ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

LÊ QUANG HÙNG

KHAI PHÁ TRI THỨC SONG NGỮ VÀ ỨNG DỤNG TRONG DỊCH MÁY ANH – VIỆT

Chuyên ngành: Khoa học máy tính
Mã số: 62 48 01 01

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Hà Nội - 2016

Công trình được hoàn thành tại: Trường Đại học Công nghệ , Đại học Quốc gia Hà Nội.

Người hướng dẫn khoa học:

1. PGS.TS. Lê Anh Cường
2. PGS.TS. Huỳnh Văn Nam

Phản biện 1: PGS.TS. Nguyễn Kim Anh

Phản biện 2: TS. Nguyễn Đức Dũng

Phản biện 3: TS. Lê Hồng Phương

Luận án đã được bảo vệ trước Hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ họp tại Trường Đại học Công nghệ , Đại học Quốc gia Hà Nội vào hồi 9 giờ ngày 12 tháng 01 năm 2016.

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

Mở đầu

1. Tính cấp thiết của luận án

Ý tưởng về dịch máy (machine translation - MT) ra đời từ năm 1949. Từ đó đến nay, sau hơn 60 năm nghiên cứu và phát triển, các dịch vụ dịch máy bây giờ đã trở nên phổ biến rộng rãi. Hiện nay, dịch máy dựa trên cách tiếp cận thống kê đang là một hướng phát triển đầy tiềm năng bởi những ưu điểm vượt trội so với các cách tiếp cận khác. Đối với một hệ thống dịch máy thống kê (statistical machine translation - SMT), chất lượng dịch tỷ lệ thuận với số lượng và chất lượng của ngữ liệu song ngữ được sử dụng để xây dựng hệ thống dịch. Tuy nhiên, ngữ liệu song ngữ hiện vẫn còn hạn chế cả về kích thước lẫn chất lượng, ngay cả đối với các ngôn ngữ chính. Ngoài ra, đối với các cặp ngôn ngữ có nhiều khác biệt về cấu trúc ngữ pháp (ví dụ, Anh - Việt), vấn đề về chất lượng dịch đang là thách thức đối với các nhà nghiên cứu về dịch máy trong nhiều năm qua. Vì vậy, các nghiên cứu nhằm khai thác thêm ngữ liệu song ngữ và phát triển các phương pháp hiệu quả hơn dựa trên ngữ liệu hiện có để tăng chất lượng dịch cho SMT là những vấn đề cấp thiết và mang tính thời sự trong lĩnh vực xử lý ngôn ngữ tự nhiên hiện nay. Điều này là động lực để chúng tôi lựa chọn nghiên cứu về đề tài "Khai phá tri thức song ngữ và ứng dụng trong dịch máy Anh - Việt".

2. Mục tiêu của luận án

Trong luận án này, chúng tôi đặt ra hai mục tiêu chính:

- Thứ nhất, nghiên cứu đề xuất một số phương pháp để khai thác tri thức song ngữ nhằm bổ sung nguồn ngữ liệu cho SMT.
- Thứ hai, nghiên cứu đề xuất một số phương pháp để làm tăng chất lượng dịch cho SMT dựa trên ngữ liệu hiện có.

3. Đóng góp của luận án

- Đề xuất một số phương pháp để xây dựng ngữ liệu song ngữ cho dịch máy thống kê từ Web và sách điện tử song ngữ. Đối với nguồn từ Web, chúng tôi đề xuất hai phương pháp thiết kế các đặc trưng dựa trên nội dung: sử dụng *cognate* và sử dụng các phân đoạn dịch. Đối với nguồn từ sách điện tử,

chúng tôi đề xuất phương pháp dựa trên nội dung, sử dụng một số mẫu liên kết giữa các khối văn bản trong hai ngôn ngữ để rút trích các câu song ngữ.

- Đề xuất một số cải tiến đối với mô hình gióng hàng IBM theo cách tiếp cận dựa trên ràng buộc, bao gồm: ràng buộc neo, ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ. Những cải tiến này đã giúp nâng cao chất lượng dịch cho hệ thống dịch máy thống kê Anh - Việt.
- Đề xuất phương pháp xác định cụm từ song ngữ cho dịch máy thống kê. Trước hết, chúng tôi sử dụng tập các mẫu cú pháp ở một ngôn ngữ để phát hiện cụm từ nguồn. Sau đó, chúng tôi tìm bản dịch của cụm từ nguồn sử dụng mô hình gióng hàng từ ràng buộc. Các cụm từ song ngữ này đã được ứng dụng vào việc nâng cao chất lượng dịch cho dịch máy thống kê Anh - Việt.

Các nội dung và kết quả nghiên cứu trình bày trong luận án (từ Chương 2 đến Chương 4) đã được công bố trong 8 công trình. Trong đó, 1 bài báo ở tạp chí quốc tế có phản biện, được xuất bản bởi IGI Global; 4 báo cáo trong kỷ yếu của hội nghị quốc tế có phản biện, được xuất bản bởi IEEE và Springer; 2 báo cáo trong kỷ yếu của hội thảo quốc gia có phản biện và 1 bài báo ở tạp chí trong nước có phản biện.

4. Bố cục của luận án

Ngoài phần mở đầu và kết luận, luận án được tổ chức thành 4 chương:

- **Chương 1.** Giới thiệu tổng quan về các vấn đề nghiên cứu trong luận án. Chúng tôi phân tích, đánh giá các công trình nghiên cứu liên quan; nêu ra một số vấn đề còn tồn tại mà luận án sẽ tập trung giải quyết; xác định nội dung nghiên cứu của luận án.
- **Chương 2.** Trình bày nội dung, kết quả nghiên cứu về xây dựng ngữ liệu song ngữ cho dịch máy thống kê.
- **Chương 3.** Trình bày nội dung, kết quả nghiên cứu về một số cải tiến mô hình IBM để gióng hàng từ cho dịch máy thống kê.
- **Chương 4.** Trình bày nội dung, kết quả nghiên cứu về xác định cụm từ song ngữ cho dịch máy thống kê.

Chương 1

Tổng quan

1.1 Khai phá tri thức song ngữ

Nhiệm vụ của khai phá tri thức song ngữ là tự động tìm ra các thành phần có ngữ nghĩa tương ứng trong các văn bản ở hai ngôn ngữ khác nhau. Tri thức song ngữ gồm nhiều khía cạnh: song ngữ về từ, song ngữ về cụm từ, song ngữ về cấu trúc, vv.

1.1.1 Xây dựng ngữ liệu song ngữ

Ngữ liệu song ngữ là tập hợp các văn bản song ngữ. Web là nguồn cơ sở dữ liệu khổng lồ chứa các tài liệu đa ngôn ngữ, nguồn dữ liệu này được sử dụng cho các ứng dụng xử lý văn bản song ngữ. Ngoài ra, nhiều sách điện tử song ngữ chứa một số lượng lớn các văn bản song ngữ được dịch cẩn thận. Đây là nguồn dữ liệu rất tiềm năng để bổ sung ngữ liệu song ngữ cho SMT, đặc biệt đối với các cặp ngôn ngữ còn hạn chế về ngữ liệu song ngữ như Anh - Việt, Nhật - Việt, vv.

1.1.2 Gióng hàng văn bản

1.1.2.1 Gióng hàng đoạn/câu

Nhiệm vụ của gióng hàng đoạn/câu là liên kết các đoạn/câu trong một văn bản ở ngôn ngữ này với các đoạn/câu là bản dịch tương ứng của nó trong một văn bản

ở ngôn ngữ khác.

1.1.2.2 Gióng hàng từ

Gióng hàng từ là một nhiệm vụ xác định sự tương ứng giữa các từ trong một văn bản song ngữ. Đây là bước đầu tiên trong hầu hết các cách tiếp cận hiện tại của SMT. Chất lượng của gióng hàng từ đóng vai trò rất quan trọng cho sự thành công của một hệ thống SMT.

1.1.3 Xác định cụm từ song ngữ

Các cụm từ song ngữ hữu ích cho nhiều nhiệm vụ của xử lý ngôn ngữ tự nhiên như truy xuất thông tin liên ngữ, phân tích cú pháp, khai phá văn bản và đặc biệt là cho MT. Trong các hệ thống SMT, chất lượng của các bản dịch phụ thuộc chủ yếu vào chất lượng của các cặp cụm từ song ngữ được rút trích từ ngữ liệu song ngữ.

1.2 Sơ lược về dịch máy

Không lâu sau khi những chiếc máy tính điện tử đầu tiên ra đời, Warren Weaver (1949) đưa ra ý tưởng rằng, có thể một ngày nào đó máy tính nhận đầu vào là một tài liệu viết bằng một số ngôn ngữ nào đó (ngôn ngữ nguồn) và tự động tạo ra một tài liệu tương đương viết bằng một số ngôn ngữ khác (ngôn ngữ đích) - một nhiệm vụ mà bây giờ chúng ta gọi là MT. Từ đó đến nay, sau hơn 60 năm nghiên cứu và phát triển, các dịch vụ MT bây giờ đã trở nên phổ biến rộng rãi và được sử dụng miễn phí.

1.3 Dịch máy thống kê

1.3.1 Mô hình hóa bài toán

Brown và cộng sự (1993) sử dụng quy tắc Bayes để xây dựng công thức tính xác suất dịch câu nguồn $\mathbf{f}$ sang câu đích $\mathbf{e}$ như sau:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} Pr(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e}} \frac{Pr(\mathbf{f}|\mathbf{e})Pr(\mathbf{e})}{Pr(\mathbf{f})} = \arg\max_{\mathbf{e}} Pr(\mathbf{f}|\mathbf{e})Pr(\mathbf{e}) \quad (1.1)$$

Trong đó, $Pr(\mathbf{e})$ là mô hình ngôn ngữ và $Pr(\mathbf{f}|\mathbf{e})$ là mô hình dịch.

1.3.2 Mô hình ngôn ngữ

Một cách hình thức, mô hình ngôn ngữ là một hàm nhận tham số đầu vào là một câu và trả về xác suất của câu thuộc ngôn ngữ. Mô hình ngôn ngữ sẽ hỗ trợ các quyết định khó khăn về trật tự từ (word order) và dịch từ (word translation) Phương pháp hàng đầu cho các mô hình ngôn ngữ là mô hình ngôn ngữ *n-gram*.

1.3.3 Mô hình dịch

Mô hình dịch (translation model) giúp tính toán xác suất có điều kiện $Pr(\mathbf{f}|\mathbf{e})$. Xác suất này được ước lượng từ ngữ liệu song ngữ của cặp ngôn ngữ nguồn - đích.

1.3.3.1 Mô hình dịch dựa trên từ

Mô hình dịch dựa trên từ là thế hệ đầu tiên của SMT, được nghiên cứu và phát triển bởi IBM. Mô hình dịch này dựa trên sự tương ứng của các từ theo tương ứng một một. Mô hình dịch dựa trên đơn vị từ không cho kết quả tốt trong trường hợp kết nối *nhiều-1* hoặc *nhiều-nhiều* với trật tự các từ trong câu tương ứng là khác nhau. Khi đó, mô hình dựa trên đơn vị cụm từ được đề xuất để giải quyết vấn đề này.

1.3.3.2 Mô hình dịch dựa trên cụm từ

Cách tiếp cận hiện thành công nhất với SMT là sử dụng cách dịch theo cụm từ. Ở đây, cụm từ là chuỗi các từ liền kề nhau không nhất thiết là cụm từ trong ngôn ngữ học. Trong phương pháp này, câu đầu vào được chia thành một chuỗi các cụm từ; những cụm từ được ánh xạ *một-một* đến các cụm từ đầu ra, có thể được sắp xếp lại thứ tự các cụm từ. Thông thường, các mô hình cụm từ được ước lượng từ ngữ liệu song ngữ đã được gióng hàng từ. Tất cả các cặp cụm từ nhất quán với gióng hàng từ sẽ được rút trích và gán với một xác suất tương ứng.

1.3.3.3 Mô hình dịch dựa trên cú pháp

Khác với hai mô hình dịch dựa trên từ và cụm từ như đã trình bày ở trên, mô hình dịch dựa trên cú pháp sử dụng thông tin về cú pháp ngôn ngữ. Các mô hình dịch dựa trên cú pháp rất đa dạng, sử dụng các hình thức và đặc trưng ngữ pháp khác nhau. Một số cách tiếp cận thực hiện phân tích cú pháp cho câu nguồn (tree to string - dịch từ cây cú pháp sang chuỗi), một số khác tạo ra cây cú pháp khi sinh ra câu đích (string to tree - dịch từ chuỗi sang cây cú pháp) và một số kết hợp cả hai (tree to tree - dịch từ cây cú pháp sang cây cú pháp).

1.3.4 Giải mã

Mục tiêu của giải mã là tìm bản dịch với số điểm tốt nhất. Trong quá trình giải mã, chúng ta xây dựng bản dịch theo từng từ một, từ đầu đến cuối. Bộ giải mã trong mô hình SMT thường áp dụng các thuật toán tìm kiếm tối ưu. Thuật toán mà bộ giải mã thường áp dụng A*, một kỹ thuật tìm kiếm chuẩn trong trí tuệ nhân tạo.

1.3.5 Đánh giá chất lượng dịch

Có một số phương pháp đánh giá tự động chất lượng dịch như BLEU, NIST và TER. Trong đó, phương pháp BLEU được sử dụng phổ biến nhất. Ý tưởng chính của phương pháp này 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. Việc so sánh được thực hiện dựa vào kết quả thống

kê sự trùng khớp của các *n-gram* trong hai bản dịch có tính đến thứ tự của chúng trong câu.

1.4 Thảo luận

Từ những phân tích, đánh giá các nghiên cứu liên quan ở trên, chúng tôi nhận thấy một số vấn đề còn tồn tại, cụ thể như sau: Thứ nhất, đối với bài toán xây dựng ngữ liệu cho SMT, chúng ta có thể khai thác từ hai nguồn: Web và sách điện tử song ngữ. Thứ hai, gióng hàng từ đóng vai trò rất quan trọng cho sự thành công của một hệ thống SMT. Sử dụng thêm các nguồn tri thức bên ngoài như thông tin về từ vựng, thông tin về cú pháp là thật sự cần thiết để cải thiện chất lượng của gióng hàng. Thứ ba, các cụm từ song ngữ được sử dụng để bổ sung nguồn tri thức song ngữ cho các hệ thống SMT. Bouamor và cộng sự (2012) đã chỉ ra rằng, các cụm từ song ngữ được sử dụng để cải thiện chất lượng dịch cho SMT.

Chương 2

Xây dựng ngữ liệu song ngữ cho dịch máy thống kê

2.1 Rút trích văn bản song ngữ từ Web

2.1.1 Thu thập dữ liệu

Để thực hiện việc thu thập các tài liệu HTML từ Web, chúng tôi sử dụng công cụ *Teleport-Pro*. Ở đây, chúng tôi chọn các URL từ ba web-site: BBC, VietnamPlus và VOA News.

2.1.2 Thiết kế các đặc trưng dựa vào nội dung

2.1.2.1 Sử dụng cognate

Phương pháp này sử dụng các từ cùng nguồn gốc (cognate) hay còn gọi là các từ bất biến giữa hai ngôn ngữ. Với một cặp văn bản ($Etext$, $Vtext$), trong đó: $Etext$ là viết tắt của văn bản tiếng Anh và $Vtext$ là viết tắt của văn bản tiếng Việt, chúng tôi xác định các tập T_1 và T_2 chứa các *cognate* ở trong $Etext$ và $Vtext$. Độ tương tự về *cognate* của $Vtext$ với $Etext$ được xác định theo công thức (2.1)[1].

$$sim_{cognate}(Etext, Vtext) = \frac{|T_1 \cap T2|}{|T_1|} \tag{2.1}$$

[1]Lưu ý, theo cách tính của chúng tôi $sim_{cognate}(Etext, Vtext) \neq sim_{cognate}(Vtext, Etext)$

2.1.2.2 Sử dụng các phân đoạn dịch

Ký hiệu $Epage$, $Etext$, $Vpage$ và $Vtext$ lần lượt là trang web tiếng Anh, nội dung của trang web tiếng Anh, trang web tiếng Việt, nội dung của trang web tiếng Việt. Khi đó, $Etext$ được biểu diễn như là một chuỗi các đoạn $pe_1 pe_2 \dots pe_n$ và $Vtext$ được biểu diễn như là một chuỗi các đoạn $pv_1 pv_2 \dots pv_m$. Trong đó, pe_i và pv_j tương ứng là các đoạn trong văn bản tiếng Anh và tiếng Việt. Chúng tôi thiết kế hàm $Similarity_{paragraph}(pe, pv)$ để đo mối quan hệ dịch giữa pe và pv. Như vậy, đối với mỗi pe_i chúng ta cần tìm pv_j thích hợp nhất được ký hiệu như trong công thức (2.2).

$$pv_j = \arg\max_{pv_k} Similarity_{paragraph}(pe_k, pv_i),\ \mathrm{k} = 1,\ \dots,\ \mathrm{n} \tag{2.2}$$

2.1.3 Thiết kế các đặc trưng dựa vào cấu trúc

Quá trình phân tích cấu trúc được thực hiện theo hai bước. Tại bước đầu tiên, hai trang web là cặp ứng viên được phân tích thông qua một bộ phân tích thẻ HTML. Ở bước thứ hai, chúng tôi thực hiện gióng hàng các thẻ thu được ở bước 1.

2.1.4 Mô hình hóa bài toán phân loại

Mỗi cặp ứng viên của trang web song ngữ được biểu diễn bởi một véc-tơ đặc trưng. Gọi $F = \{f_1, f_2, ..., f_m\}$ là tập đặc trưng, $D = \{d_1, d_2, ..., d_n\}$ là tập chứa tất cả các cặp ứng viên và $C = \{0, 1\}$ là tập các loại (0: không song ngữ, 1: song ngữ). Khi đó, mỗi cặp ứng viên $d_i \in D$ được biểu diễn bởi véc-tơ đặc trưng $d_i = (f_{1i}, f_{2i}, ..., f_{mi})$. Chúng tôi gắn nhãn cho chúng là 1 hoặc 0 nếu mỗi cặp tương ứng là song ngữ hoặc không song ngữ. Bằng cách này, chúng ta sẽ có được dữ liệu huấn luyện. Ở đây, chúng tôi sử dụng thuật toán SVM để huấn luyện hệ thống phân loại. Đối với một cặp trang web mới, đầu tiên chúng tôi rút trích tập đặc trưng F để có thể biểu diễn nó như là một véc-tơ. Véc-tơ này đi qua hệ thống phân loại và nhận được kết quả là 1 hoặc 0.

2.2 Rút trích câu song ngữ từ sách điện tử

2.2.1 Tiền xử lý

Sách điện tử ban đầu ở định dạng PDF sẽ được chuyển đổi sang định dạng Text. Sau đó, chúng tôi tiến hành khôi phục lại ranh giới giữa các đoạn. Tiếp theo, chúng tôi sử dụng một hệ thống SMT để dịch văn bản trong sách tiếng Anh sang tiếng Việt.

2.2.2 Đo độ tương tự

Giả sử chúng ta đang làm việc với sách điện tử song ngữ Anh - Việt. Sách tiếng Anh $\boldsymbol{E}$ chứa I khối (văn bản) $ue_1, ..., ue_I$ và sách tiếng Việt $\boldsymbol{V}$ chứa J khối $uv_1, ..., uv_J$. Gọi $\boldsymbol{T}$ là bản dịch tiếng Việt của $\boldsymbol{E}$ và ut_i là bản dịch tiếng Việt của khối ue_i (trong $\boldsymbol{E}$). Gọi $S_n(ut_i)$ và $D_n(uv_j)$ lần lượt là các tập $n-gram$ của các khối ut_i và uv_j. Độ tương tự giữa các khối ut_i và uv_j được định nghĩa như trong công thức (3.2).

$$Similarity_n(ut_i, uv_j) = \frac{|S_n(ut_i) \cap D_n(uv_j)|}{|S_n(ut_i) \cup D_n(uv_j)|} \tag{2.3}$$

Trong công thức này, $Similarity_n(ut_i, uv_j)$ là độ tương tự giữa hai khối văn bản ut_i và uv_j khi phân chia theo n, $0 \leq Similarity_n(ut_i, uv_j) \leq 1$.

2.2.3 Gióng hàng đoạn

Chúng tôi tính toán độ tương tự của các khối theo các mẫu $1-1$, $1-2$, $1-3$, $2-1$ và $3-1$ bằng cách sử dụng hàm $Similarity_n(ut_i, uv_j)$ như trong công thức (**??**). Sau đó, cặp khối (u_s, u_t) có độ tương tự lớn nhất sẽ được chọn theo công thức (3.3).

$$(u_s, u_t) = \arg\max \begin{cases} Similarity_n(pt_i, pv_j) \\ Similarity_n(pt_i, pv_j pv_{j+1}) \\ Similarity_n(pt_i, pv_j pv_{j+1} pv_{j+2}) \\ Similarity_n(pt_i pt_{i+1}, pv_j) \\ Similarity_n(pt_i pt_{i+1} pt_{i+2}, pv_j) \end{cases} \tag{2.4}$$

2.2.4 Gióng hàng câu

Nhiệm vụ của chúng ta là cần tìm ra câu ở vị trí thứ x ở trong đoạn pe là dịch của câu ở vị trí thứ y ở trong đoạn pv. Cặp câu (se_x, sv_y) có độ tương tự lớn nhất sẽ được lựa chọn như trong công thức (2.5).

$$(se_x, sv_y) = \arg\max \begin{cases} Similarity_n(st_i, sv_j) \\ Similarity_n(st_i, sv_{j+1}) \\ Similarity_n(st_i, sv_{j+2}) \\ Similarity_n(st_{i+1}, sv_j) \\ Similarity_n(st_{i+2}, sv_j) \\ Similarity_n(st_i, sv_j sv_{j+1}) \\ Similarity_n(st_i, sv_j sv_{j+1} sv_{j+2}) \\ Similarity_n(st_i st_{i+1}, sv_j) \\ Similarity_n(st_i st_{i+1} st_{i+2}, sv_j) \end{cases} \tag{2.5}$$

2.3 Thực nghiệm

2.3.1 Thực nghiệm về rút trích văn bản song ngữ từ Web

Chúng tôi tải về 64.323 trang web từ ba web-site: BBC, VOA, VietnamPlus. Tiếp theo, chúng tôi tạo ra các cặp ứng viên từ nguồn dữ liệu thu thập được sử dụng một số ngưỡng: $sim_{cognate} > 0,5$ và $distance_{date} \leq 1$. Từ đó, chúng tôi nhận được 1.170 cặp ứng viên. Tiếp theo, chúng tôi thiết kế các đặc trưng về nội dung và cấu trúc cho tất cả các cặp ứng viên như trình bày ở các phần trước. Các kết quả thực nghiệm cho thấy, hai phương pháp chúng tôi đề xuất đạt được kết quả tốt hơn (độ chính xác 88,2% và 90,0%) so với phương pháp sử dụng các đặc trưng dựa vào cấu trúc trang web của Resnik (độ chính xác 44,4%) và phương pháp sử dụng từ điển của Ma (độ chính xác 65,2%).

2.3.2 Thực nghiệm về rút trích câu song ngữ từ sách điện tử

Chúng tôi sử dụng bốn cuốn sách điện tử song ngữ Anh - Việt làm dữ liệu thực nghiệm. Để đo độ tương tự giữa hai khối văn bản (ut_i and uv_j), chúng tôi sử dụng

công thức (3.2) với $n = 1$. Chúng tôi chọn ngẫu nhiên 200 mẫu (của đoạn) từ dữ liệu thực nghiệm để đánh giá hiệu suất của phương pháp đã đề xuất. Kết quả thực nghiệm đạt được độ chính xác là 97%. Chúng tôi thiết kế bộ dữ liệu gồm 40 đoạn song ngữ có chứa 202 câu song ngữ. Phương pháp của Gale được sử dụng như phương pháp baseline. Phương pháp chúng tôi đã đạt được điểm số cao hơn trên cả hai độ đo *precision* và *recall*.

2.3.3 Thực nghiệm về bổ sung ngữ liệu song ngữ cho dịch máy

Chúng tôi bổ sung 21.072 câu song ngữ Anh - Việt (từ nguồn ngữ liệu song ngữ xây dựng được) vào hệ thống SMT Anh - Việt được xây dựng trên tập dữ liệu huấn luyện gồm 90.000 câu song ngữ Anh - Việt cho mô hình dịch và 100.000 câu tiếng Việt cho mô hình ngôn ngữ. Kết quả, chất lượng dịch tăng lên 3% (tương đương với 0,6 điểm BLEU) so với hệ thống SMT ban đầu.

2.4 Kết luận chương

Chúng tôi đã trình bày các nội dung, kết quả nghiên cứu về xây dựng ngữ liệu song ngữ cho SMT. Trong nghiên cứu của chúng tôi, ngữ liệu song ngữ được khai thác từ Web và sách điện tử song ngữ. Các kết quả đạt được cho thấy, chúng tôi có thể đạt được ngữ liệu song ngữ Anh - Việt đủ để xây dựng một hệ thống SMT thông qua việc khai thác ngữ liệu song ngữ từ hai nguồn này.

Chương 3

Gióng hàng từ cho dịch máy thống kê

3.1 Cơ sở lý thuyết

3.1.1 Định nghĩa từ

Theo Diệp Quang Ban, về mặt nghiên cứu chung, người ta gặp không ít khó khăn trong việc xác định và nêu định nghĩa từ. Về mặt ngữ pháp, có thể hiểu từ là đơn vị nhỏ nhất có nghĩa và hoạt động tự do trong câu.

3.1.2 Định nghĩa bài toán gióng hàng từ

Cho câu **f** ở ngôn ngữ nguồn (câu nguồn) chứa J từ $f_1, ..., f_J$ và câu **e** ở ngôn ngữ đích (câu đích) chứa I từ $e_1, ..., e_I$, chúng tôi định nghĩa liên kết $l = (i, j)$ tồn tại nếu e_i và f_j là dịch (hoặc dịch một phần) của nhau. Khi đó, một gióng hàng từ **a** (giữa **f** và **e**) là một ánh xạ từ các vị trí từ trong **f** đến các vị trí từ trong **e**:

$$\mathbf{a} : j \to i, \text{ với } j = 1, ..., J \text{ và } i = 0, ..., I \tag{3.1}$$

3.1.3 Các mô hình IBM

Cho câu nguồn $\mathbf{f} = f_1, f_2, \ldots f_J$ với độ dài J, câu đích $\mathbf{e} = e_1, e_2, \ldots e_I$ với độ dài I và tập hợp các gióng hàng từ $\mathbf{a}$. Khi đó, với mô hình IBM 1 xác suất $P(\mathbf{f}, \mathbf{a}|\mathbf{e})$ được tính theo công thức (3.2).

$$Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \frac{\varepsilon}{(I+1)^J} \prod_{j=1}^{J} t(f_j|e_{a_j}) \tag{3.2}$$

Khi mô hình IBM 1 được cải tiến, các kết quả thu được từ mô hình này sẽ chuyển đến các mô hình IBM cao hơn (IBM 2-5). Vì vậy, về tổng thể, nó sẽ cải tiến các mô hình IBM.

3.1.4 Thuật toán cực đại kỳ vọng cho mô hình IBM 1

Thuật toán EM cho mô hình IBM 1 bao gồm hai bước: (i) bước E: áp dụng mô hình đến dữ liệu, các xác suất gióng hàng được tính toán từ các tham số mô hình; (ii) bước M: ước lượng mô hình từ dữ liệu, giá trị của các tham số được ước lượng lại dựa trên các xác suất gióng hàng và dữ liệu.

3.2 Một số cải tiến mô hình IBM 1 theo cách tiếp cận dựa trên ràng buộc

3.2.1 Cải tiến mô hình IBM 1 sử dụng ràng buộc neo

Gióng hàng giữa hai từ trong một điểm neo được tạo ra bằng cách thiết lập xác suất dịch bằng không ở vị trí đó cho tất cả các từ khác. Chúng tôi lựa chọn những từ không được dịch và nó cùng xuất hiện trong cặp câu song ngữ (ví dụ: chữ viết tắt, chữ số,...). Ngoài ra, chúng tôi sử dụng thêm các cặp từ song ngữ (từ dữ liệu huấn luyện). Chúng tôi định nghĩa danh sách L là tập hợp các cặp từ song ngữ như sau:

$$L = \{(f_j, e_i)|t(f_j|e_i) > \alpha, count(f_j, e_i) > \beta\}. \tag{3.3}$$

Ở đây, e_i là từ ở ngôn ngữ nguồn, f_j là từ ở ngôn ngữ đích và α, β là các ngưỡng được xác định trước.

3.2.2 Cải tiến mô hình IBM 1 sử dụng ràng buộc về vị trí của từ

Ràng buộc về vị trí của từ giới hạn phạm vi gióng hàng giữa các từ trong một cặp câu song ngữ. Với mỗi cặp từ (f,e) trong cặp câu (**f**,**e**), chúng tôi gán trọng số cao hơn nếu ràng buộc về vị trí của từ được thỏa mãn và trọng số thấp hơn trong trường hợp ngược lại. Tức là, xác suất gióng hàng giữa f và e được nhân với trọng số λ khi ràng buộc được thỏa mãn và nhân với $(1 - \lambda)$ nếu ràng buộc không thỏa mãn.

3.2.3 Cải tiến mô hình IBM 1 sử dụng ràng buộc về từ loại

Ký hiệu R là tập hợp các quan hệ về POS giữa tiếng Anh và tiếng Việt, như sau:

$$R = \{(x \rightarrow y) | x \in X, y \in Y\} \tag{3.4}$$

Trong đó, X và Y tương ứng là tập chứa các thẻ POS của tiếng Anh và tiếng Việt. Ở đây, ràng buộc POS đòi hỏi mỗi từ nguồn f_j chỉ gióng hàng với các các từ đích e_i có cùng quan hệ về POS. Ký hiệu $P(f_j)$, $P(e_i)$ tương ứng với thẻ POS của từ nguồn f_j và từ đích e_i. Khi đó, một cặp từ (f_j, e_i) thỏa mãn ràng buộc POS nếu $P(f_j) \rightarrow P(e_i) \in R$.

3.2.4 Cải tiến mô hình IBM 1 sử dụng ràng buộc về cụm từ

Giả sử rằng, chúng ta có cặp câu (**f**,**e**) trong ngữ liệu song ngữ so khớp với mẫu cú pháp song ngữ tại vị trí (j_1, j_2) ở câu nguồn và (i_1, i_2) ở câu đích. Bây giờ, chúng tôi tách mỗi câu thành ba phần $\mathbf{f} = \overline{f_1}, \overline{f_2}, \overline{f_3}$ và $\mathbf{e} = \overline{e_1}, \overline{e_2}, \overline{e_3}$. Ở đây, ràng buộc về cụm từ yêu cầu mỗi từ f_j trong cụm từ nguồn $\overline{f_2}$ chỉ gióng hàng với các từ e_i trong cụm từ đích $\overline{e_2}$. Tương tự, các từ ngoài cụm từ nguồn gióng hàng với các từ ngoài cụm từ đích.

3.2.5 Kết hợp các ràng buộc

Gọi $C = \{c_1, c_2, ..., c_K\}$ là tập các ràng buộc. Cặp từ (f, e) (trong cặp câu (**f**,**e**)) được gọi là thỏa mãn ràng buộc nếu nó thỏa mãn một ràng buộc $c_k \in C$ bất kỳ, $1 \leq k \leq K$ (tức là, thỏa mãn ràng buộc c_1 hoặc c_2, ..., hoặc c_K). Gọi $E_C = \{e_1, e_2, ..., e_n\}$ là tập hợp các từ trong **e** thỏa mãn ràng buộc. Khi đó, hàm c được định nghĩa lại như sau:

$$c(f|e; \mathbf{f}, \mathbf{e}, C) = \frac{t(f|e)}{\sum_{e_k \in E_C} t(f|e_k)} \sum_{j=1}^{J} \delta(f, f_j) \sum_{i=0}^{I} \delta(e, e_i) \tag{3.5}$$

Về cơ bản, việc ước lượng xác suất dịch $t(f|e)$ và tích hợp tập ràng buộc C vào thuật toán EM cho mô hình IBM 1 được thực hiện tương tự như với các ràng buộc chúng tôi đã trình bày ở trên.

3.3 Thực nghiệm

Quá trình thực nghiệm, đánh giá về gióng hàng từ được thực hiện trên hệ thống SMT Anh - Việt (dịch từ tiếng Anh sang tiếng Việt). Chúng tôi thiết kế bốn tập dữ liệu huấn luyện lần lượt chứa 60.000, 70.000, 80.000 và 90.000 câu song ngữ Anh - Việt. Tập dữ liệu gồm 1.000 câu song ngữ Anh - Việt được sử dụng để đánh giá chất lượng dịch.

3.3.1 Kết quả thực nghiệm với ràng buộc neo và ràng buộc về vị trí của từ

Mô hình IBM được cải tiến với việc sử dụng hai ràng buộc này đã đạt được điểm BLEU cao hơn so với mô hình IBM gốc trên cả bốn tập dữ liệu huấn luyện. Cụ thể, điểm BLEU tăng trung bình 0,67 điểm với ràng buộc neo và 1,48 điểm với ràng buộc về vị trí của từ. Ngoài ra, so với Giza++, tính trung bình trên cả bốn tập dữ liệu, phương pháp của chúng tôi đạt được điểm BLEU cao hơn 0,28 điểm khi sử dụng ràng buộc neo và 1,08 điểm khi sử dụng ràng buộc về vị trí của từ.

3.3.2 Kết quả thực nghiệm với ràng buộc từ loại

Sử dụng ràng buộc về từ loại đạt được điểm BLEU cao hơn trên tất cả các tập dữ liệu huấn luyện so với mô hình IBM gốc và Giza++. Cụ thể, khi sử dụng ràng buộc về từ loại điểm BLEU tăng trung bình 0,98 điểm, tương đương với việc chất lượng MT tăng 4,31% so với mô hình IBM gốc. Ngoài ra, so với sử dụng Giza++, phương pháp dùng ràng buộc từ loại đạt được chất lượng dịch tốt hơn 2,50%.

3.3.3 Kết quả thực nghiệm với ràng buộc cụm từ

Kết quả thực nghiệm cho thấy, cải tiến của chúng tôi đạt được điểm BLEU cao hơn so với mô hình IBM gốc trên tất cả các tập dữ liệu huấn luyện. Cụ thể, điểm BLEU tăng trung bình $0,45$ điểm so với mô hình IBM gốc không sử dụng ràng buộc. So sánh với Giza++, phương pháp dùng ràng buộc cụm từ đạt được điểm BLEU cao hơn trung bình $0,05$ điểm.

3.3.4 Kết quả thực nghiệm về kết hợp ràng buộc

Khi chúng tôi kết hợp ràng buộc về vị trí của từ với ràng buộc về từ loại, chất lượng dịch tốt hơn so với việc sử dụng riêng lẻ từng ràng buộc. Cụ thể, so với mô hình IBM gốc điểm BLEU tăng trung bình 1,63 điểm khi kết hợp ràng buộc, tương đương với việc chất lượng MT tăng 7,16% với độ tin cậy $p \leq 0,0007$. So với việc sử dụng Giza++, phương pháp kết hợp ràng buộc này đạt được điểm BLEU cao hơn trung bình 1,23 điểm với độ tin cậy $p \leq 0,0034$.

3.4 Kết luận chương

Chúng tôi đã đề xuất một số cải tiến mô hình IBM 1 theo cách tiếp cận dựa trên ràng buộc, cụ thể là: ràng buộc neo, ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ. Các ràng buộc này sau đó được sử dụng để ước lượng các tham số của mô hình trong thuật toán EM. Kết quả thực nghiệm cho thấy các cải tiến của chúng tôi cải thiện hiệu suất dịch cho hệ thống SMT Anh - Việt.

Chương 4

Xác định cụm từ song ngữ cho dịch máy thống kê

4.1 Bài toán rút trích cụm từ song ngữ

Cho một cụm từ *pe* ở ngôn ngữ nguồn (tiếng Anh) và một cụm từ *pv* ở ngôn ngữ đích (tiếng Việt). Chúng tôi định nghĩa một cặp cụm từ $p = (pe, pv)$ là một cụm từ song ngữ nếu cụm từ nguồn *pe* và cụm từ đích *pv* là bản dịch của nhau, tức là, không có bổ sung từ trong cụm từ đích mà không thể tìm thấy từ tương ứng trong cụm từ nguồn và ngược lại. Cho ngữ liệu $C = \{(\mathbf{f}^{(l)}, \mathbf{e}^{(l)})\}$ chứa các câu song ngữ Anh - Việt. Trong đó, $1 \leq l \leq N$ và N là kích thước của ngữ liệu. Bài toán đặt ra ở đây là tìm và rút trích các cụm từ song ngữ trong ngữ liệu C.

4.2 Phương pháp rút trích cụm từ song ngữ

Phương pháp của chúng tôi mở rộng ý tưởng của Vogel về gióng hàng từ ràng buộc. Trong phần này, chúng tôi sẽ trình bày chi tiết ba bước chính để rút trích các cụm từ song ngữ như sau: (i) xác định cụm từ, (ii) tìm cụm từ đích và (iii) rút trích cụm từ song ngữ.

4.2.1 Xác định cụm

Chúng tôi sử dụng các mẫu cú pháp được xác định trước để phát hiện và rút trích các cụm từ song ngữ từ ngữ liệu song ngữ Anh - Việt. Giả sử chúng ta có một cặp câu (**f**,**e**) từ ngữ liệu song ngữ so khớp với một cặp mẫu cú pháp tại các vị trí (j_1, j_2) trong câu nguồn và (i_1, i_2) trong câu đích. Từ đó, chúng tôi rút trích các cụm từ nguồn $pe = f_{j_1}...f_{j_2}$ và cụm từ đích $pv = e_{i_1}...e_{i_2}$. Trong trường hợp chỉ so khớp ở một phía (trong câu **f** hoặc **e**), ta xác định cụm từ này (chúng tôi gọi là cụm từ nguồn) và tìm cụm từ còn lại (chúng tôi gọi là cụm từ đích).

4.2.2 Tìm cụm từ đích

Giả sử, cho cặp câu (**f**,**e**) và cụm nguồn $pe = f_{j_1}...f_{j_2}$, chúng tôi cần tìm một chuỗi các từ $e_{i_1}...e_{i_2}$ trong câu đích, là bản dịch của cụm từ nguồn. Để thực hiện công việc này, chúng tôi sử dụng mô hình gióng hàng từ ràng buộc được mô tả trong công thức (4.1).

$$\begin{aligned} Pr_{i_1,i_2}(\mathbf{f}|\mathbf{e}) = \prod_{j=1}^{j_1-1} \sum_{i \notin (i_1..i_2)} \frac{1}{I-k} t(f_j|e_i) \\ \times \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i2} \frac{1}{k} t(f_j|e_i) \\ \times \prod_{j=j_2+1}^{J} \sum_{i \notin (i_1..i_2)} \frac{1}{I-k} t(f_j|e_i) \end{aligned} \tag{4.1}$$

Ranh giới i_1 và i_2 của cụm từ pv trong câu đích được xác định bởi công thức (4.2).

$$(i_1, i_2) = \arg\max_{i_1,i_2} \{Pr_{i_1,i_2}(\mathbf{f}|\mathbf{e})\} \tag{4.2}$$

4.2.3 Rút trích cụm từ

Chúng tôi thực hiện rút trích các ứng viên của cụm từ song ngữ, như sau:

- Ước lượng xác suất $t(f|e)$
- Với mỗi cặp câu $(\mathbf{f}^{(l)}, \mathbf{e}^{(l)}), 1 \leq l \leq N$[1]:

[1] N là kích thước của ngữ liệu.

– Với mỗi cặp mẫu cú pháp trong tập các mẫu cú pháp được xác định trước:

  * Nếu một cặp mẫu cú pháp được so khớp thì (pe, pv) là một ứng viên của cụm từ song ngữ.
  * Ngoài ra, nếu một mẫu cú pháp trong ngôn ngữ nguồn được so khớp thì rút trích cụm từ nguồn pe và tìm kiếm cụm từ đích pv dùng công thức (4.2).

Để lọc cụm từ song ngữ (loại bỏ các cụm sai), chúng tôi tính xác suất dịch cụm từ bằng cách sử dụng tần suất tương đối:

$$Pr(pv|pe) = \frac{N(pv, pe)}{N(pe)} \tag{4.3}$$

Trong công thức (4.3), pe và pv lần lượt là cụm từ nguồn và đích. $N(pe, pv)$ là số lần cụm pe được dịch bởi pv và $N(pe)$ là số lần pe xuất hiện trong ngữ liệu. Để tăng độ tin cậy, chúng tôi sử dụng giá trị nhỏ nhất của hai tần suất tương đối như là xác suất dịch cụm từ, như thể hiện trong công thức (4.4).

$$Pr(pv|pe) = min(Pr(pv|pe), Pr(pe|pv)) \tag{4.4}$$

4.3 Tích hợp cụm từ song ngữ vào dịch máy

Chúng tôi tích hợp các cụm từ song ngữ sau khi được rút trích từ ngữ liệu vào hệ thống SMT Anh - Việt theo hai cách: (i) xây dựng thêm một bảng cụm từ từ các cụm từ song ngữ được rút trích tự động và (ii) sử dụng các cụm từ song ngữ được rút trích tự động như là cặp câu song ngữ và thêm chúng vào dữ liệu huấn luyện, sau đó huấn luyện lại mô hình dịch.

4.4 Thực nghiệm

4.4.1 Thực nghiệm về rút trích cụm từ song ngữ

4.4.1.1 Cài đặt thực nghiệm

Các thực nghiệm về rút trích cụm từ song ngữ được thực hiện trên 5.000 câu song ngữ Anh - Việt. Để gán nhãn từ loại cho dữ liệu thực nghiệm, chúng tôi sử dụng các bộ công cụ: *vnTagger* cho văn bản tiếng Việt và *posTagger-1.0* cho văn bản tiếng Anh. Chúng tôi xây dựng một tập hợp các cặp mẫu cú pháp tiếng Anh và tiếng Việt, tập này bao gồm 10 cặp mẫu.

4.4.1.2 Kết quả thực nghiệm

Theo kết quả từ các thực nghiệm, chúng tôi thấy rằng với ngưỡng $\theta = 0,25$ chúng tôi đạt được kết quả tốt nhất, trong đó sự cân bằng giữa *precision* và *recall* được đảm bảo. Ngoài ra, chúng tôi so sánh giữa phương pháp đề xuất với phương pháp so khớp mẫu cú pháp ở hai phía *baseline*. Kết quả, chúng tôi đã đạt được điểm số cao hơn trên cả hai độ đo *precision* và *recall*. Điểm F_{score} của phương pháp chúng tôi là $36,07$ trong khi F_{score} của *baseline* là $20,07$. Phương pháp chúng tôi tăng $79,72\%$ điểm F_{score} khi so sánh với *baseline*. Các kết quả này đã cho thấy phương pháp đề xuất của chúng tôi là rất hiệu quả.

4.4.2 Thực nghiệm về tích hợp cụm từ song ngữ vào dịch máy

4.4.2.1 Cài đặt thực nghiệm

Chúng tôi sử dụng 200.000 câu song ngữ Anh - Việt được thu thập từ các *web-site* và sách điện tử song ngữ. Hệ thống SMT Anh - Việt dựa trên cụm từ được xây dựng với các thành phần như sau: (i) Mô hình ngôn ngữ với công cụ SRILM: Chúng tôi xây dựng mô hình ngôn ngữ *3-gram* sử dụng kỹ thuật làm trơn Kneyser-Ney trên ngữ liệu 1.430.177 câu tiếng Việt chứa 22.056.253 từ và 317.028 từ vựng; (ii) Mô hình dịch và giải mã sử dụng công cụ MOSES. Tập dữ liệu bao gồm 1.000 cặp câu được sử dụng để đánh giá chất lượng dịch theo độ đo BLEU.

4.4.2.2 Kết quả thực nghiệm

Chúng tôi tích hợp các cụm từ song ngữ được rút trích tự động từ hai tập ngữ liệu: 100.000 và 200.000 câu song ngữ vào hệ thống SMT Anh - Việt. Chất lượng dịch tăng tương ứng là 0,35 và 0,41 điểm BLEU khi thêm cụm từ vào dữ liệu huấn luyện (sau đó huấn luyện lại mô hình dịch) và xây dựng thêm một bảng cụm từ (từ các cụm từ song ngữ được rút trích tự động). Ngoài ra, kết hợp giữa hai phương pháp đạt được kết quả cao hơn với điểm BLEU tăng 0,53.

4.5 Kết luận chương

Chúng tôi đã trình bày phương pháp dựa trên cách tiếp cận lai để rút trích cụm từ song ngữ từ ngữ liệu song ngữ Anh - Việt và ứng dụng cho SMT. Phương pháp của chúng tôi kết hợp giữa các mẫu cú pháp được xác định trước và xác suất dịch cụm từ để rút trích các cụm từ song ngữ. Bằng cách sử dụng các mẫu cú pháp ở một phía và áp dụng mô hình gióng hàng từ ràng buộc để tìm bản dịch của cụm từ nguồn, chúng tôi có thể rút trích nhiều cụm từ song ngữ hơn. Các kết quả thu được đã cho thấy hiệu quả của đề xuất này. Khi tích hợp các cụm từ song ngữ được rút trích tự động vào hệ thống SMT, chất lượng dịch đã cải thiện đáng kể.

Kết luận

Luận án chúng tôi tập trung vào việc khai phá tri thức song ngữ và ứng dụng trong dịch máy Anh- Việt. Chúng tôi đã đề xuất một số phương pháp để xây dựng ngữ liệu song ngữ cho dịch máy thống kê, đưa ra một số cải tiến mô hình IBM để gióng hàng từ cho dịch máy thống kê và xác định cụm từ song ngữ cho dịch máy thống kê. Các đóng góp chính của luận án có thể được tóm tắt như sau:

Thứ nhất, chúng tôi đã đề xuất một số phương pháp để xây dựng ngữ liệu song ngữ cho SMT. Cụ thể, chúng tôi khai thác từ hai nguồn: Web và sách điện tử song ngữ. Đối với nguồn từ Web, chúng tôi rút trích các văn bản song ngữ từ các trang web song ngữ Anh - Việt; đưa ra hai phương pháp thiết kế các đặc trưng dựa trên nội dung: dựa trên *cognate* và dựa trên việc xác định các phân đoạn dịch. Các phương pháp chúng tôi đề xuất đạt được kết quả tốt hơn (độ chính xác 88,2% và 90,0%) so với phương pháp sử dụng các đặc trưng dựa vào cấu trúc trang *web* (độ chính xác 44,4%) và phương pháp sử dụng từ điển (độ chính xác 65,2%). Đối với nguồn từ sách điện tử song ngữ, chúng tôi sử dụng một số mẫu liên kết giữa các khối văn bản trong hai ngôn ngữ để rút trích các câu song ngữ. Các thực nghiệm về rút trích câu song ngữ từ sách điện tử theo phương pháp chúng tôi đề xuất đã đạt được $95,0\%$ theo độ đo F_{score}.

Thứ hai, chúng tôi đã đề xuất một số cải tiến đối với mô hình IBM 1 theo cách tiếp cận dựa trên ràng buộc, bao gồm: ràng buộc neo, ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ. Với mỗi ràng buộc, chúng tôi đưa ra phương pháp tổng quát để tích hợp nó vào thuật toán EM trong quá trình ước lượng tham số của mô hình. Việc cải tiến này giúp nâng cao chất lượng dịch cho các hệ thống SMT. Cụ thể, với phương pháp kết hợp ràng buộc, chất lượng MT tăng 7,16% so với mô hình IBM gốc và tăng 5,31% so với sử dụng Giza++.

Thứ ba, chúng tôi đã đề xuất phương pháp rút trích cụm từ song ngữ từ ngữ liệu song ngữ, sử dụng các mẫu cú pháp kết hợp với gióng hàng cụm từ. Các cụm từ song ngữ này được ứng dụng vào việc tăng chất lượng SMT. Các thực nghiệm được thực hiện trên hệ thống SMT Anh - Việt cho thấy phương pháp xác định cụm từ song ngữ như chúng tôi đưa ra đạt được chất lượng dịch tốt hơn so với không xử lý cụm từ, cụ thể trong trường hợp tốt nhất điểm BLEU tăng 0,53.

Danh mục công trình khoa học của tác giả liên quan đến luận án

[1] Le Quang Hung and Le Anh Cuong (2010), "Extracting parallel texts from the web", *Proceedings of the Second International Conference on Knowledge and Systems Engineering, IEEE Computer Society*, pages 147-151.

[2] Le Quang Hung and Le Anh Cuong (2012), "Improving Word Alignment for Statistical Machine Translation Based on Constraints", *Asian Language Processing (IALP), International Conference on, IEEE Computer Society*, pages 113-116.

[3] Le Quang Hung and Le Anh Cuong (2012), "Statistical Word Alignment with Part-of-Speech Constraint", *Kỷ yếu hội thảo Quốc gia lần thứ XV "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông"*, trang 410-416.

[4] Quang-Hung LE, Duy-Cuong NGUYEN, Duc-Hong PHAM, Anh-Cuong LE, and Van-Nam HUYNH (2013), "Paragraph Alignment for English-Vietnamese Parallel E-Books", In *Knowledge and Systems Engineering, Springer International Publishing*, pages 251-259.

[5] Quang-Hung LE, Anh-Cuong LE, and Van-Nam HUYNH (2013), "Parallel phrase extraction from English-Vietnamese parallel corpora", In *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on*, pages 175-179.

[6] Le Quang Hung and Le Anh Cuong (2013), "An effective method to sentence alignment for the English-Vietnamese parallel e-book", *Kỷ yếu hội thảo Quốc gia lần thứ XVI "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông"*, trang 12-16.

[7] Le Quang Hung (2014), "A new approach to extract parallel corpus", *Tạp chí khoa học Trường Đại học Quy Nhơn*, Số 4, Tập VIII, trang 12-24.

[8] Quang-Hung LE and Anh-Cuong LE (2014), "Syntactic pattern based Word Alignment for Statistical Machine Translation", *The International Journal of Knowledge and Systems Science (IJKSS), IGI Global Publishing*, Volume 5 Issue 3, pages 36-45.