XÂY DỰNG CÁC CẶP CÂU HỎI - CÂU TRẢ LỜI CHẤT LƯỢNG CAO TỪ CÁC TRANG WEB HỎI ĐÁP CỘNG ĐỒNG

Nguyễn Văn Tú[1], Lê Anh Cường[2], Nguyễn Hà Nam[3]
[1]Trường Đại học Tây Bắc
[2]Trường Đại học Tôn Đức Thắng
[3]Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội

***Tóm tắt*:** Các trang web hỏi đáp cộng đồng có chứa một lượng lớn thông tin hỏi - đáp có giá trị sinh ra bởi những người sử dụng. Trong các trang web hỏi đáp cộng đồng, người dùng có thể gửi các câu hỏi, trả lời các câu hỏi của người khác, và cung cấp thông tin phản hồi cho những câu hỏi/câu trả lời. Trong nghiên cứu này chúng tôi sử dụng tiếp cận học máy nhằm xây dựng các cặp câu hỏi - câu trả lời chất lượng cao từ các trang web hỏi đáp cộng đồng. Các cặp câu hỏi - câu trả lời này sẽ được sử dụng làm nguồn dữ liệu cho các hệ thống hỏi đáp tự động. Chúng tôi thực hiện trích rút những đặc trưng quan trọng từ mỗi luồng hỏi đáp cũng như thông tin của người gửi câu trả lời và xây dựng mô hình phân loại để xác định được các cặp câu hỏi - câu trả lời có ý nghĩa. Các kết quả thực nghiệm trên bộ dữ liệu cung cấp bởi SemEval 2015 cho thấy những đề xuất của chúng tôi sẽ mang lại kết quả cao.

Từ khóa: Hỏi đáp cộng đồng, phân loại, Support Vector Machines, hệ thống hỏi đáp tự động.

I. TỔNG QUAN

Trong lĩnh vực xử lý ngôn ngữ tự nhiên và truy xuất thông tin, vấn đề hỏi - đáp đã thu hút nhiều sự chú ý trong những năm qua. Tuy nhiên, các nghiên cứu về hỏi - đáp chủ yếu tập trung vào việc tìm câu trả lời chính xác cho câu hỏi factoid được đưa ra trong các tài liệu liên quan. Các đánh giá nổi tiếng nhất về nhiệm vụ hỏi - đáp factoid là hội nghị truy hồi văn bản (Text REtrieval Conference-TREC). Các câu hỏi và câu trả lời được phát hành bởi TREC đã trở thành nguồn dữ liệu quan trọng cho các nhà nghiên cứu trong việc nghiên cứu xây dựng các hệ thống hỏi đáp tự động [1]. Tuy nhiên, khi phải đối mặt với các câu hỏi non-factoid như các câu hỏi về lý do tại sao, như thế nào, hoặc những gì về… hầu như không có hệ thống hỏi đáp tự động nào làm việc tốt.

Các cặp câu hỏi - câu trả lời do người dùng tạo ra chắc chắn sẽ rất quan trọng để giải quyết vấn đề trả lời các câu hỏi non-factoid. Rõ ràng, những cặp câu hỏi - câu trả lời tự nhiên thường được tạo ra trong quá trình giao tiếp của con người thông qua phương tiện truyền thông xã hội Internet, trong đó chúng tôi đặc biệt quan tâm tới các trang web hỏi đáp dựa vào cộng đồng. Các trang web hỏi đáp dựa vào cộng đồng cung cấp nền tảng mà ở đó người dùng có thể đặt câu hỏi, cung cấp câu trả lời và các thông tin phản hồi (ví dụ, bằng cách biểu quyết hoặc cho ý kiến) cho những câu hỏi/câu trả lời và câu trả lời tốt nhất sẽ được lựa chọn để hiển thị cho người dùng.

Trong bài báo này, chúng tôi sử dụng tiếp cận học máy nhằm xây dựng các cặp câu hỏi - câu trả lời có chất lượng cao từ các dữ liệu hỏi đáp thu thập từ các trang web hỏi đáp cộng đồng. Các cặp câu hỏi - câu trả lời này có thể được sử dụng làm nguồn dữ liệu cho các hệ thống hỏi đáp tự động. Để xây dựng các cặp câu hỏi - câu trả lời chất lượng từ các trang web hỏi đáp cộng đồng, trong bài báo này

Tác giả liên hệ: Nguyễn Văn Tú
Email: tuspttb@gmail.com
Đến tòa soạn: 25/10/2016, chỉnh sửa: 28/12/2016, chấp nhận đăng: 1/1/2017

chúng tôi đề xuất sử dụng sự kết hợp của nhiều loại đặc trưng quan trọng trích rút từ mỗi luồng hỏi đáp cũng như thông tin của người gửi câu trả lời và xây dựng mô hình phân loại để xác định được các cặp câu hỏi - câu trả lời có ý nghĩa.

Để thực hiện những đề xuất của mình, chúng tôi đã sử dụng tập dữ liệu cung cấp bởi SemEval 2015 trong các thực nghiệm. Chúng tôi tiến hành đánh giá thử nghiệm rộng rãi để chứng minh tính hiệu quả của phương pháp tiếp cận của chúng tôi. Các kết quả thực nghiệm của chúng tôi đã cho thấy phương pháp tiếp cận mà chúng tôi đề xuất có thể xây dựng được bộ dữ liệu là các cặp câu hỏi - câu trả lời chất lượng cao để làm nguồn dữ liệu phục vụ cho các hệ thống hỏi đáp tự động.

Phần còn lại của bài báo này được tổ chức như sau: phần II trình bày về các nghiên cứu liên quan, phần III trình bày về bài toán xây dựng các cặp câu hỏi - câu trả lời chất lượng cao từ các trang web hỏi đáp cộng đồng, phần IV trình bày về thuật toán phân loại và các độ đo đánh giá hiệu suất của bộ phân loại, phần V chúng tôi thực hiện trích rút các đặc trưng quan trọng để phân loại câu hỏi, các thực nghiệm và kết quả được trình bày trong phần VI và cuối cùng là kết luận và hướng phát triển được trình bày trong phần VII.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Giá trị của các cặp câu hỏi - câu trả lời được sinh ra một cách tự nhiên đã không được những nhà nghiên cứu xây dựng hệ thống hỏi đáp tự động quan tâm cho đến tận những năm gần đây. Các nghiên cứu xây dựng hệ thống hỏi đáp ban đầu chủ yếu tập trung vào trích xuất các cặp câu hỏi - câu trả lời từ các câu hỏi được hỏi thường xuyên (FAQ) [2] hoặc dịch vụ đối thoại cuộc gọi trung tâm [3]. Các nghiên cứu gần đây đã tập trung khai thác nguồn thông tin hỏi đáp do người dùng cung cấp thông qua các trang web hỏi đáp cộng đồng. Bởi vì người dùng có quyền tự do trong việc gửi câu hỏi/câu trả lời trên các trang web hỏi đáp cộng đồng, cho nên có một số lượng lớn các câu trả lời không phù hợp hay liên quan cho các câu hỏi. Điều này là thực sự khó khăn để phát hiện các cặp câu hỏi - câu trả lời có ý nghĩa trong các trang web hỏi đáp cộng đồng.

Các nghiên cứu gần đây trong việc đánh giá chất lượng của các câu trả lời cung cấp bởi các trang web hỏi đáp cộng đồng thường thông qua các đặc trưng biểu diễn văn bản của câu hỏi - câu trả lời như là độ dài của câu hỏi, độ dài của câu trả lời, tỷ lệ độ dài giữa câu hỏi và các câu trả lời của nó, các độ đo tương tự giữa câu hỏi và câu trả lời [4, 5, 6]. Các đặc trưng thông dụng khác sử dụng trong phân tích chất lượng câu trả lời là sử dụng độ đo phổ biến và tương tác xã hội [4, 7] chẳng hạn như số lượng câu trả lời của người trả lời.

Tiếp cận khác là sử dụng sự kết hợp của các đặc trưng như các đặc trưng từ vựng, các đặc trưng cú pháp, thông tin người sử dụng [5]. Để nhận ra các câu trả lời chất lượng cao, Hu [8] học kết hợp sự biểu diễn cho mỗi cặp câu hỏi - câu trả lời bởi lấy cả các đặc trưng văn bản và phi văn bản như là đầu vào của mô hình. Surdeanu [9] đề xuất một cách tiếp cận khác để nhận ra các câu trả lời chất lượng cao là xếp hạng các câu trả lời lấy từ trang web hỏi đáp cộng đồng Yahoo!Answers và chọn các câu trả lời có thứ hạng cao như là các câu trả lời tốt nhất cho câu hỏi.

III. BÀI TOÁN XÂY DỰNG CÁC CẶP CÂU HỎI - CÂU TRẢ LỜI CHẤT LƯỢNG CAO TỪ CÁC TRANG WEB HỎI ĐÁP CỘNG ĐỒNG

Việc xây dựng các cặp câu hỏi - câu trả lời chất lượng cao từ các trang web hỏi đáp cộng đồng là nhằm tìm ra được các câu trả lời có ý nghĩa cho mỗi câu hỏi tương ứng trong một tập rất lớn các luồng hỏi - đáp. Vì vậy, trong nghiên cứu này chúng tôi coi vấn đề xây dựng các cặp câu hỏi - câu trả lời chất lượng cao từ các trang web hỏi đáp cộng đồng như là một vấn đề phân loại các cặp câu hỏi - câu trả lời và được phát biểu như sau:

Cho một tập Q các câu hỏi, ở đó mỗi câu hỏi $q_i \in Q$ có một tập các câu trả lời ứng viên $\{a_{i1}, a_{i2}, \ldots, a_{in}\}$ (n = 1, 2, …). Việc phân loại các cặp câu hỏi-câu trả lời cho câu hỏi q_i chính là gán nhãn cho các câu trả lời $\{a_{i1}, a_{i2}, \ldots, a_{in}\}$ các nhãn tương ứng là $\{l_{i1}, l_{i2}, \ldots, l_{in}\}$ trong đó l_{ij} = "Good" nếu a_{ij} là câu trả lời đúng cho câu hỏi q_i, l_{ij} = "Potential" nếu a_{ij} không phải là một câu trả lời đúng cho câu hỏi q_i nhưng

có chứa những thông tin cho câu trả lời mà câu hỏi q_i cần, l_{ij} = "Bad" nếu a_{ij} là câu trả lời không liên quan đến câu hỏi q_i.

IV. THUẬT TOÁN PHÂN LOẠI

A. Thuật toán phân loại

Có nhiều bộ phân loại khác nhau đã được sử dụng để phân loại các dữ liệu văn bản như: Support Vector Machine (SVM), Naive Bayes, Maximum Entropy Models, Sparse Network of Winnows, … Tuy nhiên trong các bộ phân loại đó thì Support Vector Machine được xem là hiệu quả hơn cả [10, 11, 12]. Trong vấn đề phân loại các cặp câu hỏi - câu trả lời, mỗi cặp câu hỏi - câu trả lời được coi như là một văn bản và được biểu diễn trong mô hình không gian vectơ có số chiều rất lớn, điều này có thể được phân loại tốt bởi Support Vector Machine. Chính vì vậy trong nghiên cứu của mình, chúng tôi sử dụng bộ phân loại Support Vector Machine với hàm nhân tuyến tính.

B. Hiệu suất của phân loại

Để đánh giá hiệu suất của việc phân loại các cặp câu hỏi - câu trả lời, chúng tôi sử dụng các độ đo *precision*, *recall*, F_1*-measure, accuracy* được định nghĩa như dưới đây. Để ước lượng các độ đo này có thể dựa vào bảng I:

Bảng I. *Các kết quả dự đoán của phân loại*

	Label $y^* = +1$	Label $y^* = -1$
Prediction $f(x^*) = +1$	*TP*	*FP*
Prediction $f(x^*) = -1$	*FN*	*TN*

Mỗi ô trong bảng đại diện cho một trong bốn kết quả đầu ra có thể của một dự đoán $f(x^*)$.

Trong đó:

TP (True Positive): số lượng các cặp câu hỏi - câu trả lời positive được phân loại đúng.

TN (True Negative): số lượng các cặp câu hỏi - câu trả lời negative được phân loại đúng.

FP (False Positive): số lượng các cặp câu hỏi - câu trả lời positive bị phân loại sai.

FN (False Negative): số lượng các cặp câu hỏi - câu trả lời negative bị phân loại sai.

Precision được định nghĩa như là xác suất mà một dữ liệu phân loại là $f(x^*) = +1$ là một phân loại đúng. Nó có thể được ước lượng như sau:

$$Precision \qquad p = \frac{TP}{TP + FP} \tag{1}$$

Recall được định nghĩa như là xác suất mà một dữ liệu với nhãn là $y^* = +1$ đã được phân loại đúng. Nó có thể được ước lượng như sau:

$$Recall \qquad r = \frac{TP}{TP + FN} \tag{2}$$

$$F_1 - measure = \frac{2 * p * r}{p + r} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

V. CÁC ĐẶC TRƯNG TRONG PHÂN LOẠI

Để phân loại các cặp câu hỏi - câu trả lời chúng tôi đã thực hiện trích rút các loại đặc trưng quan trọng được trình bày dưới đây.

A. Các đặc trưng từ vựng

Đặc trưng n-gram

Các đặc trưng n-gram của một cặp câu hỏi - câu trả lời được trích rút dựa trên ngữ cảnh của các từ của câu, nghĩa là, các từ đó xuất hiện trong một cặp câu hỏi - câu trả lời. Mỗi cặp câu hỏi - câu trả lời x được biểu diễn giống như sự biểu diễn tài liệu trong mô hình không gian vectơ như sau:

$$x = (x_1, x_2, \ldots, x_N) \tag{5}$$

trong đó: x_i là tần số xuất hiện của từ i trong x và N là tổng số các từ trong x. Do tính thưa thớt của các đặc trưng, chỉ các đặc trưng có giá trị khác không mới được giữ lại trong vectơ đặc trưng. Bởi vậy các cặp câu hỏi - câu trả lời cũng được biểu diễn dưới hình thức sau:

$$x = \{(t_1, f_1), \ldots, (t_p, f_p)\} \tag{6}$$

trong đó: t_i là từ thứ i trong x và f_i là tần số xuất hiện của t_i trong x.

Để trích rút các đặc trưng n-gram, bất kỳ n từ liên tiếp nào trong một cặp câu hỏi - câu trả lời đều được coi là một đặc trưng. Bảng II là danh sách một số đặc trưng n-gram của câu hỏi "How many Grammys did Michael Jackson win in 1983 ?".

Bảng II. *Ví dụ về một số đặc trưng n-gram*

Tên đặc trưng	Đặc trưng
Unigram	{(How, 1) (many, 1) (Grammys, 1) (did, 1) (Michael, 1) (Jackson, 1) (win, 1) (in, 1) (1983, 1) (?, 1)}
Bigram	{(How-many, 1) (many-Grammys, 1) (Grammys-did, 1) (did-Michael, 1) (Michael-Jackson, 1) …(1983-?, 1)}
Trigram	{(How-many-Grammys, 1) (many-Grammys-did, 1) …(in-1983-?, 1)}

Số lượng các từ trong câu hỏi, số lượng các từ trong câu trả lời

Để phân loại các cặp câu hỏi - câu trả lời có thể dựa vào các đặc trưng là số lượng các từ trong câu hỏi, số lượng các từ trong câu trả lời. Từ quan sát dữ liệu thực tế chúng tôi thấy rằng các cặp câu hỏi - câu trả lời có số lượng các từ ít hơn 10 thường là các cặp câu hỏi - câu trả lời không có ý nghĩa.

Số lượng câu (sentence) trong mỗi câu trả lời

Đây là một đặc trưng quan trọng trong việc phân loại các cặp câu hỏi - câu trả lời. Thông thường các câu trả lời có nhiều sentence thường mang thông tin trả lời đầy đủ hơn cho câu hỏi.

Tỷ lệ giữa số lượng câu (sentence) của câu trả lời và câu hỏi

Trong nghiên cứu này chúng tôi sử dụng đặc trưng là tỷ lệ giữa số lượng câu (sentence) của câu trả lời và câu hỏi.

Chồng chéo n-gram từ giữa câu hỏi và câu trả lời

Khi trả lời một câu hỏi nào đó trên các trang web hỏi đáp cộng đồng, người sử dụng thường có xu hướng sử dụng lại một số từ ở câu hỏi trong câu trả lời của họ. Vì vậy nếu trong câu trả lời có chứa từ hoặc cụm từ của câu hỏi thì câu trả lời đó có khả năng là một câu trả lời tốt cho câu hỏi. Để tính toán sự chồng chéo giữa câu hỏi và câu trả lời, chúng tôi thực hiện loại bỏ các stopword trong mỗi câu hỏi, câu trả lời sau đó mới tính toán sự chồng chéo từ sử dụng n-gram từ (n=1, 2, 3).

B. Các đặc trưng đo sự giống nhau giữa câu hỏi và câu trả lời

Để xây dựng các đặc trưng này, chúng tôi thực hiện loại bỏ các từ stopword trong mỗi câu hỏi và câu trả lời. Các câu hỏi và câu trả lời sau đó được biểu diễn dưới dạng vectơ (bag-of-word). Để tính toán sự giống nhau giữa câu hỏi và các câu trả lời của nó, chúng tôi sử dụng 5 độ đo khác nhau: euclidean, manhattan, minkowski, cosine, jaccard. Bảng III là một ví dụ về việc tính toán các đặc trưng đo sự giống nhau này.

Bảng III. *Ví dụ về các đặc trưng đo sự giống nhau*

Câu hỏi	Câu trả lời	Các độ đo	Các giá trị độ đo
Massage oil. Where I can buy good oil for massage?	You might be able to find Body Massage Oil in Body Shop at Landmark or City Centre, and if they do have it there, ...	euclidean	5.196152
		manhattan	25
		minkowski	3.141
		cosine	0.405062
		jaccard	1.0

C. Đặc trưng dựa trên thông tin người dùng

Số lượng câu trả lời của người trả lời

Số lượng câu trả lời của người trả lời chính là thông tin về tổng số câu trả lời của người trả lời trong toàn tập dữ liệu. Chúng tôi nhận thấy rằng những người thường xuyên trả lời các câu hỏi của người khác thì câu trả lời của họ thường mang độ chính xác cao hơn so với những câu trả lời của những người ít trả lời. Chính vì vậy trong nghiên cứu này chúng tôi sử dụng số lượng câu trả lời của người trả lời như là một đặc trưng dùng để phân loại các cặp câu hỏi - câu trả lời.

D. Các đặc trưng dựa trên sự biểu diễn vectơ từ

Chúng tôi sử dụng sự biểu diễn vectơ từ để mô hình hóa mối quan hệ ngữ nghĩa giữa câu hỏi và các câu trả lời của nó. Chúng tôi chọn mô hình word2vec[2] đề xuất bởi Mikolov [13, 14] để tính toán độ tương tự ngữ nghĩa giữa câu hỏi và câu trả lời. Word2vec biểu diễn các từ dưới dạng một phân bố quan hệ với các từ còn lại. Giả sử ta có một vectơ có số chiều 100. Khi đó, mỗi từ được biểu diễn bằng một vectơ có các phần tử mang giá trị là phân bố quan hệ của từ này đối với các từ khác trong từ điển. Trong bài báo này chúng tôi sử dụng tập dữ liệu từ Qatar Living (English)[3] để huấn luyện mô hình word2vec với các vectơ có số chiều là 200.

Độ tương tự ngữ nghĩa giữa câu hỏi và câu trả lời

Các câu hỏi và câu trả lời được phân tích thành các từ tố và biểu diễn dưới dạng các vectơ từ sử dụng mô hình huấn luyện word2vec. Đối với việc tính toán độ tương tự chúng tôi sử dụng tính toán độ tương tự giữa các thành phần của câu hỏi với câu trả lời: giữa tiêu đề (QSubject) của câu hỏi với câu trả lời, giữa phần mô tả của câu hỏi (QBody) với câu trả lời, giữa câu hỏi (Qsubject + QBody) với câu trả lời. Bảng IV là một ví dụ về việc tính toán độ tương tự ngữ nghĩa giữa câu hỏi và câu trả lời.

Bảng IV. Ví dụ về tính toán độ tương tự ngữ nghĩa

Câu hỏi		Câu trả lời	Độ tương tự
Qsubject	Massage oil.	You might be able to find Body Massage Oil in Body Shop at Landmark or City Centre, and if they do have it there, ...	0.2692716
QBody	Where I can buy good oil for massage?		0.7076797
Qsubject +QBody	Massage oil. Where I can buy good oil for massage?		0.6686702

Gióng từ giữa câu hỏi và câu trả lời

Các câu hỏi và câu trả lời được phân tích thành các từ tố và biểu diễn dưới dạng các vectơ từ sử dụng mô hình huấn luyện word2vec. Mỗi từ t_k trong câu hỏi sau đó sẽ được gióng với tất cả các từ trong câu trả lời và lựa chọn độ tương tự vectơ lớn nhất như công thức dưới đây:

$$score(t_k) = \max_{1 \le h \le m} (word2vec\_sim(t_k, b_h)) \quad (7)$$

Trong đó:

m- số từ trong câu hỏi;

t_k- sự biểu diễn vectơ của từ thứ k trong câu hỏi;

b_h- sự biểu diễn vectơ của từ thứ h trong câu trả lời;

$word2vec\_sim(t_k, b_h)$- độ tương tự cosin giữa hai sự biểu diễn vectơ từ của t_k và b_h.

Điểm số tương tự giữa câu hỏi và câu trả lời được tính toán như sau:

$$score(a_i) = \frac{\sum_{k=1}^{n} score(t_k)}{n} \quad (8)$$

Trong đó: n là số lượng các từ trong câu hỏi.

Độ tương tự ngữ nghĩa giữa câu trả lời và loại của câu hỏi (QCategory)

Chúng tôi cũng sử dụng độ tương tự ngữ nghĩa giữa mỗi câu trả lời với loại (QCategory) của câu hỏi tương ứng của nó. Trong tập dữ liệu làm thực nghiệm ở phần VI, các câu hỏi đã được phân vào một trong 27 loại khác nhau. Bảng V là một ví dụ về việc tính toán độ tương tự ngữ nghĩa giữa câu trả lời và các loại của câu hỏi.

Bảng V. *Ví dụ về tính toán độ tương tự ngữ nghĩa*

Câu trả lời	Loại câu hỏi (QCategory)	Độ tương tự ngữ nghĩa
You might be able to find Body Massage Oil in Body Shop at Landmark or City Centre, and if they do have it there, ...	Beauty and Style	0.1182937
	Electronics	0.2048591
	Doha Shopping	0.3174826
	Cars	0.0705854

VI. CÁC THỰC NGHIỆM VÀ ĐÁNH GIÁ

Trong phần này chúng tôi sẽ thực hiện việc xây dựng các thực nghiệm sử dụng thuật toán phân loại SVM và các đặc trưng chúng tôi đề xuất đã được trình bày trong phần V.

[2] https://code.google.com/p/word2vec

[3] http://alt.qcri.org/semeval2015/task3/index.php?id=data-and-tools

A. Tập dữ liệu

Trong các trang web hỏi đáp cộng đồng, mỗi câu hỏi thường chứa một tiêu đề hỏi và một đoạn văn bản ngắn mô tả về nội dung hỏi được đưa ra bởi người hỏi. Phần tiêu đề hỏi và phần mô tả được coi như là một câu hỏi duy nhất gồm nhiều câu [15].

Để thực hiện các thực nghiệm của mình, chúng tôi đã sử dụng tập dữ liệu từ SemEval 2015[4]. Tập dữ liệu này được trích rút từ các trang web hỏi đáp cộng đồng, bao gồm các câu hỏi và mỗi câu hỏi gồm một tập các câu trả lời tương ứng. Tất cả các cặp câu hỏi - câu trả lời đều được trình bày bằng ngôn ngữ tiếng Anh. Tập dữ liệu này bao gồm 3 tập con: train - tập dữ liệu dùng để huấn luyện mô hình phân loại, dev - tập dữ liệu dùng để đánh giá mô hình phân loại và test - tập dữ liệu dùng để kiểm tra tính hiệu quả của mô hình phân loại. Bảng VI trình bày một số thống kê trên tập dữ liệu này.

Bảng VI. *Thống kê tập dữ liệu được sử dụng*

Tập dữ liệu	Số câu hỏi	Số câu trả lời	Số câu trả lời trung bình của mỗi câu hỏi
Train	2270	11503	5.07
Dev	255	1178	4.62
Test	317	1526	4.81

B. Các thực nghiệm

Thực nghiệm 1:

Bảng VII. *Kết quả phân loại sử dụng các đặc trưng từ vựng*

Tập dữ liệu	*Accuracy*	*Precision*	*Recall*	*F_1-measure*
Dev	56.37%	49.64%	48.62%	47.91%
Test	61.53%	48.03%	47.72%	47.73%

Trong thực nghiệm này chúng tôi muốn kiểm tra tính hiệu quả của việc sử dụng các đặc trưng từ vựng như được trình bày trong mục V.A. Các đặc trưng từ vựng này bao gồm: đặc trưng Unigram, số từ trong câu hỏi, số từ trong câu trả lời, số lượng câu (sentence) trong câu trả lời, tỷ lệ giữa số câu của câu trả lời và câu hỏi, chồng chéo n-gram từ giữa câu hỏi và câu trả lời. Bảng VII trình bày các kết quả của thực nghiệm này.

[4] http://alt.qcri.org/semeval2015/task3/

Thực nghiệm 2:

Thực nghiệm thứ 2 này chúng tôi sử dụng các đặc trưng tính toán sự giống nhau giữa câu hỏi và câu trả lời. Để tính toán được các độ đo sự giống nhau giữa câu hỏi và câu trả lời, chúng tôi thực hiện: (1) loại bỏ các từ stopword trong mỗi câu hỏi và câu trả lời; (2) biểu diễn mỗi câu hỏi và câu trả lời dưới dạng các bag-of-word; (3) sử dụng các độ đo euclidean, manhattan, minkowski, cosine, jaccard để tính toán độ tương tự giữa câu hỏi và các câu trả lời của nó. Kết quả của thực nghiệm này được trình bày trong bảng VIII.

Bảng VIII. *Kết quả phân loại sử dụng các đặc trưng tính toán độ tương tự*

Tập dữ liệu	*Accuracy*	*Precision*	*Recall*	*F_1-measure*
Dev	54.84%	41.85%	42.13%	41.96%
Test	57.93%	41.25%	42.15%	41.57%

Thực nghiệm 3:

Thực nghiệm này được thực hiện với việc sử dụng đặc trưng trích rút từ thông tin người sử dụng (những người gửi câu hỏi, câu trả lời). Kết quả của thực nghiệm được trình bày trong bảng IX.

Bảng IX. *Kết quả phân loại sử dụng đặc trưng trích rút từ thông tin người dùng*

Tập dữ liệu	*Accuracy*	*Precision*	*Recall*	*F_1-measure*
Dev	61.63%	34.29%	38.89%	30.94%
Test	66.32%	35.36%	85.34%	30.44%

Thực nghiệm 4:

Trong thực nghiệm này chúng tôi sử dụng các đặc trưng tính toán độ giống nhau về mặt ngữ nghĩa giữa các thành phần của câu hỏi với câu trả lời. Để tính toán độ tương tự ngữ nghĩa giữa các thành phần của câu hỏi và câu trả lời, chúng tôi sử dụng các tập dữ liệu đã được loại bỏ các từ stopword và tập dữ liệu gốc (chưa loại bỏ các từ stopword). Tuy nhiên khi thực nghiệm phân loại chúng tôi thấy rằng việc sử dụng tập dữ liệu đã loại các từ stopword cho kết quả phân loại thấp hơn việc sử dụng tập dữ liệu gốc. Vì vậy chúng tôi quyết định chỉ sử dụng tập dữ liệu gốc cho việc tính toán độ

tương tự ngữ nghĩa. Kết quả phân loại của thực nghiệm 4 được trình bày trong bảng X.

Bảng X. Kết quả phân loại sử dụng các đặc trưng tính toán độ tương tự ngữ nghĩa

Tập dữ liệu	*Accuracy*	*Precision*	*Recall*	*F_1-measure*
Dev	60.61%	43.42%	52.32%	45.16%
Test	59.90%	46.83%	46.27%	46.38%

Thực nghiệm 5:

Trong thực nghiệm này chúng tôi thực hiện phân loại các cặp câu hỏi - câu trả lời bằng cách kết hợp tất cả các loại đặc trưng đã được thực hiện trong các thực nghiệm trên. Các kết quả phân loại của thực nghiệm này được trình bày trong bảng XI.

Bảng XI. *Kết quả phân loại sử dụng sự kết hợp của nhiều loại đặc trưng*

Tập dữ liệu	**Accuracy**	**Precision**	**Recall**	**F_1-measure**
Dev	65.62%	52.92%	56.88%	54.41%
Test	69.72%	50.91%	62.87%	53.84%

Từ các kết quả của các thực nghiệm trên chúng tôi nhận thấy rằng việc phân loại các cặp câu hỏi - câu trả lời trong các hệ thống hỏi đáp cộng đồng cần sự kết hợp của nhiều loại đặc trưng khác nhau để cho kết quả tốt. Các đặc trưng về từ vựng đóng một vai trò quan trọng trong nhiệm vụ này. Điều này là do các câu trả lời của người dùng thường được viết một cách tự do, không theo một cấu trúc nhất định, có nhiều câu trả lời trình bày sai cấu trúc cú pháp hoặc chứa những từ không liên quan đến câu hỏi. Các kết quả từ thực nghiệm 4 cho thấy việc trích rút các đặc trưng dựa trên sự biểu diễn vectơ từ (ở đây là word2vec) cũng có ý nghĩa quan trọng trong việc phân loại các cặp câu hỏi - câu trả lời. Việc huấn luyện lại mô hình word2vec và sử dụng nó trong việc tính toán độ tương tự ngữ nghĩa giữa các thành phần của câu hỏi với câu trả lời, giữa câu trả lời với các loại của câu hỏi đã cho kết quả phân loại cao. Trong thực nghiệm 5 chúng tôi đã thực hiện việc kết hợp của nhiều loại đặc trưng khác nhau và đã đạt được kết quả phân loại cao nhất trong tất cả các độ đo mà chúng tôi sử dụng. Điều này cũng chứng minh rằng vấn đề phân loại các cặp câu hỏi - câu trả lời trong các trang web hỏi đáp cộng đồng cần sự kết hợp của nhiều loại đặc trưng khác nhau.

Chúng tôi cũng thực hiện so sánh các kết quả nghiên cứu của chúng tôi với các kết quả nghiên cứu của các tác giả khác. Các nghiên cứu mà chúng tôi sử dụng để so sánh ở đây cũng sử dụng tập dữ liệu từ SemEval 2015 và sử dụng cùng số lớp phân loại. Bảng XII trình bày một số kết quả nghiên cứu của các tác giả khác để so sánh với các kết quả của chúng tôi trong vấn đề phân loại các cặp câu hỏi - câu trả lời.

Bảng XII. *So sánh với các kết quả nghiên cứu khác*

Nghiên cứu của tác giả	***F_1-measure***	***Accuracy***
Massimo Nicosia (2015)[6]	53.74%	70.50%
Liang Yi (2015)[16]	53.47%	70.55%
Xiaoqiang Zhou (2015)[17]	49.60%	67.86%
Yonatan Belinkov (2015)[18]	49.10%	66.45%
Amin Heydari (2015)[19]	47.34%	56.83%
Vo (2015)[20]	47.32%	69.13%
Ivan Zamanov (2015)[21]	46.07%	62.35%
Nghiên cứu của chúng tôi	53.84%	69.72%

Từ bảng so sánh cho thấy nghiên cứu của chúng tôi cho kết quả phân loại cao nhất về độ đo *F_1-measure*.

VII. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã trình bày những đề xuất của chúng tôi trong việc xây dựng các cặp câu hỏi - câu trả lời chất lượng cao từ các dữ liệu thu thập trên các trang web hỏi đáp cộng đồng. Chúng tôi đã thực hiện trích rút nhiều loại đặc trưng khác nhau từ các đặc trưng từ vựng, các đặc trưng dựa trên sự tính toán độ tương tự giữa câu hỏi và câu trả lời, các đặc trưng dựa trên sự biểu diễn vectơ từ (ở đây là word2vec) và sử dụng bộ phân loại Support Vector Machines để phân loại các cặp câu hỏi - câu trả lời. Các kết quả của thực nghiệm cho thấy bộ phân loại đã đạt kết quả phân loại với độ đo *F_1-measure* cao nhất là 53.84% khi sử dụng sự kết hợp của nhiều loại đặc trưng. Từ các kết quả nghiên cứu trên, chúng tôi đã xây dựng được một bộ dữ liệu bao gồm các cặp câu

hỏi - câu trả lời chất lượng để phục vụ làm nguồn dữ liệu cho việc xây dựng các hệ thống hỏi đáp tự động. Các nghiên cứu tiếp theo chúng tôi sẽ thực hiện xây dựng các cặp câu hỏi - câu trả lời có ý nghĩa từ nhiều nguồn hỏi đáp cộng đồng khác nhau để làm phong phú thêm nguồn dữ liệu hỏi đáp phục vụ xây dựng các hệ thống hỏi đáp tự động.

TÀI LIỆU THAM KHẢO

[1] Zeyi Wen, Rui Zhang, Kotagiri Ramamohanarao. Enabling Precision/Recall Preferences for Semi-supervised SVM Training, CIKM'14, pp. 421-430, 2014.

[2] Valentin Jijkoun and Maarten de Rijke. Retrieving answers from frequently asked questions pages on the web. In CIKM '05, pp. 76-83, 2005.

[3] Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. Bridging the lexical chasm: Statistical approaches to answer-f nding. In Proceedings of SIGIR, pp. 192-199, 2000.

[4] C. Shah, J. Pomerantz. Evaluating and predicting answer quality in community QA. In Proceedings of SIGIR, 2010.

[5] H. Toba, Z. Y. Ming, M. Adriani, T. Chua. Discovering high quality answers in community question answering archives using a hierarchy of classif ers. Information Sciences 261, pp. 101-115, 2014.

[6] Massimo Nicosia1, Simone Filice, et al. QCRI: Answer Selection for Community Question Answering – Experiments for Arabic and English. In Proceedings of SemEval, pp. 203-209, 2015.

[7] J. Lou, Y. Fang, K.H. Lim, J.Z. Peng. Contributing high quantity and quality knowledge to online q&a communities. Journal of the American Society for Information Science and Technology 64(2), pp. 356-371, 2013.

[8] H. Hu, B. Liu, B. Wang, M. Liu, X. Wang. Multimodal DBN for predicting high-quality answers in cQA portals. In Proceedings of ACL, pp. 843-847, 2013.

[9] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers on large online QA collections. In Proceedings of ACL-08: HLT. Association for Computational Linguistics, pp. 719-727, 2008.

[10] Zhiheng Huang, Marcus Thint, and Zengchang Qin. Question classif cation using head words and their hypernyms. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP '08), pp. 927-936, 2008.

[11] Vijay Krishnan, Sujatha Das, and Soumen Chakrabarti. Enhanced answer type inference from questions using sequential models. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, pp. 315-322, 2005.

[12] Babak Loni, Gijs van Tulder, Pascal Wiggers, David M.J. Tax, and Marco Loog. Question classif cation with weighted combination of lexical, syntactical and semantic features. In Proceedings of the 15th international conference of Text, Dialog and Speech, pp. 243-250, 2011.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean. (2013a) Eff cient Estimation of Word Representations in Vector Space. CoRR, abs/1301.3781.

[14] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. (2013b) Distributed Representations of Words and Phrases and their Compositionality. CoRR, abs/1310.4546.

[15] Vinay Pande, Tanmoy Mukherjee, Vasudeva Varma. Summarizing Answers For Community Question Answer Services, The International Conference of the German Society for Computational Linguistics and Language Technology, pp. 151-161, 2013.

[16] Liang Yi, Jianxiang Wang, Man Lan. ECNU: Using Multiple Sources of CQA-based Information for Answer Selection and YES/NO Response Inference. In Proceedings of SemEval, pp. 236-241, 2015.

[17] Xiaoqiang Zhou Baotian Hu Jiaxin Lin Yang Xiang Xiaolong Wang. ICRC-HIT: A Deep Learning based Comment Sequence Labeling System for Answer Selection Challenge. In Proceedings of SemEval, pp. 210-214, 2015.

[18] Yonatan Belinkov, Mitra Mohtarami, Scott Cyphers, James Glass. VectorSLU: A Continuous Word Vector Approach to Answer Selection in Community Question Answering Systems. In Proceedings of SemEval, pp. 282-287, 2015.

[19] Amin Heydari, Alashty Saeed Rahmani Meysam Roostaee Mostafa Fakhrahmad. Shiraz: A Proposed List Wise Approach to Answer Validation. In Proceedings of SemEval, pp. 220-225, 2015.

[20] Ngoc Phuoc An Vo, Simone Magnolini, Octavian Popescu. FBK-HLT: An Application of Semantic Textual Similarity for Answer Selection in Community Question Answering. In Proceedings of SemEval, pp. 231-235, 2015.

[21] Ivan Zamanov, Nelly Hateva, et al. Voltron: A Hybrid System For Answer Validation Based On Lexical And Distance Features. In Proceedings of SemEval, pp. 242-246, 2015.

CONSTRUCTING HIGH-QUALITY QUESTION-ANSWER PAIRS FROM COMMUNITY QUESTION ANSWERING SITES

Abstract: Community Question Answering (cQA) sites that contains a large amount of valuable information generated by the users. In cQA sites, users can post questions, answer other people's questions and provide feedback to the questions / answers. In this paper, we use machine learning approach to constructing high-quality question - answer pairs from community question answering sites. These question - answer pairs will be used as the data source for the automatic question answering systems. We extracted important features from each question-answer thread as well as the users information and build classif cation model to identify the meaningful question - answer pairs. The experimental results on the data provided by SemEval 2015 showed that our proposal will bring good results.

Keywords: Community Question Answering, classif cation, Support Vector Machines, Automatic Question Answering system.

Nguyễn Văn Tú tốt nghiệp cử nhân tại khoa Toán trường Đại học Sư phạm Thái Nguyên ngành Sư phạm tin năm 2005, tốt nghiệp thạc sĩ tại trường Đại học Sư phạm Hà Nội năm 2009. ThS. Nguyễn Văn Tú hiện đang làm nghiên cứu sinh tại trường Đại học Công nghệ và làm việc tại trường Đại học Tây Bắc. Hướng nghiên cứu bao gồm: Các kỹ thuật học máy, xử lý ngôn ngữ tự nhiên.

Lê Anh Cường tốt nghiệp cử nhân và thạc sĩ Công nghệ Thông tin tại trường Đại học Công nghệ, Đại học Quốc gia Hà Nội vào năm 1998 và 2001, và nhận bằng tiến sĩ tại Trường Khoa học thông tin - Viện Khoa học và Công nghệ tiên tiến Nhật Bản (Japan Advanced Institute of Science and Technology) vào năm 2007. Hiện nay, PGS. TS Lê Anh Cường đang là giảng viên tại khoa Công nghệ thông tin, trường Đại học Tôn Đức Thắng. Lĩnh vực nghiên cứu bao gồm: xử lý ngôn ngữ tự nhiên, khai phá văn bản, học máy.

Nguyễn Hà Nam tốt nghiệp cử nhân tại trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội năm 2001, nhận bằng thạc sĩ tại trường Đại học Chungwoon, Hàn Quốc năm 2003 và tiến sĩ tại trường Đại học Hàng không, Hàn Quốc năm 2007. Hiện nay, PGS. TS Nguyễn Hà Nam đang là giảng viên tại khoa Công nghệ thông tin, trường Đại học Công nghệ - Đại học Quốc gia Hà Nội. Lĩnh vực nghiên cứu bao gồm: trí tuệ nhân tạo, khai phá dữ liệu, học máy, phân tích thống kê, cơ sở dữ liệu, kho dữ liệu và OLAP.