(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) **1-0052679**

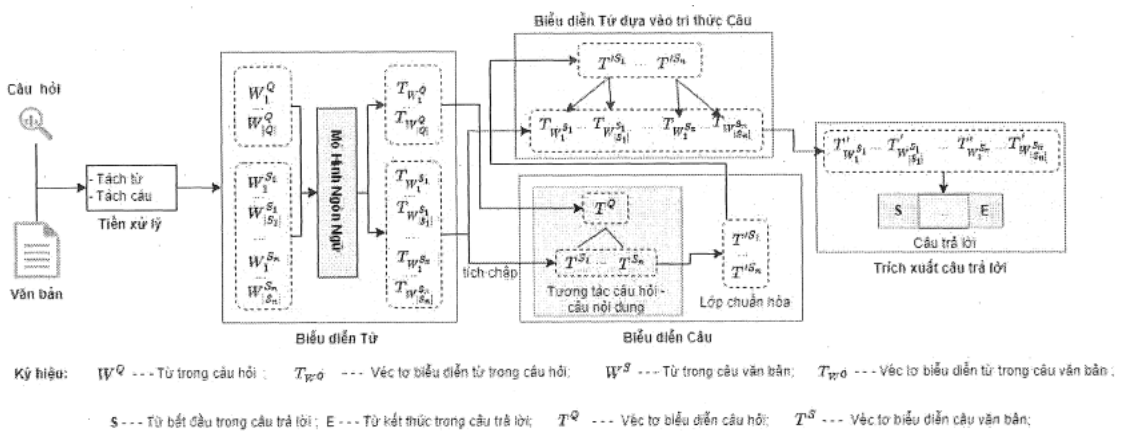
(51)[2022.01] **G06F 40/20**; G06N 3/00; G06F 11/00 (13) **B**

(21) 1-2023-03989 (22) 16/06/2023
(45) 27/10/2025 451 (43) 25/09/2023 426A
(73) TẬP ĐOÀN CÔNG NGHIỆP - VIỄN THÔNG QUÂN ĐỘI (VN)
Lô D26 Khu đô thị mới Cầu Giấy, phường Yên Hòa, quận Cầu Giấy, thành phố Hà Nội
(72) BÙI KHẮC HOÀI NAM (VN); PHAN TUẤN ANH (VN); NGUYỄN THỊ MINH ÁNH (VN); NGUYỄN THẾ LÂM (VN); BÙI CHÍ MINH (VN); NGUYỄN NGỌC DŨNG (VN).
(74) Công ty TNHH NACILAW (NACILAW)

(54) PHƯƠNG PHÁP TRÍCH XUẤT NỘI DUNG TRẢ LỜI BẰNG PHƯƠNG PHÁP ĐỌC HIỂU DỰA TRÊN TRI THỨC CÂU CHO VĂN BẢN PHÁP LUẬT VIỆT NAM

(21) 1-2023-03989

(57) Sáng chế đề xuất phương pháp mới cho hệ thống hỏi đáp sử dụng phương pháp đọc hiểu, tập trung cho văn bản pháp luật Việt Nam. Cụ thể, có hai vấn đề chính trong việc xây dựng hệ thống hỏi đáp cho văn bản pháp luật Việt Nam, bao gồm: i) tài nguyên dữ liệu ít cho việc áp dụng các mô hình ngôn ngữ để biểu diễn văn bản; ii) các mô hình hiện đại cho phần đọc hiểu chỉ tập trung vào việc biểu diễn Từ trong văn bản, tận dụng tri thức từ các đơn vị ngữ nghĩa khác trong văn bản (ví dụ Câu) chưa được khai thác. Do đó, sáng chế tập trung vào việc cung cấp một bộ dữ liệu mẫu để huấn luyện mô hình cho hệ thống hỏi đáp, tập trung vào văn bản pháp luật Việt Nam. Ngoài ra, sáng chế đề xuất mô hình mạng nơ-ron mới để biểu diễn nội dung văn bản, bằng cách bổ sung thông tin từ tri thức của Câu cho việc biểu Từ, từ đó trích xuất ra cặp Từ (bắt đầu – kết thúc) với độ chính xác cao hơn.



Hình 2

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp trích xuất nội dung trả lời bằng phương pháp đọc hiểu dựa trên tri thức câu cho văn bản pháp luật Việt Nam. Cụ thể, phương pháp được đề cập trong sáng chế cải thiện tính hiệu quả khi sử dụng phương pháp đọc hiểu (machine reading comprehension, MRC) cho hệ thống hỏi đáp trong việc xử lý ngôn ngữ tự nhiên.

Tình trạng kỹ thuật của sáng chế

Hỏi đáp tự động là một trong những bài toán nâng cao trong NLP. Trong đó, hướng tiếp cận dựa trên phương pháp đọc hiểu (machine reading comprehension, MRC) hiện đang được nghiên cứu và phát triển mạnh mẽ trong những năm gần đây. Phương pháp này dựa vào cơ chế tập trung (attention mechanism) giữa câu hỏi và nội dung văn bản cho trước để trích xuất cụm từ trong đoạn văn bản là câu trả lời.

Hiện nay, các mô hình tiên tiến trên thế giới cho hệ thống hỏi đáp sử dụng phương pháp đọc hiểu, tập trung vào tận dụng tri thức của các mô hình ngôn ngữ (language models) như BERT và RoBERTa để biểu diễn véc tơ của từ (Encoder). Hình 1 biểu diễn hệ thống MRC cơ bản dựa vào kiến trúc mô hình ngôn ngữ để trích xuất nội dung câu trả lời cho bài toán hỏi-đáp. Cụ thể, cách tiếp cận này đã đạt được hiệu quả trên hầu hết các tập dữ liệu mở về bài toán hỏi đáp như SQUAD, NewQA, và Natural Question. Tuy nhiên, hầu hết các phương pháp hiện tại tập trung vào các câu hỏi về nội dung kiến thức chung (open domain), dữ liệu được thu thập chủ yếu từ Wikimedia. Do đó, việc xây dựng mô hình cho những bộ dữ liệu cụ thể theo từng chủ đề cụ thể (specific domain) vẫn đang là nghiên cứu mở trong lĩnh vực này. Hơn nữa, việc áp dụng các phương pháp nêu trên vào văn bản tiếng Việt vẫn còn đang gặp nhiều khó khăn, nguyên nhân chính là vì việc hạn chế về tài nguyên ngôn ngữ (low resource language).

Ngoài ra, các phương pháp hiện tại vẫn chỉ đang tập trung vào việc sử dụng mô hình ngôn ngữ để biểu diễn nội dung văn bản và câu hỏi ở mức từ (word-level). Việc tận dụng tri thức từ các đơn vị ngữ nghĩa khác trong văn bản (ví dụ như mức câu (sentence-level)), vẫn

chưa được khai thác. Do đó, dựa vào quan sát rằng, nội dung câu trả lời thường được trích xuất từ một câu trong văn bản, sáng chế này đưa ra đề xuất tận dụng thêm tri thức của các câu trong văn bản để bổ sung ngữ nghĩa cho từ, trước khi trích xuất ra nội dung câu trả lời.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất giải pháp cho bài toán hỏi đáp văn bản pháp luật Việt Nam. Cụ thể, sáng chế tập trung vào hai vấn đề chính, bao gồm: chuẩn bị dữ liệu văn bản pháp luật Việt Nam cho bài toán hỏi đáp, đưa ra giải pháp đọc hiểu mới, từ đó cải thiện tính hiệu quả trong việc trích xuất nội dung câu trả lời.

Cụ thể, sáng chế này cung cấp bộ dữ liệu mẫu về văn bản pháp luật Việt Nam, phục vụ cho bài toán hỏi đáp. Đây là bộ dữ liệu tiếng Việt đầu tiên về văn bản pháp luật Việt Nam, sử dụng cho bài toán hỏi đáp bằng phương pháp đọc hiểu.

Ngoài ra, sáng chế đề xuất mô hình mới để giải quyết bài toán MRC sử dụng thêm thông tin tri thức của câu trong văn bản, với ý tưởng chính dựa vào quan sát là nội dung câu trả lời luôn nằm trong phạm vi một câu. Do đó, mô hình có thể tận dụng thông tin của câu để bổ sung cho thông tin của từ, nâng cao khả năng xác định chính xác được từ đầu tiên (start point) và từ cuối cùng (end point), là cơ sở để trích xuất nội dung câu trả lời. Để đạt được mục đích trên, sáng chế bao gồm các bước:

Bước 1: thu thập dữ liệu văn bản pháp luật Việt Nam;

Bước 2: tách điều trong văn bản pháp luật Việt Nam;

Bước 3: xây dựng bộ dữ liệu về văn bản pháp luật Việt Nam sử dụng phương pháp đọc hiểu cho bài toán hỏi-đáp;

Bước 4: xây dựng mô hình hỏi đáp bằng phương pháp đọc hiểu sử dụng tri thức của câu trong văn bản;

Bước 5: trích xuất nội dung câu trả lời.

Mô tả vắn tắt các hình vẽ

Hình 1 là hình vẽ mô tả các bước thực hiện của mô hình tiêu chuẩn cho bài toán hỏi đáp sử dụng phương pháp MRC;

Hình 2 là hình vẽ mô tả các bước thực hiện của mô hình đề xuất cho bài toán hỏi đáp

sử dụng phương pháp MRC.

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 2 mô tả các bước xử lý khi câu hỏi và văn bản được đưa vào như sau: i) tiền xử lý văn bản, sử dụng tách câu và tách từ trong câu hỏi và văn bản; ii) tập các từ trong câu hỏi và câu trả lời sau khi tách sẽ được biểu diễn dưới dạng véc-tơ bằng mô hình ngôn ngữ; iii) câu hỏi và các câu trong văn bản sau đó được biểu diễn sử dụng phương pháp tích chập trung bình (average pooling) dựa vào tập các từ thuộc câu đó; iv) các véc-tơ biểu diễn các câu văn bản sau đấy được bổ sung thông tin của câu hỏi bằng sử dụng cơ chế tập trung (attention mechanism), với mục tiêu đưa ra những biểu diễn mới cho các câu văn bản theo nội dung của câu hỏi. Theo đó, các câu văn bản có trọng số tập trung (attention score) cao, biểu diễn các câu đó sẽ được bổ sung nhiều thông tin của câu hỏi; iv) biểu diễn mới của câu văn bản sẽ được đưa vào bổ sung cho biểu diễn từ sử dụng các phương pháp tích hợp hai véc-tơ, mục tiêu là đưa ra biểu diễn mới của các từ, được bổ sung thông tin của câu hỏi; v) các biểu diễn từ sau đó được tính toán để trích xuất ra kết quả là các cặp từ bao gồm từ bắt đầu và từ kết thúc; vi) cặp từ thỏa mãn một số yêu cầu với kết quả cao nhất được chọn để trích xuất nội dung câu trả lời.

Cụ thể hơn, phương pháp trích xuất nội dung trả lời bằng phương pháp đọc hiểu dựa trên tri thức câu cho văn bản pháp luật Việt Nam bao gồm các bước:

Bước 1: thu thập dữ liệu văn bản pháp luật Việt Nam;

Tại bước này, tập trung vào thu thập dữ liệu là các văn bản về pháp luật Việt Nam. Việc thu thập được tiến hành tự động nhờ liên kết với nguồn dữ liệu từ các trang công bố văn bản pháp luật của quốc hội, chính thủ, bộ tư pháp.... được công bố trên cơ sở dữ liệu quốc gia về văn bản pháp luật. Cụ thể, khoảng 550 văn bản pháp luật được thu thập từ bước này.

Bước 2: tách điều trong văn bản pháp luật Việt Nam;

Tại bước này, 550 văn bản được xử lý tách thành các đoạn văn, nội dung mỗi đoạn văn là các điều luật trong văn bản. Việc tách điều được xác định bằng giả định rằng một điều được

bắt đầu với từ "Điều", theo sau là các con số chỉ mục số điều và dấu '.' , đây chính là dấu hiệu để nhận biết vị trí đầu tiên của điều. Bước này được xử lý tự động bằng ngôn ngữ lập trình Python, được lập trình bởi nhóm tác giả sáng chế. Cụ thể, khoảng 6000 đoạn văn bản là các điều luật được tạo ra từ bước này.

Bước 3: xây dựng bộ dữ liệu về văn bản pháp luật Việt Nam cho bài toán hỏi-đáp sử dụng phương pháp đọc hiểu;

Đầu vào cho bài toán MRC bao gồm câu hỏi và đoạn văn bản, đầu ra là nội dung câu trả lời cho câu hỏi, được trích xuất trong nội dung đoạn văn bản. Do đó, tại bước này, nhóm tác giả sáng chế xây dựng bộ dữ liệu VTCC-ViLegalDoc (VTCC-Vietnamese Legal Document), với khoảng 10000 mẫu dữ liệu. Mỗi mẫu dữ liệu bao gồm cặp câu hỏi - câu trả lời, được tạo ra từ nội dung của các đoạn văn bản là các điều luật của văn bản pháp luật Việt Nam. Cụ thể, bộ dữ liệu VTCC-ViLegalDoc được xây dựng để huấn luyện mô hình đề xuất trong sáng chế. Đây là bộ dữ liệu văn bản pháp luật Việt Nam đầu tiên được xây dựng cho bài toán MRC, với các đặc điểm sau:

- Ứng với mỗi đoạn văn bản, tiến hành gán nhãn từ một đến hai câu hỏi. Nội dung câu hỏi được tạo ra dựa vào nội dung trong đoạn văn;

- Nội dung câu trả lời cho mỗi câu hỏi là đoạn từ (text span) được trích xuất từ nội dung đoạn văn bản dùng để đặt câu hỏi;

- Việc tiến hành kiểm tra lỗi cặp câu hỏi – câu trả lời được tiến hành theo hai bước: i) bước 1 người gán nhãn tự kiểm tra lại (self-checking); ii) bước 2 tiến hành kiểm tra chéo giữa các thành viên tham gia gán nhãn (cross-checking). Một cặp câu hỏi – câu trả lời được xác định là lỗi nếu bị rơi vào một trong ba trường hợp sau: i) sai chính tả; ii) câu hỏi không rõ ràng; iii) câu trả lời không đúng với câu hỏi;

- Có khoảng 10 người tham gia quá trình gán nhãn cho bộ dữ liệu VTCC-ViLegalDoc, được thực hiện trong vòng 02 tháng;

- Bộ dữ liệu VTCC-ViLegalDoc bao gồm khoảng 10000 cặp câu hỏi – câu trả lời.

Bước 4: xây dựng mô hình hỏi đáp bằng phương pháp đọc hiểu sử dụng tri thức của câu trong văn bản.

Bộ dữ liệu VTCC-ViLegalDoc được sử dụng để đưa vào huấn luyện mô hình đề xuất

trong sáng chế. Cụ thể, 90% được sử dụng để huấn luyện mô hình (train dataset) và 10% còn lại được sử dụng để đánh giá chất lượng của giải pháp (test dataset). Ý tưởng chính của phương pháp đề xuất là sử dụng thông tin của câu để bổ sung thông tin cho từ trước khi trích xuất cặp từ bắt đầu và kết thúc, làm đầu ra cho mô hình. Cụ thể, mô hình đề xuất dựa vào kiến trúc mô hình ngôn ngữ để biểu diễn thông tin ở mức từ (word-level). Biểu diễn thông tin các từ sau đấy được sử dụng để biểu diễn thông tin của mức câu (sentence-level). Dựa vào biểu diễn thông tin của mức câu (sentence-level) để cập nhập lại biểu diễn của mức từ (word-level). Bằng cách này, mô hình đề xuất có khả năng học được biểu diễn từ nằm trong câu chứa nội dung trả lời, sẽ mang nhiều thông tin biểu diễn nội dung câu hỏi hơn các từ nằm ở các câu khác, từ đó sẽ trích xuất nội dung câu trả lời chính xác hơn.

Bước 5: trích xuất nội dung câu trả lời;

Tại bước này, đưa ra cặp từ bao gồm từ bắt đầu và từ kết thúc với điểm số (score) cao nhất. Số điểm của cặp từ này được tính bởi hai mươi từ có số điểm là từ bắt đầu cao nhất và hai mươi từ có số điểm là từ kết thúc cao nhất. Cụ thể hai mươi cặp từ lần lượt được tính toán dựa vào các nguyên tắc sau: i) điểm số của cặp từ được tính bắt tổng xác suất của từ bắt đầu và từ kết thúc; ii) nếu vị trí của từ đầu tiên và từ từ kết thúc đều bằng 0, thì câu trả lời là rỗng (tức là câu hỏi không có câu trả lời); iii) nếu vị trí của từ kết thúc đứng trước từ bắt đầu, thì bỏ qua không xem xét cặp từ này. Hệ thống sẽ lựa chọn cặp từ có điểm số cao nhất làm đầu ra của quá trình trích xuất câu trả lời. Kết thúc bước này sẽ thu được nội dung câu trả lời đáp ứng mục đích đề ra của sáng chế.

Ví dụ thực hiện sáng chế

Giải pháp đã được đưa vào thử nghiệm trên tập dữ liệu thử nghiệm (text dataset) VTCC-ViLegalDoc, với khoảng 1000 mẫu dữ liệu cặp câu hỏi-câu trả lời. Ngoài ra, để đánh giá khả năng của mô hình đề xuất trên bộ dữ liệu về kiến thức chung, bộ dữ liệu hỏi đáp tiếng Việt phổ biến nhất là UIT-ViQuAD, được gán nhãn từ 176 bài viết trên Wikipedia tiếng Việt; được sử dụng để thực hiện thí nghiệm.

Về thông số đánh giá hiệu quả, sử dụng hai thông số: điểm F1 (F1 score) và giá trị EM (exact matching), đây là những độ đo phổ biến cho các bài toán về tự động trích xuất nội dung câu trả lời trong bài toán hỏi đáp. Các thông số đánh giá trên được mô tả cụ thể như sau:

- F1: là độ đo thông dụng cho bài toán phân loại, và được sử dụng rộng rãi trong bài toán hỏi đáp. Độ đo này phù hợp trong trường hợp chúng ta đánh giá ngang bằng giữa độ đo lường tính "chuẩn xác" (Precision) và độ đo lường tính "hữu dụng" (Recall). Trong đó độ đo lường tính "chuẩn xác" được định nghĩa là tỉ lệ số từ mô hình chung trên tổng số từ mô hình dự đoán và độ đo lường tính "hữu dụng" là tỉ lệ số từ chung trên tổng số từ trên tập gán nhãn. Cụ thể, điểm F1 được tính toán như sau:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- EM: đối với mỗi cặp câu hỏi-trả lời, nếu ký tự của mô hình dự đoán giống với ký tự của câu trả lời (được gán nhãn) thì EM = 1, ngược lại EM = 0. Có thể thấy đây là chỉ số khá nghiêm ngặt, vì trong trường hợp mô hình dự đoán bị lệch một ký tự dẫn đến kết quả bằng 0. Cụ thể, công thức tính tỉ lệ EM (EMR) được tính như sau:

$$EM = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(\mathrm{Y}_i == \mathrm{Z}_i)$$

Trong đó, n là tổng số mẫu, Y_i là giá trị nhãn của mẫu i, Z_i là giá trị dự đoán nhãn của mẫu i.

Về phương pháp đánh giá so sánh, tiến hành thực nghiệm phương pháp phổ biến nhất ở thời điểm hiện tài cho bài toán MRC đối với tiếng Việt là sử dụng mô hình ngôn đa ngữ XLM-RoBERTa, để so sánh kết quả với phương pháp đề xuất.

Kết quả đánh giá việc tóm tắt văn bản được đo lại như sau:

Dữ liệu	Phương pháp	EM	F1
UIT-ViQuAD	XLM-RoBERTa	65,89	85,09
	Phương pháp đề xuất	67,48	86,17
VTCC-ViLegalDoc	XLM-RoBERTa	67,63	88,17
	Phương pháp đề xuất	69,84	89,77

Kết quả cho thấy, mô hình sáng chế đưa ra cho kết quả cao hơn so với mô hình phổ biến nhất hiện này của bài toán hỏi-đáp sử dụng phương pháp đọc hiểu.

Hiệu quả đạt được của sáng chế

Ưu điểm đặc biệt liên quan đến sáng chế này đó là xây dựng hệ thống đọc hiểu cho bài toán hỏi-đáp liên quan đến văn bản pháp luật Việt Nam. Cụ thể, các hệ thống hỏi-đáp hiện tại áp dụng phương pháp đọc hiểu đều sử dụng cho việc hỏi đáp kiến thức chung, dữ liệu được thu thập từ các trang Wikipedia tiếng Việt. Do đó, sáng chế này đưa ra bộ dữ liệu VTCC-ViLegalDoc, phục vụ cho vấn đề hỏi đáp liên quan đến văn bản pháp luật Việt Nam. Ngoài ra, sáng chế đề xuất mô hình mới cho bài toán hỏi đáp sử dụng phương pháp đọc hiểu bằng cách đào sâu khai thác mối quan hệ thông tin giữa các thành phần trong văn bản, từ đó làm giàu thông tin biểu diễn của văn bản với độ chính xác cao hơn. Cụ thể, với phương pháp đề xuất này, có thể trích xuất ra nội dung câu trả lời với độ chính xác cao hơn so với mô hình phổ biến nhất hiện tại trong lĩnh vực nghiên cứu này.

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 trích xuất nội dung trả lời bằng phương pháp đọc hiểu dựa trên tri thức câu cho văn bản pháp luật Việt Nam bao gồm các bước:

bước 1: thu thập dữ liệu văn bản pháp luật Việt Nam;

tại bước này, tập trung vào thu thập dữ liệu là các văn bản về pháp luật Việt Nam; việc thu thập được tiến hành tự động nhờ liên kết với nguồn dữ liệu từ các trang công bố văn bản pháp luật của quốc hội, chính thủ, bộ tư pháp…. được công bố trên cơ sở dữ liệu quốc gia về văn bản pháp luật; cụ thể, khoảng 550 văn bản pháp luật được thu thập từ bước này;

bước 2: tách điều trong văn bản pháp luật Việt Nam;

tại bước này, 550 văn bản được xử lý tách thành các đoạn văn, nội dung mỗi đoạn văn là các điều luật trong văn bản; việc tách điều được xác định bằng giả định rằng một điều được bắt đầu với từ "điều", theo sau là các con số chỉ mục số điều và dấu '.' , đây chính là dấu hiệu để nhận biết vị trí đầu tiên của điều; bước này được xử lý tự động bằng ngôn ngữ lập trình python, được lập trình bởi nhóm tác giả sáng chế; cụ thể, khoảng 6000 đoạn văn bản là các điều luật được tạo ra từ bước này;

bước 3: xây dựng bộ dữ liệu về văn bản pháp luật Việt Nam cho bài toán hỏi-đáp sử dụng phương pháp đọc hiểu;

đầu vào cho bài toán MRC bao gồm câu hỏi và đoạn văn bản, đầu ra là nội dung câu trả lời cho câu hỏi, được trích xuất trong nội dung đoạn văn bản; do đó, tại bước này, nhóm tác giả sáng chế xây dựng bộ dữ liệu với khoảng 10000 mẫu dữ liệu; mỗi mẫu dữ liệu bao gồm cặp câu hỏi - câu trả lời, được tạo ra từ nội dung của các đoạn văn bản là các điều luật của văn bản pháp luật Việt Nam; cụ thể, bộ dữ liệu được xây dựng để huấn luyện mô hình đề xuất trong sáng chế là bộ dữ liệu văn bản pháp luật Việt Nam đầu tiên được xây dựng cho bài toán MRC, với các đặc điểm sau:

ứng với mỗi đoạn văn bản, tiến hành gán nhãn từ một đến hai câu hỏi; nội dung câu hỏi được tạo ra dựa vào nội dung trong đoạn văn;

nội dung câu trả lời cho mỗi câu hỏi là đoạn từ (text span) được trích xuất từ nội dung đoạn văn bản dùng để đặt câu hỏi;

việc tiến hành kiểm tra lỗi cặp câu hỏi – câu trả lời được tiến hành theo

hai bước: i) bước thứ nhất: người gán nhãn tự kiểm tra lại (self-checking); ii) bước thứ hai: tiến hành kiểm tra chéo giữa các thành viên tham gia gán nhãn (cross-checking); một cặp câu hỏi – câu trả lời được xác định là lỗi nếu bị rơi vào một trong ba trường hợp sau: i) sai chính tả; ii) câu hỏi không rõ ràng; iii) câu trả lời không đúng với câu hỏi;

có khoảng mười người tham gia quá trình gán nhãn cho bộ dữ liệu và được thực hiện trong vòng hai tháng;

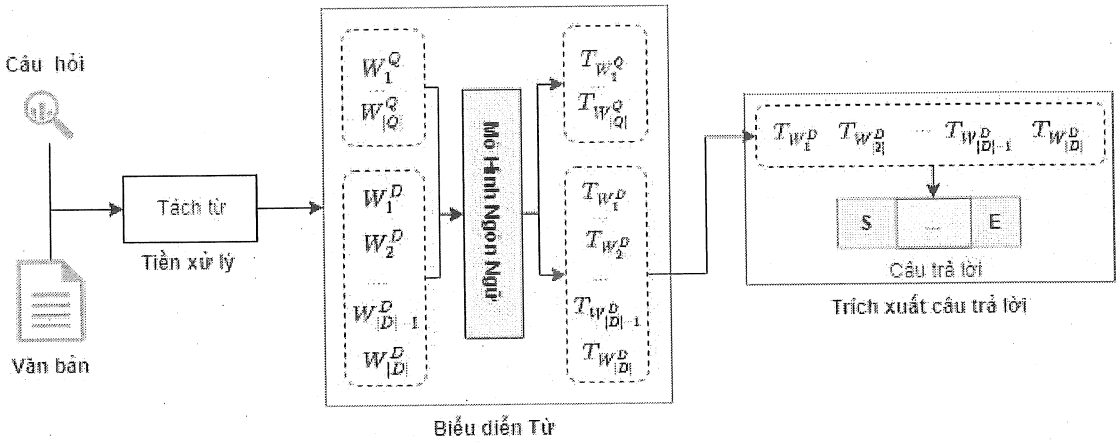
bộ dữ liệu bao gồm khoảng 10000 cặp câu hỏi – câu trả lời;

bước 4: xây dựng mô hình hỏi đáp bằng phương pháp đọc hiểu sử dụng tri thức của câu trong văn bản;

bộ dữ liệu được sử dụng để đưa vào huấn luyện mô hình đề xuất trong sáng chế; cụ thể, 90% được sử dụng để huấn luyện mô hình (train dataset) và 10% còn lại được sử dụng để đánh giá chất lượng của giải pháp (test dataset); cụ thể, mô hình đề xuất dựa vào kiến trúc mô hình ngôn ngữ để biểu diễn thông tin ở mức từ (word-level); biểu diễn thông tin các từ sau đấy được sử dụng để biểu diễn thông tin của mức câu (sentence-level); dựa vào biểu diễn thông tin của mức câu (sentence-level) để cập nhập lại biểu diễn của mức từ (word-level); bằng cách này, mô hình đề xuất có khả năng học được biểu diễn từ nằm trong câu chứa nội dung trả lời, sẽ mang nhiều thông tin biểu diễn nội dung câu hỏi hơn các từ nằm ở các câu khác, từ đó sẽ trích xuất nội dung câu trả lời chính xác hơn;

bước 5: trích xuất nội dung câu trả lời;

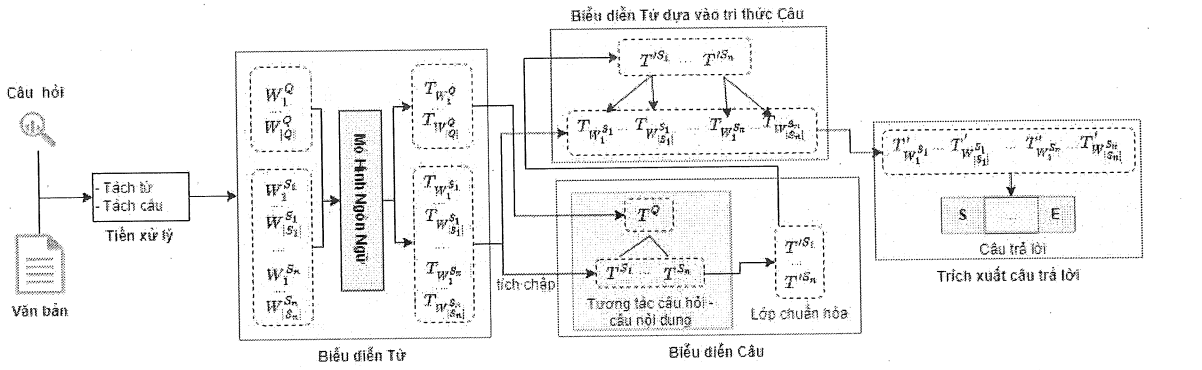
tại bước này, đưa ra cặp từ bao gồm từ bắt đầu và từ kết thúc với điểm số (score) cao nhất. Số điểm của cặp từ này được tính bởi hai mươi từ có số điểm là từ bắt đầu cao nhất và hai mươi từ có số điểm là từ kết thúc cao nhất; cụ thể hai mươi cặp từ lần lượt được tính toán dựa vào các nguyên tắc sau: i) điểm số của cặp từ được tính bắt tổng xác suất của từ bắt đầu và từ kết thúc; ii) nếu vị trí của từ đầu tiên và từ từ kết thúc đều bằng 0, thì câu trả lời là rỗng (tức là câu hỏi không có câu trả lời); iii) nếu vị trí của từ kết thúc đứng trước từ bắt đầu, thì bỏ qua không xem xét cặp từ này; hệ thống sẽ lựa chọn cặp từ có điểm số cao nhất làm đầu ra của quá trình trích xuất câu trả lời; kết thúc bước này sẽ thu được nội dung câu trả lời đáp ứng mục đích đề ra của sáng chế.



Ký hiệu: W^Q --- Từ trong câu hỏi ; T_{W^Q} --- Véc tơ biểu diễn từ trong câu hỏi; W^D --- Từ trong câu văn bản;

T_{W^D} --- Véc tơ biểu diễn từ trong câu văn bản ; S --- Từ bắt đầu trong câu trả lời ; E --- Từ kết thúc trong câu trả lời;

Hình 1



Ký hiệu: W^Q --- Từ trong câu hỏi ; T_{W^Q} --- Véc tơ biểu diễn từ trong câu hỏi; W^S --- Từ trong câu văn bản; T_{W^S} --- Véc tơ biểu diễn từ trong câu văn bản ;

S --- Từ bắt đầu trong câu trả lời ; E --- Từ kết thúc trong câu trả lời; T^Q --- Véc tơ biểu diễn câu hỏi; T^S --- Véc tơ biểu diễn câu văn bản;

Hình 2