ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

PHẠM NGHĨA LUÂN

NGHIÊN CỨU THÍCH ỨNG MIỀN TRONG DỊCH MÁY THỐNG KÊ ANH - VIỆT

Chuyên ngành: Hệ thống thông tin

Mã số: 9480104.01

TÓM TẮT LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

Hà Nội – 2022

Công trình được hoàn thành tại
Trường Đại học Công nghệ, Đại học Quốc Gia Hà Nội

Người hướng dẫn khoa học:
1. TS. Nguyễn Văn Vinh
2. TS. Phạm Việt Thắng

Phản biện 1:.....

Phản biện 2:.....

Phản biện 3:.....

Luận án sẽ được bảo vệ trước Hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ

họp tại.....

vào hồi.....giờ.....ngày.....tháng.....năm.....

Có thể tìm hiểu luận án tại:
- Thư viện Quốc gia Việt Nam
- Trung tâm Thông tin - Thư viện, Đại học Quốc gia Hà Nội

MỤC LỤC

Chương 1. MỞ ĐẦU

1.1. Đặt vấn đề

Ngày nay, nhu cầu trao đổi thông tin giữa các quốc gia, các nền văn hóa ngày càng tăng làm cho nhu cầu dịch thuật trở nên cần thiết. Quá trình dịch thủ công bởi con người cho chất lượng cao nhưng tốc độ chậm, năng suất thấp và chi phí lớn mà không thể tái sử dụng. Hơn nữa, một phiên dịch viên dù giỏi đến đâu cũng không thể dịch tốt được tất cả các lĩnh vực, các ngôn ngữ khác nhau. Vì vậy, hệ thống dịch tự động bằng máy tính là cần thiết để trợ giúp cho quá trình dịch thuật.

Hiện nay có nhiều sản phẩm dịch tự động được thương mại và sử dụng phổ biến như (hệ dịch *Google Translate*[1] của Google, Bing Translator [2] của Microsoft,...) và mang lại kết quả nổi bật. Tuy nhiên, các mô hình dịch máy thường dịch sai khi dịch các từ, cụm từ hoặc các câu thuộc lĩnh vực, chủ đề khác với chủ đề của các câu được sử dụng huấn luyện mô hình, ví dụ các câu dịch thuộc lĩnh vực thể thao nhưng các câu được sử dụng để đào tạo mô hình dịch máy thuộc lĩnh vực y tế. Do đó, để đạt được chất lượng dịch cao trong một lĩnh vực nhất định, chúng ta phải điều chỉnh mô hình dịch máy cho lĩnh vực cụ thể đó. Các nghiên cứu về thích ứng miền trong dịch máy chủ yếu theo hai hướng tiếp cận chính là (1) các kĩ thuật để cải tiến mô hình và (2) các kĩ thuật để tăng cường, cải tiến chất lượng của dữ liệu huấn luyện.

Hiện nay, nghiên cứu về thích ứng miền trong dịch máy thống kê Anh-Việt vẫn còn một số tồn tại, thách thức:

- Thiếu tài nguyên song ngữ, chưa tận dụng được hết các dạng tài nguyên, dữ liệu song ngữ miền hạn chế về số lượng, chất lượng.
- Các nghiên cứu chủ yếu áp dụng cho các cặp ngôn ngữ phổ biến, chưa có nhiều nghiên cứu cho cặp ngôn ngữ Anh-Việt.

Nhằm góp phần giải quyết các vấn đề nêu trên, nghiên cứu sinh đã chọn đề tài *"Nghiên cứu thích ứng miền trong dịch máy thống kê Anh-Việt"*.

1.2. Mục tiêu của luận án

Mục tiêu chung: đề xuất các giải pháp để cải tiến chất lượng hệ thống dịch máy thống kê với cặp ngôn ngữ Anh-Việt. Các mục tiêu cụ thể gồm:

- Đề xuất được các giải pháp nâng cao chất lượng dịch theo miền của hệ dịch thống kê cho cặp ngôn ngữ Anh–Việt;
- Nghiên cứu đề xuất các phương pháp tăng cường thêm dữ liệu song ngữ để huấn luyện, cải thiện chất lượng dịch máy thống kê;
- Nghiên cứu các hệ thống dịch thống kê đã có như Moses, dịch máy mạng nơ-ron, các phương pháp tích hợp tri thức ngôn ngữ, đề xuất các phương pháp mới, thực nghiệm.

[1] https://translate.google.com/
[2] https://www.bing.com/translator

1.3. Đóng góp chính của luận án

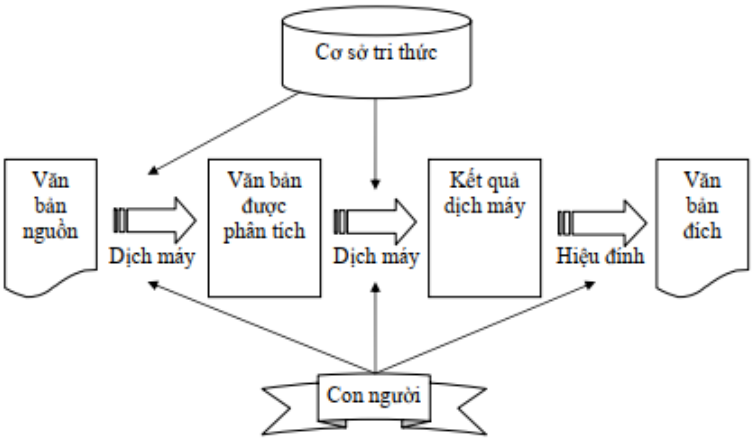
Luận án có ba đóng góp chính, cụ thể:

- Đề xuất phương pháp tinh chỉnh bảng dịch cụm từ (phrase-table) theo định hướng miền dựa vào phân loại miền các cụm từ trong bảng cụm từ, từ đó điều chỉnh, cập nhật lại giá trị xác suất của các cụm từ theo hướng ưu tiên hơn trong miền đích [4].
- Đề xuất phương pháp sinh tự động dữ liệu song ngữ cho dịch máy sử dụng kĩ thuật dịch ngược để tận dụng nguồn dữ liệu đơn ngữ có sẵn [5].
- Đề xuất phương pháp nâng cao chất lượng, hiệu quả của phương pháp sinh tự động dữ liệu song ngữ cho dịch máy với giải pháp tiền xử lý, giảm nhiễu cho các văn bản đầu vào cho dịch ngược, quá trình tiền xử lý này được thực hiện bởi mô hình sửa lỗi chính tả, ngữ pháp [7] nhờ đó kết quả đầu ra của dịch ngược tốt hơn [8].

Chương 2. KIẾN THỨC CƠ SỞ

2.1. Tổng quan về dịch máy

Dịch máy (Machine Translation), còn được gọi là dịch tự động, có lịch sử phát triển lâu đời. Khái niệm dịch máy được nhiều tác giả định nghĩa, tuy có một vài điểm khác biệt nhưng hầu hết đều tương đương với định nghĩa sau: Dịch máy là một hệ thống sử dụng máy tính để chuyển đổi văn bản được viết trong ngôn ngữ tự nhiên này thành bản dịch tương ứng trong ngôn ngữ tự nhiên khác.



Hình 2.1: *Mô tả hệ thống dịch máy*

Hình 2.1 mô tả hệ thống dịch máy, đầu vào là một văn bản trong ngôn ngữ nguồn, quá trình dịch chia thành hai giai đoạn. Giai đoạn một, văn bản được phân tích thành các thành phần. Giai đoạn hai, các thành phần được dịch thành văn bản ở ngôn ngữ đích. Kết quả dịch có thể được hiệu đính bởi con người để có bản dịch tốt.

2.2. Dịch máy thống kê

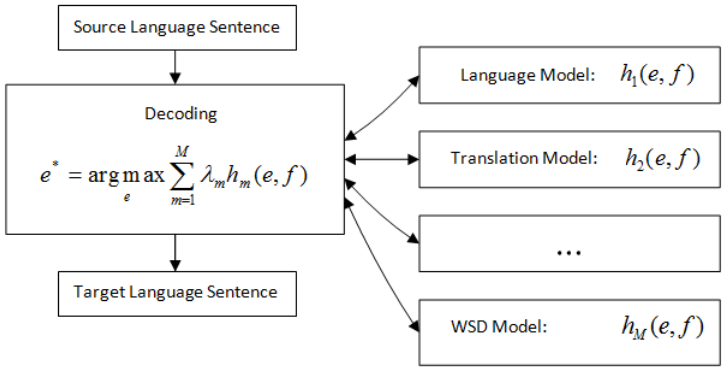
2.2.1. Cơ sở toán học

Dịch máy thống kê được Brown và cộng sự đề xuất năm 1990. Hình 2.2 mô tả kiến trúc cơ bản của một hệ dịch máy thống kê, trong đó:

- f là câu nguồn gồm j từ ($f = f_1, ..., f_j$).
- e như một câu đích gồm i từ ($e = e_1, ..., e_i$).

Giả sử câu nguồn f là tiếng Pháp, câu đích e là tiếng Anh thì của câu f phù hợp nhất có thể tìm được xác định qua tìm kiếm các câu tiếng Anh e để cực đại hóa điều kiện xác suất $p(e|f)$, được mô tả như 2.1:

$$e_{best} = arg\max_{e} p(e|f) \tag{2.1}$$



Hình 2.2: *Kiến trúc cơ bản của hệ thống dịch máy thống kê*

Áp dụng quy tắc Bayes, chia quá trình này thành hai mô hình: mô hình ngôn ngữ $p(e)$ và mô hình dịch $p(f|e)$ như công thức 2.2.

$$arg\max_{e} p(e|f) = arg\max_{e} \frac{p(e) \times p(f|e)}{p(f)} \tag{2.2}$$

Do $p(f)$ là độc lập với e, biểu thức có thể được viết như công thức 2.3:

$$arg\max_{e} p(e|f) = arg\max_{e} p(e) \times p(f|e) \tag{2.3}$$

2.2.2. Mô hình ngôn ngữ

Mô hình ngôn ngữ giúp hệ dịch xác định độ chính xác của trật tự từ, các hệ thống hiện nay thường tính toán sử dụng mô hình ngôn ngữ $n-gram$.

Mô hình ngôn ngữ $n-gram$ tính xác suất xuất hiện của một từ dựa trên $n-1$ từ đứng trước nó trong câu. Với câu s gồm chuỗi các từ $w_1, w_2, ...w_n$, xác suất trong mô hình ngôn ngữ được tính như sau:

Xác suất unigram:

$$p(w_1) = \frac{\Sigma w_1}{\sum_{i=1}^{n} w_i} \tag{2.4}$$

Xác suất bigram:

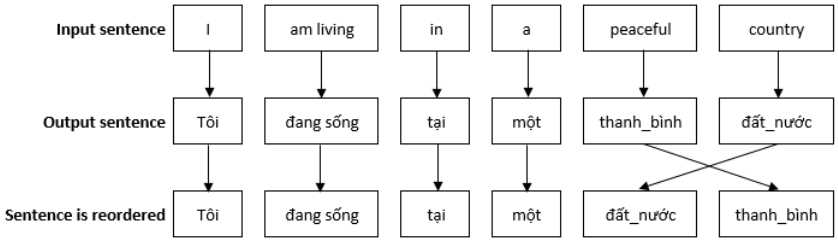
$$p(w_1|w_2) = \frac{\Sigma w_1 w_2}{\Sigma w_1} \tag{2.5}$$

Xác suất trigram:

$$p(w_3|w_1w_2) = \frac{\Sigma w_1w_2w_3}{\Sigma w_1w_2} \tag{2.6}$$

2.2.3. Dịch máy thống kê dựa vào cụm từ

Hình 2.3 mô tả quá trình dịch dựa vào cụm từ. Câu đầu vào được tách thành chuỗi các từ liên tiếp. Mỗi từ hoặc cụm từ trong ngôn ngữ nguồn được dịch tương ứng thành một từ hoặc cụm từ trong ngôn ngữ đích.



Hình 2.3: *Ví dụ minh họa quá trình dịch dựa trên cụm từ*

Mô hình dịch thống kê dựa vào cụm từ dựa trên mô hình kênh nhiễu, sử dụng quy tắc Bayes để xác định xác suất dịch để dịch một câu đầu vào f thành câu đầu ra e ở một ngôn ngữ khác. Bản dịch tốt nhất cho câu đầu vào f được mô tả theo công thức 2.7:

$$e = arg\max_e p(e)p(e|f) \tag{2.7}$$

Công thức trên bao gồm hai thành phần:

- Mô hình ngôn ngữ ấn định xác suất *p(e)*.
- Mô hình dịch *p(e|f)*.

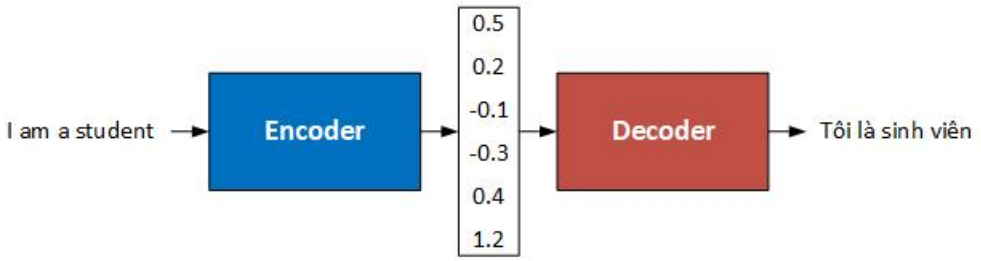
Mô hình ngôn ngữ được huấn luyện sử dụng dữ liệu đơn ngữ ở ngôn ngữ đích, mô hình dịch được huấn luyện sử dụng dữ liệu song ngữ.

2.3. Dịch máy mạng nơ-ron

2.3.1. Kiến trúc Encoder - Decoder

Đây là kiến trúc đầu tiên của hệ thống dịch máy mạng nơ-ron (NMT), đặt nền móng cho các hệ thống sau này. Kiến trúc này gồm hai thành phần là bộ mã hóa (encoder) và bộ giải mã (decoder), được mô tả như Hình 2.4.

Hệ dịch NMT sử dụng bộ mã hóa để đọc toàn bộ câu nguồn và mã hóa nó thành một vectơ biểu diễn ý nghĩa của câu. Sau đó, bộ giải mã sử dụng vectơ này để sinh câu dịch tương ứng trong ngôn ngữ đích.



Hình 2.4: *Kiến trúc mã hóa – giải mã (encoder – decoder). Bộ mã hóa chuyển một câu nguồn thành một vecto có nghĩa, sau đó bộ giải mã sẽ giả mã vecto này để tạo ra bản dịch.*

Bộ mã hóa và bộ giải mã đều được cấu tạo từ hai lớp RNN cùng chiều chồng lên nhau, ký hiệu $< s >$ và $< /s >$ sử dụng để báo hiệu bắt đầu và kết thúc quá trình giải mã.

Bộ mã hóa đọc câu nguồn là một dãy các vectơ $x = (x_1, ..., x_n)$ một vector c. Phương pháp phổ biến nhất là sử dụng một mạng nơ-ron hồi quy, sao cho:

$$h_t = f(x_t, h_{t-1}) \tag{2.8}$$

và

$$c = q(h_1, ..., h_T) \tag{2.9}$$

trong đó, h_t là trạng thái ẩn tại thời điểm t, và c là véc tơ ngữ cảnh được sinh ra từ dãy của các trạng thái ẩn. f và q là các hàm phi tuyến.

Bộ giải mã thường được huấn luyện để dự đoán từ tiếp theo y_T khi biết véc tơ ngữ cảnh c và tất cả các từ đã được sinh ra trước đó $y_1, ..., y_{T-1}$. Nói cách khác, bộ giải mã định nghĩa một xác suất cho câu dịch y bằng cách ước lượng hàm phân phối xác suất có điều kiện sau:

$$p(y) = \prod_{t=1}^{T} (p(y_t | y_1, ..., y_{t-1}, c) \tag{2.10}$$

trong đó, $y = (y_1, ..., y_T)$. Với một RNN, mỗi xác suất có điều kiện được mô hình hóa như sau:

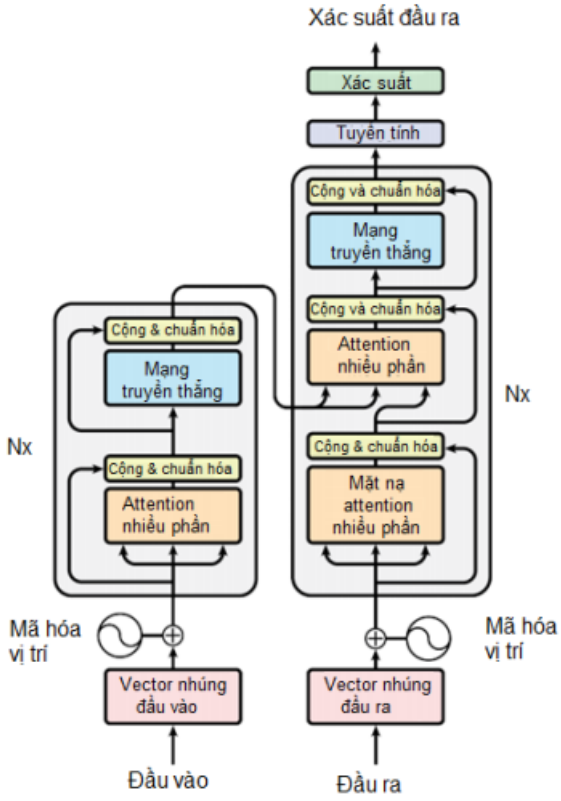
$$p(y_t | y_1, ..., y_{t-1}, c) = g(y_{t-1}, s_t, c) \tag{2.11}$$

trong đó g là hàm phi tuyến để sinh ra xác suất của y_t, và s_t là trạng thái ẩn của mạng nơ-ron hồi quy, sau này g thường được dùng bởi LSTM. Thông thường một RNN được sử dụng cho cả bộ mã hóa và giải mã.

2.3.2. Kiến trúc Transformer

Kiến trúc Transformer được đề xuất bởi nhóm nghiên cứu của Google AI (Vaswani và cộng sự, 2017), có thể coi như là một mô hình mở rộng của mô hình mã hóa - giải mã với attention.

Hai thành phần mã hóa và giải mã trong mô hình Transformer đều sử dụng self-attention nhiều tầng, mã hóa vị trí, các tầng kết nối với nhau toàn bộ (fully connected) như Hình 2.5.



Hình 2.5: *Kiến trúc Transformer*

Về cơ bản, bộ mã hóa gồm N tầng giống nhau xếp chồng lên nhau, mỗi tầng có 2 tầng con. Tầng con thứ nhất là cơ chế self-attention nhiều phần (multi-head), tầng con thứ 2 là mạng truyền thẳng đầy đủ (fully connected feed-forward). Ngoài ra, có thể thêm kĩ thuật kết nối dư (residual connection), theo sau bởi 1 tầng chuẩn hóa (normalization layer). Bộ giải mã cũng gồm N tầng giống nhau xếp chồng. Tại mỗi tầng, bên cạnh 2 tầng con giống như bộ mã hóa, bộ giải mã chèn thêm 1 tầng con ở giữa, cái thể hiện multi-head attention để có thể mô hình khóa được các thông tin cần thiết của câu nguồn tại mỗi thời điểm giải mã.

2.4. Đánh giá chất lượng dịch máy

2.4.1. Đánh giá dựa vào con người

Phương pháp dựa vào con người cho đánh giá tốt nhất đối với chất lượng của bản dịch, tuy nhiên cách đánh giá này mất nhiều thời gian và tốn kém.

2.4.2. Đánh giá tự động: BLEU

Độ đo được sử dụng phổ biến để đánh giá tự động chất lượng của dịch máy là BiLingual Evaluation Understudy Score, viết tắt là BLEU, do Papineni đề xuất năm 2002. Ý tưởng chính là so sánh bản dịch tự động với bản dịch chuẩn do người dịch, được xác định dựa trên số lượng

$n-gram$ giống nhau giữa bản dịch của câu nguồn với các câu tham chiếu tương ứng, có xét tới yếu tố độ dài của câu, được định nghĩa như công thức 2.12.

$$BLEUscore = BP.e^{(\sum_{i=1}^{n} w_i \log p_i)} \tag{2.12}$$

Trong đó:

p_i: Giá trị trung bình của độ chính xác n-gram được thay đổi.

w_i: Trọng số tích cực.

BP (Brevity Penalty): Phạt ngắn dùng để phạt các bản dịch quá vắn tắt. Phạt ngắn được tính toán trên toàn bộ kho ngữ liệu theo công thức 2.13

$$BP = \begin{cases} 1 & if \quad c > r \\ e^{1-\frac{r}{c}} & if \quad c < r \end{cases} \tag{2.13}$$

Trong đó:

c: độ dài của bản dịch.

r: độ dài của kho ngữ liệu tham chiếu.

Điểm BLEU nhận giá trị trong khoảng [0,1] để đo mức độ tương tự của văn bản được dịch bằng máy với tập dữ liệu chất lượng cao do chuyên gia dịch.

2.5. Thích ứng miền trong dịch máy thống kê

Phương pháp thích ứng miền có thể cải thiện chất lượng dịch máy trên một miền dữ liệu cụ thể nào đó, có thể mô tả như sau:

Gọi S là tập hợp các câu đầu vào và T là tập hợp các câu dịch tương ứng của S ở đầu ra của hệ thống dịch máy. Trong dịch máy thống kê, hệ thống thực hiện học hàm $f : S \longrightarrow T$. Để huấn luyện một hệ thống dịch máy thống kê cần có tập dữ liệu huấn luyện $D\{(s_n, t_n) \in S \times T\}$. Các mẫu huấn luyện (s_n, t_n) là độc lập và tuân theo phân phối p. Mô hình dịch máy thống kê được huấn luyện để xấp xỉ t_n với $f(s_n)$ đối với mọi $(s_n, t_n) \in D$. Thông thường mô hình kiểm thử trên tập D^A cũng tuân theo phân phối p.

Vấn đề ở đây là tệp kiểm thử D^A thường được lấy từ một miền khác, vì vậy mà có phân phối p^A khác với p. Trong khi một mô hình có thể xấp xỉ t_n rất tốt bởi $f(s_n)$ đối với mọi $(s_n, t_n) \in D$, nhưng thường là không tốt với các mẫu từ dữ liệu kiểm thử D^A lấy từ một miền khác. Nếu D và D^A là rất khác nhau, hàm xấp xỉ sẽ không cho kết quả như mong đợi.

2.6. Kết luận chương 2

Chương này trình bày tổng quan về dịch máy và các kiến thức liên quan.

Chương 3. PHƯƠNG PHÁP TINH CHỈNH BẢNG DỊCH CỤM TỪ

Chương này trình bày đề xuất *phương pháp tinh chỉnh bảng cụm từ (phrase-table) trong hệ dịch máy thống kê dựa trên cụm từ (PBSMT)* để cải tiến chất lượng hệ dịch.

3.1. Giới thiệu

Mô hình dịch là mô hình quan trọng nhất trong hệ thống PBSMT, ảnh hưởng và quyết định tới chất lượng của bản dịch. Chương này luận án trình bày đề xuất phương pháp thích ứng mô hình dịch bằng cách tinh chỉnh bảng dịch cụm từ theo hướng ưu tiên hơn trong miền đích. Các thực nghiệm được thực hiện trên miền chung và miền pháp luật của cặp ngôn ngữ Anh-Việt, theo chiều từ tiếng Anh sang tiếng Việt.

3.2. Các nghiên cứu liên quan

Có nhiều nghiên cứu về thích ứng miền đã được đề xuất, các nghiên cứu chủ yếu tiếp cận theo hai hướng chính là: (1) tăng cường, nâng cao chất lượng dữ liệu và (2) cải tiến mô hình.

Có nhiều nghiên cứu đề xuất cải tiến chất lượng dịch máy với các phương pháp, kĩ thuật nhằm cải tiến bảng cụm từ. Có thể kể tới một số nghiên cứu như: đề xuất của (Hua Wu và cộng sự, 2008) xây dựng từ điển miền và tích hợp trực tiếp vào bảng cụm từ; đề xuất (Passban và cộng sự, 2016) và (Nguyen Minh-Thuan, 2018) can thiệp trực tiếp vào bảng cụm từ để làm giàu thêm thông tin miền; đề xuất của (Pratyush và cộng sự, 2010) huấn luyện nhiều mô hình dịch máy riêng lẻ miền, thực hiện phân loại miền các câu cần dịch để lựa chọn mô hình phù hợp.

3.3. Phân loại văn bản

Phân loại văn bản là quá trình gán nhãn các văn bản ngôn ngữ tự nhiên vào một hoặc nhiều lớp từ tập các lớp hữu hạn cho trước.

3.3.1. Entropy cực đại cho phân loại văn bản

Phân loại entropy cực đại là phân loại xác suất thuộc loại mô hình hàm mũ, thường được sử dụng để phân loại văn bản, được mô tả theo công thức sau:

$$p(y|x) = \frac{\exp(\sum_k \lambda_k f_k(x,y))}{\sum \exp(\sum_k \lambda_k f_k(x,z))} \tag{3.1}$$

trong đó λ_k là các tham số mô hình và f_k là các đặc trưng của mô hình [0].

Chúng tôi đã huấn luyện mô hình phân loại xác suất với 2 lớp là pháp luật và Chung. Sau khi huấn luyện, mô hình phân loại được sử dụng để phân loại danh sách các cụm từ trong bảng cụm từ ở phía đích, chúng tôi coi những cụm từ này nằm trong miền chung ở phần đầu. Đầu

ra của nhiệm vụ phân loại là xác suất của cụm từ trong mỗi miền *(P(pháp luật) và P(chung))*, một số kết quả của nhiệm vụ phân loại như trong Hình **??**.

3.4. Phương pháp tinh chỉnh bảng dịch cụm từ

3.4.1. Bảng dịch cụm từ

Quá trình dịch máy theo đơn vị cụm từ được như mô tả như Hình 2.3, kiến trúc hệ dịch máy thống kê dựa vào cụm từ được mô tả như Hình 2.2. Chất lượng bản dịch phụ thuộc vào chất lượng mô hình dịch (bảng dịch cụm từ), bảng cụm từ là một tệp, trên mỗi dòng chứa các xác suất dịch của một cụm từ nguồn f thành một cụm từ đích e. Bảng cụm từ được sinh ra bắt đầu từ quan hệ gióng hàng từ *(word alignment)* giữa mỗi cặp câu trong ngữ liệu song ngữ, sau đó trích xuất các cặp cụm từ phù hợp, được mô tả như thuật toán trong Hình 3.1, với f là nguôn ngữ nguồn, e là ngôn ngữ đích.

```
Input: word alignment A for sentence pair (e,f)
Output: set of phrase pairs BP
 1: for e_start = 1 ... length(e) do
 2:   for e_end = e_start ... length(e) do
 3:     // find the minimally matching foreign phrase
 4:     (f_start,f_end) = ( length(f), 0 )
 5:     for all (e,f) ∈ A do
 6:       if e_start ≤ e ≤ e_end then
 7:         f_start = min( f, f_start )
 8:         f_end = max( f, f_end )
 9:       end if
10:     end for
11:     add extract(f_start,f_end,e_start,e_end) to set BP
12:   end for
13: end for
function extract(f_start,f_end,e_start,e_end)
 1: return {} if f_end == 0 // check if at least one alignment point
 2: // check if alignment points violate consistency
 3: for all (e,f) ∈ A do
 4:   return {} if e < e_start or e > e_end
 5: end for
 6: // add pharse pairs (incl. additional unaligned f)
 7: E = {}
 8: f_s = f_start
 9: repeat
10:   f_e = f_end
11:   repeat
12:     add phrase pair (e_start .. e_end, f_s .. f_e) to set E
13:     f_e++
14:   until f_e aligned
15:   f_s--
16: until f_s aligned
17: return E
```

Hình 3.1: *Thuật toán rút trích bảng cụm từ*

Sau đó, điểm cụm từ cho mỗi cặp cụm từ được xác định bằng cách ước tính xác suất căn cứ vào tần suất tương đối *(relative frequencies)* của chúng theo công thức 3.2.

$$\phi(\overline{f}|\overline{e}) = \frac{count(\overline{e}, \overline{f})}{\sum_{\overline{f}_i} count(\overline{e}, \overline{f}_i)} \tag{3.2}$$

Trong bảng cụm từ có bốn điểm cụm từ: (1) Xác suất dịch cụm từ theo chiều ngược $\phi(f|e)$;

(2) Trọng số từ vựng theo chiều ngược $lex(f|e)$; (3) Xác suất dịch cụm từ theo chiều xuôi $\phi(e|f)$; (4) Trọng số từ vựng theo chiều xuôi $lex(e|f)$. Bảng cụm từ như Hình 3.2.

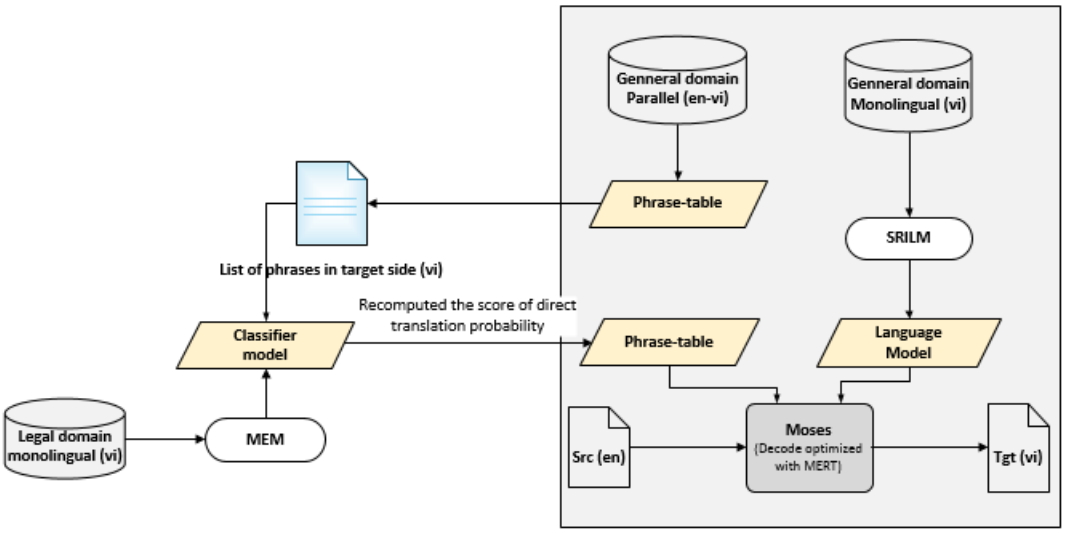
```
discrimination against ||| phân_biệt đối_xử với những người ||| 1 0.0283582 0.05 3.79301e-06 ||| 0-0 0-1 1-1 1-2 ||| 1
discrimination against ||| phân_biệt đối_xử với những ||| 1 0.0283582 0.05 0.000480547 ||| 0-0 0-1 1-1 1-2 ||| 1 20 1 |
discrimination against ||| phân_biệt đối_xử với ||| 0.470588 0.0283582 0.4 0.0197545 ||| 0-0 0-1 1-1 1-2 ||| 17 20 8 ||
discrimination against ||| phân_biệt đối_xử đối_với ||| 0.666667 0.0321897 0.2 0.0192058 ||| 0-0 0-1 1-1 1-2 ||| 6 20 4
discrimination and ||| phân_biệt đối_xử và ||| 0.142857 0.409019 0.142857 0.175113 ||| 0-0 0-1 1-2 ||| 7 7 1 ||| |||
discrimination and ||| sự phân_biệt đối_xử và ||| 1 0.273486 0.285714 0.00742903 ||| 0-0 0-1 0-2 1-3 ||| 2 7 2 ||| |||
discrimination between ||| phân_biệt đối_xử giữa các ||| 1 0.236418 0.333333 0.0219022 ||| 0-0 0-1 1-2 ||| 1 3 1 ||| ||
discrimination between ||| phân_biệt đối_xử giữa ||| 1 0.236418 0.666667 0.156906 ||| 0-0 0-1 1-2 ||| 2 3 2 ||| |||
discrimination in the ||| phân_biệt đối_xử trong ||| 0.4 0.0408001 1 0.0687546 ||| 0-0 0-1 1-2 ||| 5 2 2 ||| |||
discrimination in ||| phân_biệt đối_xử liên_quan đến hiv/aids ||| 0.2 0.0145789 0.166667 9.7585e-10 ||| 0-0 0-1 1-3 |||
discrimination in ||| phân_biệt đối_xử liên_quan đến ||| 0.2 0.0145789 0.166667 2.02922e-06 ||| 0-0 0-1 1-3 ||| 5 6 1 |
discrimination in ||| phân_biệt đối_xử trong ||| 0.6 0.211027 0.5 0.0687546 ||| 0-0 0-1 1-2 ||| 5 6 3 ||| |||
discrimination on the basis of sexual orientation ||| phân_biệt đối_xử ||| 0.0126582 5.41373e-17 1 0.213407 ||| 0-0 0-1
discrimination on the basis of sexual ||| phân_biệt đối_xử ||| 0.0126582 1.86039e-12 1 0.213407 ||| 0-0 0-1 ||| 79 1 1
discrimination on the basis of ||| phân_biệt đối_xử ||| 0.0126582 3.19106e-08 0.5 0.213407 ||| 0-0 0-1 ||| 79 2 1 ||| |
discrimination on the basis ||| phân_biệt đối_xử ||| 0.0126582 2.50499e-07 0.5 0.213407 ||| 0-0 0-1 ||| 79 2 1 ||| |||
discrimination on the ||| phân_biệt đối_xử ||| 0.0126582 0.000687619 0.5 0.213407 ||| 0-0 0-1 ||| 79 2 1 ||| |||
discrimination on ||| phân_biệt đối_xử trên ||| 1 0.0905388 0.5 0.0224057 ||| 0-0 0-1 1-2 ||| 2 4 2 ||| |||
discrimination on ||| phân_biệt đối_xử ||| 0.0126582 0.00355651 0.25 0.213407 ||| 0-0 0-1 ||| 79 4 1 ||| |||
discrimination reduction , ||| phân_biệt đối_xử , ||| 0.0714286 0.00408537 1 0.0885124 ||| 0-0 0-1 1-1 2-2 ||| 14 1 1 |
discrimination reduction ||| phân_biệt đối_xử ||| 0.0126582 0.00480614 1 0.108149 ||| 0-0 0-1 1-1 ||| 79 1 1 ||| |||
discrimination ||| và phân_biệt đối_xử đối_với ||| 0.166667 0.43976 0.0119048 3.07062e-06 ||| 0-1 0-2 ||| 6 84 1 ||| ||
discrimination ||| sự phân_biệt ||| 0.105263 0.249804 0.0238095 0.0213407 ||| 0-0 0-1 ||| 19 84 2 ||| |||
discrimination ||| sự phân_biệt đối_xử ||| 0.833333 0.29404 0.0595238 0.00905361 ||| 0-0 0-1 0-2 ||| 6 84 5 ||| |||
discrimination ||| kỳ_thị ||| 0.05 0.0483871 0.0238095 0.0181818 ||| 0-0 ||| 40 84 2 ||| |||
discrimination ||| phân_biệt ||| 0.243902 0.497006 0.119048 0.50303 ||| 0-0 ||| 41 84 10 ||| |||
discrimination ||| phân_biệt đối_xử ||| 0.620253 0.43976 0.583333 0.213407 ||| 0-0 0-1 ||| 79 84 49 ||| |||
discrimination ||| tình_trạng phân_biệt đối_xử ||| 0.333333 0.43976 0.0119048 0.000130925 ||| 0-1 0-2 ||| 3 84 1 ||| ||
```

Hình 3.2: *Bảng cụm từ trong hệ dịch máy thống kê dựa vào cụm*

3.4.2. Phương pháp tinh chỉnh bảng dịch cụm từ

Điểm cụm từ là phần quan trọng nhất trong bảng cụm từ, nó ước tính trọng số cho các cặp cụm từ dựa trên một kho ngữ liệu song ngữ lớn. Do đó, trong các ngôn ngữ ít phổ biến và ít tài nguyên, ước tính thường không chính xác. Để giải quyết vấn đề này, chúng tôi đề xuất giải pháp tinh chỉnh bảng cụm từ theo hướng ưu tiên miền, chúng tôi chỉ sử dụng xác suất dịch cụm từ trực tiếp ϕ(e|f) của bảng cụm từ, giả thuyết dịch có xác suất cao hơn giá trị ϕ(e|f) thì giả thuyết dịch đó được ưu tiên dịch hơn giả thuyết khác. Chúng tôi sử dụng mô hình phân loại xác suất miền của cụm từ trong bảng cụm từ, sau đó chúng tôi tính lại xác suất dịch của cụm từ ϕ(e|f). Đề xuất được minh họa như Hình 3.3, quá trình gồm năm bước như sau:

- **Bước 1.** Huấn luyện mô hình phân loại miền cho văn bản, mục tiêu để xác định một cụm từ trong bảng cụm từ thuộc lớp pháp luật hay lớp chung.
- **Bước 2.** Huấn luyện một hệ thống PBSMT ban đầu sử dụng dữ liệu song ngữ thuộc miền chung, chiều dịch từ tiếng Anh sang tiếng Việt.
- **Bước 3.** Rút trích cụm từ ở phía đích trong bảng cụm từ của hệ thống PBSMT được huấn luyện ở Bước 2, tiến hành phân loại miền đối với các cụm từ này sử dụng mô hình phân loại được huấn luyện ở Bước 1.
- **Bước 4.** Tinh chỉnh bảng dịch cụm từ, cập nhật lại xác suất dịch ϕ (e|f) theo hướng ưu tiên miền.
- **Bước 5.** Sử dụng bảng dịch cụm từ đã được tinh chỉnh để dịch văn bản thuộc miền luật.



Hình 3.3: *Phương pháp tinh chỉnh bảng dịch cụm từ.*

3.5. Thực nghiệm

3.5.1. Dữ liệu

Thực nghiệm sử dụng ngữ liệu song ngữ Anh-Việt từ *hội nghị IWSLT năm 2015*[1] cho đánh giá hệ thống dịch máy, Thống kê chi tiết cho các tập dữ liệu được đưa ra trong Bảng 5.2.

Các tập dữ liệu		**Ngôn ngữ**	
		Tiếng Anh	**Tiếng Việt**
Training	Sentences	131019	
	Average Length	15.93	15.58
	Words	1946397	1903504
	Vocabulary	40568	28414
Dev	Sentences	745	
	Average Length	16.61	15.97
	Words	12397	11921
	Vocabulary	2230	1986
General_test	Sentences	1080	
	Average Length	16.25	15.97
	Words	17023	16889
	Vocabulary	2701	2759
Legal_test	Sentences	500	
	Average Length	15.21	15.48
	Words	7605	7740
	Vocabulary	1530	1429

Bảng 3.1: *Thống kê ngữ liệu song ngữ Anh-Việt*

Dữ liệu ngoài miền: dữ liệu đơn ngữ miền pháp luật trong tiếng Việt, được thu thập từ tài liệu, từ điển chuyên ngành, được gắn nhãn thủ công gồm hai lớp lớp pháp luật và lớp chung. Ngoài ra, chúng tôi sử dụng thêm 500 câu song ngữ miền pháp luật.

[1]https://wit3.fbk.eu/2015-01

3.5.2. Tiền xử lý

Chúng tôi đã tiến hành tiền xử lý theo hai bước: (1) Làm sạch dữ liệu, giữ lại các câu có độ dài trong khoảng [1-80] và (2) Tách từ cho câu.

3.5.3. Các thực nghiệm

Các hệ thống thử nghiệm gồm:

- **Baseline_SMT:** Hệ thống dịch máy SMT cơ sở dựa trên cụm, được huấn luyện với dữ liệu song ngữ miền chung.
- **Adaptation_SMT:** Là hệ thống Baseline_SMT sau khi bảng dịch cụm từ được tinh chỉnh hướng miền.
- **Baseline_NMT:** Hệ dịch NMT cơ sở để so sánh bổ sung với Baseline_SMT.

3.5.4. Kết quả thực nghiệm

Kết quả thực nghiệm thể hiện trong Bảng 3.2, cho thấy hệ thống SMT được huấn luyện trên miền chung nếu miền kiểm tra khác miền huấn luyện thì chất lượng bản dịch sẽ giảm xuống. Trong các thực nghiệm này, điểm BLEU đã giảm 2,5 điểm từ 31,3 xuống 28,8. Hệ thống Adaptation_SMT được thích ứng theo đề xuất đã cải thiện được chất lượng của hệ thống dịch. Trong các thử nghiệm này, điểm BLEU được cải thiện từ 28,8 lên 29,7 từ 0,9 điểm.

Hệ thống	**BLEU(%)**	**Mô tả**
Baseline_SMT	31.3	Áp dụng trên tập General_test
Baseline_SMT	28.8	Áp dụng trên tập Legal_test
Adaptation_SMT	29.7	Áp dụng trên tập Legal_test
Baseline_NMT	30.1	Áp dụng trên tập General_test
Baseline_NMT	20.9	Áp dụng trên tập Legal_test

Bảng 3.2: *Thực nghiệm tinh chỉnh bảng dịch cụm từ*

3.6. Kết luận chương 4

Mục này tổng kết các kết quả nghiên cứu ở Chương 3.

Chương 4. PHƯƠNG PHÁP SINH TỰ ĐỘNG DỮ LIỆU SONG NGỮ CHO DỊCH MÁY

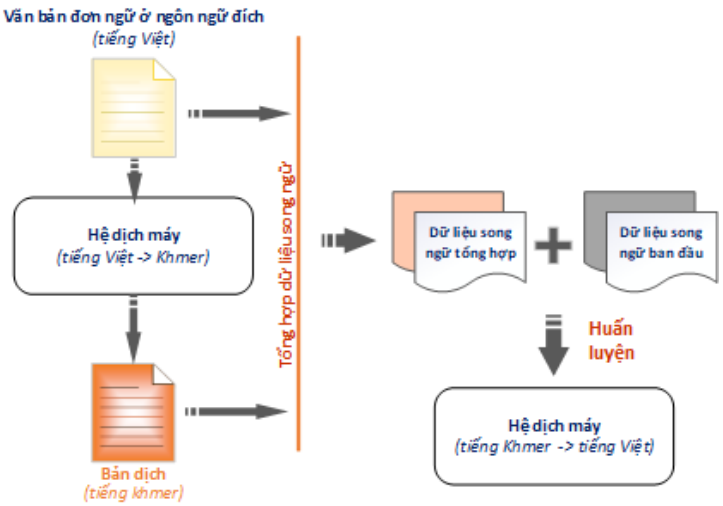
4.1. Giới thiệu

Dữ liệu song ngữ đóng vai trò rất quan trọng trong dịch máy. Tuy nhiên, dữ liệu song ngữ thường hiếm, chi phí xây dựng lớn. Trong khi đó dữ liệu đơn ngữ có sẵn nên đã có nhiều nghiên cứu sử dụng dữ liệu đơn ngữ để cải thiện chất lượng dịch.

4.2. Phương pháp dịch ngược

Dịch ngược là phương pháp chỉ sử dụng dữ liệu đơn ngữ để tổng hợp, sinh ra dữ liệu song ngữ, có thể phát biểu như sau:

Cho một tập dữ liệu song ngữ đã được gióng hàng câu $D = (X_n, Y_n)N$ và một tập dữ liệu đơn ngữ trong ngôn ngữ đích $T = (Y_m)M$, quá trình dịch ngược lần lượt được thực hiện sau và được mô tả như Hình 4.1:



Hình 4.1: *Mô tả phương pháp dịch ngược*

1. Đầu tiên, một hệ thống dịch ngược $NMT_{Y->X}$ được huấn luyện với tập dữ liệu song ngữ D.

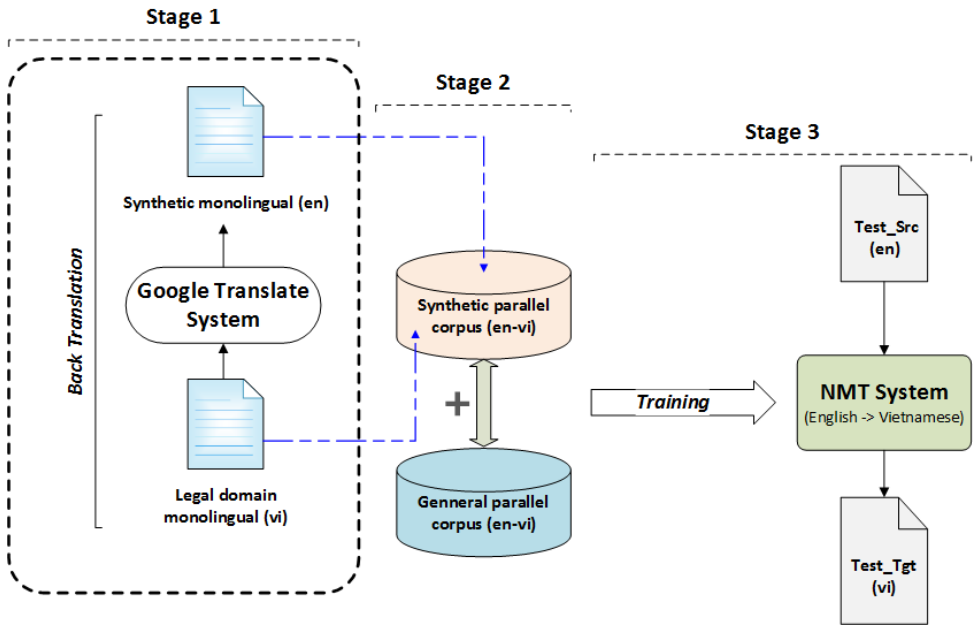
2. Sau đó, với hệ thống dịch $NMT_{Y->X}$, tập dữ liệu đơn ngữ trong ngôn ngữ đích T được dịch ngược lại thành các bản dịch trong ngôn ngữ nguồn $S = (X_m)M_{m=1}$, sau đó tập dữ liệu S được ghép nối với T, tạo thành một tập dữ liệu giả song ngữ $D_{syn} = (X_m, Y_m)M_{m=1}$.

3. Thứ ba, tập dữ liệu giả song D_{syn} và tập dữ liệu song ngữ ban đầu D được kết hợp để huấn luyện một hệ thống dịch máy $NMT_{Y->X}$.

4.3. Phương pháp đề xuất

Hiện nay, dịch máy đã phát triển về mức độ tinh vi cũng như khả năng tiếp cận, có một số dịch vụ dịch trực tuyến khác nhau như Google Translate[1], Bing Microsoft Translator [2], Babylon Translator [3], Facebook Machine Translation, v.v. Google Translate là một trong những ứng dụng được sử dụng nhiều nhất vì tính tiện lợi của nó.

Để tận dụng lợi thế của Google Translate về mặt dữ liệu và sự có sẵn của dữ liệu đơn ngữ, chúng tôi đề xuất sử dụng phương pháp sinh tự động dữ liệu song ngữ cho dịch máy sử dụng Google Translate.Đề xuất gồm ba giai đoạn sau và được mô tả như Hình **??**.



Hình 4.2: *Mô tả phương pháp đề xuất*

- Giai đoạn 1: sử dụng Google Translate để dịch dữ liệu đơn ngữ của miền sang tiếng Việt.
- Giai đoạn 2: tổng hợp ngữ liệu song song bằng cách kết hợp dữ liệu đơn ngữ miền đầu vào với bản dịch đầu ra ở giai đoạn 1. Tiếp theo, chúng tôi kết hợp ngữ liệu song song tổng hợp với ngữ liệu song song ban đầu được cung cấp bởi hội nghị IWSLT2015.
- Giai đoạn 3: với kho ngữ liệu song song hỗn hợp ở giai đoạn 2, chúng tôi tiến hành đào tạo hệ thống NMT và đánh giá chất lượng bản dịch trong miền pháp lý và miền tổng quan.

4.4. Thực nghiệm

4.4.1. Dữ liệu

- Sử dụng dữ liệu song ngữ Anh-Việt được cung cấp bởi hội nghị IWSLT2015 để huấn luyện hệ dịch cơ sở, thống kê chi tiết dữ liệu trong Bảng 5.2.

[1]https://translate.google.com
[2]https://www.bing.com/translator
[3]https://translation.babylon-software.com/

- Để sinh dữ liệu song ngữ, sử dụng 100k câu đơn ngữ miền luật tiếng Việt.
- Để đánh giá chất lượng, sử dụng 500 cặp câu trong miền luật và miền chung.

Data Sets		**Language**	
		English	**Vietnamese**
Training	Sentences	**133316**	
	Average Length	16.62	16.68
	Words	1952307	1918524
	Vocabulary	40568	28414
Val	Sentences	**1553**	
	Average Length	16.21	16.97
	Words	13263	12963
	Vocabulary	2230	1986
General_test	Sentences	**1246**	
	Average Length	16.15	15.96
	Words	18013	16989
	Vocabulary	2708	2769
Legal_test	Sentences	**500**	
	Average Length	15.21	15.48
	Words	7605	7740
	Vocabulary	1530	1429

Bảng 4.1: *Thống kê tổng hợp các tập dữ liệu: Anh-Việt*

4.4.2. Tiền xử lý

Sử dụng các tập lệnh trong Moses cho tiếng Anh và công cụ vnTokenizer để phân đoạn từ cho tiếng Việt để tách từ, sử dụng các scripts trong Moses để làm sạch dữ liệu, giữ lại các câu có độ dài trong khoảng [1-80].

4.4.3. Kết quả thực nghiệm

Thử nghiệm với các kịch bản:

- **Baseline:** Hệ thống được huấn luyện chỉ sử dụng dữ liệu IWSLT2015.
- **Synthetic:** Hệ thống được huấn luyện chỉ sử dụng dữ liệu tổng hợp, gồm 100k cặp câu.
- **Baseline_Syn50:** Hệ thống sử dụng dữ liệu IWSLT2015 kết hợp 50k cặp câu song ngữ tổng hợp.
- **Baseline_Syn100:** Hệ thống sử dụng dữ liệu IWSLT2015 kết hợp 100k cặp câu song ngữ tổng hợp.

Các hệ thống NMT được đánh giá trong miền chung và miền pháp luật. Kết quả thử nghiệm trong Bảng 4.2 và Bảng 4.3.

SYSTEM	BLEU SCORE
Baseline	**25.43**
Baseline_Syn50	**27.74**
Baseline_Syn100	**27.68**
Synthetic	**21.42**
Google Translate	**46.47**

Bảng 4.2: *Kết quả thử nghiệm của các hệ thống trong miền tổng quan.*

SYSTEM	BLEU SCORE
Baseline	**19.23**
Baseline_Syn50	**30.61**
Baseline_Syn100	**32.88**
Synthetic	**31.98**
Google Translate	**32.05**

Bảng 4.3: *Kết quả thử nghiệm của các hệ thống trong miền pháp lý.*

Theo Bảng 4.2 và Bảng 4.3, Baseline NMT đạt 25,43 điểm BLEU trong miền chung nhưng giảm xuống còn 19,23 trong miền pháp luật. Sau khi áp dụng dịch ngược, kết quả được cải thiện so với hệ thống cơ sở, phương pháp đề xuất cải thiện chất lượng bản dịch trong miền pháp luật lên 13,65 điểm BLEU và 2,25 điểm BLEU trong miền chung.

Đồ thị **??** so sánh chất lượng bản dịch khi dịch trong miền pháp luật và miền chung. Ở miền chung, điểm BLEU của Google Translate là 46,47 điểm, hệ thống cơ sở là 25,43 điểm và điểm BLEU của hệ thống của chúng tôi cao hơn hệ thống cơ sở, lần lượt là 27,68 và 27,74 điểm. Trong lĩnh vực pháp luật, điểm BLEU của Google Translate là 32,05 điểm, hệ thống cơ sở là 19,23 điểm và điểm BLEU của hệ thống của chúng tôi cao hơn hệ thống cơ sở, lần lượt đạt 31,98, 32,61 và 32,88 điểm. Như vậy, việc sử dụng Google Translate cho cặp ngôn ngữ Anh-Việt để sinh tự động dữ liệu song ngữ trong miền pháp luật có thể nâng cao chất lượng dịch của hệ thống dịch máy.

4.5. Kết luận chương 4

Mục này tổng kết các kết quả nghiên cứu ở Chương 4.

Chương 5. CẢI TIẾN CHẤT LƯỢNG CỦA PHƯƠNG PHÁP SINH TỰ ĐỘNG DỮ LIỆU SONG NGỮ

5.1. Giới thiệu

Gần đây, dịch ngược là một cách tiếp cận nổi bật để tăng cường, bổ sung thêm dữ liệu song ngữ. Để thực hiện cần xây dựng một hệ thống dịch máy ngược, việc này khó đối với các cặp ngôn ngữ có nguồn tài nguyên ít như cặp Anh-Việt do không đủ dữ liệu song ngữ để huấn luyện một hệ thống dịch máy ngược đủ tốt.

Google Translate là một hệ thống dịch máy nổi tiếng cho nhiều cặp ngôn ngữ, độ chính xác phụ thuộc vào từng cặp ngôn ngữ và từng miền dịch. chúng tôi muốn tận dụng những lợi thế của ứng dụng Google Translate để tăng cường dữ liệu song ngữ thay vì phải huấn luyện một hệ thống dịch máy ngược.

Trên thực tế, phương pháp sinh tự động dữ liệu song ngữ dùng kĩ thuật dịch ngược cho kết quả không tốt nếu văn bản đầu vào chứa các lỗi về chính tả, lỗi câu. Để giải quyết vấn đề này, chúng tôi đề xuất giải pháp cải tiến chất lượng của phương pháp trên.

5.2. Phương pháp đề xuất

Chúng tôi đề xuất sử dụng mô hình sửa lỗi ngữ pháp cho tiếng Việt để chuẩn hóa, sửa một số lỗi phổ biến của dữ liệu đầu vào cho Google Translate. Chúng tôi đã kết hợp mô hình sửa lỗi ngữ pháp với phương pháp dịch ngược (sử dụng Google Translate) để tạo bổ sung thêm nguồn dữ liệu song song chất lượng.

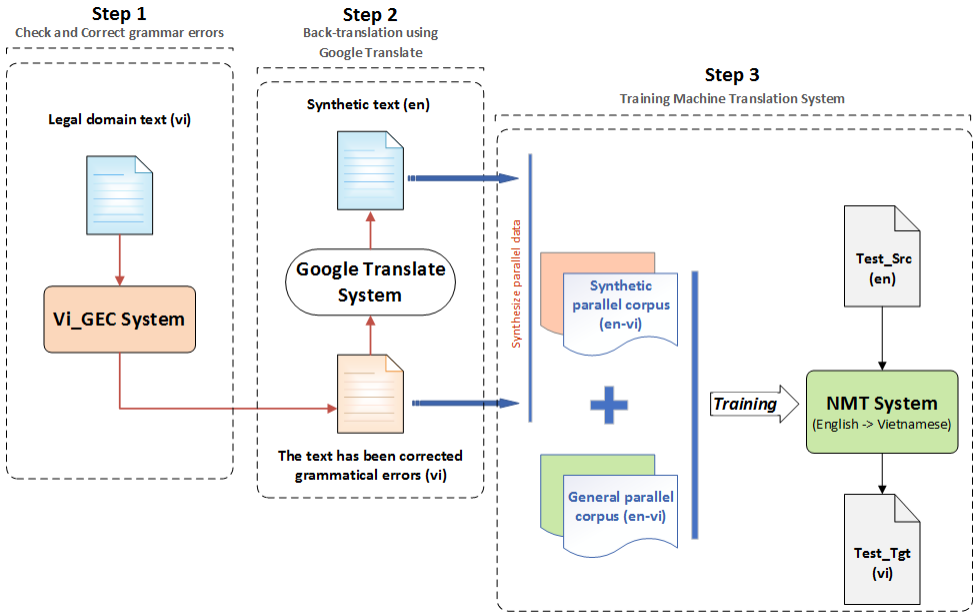
Phương pháp đề xuất gồm ba bước như sau và được mô tả trong Hình 5.1.

- **Bước 1.** Kiểm tra và sửa lỗi ngữ pháp.
- **Bước 2.** Dịch ngược sử dụng Google Translate.
- **Bước 3.** Huấn luyện hệ thống dịch máy nơ-ron Anh-Việt.

5.3. Thực nghiệm

Thử nghiệm được tiến hành với dữ liệu thuộc miền chung và miền pháp luật. Chúng tôi đã huấn luyện mô hình kiểm tra, sửa lỗi ngữ pháp cho tiếng Việt và một mô hình dịch NMT sử dụng kiến trúc Transformer với ba kịch bản:

1. Chỉ từ ngữ liệu song song IWSLT2015.
2. Chỉ từ dữ liệu tổng hợp.
3. Sử dụng hỗn hợp dữ liệu song ngữ ban đầu và dữ liệu tổng hợp.



Hình 5.1: *Mô tả đề xuất cải tiến chất lượng của sinh tự động dữ liệu song ngữ*

5.3.1. Dữ liệu

Các thử nghiệm thực hiện cho cặp ngôn ngữ Anh-Việt, trong miền pháp luật và miền chung. Thống kê chi tiết các bộ dữ liệu được trình bày trong Bảng 5.1 và Bảng 5.2.

Dữ liệu		**Tiếng Việt**	
		Sai ngữ pháp	**Đúng ngữ pháp**
Training	Sentences	**271822**	
	Average Length	21.1	20.8
	Words	5735444	5653897
Validation	Sentences	**29895**	
	Average Length	21.9	21.8
	Words	654700	651711
Test	Sentences	**15879**	
	Average Length	21.8	21.6
	Words	346162	342986

Bảng 5.1: *Thống kê dữ liệu huấn luyện hệ thống sửa lỗi ngữ pháp.*

Dữ liệu để cắt tỉa bảng cụm từ: Để xây dựng bảng cụm từ, chúng tôi sử dụng khoảng 10000 câu song ngữ trong miền pháp lý, độ dài của các câu này từ 40 đến 256. Mô tả chi tiết tập dữ liệu này được trình bày trong Bảng 5.3.

5.3.2. Tiền xử lý

Sử dụng tập lệnh trong bộ công cụ Moses cho tiếng Anh và VnTokenizer cho tách từ tiếng Việt, làm sạch dữ liệu, giữ lại các cặp câu có độ dài lớn hơn 40 token.

Dữ liệu		Ngôn ngữ	
		tiếng Anh	**tiếng Việt**
Training	Sentences	**133316**	
	Average Length	16.62	16.68
	Words	1952307	1918524
	Vocabulary	40568	28414
Val	Sentences	**1553**	
	Average Length	16.21	16.97
	Words	13263	12963
	Vocabulary	2230	1986
General_test	Sentences	**1246**	
	Average Length	16.15	15.96
	Words	18013	16989
	Vocabulary	2708	2769
Legal_test	Sentences	**500**	
	Average Length	15.21	15.48
	Words	7605	7740
	Vocabulary	1530	1429

Bảng 5.2: *Thống kê tổng hợp các bộ dữ liệu: Anh-Việt*

Dữ liệu		Ngôn ngữ	
		tiếng Anh	**tiếng Việt**
Training	Sentences	**10047**	
	Average Length	51.96	52.89
	Words	522072	531424

Bảng 5.3: *Thống kê chi tiết các tập dữ liệu để lược bỏ, cắt tỉa bảng cụm từ*

5.3.3. Kết quả thực nghiệm

Tiến hành thực hiện các kịch bản thử nghiệm như được mô tả ở trên. Hệ thống Vi_GEC và hệ thống Spell+Vi_GEC được đánh giá sử dụng độ đo BLEU. Kết quả được hiển thị trong Bảng 5.4. Điểm BLEU của hệ thống Spell+Vi_GEC là 92,18 và của hệ thống Vi_GEC là 89,70. Do đó, hệ thống Spell+Vi_GEC được chọn cho tất cả các thử nghiệm.

Hệ thống	**BLEU**
Vi_GEC	89.70
Spell+Vi_GEC	**92.18**

Bảng 5.4: *Điểm BLEU: hệ thống Vi_ GEC vs hệ thống Spell+Vi_ GEC*

Bảng 5.5 thể hiện điểm BLEU của các hệ thống NMT không sử dụng Spell+Vi_GEC cho Google Translate và Bảng 5.6 thể hiện kết quả của các hệ thống có sử dụng.

Bảng 5.5 cho thấy kết quả khi Spell+Vi_GEC không được sử dụng để sửa lỗi ngữ pháp, lỗi chính tả cho văn bản trước khi áp dụng phương pháp dịch ngược. Khi dịch trong miền chung, hệ thống Baseline đạt 28,3 điểm BLEU. Do hệ thống Baseline được huấn luyện với dữ liệu song

Hệ hống	BLEU	
	General	**Legal**
Baseline	28.3	**19.83**
Baseline_Syn50	29.80	32.91
Baseline_Syn100	29.60	**33.76**
Synthetic	22.63	32.80
Google Translate	46.47	32.05

Bảng 5.5: *Kết quả thử nghiệm khi áp dụng mô hình Spell+Vi_GEC*

ngữ ở miền chung nên khi dịch trong miền pháp luật, điểm BLEU giảm xuống còn 19,83. Sử dụng phương pháp dịch ngược cải thiện kết quả dịch khi so sánh với hệ thống Baseline.

Đặc biệt, hệ thống Baseline_Syn100 đã cải thiện chất lượng dịch trong lĩnh vực pháp luật tới 13,93 điểm BLEU so với hệ thống Baseline. Đây là hệ thống đạt được điểm BLEU cao nhất so với các hệ thống khác *(điểm BLEU là 33,91 và 32,80 tương ứng với hệ thống Baseline_Syn50 và hệ thống tổng hợp).*

Hệ thống	BLEU	
	General	**Legal**
Baseline_Syn50$_{\textbf{Spell+Vi_GEC}}$	29.04	33.86
Baseline_Syn100$_{\textbf{Spell+Vi_GEC}}$	30.70	**36.20**
Synthetic$_{\textbf{Spell+Vi_GEC}}$	23.00	34.06
Baseline_Syn50_Pru_**Spell+Vi_GEC**	x	33.6
Baseline_Syn100_Pru$_{\textbf{Spell+Vi_GEC}}$	x	**36.70**

Bảng 5.6: *Kết quả thực nghiệm của các hệ hống khi áp dụng Spell+Vi_GEC*

Khi áp dụng mô hình Spell+Vi_GEC, chất lượng bản dịch trên *(Baseline_Syn50, Baseline_Syn100 và Synthetic)* được cải thiện. Đặc biệt, Baseline_Syn100$_{Spell+Vi_GEC}$ đạt BLEU cao nhất là 36,20, cải thiện 2,44 điểm BLEU so với chính nó khi không sử dụng Spell+Vi_GEC. Hơn nữa, khi chúng tôi thêm phần tăng dữ liệu bằng cách lược bớt, cắt tỉa bảng cụm từ, điểm BLEU đã tăng thêm 0,5 cho hệ thống Baseline_Syn100$_{Spell+Vi_GEC}$.

Bảng 5.6 so sánh chất lượng của hệ thống dịch máy NMT khi áp dụng mô hình Spell+Vi_GEC.

Kết quả thử nghiệm cho thấy phương pháp này tuy đơn giản, hiệu quả, điểm BLEU tăng **2,94** so với hệ thống không áp dụng mô hình Spell+Vi_GEC *(BLEU tăng từ 33,76 lên 36,70)* và cải thiện **16,87** điểm BLEU so với hệ thống Baseline *(điểm BLEU từ 19,83 lên đến 36,70).*

5.4. Kết luận chương 5

Mục này tổng kết các kết quả nghiên cứu ở Chương 5.

Chương 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1. Các đóng góp của luận án

Luận án có ba đóng góp chính:

- Đề xuất phương pháp phân lớp miền các cụm từ trong ngôn ngữ đích.
- Đề xuất phương pháp xây dựng dữ liệu song ngữ tự động cho dịch máy:
  - Sử dụng phương pháp dịch ngược, khai thác các ưu điểm về mặt ngữ liệu và công nghệ của hệ dịch Google Translate.
  - Thực hiện phân tích ảnh hưởng của các lỗi văn bản trong quá trình sinh dữ liệu giả song ngữ, từ đó đề xuất phương pháp cải tiến, chuẩn hóa văn bản đầu vào. .
- Đề xuất phương pháp cải tiến chất lượng của phương pháp tự động sinh dữ liệu song ngữ.

6.2. Hướng phát triển

Để hoàn thiện các giải pháp thích ứng miền trong dịch máy và giúp hệ thống đạt chất lượng tốt hơn, trong thời gian tới chúng tôi sẽ tiếp tục tập trung nghiên cứu một số nội dung chính sau đây:

- Tiếp tục nghiên cứu, cải tiến phương pháp thích ứng miền nhằm đạt hiệu quả cao hơn.
- Mở rộng xây dựng kho ngữ liệu lớn ở các lĩnh vực khác nhau, có độ phủ tốt và triển khai đánh giá, phân tích và so sánh.
- Nghiên cứu các mô hình dịch máy đa miền, thích ứng miền đối với mô hình dịch máy đa ngữ.

DANH MỤC CÔNG TRÌNH KHOA HỌC

1. Nguyễn Quang Huy, Nguyễn Văn Vinh, **Phạm Nghĩa Luân**, Nguyễn Quỳnh Anh (2014). "Nghiên cứu phương pháp dóng hàng câu cho cặp ngôn ngữ Anh – Việt". *Hội thảo quốc gia lần thứ XVII: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, trang 188-195.

2. **Phạm Nghĩa Luân**, Nguyễn Văn Vinh, Nguyễn Quang Huy (2015). "Một phương pháp thích ứng miền cho dịch máy thống kê". *Hội thảo quốc gia lần thứ XVIII: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, trang 174-180.

3. Viet Tran Hong, Huyen Vu Thuong, Trung Le Tien, **Luan Nghia Pham** and Vinh Nguyen Van (2015). "The English - Vietnamese Machine Translation System for IWSLT 2015". *In Proceedings of the 12th International Workshop on Spoken Language Translation*, pp. 80-83. (SCOPUS)

4. **Nghia Luan Pham** and Van Vinh Nguyen (2019). "Adaptation in Statistical Machine Translation for low-resource domains in English-Vietnamese language". *In VNU Journal of Science: Computer Science and Communication Engineering*, [S.l.], v.36, n.1. ISSN 2588-1086. DOI:https://doi.org/10.25073/2588-1086/vnucsce.231.

5. **Nghia Luan Pham** and Van Vinh Nguyen (2019). "Adapting Neural Machine Translation for English-Vietnamese using Google Translate system for Back-translation". *In the 33rd Pacific Asia Conference on Language, Information and Computation*, pp. 567-575. (SCOPUS)

6. **Phạm Nghĩa Luân**, Nguyễn Văn Vinh (2019). "Thích ứng miền trong dịch máy nơ ron cho cặp ngôn ngữ Anh - Việt". *Hội nghị khoa học quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR 2019)*, pp.436-442.

7. **Nghia Luan Pham**, Tien Ha Nguyen and Van Vinh Nguyen (2019). "Grammatical error correction for Vietnamese using Machine Translation". *In 16th International Conference of the Pacific Association for Computational Linguistics*, pp.505-512. ISBN 978-981-15-6167-2. DOI: https://doi.org/10.1007/978-981-15-6168-9_41. (SCOPUS)

8. **Nghia Luan Pham** and Van Vinh Nguyen and Thang Viet Pham (2022). "Back-translation for data augmentation in Neural Machine Translation" - Submited.