ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

TRẦN THỊ THU HUYỀN

CHUYỂN NGỮ TỰ ĐỘNG TỪ TIẾNG NHẬT SANG TIẾNG VIỆT

LUẬN VĂN THẠC SĨ

Hà Nội – 2017

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

TRẦN THỊ THU HUYỀN

CHUYỂN NGỮ TỰ ĐỘNG TỪ TIẾNG NHẬT SANG TIẾNG VIỆT

Ngành: Công nghệ Thông tin
Chuyên ngành: Kỹ thuật Phần mềm
Mã số: 60480103

LUẬN VĂN THẠC SĨ

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS. TS. Nguyễn Phương Thái

Hà Nội - 2017

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này là kết quả nghiên cứu của tôi, được thực hiện dưới sự hướng dẫn của PGS. TS. Nguyễn Phương Thái. Các nội dung được trích dẫn từ các nghiên cứu của các tác giả khác mà tôi trình bày trong luận văn này đã được ghi rõ nguồn trong phần tài liệu tham khảo.

Người thực hiện

Trần Thị Thu Huyền

LỜI CẢM ƠN

Trước hết, tôi xin chân thành cảm ơn PGS. TS. Nguyễn Phương Thái, Thầy đã trực tiếp hướng dẫn, nhiệt tình hỗ trợ và tạo điều kiện tốt nhất cho tôi thực hiện luận văn.

Tôi xin gửi lời cảm ơn đến tất cả các Thầy/Cô ở Khoa Công nghệ Thông tin, trường Đại học Công nghệ, Đại học Quốc gia Hà Nội đã giảng dạy và giúp đỡ tôi trong quá trình học tập và nghiên cứu ở trường.

Cuối cùng, tôi cũng xin gửi lời cảm ơn tới những người thân trong gia đình, bạn bè đã luôn bên cạnh động viên, ủng hộ tôi trong thời gian đi học.

Phần thực nghiệm của luận văn sử dụng kho ngữ liệu song ngữ của đề tài "Xây dựng hệ thống dịch tự động hỗ trợ việc dịch các tài liệu giữa tiếng Việt và tiếng Nhật nhằm giúp các nhà quản lý và các doanh nghiệp Hà Nội tiếp cận và làm việc hiệu quả với thị trường Nhật Bản".

Do kinh nghiệm và kiến thức còn hạn chế, tôi rất mong các Thầy/Cô và anh chị, bạn bè đóng góp thêm những ý kiến quý báu để tôi có thể hoàn thiện thêm luận văn.

Người thực hiện

Trần Thị Thu Huyền

MỤC LỤC

BẢNG KÝ HIỆU CÁC CHỮ VIẾT TẮT

BLEU	**B**iLingual **E**valuation **U**nderstudy	Đánh giá dưới dạng song ngữ
EM	**E**stimation **M**aximization	Ước lượng cực đại
MLE	**M**aximum **L**ikelihood **E**stimation	Ước lượng khả năng cực đại
MT	**M**achine **T**ranslation	Dịch máy
NMT	**N**eural **M**achine**T**ranslation	Dịch máy mạng nơ ron
OCR	**O**ptical **C**haracter **R**ecognition	Nhận dạng kí tự thị giác
RBMT	**R**ule-**b**ased **M**achine **T**ranslation	Dịch máy dựa trên nguyên tắc
SMT	**S**tatistical **M**achine **T**ranslation	Dịch máy thống kê

DANH MỤC CÁC HÌNH VẼ

DANH MỤC BẢNG

MỞ ĐẦU

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. Một vấn đề đặt ra cho việc dịch giữa các cặp ngôn ngữ là dịch chính xác tên riêng và các thuật ngữ kỹ thuật. Đối với các ngôn ngữ có hệ thống bảng chữ cái và âm thanh tương tự nhau (như tiếng Tây Ban Nha và tiếng Anh) thì không phải là vấn đề lớn nhưng với những ngôn ngữ có hệ thống chữ viết rất khác nhau thì đây là một thách thức đối với cả thông dịch viên và máy dịch.

Trước đây đã có nhiều nghiên cứu về việc Chuyển ngữ giữa các cặp ngôn ngữ khác nhau như tiếng Anh – tiếng Nhật/Trung/Hàn/Nga/Ả rập, Urdu - Ấn Độ - tiếng Anh,… sử dụng các mô hình, phương thức, cách tiếp cận khác nhau. Tuy nhiên, cho tới thời điểm này chưa có nghiên cứu nào về Chuyển ngữ giữa ngôn ngữ tiếng Nhật – tiếng Việt. Từ đó đưa ra cho chúng ta một bài toán về việc chuyển ngữ giữa cặp ngôn ngữ Nhật – Việt được xây dựng và phát triển dựa trên các nghiên cứu trước. Vì vậy, tôi lựa chọn thực hiện đề tài "Chuyển ngữ tự động từ tiếng Nhật sang tiếng Việt".

Mục tiêu nghiên cứu là chuyển phiên âm từ tiếng Nhật sang tiếng Việt để dịch những từ tiếng Nhật có phiên âm tiếng Nhật tương ứng với phiên âm tiếng Việt của từ tiếng Việt và việc dịch ở đây không dựa vào nghĩa của từ mà dựa vào phiên âm của từ đó. Nghiên cứu này tập trung về việc chuyển ngữ tên riêng và các từ không xác định (unknown) giữa cặp ngôn ngữ này.

CHƯƠNG 1. GIỚI THIỆU CHUNG

1.1. Đặc trưng ngôn ngữ tiếng Việt, tiếng Nhật

Ngôn ngữ là một hệ thống âm thanh đặc biệt, là phương tiện giao tiếp cơ bản và quan trọng nhất của các thành viên trong một cộng đồng người; ngôn ngữ đồng thời cũng là phương tiện phát triển tư duy, truyền đạt truyền thống văn hóa - lịch sử từ thế hệ này sang thế hệ khác. Cái ngôn ngữ dùng để giao tiếp và truyền đạt tư tưởng ấy, ngay từ đầu đã là ngôn ngữ thành tiếng, ngôn ngữ âm thanh. Các nhà khoa học gọi mặt âm thanh của ngôn ngữ là ngữ âm (Phonetic).

Âm thanh ngôn ngữ (còn gọi là ngữ âm) là toàn bộ các âm, các thanh, các kết hợp âm thanh và ngôn điệu mang những ý nghĩa nhất định, tạo thành cấu trúc ngữ âm của một ngôn ngữ.

Âm thanh ngôn ngữ là hình thức biểu đạt tất yếu của ngôn ngữ, là cái vỏ vật chất tiện lợi nhất của ngôn ngữ. Về một phương diện nào đó, nếu coi ngôn ngữ bao gồm hai mặt: mặt biểu hiện và mặt được biểu hiện, thì cũng có thể coi ngữ âm là mặt biểu hiện còn từ vựng và ngữ pháp là mặt được biểu hiện của ngôn ngữ.

1.1.1. Tiếng Việt

1.1.1.1. Đặc điểm tiếng Việt

Tiếng Việt thuộc ngôn ngữ đơn lập, tức là mỗi một tiếng (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.

1.1.1.2. 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. Hệ thống âm vị tiếng Việt phong phú và có tính cân đối, tạo ra tiềm năng của ngữ âm tiếng Việt trong việc thể hiện các đơn vị có nghĩa.

1.1.1.2.1. Âm tố

Âm tố là đơn vị ngữ âm nhỏ nhất trong lời nói. Có 3 loại âm tố là **nguyên âm**, **phụ âm**, **bán âm** (bán nguyên âm hay bán phụ âm).[2]

Nguyên âm có đặc điểm là khi phát âm không bị luồng hơi cản lại, ví dụ âm a, u, i, e, o,… (xem **Bảng âm vị nguyên âm**).

Phụ âm có đặc điểm là khi phát âm thì luồng hơi bị cản lại, ví dụ âm p, b, t, m, n,…(xem thêm **Bảng âm vị phụ âm**).

Bán âm có đặc điểm giống nguyên âm về mặt cấu tạo, và giống phụ âm về mặt chức năng (nên còn được gọi là bán nguyên âm hay bán phụ âm), ví dụ /u/ (ngắn), /i/ (ngắn) (xem thêm **Bảng âm vị nguyên âm**).

1.1.1.2.2. Âm vị

Âm vị là đơn vị tối thiểu của hệ thống ngữ âm của một ngôn ngữ dùng để cấu tạo và phân biệt vỏ âm thanh của các đơn vị có nghĩa của ngôn ngữ.

- **Phân biệt âm tố với âm vị - Biến thể của âm vị:**
  - Âm vị là một đơn vị trừu tượng còn âm tố là một đơn vị cụ thể. Âm vị được thể hiện ra bằng các âm tố và âm tố là sự thể hiện của âm vị.
  - Những âm tố cùng thể hiện một âm vị được gọi là các biến thể của âm vị.
- Tiếng Việt có 16 âm vị là *nguyên âm* (trong đó có 13 nguyên âm đơn, 3 nguyên âm đôi) và 2 âm vị là bán nguyên âm*. Trong 16 âm vị nguyên âm và 2 âm vị bán nguyên âm thì có 17 cách đọc (phát âm), và được ghi lại bằng 20 chữ viết. 20 chữ viết này được hình thành từ 12 chữ cái (con chữ). [2]

Bảng 1.1. Bảng âm vị nguyên âm

Stt	Âm vị	Chữ viết	Đọc	Chữ cái (con chữ)	Ví dụ
1	|i|	*i, y*	i	i, y	*im lim, ý chí*
2	|e|	*ê*	ê	ê	*ê chề, êm đềm*
3	|ɛ|	*e*	e	e	*e dè, e thẹn*
4	|ɛ̆|	*ɑ*	a	a	*anh ách, xanh xanh*
5	|a|	*ɑ*	a	a	*a ha, la đà*
6	|ă|	*ă/ɑ*	ă/a	ă/a	*ăn năn, ăn chặn/rau đay*
7	|ɤ|	*ơ*	ơ	ơ	*bơ phờ, tờ mờ*
8	|ɤ̆|	*â*	â	â	*ân cần, lấn bấn*
9	|ɯ|	*ư*	ư	ư	*từ từ, lữ thứ*
10	|o|	*ô*	ô	ô	*ô hô, hồ đồ*
11	|ɔ|	*o/oo*	o	o/o+o	*co ro, lò dò/xoong*
12	|ɔ̆|	*o*	o	o	*vòng lọng, tóc, học*
13	|u|	*u*	u	u	*tu hú, lù mù*
14	|i̯e|	*ia/ya* *iê/yê*	ia iê	i+a/y+a i+ê/y+ê	*kia kìa/khuya* *yêu chiều*
15	|u̯o|	*uô/ua*	uô/ua	u+ô, u+a	*tuốt tuồn tuột/ tua rua*
16	|ɯ̯ɤ|	*ươ/ưa*	ươ/ưa	ư+a, ư+ơ	*lướt thướt/ lưa thưa*
17	|i̯|	*i/y*	i	i/y	*tai tái/cày cấy*
18	|u̯|	*o/u*	o/u	o/u	*toán, đào hào/tuần, đau*

* Trong 2 âm vị bán nguyên âm |i̯| và |u̯| thì |u̯| vừa đóng vai trò là âm đệm (viết "o" trong *toán, toàn, xoan*..., viết "u" trong *tuần, tuấn, quẩn*...), vừa đóng vai trò âm cuối (viết "o" trong *đào hào, báo cáo, táo*..., viết "u" trong *đau, rau câu*...), còn |i̯| đóng vai trò âm cuối.

- Tiếng Việt có 23 âm vị là **phụ âm**. Tương ứng với 23 âm vị phụ âm thì có 24 cách đọc (phát âm), và được ghi lại bằng 27 chữ viết. 27 chữ viết này được hình thành từ 19 chữ cái (con chữ).

Bảng 1.2. Bảng âm vị phụ âm

STT	Âm vị	Chữ viết	Đọc	Chữ cái (con chữ)	Ví dụ
1	| f |	ph	phờ	p+h	*phổi, pháo*
2	| t' |	th	thờ	t+h	*thu, thôi*
3	| ʈ |	tr	trờ	t+r	*trăng, trời*
4	| z |	gi/d	gi/dê	g+i/d	*giếng, dao*
5	| c |	ch	chờ	c+h	*chơi, cho, chuộng*
6	| ɲ |	nh	nhờ	n+h	*nhà, nhảy, những*
7	| ŋ |	ng/ngh	ngờ	n+g/n+g+h	*ngành, người, nghĩ, nghề*
8	| χ |	kh	khờ	k+h	*khuya, không*
9	| ɣ |	g/gh	gờ	g/g+h	*gà, gọi, ghi, ghe*
10	| k |	c/q/k	xê/quy/ca	c/q/k	*cà kê, cá quả*
11	| t |	t	tê	t	*ta, tôi, tức*
12	| ʐ |	r	e-rờ	r	*rổ, rá*
13	| h |	h	hát	h	*hoa, học hành*
14	| ɓ |	b	bê	b	*bằng, bơi, biết*
15	| m |	m	em-mờ	m	*miệng, môi, mắt, mũi*
16	| v |	v	vê	v	*vui, vắng, vụt*
17	| ɗ |	đ	đê	đ	*đang, đợi, đói*
18	| n |	n	en-nờ	n	*năm, nàng, nên*
19	| l |	l	e-lờ	l	*lên, lòng, lợi*
20	| s |	x	ích-xì	x	*xuống, xưa*
21	| p |	p	pê	p	*bắp, bịp, chắp*
22	| ʂ |	s	ét-sì	s	*say sưa, sắp sửa*
23	| ʔ |	zero	zero	zero	*ăn uống, ỉ eo, ồn ào*

Những âm tiết không có âm đầu (như: *âm, êm, oai, uyên*) khi phát âm được bắt đầu bằng động tác khép kín khe thanh, sau đó mở ra đột ngột gây nên một tiếng bật. Động tác khép kín ấy có giá trị như một phụ âm và người ta gọi là âm tắc thanh hầu, kí hiệu: /?/.

1.1.1.2.3. Tiếng

Khi người Việt phát âm các âm tiết để tạo nên chuỗi lời nói khi giao tiếp cụ thể, đơn vị được dùng trong chuỗi lời nói là "tiếng". **Tiếng** trong tiếng Việt thường được hiểu là *âm tiết*, về mặt là đơn vị có nghĩa, dùng trong chuỗi lời nói.

Trên chữ viết, mỗi tiếng được ghi thành một **chữ**. Tiếng có thể trực tiếp hay gián tiếp gắn liền với một ý nghĩa nhất định và không thể chia ra thành những đơn vị có nghĩa nhỏ hơn nữa. Vì vậy có thể hiểu *tiếng* trùng với *hình vị* và từ: *ăn, nói, đi, đứng, và, sẽ,…* là những tiếng trong tiếng Việt.

1.1.1.2.4. Hình vị

Hình vị thường có hình thức cấu tạo một âm tiết, tức là mỗi ***hình vị*** trùng với ***âm tiết***, trên chữ viết mỗi hình vị được viết thành một ***chữ***. Hình vị trong tiếng Việt có thể một mình đóng vai trò như một từ cũng có thể làm thành tố cấu tạo từ, nhưng nó chỉ được phân xuất ra nhờ phân tích bản thân các từ.

Ví dụ trong phát ngôn "***Ngày mai tôi nghỉ học***" sẽ có 5 hình vị có ý nghĩa là "***ngày*** / ***mai*** / ***tôi*** / ***nghỉ*** / ***học***".

1.1.1.3. 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.

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.

1.1.1.4. Ngữ pháp

Từ của tiếng Việt không biến đổi hình thái. Đặc điểm này sẽ chi phối các đặc điểm ngữ pháp khác. Khi từ kết hợp từ thành các kết cấu như ngữ, câu, tiếng Việt rất coi trọng phương thức trật tự từ và hư từ.

Trật tự chủ ngữ đứng trước, vị ngữ đứng sau là trật tự phổ biến của kết cấu câu tiếng Việt.

Phương thức hư từ cũng là phương thức ngữ pháp chủ yếu của tiếng Việt. Nhờ hư từ mà tổ hợp "anh của em" khác với tổ hợp "anh và em", "anh vì em".

Ngoài trật tự từ và hư từ, tiếng Việt còn sử dụng phương thức ngữ điệu. Ngữ điệu giữ vai trò trong việc biểu hiện quan hệ cú pháp của các yếu tố trong câu, nhờ đó nhằm đưa ra nội dung muốn thông báo. Trên văn bản, ngữ điệu thường được biểu hiện bằng dấu câu.

Qua một số đặc điểm nổi bật vừa nêu trên đây, chúng ta có thể hình dung được phần nào bản sắc và tiềm năng của tiếng Việt.

1.1.2. Tiếng Nhật

1.1.2.1. Hệ thống bảng chữ cái tiếng Nhật

Khác với tiếng Việt, tiếng Anh và hầu hết các ngôn ngữ khác, tiếng Nhật có 3 bảng chữ là chữ mềm (hiragana), chữ cứng (katakana) và chữ Hán (kanji). Hệ thống các bảng chữ cái này được sử dụng linh hoạt, tức là trong một câu tiếng Nhật có thể được kết hợp từ chữ của **cả 3 bảng chữ cái** trên.

- Chữ Hán để thể hiện ý nghĩa của câu

- Chữ Hiragana được dùng làm chức năng ngữ pháp, có nghĩa là Hiragana được sử dụng để biểu thị mối quan hệ, chức năng trong câu của các chữ Hán.

Ví dụ, chữ Hán "thực" (食), thêm Hiragana vào, ta sẽ có 食べる nghĩa là "ăn", 食べている là "đang ăn", 食べたい là "muốn ăn", 食べた là "Đã ăn",… Vì vậy, tất cả các trợ từ trong tiếng Nhật đều là hiragana.

- Katakana được tạo thành từ các nét thẳng, nét cong và nét gấp khúc, khác với Hiragana với những đường nét mềm dẻo, uốn lượn. Nó thường dùng để:

+ Phiên âm những từ có nguồn gốc từ nước ngoài (gọi là *gairaigo*). Ví dụ, "television" (Tivi) được viết thành "テレビ" (terebi).

+ Viết tên các quốc gia, tên người hay địa điểm của nước ngoài. Ví dụ, tên "Việt Nam" được viết thành "ベトナム" (Betonamu).

+ Viết từ ngữ trong khoa học – kỹ thuật, như tên loài động vật, thực vật, tên sản vật, hoặc tên của các công ty.

+ Nhấn mạnh, đặc biệt đối với các ký hiệu, quảng cáo, áp phích. Ví dụ, chúng ta có thể sẽ nhìn thấy chữ "ココ" – *koko* – ("ở đây") hay ゴミ *gomi* ("rác").

Hình 1.1. Bảng chữ cái Katakana

- Âm đục:

  Katakana có âm đục được kí hiệu bằng cách thêm dấu "tenten".

ガ ga	ギ gi	グ gu	ゲ ge	ゴ go
ザ za	ジ ji	ズ zu	ゼ ze	ゾ zo
ダ da	ヂ (ji)	ヅ (zu)	デ de	ド do
バ ba	ビ bi	ブ bu	ベ be	ボ bo
パ pa	ピ pi	プ pu	ペ pe	ポ po

- Âm ghép:

Katakana cũng có âm ghép. Các chữ “ヤ”, “ユ”, “ヨ” sẽ được viết nhỏ lại thành “ゃ”, “ゅ”, “ょ”.

nguyên âm đôi		
ya	yu	yo
キャ kya	キュ kyu	キョ kyo
シャ sha	シュ shu	ショ sho
チャ cha	チュ chu	チョ cho
ニャ nya	ニュ nyu	ニョ nyo
ヒャ hya	ヒュ hyu	ヒョ hyo
ミャ mya	ミュ myu	ミョ myo
リャ rya	リュ ryu	リョ ryo
ギャ gya	ギュ gyu	ギョ gyo
ジャ ja	ジュ ju	ジョ jo
ヂャ (ja)	ヂュ (ju)	ヂョ (jo)
ビャ bya	ビュ byu	ビョ byo
ピャ pya	ピュ pyu	ピョ pyo

- Âm ngắt:

Âm ngắt của Katakana cũng có cách phát âm giống như Hiragana và chữ“ツ” được viết nhỏ lại thành “ッ”.

Ví dụ: ロマンチック romanti**kk**u : lãng mạn (romantic).

- Trường âm

Trường âm của Katakana thì tất cả đều biểu diễn bằng dấu 「ー」

Ví dụ: インターネット int**aa**netto : Internet

1.1.2.2. Ngữ âm

- Âm tiết 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.

Khác với tiếng Việt, âm tiết trong tiếng Nhật hầu hết đều không mang nghĩa. Tuy nhiên, cũng có số lượng rất nhỏ những từ được cấu tạo bởi 1 âm tiết và âm tiết mang ý nghĩa của từ đó. Ví dụ: "ki" có nghĩa là cái cây, "e" có nghĩa là bức tranh,...

- 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/. Ngoài ra còn có hai âm đặc biệt là âm mũi (N) và âm ngắt (Q).

- 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.

1.1.2.3. Từ vựng

Tiếng Nhật là một ngôn ngữ có một vốn từ vựng rất lớn và vô cùng phong phú, điều này được thể hiện ở một số mặt:

- Thứ nhất, tính nhiều tầng lớp của vốn từ vựng.

+ Lớp từ gốc Hán (Kango) được vay mượn từ Trung, chiếm hơn 60% vốn từ vựng và chủ yếu là các danh từ, đặc biệt là danh từ biểu thị các khái niệm trừu tượng như ***tetsugaku*** (triết học), ***shugi*** (chủ nghĩa), ...

+ Lớp từ gốc Nhật chủ yếu bao gồm các danh từ, động từ, tính từ thuộc lĩnh vực ngôn ngữ đời sống sinh hoạt hàng ngày và nhóm các trợ từ biểu thị các kiểu ý nghĩa ngữ pháp (trợ từ cách, liên từ, thán từ, trợ động từ...). Nhóm từ ngoại lai (Gairaigo) là những từ vay mượn từ các ngôn ngữ khác mà chủ yếu là tiếng Anh, Pháp, Đức,...

Để phân biệt với nhóm từ gốc Hán và từ thuần Nhật, nhóm từ ngoại lai được viết bằng chữ Katakana. Tuy nhiên, những từ ngoại lai đầu tiên xuất hiện ở Nhật Bản vào thế kỷ thứ 16 là các từ tiếng Bồ Đào Nha như: ***tabako*** (thuốc lá), ***tempura*** (món tẩm bột rán)... trải qua một thời gian dài đã được coi như những từ thuần Nhật nên chúng đều được viết bằng chữ Hiragana.

-Thứ hai, khả năng kết hợp các từ với nhau để tạo ra từ mới là rất lớn.

1.1.2.4. Ngữ pháp

- Đặc điểm nổi bật nhất là trật tự câu hoàn toàn đảo lộn so với các ngôn ngữ khác như tiếng Việt, Anh, Trung... Trong đó, vị ngữ đứng cuối câu là một nguyên tắc bất dịch.

- Ngữ pháp tiếng Nhật giống với các ngôn ngữ biến hình như tiếng Anh, Nga, Pháp..., động từ và tính từ trong tiếng Nhật có sự biến đổi về mặt hình thức bằng cách ghép thêm tiếp vĩ ngữ để tạo thành thời, thể, trạng thái..., nhưng không biểu hiện ngôi và số.

- Trong hội thoại, các ngôi nhân xưng, đặc biệt là chủ ngữ thường được giản lược một cách tối đa có thể. Chỉ cần nhìn vào dạng thức của động từ cũng có thể phân biệt được ai là chủ thể của lời nói, ai là đối tượng giao tiếp và mối quan hệ xã hội giữa họ.

- Kính ngữ cũng là một phạm trù ngữ pháp quan trọng của tiếng Nhật.

+ Các phương tiện biểu thị kính ngữ trong tiếng Nhật bao gồm từ vựng và ngữ pháp, song phương tiện ngữ pháp chiếm tỉ lệ khá lớn.

+ Có ba dạng chính là: dạng thức kính trọng, dạng lịch sự và dạng khiêm tốn.

1.2. Bài toán dịch máy và dịch thống kê dựa vào cụm từ

1.2.1. Bài toán dịch máy

Lịch sử ra đời của dịch máy (MT) đã trải qua hơn 60 năm, ngay sau khi những chiếc máy tính đầu tiên được người Anh dùng để giải mã trong chiến tranh Thế giới thứ II [5]. Các phương pháp bắt nguồn từ các nguyên tắc về ngôn ngữ cũng được nghiên cứu. Trong những năm 1970, việc xây dựng các hệ thống thương mại đầu tiên được đưa ra và cùng với sự ra đời của máy tính cá nhân, các dịch giả chuyển sang sử dụng các công cụ ghi nhớ dịch thì bài toán MT coi như một ứng dụng thực tế. Hiện nay, xu hướng phổ biến là hướng tới các phương pháp dựa vào dữ liệu, đặc biệt là các phương pháp thống kê.

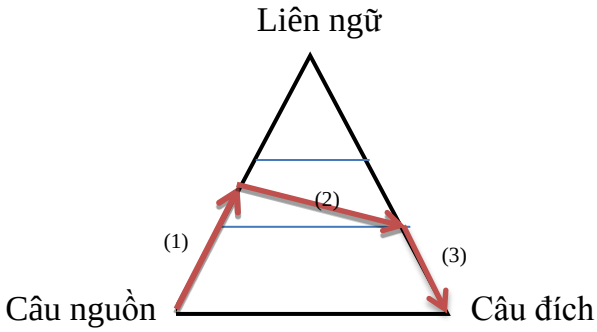
Ta có thể hiểu MT là việc dịch tự động, nó là quá trình mà phần mềm máy tính dịch văn bản từ một ngôn ngữ (ngôn ngữ nguồn) sang một ngôn ngữ khác (ngôn ngữ đích).

Để thực hiện bất kỳ việc dịch nào bởi dịch giả hay dịch tự động thì ý nghĩa của văn bản trong ngôn ngữ nguồn phải được khôi phục đầy đủ trong ngôn ngữ đích, tức là bản dịch. Nhìn bề ngoài có vẻ đơn giản nhưng quá trình dịch rất phức tạp. Việc dịch không chỉ là sự thay thế từ với từ mà dịch giả cần phải giải thích và phân tích tất cả các yếu tố trong văn bản và xem xét các từ có ảnh hưởng như thế nào trong câu và toàn văn bản. Điều này đòi hỏi dịch giả có sự hiểu biết sâu rộng về ngữ pháp, cú pháp, ngữ nghĩa… trong ngôn ngữ nguồn và ngôn ngữ đích, cũng như am hiểu về cách sử dụng câu từ ở mỗi vùng miền địa phương khác nhau.

Việc dịch thực hiện bởi dịch giả và máy tính đều có những khó khăn và thách thức. Ví dụ, không thể có hai dịch giả khác nhau cùng tạo ra một bản dịch giống hệt nhau của cùng một văn bản trong cùng một cặp ngôn ngữ và cũng cần phải chỉnh sửa một vài lần thì mới có thể đáp ứng yêu cầu của khách hàng. Nhưng khó khăn hơn cả là MT có thể tạo ra các bản dịch chất lượng có thể được sử dụng công khai, rộng rãi.

Thực hiện nghiên cứu MT không giới hạn việc dịch tự động một cách hoàn toàn và chất lượng dịch tốt. Hay nói cách khác, công nghệ MT phát triển tỉ lệ thuận với chất lượng dịch.

Quá trình MT nói chung được thể hiện theo mô hình tam giác như hình 1.2 sau:



Hình 1.2. Tam giác thể hiện quá trình dịch máy

Phía trái của tam giác mô tả câu ở ngôn ngữ nguồn; phía bên phải ở ngôn ngữ đích. Các mức khác nhau bên trong tam giác biểu diễn chiều sâu của việc phân tích của câu nguồn, ví dụ như phân tích cú pháp hoặc ngữ nghĩa. Hiện tại, ta không thể tách phân tích cú pháp và ngữ nghĩa của một câu, nhưng giả thuyết là ta có thể phân tích sâu hơn và hơn nữa một câu đã được đưa ra. Mũi tên đỏ đầu tiên (1) thể hiện sự **phân tích** câu ở ngôn ngữ nguồn. Từ câu hiện tại là một chuỗi các từ, chúng ta có thể xây dựng một sự thể hiện bên trong tương ứng với mức độ chúng ta có thể phân tích câu.

Ví dụ, ở mức độ mà chúng ta có thể xác định các phần của lời nói của mỗi từ (danh từ, động từ,…), và trên một từ khác chúng ta có thể kết nối các từ: ví dụ, cụm danh từ là chủ ngữ của động từ.

Khi việc phân tích kết thúc, câu được "**chuyển đổi**" bằng tiến trình thứ hai (2) thành việc thể hiện bằng chiều sâu tương đương hoặc ít hơn một chút về ngôn ngữ mục tiêu. Sau đó, tiến trình thứ ba (3) được gọi là "**sinh**", tạo ra câu đích từ việc biểu diễn bên trong đó, tức là một chuỗi các từ có ý nghĩa trong ngôn ngữ đích. Ý tưởng của việc biểu diễn theo hình tam giác trên là ta càng phân tích ngôn ngữ nguồn sâu hơn hoặc ở mức cao hơn thì giai đoạn **chuyển đổi** càng nhỏ hơn/đơn giản hơn. Cuối cùng, nếu chúng ta có thể chuyển đổi một ngôn ngữ nguồn thành một sự thể hiện "**liên ngữ**" chung trong quá trình phân tích này thì chúng ta sẽ không cần thực hiện bất kỳ việc **chuyển đổi** nào - và chúng ta chỉ cần tiến trình **phân tích** và **sinh** cho mỗi ngôn ngữ để dịch từ ngôn ngữ bất kỳ nào đó sang ngôn ngữ khác.

Các công nghệ chính sử dụng cho việc dịch văn bản: SMT, RBMT và NMT.

- RBMT là công nghệ cũ nhất, dựa trên vô số các quy tắc ngôn ngữ được xây dựng và hàng triệu bộ từ điển song ngữ cho mỗi cặp ngôn ngữ.
    - Phần mềm phân tích cú pháp văn bản và tạo ra một biểu diễn quá độ từ đó tạo ra văn bản trong ngôn ngữ đích. Quá trình này yêu cầu các thuật ngữ đa dạng với các thông tin về hình thái, cú pháp và ngữ nghĩa, cùng các bộ quy tắc rộng rãi. Phần mềm sử dụng các bộ quy tắc phức tạp và sau đó chuyển cấu trúc ngữ pháp của ngôn ngữ nguồn sang ngôn ngữ đích.
    - Trong hầu hết các trường hợp, có hai bước: đầu tiên là một khoản đầu tư ban đầu làm tăng đáng kể chất lượng dịch với chi phí giới hạn; sau đó đầu tư liên tục

để nâng cao chất lượng. Mặc dù RBMT giúp các doanh nghiệp đạt chất lượng nhưng quá trình cải tiến chất lượng có thể tốn kém.

- SMT là công nghệ được ứng dụng rộng rãi hiện nay, để dịch văn bản tự động có sử dụng các mô hình dịch thống kê có các tham số bắt nguồn từ việc phân tích các ngữ liệu đơn ngữ và song ngữ, việc học máy phụ thuộc vào bộ dữ liệu các bản dịch trước đó, hay còn gọi là bộ nhớ dịch.
    - Xây dựng mô hình dịch thống kê là một quá trình nhanh chóng, nhưng công nghệ này dựa chủ yếu vào các bộ ngữ liệu đa ngôn ngữ hiện có. Về mặt lý thuyết, có thể đạt được ngưỡng chất lượng nhưng hầu hết các doanh nghiệp không có số lượng ngữ liệu lớn như vậy để xây dựng các mô hình dịch cần thiết.
    - SMT cần CPU (Central Processing Units – bộ vi xử lý trung tâm) chuyên sâu và một cấu hình phần cứng phong phú để chạy các mô hình dịch cho mức hiệu suất trung bình.
- NMT là công nghệ mới được phát triển gần đây, nó cũng huấn luyện các bộ nhớ dịch như SMT, nó sử dụng học sâu (deep learning) và có thể cả dữ liệu huấn luyện lớn hơn để xây dựng mạng nơ ron nhân tạo. Nó đòi hỏi chạy trên GPU (Graphics Processing Units – bộ xử lý đồ họa) mạnh mẽ.

Theo Koehn [11], vào những năm 1980 – 1990, ngay trong đợt cuối nghiên cứu về mạng nơ ron, dịch máy đã được các nhà nghiên cứu khám phá ra các phương pháp này. Trên thực tế, các mô hình đề xuất bởi Forcada và Ñeco (1997) và Castaño cùng cộng sự (1997) được coi là tương tự như các cách tiếp cận dịch máy mạng nơ ron hiện nay. Tuy nhiên, không có mô hình nào được huấn luyện với kích thước dữ liệu đủ lớn để đưa ra các kết quả hợp lý. Sự tính toán phức tạp gây khó khăn, vượt xa các nguồn lực của thời đó, do đó ý tưởng này đã bị bỏ rơi trong gần hai thập niên.

Trong thời gian đó, các cách tiếp cận kênh-nguồn như dịch máy thống kê dựa vào cụm từ phát triển mạnh mẽ, đưa dịch máy trở thành công cụ hữu ích cho nhiều ứng dụng.

Sự hồi sinh của các phương pháp mạng nơ ron bắt đầu với việc tích hợp các mô hình ngôn ngữ nơ ron vào các hệ thống dịch máy thống kê truyền thống. Nghiên cứu tiên phong của Schwenk (2007) cho thấy những cải tiến lớn trong các chiến dịch đánh giá chung.

Ngoài việc sử dụng trong các mô hình ngôn ngữ, các phương pháp mạng nơ-ron được đưa vào các thành phần khác của dịch máy thống kê truyền thống, chẳng hạn như cung cấp các bảng dịch bổ sung hoặc mở rộng điểm (Schwenk, 2012; Lu và cộng sự, 2014), sắp xếp lại trật tự (Kanouchi và cộng sự, 2016, Li et al, 2014) và các mô hình sắp xếp trước (de Gispert et al, 2015), …. Ví dụ, bản dịch chung và mô hình ngôn ngữ của Devlin et al. (2014) có ảnh hưởng vì nó cho thấy những cải tiến về chất lượng lớn trên hệ thống dịch máy thống kê có tính cạnh tranh cao.

Trong một đến hai năm gần đây, các nghiên cứu của dịch máy là chủ yếu về mạng nơ ron. Tuy nhiên, phương pháp dịch máy thống kê truyền thống vẫn có nhiều ưu điểm, nhất là tính toán thống kê giúp giải quyết rõ ràng các hiện tượng như mối quan hệ giữa các từ, cụm từ trong văn bản… nên hướng nghiên cứu của luận văn tập trung về dịch máy thống kê sẽ được trình bày ở các nội dung sau đây.

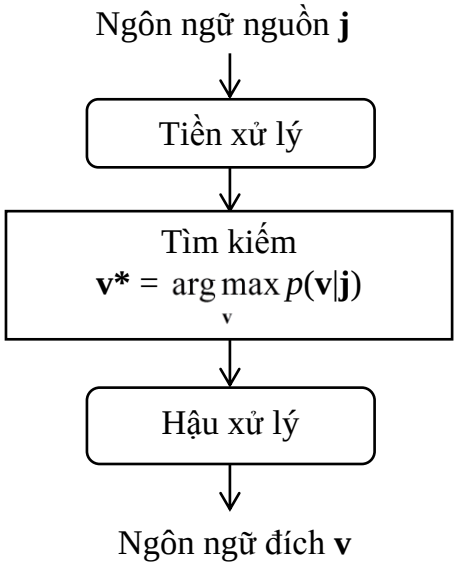
1.2.2. Dịch máy thống kê

Theo Philipp Koehn [1], vào cuối những năm 1980, ý tưởng ra đời SMT của IBM Research trong sự thành công của các phương pháp thống kê trong nhận dạng giọng nói. Bằng việc mô hình hóa nhiệm vụ dịch như một vấn đề tối ưu hóa thống kê, dự án Candide đã đặt MT trên một nền tảng toán học đã xây dựng vững chắc.

SMT đã được định nghĩa ở phần 1.2.1 như trên. Dịch máy dựa trên phương pháp thống kê tìm câu **v** ở ngôn ngữ đích ("Tiếng Việt") phù hợp nhất (có xác suất cao nhất) khi cho trước câu **j** ở ngôn ngữ nguồn ("Tiếng Nhật"), biểu diễn theo công thức (1.1).

$$\mathbf{v}^* = \arg\max_{\mathbf{v}} p(\mathbf{v}|\mathbf{j}) \qquad (1.1)$$

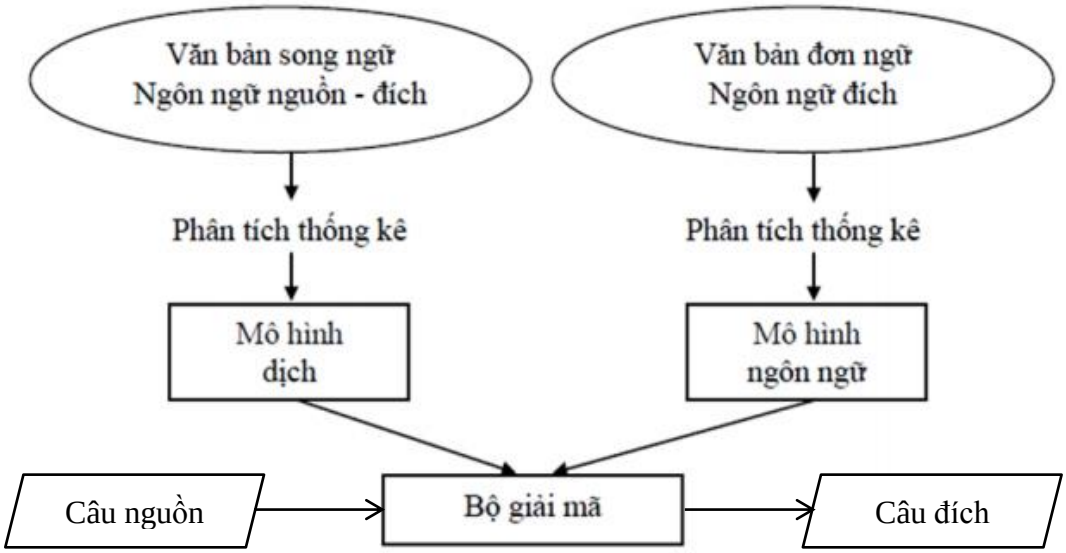
Ta có mô hình hóa bài toán MT dựa trên phương pháp thống kê như sau:



Hình 1.3. Mô hình hóa bài toán MT dựa trên phương pháp thống kê

Trong đó, bước Tìm kiếm **v*** là giai đoạn giải mã của hệ dịch máy. Ta cần tìm giá trị **v*** lớn nhất trong tập các bản dịch **v** ở ngôn ngữ đích và không gian tìm kiếm ở đây rất lớn.

- **Các thành phần của SMT:**



Hình 1.4. Các thành phần của hệ dịch máy SMT

- Mô hình dịch (translation model):

o Giúp ước lượng xác suất có điều kiện $p(\mathbf{j}|\mathbf{v})$. Xác suất này được ước lượng từ ngữ liệu song ngữ của cặp ngôn ngữ nguồn – đích.

o Có ba hướng tiếp cận chính cho mô hình dịch SMT:

- SMT dựa trên từ (Word – based SMT): được phát triển đầu tiên và đơn vị dịch là các từ. các câu trong ngôn ngữ nguồn sẽ được phân tách thành các từ và được dịch tương ứng một – một sang các từ ở câu trong ngôn ngữ đích.
- SMT dựa trên cụm từ (Phrase – based SMT): Đơn vị dịch ở đây là cụm từ, các câu được phân tách thành các cụm từ. Các cụm từ ở đây không theo nghĩa của ngôn ngữ học mà là trình tự tiếp giáp của nhiều từ trong một câu.
- SMT dựa trên cú pháp (Syntax – based SMT): dựa trên ý tưởng của việc dịch các đơn vị cú pháp (phân tích cây của câu), hơn là những từ đơn hay cụm từ (như trong dịch máy thống kê trên cơ sở cụm từ).

- Mô hình ngôn ngữ (language model): là một thành phần quan trọng của hệ thống SMT. Nó đảm bảo "trôi chảy" cho đầu ra và ảnh hưởng tới việc chọn lựa từ, sắp xếp lại trật tự từ… [5]. Về mặt toán học, nó gán cho mỗi câu một xác suất - khả năng xảy ra câu đó là thế nào trong văn bản.

- Bộ giải mã (decoder): Theo Koehn [5], các mô hình xác suất trong SMT gán điểm số cho tất cả các bản dịch có thể có của một câu đầu vào ở ngôn ngữ nguồn. Mục đích của việc giải mã là tìm bản dịch có điểm số lớn nhất. Trong quá trình giải mã, ta cấu trúc bản dịch theo từng từ với từ, từ đầu đến cuối. Các mô hình dựa trên từ

và cụm từ phù hợp với điều này, vì nó cho phép tính toán điểm số cho các bản dịch một phần (partial translation).

- **Đánh giá chất lượng dịch:**

Do có nhiều bản dịch hợp lệ cho mỗi câu đầu vào nên ta làm thế nào để đánh giá chất lượng bản dịch nào là tốt nhất. Từ đó đưa ra ý tưởng về việc định lượng chất lượng hệ thống MT. Để đánh giá chất lượng chính xác của bản dịch, ta có thể đánh giá bằng các dịch giả hoặc máy tính. Tuy nhiên, nếu bản dịch có kích thước càng lớn thì việc con người thực hiện đánh giá là không khả thi, gây mất thời gian. Hiện nay, các mô hình MT đều sử dụng phương pháp đánh giá tự động. Có một số phương pháp đánh giá tự động như BLEU, NIST…

Ở đây, tôi giới thiệu phương pháp đánh giá tự động phổ biến nhất là BLEU. Ý tưởng chính là so sánh kết quả bản dịch tự động bằng máy với các bản dịch mẫu của con người, bản MT nào càng giống với bản dịch mẫu của con người thì bản dịch đó càng chính xác.

1.2.3. Thảo luận

SMT có những ưu điểm so các phương pháp khác và đây là một hướng phát triển đầy tiềm năng trong MT.

- Dịch máy là vấn đề quyết định: Cho trước những từ trong ngôn ngữ nguồn, chúng ta phải quyết định chọn những từ trong ngôn ngữ đích. Vì vậy, nó tạo cho chúng ta một cảm giác là có thể giải quyết nó bằng định lý, phép toán thống kê. Điều đó dẫn đến cách tiếp cận thống kê được đề xuất. Từ đó ta có thể dựa vào các tính toán thống kê để giải thích các hiện tượng.

- Mối quan hệ giữa các từ, cụm từ và cấu trúc ngữ pháp thường mơ hồ. Để mô hình hóa những quan hệ này, phân phối xác suất và kỹ thuật thống kê cho phép ta giải quyết những vấn đề phụ thuộc nhau.

- Để thực hiện MT, ta nhất thiết phải kết hợp nhiều nguồn trí thức. Trong SMT, chúng ta dựa vào toán học để thực hiện kết hợp tối ưu của các nguồn trí thức.

- Trong dịch máy thống kê, trí thức dịch được học một cách tự động từ dữ 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.

- Một mô hình thống kê có thể được huấn luyện trên số lượng lớn dữ liệu và tăng dữ liệu huấn luyện sẽ cho phép các mô hình nắm bắt thêm các "hiện tượng ngôn ngữ"trong các ngôn ngữ. Do đó, khi tăng số lượng dữ liệu huấn luyện sẽ đưa ra các bản dịch có chất lượng cao hơn.

Chất lượng dịch của hệ thống SMT tỷ lệ thuận với số lượng và chất lượng của ngữ liệu song ngữ sử dụng để phục vụ hệ thống dịch. Tuy nhiên, ngữ liệu song ngữ hiện vẫn còn hạn chế cả về kích thước lẫn chất lượng. Bên cạnh đó, việc phát triển các

phương pháp giúp làm tăng chất lượng dịch dựa trên ngữ liệu hiện có đang là một vấn đề mở. Hiện nay, các nghiên cứu để làm tăng chất lượng dịch vẫn đang được tiến hành phù hợp với từng cặp ngôn ngữ.

1.3. Vấn đề tên riêng, từ mượn trong dịch máy

Như chúng ta thấy, một trong những vấn đề thường xuyên gặp phải của các hệ thống dịch máy là dịch tên riêng, thuật ngữ kỹ thuật, từ không xác định hay các từ mượn. Đối với những ngôn ngữ có hệ thống bảng chữ cái và chữ viết tương tự nhau thì việc dịch các từ này giữa các cặp ngôn ngữ đó không gặp nhiều khó khăn; tuy nhiên, với những cặp ngôn ngữ khác nhau về hệ thống chữ viết cũng như âm thanh thì đây là một thách thức đặt ra cho cả hệ thống dịch máy cũng như dịch giả bởi chúng ta không có cơ sở dữ liệu đầy đủ về những từ này.

Từ đó đưa ra bài toán cho dịch tên riêng và các từ không xác định cần được giải quyết.

1.4. Bài toán dịch tên riêng, chuyển ngữ

Ta thấy hầu hết các hệ thống chữ viết là ngữ âm, tức là chúng phiên âm các âm thanh của các ngôn ngữ, có thể là các âm tiết (như chữ Trung, chữ cái kanji tiếng Nhật) hoặc các phụ âm và nguyên âm riêng biệt (như chữ Latin, chữ Ả rập, chữ cái katakana của tiếng Nhật).

Từ khi việc dịch tên riêng là quá trình ánh xạ các chữ cái (hoặc kí tự) giữa các cặp ngôn ngữ thì nó được gọi là chuyển ngữ (transliteration).

Sau đây, tôi đưa ra một số nội dung cụ thể hơn về Chuyển ngữ.

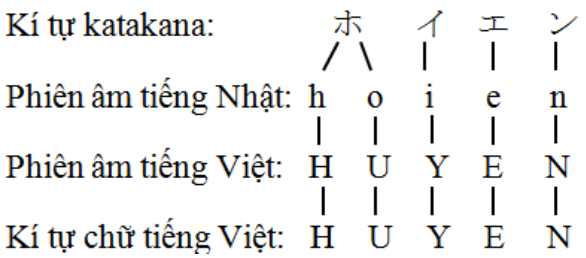
1.4.1. Khái niệm chuyển ngữ

Có nhiều khái niệm được định nghĩa cho chuyển ngữ, cụ thể như sau:

- Chuyển ngữ là việc dịch ngữ âm giữa các cặp ngôn ngữ khác nhau về hệ thống bảng chữ cái và âm thanh [7].
- Chuyển ngữ có thế hiểu là phương thức ánh xạ từ một hệ thống văn bản này thành một hệ thống văn bản khác dựa trên sự tương đồng về mặt ngữ âm. [8]

Do vậy, Chuyển ngữ tự động là quá trình chuyển đổi tự động kịch bản của một từ từ một ngôn ngữ nguồn sang ngôn ngữ đích, trong khi đó vẫn giữ cách phát âm. [12]

Ví dụ về việc chuyển ngữ tên riêng dựa trên phiên âm từ tiếng Nhật sang tiếng Việt như sau:



Hình 1.5. Chuyển ngữ từ tiếng Nhật sang tiếng Việt của tên riêng "Huyền"

Lưu ý, quá trình ánh xạ chữ cái katakana tiếng Nhật tới các phiên âm tiếng Nhật tới các phiên âm tiếng Việt tới các chữ cái tiếng Việt có thể không rõ ràng ở mỗi bước. Trong đó, việc ánh xạ chữ cái katakana tiếng Nhật sang các phiên âm tiếng Nhật được thực hiện chính xác, còn từ các âm thanh tiếng Nhật sang âm thanh tiếng Việt có thể không được ánh xạ chính xác. Ví dụ, trong tiếng Việt có những âm vị khác với tiếng Nhật, ở tiếng Việt có thể có nhưng trong tiếng Nhật lại không có. Việc ánh xạ các phiên âm tiếng Việt sang các chữ cái tiếng Việt thì chính xác bởi tiếng Việt không có sự khác biệt nhiều giữa phiên âm và chữ cái.

1.4.2. Phân biệt Chuyển ngữ (Transliteration) và Biên dịch (Translation)

Chuyển ngữ liên quan đến việc dịch một ngôn ngữ từ một hệ thống chữ viết này sang hệ thống chữ viết khác. Mặc dù nó có vẻ tương tự như biên dịch nhưng chúng là hai quá trình khác nhau với những mục tiêu rất khác nhau. Sau đây là một số khác biệt quan trọng.

- Biên dịch là chuyển đoạn văn trong một kịch bản sang đoạn văn trong kịch bản khác với ý nghĩa tương đương nhau. Biên dịch cho phép các từ trong một ngôn ngữ được hiểu bởi những người nói ngôn ngữ khác. Về cơ bản, biên dịch một từ nước ngoài liên quan đến việc giải thích ý nghĩa của nó.
- Chuyển ngữ là sự chuyển đổi dựa trên cách phát âm; nógiúp cho một ngôn ngữ dễ tiếp cận hơn một chút cho những người không quen với bảng chữ cái của ngôn ngữ đó. Chuyển ngữ tập trung vào việc phát âm hơn là ý nghĩa, nó đặc biệt hữu ích khi thảo luận về người, địa điểm và văn hóa nước ngoài. Chuyển ngữ là tìm các bảng chữ cái tương đương và không quan tâm tới ý nghĩa tương đương của từ hoặc câu.

Vì vậy, nếu chúng ta cần phải đọc văn bản trong ngôn ngữ khác và quan tâm vào việc phát âm hơn là hiểu nó thì chúng ta cần chuyển ngữ, nhưng nếu chúng ta muốn biết nó nghĩa gì thì chúng ta cần biên dịch.

1.4.3. Ứng dụng của Chuyển ngữ

Chuyển ngữ thường được sử dụng phục vụ cho các thư viện hoặc cho quá trình xử lý dữ liệu văn bản. Khi người dùng thực hiện tìm kiếm hoặc đánh chỉ mục nội dung, quá trình chuyển ngữ có thể tìm thấy những thông tin được viết bằng một bảng

chữ cái khác và trả về kịch bản của người dùng. Tính năng chuyển ngữ cũng cho phép sử dụng bàn phím để nhập một văn bản ở định dạng chữ viết này được gõ với một định dạng khác. Ví dụ, với kỹ thuật này có thể sử dụng một bàn phím qwerty để gõ văn bản với bảng chữ cái kirin. [10].

Chuyển ngữ được sử dụng phổ biến hơn là chúng ta nghĩ. Khi chúng ta đọc về tin tức quốc tế, chúng ta nên cần tới sự trợ giúp của chuyển ngữ, ví dụ có thể mọi người khá bối rối nếu các mục tin tức nằm rải rác... Hay chuyển ngữ cũng được sử dụng ở nhiều nơi khác như ở nhà hàng, chúng ta tìm kiếm thực đơn ăn uống; hoặc trong thư viện, nó cho phép mọi người để thực hiện tìm kiếm nội dung trong hệ thống chữ viết khác nhau; trong thế giới học thuật, phục vụ cho việc nghiên cứu các bài báo và trong việc học ngôn ngữ. Đồng thời, nó cũng có trong ngôn ngữ hàng ngày, những từ như ***karate*** (Nhật Bản) và ***pajamas*** (Urdu) được vay mượn bởi tiếng Anh [9].

Hay nói cách đơn giản, nhu cầu sử dụng hệ thống Chuyển ngữ thường dành cho các từ không là từ vựng (out-of-vocabulary words (OOVs)), tức là những từ mà hệ thống không dịch được. OOVs có xu hướng là các tên riêng, có thể là tên địa danh, tên người, ... Từ đó, hướng nghiên cứu và xây dựng luận văn của tôi là bài toán chuyển ngữ tự động cho tên riêng.

1.4.4. Một số khó khăn của bài toán Chuyển ngữ

Luận văn đưa ra bài toán chuyển ngữ xây dựng cho cặp ngôn ngữ Nhật – Việt và theo Kevin Knight [7] đây có thể gọi là bài toán chuyển ngữ theo hướng ngược. Từ đó, nó gặp một số khó khăn chính:

- Chuyển ngữ ngược thì khó thực hiện hơn chuyển ngữ xuôi. Có nhiều cách để viết một từ tiếng Anh như "*switch*" trong katakana, tất cả đều hợp lệ, nhưng ta không có sự linh hoạt trong hướng ngược lại. Ví dụ, ta không thể bỏ chữ "*t*" trong "*switch*", hoặc viết "*arture*" khi ta lấy nghĩa là "*archer*". Hướng xuôi phá bỏ linh hoạt với các giải pháp dựa trên từ điển, bởi vì không có từ điển nào chưa tất cả các biến thể katakana.
- Chuyển ngữ ngược khó hơn việc chuyển ngữ sang chữ Latin. Một chương trình chữ Latin thường thiết lập một phương pháp cho việc viết một kịch bản tiếng nước ngoài trong các văn bản chữ Latin.

Ví dụ, để viết chữ Latin của "アンジラ", ta cần tìm mỗi kí tự trong *Bảng 1.3. Bảng chữ cái Katakana* và các kí tự thay thế. Việc thay thế này đưa ra kí tự Latin là "*anjira*", nhưng không (dịch) là "*angela*". Việc viết chữ Latin thường xác định và có thể đảo ngược mặc dù có thể phát sinh một chút nhập nhằng, mơ hồ.

Không phải tất cả cụm từ katakana có thể được phát âm bởi chuyển ngữ ngược. Một số là cụm từ viết tắt, một số từ "lạ" khó đoán nghĩa (ví dụ: トランプ**(torampu) :** *Tú lơ khơ*), một số từ khác là từ tượng thanh và khó dịch. Những trường hợp ngoại lệ

này phải được giải quyết bởi các kỹ thuật khác hơn những nội dung sẽ trình bày trong luận văn.

1.4.5. Thuộc tính kỳ vọng của quá trình Chuyển ngữ

- Thuộc tính kì vọng nhất của một quá trình chuyển ngữ (ngược) tự động là tính chính xác;
- Có thể sử dụng cho những cặp ngôn ngữ mới như tiếng Ả rập/tiếng Anh với kết quả đạt được một cách tối thiểu, có thể tái sử dụng tài nguyên;
- Chống lại mạnh mẽ các lỗi được đưa ra bởi OCR;
- Tương thích với các tình huống nhận dạng giọng nói trong trường hợp người nói có giọng nói tiếng nước ngoài nặng;
- Có thể giữ đúng ngữ cảnh (đúng chủ đề/cú pháp), hoặc ít nhất có thể trả về một danh sách xếp hạng các bản dịch tiếng Việt có thể có.

Tóm lại, ở chương này, tôi đề cập đến hệ thống dịch máy, dịch máy thống kê và chuyển ngữ tên riêng và các từ không xác định giữa các cặp ngôn ngữ khác nhau.

Bài toán chuyển ngữ tên riêng cũng được ứng dụng trong SMT. Các phương pháp hiện tại của các hệ thống SMT tự động đánh giá dựa vào việc tính các kết hợp chính xác của chuỗi các từ có độ dài khác nhau, ví dụ: BLEU (Papineni và cộng sự, 2001) được nhắc đến ở mục **1.2.2.2**. Do đó, nếu chỉ chuyển ngữ các tên không xác định thì sẽ làm tăng hiệu năng hoạt động.

Để mở rộng các bản dịch có thể được chấp nhận thì đôi khi việc chuyển ngữ các bản dịch tham khảo được đưa vào. Nhưng ngay cả với những cái đó, việc cải thiện hiệu suất hệ thống SMT chuyển ngữ những tên không xác định thì vẫn là một nhiệm vụ khó khăn.

Trong luận văn này, tôi sử dụng hệ thống mã nguồn mở Moses (Koehn và cộng sự, 2007), SMT dựa trên cụm từ để thực hiện thực nghiệm chuyển ngữ tên riêng từ tiếng Nhật sang tiếng Việt.

Luận văn được chia làm 3 chương với bố cục các phần còn lại như sau:

Chương 2: Trình bày nội dung về dịch máy thống kê dựa vào cụm từ và mô hình chuyển ngữ không giám sát

Chương 3: Trình bày nội dung, kết quả thực nghiệm cho dịch máy và chuyển ngữ tự động.

Và cuối cùng là phần kết luận về những vấn đề đã đạt được cùng định hướng nghiên cứu tiếp theo cho luận văn.

CHƯƠNG 2. DỊCH MÁY THỐNG KÊ DỰA VÀO CỤM TỪ VÀ CHUYỂN NGỮ TỪ TIẾNG NHẬT SANG TIẾNG VIỆT

2.1. Dịch máy thống kê dựa vào cụm từ

2.1.1. Giới thiệu

Cách tiếp cận thành công trong hệ dịch máy là dịch dựa vào cụm từ, nghĩa là sử dụng cụm từ làm đơn vị. Trong phương pháp này, câu đầu vào của ngôn ngữ nguồn được chia thành một chuỗi các cụm từ, những cụm từ này được ánh xạ một – một để cho ra được các cụm từ của ngôn ngữ đích, thứ tự của các cụm từ trong ngôn ngữ đích có thể được sắp xếp lại. Thông thường các mô hình cụm từ được ước lượng từ các tập ngữ liệu song ngữ đã được gióng hàng. Tất cả các cặp cụm từ phù hợp với gióng hàng của từ đều được trích xuất. Xác suất được đưa ra dựa trên số lượng tương đối hoặc xác suất dịch từ vựng.

Ví dụ về phân chia cụm từ:

Hình 2.1. Ví dụ về việc phân cụm từ của cặp câu ngôn ngữ Nhật – Việt

Như ở ví dụ trên, câu đầu vào tiếng Nhật là “彼は上手な学生です” được tách thành các cụm từ: 彼, は, 上手, な学生です. Sau đó dịch *một – một* các cụm từ tiếng Nhật sang tiếng Việt: 彼 → Anh ấy, は → là, 上手 → giỏi, な学生です → học sinh. Cuối cùng, có thể sắp xếp lại trật tự các cụm từ tiếng Việt này theo đúng ngữ pháp.

2.1.2. Mục đích của mô hình dịch dựa trên cụm từ

Phương pháp dựa trên từ được ra đời trước và có nhiều hạn chế. Do thiếu thông tin ngữ cảnh khi xác định xác suất của các từ, nên nghĩa của từ được chọn nhiều lúc không đúng với ngữ cảnh. Ngữ nghĩa của từ khi dịch lại phụ thuộc vào các từ khác xuất hiện cùng với nó trong câu, ở đây ngôn ngữ nguồn là tiếng Nhật và các từ trong tiếng Nhật cũng phụ thuộc vào ngữ cảnh để xác định nghĩa của từ. Đôi khi ngữ nghĩa một từ của ngôn ngữ đích không đủ để diễn tả nghĩa một từ trong ngôn ngữ nguồn và ngược lại. Với mô hình dịch song ngữ dựa trên từ thì quá trình xác định nghĩa của câu đích chỉ được thực hiện bởi sự ghép từ và hoán đổi vị trí của từ theo cấu trúc cú pháp. Trong một số trường hợp người ta cần có thêm các thao tác phụ như chèn thêm từ hoặc xóa bớt từ. Thực tế các mô hình dịch theo từ không bảo đảm đúng nghĩa cho câu đích là do nó không có khả năng lưu chứa đủ các luật sinh cho tất cả các câu trong thực tế và các đặc tả chi tiết các hành vi ngữ nghĩa nhúng trong mỗi luật sinh của từng ngữ cảnh cụ thể.

Giải pháp: Để khắc phục những hạn chế trên, một phương pháp MT mới được phát triển là SMT dựa trên cụm từ. Điều này cho phép hệ thống dịch các cụm từ tránh tình trạng dịch word-by-word. Vì có trường hợp một từ trong ngôn ngữ tiếng Việt có nhiều hơn một nghĩa trong ngôn ngữ tiếng Việt.

2.1.3. Định nghĩa bài toán

Nhiệm vụ của một hệ thống SMT là mô hình xác suất dịch $p(\mathbf{v}|\mathbf{j})$, trong đó câu ở ngôn ngữ nguồn **j** được dịch sang câu ở ngôn ngữ đích **v**. Brown và cộng sự [2] đã sử dụng luật Bayes để tính xác suất dịch câu ở ngôn ngữ nguồn **j** sang câu ở ngôn ngữ đích **v** như sau:

$$\begin{aligned} \mathbf{v}^* &= \arg\max_{\mathbf{v}} p(\mathbf{v}|\mathbf{j}) \\ &= \arg\max_{\mathbf{v}} \frac{p(\mathbf{j}|\mathbf{v})\,p(\mathbf{v})}{p(\mathbf{j})} \\ &= \arg\max_{\mathbf{v}} p(\mathbf{j}|\mathbf{v})p(\mathbf{v}) \end{aligned} \qquad (2.1)$$

Trong đó: $p(\mathbf{v})$ là mô hình ngôn ngữ và $p(\mathbf{j}|\mathbf{v})$ là mô hình dịch. Mô hình ngôn ngữ $p(\mathbf{v})$ được ước lượng từ ngữ liệu ở ngôn ngữ đích (ngữ liệu đơn ngữ) và mô hình dịch $p(\mathbf{j}|\mathbf{v})$ được ước lượng từ ngữ liệu song ngữ từ cặp ngôn ngữ Nhật – Việt.

2.1.4. Mô hình dịch

Trong phương pháp này, câu đầu vào được chia thành một chuỗi các cụm từ; những cụm từ được ánh xạ *1-1* đến các cụm từ của câu đầu ra, có thể được sắp xếp lại thứ tự các cụm từ. Chất lượng của bản dịch trong dịch thống kê dựa trên cụm từ phụ thuộc nhiều vào chất lượng của bảng dịch cụm từ (phrase table). Để xây dựng bảng dịch cụm từ đầu tiên, chúng ta tạo ra gióng hàng từ giữa mỗi cặp câu trong ngữ liệu song ngữ, sau đó trích xuất các cặp cụm từ phù hợp với gióng hàng từ.

Khi trích xuất các cặp cụm từ, chúng ta phải chọn cả những cụm từ ngắn và cụm từ dài, vì tất cả đều hữu ích. Các cặp cụm từ này được lưu giữ lại trong bảng cụm từ cùng với xác suất $\phi(\bar{j}_i|\bar{v}_i)$, trong đó:

$$\phi(\bar{j}_i|\bar{v}_i) = \frac{count(\bar{j}|\bar{v})}{\sum_{\bar{j}} count(\bar{j}|\bar{v})}$$

Theo Koehn [1], câu ngôn ngữ nguồn **j** được tách thành I cụm từ $\bar{j}_1, \bar{j}_2, ..., \bar{j}_I$. Mỗi cụm từ $\bar{j}_i$ trong **j** được dịch ra thành một cụm từ $\bar{v}_i$ tương ứng trong **v**. Quá trình này được thực hiện dựa vào phân phối xác suất $\phi(\bar{j}_i|\bar{v}_i)$. Ngoài ra, các cụm từ $\bar{v}_i$ còn được sắp xếp lại theo một thứ tự nhất định dựa trên mô hình chuyển đổi

$d(start_i - end_{i-1})$, với $start_i$ là vị trí bắt đầu của cụm từ $\bar{j}_i$ và end_{i-1} là vị trí kết thúc của cụm từ $\bar{v}_{i-1}$. Khi đó, xác suất dịch $p(\mathbf{j}|\mathbf{v})$ được tính theo công thức:

$$p(\mathbf{j}|\mathbf{v}) = \prod_{i=1}^{I} \phi(\bar{f}_i \mid \bar{e}_i) d(start_i - end_{i-1}) \quad (2.2)$$

2.1.5. Mô hình ngôn ngữ

Các mô hình ngôn ngữ *n-gram* sử dụng giả định Markov để phân tách xác suất của một câu thành tích xác suất của từng từ trong câu, cho giới hạn số lượng các từ đứng trước.

Hay, trong mô hình ngôn ngữ *n-gram*, ta muốn tính xác suất của câu $c = t_1, t_2, \ldots, t_n$ thì xác suất của câu c được phân rã thành tích của các xác suất có điều kiện. Sử dụng quy tắc dây chuyền (chain rule), xác suất của câu $p(c)$ được phân rã thành xác suất của từng từ riêng lẻ $p(t)$.

$$p(t_1, t_2, \ldots, t_n) = p(t_1)p(t_2|t_1)\ldots p(t_n| t_1, t_2, \ldots, t_{n-1}) \quad (2.3)$$

Để ước lượng các phân phối xác suất từ trong công thức (2.3), ta sử dụng xấp xỉ Markov, ta có xác suất xuất hiện của một từ t_n coi như chỉ phụ thuộc vào m từ đứng liền trước nó:

$$p(t_n|t_1, t_2, , t_{n-1}) \simeq p(t_n|t_{n-m}, \ldots, t_{n-2}, t_{n-1}) \quad (2.4)$$

Thông thường, ta chọn giá trị của m dựa trên lượng dữ liệu huấn luyện ta có. Nhiều dữ liệu huấn luyện cho phép chọn m lớn hơn. Mô hình ngôn ngữ *trigram* thường được sử dụng nhiều. Với mô hình *trigram*, ta xem xét hai từ đứng liền trước (tức $m = 2$) để dự đoán từ thứ ba. Điều này đòi hỏi thu thập số liệu thống kê trên các chuỗi gồm ba từ, nên được gọi là *3-gram* (trigram). Ngoài ra, các mô hình ngôn ngữ có thể được ước lượng với *2-gram* (bigram), *1-gram* (unigram),…

Ở đây, ta ước lượng các xác suất *trigram* là $p(t_3|t_1, t_2)$. Để thực hiện công việc này, ta đếm số chuỗi t_1, t_2 theo sau từ t_3 (ký hiệu count(t_1, t_2, t_3)) và số chuỗi t_1, t_2 được theo sau bởi các từ khác (ký hiệu $\sum_t count(t_1, t_2, t)$) trong ngữ liệu huấn luyện.

Theo MLE, ta tính:

$$p(t_3|t_1, t_2) = \frac{count(t_1, t_2, t_3)}{\sum_t count(t_1, t_2, t)} \quad (2.5)$$

2.1.6. Giải mã

Nhiệm vụ của thành phần này là tìm câu **v** ở ngôn ngữ đích sao cho tích $p(\mathbf{j}|\mathbf{v})p(\mathbf{v})$ trong công thức (2.1) đạt giá trị cực đại với mỗi câu đầu vào **j** ở ngôn ngữ nguồn.

Trước khi dịch một câu đầu vào ở ngôn ngữ nguồn, ban đầu ta tham khảo bản dịch và tìm kiếm các lựa chọn dịch thích hợp. Trong quá trình giải mã, ta lưu lại các bản dịch một phần trong một cấu trúc dữ liệu gọi là giả thuyết. Bộ giải mã đưa ra hình

thức mở rộng cho các giả thuyết đó bằng cách quyết định cụm từ dịch tiếp theo. Do sự tính toán phức tạp của bộ giải mã (NP – đầy đủ), ta cần hạn chế không gian tìm kiếm. Để thực hiện việc này, ta tái tổ hợp, dùng kỹ thuật quy hoạch động để loại các bỏ giả thuyết không là phần của bản dịch tốt nhất. Giới hạn cả trật tự từ cũng làm giảm tương đối không gian tìm kiếm. Do không gian tìm kiếm rất lớn nên bộ giải mã thường áp dụng các thuật toán tìm kiếm tối ưu. Thuật toán được đưa ra ở đây là A*, đây là một kỹ thuật tìm kiếm tiêu chuẩn trong trí tuệ nhân tạo.

Thuật toán A* khái quát như sau: tại mỗi bước mở rộng không gian tìm kiếm thì ta sử dụng các hàm ước lượng, đánh giá trọng số để kết quả tìm kiếm luôn tốt nhất có thể và tìm thấy đầu tiên.

2.1.7. Tối ưu hóa và Đánh giá

Như đã trình bày ở phần 1.2.2, phương pháp đánh giá được đưa ra là BLEU. Ở phần này tôi sẽ cụ thể hơn về cách thức.

Tổng quát, với bản MT *T* và bản dịch mẫu *S*, trước hết BLEU thống kê số lần tối thiểu các cụm *n-gram* xuất hiện trong từng cặp câu, sau đó chia cho tổng số cụm *n-gram* trong *T*. Tỉ lệ trùng khớp p_n của *T* và *S* được tính theo công thức:

$$p_n = \frac{\sum_{t \in T} \sum_{n-gram \in t} Count_{clip}(n-gram)}{\sum_{t' \in T} \sum_{n-gram' \in t'} Count_{clip}(n-gram')} \quad (2.7)$$

Trong đó, $Count_{clip}$(*n-gram*) là số lượng tối thiểu cụm n-gram có trong *S* và $Count_{clip}$(*n-gram'*) là số lượng cụm *n-gram'* có trong *T*.

Điểm BLEU đánh giá bản *T* với bản dịch mẫu *S* được tính theo công thức (2.8). trong đó, w_n và *N* lần lượt là trọng số (tổng các trọng số w_n bằng 1) và độ dài (tính theo đơn vị từ) các *n-gram* được sử dụng:

$$\text{BLEU} = \text{BP} * \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (2.8)$$

Với giá trị *BP* được tính theo công thức sau:

$$BP = \begin{cases} 1 & \text{nếu t} > \text{s} \\ e^{1-s/t} & \text{nếu t} \leq \text{s} \end{cases} \quad (2.9)$$

Trong đó, t là độ dài của bản MT *T* và s là độ dài của bản *S*.

Bản dịch nào có điểm BLEU càng cao thì độ trùng khớp giữa bản MT và bản dịch mẫu càng nhiều. Như vậy bản dịch đó càng chính xác.

2.2. Chuyển ngữ từ tiếng Nhật sang tiếng Việt

Phần này sẽ mô tả mô hình chuyển ngữ không giám sát cho những từ chưa được dịch (không xác định) ở hệ thống dịch máy.

- **Ý tưởng**: Theo Koehn [8], ta sử dụng một mô hình chuyển ngữ không giám sát dựa trên thuật toán EM để tạo bộ ngữ liệu chuyển ngữ từ dữ liệu song ngữ đã sắp xếp các từ. Từ đó sử dụng nó để huấn luyện mô hình chuyển ngữ. Koehn đưa ra ba phương pháp để tích hợp việc chuyển ngữ trong khi giải mã và ta thực hiện với hệ thống Moses. Trong luận văn, tôi áp dụng phương pháp Thay thế những từ OOV bởi từ được chuyển ngữ có xác suất cao nhất (1-best transliteration) trong giai đoạn hậu giải mã để tích hợp mô hình chuyển ngữ không giám sát vào hệ thống SMT.

Như tôi đã nói từ trước, khó khăn chính cho việc xây dựng hệ thống chuyển ngữ là thiếu dữ liệu từ các cặp dữ liệu huấn luyện cho chuyển ngữ. Tuy nhiên, bất kỳ dữ liệu song ngữ nào cũng có số lượng phù hợp các cặp từ được chuyển ngữ. Việc khai thác mô hình chuyển ngữ có thể được sử dụng để trích xuất các cặp từ như vậy từ hệ thống song ngữ. Phương pháp chuyển ngữ không giám sát giúp khai thác các cặp ngôn ngữ mà dữ liệu huấn luyện đã có sẵn.

Các bước thực hiện chuyển ngữ:

1. Khai phá chuyển ngữ
2. Huấn luyện mô hình chuyển ngữ
3. Tích hợp mô hình chuyển ngữ vào hệ thống dịch.

Sau đây, tôi mô tả cụ thể về việc triển khai các bước trên như sau.

Thứ nhất, Khai phá chuyển ngữ:

Việc khai phá chuyển ngữ sẽ tìm ra các cặp từ là chuyển ngữ của nhau và tính xác suất cho mỗi cặp từ. Mô hình khai phá gồm hai mô hình con là mô hình chuyển ngữ (transliteration model) và mô hình không chuyển ngữ (non-transliteration model). Trong đó, mô hình chuyển ngữ sẽ đưa ra các cặp từ được chuyển ngữ có xác suất cao hơn với mô hình không chuyển ngữ. Mô hình không chuyển ngữ đưa ra các cặp từ không có quan hệ liên kết kí tự nào giữa chúng.

Ta kí hiệu cặp từ giữa hai ngôn ngữ là (*e, f*).

- *Mô hình chuyển ngữ*
  - Xác suất của cặp từ là:

$$p_{tm}(e, f) = \sum_{a \in A(e,f)} \prod_{j=1}^{|a|} p(q_j) \qquad (2.10)$$

với $A(e, f)$ là tập hợp tất cả các chuỗi có thể có từ các ánh xạ kí tự;

a là một chuỗi ánh xạ bất kỳ;

q_j là một kí tự trong chuỗi ánh xạ.

- *Mô hình không chuyển ngữ*
  - Xác suất của cặp từ là:

$$p_{ntm}(e, f) = \prod_{i=1}^{|e|} p_E(e_i) \prod_{i=1}^{|f|} p_F(f_i) \qquad (2.11)$$

- Mô hình này không thay đổi trong quá trình huấn luyện dữ liệu.

- *Mô hình khai phá chuyển ngữ* (*transliteration mining model*)

Do ko biết trước cặp từ nào là chuyển ngữ của nhau trong bộ dữ liệu là các cặp từ nên ta có thể tính điểm của mỗi cặp từ theo công thức nội suy tuyến tính như sau:

$$p(e, f) = (1-\lambda) p_{tm}(e, f) + \lambda p_{ntm}(e, f) \quad (2.12)$$

Với λ là hệ số, có giá trị trong khoảng (0, 1).

Có thể hiểu xác suất được tính bởi công thức (2.12) giúp dung hòa xác suất được tính giữa hai công thức (2.10) và (2.11) và cuối cùng, xác suất được đưa ra từ công thức này là xác suất cho mỗi cặp từ.

Thứ hai, Huấn luyện mô hình chuyển ngữ không giám sát

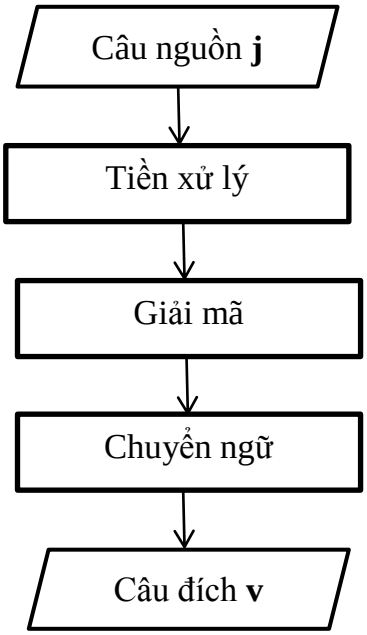
Phương pháp: Sử dụng mô hình SMT dựa trên cụm từ để học mô hình chuyển ngữ. Dữ liệu huấn luyện là các cặp từ, tôi tách thành các ký tự và học hệ thống dịch cụm từ trên các cặp ký tự.

Tôi chia ngữ liệu huấn luyện thành các kí tự, thực hiện phân cụm từ cho các cặp kí tự. Mô hình chuyển ngữ giả sử rằng thứ tự các ký tự ở từ nguồn và từ đích là không thay đổi nên tôi không sử dụng mô hình sắp xếp trật tự từ. Tôi sử dụng 4 đặc trưng cơ bản là: các đặc trưng dịch dựa trên cụm từ (dịch trực tiếp, dịch đảo cụm từ và đánh trọng số từ vựng), mô hình ngôn ngữ (được xây dựng từ phía đích của ngữ liệu chuyển ngữ đã được khai phá), điểm phạt của từ và cụm từ. Các trọng số được điều chỉnh trong một bộ gồm 1000 cặp từ được chuyển ngữ.

Thứ ba, Tích hợp chuyển ngữ vào MT

Thay thế các từ OOV ở đầu ra bởi từ được chuyển ngữ tốt nhất. Kết quả chỉ phụ thuộc vào độ chính xác của mô hình chuyển ngữ trình bày ở trên. Ngoài ra, phương pháp này bỏ qua ngữ cảnh cũng có thể dẫn tới sự chuyển ngữ không chính xác.

Khi đó, sơ đồ dịch của hệ thống MT là:



Hình 2.2. Sơ đồ dịch của hệ thống MT sau khi tích hợp chuyển ngữ

CHƯƠNG 3. THỬ NGHIỆM

3.1. Môi trường triển khai

- Phần cứng: Bộ xử lý Core i5 -3437U CPU 2.40GHz, RAM 4GB.
- Phần mềm: Hệ điều hành Ubuntu 16.04 64 bit.

3.2. Dữ liệu

- Dữ liệu đầu vào là bộ dữ liệu song ngữ Nhật – Việt, gồm gần 40000 cặp câu Nhật – Việt được thu thập từ các nguồn Wiki, TED.
- Tiền xử lý văn bản dùng công cụ tách từ để gộp các từ vào thành 1 cụm từ.
  - Công cụ tách từ tiếng Nhật: Mecab
    https://pypi.python.org/pypi/mecab-python3
  - Công cụ tách từ tiếng Việt: Vitk
    https://github.com/phuonglh/vn.vitk

3.3. Công cụ cho hệ dịch máy

3.3.1. Moses

Theo Koehn [5], Moses là một bộ công cụ MT mã nguồn mở. Nó là sự thực hiện của bộ giải mã dựa trên cụm từ bao gồm cả việc huấn luyện dữ liệu và được xây dựng bởi trường Đại học Edinburgh; sau đó, nó được phát triển trong một hội thảo của trường Đại học Johns Hopkins. Moses được sử dụng rộng rãi trong cộng đồng nghiên cứu phát triển.

Moses 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 đánh giá cho các mô hình này sử dụng huấn luyện với tối ưu lỗi và đánh giá kết quả dịch sử dụng điểm BLEU.

3.3.2. GIZA

GIZA++ là chương trình dùng để gióng hàng từ và trình tự của các từ trong bộ ngữ liệu song ngữ nhằm mục đích liên kết các mô hình phụ thuộc vào lớp từ. Nó là phương pháp gióng hàng từ không giám sát tốt nhất hiện nay.

GIZA++ là việc thực hiện mô hình dựa trên từ của IBM, nó được phát triển bởi hội thảo trường Đại học Johns Hopkins và sau đó được cải tiến bởi Franz Och (2003).

3.3.3. KenLM

Đánh giá mô hình ngôn ngữ với sự cải tiến của Kneser-Ney. Việc xây dựng dựa trên ổ đĩa: ta xác định số lượng RAM cần dùng và nó thực hiện sắp xếp dựa trên ổ đĩa khi cần thiết.

3.3.4. MERT (Minimum Error Rate Training)

MERT là việc điều chỉnh tham số với một thước đo lỗi cụ thể trong việc khai thác dữ liệu. Ta muốn điều chỉnh các tham số để ta đạt được điểm BLEU tối ưu trong bộ điều chỉnh (turing set). MERT được dùng trong Moses để tối ưu hóa hiệu năng của hệ thống dịch.

3.4. Thiết lập mặc định

Các thông số và dữ liệu được thiết lập mặc định trong quá trình huấn luyện như sau:

- **Độ dài cụm từ lớn nhất**: 3
- **Dữ liệu mô hình ngôn ngữ**: tất cả
- **N-gram cho mô hình ngôn ngữ**: 3
- **Các tham số mô hình**
  Distortion: 0.0775344
  Language Model: 0.0775344
  Translation Model: 0.110447, 0.053495, 0.0266803, 0.0686311
  WordPenalty: -0.279847
  PhrasePenalty: -0.306445
  UnknownWordPenalty: 1

3.5. Kết quả thực nghiệm

3.5.1. Dữ liệu đầu vào

	Ngôn ngữ	Số câu thực nghiệm
Dữ liệu huấn luyện	Tiếng Nhật	40000 câu
	Tiếng Việt	40000 câu
Dữ liệu điều chỉnh tham số	Tiếng Nhật	950 câu
	Tiếng Việt	950 câu
Dữ liệu đánh giá	Tiếng Nhật	1000 câu
	Tiếng Việt	1000 câu

- **Độ dài trung bình câu tiếng Nhật:** 39.3 từ.
- **Độ dài trung bình câu tiếng Việt:** 25.8 từ.

3.5.2. Quá trình xử lý dữ liệu và huấn luyện

3.5.2.1. Xử lý dữ liệu cho hệ thống MT

- Từ các tệp tin dữ liệu đầu vào, tôi tiến hành tách từ bằng việc sử dụng các công cụ tách từ đã nói ở trên.

3.5.2.2. Huấn luyện mô hình ngôn ngữ

Như đã trình bày trong các nội dung ở trên, tôi sử dụng mô hình ngôn ngữ *trigram (3-gram)* được huấn luyện từ 12481 từ tiếng Việt.

Tài liệu về KenLM đưa ra các giải thích về các tùy chọn dòng lệnh. Sau đó nhị phân các tập tin *. arpa.en sử dụng KenLM để tải nhanh hơn.

3.5.2.3. Huấn luyện mô hình dịch

Tôi thực hiện huấn luyện thông qua kích thước tập dữ liệu huấn luyện thay đổi tăng dần (với số lượng cặp câu Nhật – Việt tương ứng lần lượt là: 5000, 10000, 15000, 20000, 30000 và 40000).

Bảng 3.1. Kết quả chất lượng dịch khi tăng dần kích thước dữ liệu huấn luyện

Kích thước dữ liệu (số lượng cặp câu)	**Điểm BLEU**
5000	9.88
10000	10.02
15000	10.07
20000	11.02
30000	11.88
40000	12.39

Nhìn vào bảng 3.1 ở trên, ta dễ dàng nhận thấy với kích thước dữ liệu càng lớn thì điểm BLEU càng cao tương ứng chất lượng dịch càng tốt.

- Một số ví dụ dịch khi chưa tích hợp chuyển ngữ:

Bảng 3.2. Một số ví dụ của hệ thống dịch máy khi chưa tích hợp chuyển ngữ

STT	**Câu tiếng Nhật**	**Câu tiếng Việt**
1	ウクライナのドネツク市で炭坑の爆発で少なくとも８０人が死亡し、２０人が行方不明だと報告された。	tại thành_phố ドネツク của ukraine trong vụ nổ ở mỏ có ít_nhất 80 người chết , 20 người mất_tích và đã được báo_cáo .
2	組合の推定によると、２００６から２００７年にほぼ２５０人の鉱夫が事故で死んだ。	theo ước_tính của hiệp_hội , từ 2006 đến năm 2007 gần 250 thợ mỏ thiệt_mạng trong vụ tai_nạn .
3	ウェブ上の最大の検索エンジンGoogleはいろいろなサービスを通して毎日２億以上の問合わせを受ける。	các trang web lớn nhất của công_cụ tìm_kiếm của google thông_qua dịch_vụ nhiều hơn hai triệu mỗi ngày với phép_tính .
4	爆弾が安全に信管を外される間、20人が自宅を避難した。	quả bom đã được tháo ngòi_nổ an_toàn , có 20 người trong nhà đã được sơ_tán .

5	ジャスティン・ヤクと彼の妻も死亡が確認されている。	ジャスティン・ヤク và vợ của ông cũng được xác_nhận đã thiệt_mạng .

Nhìn vào một số câu được dịch từ hệ dịch máy như ở ví dụ trên thì ta thấy kết quả dịch của hệ thống vẫn còn tồn tại một số câu chứa những từ không xác định hay chưa được dịch. Khi đó, tôi sử dụng mô hình chuyển ngữ cho các từ này vào giai đoạn hậu giải mã của hệ thống dịch. Kết quả được trình bày ở phần tiếp theo.

3.5.2.4. Huấn luyện mô hình chuyển ngữ

- Dữ liệu được trích xuất từ bộ dữ liệu gồm 40000 cặp câu song ngữ là 12481 cặp từ dùng để huấn luyện cho mô hình chuyển ngữ. Số lượng cặp từ này được lấy theo các công thức (3.1), (3.2) và (3.3) ở chương 2.

- Hệ số $\lambda = 0.2$ được lấy trong thực nghiệm.

- Sau khi huấn luyện xong, tôi thực hiện chuyển ngữ cho các từ không xác định gồm các tên riêng (từ không có nghĩa) và các từ có nghĩa khác trong file kết quả dịch của mô hình dịch máy.

Đầu tiên, tôi thống kê số lượng các từ không xác định (không dịch được) như bảng 3.3 sau:

Bảng 3.3. Thống kê số lượng từ không xác định của hệ dịch máy dựa trên cụm từ

Từ không xác định	**Số lượng (từ)**	**Tỉ lệ (%)**
Tên riêng	708	81.1
Từ có nghĩa	165	18.9
Tổng	**873**	**100**

Từ bảng 3.3, ta thấy tổng số các từ không xác định từ hệ dịch máy là 523 từ, trong đó có 358 từ tên riêng và 165 từ có nghĩa khác.

Sau khi thống kê tổng số lượng các từ không xác định được bao gồm tên riêng và các từ có nghĩa, tôi áp dụng chuyển ngữ cho các từ này bằng mô hình chuyển ngữ không giám sát. Kết quả chuyển ngữ sẽ đưa ra những từ có thể được chuyển ngữ đúng và chuyển ngữ sai như bảng 3.4 sau:

Bảng 3.4. Thống kê kết quả chuyển ngữ cho các từ không xác định từ hệ dịch máy

Từ không xác định	**Chuyển ngữ đúng (số từ)**	**Tỉ lệ đúng (%)**	**Chuyển ngữ sai (số từ)**	**Tỉ lệ sai (%)**
Tên riêng	116	16.38	592	83.62
Từ có nghĩa	38	23.03	127	76.97
Tổng	**154**	**17.64**	**719**	**82.36**

Nhìn vào kết quả ở bảng 3.4, các từ không xác định được từ hệ dịch máy sẽ được chuyển ngữ và kết quả đầu ra là thêm một lượng các từ được chuyển ngữ đúng. Trong đó:

- Từ tên riêng được chuyển ngữ đúng: 116 từ/708 từ tên riêng được chuyển ngữ, tương ứng 16.38 % trên tổng số từ tên riêng được chuyển ngữ.
- Từ có nghĩa khác được chuyển ngữ đúng: 38 từ/165 từ có nghĩa khác được chuyển ngữ, tương ứng 23.03% trên tổng số từ có nghĩa được chuyển ngữ.
- Tổng số từ được chuyển ngữ đúng (gồm tên riêng và từ có nghĩa khác): 154 từ/873 từ không xác định, tương ứng 17.64% trên tổng số tất cả các từ không xác định từ hệ dịch máy.

Đồng thời, tôi thống kê được số lượng câu được dịch đúng và số kí tự được dịch đúng trong hệ dịch máy trước và sau khi được tích hợp chuyển ngữ như sau:

	Chưa tích hợp chuyển ngữ	**Đã tích hợp chuyển ngữ**
Số câu được dịch đúng	325/1000 (câu)	356/1000 (câu)
Số kí tự dịch đúng	231895	245387

Một số ví dụ về việc chuyển ngữ:

- ***Chuyển ngữ đúng:***
  - Tên riêng:

STT	**Tên riêng tiếng Nhật**	**Tên riêng tiếng Việt**
1	ドネツク	donetsk
2	ブレンダン・テイラー	brendan_taylor
3	アリゴテ	aligote
4	ホア	Hoa
5	ティエップ	Tiếp

  - Từ có nghĩa:

STT	**Từ tiếng Nhật**	**Từ tiếng Việt**
1	混ざっ	n
2	トウェンティ	twente
3	成	đ
4	取り壊さ	phá_huỷ
5	切ら	ệm

- ***Chuyển ngữ sai:***
    - Tên riêng:

STT	Tên riêng tiếng Nhật	Tên riêng tiếng Việt
1	ビクトル・ヤヌコビッチ	biktl_yanoucobiuc
2	ライン	line
3	ツアン	zan
4	カイン	caine
5	ホウオン	howon

    - Từ có nghĩa:

STT	Từ tiếng Nhật	Từ tiếng Việt
1	乗っ取っ	nganh
2	灯さ	ang
3	運び込む	ép
4	青白かっ	mặn
5	取り乱し	ổn

Khi đó, các câu trong ngôn ngữ đích sẽ có thêm những câu được dịch đúng và chính xác hơn.

Một số ví dụ cho việc dịch đúng khi tích hợp chuyển ngữ:

STT	Câu tiếng Nhật	Câu tiếng Việt
1	ウクライナの**ドネツク**市で炭坑の爆発で少なくとも８０人が死亡し、２０人が行方不明だと報告された。	một vụ nổ tại một mỏ than đã giết chết ít_nhất 80 người ở thành_phố donetsk , ukraina , trong khi 20 người được báo_cáo là mất_tích .
2	**ジャスティン・ヤク**と彼の妻も死亡が確認されている。	justin_yak và vợ của ông cũng được xác_nhận là đã chết .
3	**アジンホスメチル**は、第二次世界大戦中に使用された神経剤に由来する危険な神経毒である。	azinphos methyl là một chất_độc thần_kinh nguy_hiểm có nguồn_gốc từ chất_độc thần_kinh được sử_dụng trong thế_chiến thứ ii .

Như vậy, sau khi tôi tích hợp mô hình chuyển ngữ không giám sát vào hệ dịch máy thì điểm BLEU sẽ tăng từ 12.39 lên 12.57. Điểm BLEU tăng bởi kết quả được tính thêm tỉ lệ chuyển ngữ đúng cho các từ không được dịch từ hệ dịch máy. Do đó, chất lượng dịch của hệ dịch máy chính xác hơn.

Tuy nhiên, trong phần thực nghiệm của luận văn, do bị hạn chế bởi số lượng bộ dữ liệu song ngữ Nhật – Việt nên điểm BLUE chưa cao. Trong tương lai, để nâng cao chất lượng dịch cũng như chuyển ngữ thì cần phát triển thêm bộ dữ liệu song ngữ.

KẾT LUẬN

Luận văn đã trình bày những kiến thức cơ bản về bài toán chuyển ngữ, ứng dụng trong dịch máy thống kê; tìm hiểu về mô hình dịch máy thống kê dựa vào cụm từ; nghiên cứu phương pháp chuyển ngữ không giám sát và thử nghiệm cho cặp ngôn ngữ Nhật – Việt khi tích hợp chuyển ngữ và không tích hợp chuyển ngữ vào dịch máy thống kê dựa vào cụm từ. Từ đó, ta thấy việc đưa chuyển ngữ vào bài toán dịch máy là hoàn toàn hợp lý và cần thiết để kết quả dịch chính xác và tối ưu hơn.

Hướng nghiên cứu tiếp của luận văn:

- Tiếp tục xây dựng thêm bộ ngữ liệu song ngữ, nghiên cứu thêm về phương pháp chuyển ngữ không giám sát cùng các phương pháp chuyển ngữ khác để chuyển ngữ cho những tên riêng, các từ không xác định khác.
- Tích hợp chuyển ngữ vào giao đoạn giải mã để cải tiến chất lượng cũng như hiệu năng của hệ thống dịch máy.

TÀI LIỆU THAM KHẢO

Tiếng Việt:

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

[2]. VNLP – Nhóm xử lý ngôn ngữ tự nhiên cho tiếng Việt (2015), Hệ thống âm vị, http://vnlp.net/ti%E1%BA%BFng-vi%E1%BB%87t-c%C6%A1-b%E1%BA%A3n/h%E1%BB%87-th%E1%BB%91ng-am-v%E1%BB%8B/

[3]. Lê Quang Hùng (2015), *Khai phá tri thức song ngữ và ứng dụng trong dịch máy Anh – Việt*, Luận án Tiến sĩ Khoa học Máy tính, Đại học Quốc gia Hà Nội, Trường Đại học Công nghệ, Hà Nội.

[4]. Ngô Hương Lan, Hồ Hoàng Hoa (2008), Một số đặc điểm của tiếng Nhật, *Tạp chí Nghiên cứu Đông Bắc Á*, Số 7, đăng ngày 30/10/2012, trên trang http://www.inas.gov.vn/403-mot-so-dac-diem-cua-tieng-nhat.html

Tiếng Anh:

[5]. Philipp Koehn (2009), *Statistical Machine Translation*, School of Informatics, University of Edinburgh, Cambridge University Press.

[6]. David Matthews (2007), *Machine Transliteration of Proper Names*, Master of Science, School of Informatics, University of Edinburgh.

[7]. Kevin Knight, Jonathan Graehl (1998), Machine Transliteration, *Computational Linguistics*, Volume 24, Number 4, pp. 599-612

[8]. Hieu Hoang, Philipp Koehn (et.al, 2014), Integrating an Unsupervised Transliteration Model into Statistical Machine Translation, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 148–153, Gothenburg, Sweden, April 26-30 2014. © 2014 Association for Computational Linguistics.

[9]. Sarvnaz Karimi, Falk Scholer, Andrew Turpin (2011), Machine Transliteration Survey, *ACM Computing Surveys*, Vol. 43, No. 3, pp. 17:0 – 17:46, Article 17, Publication date: April 2011, DOI: 10.1145/1922649.1922654·Source: DBLP.

[10]. Hoang Gia Ngo, Nancy F. Chen, Sunil Sivadas, Bin Ma, Haizhou Li (2014), A Minimal-Resource Transliteration Framework for Vietnamese, Published in INTERSPEECH, Singapore.

[11]. Philipp Koehn (2017), *Statistical Machine Translation - Chapter 13: Neural Machine Translation*, Center for Speech and Language Processing, Department of Computer Science, Johns Hopkins University.

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