Khảo sát yếu tố ranh giới từ trong dịch thống kê Hoa-Việt

• **Trần Thanh Phước**
Trường Đại học Tôn Đức Thắng
• **Đinh Điền**
Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM
(Bài nhận ngày 04 tháng 03 năm 2015, nhận đăng ngày 12 tháng 06 năm 2015)

TÓM TẮT

Trong các ngôn ngữ đơn lập như tiếng Hoa và tiếng Việt, các từ không được phân biệt với nhau bởi khoảng trắng, một từ có thể bao gồm một hoặc nhiều từ chính tả. Việc có nên phân đoạn từ hay không trước khi cho qua hệ thống huấn luyện và dịch là vấn đề cần được xem xét. Trong bài báo này, chúng tôi sẽ tiến hành khảo sát ảnh hưởng của yếu tố ranh giới từ đến kết quả dịch thống kê Hoa-Việt. Kết quả thực nghiệm của bài báo sẽ làm cơ sở cho các hướng nghiên cứu cải tiến phân đoạn từ tiếp theo nhằm tăng hiệu suất dịch. Chúng tôi đã khảo sát trên hai trường hợp sau: không phân đoạn từ và phân đoạn từ trên kho ngữ liệu 8.000 và 12.000 cặp câu. Dựa trên kết quả thực nghiệm, chúng tôi nhận thấy rằng: ngữ liệu chưa phân đoạn từ hoặc được phân đoạn từ đều có những ưu và khuyết điểm riêng. Một hướng cải tiến mà bài báo đề xuất là tích hợp các ưu điểm của hai phương pháp này vào hệ thống dịch máy.

Từ khóa: *Dịch thống kê, ranh giới từ, phân đoạn từ, hình vị.*

GIỚI THIỆU

Không giống với các ngôn ngữ phương Tây, điển hình là tiếng Anh, các từ trong tiếng Hoa và tiếng Việt không được phân biệt bởi khoảng trắng. Một câu tiếng Hoa bao gồm một dãy các từ chính tả, kể cả dấu câu, nằm liên tiếp với nhau và không có khoảng trắng giữa các từ chính tả này. Trong tiếng Việt, các từ chính tả được phân cách với nhau bởi một khoảng trắng, các dấu câu nằm liền sau từ chính tả. Do đó, vấn đề phân đoạn từ luôn được giải quyết đầu tiên trong bài toán dịch máy từ tiếng Hoa, Việt sang ngôn ngữ khác (chủ yếu là tiếng Anh).

Một từ tiếng Hoa thường bao gồm nhiều hình vị có nghĩa, thường được chia thành ba trường hợp. Trường hợp một: nghĩa của các hình vị giống hoặc có liên quan đến nghĩa của từ chứa các hình vị đó; trường hợp hai: nghĩa của từng hình vị không liên quan gì đến nghĩa của từ chứa nó; trường hợp ba: nghĩa của các hình vị trong từ bị đảo trật tự. Thí dụ như ở Bảng 1.

Bảng 1. Nghĩa của hình vị và từ chứa hình vị ở tiếng Hoa

Từ	Âm Hán Việt	Nghĩa	Hình vị có nghĩa
教师	Giáo sư	Thầy dạy, giáo sư, thầy	教: dạy 师: thầy
花完	Hoa hoàn	Chi tiêu, xài hết	花: hoa 完: xong
放假	Phóng giả	Nghỉ phép	放: thả 假: giả
银行卡	Ngân hàng tạp	Thẻ ngân hàng	银行: ngân hàng 卡: thẻ

Từ 教师 bao gồm các hình vị có nghĩa tương đồng với nghĩa của chúng. Trong khi đó, các từ 花完, 放假 đã không còn giữ các nghĩa của các hình vị tạo nên chúng. Trường hợp còn lại, từ 银行卡 nếu không đảo trật tự thì nghĩa tiếng Việt tương ứng sẽ là "ngân hàng thẻ", trong khi đó nghĩa đúng sẽ là "thẻ ngân hàng", nghĩa của các hình vị trong từ chứa nó bị đảo trật tự. Ý nghĩa tương tự cho từ tiếng Việt; điển hình như từ "ba hoa" (nghĩa: nói nhiều, phóng đại quá sự thật) bao gồm hai hình vị: "ba" có nghĩa thông thường là "số 3, người sinh ra mình", "hoa" trong "bông hoa"; ý nghĩa hai hình vị này không liên quan gì đến từ chứa chúng.

Chính sự thay đổi về nghĩa, về trật tự của các hình vị khi chúng kết hợp với nhau để tạo nên từ ở cả tiếng Hoa và tiếng Việt nên khi dịch từ tiếng Hoa, tiếng Việt sang tiếng Anh hoặc ngược lại thì phân đoạn từ luôn là công việc được thực hiện đầu tiên (nhằm tạo tương ứng "từ" – "từ"). Nhưng liệu đối với cặp ngôn ngữ mà cả hai đều không xác định được ranh giới từ bởi khoảng trắng như tiếng Hoa và Việt thì biện pháp phân đoạn từ có cho kết quả tốt hơn? Bài báo này chúng tôi sẽ khảo sát sự ảnh hưởng của yếu tố ranh giới từ đến hiệu suất dịch thống kê Hoa-Việt. Từ cơ sở thực nghiệm, chúng tôi sẽ giới thiệu các phương pháp cải tiến phân đoạn từ nhằm nâng cao hiệu suất dịch cho cặp ngôn ngữ Hoa-Việt.

Bài báo này bao gồm các nội dung sau:

Công trình liên quan đến phân đoạn từ Hoa-Việt.

Các thử nghiệm về phân đoạn từ, không phân đoạn từ và các kết quả thực nghiệm tương ứng.

Một số thảo luận cũng như các đề xuất cải tiến phân đoạn từ.

Kết luận và hướng phát triển tiếp theo.

CÁC CÔNG TRÌNH LIÊN QUAN

Đối với các loại ngôn ngữ mà từ không được xác định bởi khoảng trắng như tiếng Hoa và Việt thì việc phân đoạn từ luôn được thực hiện đầu tiên trong các bài toán xử lý ngôn ngữ, điển hình như: nhận dạng chủ đề văn bản, dịch máy, … Có nhiều phương pháp phân đoạn từ, chủ yếu tập trung vào hai hướng: hướng dựa vào tri thức từ vựng và hướng dựa vào tri thức ngôn ngữ (phương pháp thống kê dựa vào ngữ liệu lớn). Phương pháp so khớp cực đại (Maximum Matching) dựa vào từ điển là một điển hình của phương pháp phân đoạn từ theo hướng tri thức từ vựng; phương pháp tách từ dựa vào thống kê phổ biến hiện nay cho cả tiếng Hoa và Việt là CRF (Conditional Random Fields) [1]. Với ưu thế về ngữ liệu có sẵn và không đòi hỏi người nghiên cứu phải hiểu sâu về tri thức ngôn ngữ nên các phương pháp phân đoạn từ theo hướng thống kê chiếm ưu thế hơn so với các phương pháp dựa vào từ vựng. Phân đoạn từ nhằm mục đích làm tăng hiệu suất dịch của dịch máy thống kê (statistic machine translation: SMT) là một trong những hướng nghiên cứu tiên tiến hiện nay. Từ được phân đoạn lúc này không hoàn toàn đúng với ý nghĩa của từ theo định nghĩa của các nhà ngôn ngữ, từ ở đây có thể là từ ngữ dụng lớn hơn ý nghĩa của từ ngôn ngữ hoặc từ chỉ là một hình vị có nghĩa nhỏ hơn từ ngôn ngữ. Ý nghĩa chung cuộc của việc điều chỉnh phân đoạn từ là làm sao trong ngữ liệu huấn luyện các từ được bao phủ càng nhiều càng tốt và kết quả gióng hàng càng có nhiều ánh xạ 1-1.

Để làm được điều này, đối với tiếng Hoa, theo tác giả Ruiqiang Zhang và các cộng sự [4] thì cần phải tích hợp nhiều phương pháp phân đoạn từ lại với nhau, bao gồm cả phân đoạn từ dựa vào từ điển và dựa vào thống kê. Một hướng cải tiến khác là phân đoạn từ dựa vào song ngữ của nhóm tác giả Yanju Ma và Andy Way [5]. Theo các tác giả, việc phân đoạn từ thông thường chỉ dựa vào ngữ liệu đơn ngữ mà không có bất kỳ sự quan tâm nào đến nguồn ngữ liệu song ngữ, việc phân đoạn từ ở ngôn ngữ nguồn không chắc tương ứng với phân đoạn từ ở ngôn ngữ đích, điều này dẫn đến kết quả gióng hàng từ bị sai.

Mặt khác, hầu hết các công cụ phân đoạn từ hiện nay đều được huấn luyện trên một ngữ liệu thuộc một lĩnh vực nào đó. Việc phân đoạn từ như vậy sẽ dẫn đến tính cục bộ về nghĩa của từ và chắc chắn sẽ cho kết quả không tốt khi áp dụng vào nhiều lĩnh vực khác.

Phân rã các từ tiếng Hoa thành các hình vị có nghĩa nhỏ hơn cũng là một phương pháp phổ biến hiện nay nhằm tăng hiệu suất dịch (Ming-Hong Bai, …, 2008)[6]. Nhóm tác giả đã sử dụng phương pháp học không giám sát nhằm phân rã các từ thành các hình vị nhỏ hơn với mục đích là tạo ra càng nhiều ánh xạ 1-1 càng tốt. Tiếng Hoa với các từ đa âm tiết bao gồm nhiều hơn một hình vị có nghĩa và được dịch sang nhiều từ tiếng Anh. Thí dụ: từ 教育署 được dịch sang tiếng Anh là "Department of Education" (Sở Giáo dục). Từ 署 có nghĩa thông thường là "Ban", "Khoa", "Sở" (Department), 教育 có nghĩa là "giáo dục". Việc phân đoạn từ như trên sẽ làm giảm tổng số lần đồng xuất hiện của cặp từ Hoa-Anh đồng thời tăng cường nhiều hơn gióng hàng n-1. Thí dụ: do 教育署 là một từ nên nó không đóng góp gì cho cặp 教育/Education và 署/Department. Mặt khác cặp từ "教育署/Education of Department" sẽ tạo ra gióng hàng n-1: 教育署 -> Educattion và 教育署-> Department. Do đó, nhóm tác giả đã đề xuất phân rã từ tiếng Hoa thành các hình vị có nghĩa nhỏ hơn.

Một trong những đặc điểm chung của các phương pháp cải tiến phân đoạn từ tiếng Hoa ở trên là việc cải tiến được áp dụng cho cặp ngôn ngữ Hoa-Anh hoặc Anh-Hoa. Tiếng Anh với khoảng trắng là dấu hiệu cho biết ranh giới từ, đây là điểm vô cùng quan trọng cho các phương pháp cải tiến. Điển hình như phương pháp phân rã từ thành các hình vị nhỏ hơn, không phải bất kỳ từ tiếng Hoa nào bao gồm nhiều hình vị có nghĩa cũng đều được phân rã, chỉ có những từ gióng hàng với nhiều từ tiếng Anh mới là ứng viên cần phân rã. Thí dụ: "washing machine/洗衣机" có thể được tách thành 洗衣 và 机 tương ứng với "washing" và "machine", nhưng "heater/暖气机" thì không được tách thành 暖气 và 机, vì từ này được gióng hàng với một từ tiếng Anh duy nhất "heater".

Cũng giống như tiếng Hoa, phân đoạn từ tiếng Việt cũng được các nhà nghiên cứu quan tâm và cài đặt thử nghiệm; hướng tiếp cận dựa vào thống kê cũng chiếm ưu thế hơn so với hướng dựa vào từ điển. Nhóm tác giả Cam-Tu Nguyen [8] cũng thực hiện phân đoạn từ tiếng Việt dựa vào CRF (như tiếng Hoa). Trong khi đó, nhóm VCL của Đinh Điền [8] thì phân đoạn từ theo phương pháp SVM và sau này là phân đoạn từ dựa vào ngữ dụng. Chúng tôi sử dụng phân đoạn từ tiếng Hoa theo CRF và tiếng Việt theo ngữ dụng của nhóm VCL trong việc phân đoạn từ tiếng Việt của kho ngữ liệu thử nghiệm.

CÁC THỬ NGHIỆM RANH GIỚI TỪ TRONG DỊCH THỐNG KÊ HOA-VIỆT

Ngữ liệu thử nghiệm

Kho ngữ liệu song ngữ thử nghiệm của chúng tôi bao gồm 12.000 cặp câu được chúng tôi tổng hợp từ các sách giáo khoa đàm thoại tiếng Hoa và các diễn đàn tiếng Hoa online. Văn bản trong kho ngữ liệu chủ yếu là văn bản giao tiếp phổ thông, chiều dài của các câu tương đối ngắn, bình quân khoảng 10 từ trong một câu. Chất lượng kho ngữ liệu khá sạch, nội dung ngữ liệu đồng nhất và trải đều trong 12.000 câu. Chúng tôi tiến hành thử nghiệm trên hai kho ngữ liệu: 8.000 và 12.000 cặp câu. Trong cả hai kho ngữ liệu, chúng tôi sử dụng 90 % tổng số câu cho huấn luyện (training), 5 % số câu dành cho kiểm tra (testing) và 5 % số câu còn lại dành cho điều chỉnh tham số (developing). Ngữ liệu huấn luyện (các câu dành cho huấn luyện và điều chỉnh tham số) được huấn luyện bằng công cụ Moses với các tham số mặc định (SMT Baseline). Chúng tôi sử dụng bộ ngữ liệu này để thực hiện hai thử nghiệm: ngữ liệu chưa được phân đoạn từ và ngữ liệu được phân đoạn từ.

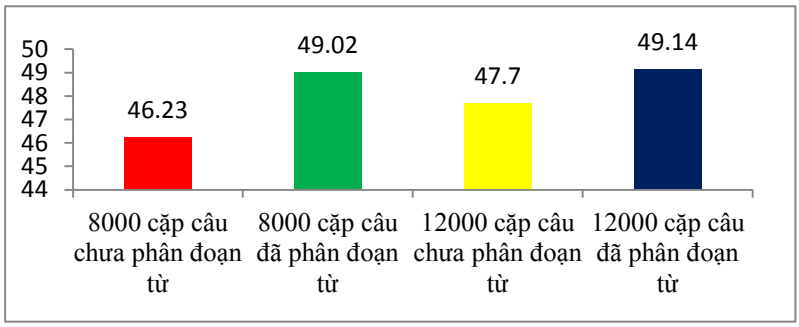
Thiết lập ngữ liệu và tiến hành thử nghiệm

Đối với ngữ liệu chưa phân đoạn từ, chúng tôi xem các ký tự tiếng Hoa và từ chính tả tiếng Việt như những đơn vị độc lập. Chúng tôi tiến hành chèn một khoảng trắng vào giữa các ký tự tiếng Hoa; ví dụ: "明天我去她家。" (Ngày mai tôi đi đến nhà cô ấy.) ->"明天我去她家。". Đối với tiếng Việt, chúng tôi thực hiện chèn khoảng trắng vào giữa các từ chính tả với các dấu câu (. , ? ! :); ví dụ câu: "Tôi nói tiếng Hoa, bạn có hiểu không?"-> "Tôi nói tiếng Hoa , bạn có hiểu không ?".

Trường hợp ngữ liệu được phân đoạn từ, chúng tôi tiến hành phân đoạn từ tiếng Hoa bằng công cụ Stanford Chinese Segmenter[1] do nhóm tác giả Pi-Chuan Chang, Huihsin Tseng và Galen Andrew thuộc trường Đại học Stanford thực hiện. Đây là công cụ phân đoạn từ được cài đặt theo phương pháp CRF (Conditional Random Field), được sử dụng phổ biến hiện nay. Đối với tiếng Việt, chúng tôi phân đoạn từ bằng công cụ của nhóm VCL. Khái niệm từ trong công cụ này là từ theo ngữ dụng, ứng dụng hiệu quả trong dịch máy thống kê.

Sau khi thiết lập ngữ liệu thử nghiệm, chúng tôi tiến hành huấn luyện và dịch thống kê bằng công cụ Moses cho cả hai trường hợp phân đoạn từ và chưa phân đoạn từ trên ngữ liệu 8.000 và 12.000 cặp câu. Điểm BLEU cho từng trường hợp như sau:

Hình 1. Kết quả thử nghiệm



[1] Download tại địa chỉ:

http://nlp.stanford.edu/software/segmenter.shtml

THẢO LUẬN

Qua hai thử nghiệm, chúng tôi nhận thấy rằng: khi tăng số câu trong kho ngữ liệu thì chất lượng dịch của cả hai trường hợp tăng theo. Thật vậy, ngữ liệu tăng đồng nghĩa với việc vốn từ trong "từ điển" gióng hàng phong phú hơn, bao quát hơn; hệ thống có khả năng nhận dạng nhiều từ hơn trong quá trình giải mã và cuối cùng hệ thống sẽ cho kết quả dịch tốt hơn.

Bên cạnh đó, chất lượng dịch máy sẽ tăng rõ rệt khi chúng tôi tiến hành phân đoạn từ cho hai kho ngữ liệu 8.000 và 12.000 cặp câu. Sau khi phân đoạn từ, hệ dịch cho kết quả với các từ được dịch đúng nghĩa hơn so với trường hợp chưa phân đoạn từ. Sau đây là bốn trường hợp điển hình trong 600 câu kiểm tra trong bộ dữ liệu thử nghiệm của chúng tôi (Bảng 2 và Bảng 3).

Bảng 2. Nghĩa của hình vị và từ chứa chúng

STT	Từ tiếng Hoa	Nghĩa đúng	Câu dịch chưa phân đoạn từ	Câu dịch đã phân đoạn từ
1	是的	Đúng (true, yes)	là của (is of)	đúng
2	花完	xài hết (spend all)	hoa xong (flower end)	xài hết
3	假使	giả sử (if)	Phép làm cho (holliday mak)	giả sử
4	银行卡	thẻ ngân hàng (bank card)	Ngân hàng thẻ (card bank)	thẻ ngân hàng

Bảng 3. Kết quả gióng hàng của hình vị, từ

STT	Chưa phân đoạn từ	Đã phân đoạn từ
1	是 -> là (is) 的 -> của (of)	是的 -> đúng (true, yes)
2	花 -> hoa (flower) 完 -> xong (end)	花完 -> xài hết (spend all)
3	假 -> phép (holliday) 使 -> làm cho (make)	假使 -> giả sử (if)
4	银行 -> ngân hàng (bank) 卡 -> thẻ (card)	银行卡 -> thẻ ngân hàng (bank card)

Có hai lý do cho ưu điểm này: một là do nghĩa của các hình vị không giống hoặc không liên quan đến nghĩa của từ chứa chúng (trường hợp 1, 2 và 3); hai là do trật tự của các hình vị trong từ bị đảo. Hai điều này đã làm giảm chất lượng dịch khi chưa phân đoạn từ. Ở từ số 1, từ 是/ là, từ 的/ của; trong khi đó, từ 是的 có nghĩa là "đúng". Ý nghĩa này không liên quan gì đến ý nghĩa "là", "của" ở các hình vị. Tương tự ở trường hợp 2, từ 花/hoa (trong bông hoa), từ 完/ xong. Kết hợp với nhau, 花完 tạo thành một từ duy nhất có nghĩa là "xài hết"; "hoa xong" không liên quan gì đến "xài hết". Ở trường hợp 3, từ 假 được gióng hàng là "phép" (放假 ->"nghỉ phép"), từ 使 có nghĩa là "khiến", "làm cho". Từ 假使 /"giả sử" (hoặc "nếu như") khác biệt hoàn toàn với nghĩa của hai hình vị 假,使 ("phép làm cho").

Khác với trường hợp 1, 2 và 3; ở trường hợp 4 nghĩa của các hình vị so với nghĩa của từ chứa các hình vị đó giống nhau, sự khác biệt ở đây là trật tự của các hình vị khi được dịch sang ngôn ngữ đích. Hai hình vị 银行/ "ngân hàng" và 卡 /"thẻ" dịch thành "ngân hàng thẻ"; trong khi đó từ "银行卡" dịch đúng sẽ là "thẻ ngân hàng". Ý nghĩa của "ngân hàng thẻ" và "thẻ ngân hàng" gần như nhau, chỉ khác trật tự từ.

Cũng từ kết quả thực nghiệm, chúng tôi nhận thấy rằng: hệ dịch đã qua phân đoạn từ mặc dù chất lượng chung cuộc tốt hơn so với chưa phân đoạn từ nhưng kết quả dịch lại xuất hiện nhiều từ mới hơn (hệ dịch không dịch được các từ này). Nguyên nhân chính của trường hợp này là do trong 600 câu thử nghiệm của trường hợp đã phân đoạn từ (bộ ngữ liệu 12.000 cặp câu) có nhiều từ mới và hệ thống không dịch được các từ này (do các từ mới này không tồn tại trong 10.800 câu huấn luyện). Tổng số từ ở trường hợp phân đoạn từ trong 10.800 cặp câu huấn luyện chắc chắn sẽ nhỏ hơn tổng số từ ở trường hợp chưa phân đoạn từ. Điều này dẫn đến số lượng cặp gióng hàng từ trong ngữ liệu song ngữ ở trường hợp phân đoạn từ sẽ ít hơn trường hợp chưa phân đoạn từ. Nói cách khác, nếu đứng ở khái niệm "từ" thì ngữ liệu huấn luyện chưa có nhưng các "hình vị" tạo nên "từ" đó thì lại có trong ngữ liệu huấn luyện. Điển hình như hai trường hợp sau trong 600 câu thử nghiệm (Bảng 4).

Bảng 4. Trường hợp dịch chưa phân đoạn từ cho kết quả tốt hơn dịch phân đoạn từ

STT	Câu tiếng Hoa	Kết quả dịch chưa phân đoạn từ	Kết quả dịch đã phân đoạn từ
1	气车在外边，我们送你到酒店	tiết xe ở bên ngoài , chúng tôi đưa ông đến khách sạn	气车 ở bên ngoài , chúng tôi đưa ông đến khách sạn
2	我走进病房看他	tôi đi vào phòng bệnh thăm anh ta	tôi 走进 phòng bệnh thăm anh ta

Ở câu số 1, từ 气车 (nghĩa đúng: "xe bus") bao gồm hai hình vị 气 và 车 (đều tồn tại trong ngữ liệu huấn luyện chưa phân đoạn từ), trong đó 气 có nghĩa là "tiết" (trong 天气: "thời tiết") có xác suất cao nhất, 车 có nghĩa phổ biến nhất là "xe'. Do đó, ở trường hợp chưa phân đoạn từ thì 气车 dịch là "tiết xe". Sau khi phân đoạn từ, từ 气车 được xem là một từ độc lập, do trong kho ngữ liệu huấn luyện chưa xuất hiện từ này nên hệ thống không dịch được từ 气车. Rốt cuộc, dù dịch sai từ 气 nhưng lại dịch đúng từ 车, kết quả dịch chưa phân đoạn từ của từ 气车 vẫn tốt hơn so với trường hợp phân đoạn từ (không dịch được, xem 气车 là từ mới). Tương tự cho câu số 2, từ 走进 ("đi vào"), ở trường hợp chưa phân đoạn từ thì hệ dịch cho kết quả chính xác (走: đi 进: vào); trong khi hệ dịch đã phân đoạn từ thì xem 走进 là từ mới.

Chúng tôi vừa trình bày các ưu và nhược điểm của hai trường hợp phân đoạn từ và không phân đoạn từ trong dịch thống kê Hoa-Việt trên hai bộ ngữ liệu với số câu tăng dần (8.000 và 12.000 cặp câu). Ở cả hai trường hợp đều có những ưu và khuyết điểm riêng. Các hướng cải tiến mà bài báo đề xuất tiếp theo sẽ dựa vào việc giảm bớt các nhược điểm và tăng cường ưu điểm của từng trường hợp. Như phần trên đã đề cập, chất lượng của hệ dịch sẽ tỉ lệ thuận với số lượng câu trong kho ngữ liệu dù là dịch đã phân đoạn từ hay chưa phân đoạn từ. Do đó, bài toán bổ sung số câu cho kho ngữ liệu là cần thiết và quan trọng nhằm làm phong phú "từ điển" gióng hàng của hệ thống. Hai vấn đề cần quan tâm cho bài toán bổ sung ngữ liệu đó là: số câu trong kho ngữ liệu như thế nào là đủ và nguồn ngữ liệu sẽ được thu thập từ đâu. Rất khó để xác định ngữ liệu bao nhiêu là đủ, bao nhiêu ngữ liệu để hệ thống có khả năng nhận diện tất cả các từ khi giải mã. Chúng ta chỉ mong muốn là kho ngữ lớn và sạch. Bên cạnh đó, vấn đề thu thập ngữ liệu lớn cho hệ dịch thống kê cũng là một thử thách. Bài toán thu thập ngữ liệu tự động từ các web song ngữ được đặt ra. Tuy nhiên, chất lượng ngữ liệu được lấy từ các web song ngữ thường thấp, cần có sự can thiệp tri thức để tăng chất lượng cho kho ngữ liệu này.

Một ưu điểm khác của dịch phân đoạn từ so với dịch chưa phân đoạn từ là dịch đúng nghĩa hơn. Một vấn đề cần quan tâm khi chúng ta phân đoạn từ cho ngữ liệu đó là: các gióng hàng giữa các hình vị có

nghĩa trong từ sẽ bị mất đi, thí dụ: gióng hàng "washing machine /洗衣机" sẽ không đóng góp gì cho cặp washing/洗衣 và machine/机. Đây cũng là lý do giải thích tại sao hệ dịch phân đoạn từ cho ra kết quả với nhiều từ không dịch được (từ mới). Giải quyết vấn đề này, nhóm tác giả ở công trình [6] đã phân rã từ tiếng Hoa thành các hình vị nhỏ hơn. Tuy nhiên, cặp ngôn ngữ mà nhóm tác giả này áp dụng là Anh-Hoa, nơi đó ranh giới từ ở tiếng Anh (mặc nhiên là khoảng trắng) là cơ sở để nhóm tác giả phân rã các từ tiếng Hoa. Cặp ngôn ngữ của chúng tôi là Hoa-Việt, khoảng trắng không cho biết ranh giới từ; do đó, chúng tôi không thể áp dụng việc phân rã từ dựa vào ngôn ngữ nguồn (tiếng Anh) như nhóm tác giả trên.

Chúng tôi đề nghị phân đoạn từ cho hệ dịch Hoa-Việt. Để tránh trường hợp bỏ mất các gióng hàng từ giữa các hình vị có nghĩa khi phân đoạn từ, chúng tôi sẽ tiến hành phân rã từ thành các hình vị có nghĩa nhỏ hơn. Một thách thức khi phân rã một từ thành các hình vị nhỏ hơn là có thể nghĩa của các hình vị không có liên quan gì đến nghĩa của từ chứa nó. Do đó, theo chúng tôi, chỉ những từ mà các hình vị tạo ra nó có nghĩa giống hoặc liên quan đến nó chúng tôi mới tiến hành phân rã. Thí dụ: với từ 走进 (đi vào) gồm hai hình vị 走/đi và 进/vào; nghĩa của hai hình vị gộp lại cũng chính là nghĩa của từ; do đó, từ này sẽ được phân rã thành hai hình vị. Tuy nhiên, đối với từ 是的 (đúng) cũng gồm hình vị 是/là và 的/của; nghĩa của hai hình vị này không có liên quan gì đến từ 是的. Do đó, từ 是的 sẽ không được phân rã.

Chúng tôi cũng đề xuất một hướng tiếp cận mới nhằm dịch lại các từ mới (unknown word: UKW) tiếng Hoa cho trường hợp dịch phân đoạn từ. Bản thân chữ Trung Quốc được phát âm khác nhau, ngay tại Trung Quốc, tuỳ từng vùng mà có nhiều giọng hoặc âm đọc khác nhau, như tiếng Quảng Đông, tiếng Phúc Kiến, tiếng Triều Châu, tiếng Bắc Kinh, ... Các nước lân cận như Triều Tiên có cách đọc riêng của người Triều Tiên, gọi là Hán-Triều (漢朝); người Nhật có cách đọc riêng của người Nhật, gọi là Hán-Hòa (漢和); và người Việt có cách đọc của mình gọi là Hán-Việt (漢越). Như vậy, âm Hán Việt là cách đọc tiếng Hán (tiếng Hoa) của người Việt. Thông thường, một ký tự tiếng Hoa sẽ có một âm Hán Việt và một nghĩa Thuần Việt, một số trường hợp âm Hán Việt cũng chính là nghĩa Thuần Việt. Thí dụ: ký tự 水 có âm Hán Việt là "thủy", Thuần Việt là "nước"; ký tự 东 có âm Hán Việt là "đông" (trong "đông tây nam bắc"), Thuần Việt cũng có nghĩa là "đông". Một điểm đặc biệt nữa giữa tiếng Hoa và Tiếng Việt đó là các tên riêng thuộc tên người, tên tổ chức và địa danh của tiếng Hoa sẽ được dịch sang tiếng Việt chính là âm Hán Việt của chúng. Quan hệ đặc biệt về nghĩa giữa tiếng Hoa và tiếng Việt là cơ sở quan trọng cho phương pháp dịch lại từ mới của chúng tôi trong tương lai.

KẾT LUẬN

Trong bài báo này, chúng tôi đã tiến hành khảo sát ảnh hưởng của yếu tố ranh giới từ đến kết quả dịch thống kê Hoa-Việt. Với kho ngữ liệu sạch, dữ liệu phân bố đồng đều thì khi tăng số câutrong kho ngữ liệu chất lượng dịch sẽ tăng theo. Hiệu suất dịch máy cũng sẽ tăng đáng kể nếu như ngữ liệu được phân đoạn từ. Bài báo cũng đã trình bày lại một số cải tiến phân đoạn từ nhằm tăng hiệu suất dịch thống kê Hoa-Anh-Hoa. Hai cải tiến đáng kể đó là phân đoạn từ dựa vào song ngữ và phân rã từ thành các hình vị có nghĩa nhỏ hơn. Dựa vào kết quả thực nghiệm cho hệ dịch thống kê Hoa-Việt, chúng tôi cũng đã đề xuất một số phương pháp cải tiến phân đoạn từ cũng như dịch lại từ mới nhằm tăng chất lượng dịch máy thống kê Hoa-Việt.

Trong nghiên cứu tiếp theo, chúng tôi sẽ tiến hành hiện thực hóa phương pháp cải tiến của mình cho hệ dịch Hoa-Việt, giúp cho chất lượng dịch của hệ thống dịch Hoa-Việt ngày càng tốt hơn.

LỜI CẢM ƠN: *Bài báo này được thực hiện dưới sự tài trợ của quỹ NAFOSTED và Trung tâm ngữ liệu đa ngữ Kim Từ Điển.*

Surveying word boundary factor in Chinese - Vietnamese statistical machine translation

• **Tran Thanh Phuoc**
Ton Duc Thang University
• **Dinh Dien**
University of Science,VNU- HCM

ABSTRACT

In isolating languages such as Chinese and Vietnamese, words are not separated by spaces, a word can include one or more spelling words. Segmenting word or not before training and translating process is a problem that need to be considered. In this paper, we will survey the effect of word boundary factor in the translation result of Chinese-Vietnamese statistical machine translation (SMT). The experimental result of this paper will be the basis for word segmentation improvement in future research which increase machine translation performance. We surveyed on two experiments: word segmentation (WS) and word un-segmentation (WUS) on the corpus of 8,000 and 12,000 sentence pairs. Based on the experimental results, we found that both of WS corpus and WUS corpus have their own advantages and defects. We propose integrating the advantages of these two methods in SMT.

Keywords*: Statistical machine translation, word boundary, word segmentation, morpheme, spelling word.*

TÀI LIỆU THAM KHẢO

[1]. P.C. Chang, M. Galley, C.D. Manning, Optimizing Chinese word segmentation for machine translation performance, in ACL Proceeding of the third workshop on statistical machine, translation, 224-232 (2008).

[2]. T.P. Tran, D. Dinh, Identifying and reodering prepositions in Chinese-Vietnamese machine translation, First International Workshop on Vietnamese language and speech processing (VLSP), In conjunction with 9th IEEE-RIVF conference on Computing and Communication Technologies *(RIVF 2012)*, 41-46 (2012).

[3]. T.T. Phước, Đ. Điền, Xử lý câu hỏi chính phản trong dịch thống kê Hoa-Việt, CS2602, Chuyên san Các công trình nghiên cứu, phát triển và ứng dụng công nghệ thông tin và truyền thông, 27, Bộ Thông tin và Truyền thông, 71-78 (2012).

[4]. R. Zhang, K. Yasuda, E. Sumita, Improved statistical machine translation by multiple Chinese word segmentation, ACL 2008, Third workshop on SMT, 216-223 (2008).

[5]. Y. Ma, A. Way, Bilingually motivated word segmentation for SMT, *In: EACL 2009* Workshop on Computational Approaches to Semitic Languages, 31 March, Athens, *Greece*, 549-557 (2009).

[6]. M.H. Bai, K.J. Chen, J.S. Chang, Improving word alignment by adjusting Chinese word segmentation, in Proceedings of the Third International Joint Conference on Natural Language Processing: I, 2008, India, 249-256 (2008).

[7]. C.T. Nguyen, T.K. Nguyen, X.H. Phan, L.M. Nguyen, Q.T. Ha, Vietnamese word

segmentation with CRFs and SVMs, Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC20), *China*, 215-222 (2006).

[8]. D. Dien, V. Thuy, A maximum entropy approach for Vietnamese word segmentation, in Research, Innovation and Vision for the Future, 248-253 (2006).