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

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN TUẤN ANH

CHUYỂN NGỮ TỰ ĐỘNG TỪ TIẾNG VIỆT SANG TIẾNG NHẬT

Ngành : Công nghệ thông tin
Chuyên ngành : Kỹ thuật phần mềm
Mã số : 60480103

LUẬN VĂN THẠC SỸ

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS NGUYỄN PHƯƠNG THÁI

Hà Nội - 2017

LỜI CAM ĐOAN

Tôi xin cam đoan các kết quả nghiên cứu, thực nghiệm được trình bày trong luận văn này do tôi thực hiện dưới sự hướng dẫn của Phó giáo sư, Tiến sĩ Nguyễn Phương Thái.

Tất cả những tham khảo từ các nghiên cứu liên quan đều được nêu nguồn gốc một cách rõ ràng từ danh mục tài liệu tham khảo của luận văn. Trong luận văn, không có việc sao chép tài liệu, công trình nghiên cứu của người khác mà không chỉ rõ về tài liệu tham khảo.

TÁC GIẢ LUẬN VĂN

Nguyễn Tuấn Anh

LỜI CẢM ƠN

Trước tiên, tôi xin gửi lời cảm ơn sâu sắc nhất đến thầy giáo, Phó giáo sư, Tiến sĩ Nguyễn Phương thái đã tận tình hướng dẫn tôi trong suốt quá trình thực hiện luận văn tốt nghiệp.

Tôi xin bày tỏ lời cảm ơn chân thành tới trường Đại học Công Nghệ - ĐHQG Hà Nội và những thầy cô giáo tôi đã giảng dạy, truyền thụ kiến thức trong thời gian qua.

Cuối cùng, tôi xin cảm ơn tất cả gia đình, bạn bè đã luôn động viên giúp đỡ tôi trong thời gian nghiên cứu đề tài. Tuy đã có những cố gắng nhất định nhưng do thời gian và trình độ có hạn nên luận văn còn nhiều thiếu sót và hạn chế. Kính mong nhận được sự góp ý của thầy cô và các bạn.

TÁC GIẢ LUẬN VĂN

Nguyễn Tuấn Anh

MỤC LỤC

Danh mục hình vẽ

Danh mục bảng

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 … 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 đi vào đời sống con người, được ứng dụng rộng rãi. 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 riêng và các thuật ngữ kỹ thuật (các từ không xác định). 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.

Việc dịch các từ không xác định là một vấn đề khó do các ngôn ngữ thường khác nhau về bảng chữ cái và cách phát âm. Các từ này thường được chuyển ngữ, tức là, thay thế bằng khoảng ngữ âm gần đúng. Ví dụ: "Nguyễn Thu Trang" trong tiếng Việt xuất hiện dưới dạng " グエン テゥー チャン" (Guen tuu chan) trong tiếng Nhật.

1.1. Đặc điểm ngôn ngữ tiếng Việt và tiếng Nhật

1.1.1. Đặc điểm ngôn ngữ tiếng Việt[16]

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... Hiện nay, đây là phương thức chủ yếu để sản sinh ra các đơn vị từ vựng. Theo phương thức này, tiếng Việt triệt để sử dụng các yếu tố cấu tạo từ thuần Việt hay vay mượn từ các ngôn ngữ khác để tạo ra các từ, ngữ mới, ví dụ: tiếp thị, karaoke , xa lộ thông tin, siêu liên kết văn bản, truy cập ngẫu nhiên, …

Vốn từ vựng tối thiểu của tiếng Việt phần lớn là các từ đơn tiết (một âm tiết, một tiếng). Sự linh hoạt trong sử dụng, việc tạo ra các từ ngữ mới một cách dễ dàng đã tạo điều kiện thuận lợi cho sự phát triển vốn từ, vừa phong phú về số lượng, vừa đa

dạng trong hoạt động. Cùng một sự vật, hiện tượng, một hoạt động hay một đặc trưng, có thể có nhiều từ ngữ khác nhau biểu thị.

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 [3]

Ngữ âm[17]

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). Tiếng Nhật có số lượng âm tiết không lớn, có tất cả 112 dạng âm tiết. Trong số này, có 21 dạng âm tiết chỉ xuất hiện trong các từ được vay mượn từ nước ngoài.

Nếu như trong tiếng Việt, có rất nhiều từ được cấu tạo bởi một âm tiết, và mỗi âm tiết đều mang ý nghĩa nhất định, VD: bàn, trà, bạn, đèn..., thì đối với tiếng Nhật, phần lớn các từ được cấu tạo từ hai âm tiết trở lên và mỗi một âm tiết thường không mang ý nghĩa nào cả. VD: từ “hay” - “omoshiroi” có 5 âm tiết /o/mo/shi/ro/i, khó có thể tìm thấy ý nghĩa của mỗi âm tiết này. Cũng có những từ được cấu tạo bởi 1 âm tiết và trong trường hợp này, âm tiết mang ý nghĩa của từ đó, VD: “ki” có nghĩa là cái cây, “e” có nghĩa là bức tranh, “te” có nghĩa là cái tay... nhưng những từ như vậy chiếm số lượng rất nhỏ trong vốn từ vựng tiếng Nhật.[15]

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).

Trong tiếng Nhật, trọng âm cũng giữ một vị trí khá quan trọng. Trọng âm được thể hiện chủ yếu bằng độ cao khi phát âm, và nhờ có trọng âm mà nhiều từ đồng âm khác nghĩa được phân biệt. Ví dụ như từ "hashi" nếu phát âm cao ở âm tiết thứ nhất thì có nghĩa là "đôi đũa", nếu phát âm cao ở âm tiết thứ hai thì lại có nghĩa là "cây cầu". Tuy nhiên, các phương ngữ lại có sự phân bố trọng âm không giống nhau. Vì vậy, phương ngữ Tokyo đã được lấy làm ngôn ngữ chuẩn.

Từ vựng

Tiếng Nhật là một ngôn ngữ có một vốn từ vựng lớn và 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), kereraisu (cơm cà ri), uirusu (vi-rút).....

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 - 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

Theo các nhà khoa học, thế giới hiện nay có ít nhất 7099 ngôn ngữ khác nhau, với một số lượng ngôn ngữ lớn như vậy đã dẫn đến rất nhiều khó khăn, tốn kém trong việc trao đổi thông tin giữa các nước trên thế giới. Vì những khó khắn đó người ta đã phải dùng đến một đội ngũ phiên dịch viên khổng lồ, để dịch các văn bản, tài liệu, lời nói từ tiếng nước này sang tiếng nước khác. Để cải thiện vấn đề trên, người đã đề xuất thiết kế các mô hình tự động. Ngay từ những ngày đầu tiên xuất hiện máy vi tính, con người đã tiến hành nghiên cứu về dịch máy.

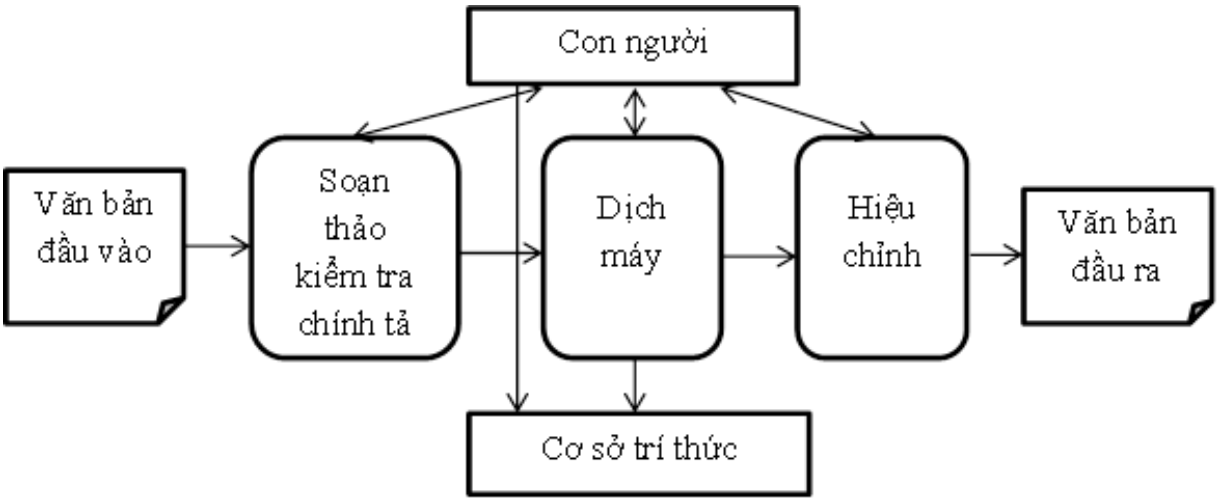
Dịch máy được coi là một trong những bài toán có ý nghĩa ứng dụng cao. Điều này là do dịch máy tiết kiệm thời gian, tiền bạc và công sức. Tuy nhiên, một hệ thống dịch máy không thể thay thế hoàn toàn công việc của người dịch vì máy không thể sản xuất ra bản dịch chất lượng cao hoàn toàn tự động. Do đó, hệ thống vẫn cần sự tương tác của con người trước, trong và sau quá trình dịch.

c. Sơ đồ tổng quan của một hệ dịch máy[1]

Đầu vào của một hệ dịch máy là một văn bản ở ngôn ngữ nguồn. Văn bản này có thể thu được từ một hệ soạn thảo hay một hệ nhận dạng chữ viết, lời nói. Sau đó

văn bản có thể được chỉnh sửa lại nhờ khối soạn thảo, kiểm tra chính tả, trước khi đưa vào máy dịch.

Phần mềm 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]

1.2.2 Mô hình dịch máy thống kê

Bài toán dịch máy đã được phát triển từ thập kỷ 50 và được phát triển mạnh mẽ từ thập kỷ 80. Hiện nay, có rất nhiều hệ dịch máy thương mại nổi tiếng trên thế giới như Systrans, Kant, … hay những hệ dịch máy mở tiêu biểu như hệ dịch của Google hỗ trợ hàng chục cặp ngôn ngữ phổ biến như Anh-Pháp, Anh-Trung, Anh-Nhật, … Các cách tiếp cận dịch máy chính bao gồm dịch dựa vào luật và dịch dựa vào xác suất thống kê. Các hệ dịch máy này đã đạt được kết quả khá tốt với những cặp ngôn ngữ tương đồng nhau về chữ cái và phát âm như các cặp ngôn ngữ Anh – Việt, Đức-Anh, … nhưng còn gặp nhiều hạn chế đối với các cặp ngôn ngữ có cú pháp khác nhau như Anh-Trung, Việt-Nhật, …

Hiện nay, các nghiên cứu để làm tăng chất lượng hệ dịch vẫn đang được tiến hành phù hợp với đặc điểm của các cặp ngôn ngữ. Ngoài ra, phương pháp dịch dựa trên mạng nơ-ron cũng là một hướng tiếp cận mới đang được phát triển mạnh với nhiều bước đột phá.

a. Khảo sát phương pháp dịch máy thống kê

Dịch máy thống kê dựa trên từ có nguồn gốc từ nghiên cứu của Brown (1993) người đã phát triển một mô hình kênh nhiễu dựa trên từ được dịch giống như bài báo của Knight và Graehl (1997) về mô hình chuyển ngữ.

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.

Dịch máy dựa trên phương pháp thống kê sẽ tìm câu ngôn ngữ đích e phù hợp nhất (có xác suất cao nhất) khi cho trước câu ngôn ngữ nguồn f.

$$\hat{e} = argmax_e p(e|f)$$

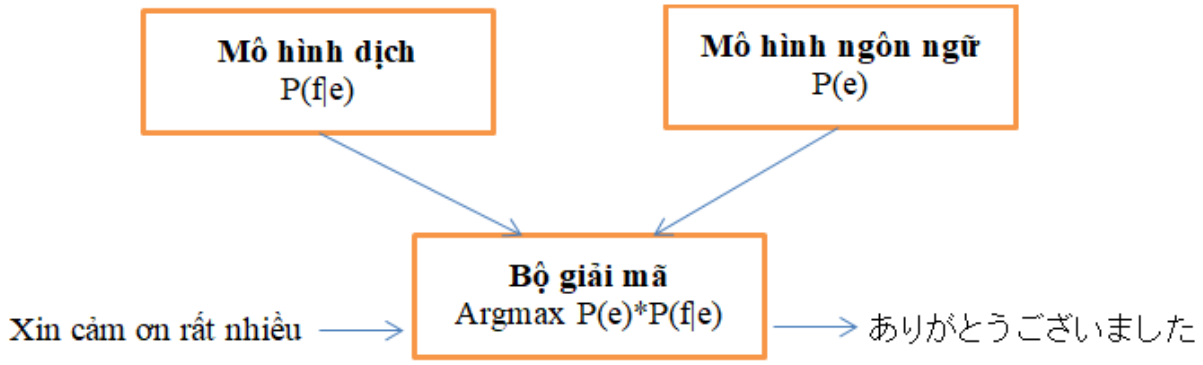
Mô hình dịch máy được Brown áp dụng vào bài toán như sau:

Giả sử cho câu tiếng Việt $f_1^J = f_1 \dots f_j \dots f_J$ cần dịch sang câu tiếng Nhật $e_1^I = e_1 \dots e_i \dots e_I$. Brown dựng lên mô hình kênh nhiễu với e là đầu vào bộ mã hoá (Encoder), qua kênh nhiễu được chuyển hoá thành f và sau đó, gửi f đến bộ giải mã (Decoder). Như vậy, trong các câu tiếng Nhật, ta chọn câu sao cho xác suất hậu nghiệm $\Pr(e_1^I|f_1^J)$ là lớn nhất, theo luật quyết định Bayes:

$$\hat{e}_1^I = \arg\max_{e_1^I} \Pr(e_1^I|f_1^J)$$

$$= \arg\max_{e_1^I} \{\Pr(e_1^I).\Pr(f_1^I|e_1^J)\}$$

Như vậy, ta có thể xây dựng mô hình chung của hệ dịch máy bằng phương pháp thống kê theo hình 1.2 như sau:



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

Mô hình ngôn ngữ thường được giải quyết bằng mô hình n-gram và mới đây là mô hình neuron.

Pha giải mã thường được giải quyết bằng các thuật toán Search như Viterbi Beam, A* stack, Graph Model.

Trong mô hình dịch, vấn đề trọng tâm của việc mô hình hoá xác suất dịch $\Pr(f_1^J|e_1^I)$ là việc định nghĩa sự tương ứng giữa các từ của câu nguồn với các từ của câu đích. Mô hình thực hiện việc đó gọi là mô hình gióng hàng từ.

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

Bước đầu tiên là tập hợp ngữ liệu huấn luyện. Ở đây, chúng ta cần thu thập các văn bản song ngữ, thực hiện việc dóng hàng câu và trích lọc ra các cặp câu phù hợp. Trong bước thứ hai, chúng ta thực hiện huấn luyện tự động hệ thống dịch máy. Đầu ra của bước này là hệ thống dịch máy có hiệu lực.

Tiếp theo hệ thống dịch máy được kiểm tra và việc phân tích lỗi được thực hiện. Dựa vào kiến trúc của hệ thống dịch máy thống kê, chúng ta có thể phân biệt các kiểu lỗi khác nhau: lỗi tìm kiếm, lỗi mô hình, lỗi huấn luyện, lỗi corpus huấn luyện và lỗi tiền xử lý.

Mô hình tốt hơn: Ở đây, mục tiêu là phải phát triển mô hình mà mô hình này mô tả càng nhiều các thuộc tính của ngôn ngữ tự nhiên và các tham số tự do của nó có thể được ước lượng từ ngữ liệu huấn luyện.

Huấn luyện tốt hơn: Thuật toán huấn luyện thường dựa vào cách tiếp cận hợp lý cực đại. Thông thường, các thuật toán huấn luyện thường cho ta kết quả là tốt ưu địa phương. Do vậy, để làm tốt việc huấn luyện này, cần xây dựng các thuật toán mà kết quả tối ưu địa phương thường gần với tối ưu toàn cục.

Tìm kiếm tốt hơn: Lỗi tìm kiếm xuất hiện nếu thuật toán không tìm kiếm ra câu dịch của câu nguồn. Vì vậy, chỉ có các cách tìm kiếm gần đúng để tìm ra câu dịch. Thuật toán hiệu quả là thuật toán mà cân bằng giữa chất lượng và thời gian.

Nhiều ngữ liệu huấn luyện hơn: Chất lượng dịch càng tăng khi kích thước của ngữ liệu huấn luyện càng lớn. Quá trình học của hệ thống dịch máy sẽ cho biết kích thước của ngữ liệu huấn luyện là bao nhiêu để thu được kết quả khả quan.

Tiền xử lý tốt hơn: Hiện tượng ngôn ngữ tự nhiên khác nhau là rất khó xử lý ngay cả trong cách tiếp cận thống kê tiên tiến. Do đó để cho việc sử dụng cách tiếp cận thống kê được tốt thì trong bước tiền xử lý, chúng ta làm tốt một số việc như: loại bỏ các kí hiệu không phải là văn bản, đưa các từ về dạng gốc của nó, ...

c. Ưu điểm của phương pháp dịch thống kê[1]

Cách tiếp cận thống kê có những ưu điểm sau:

Mối quan hệ giữa đối tượng ngôn ngữ như từ, cụm từ và cấu trúc ngữ pháp thường yếu và mơ hồ. Để mô hình hóa những phụ thuộc này, chúng ta cần một công thức hóa như đưa ra phân phối xác suất mà nó có thể giải quyết với những vấn đề phụ thuộc lẫn nhau.

Để thực hiện dịch máy, chúng ta nhất thiết phải kết hợp nhiều nguồn tri thức. Trong dịch thống kê, chúng ta dựa vào toán học để thực hiện kết hợp tối ưu của các nguồn tri thức.

Trong dịch máy thống kê, tri thức dịch được học một cách tự động từ ngữ liệu huấn luyện. Với kết quả như vậy, việc phát triển một hệ dịch dựa vào thống kê sẽ rất nhanh so với hệ dịch dựa vào luật.

Dịch máy thống kê khá phù hợp với phần mềm nhúng mà ở đây dịch máy là một phần của ứng dụng lớn hơn.

Việc đưa ra khái niệm "chính xác" của mối quan hệ ngữ pháp, ngữ nghĩa, văn phong là khó. Vì vậy, việc hình thức hóa vấn đề này càng chính xác càng tốt không thể dựa vào sự ràng buộc bởi các luật mô tả chúng. Thay vào đó, trong cách tiếp cận thống

kê, các giả định mô hình được kiểm định bằng thực nghiệm dựa vào ngữ liệu huấn luyện.

1.2.3. Tiếp cận dịch máy dựa trên cụm từ phân cấp

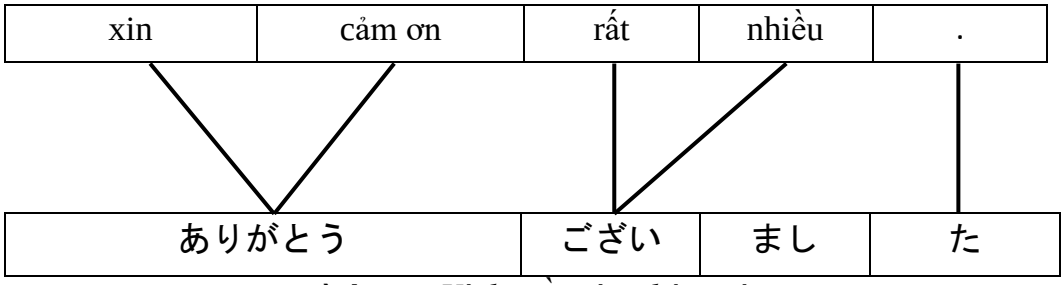
a. Các nghiên cứu đã được công bố

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, tuỳ vào đặc điểm của ngôn ngữ, như cặp ngôn ngữ Việt – Nhật cũng giống với cặp ngôn ngữ Anh-Trung, Anh-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. Trong quá trình dịch, kết nối từ tiếng Việt tương ứng với từ tiếng Nhật có thể là 1-1, 1-không, 1-nhiều, nhiều-1 hoặc nhiều-nhiều. 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 do Koehn và cộng sự (2003) phát triển 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

Phương pháp dịch máy thống kê dựa trên đơn vị cụm từ là phương pháp mới được phát triển, có một số mô hình đã được xây dựng và chất lượng được đánh giá là khá cao khi áp dụng cho các cặp ngôn ngữ như Anh-Trung, Anh-Arab, ... Tuy chất

lượng có tốt hơn mô hình dịch thống kê dựa trên đơn vị từ, mô hình dịch thống kê dựa trên cụm từ vẫn chưa giải quyết được một số vấn đề như ngữ pháp, khả năng lựa chọn cụm từ với tính chính xác cao, dịch tên, lượng từ vựng có hạn và các hạn chế chuyển đổi cú pháp.

Gióng hàng từ

Hiện nay, rất nhiều cách tiếp cận khác ra đời nhằm cải thiện chất lượng của hệ dịch, tích hợp thêm các thông tin ngôn ngữ như tiến hành tiền xử lý, sử dụng các thông tin về ngữ pháp để chuyển đổi câu ngôn ngữ nguồn f về một dạng f' gần với ngôn ngữ đích trước khi thực hiện việc gióng hàng từ

Giả sử, cho một chuỗi câu ngôn ngữ tiếng Việt f, mô hình sẽ cung cấp cho chúng ta xác suất p(e|f) của một câu tiếng Nhật e. Định lý Bayes được áp dụng cho phép chúng ta mô hình hóa xác suất bản dịch p(f|e) , đảm bảo rằng tiếng Nhật được tạo ra là một bản dịch phù hợp của câu tiếng Việt, và câu tiếng Nhật p(e) đảm bảo chuỗi tiếng Nhật đầu ra lưu loát:

$$p(e|f) = \frac{p(e)p(f|e)}{p(f)}$$

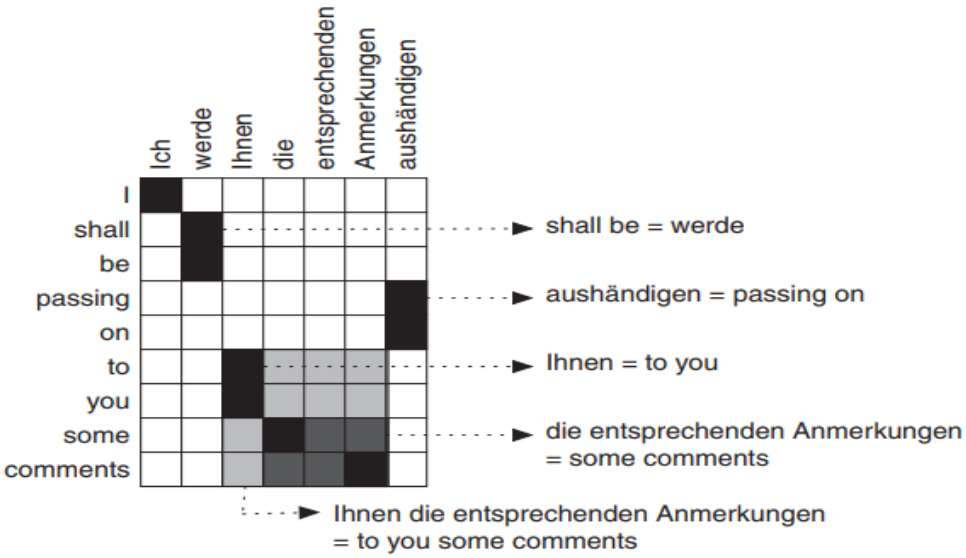
Xác suất của câu tiếng Việt p(f) có thể được loại bỏ vì nó là hằng số và sẽ không có bất kỳ ảnh hưởng nào đối với việc tìm kiếm câu tiếng Nhật e, tối đa hoá phương trình p(e)p(f|e):

$$\hat{e} = argmax_e p(e)p(f|e)$$

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

Trong phần này, chúng ta sẽ mô tả thiết kế và thực hiện mô hình dịch máy dựa trên cụm từ phân cấp và báo cáo về các thử nghiệm chứng minh rằng các cụm từ phân cấp thực sự cải thiện bản dịch.

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.

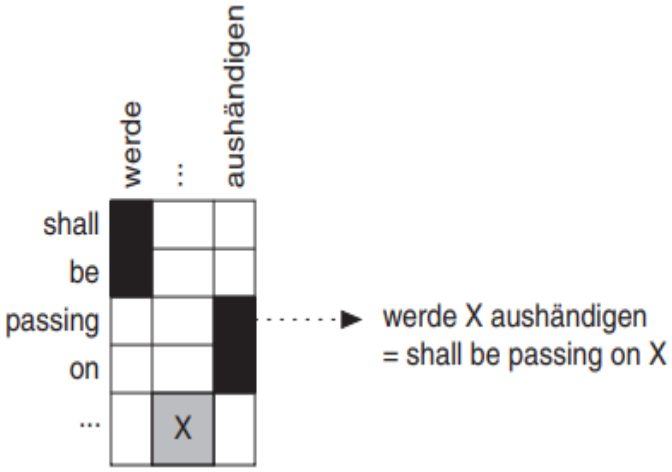


Hình 1.4: Trích xuất các quy tắc dịch cụm từ truyền thống

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:

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

Bây giờ chúng ta muốn xây dựng các quy tắc dịch phức tạp hơn, bao gồm cả các ký tự kết thúc và không kết thúc ở phía bên phải của quy tắc. Chúng ta học các quy tắc này như khái quát hóa các quy tắc từ ngữ truyền thống.



Hình 1.5: Trích xuất quy tắc dịch cụm từ phân cấp

Chúng tôi muốn học một quy tắc dịch cho cụm từ phức tạp của tiếng Đức "werde aushändigen". Tuy nhiên, các từ tiếng Đức "werde" và "aushändigen" không nằm cạnh nhau, chúng cách nhau bởi những từ can thiệp. Trong các mô hình dịch cụm từ truyền thống, chúng ta không thể học một quy tắc dịch chỉ có hai từ tiếng Đức như thế này, vì các cụm từ trong các mô hình truyền thống là các chuỗi từ liền kề nhau. Một quy tắc có chứa "werde" và "aushändigen" cũng bao gồm tất cả các từ can thiệp:

$$Y \rightarrow \begin{array}{l} \textit{werde Ihnen die entsprechenden Anmerkungen aushändigen} \\ \mid \textit{shall be passing on to you some comments} \end{array}$$

Bây giờ chúng ta thay thế các từ can thiệp bằng ký tự X. Tương ứng, ở phía tiếng Anh, chúng ta thay thế chuỗi từ tiếng Anh gióng hàng với những từ tiếng Đức can thiệp bằng ký tự X. Chúng ta tiếp tục trích ra quy tắc dịch

$Y \rightarrow \textit{werde X aushändigen} \mid$ shall we passing on X

Quy tắc này là một ngữ pháp đồng bộ với một hỗn hợp các ký tự X và các lý tự kết thúc (các từ) phía bên phải. Nó gói gọn một cách độc đáo kiểu sắp xếp lại khi tham gia dịch các cụm động từ tiếng Đức sang tiếng Anh.

Lưu ý rằng chúng tôi chưa giới thiệu bất kỳ ràng buộc cú pháp nào khác với nguyên tắc là ngôn ngữ đệ quy, và loại quy tắc dịch phân cấp phản ánh tính chất này. Trước tiên chúng ta phải xác định chính xác phương pháp trích xuất các quy tắc dịch theo cấp bậc.

Cho một chuỗi đầu vào **f** = $(f_1 \ldots, f_{l_f})$ và chuỗi đầu ra **e** = $(e_1, \ldots, e_{l_e})$ và một ánh xạ gióng hàng từ A, chúng ta trích xuất tất cả các cặp cụm từ $(\bar{e}, \bar{f})$ phù hợp với gióng hàng từ:

$$\begin{aligned} (\bar{e}, \bar{f}) \text{ phù hợp với A} \leftrightarrow \\ \forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f} \\ \text{AND } \forall f_j \in \bar{f} : (e_i, f_j) \in A \rightarrow e_i \\ \text{AND } \exists e_i \in \bar{e}, f_i \in \bar{f} : (e_i, f_j) \in A \end{aligned}$$

Cho P là tập hợp của tất cả các cặp cụm từ được trích xuất $(\bar{e}, \bar{f})$. Bây giờ chúng ta xây dựng các cặp từ phân cấp từ các cặp từ hiện tại. Nếu tồn tại một cặp cụm từ hiện tại $(\bar{e}, \bar{f}) \in$ P sẽ chứa một cặp cụm nhỏ hơn khác $(\bar{e}_{SUB}, \bar{f}_{SUB}) \in$ P, chúng ta sẽ thay thế cặp cụm nhỏ hơn bằng mộ ký tự X và thêm cặp cụm từ tổng quát hơn vào bộ P:

$$\begin{aligned} \text{if } (\bar{e}, \bar{f}) \in P \text{ AND } (\bar{e}_{\text{SUB}}, \bar{f}_{\text{SUB}}) \in P \\ \text{AND } \bar{e} = \bar{e}_{\text{PRE}} + \bar{e}_{\text{SUB}} + \bar{e}_{\text{POST}} \\ \text{AND } \bar{f} = \bar{f}_{\text{PRE}} + \bar{f}_{\text{SUB}} + \bar{f}_{\text{POST}} \\ \text{AND } \bar{e} \neq \bar{e}_{\text{SUB}} \text{ AND } \bar{f} \neq \bar{f}_{\text{SUB}} \\ \text{add } (e_{\text{PRE}} + \text{X} + e_{\text{POST}}, f_{\text{PRE}} + \text{X} + f_{\text{POST}}) \text{ to } P \end{aligned}$$

Tập hợp các cặp cụm từ phân cấp là kết thúc theo cơ chế mở rộng này. Lưu ý rằng nhiều thay thế của cụm từ nhỏ hơn cho phép tạo ra ánh xạ dịch với nhiều ký hiệu X. Điều này cho phép chúng tôi xây dựng các quy tắc dịch hữu ích như:

$$Y \rightarrow X_1 X_2 | X_2 \textit{ of } X_1$$

Một lưu ý về sự phức tạp của các quy tắc phân cấp được trích ra từ một cặp câu: vì một quy tắc có thể ánh xạ bất kỳ tập con nào của các từ đầu vào (có các ký hiệu không phải là ký tự đại diện cho các khoảng trống), có thể sử dụng một số quy tắc lũy thừa. Để tránh các bộ quy tắc có quy mô không thể quản lý và để giảm độ phức tạp

giải mã, chúng tôi thường muốn đặt các giới hạn về các quy tắc có thể có. Ví dụ, các giới hạn:

- Tối đa 2 kí hiệu không xác thực X
- Ít nhất một nhưng tối đa năm từ cho mỗi ngôn ngữ
- Khoảng tối đa 15 từ (tính cả khoảng trống)

Hạn chế các ký hiệu X làm giảm độ phức tạp của quy tắc trích xuất từ lũy thừa đến đa thức. Thông thường, chúng tôi cũng không cho phép các quy tắc có các ký hiệu X nằm cạnh nhau trong cả hai ngôn ngữ.

Mô hình dịch dựa trên cụm từ phân cấp, nhưng không có cú pháp rõ ràng, đã được chứng minh là vượt trội các mô hình dịch dựa trên cụm từ truyền thống trên một số cặp ngôn ngữ. Chúng dường như giải thích việc sắp xếp lại các từ và cụm từ nhất định, đặc biệt trong trường hợp các cụm từ không liên tục.

Chúng tôi đã trình bày một phương pháp học ngữ pháp đồng bộ dựa trên phương pháp mô hình dịch dựa trên cụm từ. Bắt đầu với một gióng hàng từ và chú thích cây phân tích cú pháp cho một hoặc cả hai ngôn ngữ, chúng tôi trích ra các quy tắc ngữ pháp.

Mô hình cụm từ phân cấp có ngữ pháp không xây dựng trên bất kỳ chú thích cú pháp nào. Nếu chú thích cú pháp chỉ tồn tại cho một trong các ngôn ngữ, chúng ta gọi nó là các quy tắc bán cú pháp. Đối với chú thích cú pháp cho trước, chúng ta cần phải tìm nút quản lý của mỗi cụm từ để có một nhãn không xác định duy nhất ở phía bên trái của các quy tắc. Các quy tắc được ghi bằng các phương pháp tương tự với các quy tắc được sử dụng cho các bản dịch cụm từ trong mô hình cụm từ (nghĩa là, xác suất có điều kiện của đầu ra, cho phía đầu vào).

1.2.4 Mô hình ngôn ngữ

Xác suất của một câu tiếng Nhật p(e) được tính bằng cách sử dụng một mô hình ngôn ngữ thống kê. Câu tiếng Nhật e được đại diện bởi chuỗi các từ $e_1, e_2, \dots, e_{l_e}$và xác suất của nó được phân tách bằng cách sử dụng luật chuỗi:

$$p(e) = p(e_1)p(e_2|e_1) \dots p(e_{l_e}|e_1, e_2, \dots, e_{l_e-1})$$

Trong cách dịch của chúng ta, có một tập hợp các từ và ta muốn lấy chúng ra theo một thứ tự hợp lý. Nhưng giả sử rằng chúng ta có nhiều tập hợp khác nhau, tương ứng là tập các nghĩa của cách dịch các từ ở tập hợp trên. Chúng ta có thể tìm thứ tự từ tốt nhất của mỗi tập hợp nhưng làm thế nào để chúng ta chọn câu của ngôn ngữ đích hợp lý nhất. Câu trả lời là chúng ta sử dụng mô hình n-gram, gán xác suất cho bất kì một dãy các từ có thể hiểu được. Sau đó chúng ta chọn ra dãy có thể nhất (xác suất cao nhất).

Khi chiều dài ngữ cảnh của một cụm từ tăng lên khả năng để nhìn thấy trước từ sau đó trong cụm giảm xuống. Để ước tính chính xác các tham số của mô hình chúng ta sử dụng giả định Markov cho biết rằng xác suất của một chuỗi nhất định có thể được

ước lượng tốt từ một lịch sử giới hạn. Thông thường, hai từ trước trong một câu được sử dụng để tạo thành một mô hình ngôn ngữ trigram:

$$p(e_1)p(e_2|e_1)\dots p\left(e_{l_e}|e_1,e_2,\dots,e_{l_e-1}\right) \approx\ p(e_1)p(e_2|e_1)\dots p\left(e_{l_e}|e_{l_e-2},e_{l_e-1}\right)$$

$$p(e) \approx \prod_{i=1}^{l_e} p(e_i|e_{i-2},e_{i-1})$$

Các xác suất được ước lượng thông qua các ước lượng khả năng tối đa, những ước tính này thường được làm phẳng để đảm bảo tất cả các chuỗi có thể có một xác suất không bằng không:

$$\mathrm{p}(\mathrm{e}_3|\mathrm{e}_1,\mathrm{e}_2) = \frac{\mathrm{count}(\mathrm{e}_1,\mathrm{e}_2,\mathrm{e}_3)}{\mathrm{count}(\mathrm{e}_1,\mathrm{e}_2)}$$

Như vậy, ta có thể coi toàn bộ các chủ đề về gán xác suất cho một câu được gọi là mô hình ngôn ngữ. Mô hình ngôn ngữ không chỉ có ích cho thứ tự các từ mà còn có ích cho việc chọn nghĩa giữa các cách dịch khác nhau.

1.2.5. Giới thiệu dịch máy mạng nơ-ron

Dịch máy mạng Nơ-ron là một phương pháp tiếp cận gần đây đang được sử dụng trong dịch máy được đề xuất bởi Kalchbrenner và Blunsom (2013). Không giống như hệ thống dịch dựa trên xác suất thống kê dựa vào từ, cụm từ bao gồm nhiều phần nhỏ được điều chỉnh riêng biệt, các phiên dịch máy mạng Nơ-ron cố gắng xây dựng và đào tạo một mạng nơ-ron lớn có thể đọc một câu và cho kết quả là một bản dịch chính xác.

Hầu hết các mô hình dịch máy mạng nơ-ron đều gồm bộ mã hóa-giải mã với bộ mã hoá và bộ giải mã cho mỗi ngôn ngữ hoặc liên quan đến một bộ mã hóa ngôn ngữ cụ thể được áp dụng cho mỗi câu có đầu ra sau đó được so sánh. Một mạng nơ ron mã hóa sẽ đọc và mã hóa câu nguồn thành một vec-tơ có độ dài cố định. Một bộ giải mã sau đó xuất ra một bản dịch từ vec-tơ mã hoá. Toàn bộ hệ thống mã hoá-giải mã, bao gồm bộ mã hóa và bộ giải mã cho một cặp ngôn ngữ, cùng nhau huấn luyện để tối đa hóa xác suất của một bản dịch chính xác.

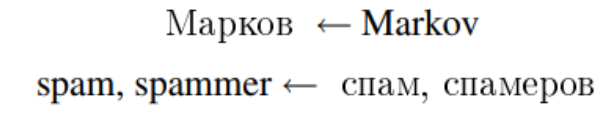
Tính năng khác biệt quan trọng nhất của phương pháp tiếp cận này từ bộ mã hóa-giải mã. Về cơ bản nó không cố mã hóa toàn bộ câu đầu vào thành một vec-tơ độ dài đơn. Thay vào đó, nó mã hóa câu đầu vào thành một dãy vec-tơ và chọn một tập con của các vec-tơ thích nghi trong khi giải mã bản dịch. Điều này giải phóng một mô hình dịch mạng Nơ-ron từ việc phải nén tất cả các thông tin của câu nguồn, bất kể độ dài của nó, thành một vec-tơ độ dài cố định. Điều này cho phép một mô hình xử lý tốt hơn với các câu dài.

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, quá trình chuyển ngữ

của một tên riêng sẽ tránh bị truy vết bởi cơ quan thực thi pháp luật và cơ quan tình báo.



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 ngữ liệu song song lớn, trong khi những tập ngữ liệu này có thể bao gồm vài triệu từ mà họ không bao giờ có thể hy vọng sẽ bao phủ hoàn chỉnh, đặc biệt là đối với các lớp từ có hiệu suất cao như tên riêng. Khi dịch một câu văn mới, hệ thống dịch máy thống kê dựa trên kiến thức thu được từ ngữ liệu được huấn luyện, nếu nó gặp một từ không nhìn thấy trong quá trình huấn luyện thì tốt nhất nó có thể thả từ đó vào danh sách chưa biết hoặc sao chép từ đó vào bản dịch và tệ nhất là dịch thất bại.

Các phương pháp tự động đánh giá hệ thống dịch máy thống kê hiện tại dựa vào việc tính toán các kết hợp chính xác của chuỗi từ có độ dài khác nhau, ví dụ Bleu. Do đó chuyển ngữ chính xác của tên riêng sẽ giúp làm tăng hiệu quả bản dịch. Các bản dịch thường có nhiều câu trả lời chấp nhận được, ví dụ như tiếng Nhật chuyển ngữ của "Merck"(tên nhà sản xuất dược phẩm) có thể là Meka, Meruka hoặc Meruku... Để cải thiện hiệu suất trong một hệ thống dịch máy thống kê, cần thiết lập phiên âm dự định thay vì phải chấp nhận bản dịch.

Để mở rộng bộ dữ liệu các bản dịch có thể chấp nhận được để từ đó đó chuyển ngữ, nhiều tài liệu tham khảo đã được cung cấp nhưng ngay cả với những cải tiến về hoạt động dịch máy thống kê thông qua việc dịch các tên riêng vẫn là một nhiệm vụ khó khăn, nhất là với ngôn ngữ Việt-Nhật.

Truy xuất thông tin ngôn ngữ chéo (CLIR) cũng có thể hưởng lợi từ việc dịch các từ không rõ ràng và tên riêng (AbdulJaleel và Larkey, 2003, Virga và Khudanpur, 2003). Theo bản chất các ứng dụng CLIR có thể xem xét tăng truy hồi nếu có sự không rõ ràng khi sử dụng chuyển ngữ không giám sát.

Trong luận văn này, chúng tôi lựa chọn và thực hiện đề tài "Chuyển ngữ tự động từ tiếng Việt sang tiếng Nhật". Kết quả đưa ra bằng sử dụng Moses cùng mô hình dịch máy thống kê dựa vào cụm từ phân cấp và các mô hình chuyển ngữ. Nhiều thí nghiệm đã được thực hiện để tìm ra các thông số tối ưu và nghiên cứu các ảnh hưởng của việc thay đổi kích cỡ của cả mô hình chuyển ngữ và phiên âm.

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, chủ yếu là từ tiếng Anh (80%). 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 xã hội ngày càng tăng lên.

1.4. Bài toán luận văn giải quyết

Chuyển ngữ tiếng Việt – Nhật là bài toán mới, chưa có đề tài được công bố rộng rãi. Trong khóa luận này chúng tôi nghiên cứu các phương pháp dịch máy từ tiếng Việt sang tiếng Nhật dựa trên xác suất thống kê. Trọng tâm luận văn sẽ đưa ra phương pháp chuyển ngữ các từ không xác định trong đó có tên riêng. Qua thực nghiệm để đánh giá chất lượng của bản dịch tiếng Nhật được cải thiện nhờ áp dụng mô hình chuyển ngữ không giám sát.

1.5. Kết cấu luận văn

Ngoài phần mở đầu và phần tài liệu 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

Trong chương này, chúng tôi trình bày phương pháp dịch máy thống kê sử dụng các cụm từ phân cấp. Mô hình dựa trên văn phạm phi ngữ cảnh (CFG) đồng bộ nhưng được học từ một bản song ngữ mà không có bất kỳ chú thích cú pháp nào. Nó có thể được xem như là sự kết hợp các ý tưởng nền tảng từ cả dịch dựa trên cú pháp và dịch dựa trên cụm từ. Chúng tôi mô tả chi tiết các phương pháp đào tạo và giải mã của hệ thống và đánh giá nó với tốc độ dịch và tính chính xác của bản dịch.

2.1. Ngữ pháp

Chúng tôi đưa ra một định nghĩa không chính thức và sau đó mô tả chi tiết cách chúng tôi xây dựng một văn phạm phi ngữ cảnh đồng bộ cho mô hình.

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 hiệu không kết thúc, cả γ và α là chuỗi kí hiệu kết thúc và kí hiệu không kết thúc, ~ là ánh xạ 1-1 giữa các biến cố γ và α. Ví dụ, ta có chuỗi tiếng Trung

"Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi"

được dịch 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:

$$X \rightarrow (yuX_1 youX_2, haveX_2 withX_1)$$
$$X \rightarrow (X_1 deX_2, theX_2 thatX_1)$$
$$X \rightarrow (X_1 zhiyi, one\ of X_1)$$

Trong đó các biến mà chúng tôi đã sử dụng các kí hiệu không kết thúc được đánh số 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:

$$X \rightarrow (Aozhou, Australia)$$
$$X \rightarrow (Beihan, North\ Korea)$$
$$X \rightarrow (shi, is)$$
$$X \rightarrow (bangjiao, deplomatic\ relations)$$
$$X \rightarrow (shaoshu\ guojia, few\ countries)$$

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

$$S \rightarrow (S_1 X_2, S_1 X_2)$$
$$S \rightarrow (X_1, X_1)$$

Một dẫn xuất văn phạm phi ngữ cảnh đồng bộ là một quá trình áp dụng luật để từ kí hiệu bắt đầu S dẫn xuất tới cặp câu song ngữ. Với trung gian là các cặp dạng câu chứa kí hiệu kết thúc và kí hiệu không kết thúc.

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

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

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

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

$\Rightarrow \langle$ 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 nguồn, e là câu đích, và ~ là một ánh xạ (nhiều - nhiều) giữa vị trí của f và vị trí của e. Các gióng hàng từ được tạo ra bằng cách chạy GIZA ++ 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. Điều này có thể thực hiện trong hai bước. Thứ nhất, chúng ta xác định các cặp cụm từ ban đầu sử dụng cùng một tiêu chí như hầu hết các hệ thống dịch dựa trên cụm từ, cụ thể là phải có ít nhất một từ bên trong một cụm từ gióng hàng với một từ bên trong chuỗi câu đích, nhưng không có từ bên trong một cụm từ có thể được gióng hàng với một từ bên ngoài cụm từ đích. Ví dụ: giả sử ngữ liệu huấn luyện của chúng tôi chứa đoạn sau:

30 duonianlai de youhao hezou
30 plus-year-past of friendly cooperation
Friendly cooperation over the last 30 years

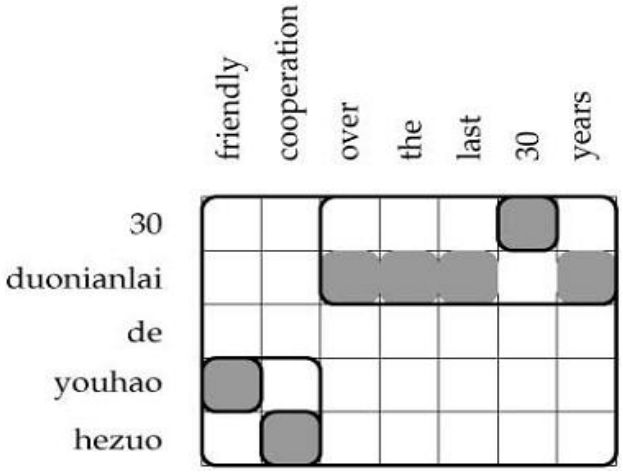
Định nghĩa 1

Cho một cặp chuỗi gióng hàng từ (f,e,~), cho f_i^j là chuỗi con của f từ vị trí i đến vị trí j, tương tự với $e_{i'}^{j'}$. Quy tắc $(f_i^j, e_{i'}^{j'})$ là viết tắt của cặp chuỗi (f,e,~) nếu:

1. $f_k \sim e_{k'}$ khi tồn tại k∈ $[i,j]$ và k' ∈ $[i',j']$
2. $f_k \sim e_{k'}$ với mọi k ∈ $[i,j]$ và k' ∈ $[i',j']$
3. $f_k \sim e_{k'}$ với mọi k ∈ $[i,j]$ và k' ∈ $[i',j']$

Thứ hai, để có được các quy tắc từ các cụm từ, chúng ta tìm các cụm từ chứa các cụm từ khác và thay các cụm từ phụ với các ký hiệu kí hiệu không kết thúc. Ví dụ, cho các cụm từ ban đầu thể hiện trong hình dưới, chúng ta có thể tạo thành quy tắc:

$$X \rightarrow (X_1 duonianlai\ de\ X_2, X_2 over\ the\ last\ X_1\ years$$



Hình 2.2: Ví dụ trích xuất ngữ pháp: Chuỗi cụm từ ban đầu

Định nghĩa 2

Bộ quy tắc (f,e,~) là bộ nhỏ nhất thỏa mãn các quy tắc sau:

1. Nếu $(f_i^j, e_{i'}^{j'})$ là cặp chuỗi mở đầu thì:

$$X \rightarrow (f_i^j, e_{i'}^{j'})$$

là quy tắc của (f,e,~)

2. Nếu $(X \rightarrow (\gamma, \alpha))$ là quy tắc của (f,e,~) và $(f_i^j, e_{i'}^{j'})$ là cặp cụm từ ban đầu sao cho $\gamma = \gamma_1 f_i^j \gamma_2$ và $\alpha = \alpha_1 e_{i'}^{j'} \alpha_2$ thì:

$$X \rightarrow (\gamma_1 X_k \gamma_2, \alpha_1 X_k \alpha_2)$$

trong đó k là chỉ dấu không được sử dụng trong γ và α là quy tắc của (f,e,~)

2.1.3. Các quy tắc khác

Quy tắc keo (Glue rules). Có quy tắc trích xuất từ ngữ liệu huấn luyện. Chúng ta có thể cho X là ký hiệu bắt đầu của ngữ pháp và dịch chuỗi mới chỉ dùng quy tắc trích xuất. Nhưng đối với tính mạnh mẽ và liên tục với các mô hình dịch dựa trên cụm từ, chúng tôi cho phép ngữ pháp chia một câu ngôn ngữ nguồn thành một chuỗi các khối và dịch một đoạn trong một thời gian. Chúng tôi chính thức hóa điều này trong

một văn phạm phi ngữ cảnh đồng bộ bằng các quy tắc gọi là các quy tắc keo, được lặp lại ở đây:

$$S \rightarrow (S_1X_2, S_1X_2)$$
$$S \rightarrow (X_1, X_1)$$

Các quy tắc này phân tích một S (ký hiệu bắt đầu) như một dãy được dịch mà không cần sắp xếp lại. Lưu ý rằng nếu chúng ta hạn chế ngữ pháp chỉ bao gồm các quy tắc keo và các cặp cụm thông thường (các quy tắc không có ký hiệu không kết thúc ở bên phải), mô hình sẽ trở thành mô hình dịch dựa trên cụm từ với bản dịch đơn âm (không có cụm từ sắp xếp lại)

Quy tắc về thực thể (Intity Rules). Cuối cùng, đối với mỗi câu được dịch, chúng tôi chạy một số mô-đun dịch chuyên ngành để dịch các con số, ngày và từng dòng trong câu, và chèn các bản dịch này vào ngữ pháp như các quy tắc mới. Các mô-đun này thường được sử dụng bởi các hệ thống dịch dựa trên cụm từ , nhưng ở đây các bản dịch có thể được đưa vào dịch dựa trên cụm từ phân cấp. Ví dụ luật:

$$X \rightarrow (X_1\ duonianlai, over\ the\ last\ X_1\ years)$$

cho phép khái quát hóa cho "years".

2.2. Mô hình

Với một câu f tiếng Việt, sẽ có một văn phạm phi ngữ cảnh đồng bộ. Nói chung, nhiều dẫn xuất f được sinh ra, và do đó nhiều bản dịch e có thể xảy ra. Bây giờ chúng ta định nghĩa một mô hình trên dẫn xuất D để dự đoán những bản dịch có nhiều khả năng hơn những bản khác.

2.2.1. Định nghĩa

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:

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

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

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

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

Trong các thử nghiệm, chúng tôi sử dụng một bộ đặc trưng tương tự như bộ đặc trưng mặc định của Pharaoh (Koehn, Och, và Marcu 2003). 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 từ 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

Cuối cùng, đối với tất cả các quy tắc, có một điểm phạt từ exp(-#T(α)), trong đó #T chỉ đếm ký tự kết thúc. Điều này cho phép mô hình học một ưu tiên chung cho các đầu ra ngắn hơn hoặc dài hơn.

2.2.3. Huấn luyện

Để ước lượng các tham số của dịch máy cụm từ và các đặc trưng trọng số, chúng ta cần phải tính toán các quy tắc trích xuất. Đối với mỗi cặp câu trong ngữ liệu huấn luyện, nói chung có nhiều hơn một dẫn xuất của cặp câu sử dụng các quy tắc trích xuất. Bởi vì chúng ta đã quan sát cặp câu nhưng không quan sát các dẫn xuất, chúng ta không biết có bao nhiêu lần mỗi dẫn xuất đã được nhìn thấy, và do đó chúng ta không thực sự biết bao nhiêu lần mỗi quy tắc đã được nhìn thấy.

Chúng tôi sử dụng phương thức tự học để giả thuyết phân phối các quy tắc có thể như thể chúng tôi đã quan sát chúng trong ngữ liệu huấn luyện, một phân phối không nhất thiết tối đa hóa khả năng ngữ liệu huấn luyện. Phương thức cho phép đếm

mỗi lần xuất hiện cặp cụm từ trích xuất. Chúng tôi cũng đếm mỗi lần xuất hiện cặp cụm ban đầu, sau đó phân phối trọng số của nó bằng nhau giữa các quy tắc thu được bằng cách loại bỏ các cụm từ phụ. Xử lý phân phối này như dữ liệu quan sát được, chúng ta sử dụng ước lượng tần suất tương đối để có được P (γ|α) và P(α|γ).

Cuối cùng, các thông số λi của mô hình logarit tuyến tính được học bởi huấn luyện tỷ lệ lỗi tối thiểu, cố gắng thiết lập các thông số để tối đa hóa điểm BLEU. Điều này cho phép văn phạm phi ngữ cảnh đồng bộ có trọng số sẵn sàng được sử dụng bởi bộ giải mã.

2.3. Giải mã

Thuật toán giải mã là vấn đề quyết định trong dịch thống kê. Và thời gian thực hiện trực tiếp ảnh hưởng tới chất lượng và tính hiệu quả. Với một thuật toán giải mã không đáng tin cậy và hiệu quả, hệ thống dịch thống kê có thể bỏ qua câu dịch tốt nhất ngôn ngữ đích của câu nguồn mặc dù nó được dự đoán đầy đủ bằng mô hình mô tả nó.

Bộ giải mã là một thuật toán CKY (Cocke-Kasami-Younger) phân tích cú pháp tìm kiếm với chùm cùng với một bộ phân tích sau f hoặc ánh xạ dẫn xuất từ tiếng Việt sang dẫn xuất tiếng Nhật. Với một câu f tiếng Việt, nó tìm ra từ tiếng Nhật có dẫn xuất tốt nhất.

Sau đây chúng ta thảo luận về các chi tiết của bộ giải mã, tập trung chú ý vào tính toán hiệu quả các xác suất mô hình tiếng Nhật cho các bản dịch có thể xảy ra, đó là thách thức kỹ thuật chính.

Thuật toán cơ bản

Dưới đây chúng ta trình bày một số trình phân tích cú pháp như là các hệ thống chứng minh suy diễn. Một trình phân tích cú pháp trong ký hiệu này định nghĩa một không gian các hạng mục trọng số, trong đó một số mục được chỉ định là các tiên đề và một số mục là mục tiêu được chỉ định (các mục cần chứng minh) và một bộ quy tắc suy luận mẫu.

$$\frac{I_1 : \omega_1 \ \ldots \ I_k : \omega_k}{I : \omega} \varphi$$

Có nghĩa là nếu tất cả các mục I_i(được gọi là các tiền đề) có thể được chứng minh, với trọng số ω_i, sau đó I (được gọi là hệ quả) được chứng minh, với trọng số ω, cung cấp điều kiện bên cạnh φ. Quá trình phân tích cú pháp phát triển một tập hợp các mục có thể kiểm chứng: Nó bắt đầu với các tiên đề, và tiến hành bằng cách áp dụng các quy tắc suy luận để chứng minh nhiều hơn và nhiều mục hơn cho đến khi một mục tiêu được chứng minh.

Ví dụ, thuật toán CKY cho văn phạm phi ngữ cảnh ở dạng chuẩn Chomsky có thể được xem như một hệ thống chứng minh suy luận có các mục có thể có một trong hai hình thức:

- [X,I,j], chỉ ra rằng một cây con được bắt nguồn từ X đã được chấp nhận kéo dài từ i đến j, hoặc

• (X → γ), nếu một quy tắc X → γ thuộc về ngữ pháp

Các tiên đề sẽ là

$$\frac{}{X \to \gamma : w} \quad (X \xrightarrow{w} \gamma) \in G$$

và các quy tắc suy luận sẽ là

$$\frac{Z \to f_{i+1} : \omega}{[Z, i, i+1] : \omega}$$

$$\frac{Z \to XY : \omega \quad [X, i, k] : \omega_1 \quad [Y, k, j] : \omega_2}{[Z, i, j] : \omega_1 \omega_2 \omega}$$

và mục tiêu sẽ là [S, 0, n], trong đó S là ký hiệu bắt đầu của ngữ pháp và n là chiều dài của chuỗi đầu vào f.

Cho một văn phạm phi ngữ cảnh đồng bộ, chúng ta có thể chuyển ngữ phía tiếng Việt sang dạng chuẩn Chomsky, và sau đó với mỗi câu, chúng ta có thể tìm phân giải tốt nhất bằng cách sử dụng CKY. Sau đó, chuyển tiếp thẳng để trở lại dạng ban đầu và ánh xạ nó vào cây tiếng Nhật tương ứng với kết quả là bản dịch đầu ra. Tuy nhiên, vì chúng ta đã giới hạn số ký hiệu không kết thúc trong các quy tắc của chúng ta là hai ký hiệu nên thuận tiện hơn khi sử dụng một thuật toán CKY đã sửa đổi hoạt động trực tiếp trên ngữ pháp mà không chuyển đổi sang dạng chuẩn Chomsky. Các tiên đề, các quy tắc suy luận, và các mục tiêu bộ giải mã cơ bản được thể hiện trong Hình 2.3. Độ phức tạp của nó là O(n^3). Ta tạm gọi là bộ phân tích cú pháp LM

$$\frac{}{X \to \gamma : w} \qquad (X \xrightarrow{w} \langle \gamma, \alpha \rangle) \in G$$

$$\frac{X \to f^{j}_{i+1} : w}{[X, i, j] : w}$$

$$\frac{Z \to f^{i_1}_{i+1} X f^{j}_{j_1+1} : w \quad [X, i_1, j_1] : w_1}{[Z, i, j] : w w_1}$$

$$\frac{Z \to f^{i_1}_{i+1} X f^{i_2}_{j_1+1} Y f^{j}_{j_2+1} : w \quad [X, i_1, j_1] : w_1 \quad [Y, i_2, j_2] : w_2}{[Z, i, j] : w w_1 w_2}$$

Goal item: $[S, 0, n]$

Hình 2.3: Các quy tắc suy luận cho bộ phân tích cú pháp LM

Phương pháp tìm kiếm thực tế được đưa ra bởi pseudoode trong hình 2.4. Nó tổ chức các thành phần đã được chứng minh thành một biểu đồ mảng có kích thước ô [X,i,j] là tập hợp các mục. Các ô được sắp xếp sao cho mọi mục xuất hiện sau các tiền đề có thể có của nó: các khoảng cách nhỏ hơn đứng trước các khoảng cách lớn hơn, và các mục X trước các mục S. Sau đó, trình phân tích cú pháp có thể tiến hành bằng cách truy cập các ô biểu đồ theo thứ tự và cố gắng để chứng minh tất cả các mục cho

mỗi ô. Bất cứ khi nào nó chứng minh một mục mới, nó sẽ thêm mục vào ô biểu đồ thích hợp. Để tái tạo các dẫn xuất sau đó, nó cũng phải lưu trữ, với mỗi mục, một bộ của các con trỏ trở lại tiền đề từ đó mục đã được suy diễn (cho tiên đề, một tuple trống được sử dụng). Nếu hai mục được thêm vào một ô tương đương, ngoại trừ trọng số hoặc các con trỏ quay trở lại, thì chúng được hợp nhất (trong ngôn ngữ giải mã dịch máy thống kê, điều này cũng được gọi là sự tái tổ hợp giả thuyết)

```
1:  procedure PARSE
2:      for all axioms (X → γ) do
3:          add (X → γ) to rchart
4:      for ℓ ← 1...n do
5:          for all i, j s.t. j − i = ℓ do
6:              if ℓ ≤ Λ then
7:                  for all items [X, i, j] : w inferable from items in rchart and chart do
8:                      add [X, i, j] : w to chart[X, i, j]
9:              if i = 0 then
10:                 for all items [S, i, j] : w inferable from items in rchart and chart do
11:                     add [S, i, j] : w to chart[S, i, j]
```

Hình 2.4: Phương pháp tìm kiếm cho bộ phân tích cú pháp LM

CHƯƠNG 3. DỊCH TÊN RIÊNG VÀ CHUYỂN NGỮ

3.1. Dịch tên riêng

Một trong những vấn đề thường gặp nhất mà dịch giả phải giải quyết là dịch chính xác tên riêng. Đối với các cặp ngôn ngữ như tiếng Tây Ban Nha-tiếng Anh, đây không phải là một thách thức lớn: cụm từ như "Antonio Gil" thường được dịch là "Antonio Gil". Tuy nhiên, tình huống phức tạp hơn đối với các cặp ngôn ngữ sử dụng bảng chữ cái và hệ thống phát âm rất khác nhau, chẳng hạn như tiếng Nhật - Anh hay tiếng Việt - Nhật. Bản dịch ngữ âm trên các cặp này được gọi là phiên âm.

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 thực thể. Tên thực thể 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).

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. Đối với các trường hợp này, chúng tôi đề nghị sử dụng mô hình ngôn ngữ không xác định ở tiếng Việt để chọn ra nghĩa phù hợp.

3.1.2. Một số nguyên tắc cần lưu ý khi chuyển tên tiếng Việt sang Katakana[17]

Nếu là nguyên âm, chuyển tương đương như sau:

Việt	Nhật
A	ア
I	イ
U	ウ
E	エ
O	オ

Bảng 3.1: Nguyên tắc chuyển ngữ nguyên âm tiếng Việt sang tiếng Nhật

Nếu là phụ âm thì chúng ta cũng chuyển các hàng tương ứng, ví dụ :

Việt	S	H
Nhật	サ(sa)	ハ(ha)
	シ(shi)	ヒ(hi)
	ス(su)	フ(fu)
	セ(se)	ヘ(he)
	ソ(so)	ホ(ho)

Bảng 3.2: Ví dụ chuyển ngữ phụ âm tiếng Việt sang tiếng Nhật

3.1.3. Phương pháp của Kevin Night (1997)

Các chú giải song ngữ chứa nhiều mục ánh xạ các cụm từ katakana trên các cụm tiếng Việt, ví dụ: "mỳ tôm" -> " ミー トム" (mitomu). Nó có thể tự động phân tích như thể các cặp đạt đủ tri thức để ánh xạ chính xác các cụm từ katakana mới mà ghép cùng nhau, và phương pháp tiếp cận này cũng sử dụng tốt cho các cặp ngôn ngữ khác. Đó là cách tiếp cận thô sơ để tìm sự tương ứng trực tiếp giữa các chữ cái tiếng Việt và các kí tự katakana, tuy nhiên nó gặp một số vấn đề.

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.

Việc phân chia bài toán của chúng ta thành 4 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. Các kỹ thuật này dựa trên xác suất và định lý Bayes.

Chúng tôi thực hiện hai thuật toán để đưa ra các bản dịch tốt nhất. đầu tiên là thuật toán đồ thị đường đi ngắn nhất Dijkstra. Thứ hai là thuật toán đường đi ngắn nhất k mà nó có thể cho chúng tôi xác định k bản dịch hiệu quả nhất với độ chính xác O(m + n log n + kn), nơi mà automat hữu hạn có trọng số chứa n trạng thái và m đối số.

Phương pháp tiếp cận đó là theo mô-đun. Chúng tôi có thể kiểm tra mỗi công cụ một cách độc lập và tin rằng các kết quả đó được kết hợp chính xác. Chúng tôi không cắt bớt, vì vậy automat hữu hạn có trọng số cuối cùng chứa tất cả các giải pháp, tuy nhiên không chắc chắn, mà tìm đường đi tốt nhất thông qua một automat hữu hạn có trọng số thay vì trình tự tốt nhất (ví dụ, cùng một chuỗi không nhận được các điểm thưởng cho việc xuất hiện nhiều hơn một lần).

3.1.4. Các mô hình xác suất

Âm tiếng Việt sang âm tiếng Nhật

Tiếp theo, chúng tôi ánh xạ các chuỗi âm tiếng Việt sang các chuỗi âm tiếng Nhật. Đây là một quá trình bị mất thông tin, như âm "R" và "L" trong tiếng Việt chuyển vào âm "r" trong tiếng Nhật, 12 nguyên âm trong tiếng Việt chuyển vào 5 nguyên âm tiếng Nhật, … chúng tôi phải đối mặt với 2 vấn đề:

1. Bản tóm tắt âm tiếng Nhật có mục đích gì?

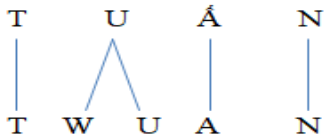
2. Chúng tôi có thể xây dựng một automat hữu hạn có trọng số để thực hiện ánh xạ các chuỗi như thế nào?

Một bản tóm tắt có mục đích rõ ràng là âm tiết tiếng Nhật được viết dưới dạng kí tự katakana của chính nó (ví dụ "ニ" tương đương "*ni*". Với cách tiếp cận này, âm "K" trong tiếng Việt tương ứng với một trong các âm カ(*ka*), キ(*ki*), ク(*ku*), ケ(*ke*) hoặc コ(*ko*), phụ thuộc vào ngữ cảnh của chúng. Không may là kí tự katakana là một âm tiết, chúng tôi sẽ không thể đưa ra khái quát rõ ràng và hữu ích, mà cụ thể âm "K" trong tiếng Việt thường tương ứng với âm "k" trong tiếng Nhật, độc lập với ngữ cảnh. Hơn nữa, sự tương ứng giữa chữ viết katakana tiếng Nhật với âm tiếng Nhật không hoàn toàn là 1-1. Vì vậy một bản tóm tắt âm thanh độc lập là nguồn tham khảo trong mọi trường hợp. Bản tóm tắt âm tiếng Nhật bao gồm 39 kí tự: 5 nguyên âm, 33 phụ âm (bao gồm nguyên âm đôi), và một âm đặc biệt (pause).

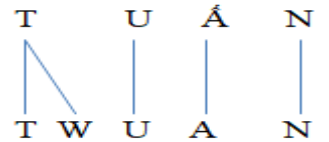
Một chuỗi âm tiếng Việt như "Hồ Chí Minh" có thể sắp xếp trên một chuỗi âm tiếng Nhật "Hochimin". Việc sắp xếp này hấp dẫn bởi các chuỗi âm tiếng Nhật luôn dài hơn chuỗi âm tiếng Việt.

Automat hữu hạn có trọng số được học tự động từ các cặp chuỗi âm 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. Sau đó áp dụng thuật toán Ước lượng tối đa hóa (estimation-maximization (EM)) để tạo xác suất kí tự nối. Thuật toán EM của chúng tôi diễn giải như sau:

1. Với mỗi cặp chuỗi âm Việt - Nhật, tính tất cả các 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. Ví dụ, có 2 cách để sắp xếp các cặp "Tuấn" <-> "twuan":



hoặc



Trong trường hợp này, sự xắp xếp bên trên 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.

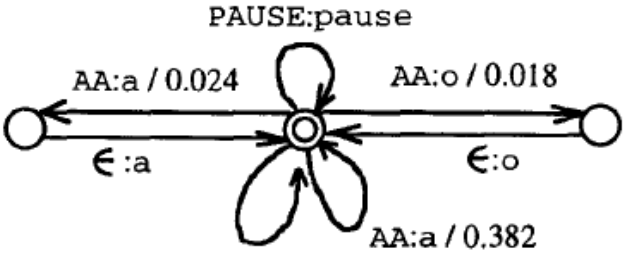
4. Với mỗi âm tiếng Việt, chuẩn hóa trọng số của các chuỗi tiếng Nhật nó kết nối tới, vì vậy tổng điểm = 1.

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 3.3: 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 thuật toán Ước lượng tối đa hóa (EM). Chỉ

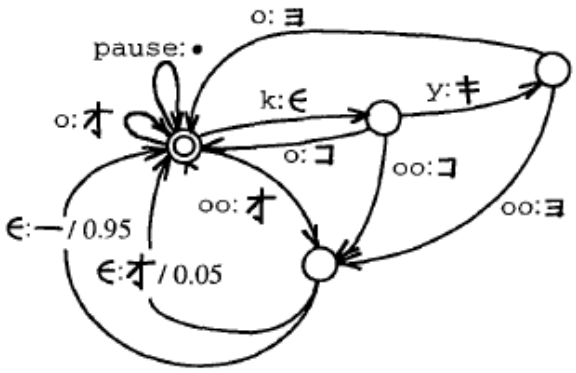
có các liên kết với xác suất điều kiện tốt hơn 1% được hiển thị, vì vậy tổng các con số có thể không = 1.

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) hơn là (t o). 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.

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". Chúng tôi kết hợp việc phân tích ngữ liệu với hướng dẫn từ sách giáo khoa tiếng Nhật (Jorden and Chaplin 1976) để chuyển lên thành nhiều biến thể chính tả và các kí tự katakana thường.

- Chuỗi âm "j i" thường được viết "ジ" nhưng thỉnh thoảng là "ヂ".
- "g u a" thường viết là "グア", nhưng thỉnh thoảng "グァ".

- “w o o” được viết bằng nhiều cách “ウオー，ウォー”, hoặc với một kí tự katakana cách viết cũ đặc biệt cho “w o”

- “y e” có thể là “エ，イエ”hoặc “イェ”.

- “w i” có thể là “ウイ” hoặc “ウィ”.

- “n y e” là một chuỗi âm hiếm gặp, nhưng được viết là “ニェ” khi nó xuất hiện.

- “t y u” hiếm gặp hơn “ch y u”, nhưng khi nó được sử dụng thì viết là “テュ”.

Sự biến đổi chính tả rõ ràng nhất trong các trường hợp mà một từ tiếng Việt như “công tắc điện” xuất hiện được chuyển ngữ khác nhau “スィッチ，スイッチ，スウィッチ” trong các từ điển khác nhau. Xử lý các biến thể này như một lớp tương đương cho phép chúng tôi học hỏi việc nối âm nói chung ngay cả khi chú giải song ngữ của chúng tôi tuân thủ một quy ước chính tả nhỏ gọn. chúng tôi không làm, tuy nhiên, tạo ra tất cả chuỗi katakana với mô hình này;

3.2. Mô hình chuyển ngữ không giám sát

Chúng tôi nghiên cứu ba phương pháp để tích hợp mô hình chuyển ngữ không giám sát vào một hệ thống dịch máy thống kê. Chúng tôi tạo ra một mô hình phiên âm từ dữ liệu song song và sử dụng nó để dịch các tên riêng. Trong các phương pháp để tích hợp chuyển ngữ, chúng tôi đã quan sát thấy những cải tiến từ điểm BLEU. Chúng tôi cũng chỉ ra rằng dữ liệu chuyển ngữ đã được khai phá cung cấp bao phủ quy tắc và chất lượng bản dịch tốt hơn so với dữ liệu chuyển ngữ theo phương pháp của Kevin Night.

3.2.1. Giới thiệu

Tất cả các hệ thống dịch máy đều tồn tại các vấn đề của tên riêng, bất kể số lượng ngữ liệu đào tạo có sẵn . Các tên riêng chủ yếu là tên các thực thể, thuật ngữ kỹ thuật hoặc các từ nước ngoài có thể được dịch sang ngôn ngữ đích bằng cách chuyển ngữ. Nhiều nghiên cứu đã cải thiện các bản dịch máy với chuyển ngữ tên các thực thể và tên riêng, cũng như hữu ích cho việc dịch các cặp ngôn ngữ liên quan chặt chẽ.

Nói chung, 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 hậu xử lý giải mã thường được sử dụng.

Trong luận văn này, chúng tôi sử dụng một mô hình chuyển ngữ không giám sát dựa trên thuật toán EM để tạo ra bộ phận phiên âm từ ngữ liệu song song được sắp xếp. Chúng tôi nghiên cứu ba phương pháp khác nhau để tích hợp chuyển ngữ trong quá trình giải mã, thực hiện trong bộ công cụ Moses.

3.2.2. Khai phá chuyển ngữ

Các khó khăn chính trong việc xây dựng một hệ thống chuyển ngữ là sự thiếu các cặp huấn luyện song ngữ sẵn có. Tuy nhiên, công bằng khi cho rằng bất kỳ dữ liệu song song nào cũng có chứa một số lượng hợp lý cặp từ đã được chuyển ngữ. Khai phá chuyển ngữ có thể được sử dụng để trích xuất các cặp từ như vậy từ hệ thống song

song. Hầu hết các kỹ thuật trước đây về khai phá chuyển ngữ thường sử dụng các phương pháp giám sát và bán giám sát. Điều này hạn chế giải pháp khai phá cho các cặp ngôn ngữ mà dữ liệu đào tạo sẵn có.

Mô hình

Mô hình khai phá chuyển ngữ là một tổng hợp của hai công thức con. Ý tưởng là công thức thứ nhất sẽ chỉ định xác suất cao hơn cho các cặp kí tự có quan hệ kí tự so với xác suất được chỉ định bởi công thức thứ 2 cho các cặp kí tự không có mối quan hệ kí tự. Xem xét một cặp từ (f, e), xác suất mô hình phiên âm cho cặp từ chuyển ngữ 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 kí tự, a là một chuỗi gióng hàng và q_j là một gióng hàng kí tự.

Với những cặp 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 hai công thức trên:

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

- λ là xác suất đầu tiên của công thức 2

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.

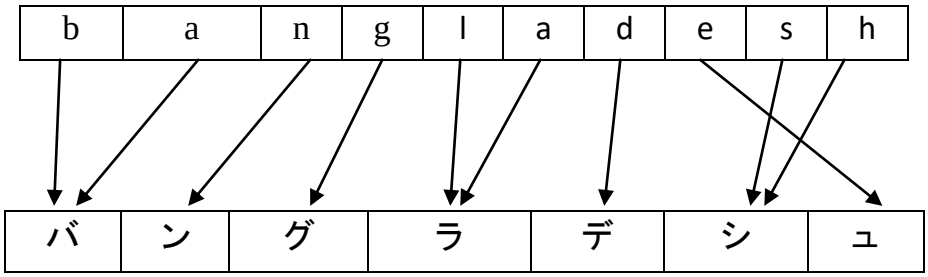
3.2.3. Mô hình chuyển ngữ

Bây giờ chúng ta có cặp từ chuyển ngữ để học một mô hình chuyển ngữ. Chúng tôi phân đoạn tập ngữ liệu huấn luyện 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 sắp xếp nào. Chúng tôi sử dụng 4 tính năng dịch cụm từ cơ bản (trực tiếp, chuyển 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 huấn luyện), và các điểm phạt từ và cụm từ.

Huấn luyện chuyển ngữ bắt đầu bằng từ không xác định được chia nhỏ thành cụm kí tự I: $f_1, f_2, \dots, f_n$... Mô hình chuyển ngữ giả sử rằng thứ tự các ký tự ở từ nguồn và từ đích là không thay đổi, chúng ta chia xác suất p(f|e) thành:

$$p(\bar{f}_1^I | \overline{fe}_1^I) = \prod_{i=1}^{I} \varphi(\bar{f}_i | \bar{e}_i)$$

Tiếp theo mỗi kí tự sẽ được chuyển ngữ sang kí tự tiếng Nhật e_i



Hình 3.1: Ví dụ gióng hàng kí tự

Kết hợp tất cả các thành phần với nhau chúng ta được:

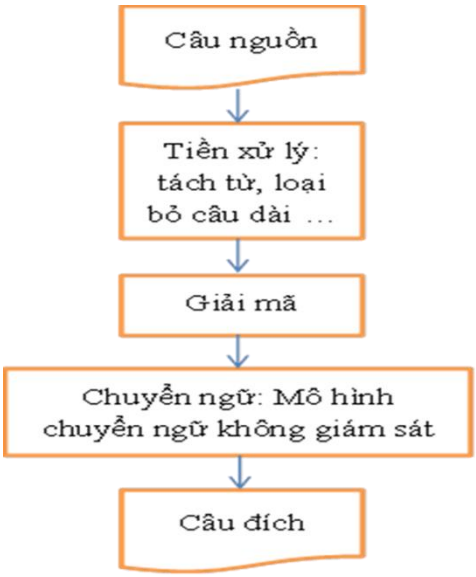
$$\hat{e} = argmax_e \prod_{i=1}^{I} \varphi(\bar{f}_i|\bar{e}_i)^{\lambda_\varphi} \prod_{i=1}^{|e|} p(e_i|e_1 \dots e_{i-1})^{\lambda}$$

3.2.4. Tích hợp với dịch máy

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

Phương pháp 1

Liên quan đến việc thay thế từ không xác định 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.



Hình 3.2 : Sơ đồ hệ dịch

Ngoài ra, nó bỏ qua bối cảnh có thể dẫn tới việc chuyển ngữ không chính xác.

Phương pháp 2

Cung cấp n bản dịch tốt nhất cho bộ giải mã đơn sử dụng một mô hình ngôn ngữ đơn và bảng chuyển ngữ cụm từ để tái ghi điểm chuyển ngữ. Chúng tôi chuyển tiếp các tính năng mô hình chuyển ngữ thứ tư được sử dụng trong hệ thống chuyển ngữ để xây dựng một cụm từ chuyển ngữ. Sử dụng thêm tính năng LM-OOV để tính số từ trong một giả thuyết không được biết đến của mô hình ngôn ngữ. Các phương

pháp làm mịn như KneserNey quy định khối lượng xác suất đáng kể cho các sự kiện không nhìn thấy, có thể khiến bộ giải mã thực hiện lựa chọn chuyển ngữ không chính xác.

Phương pháp 3

Trong Phương thức 3, chúng tôi cung cấp bảng chuyển ngữ cum từ trực tiếp vào giải mã cho phép sắp xếp các từ không xác định. Chúng tôi sử dụng tùy chọn đồ thị giải mã ngược ở Moses, cho phép nhiều bảng cụm từ chuyển đổi và các mô hình ngược. Như trong Phương thức 2, chúng ta cũng sử dụng tính năng LM-OOV trong Phương thức 3.

3.2.5. Đánh giá chất lượng dịch

Đánh giá chất lượng các bản dịch có thể được thực hiện thủ công bởi con người hoặc tự động. Mỗi phương pháp đánh giá đều có ưu nhược điểm riêng

Quá trình đánh giá thủ công cho điểm các câu dịch dựa trên sự trôi chảy và chính xác của chúng. Thế nhưng công việc đánh giá thủ công này lại tiêu tốn quá nhiều thời gian, đặc biệt khi cần so sánh nhiều mô hình ngôn ngữ, nhiều hệ thống khác nhau.

Tuy đánh giá tự động không thể phản ánh được hết mọi khía cạnh của chất lượng bản dịch, nhưng nó có thể nhanh chóng cho ta biết: chất lượng của hệ dịch ở tầm nào. Trong thực tế, điểm BLEU là độ đo chất lượng bản dịch tự động phổ biến nhất hiện nay.

BLEU tính điểm bằng cách đối chiếu kết quả dịch với tài liệu dịch tham khảo và tài liệu nguồn. Mặc dù điểm BLEU thường không thực sự tương quan với đánh giá thủ công trên các loại hệ thống khác nhau, thế nhưng vẫn có thể khá chính xác để đánh giá trên cùng một hệ thống, hoặc những hệ thống tương tự nhau.

Việc so sánh được thực hiện thông qua thống kê sự trùng khớp của các từ trong hai bản dịch tính đến thứ tự của chúng trong câu. Điểm BLEU được tính bằng công thức:

$$BLEU = BP * \frac{1}{N}\sum_{i=1}^{N} logp_i$$

với BP $= \begin{cases} 1 \text{ nếu } c > r \\ e^{(1-r/e)} \text{ nếu } c \leq r \end{cases}$

- c: độ dài bản dịch máy
- r: độ dài lớn nhất bản dịch mẫu
- N: Số lượng các bản dịch mẫu

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Chương này thảo luận về các bộ dữ liệu dùng để huấn luyện và kiểm tra các mô hình phiên âm tiếng Việt-Nhật, phần mềm Moses được sử dụng trong suốt dự án, các số liệu dùng để đánh giá và cuối cùng đưa ra mô tả các thí nghiệm.

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 được thu thập, lọc nhiễu, gióng hàng từ.

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

- Tách các từ và cụm từ trong chuỗi
- 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.
- 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:

Bộ xử lý Core2Duo T9300 2.5Ghz, RAM 2GB, HDD free 20GB

Phần mềm: Hệ điều hành Ubuntu 12.04 32 bit 4.1.2.

4.2.2. Bộ công cụ mã nguồn mở Moses

Moses là hệ dịch máy thống kê cho phép người dùng dễ dàng tạo ra mô hình dịch cho bất cứ một cặp ngôn ngữ nào. Nó bao gồm đầy đủ các thành phần để tiền xử lý dữ liệu, huấn luyện mô hình ngôn ngữ và mô hình dịch. Nó cũng bao gồm các công cụ tuning cho các mô hình này sử dụng huấn luyện với lỗi tối thiểu và đánh giá kết quả dịch sử dụng điểm BLEU.

Có thể tải về từ: http://www.statmt.org/moses/

4.2.3. GIZA ++

GIZA ++ (Och and Ney, 2003) là một phần mở rộng của chương trình GIZA (một phần của bộ công cụ SMT EGYPT) do Nhóm dịch máy thống kê phát triển trong hội thảo mùa hè năm 1999 tại Trung tâm Ngôn ngữ và Xử lý Ngôn ngữ tại Trường đại học Johns-Hopkins(CLSP / JHU) . GIZA++ mở rộng hỗ trợ của GIZA để đào tạo các mô hình IBM (Brown và cộng sự., 1993) để mô phỏng các mô hình 4 và 5. Giza được sử dụng bằng Moses để thực hiện các gióng hàng từ trên các tập ngữ liệu huấn luyện song song.

4.2.4 Mert

Việc triển khai Mert của Ashish Venugopal cho dịch máy thống kê như mô tả trong Och (2003) vàVenugopal và Vogel (2005). Nó bao gồm một số cải tiến cho phương thức đào tạo cơ bản bao gồm điều kiện ban đầu ngẫu nhiên và trật tự mẫu chuyển hoán (để giải quyết bản chất tham lam của thuật toán) và mở rộng hoặc hạn chế phạm vi các tham số động (để tăng tác động tương đối tiềm năng của chúng, hoặc

để hạn chế việc sử dụng các mô hình nhất định). Mert được sử dụng bởi Moses để tối ưu hóa hiệu năng.

4.2.5 Vitk

Công cụ phân tách từ Vitk có thể tách từ của một văn bản gồm hai triệu âm tiết tiếng Việt trong 20 giây trên một cụm ba máy tính (24 lõi, 24 GB RAM), cho độ chính xác khoảng 97%. Bộ công cụ này hướng đến khả năng xử lý dữ liệu văn bản lớn. Vì lý do này, nó sử dụng Apache Spark làm nền tảng cốt lõi. Apache Spark là một công cụ nhanh và phổ biến cho xử lý dữ liệu quy mô lớn.

Có thể tải về từ: https://github.com/phuonglh/vn.vitk

4.2.6 Mecab

Tương tự Vitk, Mecab là công cụ phân tách từ cho tiếng Nhật, độ chính xác lên đến 99% .

Có thể tải về từ: https://pypi.python.org/pypi/mecab-python3

4.3. Tiến hành thực nghiệm

4.3.1. Dữ liệu đầu vào

Dữ liệu đầu vào thu thập từ Ted và Wiki:

Dữ liệu huấn luyện	Tiếng Việt	40000 câu
	Tiếng Nhật	40000 câu
Dữ liệu điều chỉnh tham số	Tiếng Việt	950 câu
	Tiếng Nhật	950 câu
Dữ liệu đánh giá	Tiếng Việt	1000 câu
	Tiếng Nhật	1000 câu

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

Mô hình ngôn ngữ được sử dụng để đảm bảo đầu ra trôi chảy. Vì vậy nó được xây dựng bằng ngôn ngữ mục tiêu (tức là tiếng Nhật trong trường hợp này). Tài liệu KenLM cung cấp đầy đủ lời giải thích về các tùy chọn dòng lệnh, trong phạm vi luận văn sẽ xây dựng một mô hình ngôn ngữ 3-gram thích hợp.

Sau đó, chúng tôi nhị phân các tập tin *. arpa.en sử dụng KenLM để tải nhanh hơn.

Huẩn luyện mô hình dịch

Cuối cùng tới công việc chính – huấn luyện mô hình dịch. Để thực hiện việc này, chúng tôi chạy gióng hàng từ (sử dụng GIZA ++) và trích xuất cụm từ, tạo các bảng sắp xếp lại và tạo tệp cấu hình của Moses.

Huấn luyện tham số mô hình

Đây là phần chậm nhất của tiến trình. Huấn luyện tham số đòi hỏi một số lượng nhỏ dữ liệu song song, tách biệt với dữ liệu huấn luyện, vì vậy chúng tôi sử dụng một lượng dữ liệu song song gồm 950 cặp câu song ngữ Việt – Nhật.

4.3.4 Chuyển ngữ từ không xác định

Kết quả của quá trình dịch máy theo phương pháp thống kê tiếng Việt sang tiếng Nhật sẽ xuất hiện những bản dịch chứa các từ không xác định trong đó có tên riêng mà mô hình dịch không dịch được. Các từ này sẽ được chuyển ngữ bằng Phương thức 1 của mô hình chuyển ngữ không giám sát.

Phương pháp: Dùng mô hình dịch máy thống kê dựa trên cụm từ để học mô hình chuyển ngữ. Dữ liệu huấn luyện là các cặp từ trong ngữ liệu huấn luyện, chúng ta tách thành các ký tự và học hệ thống dịch cụm từ trên các cặp ký tự. Mô hình chuyển ngữ giả sử rằng thứ tự các ký tự ở từ nguồn và từ đích là không thay đổi nên chúng tôi không sử dụng mô hình sắp xếp trật tự từ (reordering model). Vì vậy, chúng tôi chỉ sử dụng 4 đặc trưng cơ bản là: đặc trưng dịch dựa trên cụm từ (phrase-translation), mô hình ngôn ngữ, điểm phạt từ và cụm (word and phrase penalties). Trọng số của các đặc trưng được được học từ 1000 cặp từ chuyển ngữ.

Dữ liệu huấn luyện mô hình chuyển ngữ: Gồm 12.260 cặp từ được trích trọn từ dữ liệu 40.000 cặp câu song ngữ.

Mô hình ngôn ngữ: 3-gram, huấn luyện từ 12.260 từ tiếng Nhật.

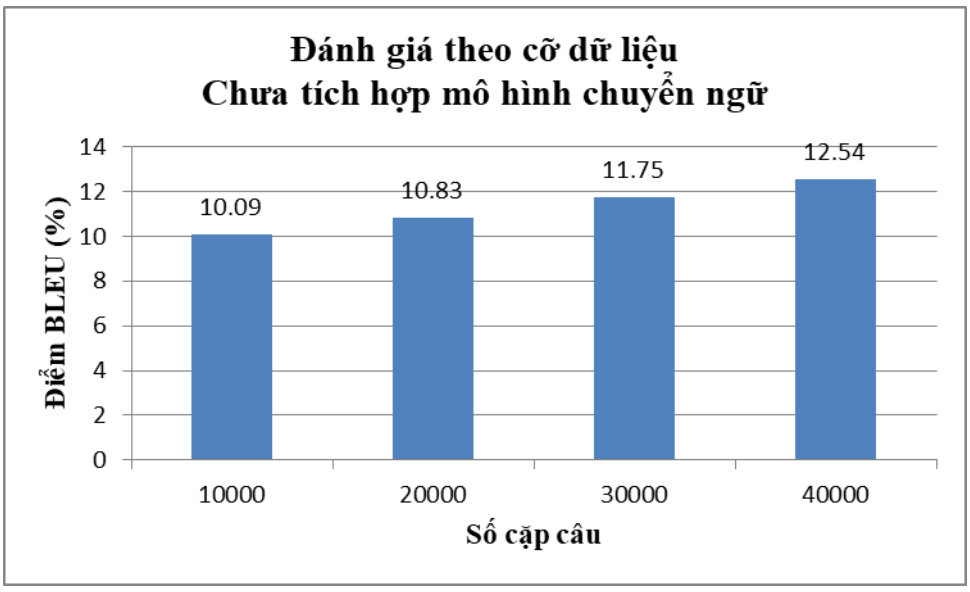
Để kiểm tra độ tốt của mô hình chuyển ngữ sau khi huấn luyện xong chúng tôi thực hiện chuyển ngữ cho các từ không xác định trong file kết quả dịch của mô hình dịch máy. Số lượng các từ không xác định của mô hình dịch máy dựa trên cụm từ phân cấp chiều Việt - Nhật (không dịch được) như sau:

- Tổng có: 2006 từ không xác định.
- Số từ mang nghĩa: 1209 từ (60.3% tổng số từ không xác định)
- Số từ không có nghĩa (tên riêng): 797 từ (39.7% tổng số từ không xác định)

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

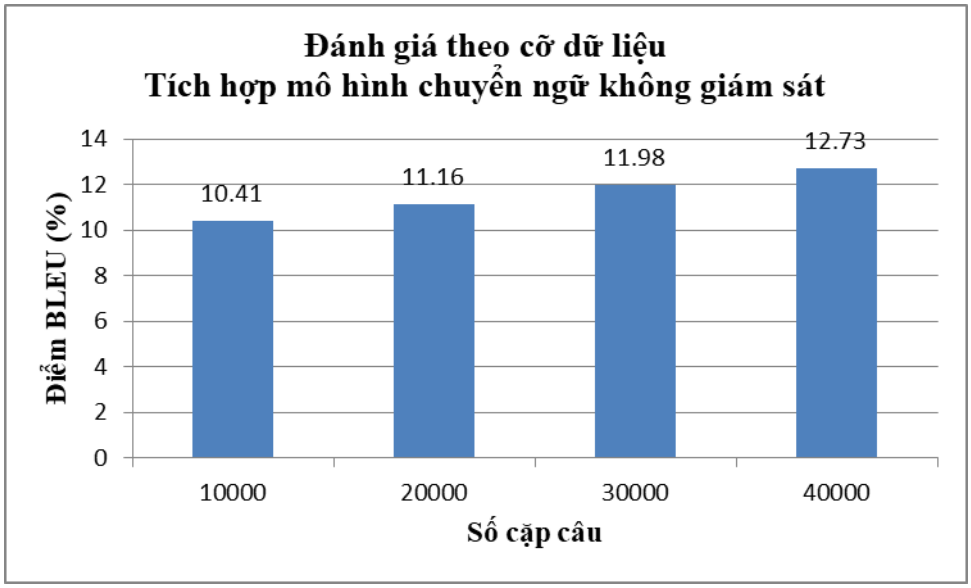
Tiếng Việt	Tiếng Nhật
alleyne đã phải nhập_viện sau khi bị bắt vì bị tức ngực .	alleyne で 逮捕 さ れ た 後 、 nhập_viện tức さ れ なけれ ば なら なかっ た ngực た 。
liên_hợp_quốc nói rằng tỷ_lệ tử_vong dân_thường ở nước này đã tăng vọt .	liên_hợp_quốc この 国 で の 民間 人 の 死亡 率 が vọt 上昇 し た と 言っ た 。
vào tháng_một năm 2011 , mandela 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 。
cùng với những vấn_đề về ánh_sáng , các hộ gia_đình cũng bị mất nước vì các trạm bơm ở mosvodokanal cũng mất điện vị sự_cố trên .	これら の と 同じ 光 の 問題 で も 、 家族 hộ mosvodokanal で も ポンプ ステーション から 電力 の 国 で の 事件 を 失っ た 。
cảnh_sát nói rằng patkar bây_giờ vẫn chưa bị bắt , nhưng báo_cáo thông_tin đầu_tiên đã được trình lên để chống lại cô vì cố_ý tự_tử .	警察 は 、 最初 の 報告 は 、 逮捕 さ れ て い た が 、 今 patkar 情報 は 自殺 を 図っ て 彼女 と 戦う ため に 提出 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 từ không xác định 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 thi_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 。
sáng thứ năm xe_buýt điện và xe_điện ở khu_vực phía nam moscow vẫn chưa hoạt_động .	バス は 木曜日 の 朝 、 モスクワ の 南部 地域 の 路面 電車 は まだ 停電 し た 。
tôi giữ quan_điểm mạnh nhất có_thể mà nói rằng điều đó là trái với lợi_ích	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 、

an_ninh của đất_nước này khi nước mỹ bị đánh_bại ở iraq .	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

Đánh giá kết quả cũng cho thấy rằng bộ phận phiên âm đã cho chất lượng bản dịch tổng thể tốt hơn so với bộ chuyển ngữ của Kevin Night. Các bản dịch tên riêng chính xác phù hợp với nguyên tắc chuyển ngữ tên riêng Việt – Nhật được mô tả trong phần 3.1.1 và 3.1.2. Mô hình chuyển ngữ không giám sát tích hợp với dịch máy thống kê dựa vào cụm từ đã được cung cấp cho cộng đồng nghiên cứu thông qua bộ công cụ của Moses.

Kết quả Áp dụng mô hình chuyển ngữ:

- Tất cả các từ: đúng 231 / 2006 từ (12%)
- Từ mang nghĩa: đúng 49 / 1209 từ (4.1%)
- Từ không có nghĩa (tên riêng): đúng 182 / 797 từ (22.8 %)

Mô hình chuyển ngữ mục đích để dịch cho những từ không có từ đối nghĩa ở ngôn ngữ đích, quá trình chuyển ngữ dựa trên việc phiên âm từ ngôn ngữ nguồn sang ngôn ngữ đích nên nó dịch tốt cho những từ không xác định trong đó có tên riêng, không tốt cho những từ đối dịch.

Ví dụ chuyển ngữ đúng:

	Tiếng Việt	**Tiếng Nhật**
Từ mang nghĩa	đối tác	パートナー
	tên lửa	ミサイル
	killing	キリン
	căn hộ	アパート
	telegraph	テレグラフ
Từ không có nghĩa (tên riêng)	dubai	ドバイ
	việt nam	ベトナム
	lê	リー
	băng cốc	バンコク
	na uy	ノルウェー

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

Ví dụ chuyển ngữ sai:

	Tiếng Việt	**Tiếng Nhật**
Từ mang nghĩa	chúc_mừng	お祈り・マャングル
	hạnh_phúc	フャ捧げる
	kỹ_thuật_viên	デジ・トゥデイ・ヴィター
	bản_quyền	ポン・直接ン
	hiệu_ứng_nhà_kính	ヒカット・ングリーンハウス
Từ không có nghĩa (tên riêng)	mâm_xôi	ムマ・ックスシェ
	hạnh_nhân	フャすばらしい
	vương_quốc_anh	ヴコーアンフ
	ấn_độ_dương	インド洋グ
	hoang_ngoc_khanh	ホアング・ンゴク・クハンフ

Bảng 4.4: Một số kết quả chuyển ngữ sai từ tiếng Việt sang tiếng Nhật tích hợp mô hình chuyển ngữ không giám sát

CHƯƠNG 5. KẾT LUẬN

Luận văn đã chỉ ra rằng một hệ thống tự động chuyển ngữ không giám sát có thể được xây dựng từ hệ thống dịch máy thống kê dựa trên cụm từ phân cấp có hiệu suất tương đương với các hệ thống hiện đại được thiết kế đặc biệt để dịch máy. Vì việc xây dựng một hệ thống không đòi hỏi có thông tin ngôn ngữ bổ sung như phát âm hoặc các ràng buộc về ngôn ngữ, nên các hệ thống cho các cặp ngôn ngữ mới có thể được phát triển nhanh chóng và rẻ tiền với đủ số lượng dữ liệu.

Một sự cải tiến đáng kể về điểm BLEU trong độ chính xác bản dịch đã được thực hiện bằng cách sử dụng mô hình chuyển ngữ không giám sát của các cặp từ không xác định tiếng Việt - Nhật được xây dựng trên dữ liệu có thước lớn hơn (40.000 cặp) so với sử dụng trong các thí nghiệm cơ bản (10.000, 20.000, 30.000 cặp).

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 tên của Kevin Night.
- Tìm hiểu phương pháp tích hợp mô hình chuyển ngữ không giám sát xử lý từ không xác định.
- Thực nghiệm mô hình dịch máy thống kê dựa trên cụm từ phân cấp, mô hình chuyển ngữ không giám sát và đánh giá kết quả giám sát và cho kết quả tương đối khả quan.

Với những kết quả đạt được trong luận văn này, trong tương lai hi vọng sẽ cải thiện được chất lượng dịch và thời gian dịch máy ngôn ngữ Việt – Nhật và từ không xác định bằng cách cập nhật các ngữ liệu đầu vào đủ lớn, giảm kích thước của bảng cụm từ, thay đổi một vài tham số để quá trình huấn luyện các mô hình được tốt hơn, cải tiến một số mô hình đảo cụm….

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Đào Ngọc Tú (2012), "Nghiên cứu vào dịch thống kê dựa vào cụm từ và thử nghiệm với cặp ngôn ngữ Anh – Việt". Luận văn Thạc sỹ, Học viện Công nghệ Bưu chính Viễn thông.

[2] 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, Đại học Quốc gia Hà Nội..

[3] Hồ Hoàng Hoa, Ngô Phương Lan, Trung tâm Nghiên cứu Tiếng Nhật.

Tiếng Anh

[4] Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio (2015). Neural Machine Translation by jointly learning to align and translate

[5] Knight, K. and Graehl, J. (1997). Machine Transliteration. Computational Linguistics, 24(4):599–612

[6] 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).

[7] 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.

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

[9] 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

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

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

[12] Koehn (2010). Statistical Machine Translation

[13] 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.

[14] 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

[15] Nadir Durrani, Hassan Sajjad, Hieu Hoang, Philipp Koehn. (2015). Integrating an Unsupervised Transliteration Model into Statistical Machine Translation

Trang web

[16] http://www.maxreading.com/sach-hay/dai-cuong-ve-tieng-viet/
[17] http://www.inas.gov.vn/403-mot-so-dac-diem-cua-tieng-nhat.html
[18] http://translate.google.com
[19] http://www.statmt.org/moses/
[20] https://github.com/phuonglh/vn.vitk
[21] https://pypi.python.org/pypi/mecab-python3