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

Chuyên ngành: Kỹ thuật Phần mềm
Mã số: 60480103

TÓM TẮT LUẬN VĂN THẠC SĨ

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 KÍ HIỆU CÁC CHỮ CÁI VIẾT TẮT

BLEU	**Bi**Lingual **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

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

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.

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

1.1.1.2.1. Âm tố

1.1.1.2.2. Âm vị

1.1.1.2.3. Tiếng

1.1.1.2.4. Hình vị

1.1.1.3. Từ vựng

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.

1.1.1.4. Ngữ pháp

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

1.1.2. Tiếng Nhật

1.1.2.1. Hệ thống bảng chữ cái tiếng Nhật

Tiếng Nhật có 3 bảng chữ là hiragana, katakana và chữ Hán (kanji). Hệ thống các bảng chữ cái này được sử dụng linh hoạt, trong một câu có thể được kết hợp từ chữ của **cả 3 bảng chữ cái** trên.

Hình 1.1. Bảng chữ cái Katakana

1.1.2.2. Ngữ âm

Âm tiết trong tiếng Nhật 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.

1.1.2.3. Từ vựng

Tiếng Nhật có một vốn từ vựng rất lớn và vô cùng phong phú, gồm nhiều tầng lớp từ vựng và chúng có khả năng kết hợp với nhau tạo ra từ mới.

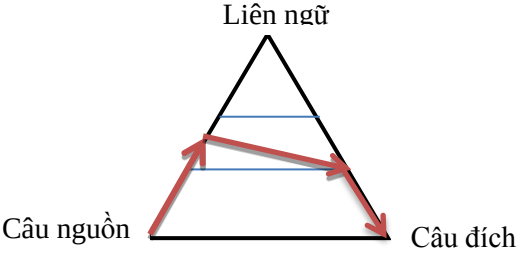
1.1.2.4. Ngữ pháp

Trong tiếng Nhật, động từ thường đứng cuối câu.

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

1.2.1. Bài toán dịch máy

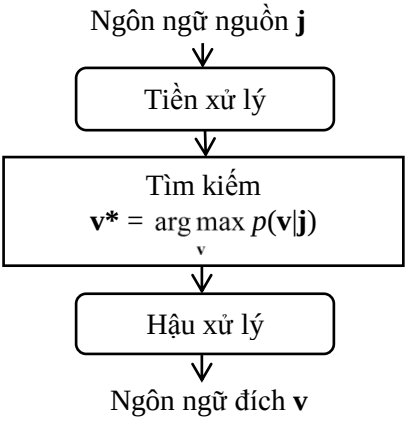
Ta có thể hiểu MT là việc dịch tự động, nó là quá trình mà phần mềm máy tính dịch văn bản từ một ngôn ngữ (ngôn ngữ nguồn) sang một ngôn ngữ khác (ngôn ngữ đích).



Hình 1.2. Tam giác thể hiện quá trình dịch máy

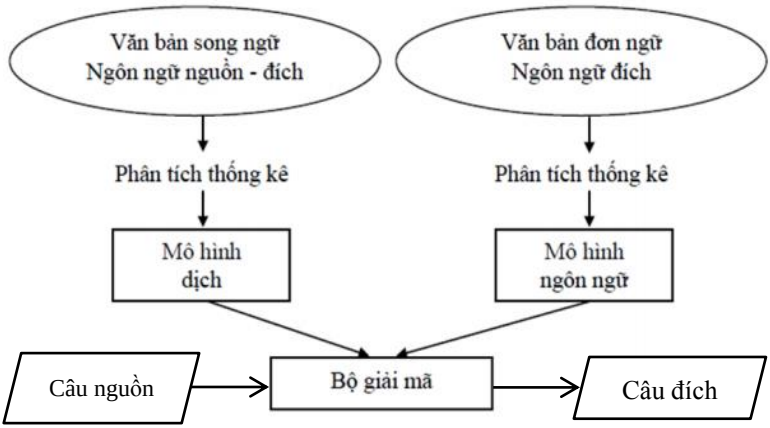
1.2.2. Dịch máy thống kê

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 sau: $v^* = \arg\max p(v|j)$ (1.1)



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

- **Các thành phần của SMT:**



Hình 1.4. Các thành phần của hệ dịch máy SMT

- **Đánh giá chất lượng dịch:**

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

Ưu điểm của SMT:

- 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.

- Mô hình hóa những mối quan hệ giữa các từ, cụm từ và cấu trúc ngữ pháp thường mơ hồ bằng phân phối xác suất và kỹ thuật thống kê.

- Trong SMT, chúng ta dựa vào toán học để thực hiện kết hợp tối ưu của các nguồn trí thức.

- Việc phát triển một hệ dịch dựa vào thống kê sẽ rất nhanh so với hệ dịch dựa vào luật.

- Tăng số lượng dữ liệu huấn luyện sẽ đưa ra các bản dịch có chất lượng cao hơn.

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 hay các từ mượn. Đối 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.

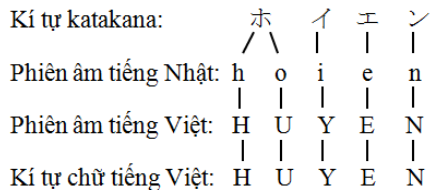
1.4. Bài toán dịch tên riêng, chuyển ngữ

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.

1.4.1. Khái niệm chuyển ngữ

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:



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

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

1.4.3. Ứng dụng của Chuyển ngữ

1.4.4. Một số khó khăn của bài toán Chuyển ngữ

1.4.5. Thuộc tính kỳ vọng của quá trình Chuyển ngữ

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.

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.

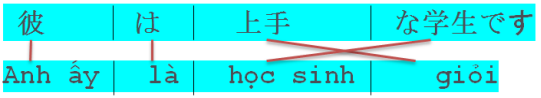
CHƯƠNG 2. DỊCH MÁY THỐNG KÊ DỰA VÀO CỤM TỪ VÀ CHUYỂN NGỮ TỪ TIẾNG NHẬT SANG TIẾNG VIỆT

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.

Ví dụ về phân chia cụm từ:



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

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

Để khắc phục những hạn chế của phương pháp SMT dựa trên 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:

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

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

Mô hình dịch (translation model) giúp ước lượng xác suất có điều kiện p(**j**|**v**). Xác suất này được ước lượng từ ngữ liệu song ngữ của cặp ngôn ngữ nguồn – đích.

2.1.5. Mô hình ngôn ngữ

Về mặt toán học, mô hình ngôn ngữ gán cho mỗi câu một xác suất - khả năng xảy ra câu đó là thế nào trong văn bản.

Mô hình ngôn ngữ *trigram* thường được sử dụng nhiều.

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.

2.1.7. Tối ưu hóa và Đánh giá

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:

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

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

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

2.2. Chuyển ngữ từ tiếng Nhật sang tiếng Việt

Phần này mô tả mô hình chuyển ngữ không giám sát cho những từ chưa được dịch 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. 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.

Các bước thực hiện chuyển ngữ:

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 và mô hình không chuyển ngữ.

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

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

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).

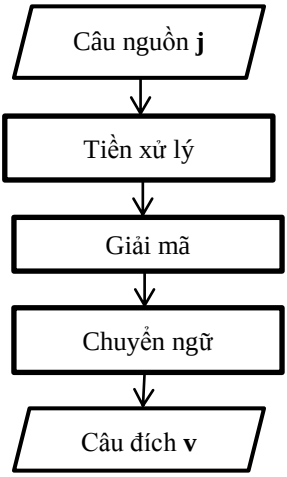
Thứ hai, Huấn luyện mô hình chuyển ngữ không giám sát

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, tách thành các ký tự và học hệ thống dịch cụm từ trên các cặp ký tự.

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 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

3.3.2. GIZA

3.3.3. KenLM

3.3.4. MERT (Minimum Error Rate Training)

3.4. Thiết lập mặc định

- **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

3.5.2.2. Huấn luyện mô hình ngôn ngữ

3.5.2.3. Huấn luyện mô hình dịch

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

- 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 .

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.

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**

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**

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	ホア	Hoa
3	ティエップ	Tiếp

  - Từ có nghĩa:

STT	Từ tiếng Nhật	Từ tiếng Việt
1	トウェンティ	twente
2	取り壊さ	phá_huỷ
3	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

  - 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

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/