(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **1-0052681**

(51)[2022.01] **G06F 40/58**

(13) **B**

(21) 1-2023-08941

(22) 14/12/2023

(45) 27/10/2025 451

(43) 26/02/2024 431

(73) TẬP ĐOÀN CÔNG NGHIỆP - VIỄN THÔNG QUÂN ĐỘI (VN)
Lô D26 Khu đô thị mới Cầu Giấy, phường Yên Hoà, quận Cầu Giấy, thành phố Hà Nội

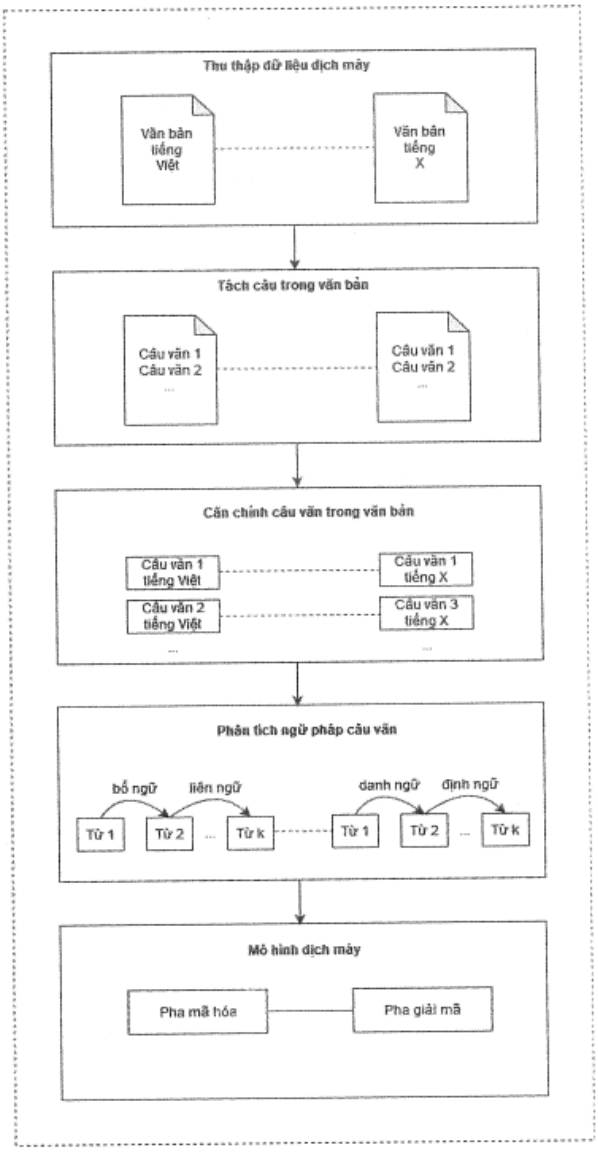
(72) Đoàn Xuân Dũng (VN); Nguyễn Ngọc Dũng (VN).

(74) Công ty TNHH NACILAW (NACILAW)

(54) PHƯƠNG PHÁP XÂY DỰNG MÔ HÌNH DỊCH MÁY SỬ DỤNG THÔNG TIN NGỮ PHÁP

(21) 1-2023-08941

(57) Sáng chế đề xuất phương pháp xây dựng mô hình dịch máy sử dụng thông tin ngữ pháp từ ngôn ngữ khác sang tiếng tiếng Việt và ngược lại. Cụ thể, sáng chế cải thiện chất lượng dịch máy bằng việc tăng cường thông tin ngữ pháp vào trong mô hình. Các mô hình dịch máy trên thị trường hiện tại đều học thông tin ngữ pháp dưới dạng các đặc trưng trong mô hình huấn luyện. Thông tin ngữ pháp sẽ không được học đầy đủ, dẫn đến các từ dịch sẽ sai văn cảnh. Do đó, sáng chế sẽ tập trung khai thác thông tin ngữ pháp trong dữ liệu huấn luyện. Từ đó, câu văn dịch sẽ chính xác, đúng văn cảnh và ngữ pháp.



Hình 1

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp dịch máy sử dụng thông tin ngữ pháp. Cụ thể là sáng chế đề cập đến phương pháp dịch máy câu văn, tập trung vào việc khai thác thông tin ngữ pháp trong câu văn.

Tình trạng kỹ thuật của sáng chế

Hiện nay, xu thế toàn cầu hóa ngày càng rõ ràng trong các lĩnh vực của cuộc sống. Cùng với đó là sự bùng nổ các ứng dụng dịch máy để phục vụ nhu cầu của con người trong việc giao tiếp và làm việc. Dịch máy là quá trình tự động dịch nội dung từ ngôn ngữ này sang ngôn ngữ khác mà không có sự can thiệp của con người trong quá trình dịch. Có nhiều phương pháp cho vấn đề này như phương pháp luật (rule-based machine translation), phương pháp dịch máy thống kê (statistical machine translation), phương pháp dịch máy sử dụng mạng nơ-ron (neural machine translation). Cùng với bước tiến vượt bậc của vi xử lý đồ họa, phương pháp mạng nơ-ron trở thành phương pháp cho kết quả dịch tốt nhất và đang được tiếp tục nghiên cứu và phát triển hiện nay.

Các nghiên cứu về dịch máy chủ yếu tập trung vào mức dịch câu văn. Các ứng dụng dịch máy mức câu văn trên thị trường vẫn còn gặp phải nhiều vấn đề như từ không đúng với văn cảnh, dịch sai các từ vựng trong các lĩnh vực chuyên môn như y tế, quân sự,… Các công nghệ dịch máy hiện tại đã mang lại những kết quả cao đối với các cặp ngôn ngữ nhiều tài nguyên. Tuy nhiên, với ngôn ngữ ít tài nguyên như Tiếng Việt, dịch máy vẫn cần phải cải thiện thêm, đặc biệt ở các lĩnh vực đặc thù.

Bản chất kỹ thuật của sáng chế

Vấn đề phân tích ngữ pháp trong câu khi dịch có vai trò quan trọng, giúp câu văn dịch sẽ trôi chảy và đúng hơn. Trong khí đó, việc học ngữ pháp thường được học tự động trong các mô hình học sâu dẫn tới cấu trúc ngữ pháp không được khai thác hết thông tin. Do đó mục đích của sáng chế là tăng cường thông tin ngữ pháp trong mô hình dịch máy.

Việc sử dụng phân tích cú pháp vào trong dịch máy là một hướng đi được nhiều nhà nghiên cứu tập trung khai thác. Phân tích cú pháp trong xử lý ngôn ngữ tự nhiên được chia

làm hai loại là phân tích cú pháp thành phần (constituency parsing) và phân tích cú pháp phụ thuộc (dependency parsing). Trong dữ liệu thực tế, phân tích cú pháp thành phần khó có thể áp dụng được vì việc xây dựng dữ liệu huấn luyện cây cú pháp thành phần rất phức tạp. Ngược lại, cây cú pháp phụ thuộc lại được áp dụng rộng rãi trong cộng đồng bởi việc xây dựng dữ liệu huấn luyện đơn giản hơn so với cây cú pháp thành phần. Do đó, sáng chế sử dụng việc phân tích cú pháp phụ thuộc để đưa vào mô hình đề xuất.

Bên cạnh việc sử dụng các công cụ phân tích cú pháp, sáng chế còn đề xuất thêm mô-đun học ngữ pháp tự động. Các từ sẽ đi kèm với một véc-tơ đại diện cho vai trò ngữ pháp của từ đó trong câu (chủ ngữ, vị ngữ, danh từ, động từ,...). Các véc-tơ đại diện này sẽ được kết nối với véc-tơ của từ qua phép toán nhân. Điều này đảm bảo cho mô hình đề xuất không bị phụ thuộc nhiều bởi các công cụ phân tích cú pháp.

Tóm lại, sáng chế đưa ra mô hình củng cố thêm việc học thông tin ngữ pháp trong câu bằng việc đưa vào thông tin ngữ pháp của câu đó thông qua các công cụ phân tích cú pháp phụ thuộc. Thông tin này sẽ được đi kèm với câu nguồn trong tập huấn luyện. Phương pháp sẽ được thiết kế kèm các véc-tơ mới để học thông tin ngữ pháp của câu đích, được khởi tạo ngẫu nhiên trong phương pháp. Tại mỗi tầng, phương pháp sẽ tính trọng số kết nối giữa véc-tơ ngữ pháp và véc-tơ từ. Sau đó, phương pháp sẽ sử dụng phép toán nhân để kết nối thông tin từ và thông tin ngữ pháp trên từng đơn vị từ của câu văn. Để đạt được mục đích trên, sáng chế bao gồm các bước:

Bước 1: thu thập dữ liệu dịch máy. Dữ liệu thu thập trên các trang tin tức dưới dạng văn bản theo nhiều loại ngôn ngữ và là bản dịch của nhau.

Bước 2: tách câu trong văn bản. Xác định ranh giới các câu trong một văn bản, làm đầu vào cho quá trình huấn luyện mô hình.

Bước 3: xây dựng thuật toán căn chỉnh câu văn giữa các văn bản. Giúp xác định chính xác các câu văn là dịch của nhau giữa các văn bản.

Bước 4: phân tích ngữ pháp của câu văn. Đưa ra thông tin mối quan hệ ngữ pháp giữa các từ trong câu văn, làm đầu vào cho mô hình dịch máy.

Bước 5: xây dựng mô hình dịch máy sử dụng thông tin ngữ pháp. Mô hình học đầy đủ thông tin ngữ pháp của câu văn đầu vào và câu văn dịch đầu ra.

Mô tả vắn tắt hình vẽ

Hình 1 là hình vẽ mô tả các bước thực hiện của phương pháp.

Hình 2 là hình vẽ mô tả chi tiết mô hình dịch máy của sáng chế.

Mô tả chi tiết sáng chế

Sáng chế được mô tả chi tiết dưới đây có dựa vào hình vẽ kèm theo, hình vẽ này nhằm mục đích minh họa các phương án của sáng chế mà không làm giới hạn phạm vi bảo hộ sáng chế.

Cụ thể, Hình 1 mô tả các bước xử lý khi văn bản được đưa vào như sau: bước 1: thu thập dữ liệu dịch máy; bước 2: tách câu trong văn bản; bước 3: xây dựng thuật toán căn chỉnh câu văn giữa các văn bản; bước 4: phân tích ngữ pháp của câu văn; bước 5: xây dựng mô hình dịch máy sử dụng thông tin ngữ pháp. Hình 2 mô tả cụ thể mô hình dịch máy trong sáng chế.

Cụ thể hơn, phương pháp dịch máy tiếng Việt sử dụng thông tin ngữ pháp được thực hiện bằng hệ thống máy vi tính bao gồm các bước:

Bước 1: thu thập dữ liệu dịch máy.

Dữ liệu huấn luyện mô hình là vấn đề then chốt để tìm lời giải cho các bài toán trong lĩnh vực xử lý ngôn ngữ tự nhiên nói chung và nhiệm vụ dịch máy nói riêng. Tuy nhiên, việc gán nhãn dữ liệu mất nhiều công sức và tốn kém về chi phí. Hơn thế nữa, nguồn dữ liệu song ngữ từ các trang tin tức rất phong phú. Do đó, tại bước này, tập trung thu thập các bài báo được công bố từ các trang tin tức uy tín có chứa ngôn ngữ tiếng Việt và các ngôn ngữ khác.

Bước 2: tách câu trong văn bản.

Các văn bản thu thập được sẽ được đưa vào giai đoạn tách câu. Các câu văn sau khi được xử lý sẽ làm dữ liệu huấn luyện cho mô hình dịch máy. Bộ tách câu là một phương pháp dựa vào các đặc điểm ngôn ngữ để xác định ranh giới các câu trong văn bản. Việc tách câu được xác định bằng giả định rằng nếu câu kết thúc bằng các dấu ".", ";", "!", "?", "..." và đằng sau có ký tự viết hoa và không phải dấu ngoặc thì đây chính là dấu hiệu để nhận biết vị trí kết thúc của một câu.

Bước 3: xây dựng thuật toán căn chỉnh câu văn giữa các văn bản.

Sau khi thu thập được các văn bản là dịch của nhau (ở bước 1) và thực hiện việc tách câu (ở bước 2), tiến hành việc xác định các câu văn là dịch của nhau giữa các cặp văn bản. Nguyên nhân là các cặp văn bản dịch của nhau thường sẽ không cùng số lượng các câu văn. Có thể có trường hợp văn bản dịch bỏ đi một số câu trong văn bản gốc, hoặc văn bản dịch tách một câu trong văn bản gốc thành hai hoặc ba câu nhỏ. Do đó, bước này tập trung vào việc xây dựng một thuật toán để tìm các câu dịch của nhau giữa các cặp văn bản. Cụ thể, các câu trong mỗi văn bản được đưa qua mô hình ngôn ngữ lớn để tạo ra véc-tơ câu nhúng. Sau đó, các công cụ truy hồi thông tin được sử dụng để với một câu văn trong văn bản gốc đưa ra một hoặc nhiều câu văn trong văn bản dịch có độ tương tự gần nhất. Các cặp câu văn tương tự này sẽ được sử dụng làm dữ liệu huấn luyện cho mô hình dịch máy.

Bước 4: phân tích ngữ pháp của câu văn.

Các câu văn trước khi đưa vào huấn luyện mô hình sẽ được phân tích cấu trúc ngữ pháp theo cú pháp phụ thuộc. Cụ thể, cú pháp phụ thuộc của một câu văn sẽ đưa ra thông tin mối quan hệ giữa các từ trong một câu văn. Mối quan hệ này được xác định bằng thông tin từ hiện tại, từ phụ thuộc và kiểu quan hệ của hai từ đó. Mỗi loại ngôn ngữ đều có một cách xác định thông tin cú pháp phụ thuộc riêng, thể hiện ở việc tách từ và kiểu quan hệ giữa các từ. Đầu tiên, các câu văn được tách từ, tức là tìm ranh giới của các từ trong câu văn. Việc tách từ sử dụng mô hình học máy LM-LSTM-CRF. Sau đó, các câu văn đã tách từ này sẽ được phân tích cú pháp phụ thuộc nhờ mô hình chú ý cặp quan hệ họ hàng sâu (Deep Biaffine Attention), để đưa ra mối quan hệ giữa các từ trong câu văn.

Bước 5: xây dựng mô hình dịch máy sử dụng thông tin ngữ pháp.

Tại bước này, các cặp câu văn dịch đầu vào và đầu ra (ở bước 3) được đưa vào mô hình có kiến trúc với pha mã hóa (Encoder) và pha giải mã (Decoder). Pha mã hóa học thông tin cú pháp phụ thuộc (ở bước 4) trong câu văn đầu vào. Mỗi từ trong pha mã hóa sẽ đi kèm thông tin từ phụ thuộc và thông tin kiểu quan hệ với từ phụ thuộc này. Hai thông tin này sẽ được thể hiện qua các ma trận từ và ma trận kiểu quan hệ để mô hình hóa thông tin câu văn. Các ma trận này được cập nhật trong quá trình huấn luyện. Pha giải mã học thông tin ngữ pháp (ở bước 4) của câu văn đầu ra. Thông tin ngữ pháp được thể hiện qua các véc-tơ ngữ pháp được tạc ra ngẫu nhiên khi tạo mô hình, đại diện cho vai trò ngữ pháp của các từ trong câu văn, như chủ ngữ, vị ngữ,... Các véc-tơ ngữ pháp này sẽ được cập nhật trong

quá trình huấn luyện. Do đó, mô hình dịch máy sẽ học được đầy đủ thông tin ngữ pháp, ở cả câu đầu vào và câu đầu ra trong bộ dữ liệu huấn luyện.

Ví dụ thực hiện sáng chế

Giải pháp đã được đưa vào thử nghiệm với bộ dữ liệu VLSP 2020 dịch tiếng Anh sang tiếng Việt với 20000 câu văn để huấn luyện, 790 câu văn được sử dụng để kiểm thử và bộ dữ liệu VLSP 2022 dịch tiếng Trung sang tiếng Việt với 300000 câu văn để huấn luyện và 1000 câu văn để kiểm thử.

Về phương pháp đánh giá hiệu quả, sử dụng thông số BLEU. Đây là độ đo phổ biến cho các bài toán về dịch máy.

Ngoài ra, tiến hành thực nghiệm phương pháp thường được sử dụng cho vấn đề dịch máy là mô hình người máy (Transformer) để so sánh kết quả với phương pháp đề xuất.

Kết quả đánh giá bộ dữ liệu VLSP 2020 từ tiếng Anh sang tiếng Việt được đo lại như sau:

Phương Pháp	BLEU
Mô hình người máy	23,9
Phương pháp đề xuất	25,4

Kết quả đánh giá bộ dữ liệu VLSP 2022 từ tiếng Trung sang tiếng Việt được đo lại như sau:

Phương Pháp	BLEU
Mô hình người máy	21,5
Phương pháp đề xuất	23,0

Kết quả cho thấy, sáng chế đưa ra cho kết quả vượt trội so với các phương pháp sử dụng mô hình người máy.

Hiệu quả đạt được của sáng chế

Ưu điểm đặc biệt liên quan đến sáng chế này đó là xây dựng một phương pháp cho bài toán dịch máy bằng cách sử dụng thông tin ngữ pháp.

Mặc dù các mô tả nêu trên chứa nhiều chi tiết cụ thể, chúng không được coi là giới hạn về phương án thực thi sáng chế, mà chỉ nhằm mục đích minh họa một số phương án thực thi được ưu tiên.

YÊU CẦU BẢO HỘ

1. Phương pháp xây dựng mô hình dịch máy sử dụng thông tin ngữ pháp bao gồm:

bước 1: thu thập dữ liệu dịch máy;

tại bước này, tập trung thu thập các bài báo được công bố từ các trang tin tức uy tín có chứa ngôn ngữ tiếng Việt và các ngôn ngữ khác bằng hệ thống máy vi tính;

bước 2: tách câu trong văn bản;

các văn bản thu thập được sẽ được đưa vào giai đoạn tách câu; các câu văn sau khi được xử lý sẽ làm dữ liệu huấn luyện cho mô hình dịch máy; bộ tách câu là một phương pháp dựa vào các đặc điểm ngôn ngữ để xác định ranh giới các câu trong văn bản; việc tách câu được xác định bằng giả định rằng nếu câu kết thúc bằng các dấu ".", ";", "!", "?", "..." và đằng sau có ký tự viết hoa và không phải dấu ngoặc thì đây chính là dấu hiệu để nhận biết vị trí kết thúc của một câu;

bước 3: xây dựng thuật toán căn chỉnh câu văn giữa các văn bản;

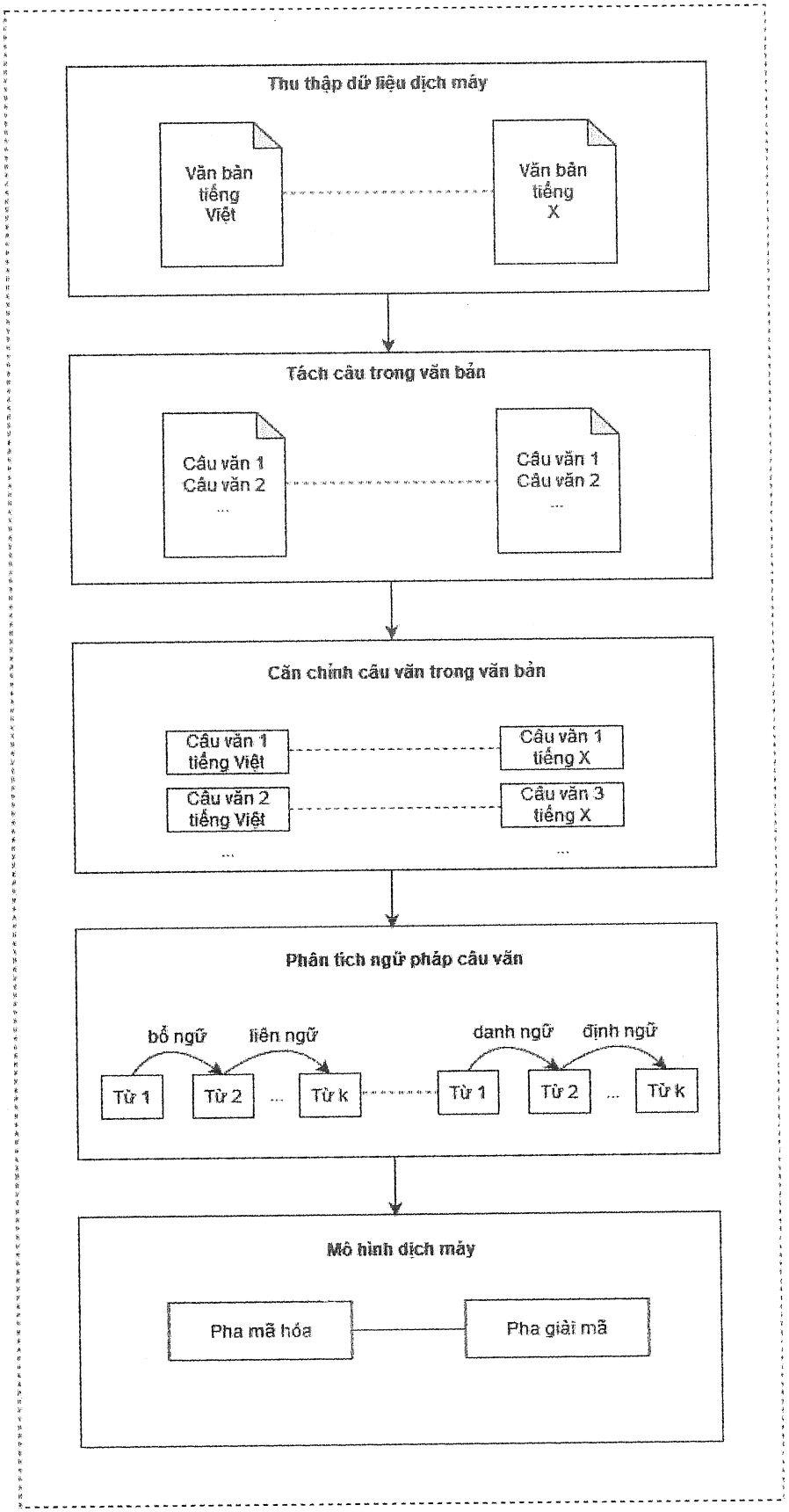
sau khi thu thập được các văn bản là dịch của nhau (ở bước 1) và thực hiện việc tách câu (ở bước 2), tiến hành việc xác định các câu văn là dịch của nhau giữa các cặp văn bản; cụ thể, các câu trong mỗi văn bản được đưa qua mô hình ngôn ngữ lớn để tạo ra véc-tơ câu nhúng; sau đó, các công cụ truy hồi thông tin được sử dụng để với một câu văn trong văn bản gốc đưa ra một hoặc nhiều câu văn trong văn bản dịch có độ tương tự gần nhất; các cặp câu văn tương tự này sẽ được sử dụng làm dữ liệu huấn luyện cho mô hình dịch máy;

bước 4: phân tích ngữ pháp của câu văn;

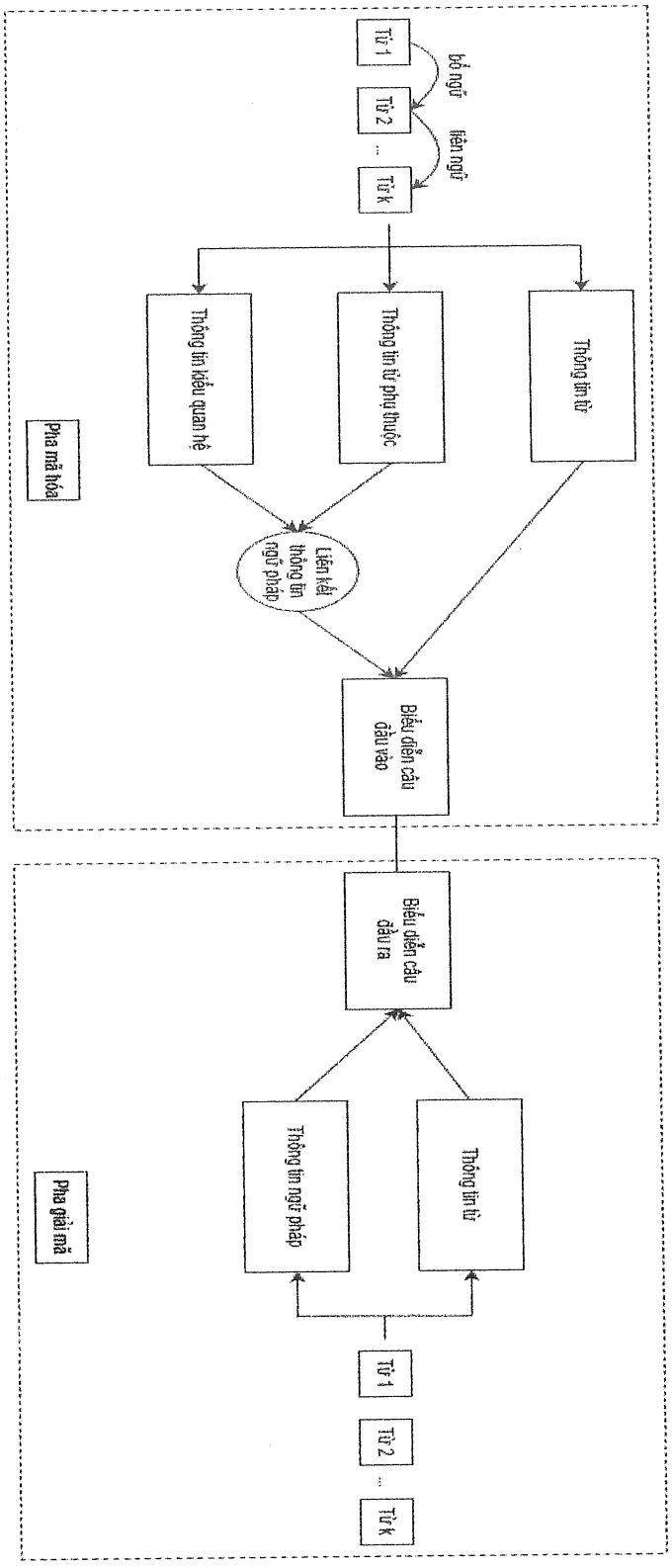
các câu văn trước khi đưa vào huấn luyện mô hình sẽ được phân tích cấu trúc ngữ pháp theo cú pháp phụ thuộc; cụ thể, cú pháp phụ thuộc của một câu văn sẽ đưa ra thông tin mối quan hệ giữa các từ trong một câu văn; mối quan hệ này được xác định bằng thông tin từ hiện tại, từ phụ thuộc và kiểu quan hệ của hai từ đó; mỗi loại ngôn ngữ đều có một cách xác định thông tin cú pháp phụ thuộc riêng, thể hiện ở việc tách từ và kiểu quan hệ giữa các từ; đầu tiên, các câu văn được tách từ, tức là tìm ranh giới của các từ trong câu văn; việc tách từ sử dụng mô hình học máy LM-LSTM-CRF; sau đó, các câu văn đã tách từ này sẽ được phân tích cú pháp phụ thuộc nhờ mô hình chú ý cặp quan hệ họ hàng sâu (deep biaffine attention), để đưa ra mối quan hệ giữa các từ trong câu văn;

bước 5: xây dựng mô hình dịch máy sử dụng thông tin ngữ pháp;

tại bước này, các cặp câu văn dịch đầu vào và đầu ra (ở bước 3) được đưa vào mô hình có kiến trúc với pha mã hóa (encoder) và pha giải mã (decoder); pha mã hóa học thông tin cú pháp phụ thuộc (ở bước 4) trong câu văn đầu vào; mỗi từ trong pha mã hóa sẽ đi kèm thông tin từ phụ thuộc và thông tin kiểu quan hệ với từ phụ thuộc này; hai thông tin này sẽ được thể hiện qua các ma trận từ và ma trận kiểu quan hệ để mô hình hóa thông tin câu văn; các ma trận này được cập nhật trong quá trình huấn luyện; pha giải mã học thông tin ngữ pháp (ở bước 4) của câu văn đầu ra; thông tin ngữ pháp được thể hiện qua các véc-tơ ngữ pháp được tạo ra ngẫu nhiên khi tạo mô hình, đại diện cho vai trò ngữ pháp của các từ trong câu văn, như chủ ngữ, vị ngữ; các véc-tơ ngữ pháp này sẽ được cập nhật trong quá trình huấn luyện; do đó, mô hình dịch máy sẽ học được đầy đủ thông tin ngữ pháp, ở cả câu đầu vào và câu đầu ra trong bộ dữ liệu huấn luyện.


Thu thập dữ liệu dịch máy
Văn bản tiếng Việt
Văn bản tiếng X
Tách câu trong văn bản
Câu văn 1
Câu văn 2
...
Câu văn 1
Câu văn 2
...
Căn chỉnh câu văn trong văn bản
Câu văn 1 tiếng Việt
Câu văn 1 tiếng X
Câu văn 2 tiếng Việt
Câu văn 3 tiếng X
...
...
Phân tích ngữ pháp câu văn
bổ ngữ
liên ngữ
danh ngữ
định ngữ
Từ 1
Từ 2
...
Từ k
Từ 1
Từ 2
...
Từ k
Mô hình dịch máy
Pha mã hóa
Pha giải mã

Hình 1


Từ 1
Từ 2
...
Từ k
bó ngữ
liên ngữ
Thông tin kiểu quan hệ
Thông tin từ phụ thuộc
Thông tin từ
Pha mã hóa
Liên kết thông tin ngữ pháp
Biểu diễn câu đầu vào
Biểu diễn câu đầu ra
Thông tin ngữ pháp
Thông tin từ
Pha giải mã
Từ 1
Từ 2
...
Từ k

Hình 2