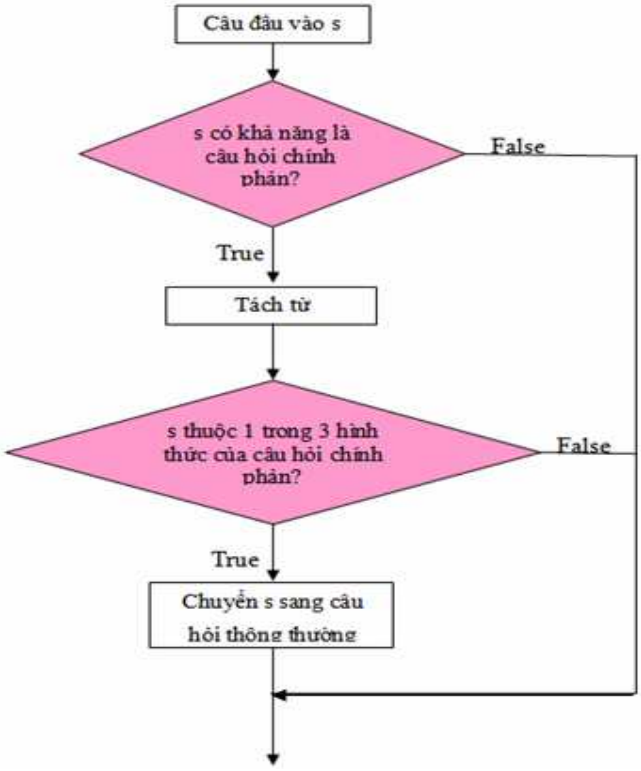
你	喜□	看	□ 影	□	?
Bạn	thích	Xem	phim	Không	?

- Phương pháp chuyển đổi (Hình 5)

Phần này trình bày mô hình tổng quát quá trình chuyển đổi câu tiếng Hoa trước khi cho qua hệ thống dịch thống kê Hoa-Việt.

Kiểm tra câu tiếng Hoa có khả năng là câu hỏi chính phản hay không?

Tìm trong câu đầu vào có chứa từ "不" hoặc "没" hay không. Nếu có và 2 từ này không phải nằm ở đầu hoặc cuối câu thì có khả năng câu này là câu hỏi chính phản.



Hình 5. Chuyển đổi câu hỏi chính phản sang câu hỏi thông thường

Tách từ câu hỏi chính phản bằng phương pháp Maximum Matching (MM)[4]

Có hai hướng tách từ phổ biết hiện nay, đó là tách từ dựa vào từ điển và tách từ dựa vào thống kê (điển hình là phương pháp tách từ CRF của nhóm tác giả PiChuan Chang, Michel Galley, Christorpher D.Manning của trường đại học Stanford). Trong nghiên cứu này, chúng tôi chỉ tách từ các câu có khả năng là câu hỏi chính phản (các câu hỏi có chứa từ "不" hoặc "没") bằng phương pháp Maximum Matching. Với từ điển đơn ngữ tiếng Hoa gồm 77.000 mục từ được định dạng bởi các chuyên gia Hoa ngữ của công ty Kim Từ Điển nên chất lượng của từ điển này rất tốt. Chất lượng của phương pháp tách từ MM

phụ thuộc vào chất lượng của từ điển nên trong nghiên cứu này chúng tôi sử dụng phương pháp truyền thống này trên từ điển sẵn có để tách từ. Qua khảo sát, với số lượng câu hỏi chính phản là 212 câu, kết quả tách từ theo thuật toán MM trên từ điển 77.000 mục từ cho kết quả chính xác 100%.

Phương pháp Maximum Matching: Còn gọi là LRMM (Left Right Maximum Matching). Ở phương pháp này , chúng ta sẽ duyệt một ngữ hoặc câu từ trái sang phải và chọn từ có nhiều âm tiết nhất có mặt trong từ điển và cứ thực hiện lặp lại như vậy cho đến hết câu.

Giả sử chúng ta có một chuỗi ký tự C_1 , C_2 , …, C_n. Chúng ta sẽ áp dụng phương pháp từ đầu chuỗi. Đầu tiên kiểm tra xem C_1 có phải là từ hay không, sau đó kiểm tra xem C_1C_2 có phải là từ hay không. Tiếp tục thực hiện như thế cho đến khi tìm được từ dài nhất.

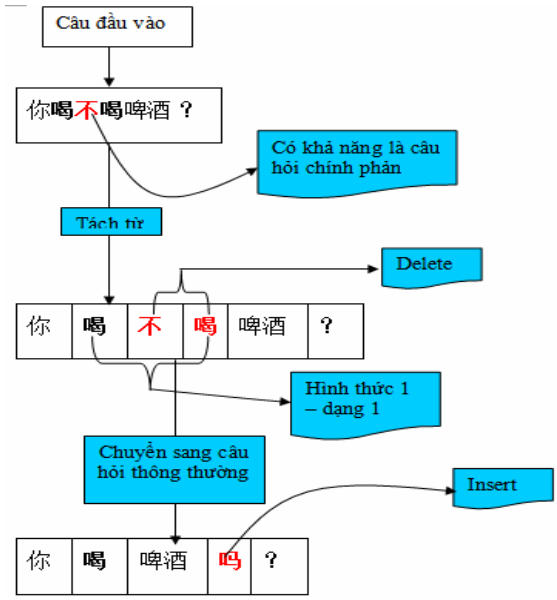
Phương pháp này thực hiện tách từ đơn giản, nhanh và chỉ cần dựa vào từ điển để thực hiện. Tuy nhiên, khuyết điểm của phương pháp này cũng chính là từ điển, nghĩa là độ chính xác khi thực hiện tách từ phụ thuộc hoàn toàn vào tính đủ, tính chính xác của từ điển.

Kiểm tra câu hỏi chính phản ở dạng nào – Chuyển sang câu hỏi thông thường.

Hình thức 1 - Dạng 1:

Quy tắc chuyển đổi cho dạng này như sau:

- Xác định câu hỏi có chứa từ "不" hoặc "没" hay không
  + Nếu có,
    - Tách từ câu hỏi này.
    - Xác định từ liền trước và từ liền sau từ "不/没" (*trong tiếng Hoa từ đứng liền sau từ "不/没" phải là động từ hoặc tính từ).* Nếu hai từ này giống nhau thì câu này là câu hỏi chính phản dạng 1 hình thức 1, ta tiến hành chuyển đổi như sau:
      - Xóa từ "不/没" và động từ/tính từ phía sau nó.
      - Thêm từ "□ " ở cuối câu (trước dấu ?)
- Ví dụ câu hỏi chính phản "你喝不喝啤酒？" được chuyển đổi như sau: (Hình 6)

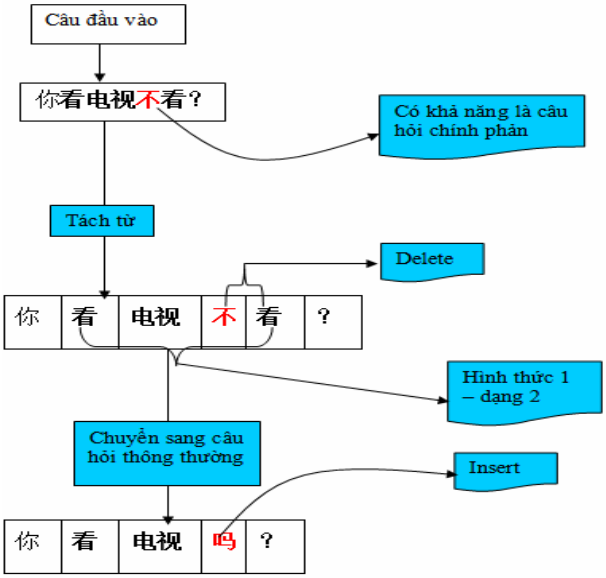


Hình 6. Chuyển đổi câu hỏi chính phản hình thức 1 – dạng 1 sang câu hỏi thông thường

Dạng 2: Mô hình chuyển đổi như sau

Quy tắc chuyển đổi cho dạng này như sau:

- Xác định câu hỏi có chứa từ "不" hoặc "没" hay không
  + Nếu có,
    - Tách từ câu hỏi này.
    - Xác định từ liền sau từ "不/没", *tìm từ trong ngữ phía trước từ "不/没"* . Nếu hai từ này giống nhau thì câu này là câu hỏi chính phản dạng 1 hình thức 1, ta tiến hành chuyển đổi như sau:
      - Xóa từ "不/没" và động từ/tính từ phía sau nó.
      - Thêm từ "□ " ở cuối câu (trước dấu ?)
- Ví dụ câu hỏi chính phản "你看□ 影不看？" được chuyển đổi như sau (Hình 7)



Hình 7. Chuyển đổi câu hỏi chính phản hình thức 1 – dạng 2 sang câu hỏi thông thường

Ở dạng 1 và dạng 2 ở hình thức 1, khi gặp một câu có khả năng là câu hỏi chính phản, để xác định câu này có thực sự là câu hỏi chính phản hay không chúng ta phải tách từ và kiểm định. Việc tách từ rất quan trọng, nếu bỏ qua bước này có khả năng chúng ta nhận diện bị sai. Ví dụ câu "学生不学□ ", nếu không tách từ thì câu này thõa mãn đặc trưng của dạng 2 hình thức 1 của câu hỏi chính phản. Trong câu này từ động từ "学" (học) đứng trước và sau từ "不" nên thuộc câu hỏi chính phản. Tuy nhiên, từ "学" trong trường hợp này không phải là từ độc lập, câu này sau khi tách từ có dạng như sau:

学生	不	学习
Học sinh	không	học bài

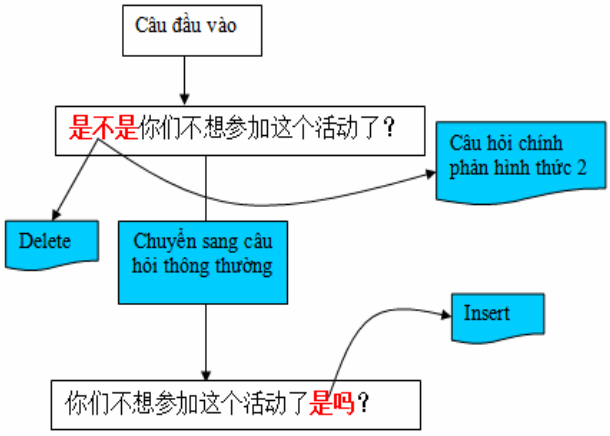
Rõ ràng sau khi tách từ, từ đứng trước và sau từ "不" không còn giống nhau và câu này không phải là câu hỏi chính phản. Qua ví dụ này cho thấy việc tách từ rất quan trọng và cần thiết trong bài toán nhận diện chính xác câu hỏi chính phản.

Hình thức 2: 是不是:

Đối với câu hỏi chứa cụm từ "是不是" chúng ta không cần tách từ câu hỏi này vì chắc chắn đây là câu hỏi chính phản thuộc hình thức 2, cụm từ này sẽ được dịch là "phải không".

Quy tắc chuyển đổi cho dạng này như sau:

- Xác định câu hỏi có chứa cụm từ "是不是"
  + Nếu có,
    - Xác định trong câu có chứa động từ hay không.
      - Nếu có: Xóa cụm từ "是不是" và thêm từ "□ " ở cuối câu.
      - Nếu không có: Xóa cụm từ "不是" và thêm từ "□ " ở cuối câu
- Ví dụ câu hỏi chính phản "是不是你□ 不想参加□ 个活□ 了？" được chuyển đổi như sau (Hình 8)
-



Hình 8. Chuyển đổi câu hỏi chính phản hình thức 2 sang câu hỏi thông thường

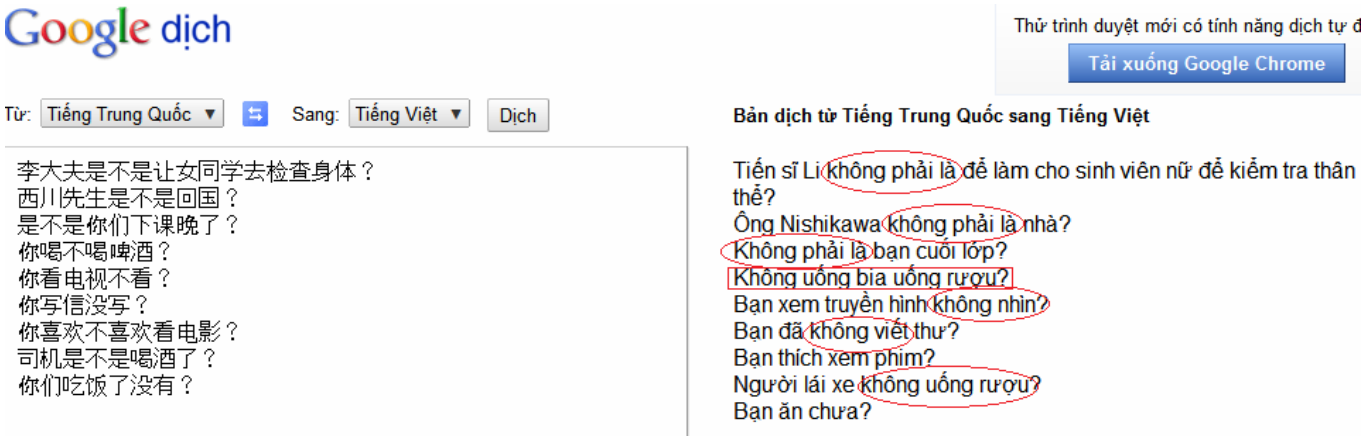
Kết quả thử nghiệm và thảo luận

Trong bài báo này chúng tôi thực hiện hai thử nghiệm sau:

- **Thử nghiệm 1:** Dùng 10.000 cặp câu song ngữ tự tạo: 9.000 dành cho huấn luyện (train), 500 dành cho kiểm tra (test) và 500 dành cho điều chỉnh tham số (dev).

Bảng 2. So sánh kết quả dịch qua ba hệ thống.

Stt	Câu hỏi chính phản	Dịch chuẩn	Dịch chưa xử lý chính phản	Dịch đã xử lý chính phản	Google dịch
1	李大夫是不是让女同学去检查身体？	Bác sĩ lý kiểm tra sức khỏe bạn học nữ phải không?	bác sĩ lý phải không phải bảo bạn học nữ kiểm tra sức khỏe ?	bác sĩ lý để bạn học nữ kiểm tra sức khỏe phải không ?	Tiến sĩ Li không phải là để làm cho sinh viên nữ để kiểm tra thân thể?
2	西川先生是不是回国？	Tây xuyên tiên sinh về nước phải không?	tây xuyên tiên sinh là không phải về nước ?	Tây xuyên tiên sinh về nước phải không ?	Ông Nishikawa không phải là nhà?
3	是不是你们下课晚了？	Các bạn tan học trễ phải không?	có phải là các bạn tan học tối rồi ?	các bạn tan học tối rồi phải không ?	Không phải là bạn cuối lớp?
4	你喝不喝啤酒？	Bạn uống bia không?	Bạn uống không uống bia ?	bạn uống bia không ?	Không uống bia uống rượu?
5	你看电视不看？	Bạn xem ti vi không?	Bạn xem xem ti vi không ?	bạn xem ti vi không ?	Bạn xem truyền hình không nhìn?
6	你写信没写？	Bạn viết thư chưa?	Bạn viết thư viết không ?	bạn viết thư chưa ?	Bạn đã không viết thư?
7	你喜欢不喜欢看电影？	Bạn thích xem phim không?	Bạn thích không thích xem phim ?	bạn thích xem phim không ?	Bạn thích xem phim?
8	司机是不是喝酒了？	Tài xế uống rượu rồi phải không?	tài xế đúng không uống rượu rồi ?	tài xế uống rượu rồi phải không ?	Người lái xe không uống rượu?
9	你们吃饭了没有？	Các bạn ăn cơm rồi chưa?	các bạn ăn cơm rồi chưa ?	các bạn ăn cơm rồi chưa ?	Bạn ăn chưa?



Hình 10. Hình chụp bản dịch google ngày 01/06/2011

+ Thử nghiệm dịch tự động SMT Baseline[5] (không xử lý câu hỏi chính phản).
+ Thử nghiệm dịch tự động SMT Baseline[5] đã qua tiền xử lý câu hỏi chính phản.
+ Trong 500 câu test có 14 câu hỏi chính phản. Điểm BLEU[5] Baseline: 46,96; Điểm BLEU có xử lý chính phản: 47,23

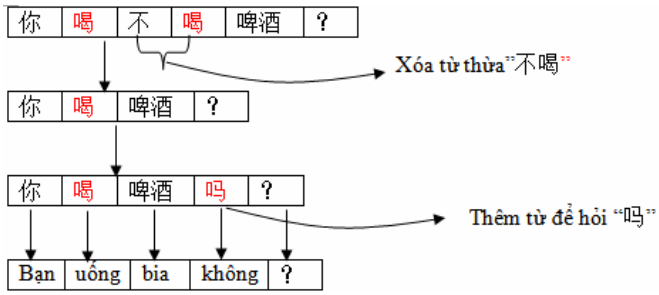
- **Thử nghiệm 2:** Sử dụng 4 câu hỏi chính phản ở bài 24 và bài ôn tập 5 trong quyển sách "301 Câu Đàm Thoại Tiếng Hoa" –Trương Văn Giới – Lê Khắc Kiều Lục dịch; Nhà xuất bản Khoa Học Xã Hội [3]. Sử dụng 5 câu hỏi chính phản ở bài 4 trong quyển sách "Giáo Trình Ngữ Pháp Tiếng Hoa" – tác giả Tô Cẩm Duy – Nhà xuất bản Trẻ [2]. Tổng

cộng 9 câu này được dịch qua ba hệ thống: Dịch Baseline dùng tool Moses[5], dịch đã xử lý chính phản (dùng tool Moses[5]) và dịch bởi Google Translator. Kết quả dịch thể hiện ở bảng 2 và bản dịch bởi google được thể hiện ở Hình 10.

Các phần *chữ màu nhạt* là phần dịch bị sai của hệ dịch cơ sở (dịch chưa xử lý câu hỏi chính phản) và câu dịch bởi Google. Lỗi chủ yếu khi dịch cấu trúc này: dịch thừa từ và thiếu từ. Ví dụ như ở câu số 4: "你喝不喝啤酒？" Trong câu này thì dịch cơ sở cho kết quả là "Bạn uống không uống bia?" và Google cho kết quả là "không uống bia uống rượu?", trong khi câu dịch đúng là "bạn uống bia không?". Ở đây, chúng tôi không đi sâu vào nguyên nhân dịch sai của hệ thống Google, vì hệ thống này phải trải qua hai lần dịch khi dịch câu Hoa–Việt; lỗi ở câu tiếng Việt bao gồm lỗi khi dịch từ Hoa sang Anh và từ Anh sang Việt. Trong phần này, chúng tôi chỉ bàn luận về nguyên nhân dịch sai của dịch cơ sở.

Trong ngữ liệu huấn luyện của hệ dịch cơ sở, các từ trong câu 4 sẽ được dịch tương ứng là: "你 (bạn) 喝 (uống) 不 (không) 喝 (uống) 啤酒 (bia)？". Do đó, hệ thống dịch cơ sở sẽ dịch câu này thành "bạn uống không uống bia?". Kết quả này không đúng với ý nghĩa của người dịch, gây ra hai lỗi: thừa và thiếu từ.

Theo như phân loại ở phần III, câu "你喝不喝啤酒？" thuộc câu hỏi chính phản dạng 1 hình thức 1 nên trước khi cho quả hệ thống dịch cơ sở, chúng tôi đã chuyển đổi câu hỏi chính phản này sang câu hỏi thông thường (Hình 9)



Hình 9. Xóa từ thừa và thêm từ để hỏi

Tương ứng với câu sau khi đã biến đổi (xóa và thêm từ), kết quả dịch có chất lượng cao hơn nhiều so với câu dịch cơ sở và Google, đúng với câu do con người dịch.

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong bài báo này, dựa trên kho ngữ liệu tự tạo, chúng tôi đã thực hiện được việc dịch tự động trực tiếp từ tiếng Hoa sang tiếng Việt mà không thông qua ngôn ngữ trung gian tiếng Anh. Bên cạnh đó, chúng tôi cũng đã khắc phục khá tốt các lỗi do câu hỏi chính phản ở ngôn ngữ tiếng Hoa gây ra và cho ra kết quả dịch gần như mong muốn.

Hướng sắp tới, chúng tôi sẽ tiếp tục khắc phục các lỗi về trật tự từ thường xảy ra ở các câu chứa các từ: "后" , "以后", "之后", từ "的"… Đồng thời nghiên cứu, thử nghiệm các phương pháp như: tách từ Hoa-Việt, gán nhãn từ loại hoa việt, phân loại văn bản,… nhằm tìm ra các phương pháp hiệu quả nhất để cải tiến dịch tự động Hoa-Việt cho ra kết quả tốt nhất.

LỜI CẢM ƠN

Đề tài được thực hiện dưới sự tài trợ của quỹ Nafosted và trường ĐH Khoa học Tự nhiên, thuộc ĐHQG Tp.HCM

TÀI LIỆU THAM KHẢO.

[1]. ĐINH ĐIỀN, *Xử lý ngôn ngữ tự nhiên*, NXB Đại Học Quốc Gia, 2006.

[2]. TÔ CẨM DUY, *Giáo trình Ngữ pháp tiếng Hoa*, Nhà Xuất Bản Trẻ, 2006.

[3]. TRƯƠNG VĂN GIỚI, LÊ KHẮC KIỀU LỤC biên dịch từ Học Viện Ngôn Ngữ Bắc Kinh, *301 Câu Đàm Thoại Tiếng Hoa tập 1 - 2*, Nhà Xuất Bản Khoa Học Xã Hội, 2009.

[4]. PAK-KWONG AND CHORKIN CHAN, *Chinese Word Segmentation based on Maximum Matching and Word Binding Force*, Department of Computer Science – The University of Hong Kong, .

[5]. PHILIPP KOEHN, *Statistical Machine Translation System – User Manual and Code Guide*, University of Edinburgh, September 29, 2010.

[6]. http:// www.jaist.ac.jp /~bao / VLSP-text/ March2008/ SP3_Baocaokythuat2008thang3.pdf, Báo cáo kỹ thuật, 2008.

[7]. Công ty Lạc Việt, Vietgle - Dịch tự động, http://www.vietgle.vn/tratu/dich-tu-dong.aspx.

[8]. HỮU LÝ, TUẤN ANH, *3000 câu đàm thoại tiếng Hoa*, NXB Giao thông vận tải, 2008.

[9]. http://www.dantiengtrung.com, tiếng Hoa online dtt.

[10]. http://www.tangthuvien.com, Tin học – công cụ dịch

Nhận bài ngày: 15/06/2011

SƠ LƯỢC VỀ TÁC GIẢ

TRẦN THANH PHƯỚC

Ngày sinh: 12-07-1981 tại Tiền Giang.

Tốt nghiệp Đại học Sư Phạm Tp. HCM năm 2006 chuyên ngành Tin học; cao học Tin học năm 2011 tại Trường Đại học Khoa Học Tự Nhiên Tp. HCM.

Tham gia giảng dạy tại khoa Công nghệ Thông tin Trường Đại học Công Nghiệp Thực Phẩm Tp. HCM.

Lĩnh vực nghiên cứu: Xử lý ngôn ngữ tự nhiên.

ĐT: 0908 350 132.

Email: phuoctt@cntp.edu.vn

ĐINH ĐIỀN

Sinh: 18-01-1966 tại Tp. HCM.

Tốt nghiệp Đại học ngành Vật lý Điện tử năm 1988 tại trường ĐH Tổng hợp Tp. HCM.; cao học Tin học năm 1996 tại trường ĐH KH Tự nhiên, Tp. HCM; Tiến sĩ Tin học năm 2003 tại trường ĐH KH Tự nhiên, Tp. HCM.; Tiến sĩ Ngôn ngữ học năm 2005 tại trường Đại học Khoa học Xã hội & Nhân văn, Tp. HCM.; Được phong Phó Giáo sư Tin học năm 2007.

Công tác giảng dạy và nghiên cứu về xử lý ngôn ngữ tự nhiên tại khoa Công nghệ Thông tin Trường Đại học Khoa học Tự nhiên Tp.HCM từ năm 1996 đến nay.

ĐT: 090 827 8207

Email: ddien@fit.hcmus.edu.vn

.