ĐẠI HỌC QUỐC GIA HÀ HỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN TUẤN ANH

CHUYỂN NGỮ TỰ ĐỘNG TỪ TIẾNG VIỆT SANG TIẾNG NHẬT

LUẬN VĂN THẠC SỸ

Hà Nội - 2017

CHƯƠNG I. GIỚI THIỆU

Hiện nay có hàng nghìn ngôn ngữ trên toàn thế giới, mỗi ngôn ngữ đều có những đặc trưng riêng về bảng chữ cái và cách phát âm. Ngày càng có nhiều những hệ thống tự động dịch miễn phí trên mạng như: systran, google translate, vietgle, vdict, … Những hệ thống này cho phép dịch tự động các văn bản với một cặp ngôn ngữ chọn trước (ví dụ dịch từ tiếng Anh sang tiếng Việt). Điều ấy cho thấy sự phát triển của dịch máy càng ngày càng tiến gần hơn đến ngôn ngữ tự nhiên của con người. Vấn đề đặt ra đối với cả dịch giả và máy dịch trong việc dịch giữa các cặp ngôn ngữ có hệ thống bảng chữ cái và cách phát âm khác nhau là dịch chính xác tên và các thuật ngữ kỹ thuật. Những đối tượng này được phiên âm, thay thế bởi những âm xấp xỉ tương đương. Việc dịch phiên âm giữa các cặp ngôn ngữ đó được gọi là Chuyển ngữ.

Thật khó để dịch các tên riêng và thuật ngữ kỹ thuật qua các ngôn ngữ với các bảng chữ cái và cách phát âm khác nhau. Các từ này thường được chuyển ngữ, tức là, thay thế bằng khoảng ngữ âm gần đúng. Ví dụ: "computer" trong tiếng Anh xuất hiện dưới dạng "konpyuutaa" trong Tiếng Nhật.

1.1. Đặc điểm ngôn ngữ tiếng Việt và tiếng Nhật

1.1.1. Đặc điểm ngôn ngữ tiếng Việt

Tiếng Việt thuộc ngôn ngữ đơn lập, tức là mỗi một âm tiết được phát âm tách rời nhau và được thể hiện bằng một chữ viết. Đặc điểm này thể hiện rõ rệt ở tất cả các mặt ngữ âm, từ vựng, ngữ pháp.

Đặc điểm ngữ âm

Trong tiếng Việt có một loại đơn vị đặc biệt gọi là "tiếng". Về mặt ngữ âm, mỗi tiếng là một âm tiết và cách viết tương đồng với phát âm. Hệ thống âm vị tiếng Việt phong phú và có tính cân đối.

Đặc điểm từ vựng

Mỗi tiếng, nói chung, là một yếu tố có nghĩa. Tiếng là đơn vị cơ sở của hệ thống các đơn vị có nghĩa của tiếng Việt. Từ tiếng, người ta tạo ra các đơn vị từ vựng khác để định danh sự vật, hiện tượng..., chủ yếu nhờ phương thức ghép và phương thức láy.

Việc tạo ra các đơn vị từ vựng ở phương thức ghép luôn chịu sự chi phối của quy luật kết hợp ngữ nghĩa, ví dụ: đất nước, máy bay, nhà lầu xe hơi, nhà tan cửa nát...

Vốn từ vựng tối thiểu của tiếng Việt phần lớn là các từ đơn tiết (một âm tiết, một tiếng).

1.1.2. Đặc điểm ngôn ngữ tiếng Nhật

Hệ thống chữ viết

Người Nhật có một bảng chữ cái đặc biệt về ngữ âm được gọi là Katakana, được sử dụng chủ yếu để viết tên nước ngoài và từ mượn. Các ký hiệu katakana được thể hiện trong Bảng 1.1, với cách phát âm tiếng Nhật của chúng. Hai ký hiệu được hiển thị ở góc dưới bên phải được sử dụng để kéo dài nguyên âm hoặc phụ âm tiếng Nhật.

ア (a)	カ (ka)	サ (sa)	タ (ta)	ナ (na)	ハ (ha)	マ (ma)	ラ (ra)
イ (i)	キ (ki)	シ (shi)	チ (chi)	ニ (ni)	ヒ (hi)	ミ (mi)	リ (ri)
ウ (u)	ク (ku)	ス (su)	ツ (tsu)	ヌ (nu)	フ (hu)	ム (mu)	ル (ru)
エ (e)	ケ (ke)	セ (se)	テ (te)	ネ (ne)	ヘ (he)	メ (me)	レ (re)
オ (o)	コ (ko)	ソ (so)	ト (to)	ノ (no)	ホ (ho)	モ (mo)	ロ (ro)
バ (ba)	ガ (ga)	パ (pa)	ザ (za)	ダ (da)	ァ (a)	ヤ (ya)	ャ (ya)
ビ (bi)	ギ (gi)	ピ (pi)	ジ (ji)	デ (de)	ィ (i)	ヨ (yo)	ョ (yo)
ブ (bu)	グ (gu)	プ (pu)	ズ (zu)	ド (do)	ゥ (u)	ユ (yu)	ュ (yu)
ベ (be)	ゲ (ge)	ペ (pe)	ゼ (ze)	ン (n)	ェ (e)	ヴ (v)	ッ
ボ (bo)	ゴ (go)	ポ (po)	ゾ (zo)	ヂ (chi)	ォ (o)	ワ (wa)	ー

Bảng 1.1: Bảng chữ cái Katakana và cách phát âm tiếng Nhật

Ngữ âm

Âm tiết trong tiếng Nhật giữ một vị trí rất quan trọng, nó vừa là đơn vị ngữ âm nhỏ nhất và vừa là đơn vị phát âm cơ bản. Mỗi âm tiết được thể hiện bằng một chữ Kana (Hiragana và Katakana). Số lượng âm tiết trong tiếng Nhật không nhiều, có tất cả 112 dạng âm tiết. Trong số này, có 21 dạng âm tiết chỉ xuất hiện trong các từ ngoại lai được vay mượn, do đó số lượng âm tiết sử dụng thường xuyên trên thực tế còn ít hơn.

Tiếng Nhật có tất cả 5 nguyên âm: /a, i, u, e, o/ và 12 phụ âm: /k, s, t, g, z, d, n, m, h, b, p, r/ một số lượng khá ít so với các ngôn ngữ khác. Ngoài ra còn có hai âm đặc biệt là âm mũi (N) và âm ngắt (Q).

Từ vựng

Có thể khẳng định rằng tiếng Nhật là một ngôn ngữ có một vốn từ vựng rất lớn và vô cùng phong phú. Sự phong phú của từ vựng tiếng Nhật trước hết được thể hiện ở tính nhiều tầng lớp của vốn từ vựng. Nhóm từ mượn được coi là những từ vay mượn từ các ngôn ngữ khác mà chủ yếu là tiếng Anh, Pháp, Đức, Tây Ban Nha, Bồ Đào Nha.... Để phân biệt với nhóm từ gốc Hán và từ thuần Nhật, nhóm từ mượn được viết bằng chữ Katakana, ví dụ như: tabako (thuốc lá), tenpura (món tẩm bột rán).....

1.2 Bài toán dịch máy và tiếp cận dịch dựa trên cụm từ phân cấp

1.2.1 Khái niệm về hệ dịch máy

a. Định nghĩa

Dịch máy (machine translation system-MT) là một lĩnh vực của ngôn ngữ học tính toán nghiên cứu việc sử dụng phần mềm để dịch văn bản hoặc bài phát biểu từ ngôn ngữ này sang ngôn ngữ khác.

b. Vai trò của dịch máy

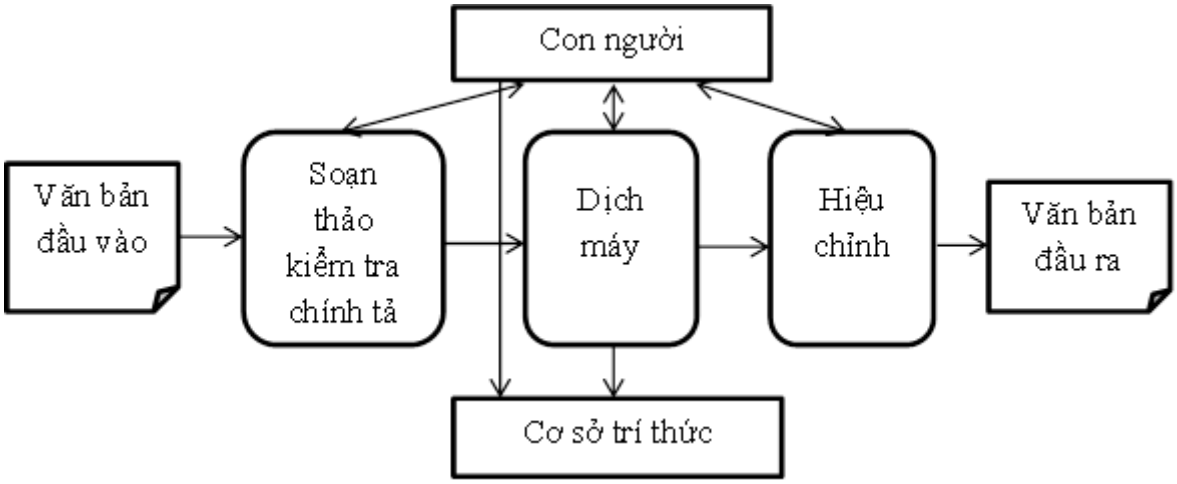
Hiện nay trên thế giới có khoảng 5650 ngôn ngữ khác nhau, với một số lượng ngôn ngữ lớn như vậy đã gây ra rất nhiều khó khăn trong việc trao đổi thông tin,.Với những khó khăn như vậy người ta đã phải dùng đến một đội ngũ phiên dịch khổng lồ, để dịch các văn bản, tài liệu, lời nói từ tiếng nước này sang tiếng nước khác. Để khắc phục những nhược

điểm trên con người đã nghĩ đến việc thiết kế một mô hình tự động trong công việc dịch, do đó ngay từ khi xuất hiện chiếc máy tính điện tử đầu tiên người ta đã tiến hành nghiên cứu về dịch máy.

c. Sơ đồ tổng quan của một hệ dịch máy

Phần dịch máy sẽ chuyển văn bản nguồn thành văn bản viết trên ngôn ngữ đích. Và cũng qua một bộ chỉnh ra để cuối cùng thu được một văn bản tương đối hoàn chỉnh.

Dưới đây là sơ đồ tổng quát của một hệ dịch máy:



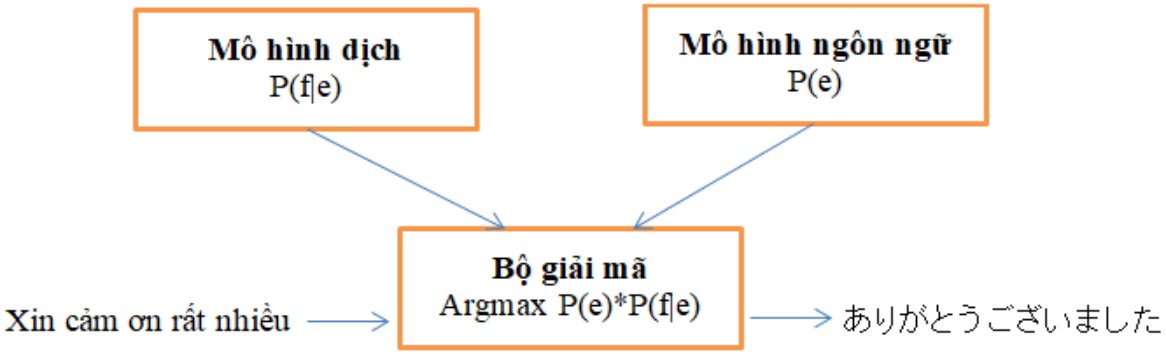
Hình 1.1: Sơ đồ tổng quan hệ dịch máy

1.2.2 Mô hình dịch máy thống kê

a. Khảo sát phương pháp dịch máy thống kê

Dịch máy dựa trên phương pháp thống kê đang là một hướng phát triển đầy tiềm năng bởi những ưu điểm vượt trội so với các phương pháp khác. Thay vì xây dựng các từ điển, các quy luật chuyển đổi bằng tay, hệ dịch này tự động xây dựng các từ điển, các quy luật dựa trên kết quả thống kê có được từ các kho ngữ liệu. Chính vì vậy, dịch máy dựa vào thống kê áp dụng được cho bất kỳ cặp ngôn ngữ nào.

Mô hình chung của hệ dịch máy bằng phương pháp thống kê như sau:



Hình 1.2: Mô hình chung hệ dịch máy thống kê Việt – Nhật

b. Chu kì phát triển của hệ thống dịch thống kê

c. Ưu điểm của phương pháp dịch thống kê

1.2.3 Tiếp cận dịch máy dựa trên cụm từ phân cấp

a. Các nghiên cứu đã được công bố

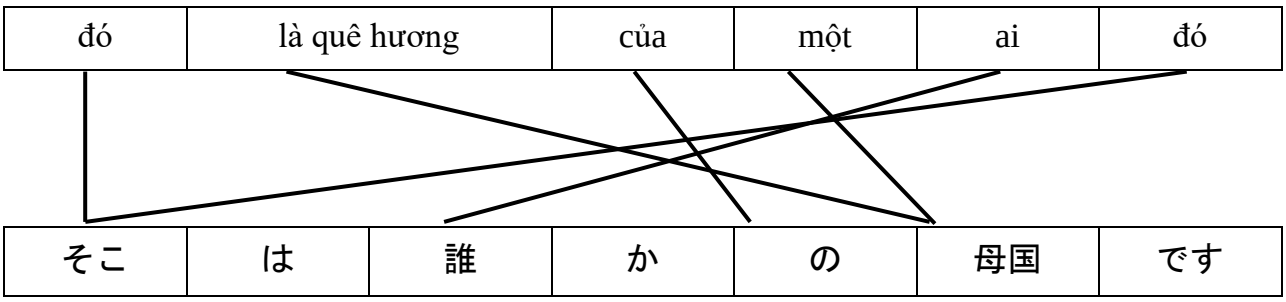
Mô hình dịch máy thống kê dựa trên cụm từ

Trong phương pháp dịch máy thống kê truyền thống dựa trên đơn vị từ, đơn vị được dịch là các từ. Số từ trong câu được dịch là khác nhau phụ thuộc vào các từ ghép, hình thái từ và thành ngữ. Tham số độ dài của chuỗi từ được dịch gọi là độ hỗn loạn, tức là số từ của ngôn ngữ đích mà từ của ngôn ngữ nguồn sinh ra. Tuy nhiên với cặp ngôn ngữ Việt – Nhật, hệ dịch phải đối mặt với khó khăn trong quá trình sắp xếp trật tự của các từ tiếng Việt tương ứng khi dịch sang câu tiếng Nhật. Mô hình dịch dựa trên đơn vị từ không cho kết quả tốt trong trường hợp kết nối nhiều-1 hoặc nhiều-nhiều với trật tự các từ trong câu tương ứng là khác nhau. Khi đó, mô hình dịch dựa trên đơn vị cụm từ phần nào đối phó với sự thiếu hụt này của mô hình dựa trên từ. Chúng ta phân rã cum từ thành cá đoạn nhỏ p(f|e) thành:

$$p\left(\bar{f}_1^I | \overline{fe}_1^I\right) = \prod_{i=1}^{I} \varphi(\bar{f}_i|\bar{e}_i)\mathrm{d}(start_i - end_{i-1} - 1)$$

Các cụm từ trong kỹ thuật này thường không theo nghĩa ngôn ngữ học mà là các cụm từ được tìm thấy bằng cách sử dụng phương pháp thống kê để trích rút từ các cặp câu.

Ví dụ:



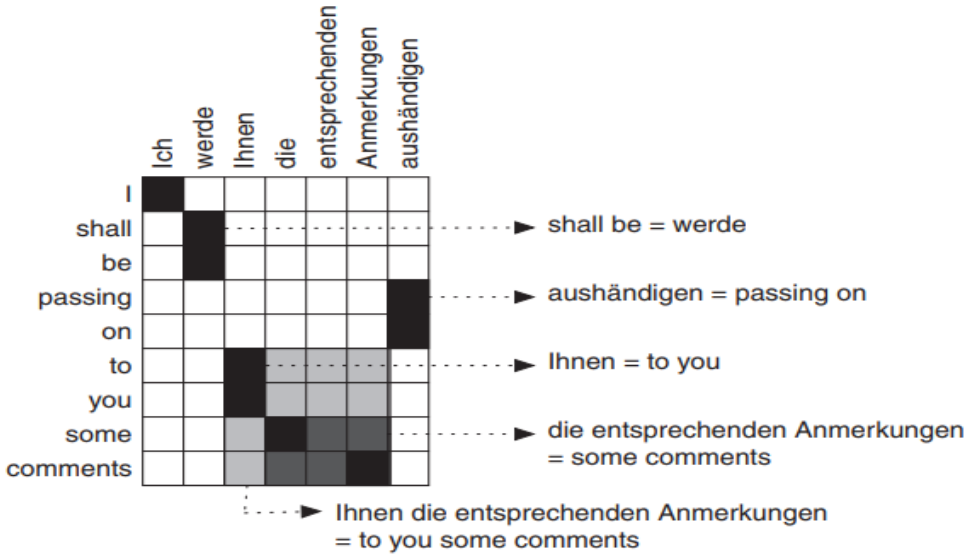
Hình 1.3: Ví dụ về gióng hàng từ

Ở đây, các cụm từ này được sinh ra dựa vào các phương pháp thống kê áp dụng trên ngữ liệu học. Trong "Introduction to Statistical Machine Translation", 2004, Koehn mô tả một cách khái quát quá trình dịch thống kê dựa trên cụm từ như sau:

- Câu nguồn được tách thành các cụm từ
- Mỗi cụm từ được dịch sang ngôn ngữ đích
- Các cụm từ đã dịch được sắp xếp lại theo một thứ tự phù hợp

b. Tiếp cận dịch máy dựa trên đơn vị cụm từ phân cấp

Xem Hình 1.4 để minh hoạ phương pháp cho các mô hình dựa trên cụm từ truyền thống. Cho một ma trận gióng hàng từ của một cặp câu song ngữ, chúng tôi trích xuất tất cả các cặp cụm từ phù hợp với gióng hàng. Những cặp cụm từ này là các quy tắc dịch trong các mô hình dựa trên cụm từ. Có nhiều cách khác nhau để ước lượng các xác suất dịch cho chúng. Ví dụ như xác suất có điều kiện φ $(\bar{e}|\bar{f})$ dựa trên tần số tương đối của cặp câu $(\bar{e}|\bar{f})$ và cụm từ $\bar{f}$ trong văn thể.



Hình 1.4: Trích xuất các quy tắc dịch cụm từ truyền thống

Tất cả các cặp cụm từ dịch máy truyền thống đều tạo thành các quy tắc cho ngữ pháp đồng bộ. Như đã thảo luận, đây là các quy tắc chỉ có các ký tự kết thúc ở phía bên phải:

$$Y \rightarrow \bar{f}|\bar{e}$$

1.3 Vấn đề tên riêng và từ mượn trong dịch máy

1.3.1 Vấn đề tên riêng

Sự quan tâm đến việc chuyển ngữ tự động tên riêng đã tăng lên trong những năm gần đây nhờ có khả năng giúp chống gian lận chuyển ngữ (The Economist Technology Quarterly, 2007), quá trình chuyển ngữ một cách chậm chạp của một tên riêng sẽ tránh bị truy vết bởi cơ quan thực thi pháp luật và cơ quan tình báo.

Марков ← Markov

spam, spammer ← спам, спамеров

Hình 1.6: Ví dụ chuyển ngữ tên riêng tiếng Nga - Anh

Khả năng chuyển ngữ tên riêng cũng có các ứng dụng trong dịch máy thống kê. Các hệ thống dịch máy thống kê được huấn luyện bằng các tập văn thể song song lớn, trong khi những tập văn thể này có thể bao gồm vài triệu từ mà họ không bao giờ có thể hy vọng sẽ có phạm vi bao phủ hoàn chỉnh, đặc biệt là đối với các lớp từ có hiệu suất cao như tên riêng.

1.3.2 Từ mượn

Theo thống kê, đến đầu những năm 1990, số lượng từ mượn chiếm 13,5% vốn từ vựng tiếng Nhật. Hiện nay, các từ mượn chiếm một vị trí quan trọng trong đời sống ngôn ngữ của người Nhật Bản. Các từ liên quan đến lĩnh vực kinh tế, chính trị và đời sống sinh hoạt hàng ngày đang tăng lên.

1.4 Bài toán luận văn giải quyết

Trong khóa luận này trình bày những vấn đề sau

- Đầu vào của bài toán là một chuỗi tiếng Việt bất kỳ
- Nghiên cứu mô hình dịch máy thống kê dựa trên cụm từ phân cấp, mô hình ngôn ngữ và giải mã để dịch các chuỗi từ tiếng Việt sang tiếng Nhật
- Sử dụng mô hình chuyển ngữ không giám sát xử lý tên riêng mà mô hình dịch không đưa ra được kết quả
- Từ đó kết quả sau chuyển ngữ se được cập nhật trở lại bản dịch ban đầu.

1.5 Kết cấu luận văn

Ngoài phần mở đầu và phần tham khảo, luận văn này được tổ chức thành 5 chương với các nội dung chính như sau:

- Chương 1: Giới thiệu
- Chương 2: Dịch máy thống kê dựa vào cụm từ phân cấp
- Chương 3: Dịch tên riêng và chuyển ngữ
- Chương 4: Thực nghiệm và đánh giá
- Chương 5: Kết luận

CHƯƠNG 2. DỊCH MÁY THỐNG KÊ DỰA TRÊN CỤM TỪ PHÂN CẤP

2.1. Ngữ pháp

Mô hình được dựa trên một văn phạm phi ngữ cảnh đồng bộ

2.1.1. Văn phạm phi ngữ cảnh đồng bộ

Trong một văn phạm phi ngữ cảnh đồng bộ các thành phần cấu trúc cơ bản được viết lại quy tắc với các cặp gióng hàng phía bên phải:

$$X \rightarrow (\gamma, \alpha, \sim)$$

Trong đó X là một kí tự không kết thúc, cả γ và α là chuỗi kí tự kết thúc và kí tự không kết thúcs, và ~ là đối xứng 1-1 giữa các biến cố kí tự không kết thúc trong γ một biến cố kí tự không kết thúc trong α. Ví dụ, ta có chuỗi tiếng Trung "Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi" được chuyển ngữ sang tiếng Anh là "Australia is one of the few countries that have diplomatic relations with North Korea". Các cặp cụm theo phân cấp có thể được biểu diễn bằng văn phạm phi ngữ cảnh đồng bộ như sau:

[Aozhou] [shi] [yu Beihan] [you] [bangjiao] [de shaoshu guojia zhiyi] [.]

[Australia] [has] [dipl. rels.] [with North Korea] [is] [one of the few countries] [.]

$$X \rightarrow \langle \text{yu } X_{[1]} \text{ you } X_{[2]}, \text{have } X_{[2]} \text{ with } X_{[1]} \rangle$$

$$X \rightarrow \langle X_{[1]} \text{ de } X_{[2]}, \text{the } X_{[2]} \text{ that } X_{[1]} \rangle$$

$$X \rightarrow \langle X_{[1]} \text{ zhiyi}, \text{one of } X_{[1]} \rangle$$

Trong đó các biến mà chúng tôi đã sử dụng các chỉ số trong hộp để chỉ ra những sự kiện không liên quan được kết nối bởi dấu "~". Các cặp cụm từ thông thường sẽ được chính thức hoá như sau:

$$X \rightarrow \langle \text{Aozhou}, \text{Australia} \rangle$$

$$X \rightarrow \langle \text{Beihan}, \text{North Korea} \rangle$$

$$X \rightarrow \langle \text{shi}, \text{is} \rangle$$

$$X \rightarrow \langle \text{bangjiao}, \text{diplomatic relations} \rangle$$

$$X \rightarrow \langle \text{shaoshu guojia}, \text{few countries} \rangle$$

Thêm hai luật để hoàn thiện ví dụ của chúng ta:

$$S \rightarrow \langle S_{[1]} X_{[2]}, S_{[1]} X_{[2]} \rangle$$

$$S \rightarrow \langle X_{[1]}, X_{[1]} \rangle$$

Một dẫn xuất văn phạm phi ngữ cảnh đồng bộ bắt đầu bằng một cặp ký hiệu bắt đầu kết nối. Tại mỗi bước, hai kết nối không liên kết được viết lại bằng cách sử dụng hai thành phần của một quy tắc. Khi biểu thị kết nối với các chỉ số đóng hộp, chúng ta phải liên tục ghi mục lục các ký hiệu mới được đưa ra từ các ký hiệu hiện có.

$\langle S_{[1]}, S_{[1]} \rangle$

$\Rightarrow \langle S_{[2]} X_{[3]}, S_{[2]} X_{[3]} \rangle$

$\Rightarrow \langle S_{[4]} X_{[5]} X_{[3]}, S_{[4]} X_{[5]} X_{[3]} \rangle$

$\Rightarrow \langle X_{[6]} X_{[5]} X_{[3]}, X_{[6]} X_{[5]} X_{[3]} \rangle$

$\Rightarrow \langle$ Aozhou $X_{[5]} X_{[3]}$, Australia $X_{[5]} X_{[3]} \rangle$

$\Rightarrow \langle$ Aozhou shi $X_{[3]}$, Australia is $X_{[3]} \rangle$

$\Rightarrow \langle$ Aozhou shi $X_{[7]}$ zhiyi, Australia is one of $X_{[7]} \rangle$

$\Rightarrow \langle$ Aozhou shi $X_{[8]}$ de $X_{[9]}$ zhiyi, Australia is one of the $X_{[9]}$ that $X_{[8]} \rangle$

$\Rightarrow \langle$ Aozhou shi yu $X_{[1]}$ you $X_{[2]}$ de $X_{[9]}$ zhiyi,
Australia is one of the $X_{[9]}$ that have $X_{[2]}$ with $X_{[1]} \rangle$

$\Rightarrow \langle$ Aozhou shi yu Beihan you $X_{[2]}$ de $X_{[9]}$ zhiyi,
Australia is one of the $X_{[9]}$ that have $X_{[2]}$ with North Korea$\rangle$

$\Rightarrow \langle$ Aozhou shi yu Beihan you bangjiao de $X_{[9]}$ zhiyi,
Australia is one of the $X_{[9]}$ that have diplomatic relations with North Korea$\rangle$

$\Rightarrow \langle$ Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi,
Australia is one of the few countries that have diplomatic relations with North Korea$\rangle$

Hình 2.1: Ví dụ trích xuất của văn phạm phi ngữ cảnh đồng bộ

2.1.2. Quy tắc trích xuất

Phần lớn ngữ pháp bao gồm các quy tắc trích xuất tự động. Quá trình trích xuất bắt đầu bằng một tập ngữ liệu huấn luyện được gióng hàng từ: một bộ ba (f, e, ~) trong đó f là một câu tiếng Việt, e là một câu tiếng Nhật, và ~ là một quan hệ nhị phân (nhiều - nhiều) giữa vị trí của f và vị trí của e. Các liên kết từ được tạo ra bằng cách chạy GIZA ++ (Och và Ney 2000) trên ngữ liệu huấn luyện theo cả hai hướng và tạo thành sự kết hợp của hai bộ gióng hàng từ.

Sau đó chúng ta trích xuất từ mỗi cặp câu đã gióng hàng từ một bộ quy tắc phù hợp với các gióng hàng. Ví dụ: giả sử ngữ liệu huấn luyện của chúng tôi chứa đoạn sau:

30 多年来　　的 友好　　合作

30 duonianlai　　de youhao hezou

30 plus-year-past　　of friendly cooperation

Friendly cooperation over the last 30 years

2.1.3. Các quy tắc khác

Quy tắc keo (Glue rules)

Quy tắc về thực thể (Intity Rules)

2.2. Mô hình

2.2.1. Định nghĩa

Theo Och và Ney (2002), chúng ta sử dụng một mô hình tuyến tính tổng quát cho các dẫn xuất D:

$$P(D) \propto \prod_i \varphi_i(D)^{\lambda_i}$$

Trong đó φi là các đặc trưng được định nghĩa trên dẫn xuất và λi có trọng số. Một trong những đặc trưng là một mô hình ngôn ngữ m-gram PLM (e); phần còn lại của các đặc trưng chúng ta sẽ định nghĩa là như là kết quả của các hàm trên các quy tắc được sử dụng trong một dẫn xuất:

$$\varphi_i(D) = \prod_{(X \to (\gamma, \alpha)) \in D} \varphi_i(X \to (\gamma, \alpha))$$

Như vậy chúng ta có thể viết lại P (D) như sau:

$$P(D) \propto P_{LM}(e)^{\lambda_{LM}} \times \prod_{i \neq LM} \prod_{(X \to (\gamma, \alpha)) \in D} \varphi_i(X \to (\gamma, \alpha))^{\lambda_i}$$

Các yếu tố khác ngoài yếu tố mô hình ngôn ngữ có thể được đưa vào một hình thức đặc biệt rõ ràng. Một văn phạm phi ngữ cảnh đồng bộ có trọng số là một văn phạm phi ngữ cảnh đồng bộ cùng với một hàm ω gán trọng số cho các quy tắc. Hàm này tạo ra một hàm trọng số trong các dẫn xuất:

$$\omega(D) = \prod_{(X \to (\gamma, \alpha)) \in D} \omega(X \to (\gamma, \alpha))$$

Nếu ta định nghĩa

$$\omega(X \to (\gamma, \alpha)) = \prod_{i \neq LM} \varphi_i(X \to (\gamma, \alpha))^{\lambda_i}$$

thì mô hình xác suất sẽ trở thành

$$P(D) \propto P_{LM}(e)^{\lambda_{LM}} \times \omega(D)$$

Rất dễ để viết các thuật toán lập trình động để tìm các bản dịch có trọng số cao nhất hoặc các bản dịch tốt nhất với một văn phạm phi ngữ cảnh đồng bộ có trọng số. Do đó vấn đề là $\omega(D)$ không bao gồm mô hình ngôn ngữ, điều này cực kỳ quan trọng đối với chất lượng bản dịch.

2.2.2. Các đặc trưng

Các quy tắc trích ra từ tệp huấn luyện có các tính năng sau:

- Các trọng số $P_\gamma(\gamma|\alpha)$ và $P_\gamma(\alpha|\gamma)$ ước lượng chất lượng những từ trong α dịch các từ trong γ (Koehn, Och, và Marcu 2003)
- Một điểm phạt exp(-1) đối với các quy tắc rút gọn, tương tự như điểm phạt cụm từ của Koehn (Koehn 2003), cho phép mô hình học cách ưu tiên các dẫn xuất dài hơn hoặc ngắn hơn

Tiếp theo, có các điểm phạt exp (-1) cho các lớp khác nhau của các quy tắc:

- Cho quy tắc keo, để mô hình có thể học một ưu tiên cho các cụm từ phân cấp trên một chuỗi kết hợp của các cụm từ

- Cho bốn loại quy tắc (số, ngày tháng, tên, từng dòng) được chèn vào bởi các mô-đun dịch chuyên ngành, để mô hình có thể học được độ tin cậy trong số đó

2.2.3. Huấn luyện

2.3. Giải mã

Thuật toán cơ bản

2.4. Đánh giá chất lượng dịch

CHƯƠNG 3. DỊCH TÊN RIÊNG VÀ CHUYỂN NGỮ

3.1. Dịch tên riêng

3.1.1. Giới thiệu

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 không xác định", ở đây chúng tôi xem từ không xác định như là một phần hiển nhiên của dịch máy và tìm cách dịch lại các không xác định này để cải tiến chất lượng dịch máy chung cuộc. Việc phân đoạn từ làm tăng chất lượng dịch chung cuộc nhưng lại xuất hiện nhiều từ không xác định ở bản dịch đích do ngữ liệu huấn luyện ở trường hợp này ít từ vựng hơn khi chưa phân đoạn từ.

Phần lớn các từ không xác định trong dịch thống kê Việt-Nhật là tên riêng. Tên riêng 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).

3.1.2. Phương pháp tiếp cận mô-đun

Sau khi những thử nghiệm ban đầu cùng các dòng này, chúng ta xây dựng một mô hình động của quá trình chuyển ngữ:

1. Một cụm từ tiếng Việt được viết ra.
2. Một máy dịch/người dịch phát âm nó bằng tiếng Việt.
3. Cách phát âm được sửa đổi để phù hợp với bản âm thanh tiếng Nhật.
4. Các âm được chuyển đổi sang katakana.
5. Katakana được viết

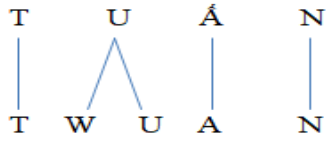
Việc phân chia bài toán của chúng ta thành 5 bài toán nhỏ. May mắn thay, có những kỹ thuật để phối hợp các giải pháp cho các bài toán nhỏ như thế. Khác với các ngôn ngữ khác trên thế giới, phát âm và cách viết tiếng Việt có sự tương đồng. Do đó chúng ta sẽ nghiên cứu bài toán 3, 4 và 5. Các kỹ thuật này dựa trên xác suất và định lý Bayes.

3.1.3. Các mô hình xác suất

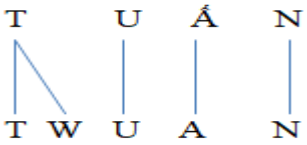
Âm tiếng Việt sang âm tiếng Nhật

Automat hữu hạn trọng số được học tự động từ các cặp chuỗi tiếng Việt - Nhật, ví dụ ((rượu nếp) <-> (mochigome)). Chúng tôi có thể tạo ra các cặp bằng cách thao tác bản chú giải thuật ngữ tiếng Việt – katakana. Chúng tôi sau đó có thể áp dụng thuật toán Ước lượng tối đa hóa (estimation-maximization (EM)) (Baum 1972; Dempster, Laird, and Rubin 1977) để tạo xác suất kí tự nối. Thuật toán EM của chúng tôi diễn giải như sau:

1. Với mỗi cặp chuỗi tiếng Việt - Nhật, tính tất cả các sự sắp xếp có thể có giữa các thành phần của chúng. Trong trường hợp của chúng tôi, một sự sắp xếp là một bản vẽ kết nối mỗi âm tiếng Việt với một hoặc nhiều âm tiếng Nhật, chẳng hạn tất cả các âm tiếng Nhật được bao phủ và không có đường đi qua. Ví dụ, có 2 cách để sắp xếp các cặp "Tuấn" <-> "twuan":



hoặc



Trong trường hợp này, sự xắp xếp bên trái bằng trực giác thích hợp hơn.

2. Với mỗi cặp, gán một trọng số bằng nhau với mỗi cách sắp xếp của chúng, như vậy tổng trọng số = 1. Trong trường hợp trên, mỗi cách sắp xếp đưa ra trọng số 0.5.

3. Mỗi âm trong âm tiếng Việt, đếm sự thể hiện của các kết nối khác nhau giữa chúng, như quan sát thấy sự sắp xếp của tất cả các cặp. mỗi sự sắp xếp đóng góp số lượng tương xứng với trọng số của nó.

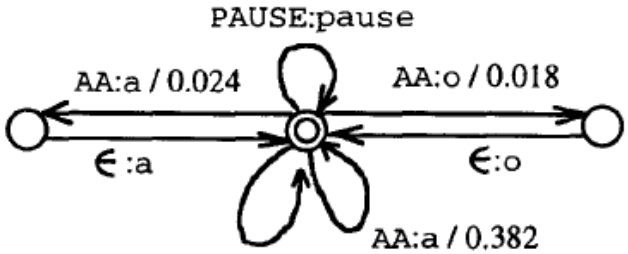
4. Với mỗi âm tiếng Việt, chuẩn hóa trọng số của các chuỗi tiếng Nhật nó kết nối tới, vì vậy tổng điểm = 1.

5. Tính lại các điểm số liên kết. mỗi liên kết được tính với kết quả của các điểm số của sự kết nối kí tự mà nó chứa.

6. Chuẩn hóa các điểm liên kết. các điểm cho mỗi cặp sắp xếp nên có tổng =1.

7. Lặp lại bước 3-6 đến khi xác suất kí tự liên kết hội tụ.

Chúng tôi sau đó xây dựng trực tiếp một mô hình automat hữu hạn có trọng số từ xác suất kí tự liên kết:



v	j	P(j|v)	v	j	P(j|v)	v	j	P(j|v)	v	j	P(j|v)
A	a aa ai	0.566 0.328 0.018	B	b bu	0.802 0.185	C	k ku	0.671 0.257	D	d j z	0.535 0.329 0.032
AO	ao oo a	0.671 0.257 0.047	G	g gu	0.598 0.304	CH	ch d chi	0.277 0.189 0.169	H	h w	0.959 0.014
I	i e	0.908 0.071	K	k ku ki	0.528 0.238 0.015	L	r ru	0.621 0.362	M	m mu	0.652 0.207
N	n	0.978	NG	ng ngu u	0.743 0.220 0.023	T	t to ch	0.462 0.305 0.043	TH	th t ch	0.418 0.303 0.043

Bảng 3.1: Ánh xạ một số âm tiếng Việt (Viết hoa) với âm tiếng Nhật (viết thường) sử dụng thật toán EM

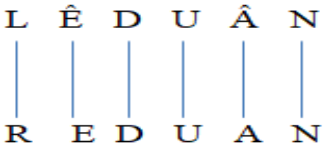
Các âm tiếng Việt (trong chữ viết hoa) với xác suất liên kết với các chuỗi âm tiếng Nhật (chữ viết thường), được học bởi ước lượng tối đa hóa (EM). Chỉ có các liên kết với xác suất điều kiện tốt hơn 1% được hiển thị, vì vậy tổng các con số có thể không = 1.

Chúng tôi cũng xây dựng các mô hình cho phép các âm tiếng Việt độc lập bị "rút đi" (ví dụ tạo ra 0 âm tiếng Nhật). tuy nhiên, các mô hình này tính toán tốn kém (nhiều sự sắp xếp hơn) và dẫn đến một số lượng lớn giả thuyết trong thành phần automat. Hơn nữa, trong việc không cho phép "nuốt", chúng tôi có thể tự động xóa hàng trăm cặp có khả năng gây hại từ tập huấn luyện của chúng tôi. Bởi vì không có sự sắp xếp nào là có thể, như các cặp bị bỏ qua bởi thuật toán học, các trường hợp như này đều phải được giải quyết bởi việc tra từ điển bằng mọi cách.

Chú ý rằng, mô hình của chúng tôi dịch mỗi âm tiếng Việt mà không liên quan đến ngữ cảnh. Chúng tôi cũng xây dựng các mô hình dựa vào ngữ cảnh, sử dụng cây quyết định mã hóa lại như automat hữu hạn có trọng số. Ví dụ, một từ âm "T" trong tiếng Việt có khả năng ra là (t o) hơn là (t). tuy nhiên, các mô hình dựa trên ngữ cảnh không thuận lợi cho việc chuyển ngữ ngược. chúng hữu ích hơn cho việc chuyển ngữ từ tiếng Việt sang tiếng Nhật.

Lê Duân

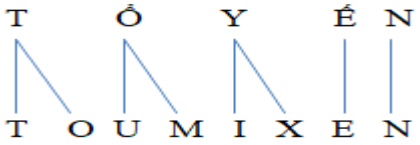
Chuỗi âm tiếng Việt:



Chuỗi âm tiếng Nhật:

Tổ yến

Chuỗi âm tiếng Việt:

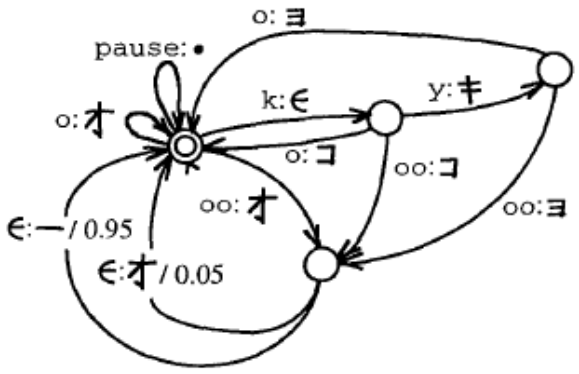


Chuỗi âm tiếng Nhật:

Hình 3.1: Giống hàng từ tiếng Việt – Nhật sử dụng thuật toán EM

Âm tiếng Nhật sang Katakana

Để liên kết các chuỗi âm tiếng Nhật như "m o o t a a" với chuỗi âm katakana như "モーター", chúng tôi thường xây dựng hai automat hữu hạn có trọng số. Kết hợp cùng nhau, chúng tạo ra một automat được tích hợp với 53 trạng thái và 303 cung, tạo ra một bản tóm tắt katakana chứa 81 kí tự, bao gồm dấu chấm phân cách (**.**). Automat đầu tiên kết hợp đơn giản nguyên âm dài tiếng Nhật với các kí tự mới aa, ii, uu, ee và oo. Automat thứ hai nối âm tiếng Nhật với các kí tự katakana. Ý tưởng cơ bản là giảm bớt toàn bộ phần âm tiết của âm thanh trước khi tạo ra bất kỳ kí tự katakana nào. Ví dụ:



Đoạn này cho thấy một sự biến thể theo chính tả trong tiếng Nhật: âm nguyên âm dài oo thường được viết với một dấu nguyên âm dài オー nhưng thi thoảng được viết với kí tự katakana lặp オオ.

3.2. Mô hình chuyển ngữ không giám sát

3.2.1. Giới thiệu

Mô hình chuyển ngữ không giám sát được đào tạo riêng rẽ nằm ngoài dòng chảy dịch máy, để thay thế các tên riêng bằng một chuyển ngữ tốt nhất trong bước tiền/hậu xử lý giải mã thường được sử dụng.

3.2.2. Khai phá chuyển ngữ

Mô hình khai phá chuyển ngữ là một tổng hợp của hai mô hình con: một chuyển ngữ và một không chuyển ngữ. Ý tưởng là mô hình chuyển ngữ sẽ chỉ định xác suất cao hơn cho các cặp chuyển ngữ so với xác suất được chỉ định bởi một mô hình không chuyển ngữ cho các cặp giống nhau. Xem xét một cặp từ (f, e), xác suất mô hình phiên âm cho cặp từ được định nghĩa như sau:

$$P_{tr}(f,e) = \sum_{a \in Align(f,e)} \prod_{j=1}^{|a|} p(q_j)$$

trong đó Align (f,e) là tập hợp của tất cả các chuỗi của gióng hàng từ, a là một chuỗi gióng hàng và q_j là một ký tự gióng hàng.

Mô hình không chuyển ngữ đề cập đến các cặp từ không có mối quan hệ ký tự. Nó được mô phỏng bằng cách nhân các kí tự nguồn và đích trong mô hình unigram:

$$P_{ntr}(f,e) = \prod_{i=1}^{|f|} P_F(f_i) \prod_{i=1}^{|e|} P_E(e_i)$$

Mô hình khai phá chuyển ngữ được định nghĩa là một phép nội suy của mô hình chuyển ngữ con và mô hình không chuyển ngữ con:

$$P(f,e) = (1 - \lambda) P_{tr}(f,e) + \lambda P_{ntr}(f,e)$$

- λ là xác suất đầu tiên của không chuyển ngữ

Mô hình không chuyển ngữ không thay đổi trong quá trình huấn luyện. Chúng tôi tính toán nó trong bước tiền xử lý. Mô hình chuyển ngữ học cách gióng hàng từ bằng cách sử dụng thuật toán EM.

3.2.3. Mô hình chuyển ngữ

Bây giờ chúng ta có cặp từ chuyển ngữ để học một mô hình chuyển ngữ. Chúng tôi phân đoạn tập ngữ liệu đào tạo thành các ký tự và tìm hiểu một hệ thống dựa trên cụm từ trên các cặp ký tự. Mô hình chuyển ngữ giả định rằng các từ nguồn và đích được tạo ra một cách đơn điệu. Do đó chúng tôi không sử dụng bất kỳ mô hình gióng hàng nào. Chúng tôi sử dụng 4 tính năng dịch cụm từ cơ bản (trực tiếp, chuyễn ngữ truy hồi, và các tính năng trọng số), tính năng mô hình ngôn ngữ (được xây dựng từ phía ngôn ngữ đích của bộ ngữ liệu để học khai phá), và các điểm phạt từ và cụm từ.

3.2.4. Tích hợp với dịch máy

Chúng tôi đã thử nghiệm ba phương thức để tích hợp chuyển ngữ, được mô tả dưới đây.

Phương thức 1

Liên quan đến việc thay thế tên riêng trong đầu ra với số lượng bản dịch tốt nhất. Thành công của Phương thức 1 chỉ phụ thuộc vào độ chính xác của mô hình chuyển ngữ. Ngoài ra, nó bỏ qua bối cảnh có thể dẫn tới việc chuyển ngữ không chính xác. Ví dụ: từ بيل dịch thành "Bill" nếu sau đó là "Clinton" và "Bell" nếu trước đó là "Alexander Graham".

Phương thức 2

Phương thức 3

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1. Chuẩn bị dữ liệu đầu vào cho hệ dịch

Dữ liệu đầu vào là dữ liệu song ngữ Việt – Nhật: Sử dụng khoảng 40.000 cặp câu Việt – Nhật

Để chuẩn bị dữ liệu để đào tạo hệ thống chuyển ngữ, chúng ta phải thực hiện các bước sau:

- Tokenisation: Tách các từ và cụm từ trong chuỗi
- Truecasing: Các từ ban đầu trong mỗi câu được chuyển đổi sang phiên bản chắc chắn nhất của chúng. Điều này giúp giảm sự thưa thớt dữ liệu.
- Cleaning: Các chuỗi dài và các chuỗi trống sẽ được gỡ bỏ vì chúng có thể gây ra vấn đề với dòng huấn luyện, và rõ ràng là những câu sai lệch sẽ bị xóa.

4.2. Công cụ tiền xử lý

4.2.1. Môi trường triển khai phần cứng

4.2.2. Bộ công cụ mã nguồn mở Moses

4.2.3. GIZA ++

4.2.4. KenLM

4.2.5 Mert

4.2.6. BLEU

4.3. Tiến hành thực nghiệm

4.3.1. Dữ liệu đầu vào

Dữ liệu huấn luyện	Tiếng Việt	40000 câu	training_seg_40k.clean.vn
	Tiếng Nhật	40000 câu	training_seg_40k.clean.jp
Dữ liệu điều chỉnh tham số	Tiếng Việt	950 câu	tuning_seg_950.clean.vn
	Tiếng Nhật	950 câu	tuning_seg_950.clean.jp
Dữ liệu đánh giá	Tiếng Việt	1000 câu	testing_seg_1k.clean.vn
	Tiếng Nhật	1000 câu	testing_seg_1k.clean.jp

4.3.2. Quá trình chuẩn bị dữ liệu và huấn luyện

Chuẩn bị dữ liệu

- Tách từ cho các file dữ liệu đầu vào
- Cắt các câu dài cho 2 tệp dữ liệu huấn luyện
- Chuyển về chữ thường

Huấn luyện mô hình ngôn ngữ

Huấn luyện mô hình dịch

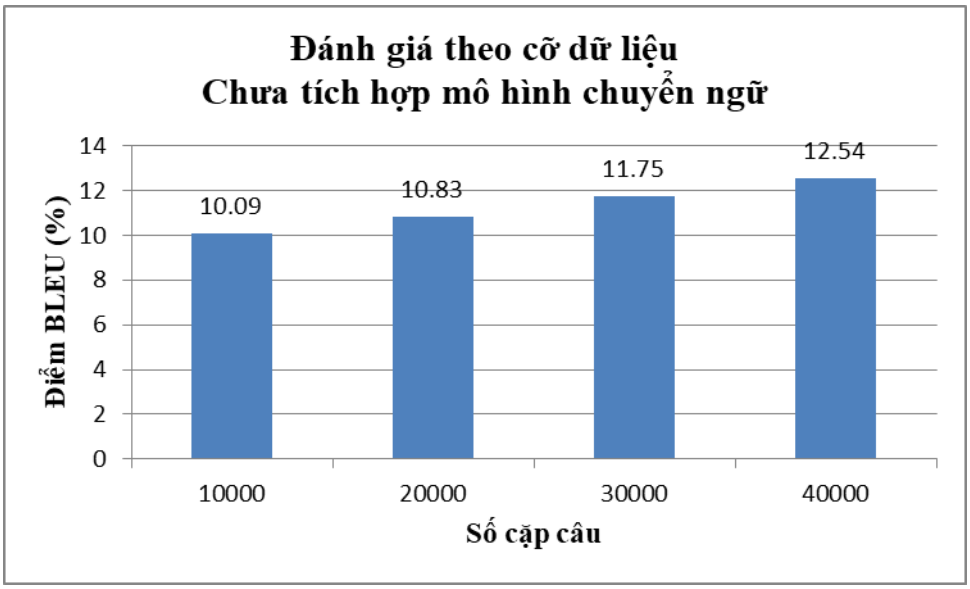
Huấn luyện tham số

4.3.3. Chạy thử

4.4. Đánh giá và phân tích kết quả theo cỡ dữ liệu huấn luyện

4.4.1. Kết quả khi chưa áp dụng mô hình chuyển ngữ

Ta thay đổi kích cỡ của ngữ liệu huấn luyện lần lượt là 10.000, 20.000,…, 40.000 cặp câu, sau đó thực hiện đánh giá chất lượng dịch dựa vào điểm BLEU. Điểm BLEU càng cao thì chất lượng dịch càng tốt.



Hình 4.1: Kết quả đánh giá chất lượng dịch khi chưa tích hợp mô hình chuyển ngữ

Tiếng Việt	Tiếng Nhật
alleyne đã phải nhập_viện sau khi bị bắt vì bị tức ngực .	alleyne で 逮捕 さ れ た 後 、 nhập_viện tức さ れ なけれ ば なら なかっ た ngực た 。
liên_hợp_quốc nói rằng tỷ_lệ tử_vong dân_thường ở nước này đã tăng vọt .	liên_hợp_quốc この 国 で の 民間 人 の 死亡 率 が vọt 上昇 し た と 言っ た 。
kể từ năm 1979 , có tất_cả bốn năm nóng hơn nhiệt_độ trung_bình ở úc .	4 年 、 1979 年 以来 、 全て が オーストラリア の 平均 より 暖かく nhiệt_độ た 。
vào tháng_một năm 2011 , mandela phải vào viện điều_trị bệnh nhiễm_trùng hô_hấp cấp_tính .	tháng_một に 病気 に điều_trị 研究所 は 、 2 0 1 1 年 、 mandela nhiễm_trùng hô_hấp cấp_tính た 。
220,000 người đã được sơ_tán khỏi các khu_vực trũng thấp ở tỉnh camaguey , trong khi 170,000 người đã được sơ_tán khỏi các tỉnh las_tunas .	220,000 camaguey 州 で 、 低 trũng 地域 から 避難 し た 人 の 人 が 170,000 las_tunas 州 から 避難 し た

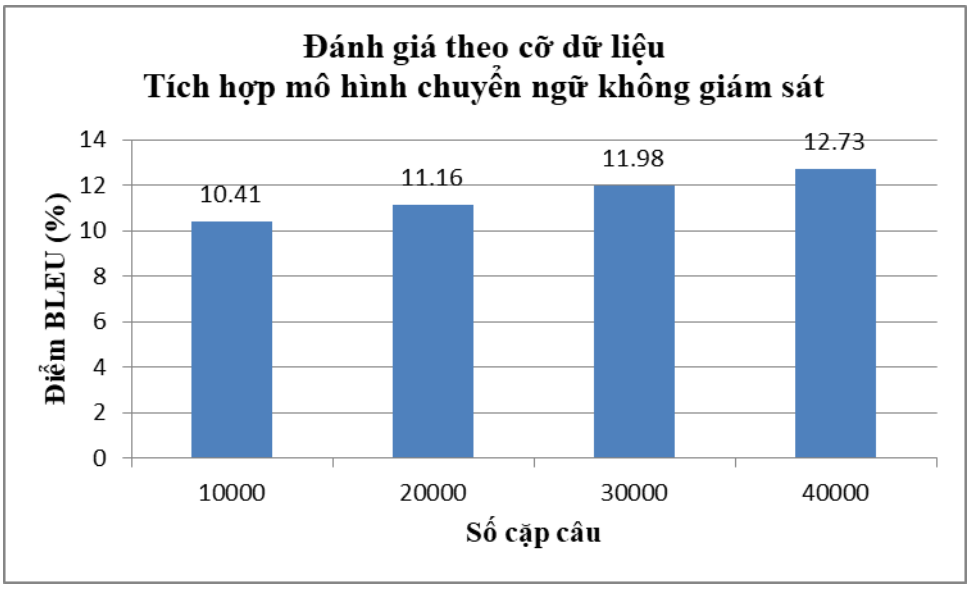
	。
lối vào moscow dành cho xe siêu_trọng sẽ bị hạn_chế đến tận 14 giờ chiều .	車 の ため の モスクワ に アクセス する siêu_trọng 14 時間 の 午後 まで に 制限 さ れ た 。

Bảng 4.1: Một số kết quả dịch từ tiếng Việt sang tiếng Nhật khi chưa tích hợp mô hình chuyển ngữ

4.4.2. Kết quả sau khi áp dụng mô hình chuyển ngữ không giám sát

Tương tự phần 4.4.1 chúng ta thay đổi kích cỡ của ngữ liệu huấn luyện lần lượt là 10.000, 20.000,…, 40.000 cặp câu, sau đó thực hiện đánh giá chất lượng dịch dựa vào điểm BLEU.

Chúng tôi đã kết hợp mô hình chuyển ngữ không giám sát vào mô hình dịch để chuyển các tên riêng mà mô hình dịch không dịch được. Chúng tôi áp dụng phương thức chuyển ngữ 1 trên cặp ngôn ngữ tiếng Việt - Nhật và cho thấy những cải tiến từ điểm BLEU 12.54 tăng lên 12.73.



Hình 4.2: Kết quả đánh giá chất lượng dịch tích hợp mô hình chuyển ngữ không giám sát

Tiếng Việt	Tiếng Nhật
đội_tuyển mỹ tuyên_bố sẽ chiến_đấu trong giải cá_nhân .	アメリカ チーム の 戦い は 個人 の シーズン の 中 で 述べ た 。
19,2 triệu đô_la úc đã được mở rộng cho uỷ_ban thể_thao úc để chạy chương_trình cộng_đồng hoạt_động sau giờ học của họ trong năm 2012/2013 với sự tài_trợ tương_tự cho 2013/2014 .	オーストラリア の スポーツ 委員 会 に 拡大 さ れ た が 、 オーストラリア の 19,2 万 ドル の 資金 援助 計画 を 実行 する ため に 、 彼ら の 活動 を 学ぶ 時間 後 に 2012/2013 年 の

	コミュニティ 2013/2014 に 類似 し た 。
tự_do ngôn_luận không_thể là lý_do để cho_phép bộ phim này , họ nói .	言論 の 自由 を 許可 する こと は でき ない 理由 は 、 この 映画 は 、 彼ら は 言っ た 。
tôi giữ quan_điểm mạnh nhất có_thể mà nói rằng điều đó là trái với lợi_ích an_ninh của đất_nước này khi nước mỹ bị đánh_bại ở iraq .	私 は 最も 強力 な 見解 を 保持 し て いる かも しれ ない と 述べ た が 、 それ は この 国 の 安全 保障 さ れ た 水 が アメリカ の 利益 と は 対照 的 に は イラク で 破っ た 。
trận động_đất này làm cho tổng_số người chết lên 30 và số người bị_thương hiện là 350 .	この 地震 は 、 合計 350 人 が 負傷 し 、 現在 の と は 30 人 が 死亡 し た 。

Bảng 4.2: Một số kết quả dịch từ tiếng Việt sang tiếng Nhật tích hợp mô hình chuyển ngữ không giám sát

Tên riêng tiếng Việt	**Tên riêng tiếng Nhật**
thủy ngân	水銀
thừa thiên huế	天空の王子
lê hoàng nam	ルプリンス
nguyễn thị diệp	グエンティディープ

Bảng 4.3: Một số kết quả dịch tên riêng tiếng Việt sang tiếng Nhật tích hợp mô hình chuyển ngữ không giám sát

CHƯƠNG 5. KẾT LUẬN

Các công việc đạt được của luận văn:

- Tìm hiểu tổng quan về hệ dịch máy đặc biệt là dịch máy thống kê dựa vào cụm từ phân cấp.
- Tìm hiểu tổng quan về mô hình chuyển ngữ không giám sát xử lý tên riêng.
- Tìm hiểu bộ công cụ mã nguồn mở Moses.
- Thử nghiệm mô hình chuyển ngữ không giám sát và cho kết quả tương đối khả quan.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Nguyễn Văn Vinh (2005). "Xây dựng chương trình dịch tự động Anh-Việt bằng phương pháp dịch thống kê". Luận văn Thạc sĩ, Đại học Công nghệ, ĐHQGHN.

[2] Nguyễn Thị Việt Thanh, 2000, Ngữ pháp tiếng Nhật. Nxb. Đại học Quốc gia Hà Nội.

Tiếng Anh

[3] Al-Onaizan, Y. and Knight, K. (2001). Translating named entities using monolingual and bilingual resources. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics

[4] D. Chiang (2005), A Hierarchical Phrase-Based Model for Statistical Machine Translation, In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05).

[5] Franz Joset Och and Hermann Ney (2002), Discriminative training and maximum entropy models for statistical machine translation, In Processdings of the 40th Anuual Meeting of the ACL, pages 295-302, Philadelphia, PA.

[6] Koehn, P (2004). Pharaoh: a beam search decoder for phrasebased. 2004.

[7] Chen, H., Yang, C., and Lin, Y. (2003). Learning Formulation and Transformation Rules for Multilingual Named Entities. Proceedings of ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models

[8] Koehn, P., et al. (2006). Moses: Open Source Toolkit for Statistical Machine Translation..

[9] Philipp Koehn, Franz Josef Och, Daniel Marcu (2003), "Statistical Phrase-Based Translation", In proceedings of NAACL.

[10] Koehn (2010). Statistical Machine Translation

[11] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38. 49 50 Bibliography Dijkstra, E. (1959). A note on two problems in connexion with graphs. Numerische Mathematik.

[12] http://translate.google.com

[13] http://www.statmt.org/moses/