Tạp chí Tin học và Điều khiển học, T.30, S.2 (2014), 127–138

DỊCH LẠI TỪ CHƯA BIẾT DẠNG BIỂU THỨC SỐ TRONG DỊCH THỐNG KÊ HOA-VIỆT

TRẦN THANH PHƯỚC[1], ĐINH ĐIỀN[2]

[1]*Khoa Công Nghệ Thông Tin, Trường Đại Học Công Nghiệp Thực Phẩm Tp. Hồ Chí Minh; phuoctt@cntp.edu.vn*
[2]*Khoa Công Nghệ Thông Tin, Trường Đại Học Khoa Học Tự Nhiên Tp. Hồ Chí Minh; ddien@fit.hcmus.edu.vn*

Tóm tắt. Ranh giới từ trong tiếng Hoa và tiếng Việt không được xác định bởi khoảng trắng. Do đó, phân đoạn từ Hoa-Việt luôn được thực hiện đầu tiên trong bài toán xử lý ngôn ngữ Hoa-Việt nói chung và trong dịch máy thống kê Hoa-Việt nói riêng. Việc phân đoạn từ làm tăng chất lượng dịch chung cuộc nhưng lại xuất hiện nhiều "từ chưa biết" (Unknown Word: UKW) ở bản dịch đích. Dạng từ chưa biết phổ biến trong hệ thống dịch Hoa-Việt đó là thực thể có tên (named entity:NE). Trong bài báo này, chúng tôi sẽ trình bày một phương pháp lai, kết hợp luật và thống kê, để dịch lại các UKW dạng thực thể có tên biểu thức số. Áp dụng phương pháp này vào trong hệ dịch thống kê Hoa-Việt, kết quả thử nghiệm cho thấy phương pháp của chúng tôi đã cải tiến đáng kể hiệu suất dịch máy thống kê Hoa-Việt.

Từ khóa. Dịch thống kê Hoa-Việt, từ chưa biết, thực thể có tên, Number expression.

Abstract. Word boundary in Chinese and Vietnamese is not defined by a space. Therefore, Chinese-Vietnamese word segmentations are always done first in Chinese-Vietnamese natural language processing problem in general and in Chinese-Vietnamese statistical machine translation in particular. The word segmentation increases the final quality of translation but it appears many unknown words (UKW) in the target translation. The type of popular unknown word in Chinese-Vietnamese translation system that is named entity (NE). In this paper, we present a hybrid method to combine statistic and rule and to re-translate number expression *NE-UKW* (*NumExp-NE-UKW*). Applying this method into Chinese-Vietnamese SMT, the experiment result shows that our method significantly improves Chinese-Vietnamese SMT performance.

Key words. Chinese-Vietnamese statistical machine translation, unknown word, named entity, number expression.

1. GIỚI THIỆU

Tập hợp từ của ngôn ngữ tự nhiên là một tập mở. Không có cách nào để chúng ta có thể thu thập được tất cả các từ của một ngôn ngữ, do từ mới được phát sinh thường xuyên trong các hoạt động như: giải thích một khái niệm mới, một phát minh mới, tên trẻ em mới sinh, các tổ chức mới thành lập, ... Trong bài toán dịch máy thống kê, chúng ta có thể kết luận rằng: ngữ liệu huấn luyện của hệ thống dịch máy dù lớn đến mức nào đi nữa cũng không thể bao phủ hết tất cả các từ của một ngôn ngữ. Do đó, thay vì tìm cách làm sao cho hệ dịch có khả năng dịch được tất cả các từ của một

ngôn ngữ để không phát sinh "từ chưa biết" (unknown word: UKW), ở đây chúng tôi xem UKW như là một phần hiển nhiên của dịch máy và tìm cách dịch lại các UKW này để cải tiến chất lượng dịch máy chung cuộc. Không giống với các ngôn ngữ phương Tây, điển hình là tiếng Anh, các từ trong tiếng Hoa 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. Do đó, vấn đề phân đoạn từ luôn được giải quyết đầu tiên trong bài toán dịch máy từ tiếng Hoa sang ngôn ngữ khác (chủ yếu là tiếng Anh). Việc phân đoạn từ làm tăng chất lượng dịch chung cuộc nhưng lại xuất hiện nhiều UKW ở bản dịch đích do ngữ liệu huấn luyện ở trường hợp này ít từ vựng hơn khi chưa phân đoạn từ [3]. Phần lớn các UKW trong dịch thống kê Hoa-Việt là UKW dạng thực thể có tên (Named Entiy: NE). NE được chia thành các loại như sau: tên người, tên tổ chức, tên địa danh và các biểu thức số (ngày, giờ, phần trăm, số, số điện thoại) [9]. Theo [10], các loại NE được phân bố như sau (Bảng 1):

Bảng 1. Sự phân bố của NE tiếng Hoa trong ngữ liệu tin tức

NE	**Tỉ lệ % trong NE**	**Tỉ lệ % trong ngữ liệu**
Tên người	15,59	1,69
Tên địa danh	23,69	2,57
Tên tổ chức	10,07	1,09
Thời gian	16,33	1,77
Số	34,43	3,72

Trong phạm vi bài báo này, chúng tôi sẽ tập trung vào giải quyết bài toán nhận dạng và dịch lại NE-UKW dạng biểu thức số (Number expression *NE-UKW*: *NumExp-NE-UKW*) trong dịch thống kê Hoa-Việt. Một từ tiếng Hoa bao gồm nhiều hình vị có nghĩa kết hợp lại với nhau (thông thường hình vị cũng là ký tự tiếng Hoa). Từ trong các *NumExp* bao gồm các ký tự số kết hợp với các từ khóa đại diện cho mỗi loại *NumExp* (trình bày chi tiết ở phần 3). Các số và các từ khóa đại diện là hữu hạn, nhưng sự kết hợp của chúng đã tạo ra tập hợp các từ *NumExp* vô cùng lớn và hệ thống dịch máy không có khả năng nhận dạng hết các *NumExp* này. Hơn nữa, khi kết hợp lại với nhau, các số và các từ khóa này không còn giữ nguyên nghĩa ban đầu của chúng, tạo ra sự nhập nhằng về nghĩa dẫn đến kết quả dịch (nếu dịch được) có thể bị sai. Trong bài báo này, dựa vào ngữ pháp hình thành nên các *NumExp* của tiếng Hoa, chúng tôi đã xây dựng nên các luật chuyển đổi nhằm dịch lại các *NumExp–NE-UKW* cho hệ dịch thống kê Hoa-Việt.

Mặt khác, với bản chất nhập nhằng vốn có của ngôn ngữ, một từ có thể có nhiều nghĩa ở nhiều ngữ cảnh khác nhau. Biểu thức số cũng không ngoại lệ. Thông thường, một biểu thức số không đầy đủ sẽ có nhiều nghĩa trong từng ngữ cảnh khác nhau. Ví dụ, từ 十分, từ này nếu là biểu thức số thì có nghĩa là "10 phút"; tuy nhiên nó lại có nghĩa là "vô cùng" nếu tồn tại trong câu "我 ***十分*** 感谢 你" (tôi ***vô cùng*** cảm ơn ông). Đối với các trường hợp này, chúng tôi đề nghị sử dụng mô hình ngôn ngữ (language model) ở tiếng Việt để chọn ra nghĩa phù hợp.

Nội dung bài báo được trình bày như sau: ở Phần 2, chúng tôi sẽ trình bày các công trình liên quan đến bài toán xử lý các UKW trong dịch máy cũng như một số hướng tiếp cận cho bài toán nhận dạng thực thể có tên tiếng Hoa. Các cấu trúc cũng như các luật chuyển đổi của các *NumExp* sẽ được trình bày ở Phần 3. Trong khi đó, ở Phần 4 chúng tôi sẽ trình bày mô hình dịch lại các *NumExp-NE-UKW* từ tiếng Hoa sang tiếng Việt. Phần 5 chúng tôi sẽ trình bày các thử nghiệm, phần thảo luận sẽ được trình bày ở Phần 6. Phần kết luận sẽ được chúng tôi trình bày ở Phần 7.

2. CÔNG TRÌNH LIÊN QUAN

Trong phần này, chúng tôi sẽ trình bày một số công trình liên quan đến nhận dạng thực thể có tên và xử lý UKW.

Nhận dạng thực thể có tên tiếng Hoa đã được các nhà nghiên cứu quan tâm và tiến hành cài đặt thử nghiệm, các hướng tiếp cận chủ yếu trong thời gian gần đây là hướng tiếp cận lai (hybrid approach), kết hợp giữa thống kê và luật. Điểm mấu chốt để nhận dạng thực thể có tên tiếng Hoa là dựa vào các từ khóa của từng loại thực thể có tên (hướng tiếp cận dựa vào luật). Ví dụ: Để nhận dạng tên người, nhóm tác giả JianFeng Gao [8] đã phân tích tên người tiếng Hoa thành mẫu: <Family name> (F) + <Given name> (G) (tương ứng với tiếng Việt là <họ> + <tên>). Trong đó, F và G có chiều dài là một hoặc hai ký tự. Nhóm tác giả chỉ xem xét các ứng viên là tên người khi thành phần F của tên người đó có mặt trong danh sách 373 từ chỉ mục F của tiếng Hoa. Bước tiếp theo, nhóm tác giả nhận dạng phần G của tên người dựa vào mô hình ngôn ngữ xác suất 2-gram (bigram model) (hướng tiếp cận dựa vào thống kê).

Cũng theo cách này, nhóm tác giả Youzheng Wu [9] cũng đã sử dụng thuật toán lai, kết hợp mô hình thống kê dựa vào phân lớp (class based statistical model) với các loại tri thức ngôn ngữ khác nhau để nhận dạng thực thể có tên. Riêng đối với tác giả Keh-Jiann Chen và các đồng sự [7], nhóm tác giả chỉ tập trung nhận dạng tên tổ chức (organization names) dựa vào phân tích hình thái từ (Morphological analysis).

NE là dạng UKW phổ biến trong dịch thống kê nói chung và dịch thống kê Hoa-Việt nói riêng. Hiện nay có rất nhiều nghiên cứu với các hướng tiếp cận khác nhau nhằm dịch lại UKW, nâng cao hiệu suất dịch máy. Dựa vào phép chính tả của từ, nhóm tác giả Joao Silva và các đồng sự [4] đã đề xuất hai phương pháp nhằm khắc phục các UKW, đó là: phát hiện từ cùng nguồn gốc (cognates' detection) và độ tương tự hợp lý (logical analogy) để dịch lại UKW. Hướng tiếp cận này được các tác giả áp dụng cho cặp ngôn ngữ biến hình, điển hình là cặp ngôn ngữ "Anh-Bồ Đào Nha".

Một hướng tiếp cận khác để xử lý UKW được thực hiện bởi tác giả Matthias Eck [5] và các đồng sự. Nhóm tác giả này đã tìm các định nghĩa của các UKW ở ngôn ngữ nguồn và dịch các định nghĩa của UKW này (thay vì dịch các UKW). Các định nghĩa của UKW sẽ được rút trích tự động từ các từ điển trực tuyến và các bách khoa toàn thư, sau đó chúng được dịch lại qua hệ thống SMT. Kết quả dịch này sẽ thay thế các UKW ở bản dịch cũ. Các tác giả đã thực nghiệm hướng tiếp cận này trên cặp ngôn ngữ "Anh-Tây Ban Nha".

Ở khía cạnh khác, tác giả Ruiqiang Zhang và đồng sự [6] đã dịch lại các UKW bằng cách phân rã các UKW thành các từ con (subwords). Nhóm tác giả đã phân rã các UKW tiếng Hoa thành các từ con và dịch dựa vào các từ con này (subword-based translation). Từ con là một đơn vị ở giữa ký tự và từ. Bên cạnh đó, nhóm tác giả còn phát hiện ra rằng, chất lượng dịch sẽ tăng đáng kể nếu áp dụng nhận dạng thực thể có tên (Named entity recognition: NER) để dịch các UKW trước khi áp dụng dịch dựa vào từ con. Hướng tiếp cận này được áp dụng cho cặp ngôn ngữ "Hoa-Anh", nơi tiếng Anh với khoảng trắng là ranh giới của một từ và khoảng trắng này cũng là một tiêu chí rất quan trọng để phân rã các từ tiếng Hoa.

Đối với dịch biểu thức số tiếng Hoa, hiện nay có rất ít các công trình nghiên cứu về vấn đề này. Gần đây nhất là công trình của nhóm tác giả Mei Tu và đồng sự [11], các tác giả thực hiện dịch các biểu thức số dựa vào luật từ các ngôn ngữ khác (điển hình là tiếng Anh) sang chữ số tiếng Hoa. Riêng đối với các công cụ dịch biểu thức số tiếng Hoa hiện nay, phần lớn các công cụ chỉ dịch số thông thường, nếu có hỗ trợ dịch biểu thức số thì thường là dịch không chính xác các câu có chứa biểu thức số. Bảng 2 liệt kê một số công cụ có dịch số tiếng Hoa trực tuyến hiện nay.

Bảng 2. Các công cụ dịch biểu thức số tiếng Hoa

Số thứ tự	**Địa chỉ website**	**Phạm vi**
1	Công cụ Mandarin [13]	Dịch số, không dịch ngày giờ, không dịch câu chứa biểu thức số
2	Bing Translator [14]	Hỗ trợ dịch Hoa-Việt thông qua ngôn ngữ trung gian tiếng Anh
3	Công cụ dịch tự động đa ngữ của công ty SmartLink [15]	Dịch tổng quát Hoa-Việt, thường dịch sai các câu chứa biểu thức số
4	Google Translator [1]	Dịch tổng quát Hoa-Việt (thông qua ngôn ngữ trung gian tiếng Anh), thường dịch sai các câu chứa biểu thức số

Trong bài báo này, chúng tôi sử dụng tập luật luật để nhận dạng và dịch lại các *NumExp-NE-UKW*. Bên cạnh đó, chúng tôi sử dụng mô hình ngôn ngữ tiếng Việt để khử nhập nhằng các biểu thức số không đầy đủ.

3. CẤU TRÚC CỦA CÁC *NUMEXP*

NumExp bao gồm các loại sau: số, số không chứa ký tự đơn vị, số thứ tự, phân số, số thập phân, ngày và giờ. Từ trong các *NumExp* bao gồm các ký tự số kết hợp với các từ khóa đại diện cho mỗi loại *NumExp*. Do đó, số đóng vai trò chủ đạo trong việc hình thành các *NumExp* của tiếng Hoa.

Giống như tiếng Việt, số trong tiếng Hoa cũng được hình thành từ sự kết hợp các ký tự tương tự như các số từ 0 đến 9 của tiếng Việt. Các ký tự số của tiếng Hoa được trình bày ở Bảng 3 và Bảng 4.

Một điểm khác biệt nhỏ giữa tiếng Hoa và Việt là các số như 100, 1.000, 10.000 hay 100.000.000 thì tiếng Hoa có từ riêng dành cho các số này (tạm gọi là ký tự đơn vị).

Bảng 3. Ký tự số tiếng Hoa (0 đến 9)

Số tiếng Hoa	零	一	二	三	四	五	六	七	八	九
Số tiếng Việt	0	1	2	3	4	5	6	7	8	9

Bảng 4. Các ký tự đơn vị của tiếng Hoa

Ký tự đơn vị	十	百	千	万	亿
Số tiếng Việt	10	100	1.000	10.000	100.000.000

Sau khi phân đoạn từ, một từ dạng NumExp trong tiếng Hoa bao gồm nhiều ký tự số học kết hợp với các từ khóa đại diện cho các *NumExp*. Sự kết hợp này đã tạo ra rất nhiều từ *NumExp* trong tiếng Hoa và hệ thống huấn luyện không thể nhận dạng hết các từ này và phát sinh UKW. Dựa vào đặc trưng của các biểu thức số, chúng tôi phân loại chúng thành bảy loại sau.

a. *NumExp* dạng số có chứa ký tự đơn vị

Đây là một trong hai loại *NumExp* cơ bản (loại còn lại là NumExp không chứa ký tự đơn vị được trình bày ở phần b tiếp theo). Hai loại này đóng vai trò hạt nhân trong cấu trúc của các *NumExp*

còn lại. Thông thường loại từ này bao gồm từ 2 ký tự số học trở lên. Gọi N là *NumExp* dạng số, N rơi vào một trong hai trường hợp sau:

- Trường hợp 1: N có giá trị tương ứng từ $0 \rightarrow 19$: chuyển đổi theo bảng 5.
- Trường hợp 2: N có giá trị từ 20 trở lên:

+ Phân tách từ N thành các cụm từ S_i, cụm từ này bao gồm ký tự đơn vị và số đứng trước chúng. Ví dụ: 四百三十 được tách thành hai cụm: 四百 và 三十.

+ Công thức chuyển đổi như sau:

$$v = \sum_1^l f(c_i) * f'(dv_i). \quad (1)$$

Trong đó v là số tiếng Việt, c_i là ký tự số $(1 <= i <= l)$, dv_i là các ký tự đơn vị với $dv_i > dv_{i+1}$, f là hàm chuyển đổi ký tự tiếng Hoa sang tiếng Việt, $f(c_i)$ có giá trị từ 1 đến 9 (Bảng 1). f' là hàm chuyển đổi ký tự đơn vị sang số tương ứng (Bảng 4), l là số cụm từ S. Như ví dụ trên, số 四百三十 -> f(四) * f'(百) + f(三) * f'(十) $= 4 * 100 + 3 * 10 = 430$.

- Ngoại lệ: Theo chiều từ trái sang phải của từ N các ký tự đơn vị có giá trị giảm dần. Nếu cụm từ cuối cùng không chứa ký tự đơn vị, chỉ chứa số thì cụm cuối cùng (ký hiệu S_l) này được tính:

$$S_l = f(c_l) * f'(dv_{l-1})/10. \quad (2)$$

Ví dụ: số 三千四 -> f(三) * f'(千) + f(四) * f'(千)/10$= 3 * 1000 + 4 * 1000/10 = 3000 + 400 = 3400$.

Bảng 5. Chuyển đổi NumExp dạng số từ $0 \rightarrow 19$ tiếng Hoa sang số tiếng Việt

Số tiếng Hoa	一	二	三	四	五	六	七	八	九	十
Số tiếng Việt	1	2	3	4	5	6	7	8	9	10
Số tiếng Hoa	十一	十二	十三	十四	十五	十六	十七	十八	十九	
Số tiếng Việt	11	12	13	14	15	16	17	18	19	

b. *NumExp* dạng số không chứa ký tự đơn vị

Dạng số này thường sử dụng để biểu thị số điện thoại, số phòng.

- Phương pháp chuyển đổi:

+ Phân rã số này thành các ký tự riêng biệt c_i.

+ Chuyển đổi từng ký tự c_i sang số tiếng Việt, công thức chuyển đổi:

$$v = f(c_1)...f(c_l) \quad (3)$$

Trong đó, v là số tiếng Việt, $c_1 \ldots c_l$ là các ký tự số tiếng Hoa, l là tổng số ký tự tiếng Hoa có trong *NumExp* điện thoại, f là hàm chuyển đổi ký tự số tiếng Hoa sang số tiếng Việt (Bảng 3).

- Ví dụ: 三九八四 -> f(三) f(九) f(八) f(四) = 3984.

c. *NumExp* dạng số thứ tự

- Cấu trúc: 第**n**, với n là *NumExp* dạng số, 第 là từ khóa đại diện.
- Chuyển đổi sang tiếng Việt:
  + Từ khóa 第 được dịch sang tiếng Việt là "thứ"
  + n được dịch sang số tiếng Việt như ở mục a (*NumExp* dạng số)
- Ví dụ: 第十一 dịch sang tiếng Việt là "thứ 11".

d. *NumExp* dạng phân số

- Cấu trúc: n_1分之n_2, n_1 và n_2 là các *NumExp* dạng số, 分之 là từ khóa đại diện.
- Chuyển đổi sang tiếng Việt:
+ Từ khóa 分之 được dịch sang tiếng Việt là "/"
+ Công thức chuyển đổi:

$$v = f(n_2)/f(n_1) \tag{4}$$

Trong đó, v là kết quả tiếng Việt, f là hàm chuyển đổi số tiếng Hoa sang tiếng Việt, n_1 và n_2 là các số dạng có chứa ký tự đơn vị, cách thức chuyển đổi giống như cách chuyển NumExp dạng số ở Mục a.
- Ví dụ: 四分之三 -> *f*(三)/*f*(四) = 3/4

e. *NumExp* dạng số thập phân

- Cấu trúc: n_1点n_2,
Với n_1, n_2 là hai số có chứa ký tự đơn vị, 点 là từ khóa đại diện.
- Chuyển đổi sang tiếng Việt:
+ Từ khóa 点 được dịch sang tiếng Việt là dấu "," (dấu phẩy trong số thập phân)
+ n_1, n_2 được dịch sang số tiếng Việt như ở Mục a.
- Ví dụ: 六点三 -> 6,3.

f. NumExp dạng ngày

- Từ tiếng Hoa dạng ngày được chia làm 3 loại: từ chỉ ngày, từ chỉ tháng và từ chỉ năm. Cấu trúc tổng quát như sau: n_1年n_2月n_3日/号, trong đó: 日/ 号, 月 và 年 là các từ khóa đại diện.
- Chuyển đổi sang tiếng Việt:

+ Cấu trúc này được đảo trật tự khi dịch sang tiếng Việt, công thức chuyển đổi:

$$v = 日/号n_3月n_2年n_1 \tag{5}$$

+ Các từ khóa 日/ 号, 月, và 年 được dịch sang tiếng Việt lần lượt là: "ngày", "tháng" và "năm".
+ n_1 và n_2 là các số chứa ký tự đơn vị, được dịch sang số tiếng Việt như ở mục a.
+ n_3 là số không chứa ký tự đơn vị, được dịch sang số tiếng Việt như ở Mục b.
- Ví dụ: 二零零五年十二月十日 -> ngày 10 tháng 12 năm 2005.
- Một *NumExp* dạng ngày chỉ có thành phần "号/ngày" có thể bị nhập nhằng về nghĩa khi dịch sang tiếng Việt. Ví dụ như từ 六号 có hai nghĩa khác nhau trong hai câu sau:
+ 1. 今天 六号, 李 老师 来 吗 ? (Hôm nay ***ngày 6***, thầy giáo Lý có đến không?)
+ 2. 请 你 过 六号 窗口 取信. (Mời bạn qua cửa sổ ***số 6*** nhận thư.)

Từ 六号 ở cả hai câu đều là các *NumExp* dạng ngày không đầy đủ, ở câu 1 nó có nghĩa là "ngày 6" nhưng ở câu 2 nó lại có nghĩa là "số 6".

g. *NumExp* dạng giờ

Một từ w được xem là giờ nếu nó thõa mãn:
- Ký tự 点 (giờ) nằm ở cuối từ w, các ký tự phía trước phải là số.

- Nếu ký tự 点 nằm ở giữa từ thì phía sau phải có ký tự 半 (rưỡi) hoặc 钟 (đồng hồ) hoặc ký tự 分 (phút), 刻 (giây). Các ký tự còn lại trong từ w phải là số.

Số trong giờ là số chứa ký tự đơn vị (*NumExp* dạng số), được chuyển đổi theo cách ở Mục a, phạm vi của số dạng giờ là nhỏ, cụ thể: giờ từ: 0 đến 24, phút và giây: từ 0 đến 60.

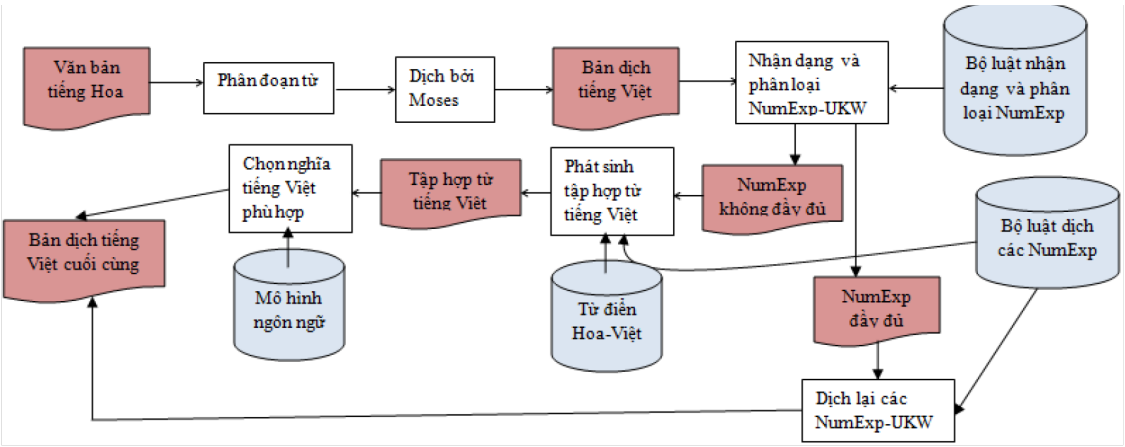
- Ví dụ: 十二点二十三分 → 12 giờ 23 phút.
- Từ 十分 nếu đứng đơn lẻ có thể bị nhập nhằng nghĩa khi dịch sang tiếng Việt. Ví dụ như ở hai câu sau, từ 十分 có hai nghĩa hoàn toàn khác nhau.

+ 1. 我 十分 感谢 你 。(Tôi vô cùng cám ơn anh.)

+ 2. 我 跑 一周 花 大概 十分 (Tôi chạy 1 vòng tốn khoảng 10 phút) Từ 十分 ở cả hai câu đều là các *NumExp* dạng ngày không đầy đủ, ở câu 1 nó có nghĩa là "vô cùng" nhưng ở câu 2 nó lại có nghĩa là "10 phút".

4. MÔ HÌNH DỊCH LẠI *NUMEXP-NE-UKW*

Mô hình của chúng tôi như sau:



Hình 1. Mô hình dịch lại *NumExp-NE-UKW*

- Chúng tôi sử dụng công cụ Stanford Chinese Segmenter[1] để phân đoạn từ cho kho ngữ liệu tiếng Hoa. Ngữ liệu tiếng Việt được chúng tôi phân đoạn bằng công cụ của nhóm chúng tôi, công cụ này được cài đặt theo hướng tiếp cận Maximum Entropy [12]. Tiếp theo, chúng tôi sử dụng công cụ Moses[2] để dịch câu tiếng Hoa đầu vào.

- Từ kết quả dịch thống kê của công cụ Moses, chúng tôi tiếp tục nhận diện các UKW tiếng Hoa và lọc ra các *NumExp-NE-UKW*. Mẫu tự Hoa-Việt khác nhau, do đó, chúng tôi dễ dàng nhận diện các *UKW* tiếng Hoa trong bản dịch tiếng Việt. Dựa vào tập luật phân loại, chúng tôi tiến hành phân loại các *NumExp-NE-UKW* này thành bảy loại tương ứng (như Phần 3. đã đề cập). Trong các loại *NumExp* thì *NumExp* dạng ngày và giờ có thể bị nhập nhằng khi chúng bị thiếu các thành phần cấu tạo, chúng tôi gọi trường hợp này là "*NumExp* không đầy đủ". Các *NumExp* còn lại được gọi là "*NumExp* đầy đủ".

- Dịch lại NumExp đầy đủ: loại này được dịch bởi bộ luật chuyển đổi.

[1]Download tại địa chỉ: http://nlp.stanford.edu/software/segmenter.shtml
[2]Download tại địa chỉ: http://www.statmt.org/moses/

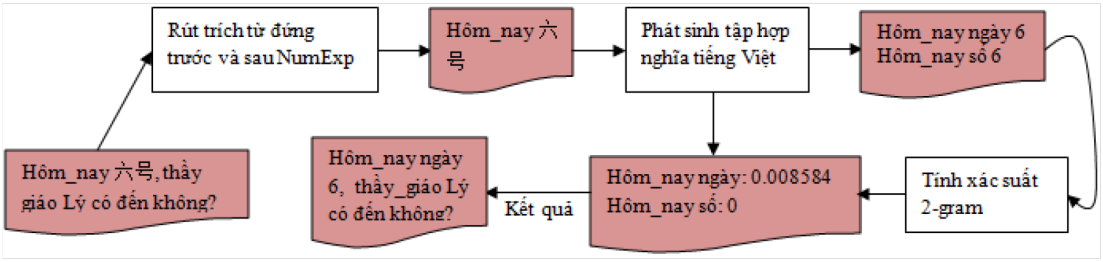
- Đối với các NumExp không đầy đủ:

+ Phát sinh tập hợp nghĩa của của các NumExp này dựa vào bộ luật chuyển đổi và từ điển Hoa-Việt.

+ Chọn nghĩa tốt nhất từ tập hợp nghĩa này dựa vào mô hình ngôn ngữ 2-gram. Gọi w_i là *NumExp-NE-UKW* trong một ngữ tiếng Việt, w_{i-1} , w_{i+1} là từ đứng trước và sau nó, nếu *NumExp* đứng ở đầu hoặc cuối một ngữ thì sẽ không có từ w_{i-1} hoặc w_{i+1} tương ứng. Nghĩa tiếng Việt tốt nhất $v(w)$ của w_i được tính theo mô hình ngôn ngữ 2-gram như sau:

$$v\,(w) = \underset{w}{^{argmax}}(p(w_i|w_{i-1}) + p(w_{i+1}|w_i)). \tag{6}$$

Chúng tôi sử dụng tổng xác suất của hai 2-gram (thay vì phải là tích) để tránh trường hợp một trong hai 2-gram có giá trị bằng 0. Hình 2 sẽ minh họa quá trình chọn nghĩa cho *NumExp* 六号 trong câu dịch tiếng Việt "Hôm_nay 六号, thầy_giáo Lý có đến không ?".



Hình 2. Minh họa quá trình dịch *NumExp*

Do từ 六号 đứng ở cuối một ngữ và được dịch thành hai từ tiếng Việt ("ngày" "6" hoặc "số" "6") nên chúng tôi chỉ tính xác suất 2-gram của từ đứng trước nó với từ "ngày" hoặc "số", cụ thể, chúng tôi tính xác xuất 2-gram cho hai cặp "Hôm_nay ngày" và "Hôm_nay số". Cụm từ "Hôm_nay ngày" có xác suất cao hơn nên câu dịch tiếng Việt được chọn sẽ là "Hôm_nay ngày 6 , thầy_giáo Lý có đến không ?".

5. THỬ NGHIỆM

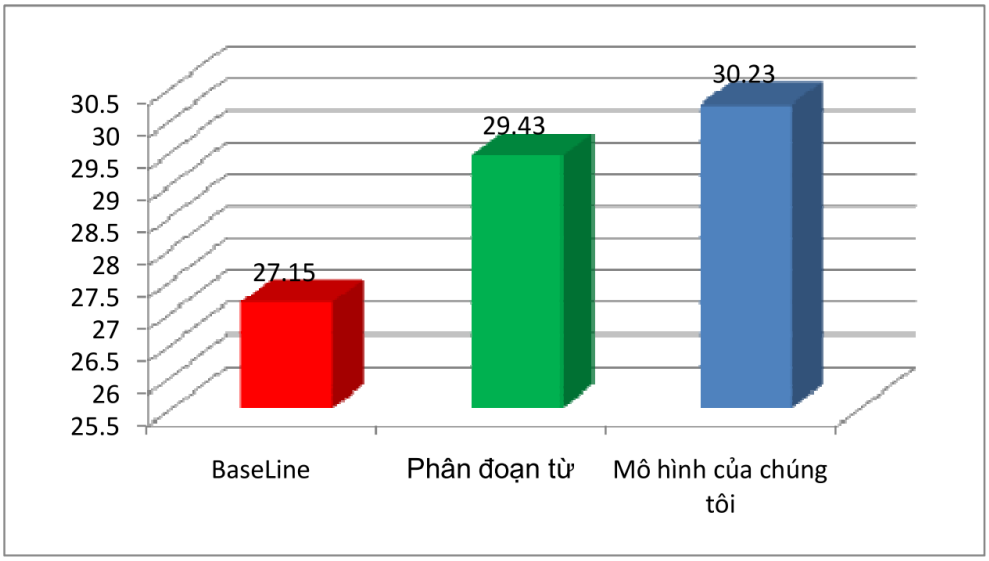
Kho ngữ liệu song ngữ thử nghiệm của chúng tôi bao gồm 20.000 cặp câu được chúng tôi tổng hợp từ các sách giáo khoa đàm thoại tiếng Hoa và các diễn đàn tiếng Hoa trực tuyến. Văn bản trong kho ngữ liệu chủ yếu là văn bản giao tiếp phổ thông, chiều dài của các câu tương đối ngắn, bình quân khoảng 10 từ trong một câu. Chất lượng kho ngữ liệu khá sạch, nội dung ngữ liệu đồng nhất và trải đều trong 20.000 câu. 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 ba thử nghiệm: dịch không phân đoạn từ, dịch phân đoạn từ, dịch lại *NumExp-NE-UKW* cho trường hợp phân đoạn từ. Bên cạnh đó, chúng tôi cũng thử nghiệm dịch các *NumExp* trong bộ ngữ liệu thử nghiệm qua hệ dịch Google Translator và Bing Translator. Thử nghiệm này không nhằm mục đích

đánh giá chất lượng dịch của hệ thống chúng tôi so với Google và Bing, vì với ngữ liệu vô cùng lớn của mình, các hệ dịch này dường như dịch được mọi câu đầu vào, ít khi phát sinh UKW. Mục đích của thử nghiệm này chỉ để chứng minh cải tiến của chúng tôi trong phạm vi hẹp là dịch văn bản tiếng Hoa có chứa các *NumExp*. Mô hình ngôn ngữ được huấn luyện từ ngữ liệu tiếng Việt gồm 212.454 cặp câu của nhóm VLSP.

Ở hệ dịch cơ sở, 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. Đối với tiếng Việt, chúng tôi thực hiện chèn khoảng trắng vào giữa các từ chính tả với các dấu câu.

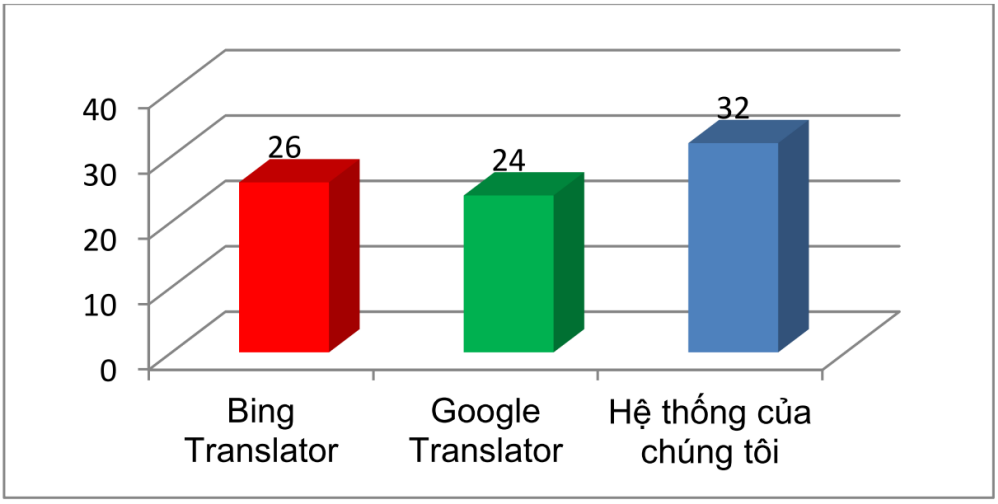
Ở thử nghiệm dịch phân đoạn từ, chúng tôi tiến hành phân đoạn từ tiếng Hoa bằng công cụ Stanford Chinese Segmenter. Đối với tiếng Việt, chúng tôi phân đoạn từ bằng công cụ của nhóm chúng tôi. Khái niệm từ trong công cụ này là từ theo ngữ dụng, ứng dụng hiệu quả trong dịch máy thống kê.

Dựa vào kết quả dịch của Moses ở trường hợp phân đoạn từ, chúng tôi tiến hành dịch lại kết quả này. Tùy vào cách chọn lựa câu thử nghiệm mà điểm BLEU có sự chênh lệch tùy theo cách chọn lựa. Sự chọn lựa bộ thử nghiệm cho ra kết quả với nhiều *NumExp-NE-UKW* chắc chắn điểm BLEU của hệ dịch phân đoạn từ được dịch bởi MOSES sẽ thấp hơn nhiều so với hệ dịch có can thiệp dịch lại UKW *NumExp* và ngược lại. Sau đây là kết quả dịch của bộ thử nghiệm được chọn lựa theo mẫu: mỗi 20 câu trong kho ngữ liệu thì 18 câu đầu dành cho huấn luyện, câu 19 dành cho điều chỉnh tham số và câu thứ 20 dành cho thử nghiệm.



Hình 3. Kết quả thử nghiệm

Bên cạnh đó, chúng tôi cũng đã tiến hành thử nghiệm dịch một số câu có chứa các *NumExp* qua 3 hệ thống: hệ thống của chúng tôi, Bing Translator và Google Translator. Chúng tôi rút trích tự động 34 câu tiếng Hoa có chứa các *NumExp* từ ngữ liệu kiểm tra (testing) của hệ thống. Số câu dịch đúng của 3 hệ thống được trình bày ở Hình 4.



Hình 4. Kết quả dịch qua 3 hệ thống

6. THẢO LUẬN

Kết quả dịch phân đoạn từ thường cho kết quả tốt hơn không phân đoạn từ (dịch cơ sở), vấn đề này chúng tôi đã trình bày ở công trình [3]. Tuy nhiên, dịch phân đoạn từ mặc dù kết quả tốt hơn so với trường hợp dịch cơ sở nhưng lại xuất hiện nhiều UKW. Kết quả dịch của trường hợp phân đoạn từ sẽ được tiếp tục dịch qua hệ thống dịch lại *NumExp* của chúng tôi. Bên cạnh đó, ngữ liệu thử nghiệm cũng được chúng tôi dịch qua hệ thống của Google và Bing Translator. Với kho ngữ liệu vô cùng lớn, hai hệ dịch này dường như dịch được tất cả các từ tiếng Hoa. Tuy nhiên, cũng giống như hệ dịch cơ sở, kết quả dịch của Google tuy ít phát sinh UKW nhưng kết quả dịch thường bị sai nghĩa. Bảng 6 trình bày bốn trường hợp cụ thể trong 34 câu thử nghiệm có chứa *NumExp*:

Cả bốn trường hợp hệ dịch cơ sở đều cho ra kết quả nhưng phần lớn kết quả lại không chính xác (trừ trường hợp 2). Do các ký tự số đều có trong ngữ liệu huấn luyện của hệ dịch cơ sở nên hệ dịch này nhận diện được các ký tự số. Tuy nhiên, do là dịch cơ sở, không có tri thức nên hệ dịch bị nhập nhằng về nghĩa ở các ký tự thời gian (点, 分, 号) cũng như không thể chuyển đổi đúng từ số tiếng Hoa sang tiếng Việt. Ví dụ: ở câu số 4 từ (号) có nghĩa là "số" với xác suất cao nhất, nhưng trong trường hợp này, từ (号) có nghĩa đúng là "ngày". Chúng tôi không bàn luận về kết quả dịch sai của Google và Bing vì hai hệ dịch này phải thông qua ngôn ngữ trung gian tiếng Anh, kết quả sai ở tiếng Việt có thể là sai cộng hưởng khi dịch từ Hoa-Anh và Anh-Việt. Ở trường hợp phân đoạn từ, số từ trong ngữ liệu của trường hợp này sẽ ít hơn so với trường hợp dịch cơ sở, dẫn đến từ điển "gióng hàng từ" cũng sẽ ít hơn, khả năng nhận dạng từ của hệ thống cũng sẽ kém hơn. Kết quả là hệ dịch phân đoạn từ phát sinh nhiều UKW.

Mặt khác, với sự phong phú của *NumExp-NE*, nên dù ngữ liệu có lớn đến mức nào đi nữa cũng rất khó có thể bao quát hết các *NumExp*, nên hệ thống không dịch được *NumExp-NE* là điều khó tránh khỏi. Kết quả dịch phân đoạn từ sau khi cho qua hệ thống của chúng tôi đã cho ra kết quả với điểm BLEU tăng lên rõ rệt, do hệ thống đã nhận dạng và dịch đúng các *NumExp*. Ở ví dụ trên, hệ thống đã dịch đúng 4 trường hợp *NumExp*.

Bảng 6. Một số câu tiếng Hoa được dịch qua 5 hệ thống

STT	Câu tiếng Hoa	Câu dịch chuẩn	Dịch cơ sở	Dịch phân đoạn từ	Dịch Google[3]	Bing Translator[3]	Dịch đã xử lý
1	今年 一九九二年，明年 一九九三	Năm nay là năm 1992, năm sau là năm 1993	Năm nay một chín chín hai năm, năm sau một chín chín ba	Năm nay 一九九二年，năm sau 一九九三	Năm nay, năm 1992, năm tới 1993	Năm 1992, tiếp theo năm 1993	Năm nay năm 1992, năm sau 1993
2	一点 过 五分 。	1 giờ 5 phút	1 giờ qua 5 phút	1 giờ qua 5 phút	Bây giờ hai mươi ba.	Ít năm điểm.	1 giờ qua 5 phút
3	我 想 改到 五月 十九日 。	Tôi muốn đổi sang ngày 19 tháng 5	Tôi muốn đổi 5 tháng 19 ngày.	Tôi muốn đổi 五 月 十九日 。	Tôi muốn thay đổi đến tháng 19.	Tôi muốn thay đổi đến 19 ngày.	Tôi muốn đổi ngày 19 tháng 5
4	不 是 五月 三号 是 五月 十号 。	Không phải ngày 3 tháng 5 mà là ngày 10 tháng 5	Không phải 5 tháng số 3 là 5 tháng số 10.	Không phải là 五 月 三号 là 五月 十号	Không phải là một ngày 3 tháng 5 là tháng 10.	Không phải trên có thể 3 ngày 10 tháng 5.	Không phải ngày 3 tháng 5 là ngày 10 tháng 5

[3] Dịch ngày 11/04/2014

7. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất phương pháp xử lý *NumExp-NE-UKW* trong dịch thống kê Hoa-Việt dựa vào tri thức ngữ pháp của chúng. Bên cạnh đó, đối với các *NumExp-NE-UKW* không đầy đủ, chúng tôi đã khử nhập nhằng nghĩa của chúng bằng mô hình ngôn ngữ 2-gram. Kết quả thực nghiệm cho thấy hệ thống của chúng tôi đã dịch rất tốt các NumExp, góp phần cải tiến đáng kể chất lượng dịch máy Hoa-Việt. Trong tương lai, chúng tôi sẽ áp dụng phương pháp này kết hợp với thống kê trên ngữ liệu lớn để xử lý các thực thể có tên còn lại, nhằm giúp cho hệ thống dịch Hoa-Việt-Hoa ngày càng nâng cao.

LỜI CẢM ƠN

Bài báo này được thực hiện dưới sự tài trợ của Trung tâm dữ liệu đa ngữ Kim Từ Điển.

TÀI LIỆU THAM KHẢO

[1] Trần Thanh Phước – Đinh 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*, số 27, Bộ Thông tin và Truyền thông, 2012.

[2] Phuoc Tran an, Dien Dinh, Identifying and reordering 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)*, Ho Chi Minh, Vietnam, 2012, 41-46.

[3] Trần Thanh Phước – Đinh Điền, Khảo sát yếu tố ranh giới từ trong dịch thống kê Hoa-Việt, *Hội nghị Khoa học lần XIII Đại Học Khoa Học Tự Nhiên TP.HCM*, 2012, 549-556.

[4] Joao Silva, Luisa Coheur, Angela Costa, Isabel Trancoso, Dealing with unknown words in statistical machine translation, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012, 3977-3981.

[5] Matthias Eck, Stephan Vogel, Alex Waibel, Communicating unknown words in machine translation, *International Conference on Language Resources and Evaluation*, 2008, 1542-1547.

[6] Ruiqiang Zhang, Eiichiro Sumita, Chinese unknown word translation by subword resegmentation, *International Joint Conference on Natural Language Processing*, 2008, 225-232.

[7] Keh-Jiann and Chao-jan Chen, Knowledge Extraction for Indentification of Chinese Organization Names, *Second Chinese Language Processing Workshop*, Hong Kong, 2000, 15-21.

[8] Jianfeng Gao, Mu Li, and Chang-Ning Huang, Improved source-channel models for chinese word segmentation, *ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 2003, 272-279.

[9] Youzheng Wu, Jun Zhao and Bo Xu, Chinese named entity recognition combining a statistical model with human knowledge, *MultiNER '03 Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition*, Volume 15, 2003, 65-72.

[10] Hua-Ping Zhang, Qun Liu, Hong-Kui Yu., Xue-Qi Cheng and Shuo Bai, Chinese named entity recognition using role model, *Computational Linguistics and Chinese Language Processing*, **8**(2) (2003) 29-60.

[11] Mei Tu, Yu Zhou and Chengqing Zong, A universal approach to translating Numerical and Time Expressions, *Proceedings IWSLT 2012, International workshop on spoken language translation*, 2012, 209-216.

[12] Dinh Dien and Vu Thuy, A maximum entropy approach for Vietnamese word segmentation, *Research, Innovation and Vision for the Future, 2006 International Conference*, Ho Chi Minh, Vietnam, 2006, 248-253.

[13] http://www.mandarintools.com/numbers.html

[14] http://www.bing.com/translator

[15] http://imtranslator.net/translation/chinese-simplified/to-vietnamese/translation/

[16] https://translate.google.com.vn/

Received on March 21, 2013
Revised on March 15, 2014