Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00088

PHÂN LOẠI CÂU HỎI TIẾNG VIỆT ỨNG DỤNG CHO HỆ THỐNG HỎI ĐÁP MỞ

Lê Thị Thanh Thùy, Nguyễn Văn Kiệt, Nguyễn Lưu Thùy Ngân

Trường Đại học Công nghệ Thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh

13520871@gm.uit.edu.vn, kietnv@uit.edu.vn, ngannlt@uit.edu.vn

TÓM TẮT: *Phân loại câu hỏi là một thành phần quan trọng trong các hệ thống hỏi đáp, đặc biệt là hệ thống hỏi đáp mở (Open-domain question answering system). Phân loại câu hỏi giúp xác định đối tượng cần tìm kiếm và phạm vi kiến thức của câu trả lời. Do đó độ chính xác của bộ phân loại câu hỏi ảnh hưởng nhiều đến chất lượng của một hệ thống hỏi đáp mở. Trong bài báo này, chúng tôi trình bày phương pháp phân loại câu hỏi tiếng Việt sử dụng kết hợp các phương pháp túi từ, từ khóa và quan hệ phụ thuộc. Chúng tôi tiến hành thử nghiệm phương pháp trên 2 bộ câu hỏi: bộ câu hỏi TREC tiếng Việt và bộ câu hỏi do chúng tôi tự xây dựng. Kết quả thử nghiệm cho ra hệ thống phân loại câu hỏi có độ chính xác ở lớp thô(Coarse) là 85.4% và lớp mịn(Fine-Grained) là 70.2%. Hệ thống cũng xây dựng được bộ dữ liệu được đặt tên là UIT-OQA. Bộ dữ liệu gồm 1,416 câu hỏi phù hợp với các nghiên cứu về phân loại câu hỏi và hệ thống hỏi đáp trên ngôn ngữ Tiếng Việt.*

Từ khóa*: Open-domainquestion answering, hệ thống hỏi đáp mở, quan hệ phụ thuộc, túi từ.*

I. GIỚI THIỆU

Hệ thống hỏi đáp là hệ thống tự động trả lời các câu hỏi dưới dạng ngôn ngữ tự nhiên được đặt ra bởi người dùng. Hệ thống này thuộc lĩnh vực rút trích thông tin và xử lý ngôn ngữ tự nhiên trong ngành khoa học máy tính. Các nghiên cứu về hệ thống hỏi đáp đã bắt đầu vào năm 1960. Từ đó, hệ thống hỏi đáp được nghiên cứu và phát triển đến nay. Hệ thống hỏi đáp được phân thành nhiều loại dựa theo nhiều tiêu chí khác nhau như: miền ứng dụng, dạng câu hỏi, nguồn dữ liệu được sử dụng, chức năng kết hợp và câu trả lời v.v... Phân loại hệ thống hỏi đáp thường dựa theo miền ứng dụng chia thành hai loại: hệ thống hỏi đáp mở và hệ thống hỏi đáp đóng.

Hệ thống hỏi đáp đóng là hệ thống tự động trả lời câu hỏi trong một miền lĩnh vực cụ thể như y học, công nghệ, lịch sử, v.v... Vì câu hỏi chỉ trong một lĩnh vực cụ thể nên nguồn dữ liệu chuyên biệt và khá dễ tìm kiếm. Đồng thời, câu trả lời được tìm kiếm trong nguồn dữ liệu cụ thể nên hệ thống thường mang lại độ chính xác cao. Tuy nhiên, cũng vì những lý do trên nên hệ thống chỉ phù hợp với người dùng cần tìm kiếm, nghiên cứu trên lĩnh vực cụ thể, nếu lĩnh vực khác thì hệ thống không trả lời được. Ngoài ra, khi người dùng muốn tìm kiếm câu trả lời cần xác định được câu hỏi mình thuộc lĩnh vực nào, từ khóa nào mới có thể tìm kiếm đúng hệ thống cần sử dụng. Do đó, hệ thống không mang tính chất ứng dụng rộng rãi.

Ngược lại với hệ thống hỏi đáp đóng, hệ thống hỏi đáp mở có thể tự động trả lời câu hỏi ở bất kỳ lĩnh vực nào. Nguồn dữ liệu tham khảo để trả lời của hệ thống thường được thu thập từ tập hợp các trang wikipedia, web hoặc là các nguồn tài liệu mang kiến thức ở nhiều lĩnh vực như sách giáo khoa, bài báo, v.v... Do nguồn dữ liệu khá rộng nên độ chính xác câu trả lời thường không cao. Tuy nhiên, hệ thống hỏi đáp mở lại phù hợp với nhiều người dùng. Vì họ có thể tìm kiếm bất kỳ câu hỏi nào mà không cần có kiến thức về lĩnh vực câu hỏi cần tìm kiếm [1].

Phân loại câu hỏi giúp cho hệ thống hỏi đáp mở giới hạn được miền kiến thức cần tìm kiếm, đặc biệt có thể giúp cho các hệ thống lựa chọn cách tìm kiếm câu trả lời phù hợp. Qua đó, hệ thống có thể dễ dàng tìm kiếm câu trả lời và tăng độ chính xác. Để giúp độ chính xác của hệ thống hỏi đáp mở được cải thiện, chúng tôi đã xây dựng hệ thống phân loại câu hỏi. Sau khi người dùng nhập vào câu hỏi dưới dạng ngôn ngữ tự nhiên, hệ thống sẽ xác định loại câu hỏi cho câu đó. Ví dụ, hệ thống phân loại của chúng tôi xác định loại câu hỏi cho câu "Ngọn núi nào cao nhất Việt Nam" là địa điểm (Location_Mountain).

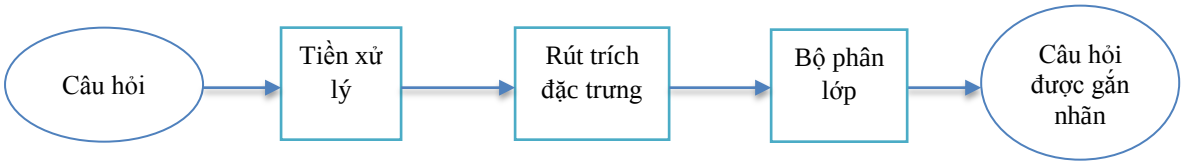
II. CÁC CÔNG TRINH LIÊN QUAN

Có rất nhiều hệ thống phân loại câu hỏi được xây dựng thành công với nhiều ngôn ngữ trên thế giới, đặc biệt là ngôn ngữ Tiếng Anh. Chẳng hạn, công trình của Tomas và Giuliano (2009) áp dụng phương pháp kernel trên bộ dữ liệu Trec cho ra phân lớp thô là 90.8%, lớp mịn là 85.6% [2]. Hay như Silva (2011) sử dụng SVM với tuyến tính kernel cho ra độ chính xác tại lớp thô, lớp mịn lần lượt là 95% và 90.8% [2].

Tuy nhiên, đối với ngôn ngữ Tiếng Việt, cho đến nay vì còn hạn chế trong dữ liệu câu hỏi nên số lượng công trình còn ít. Gần đây, nhóm tác giả xây dựng VnQCS [3] sử dụng các đặc trưng túi từ và từ khóa được rút trích từ web cho ra độ chính xác khá cao, với lớp thô là 94.1% và lớp mịn là 85.4%. Nhóm tác giả VnQCS cũng đã xây dựng nên một bộ dữ liệu tiếng Việt dựa theo bộ dữ liệu TREC tiếng Anh bằng phương pháp dịch thuật.

III. PHƯƠNG PHÁP

Trong bài báo này, chúng tôi đã quyết định thử nghiệm phương pháp túi từ và từ khóa của VNQCS kết hợp với quan hệ phụ thuộc dựa theo ý tưởng của Li Xin và cộng sự [4] trên hai bộ dữ liệu tiếng Việt. Hình 1 trình bày mô hình của hệ thống phân loại câu hỏi.



Hình 1. Mô hình hệ thống phân loại câu hỏi

Đầu vào là câu hỏi được người dùng nhập vào dưới ngôn ngữ tự nhiên. Hệ thống sẽ đưa qua bộ tiền xử lý để tách từ và loại bỏ các stop-word. Sau đó, hệ thống tiếp tục đưa câu hỏi thông qua bộ rút trích đặc trưng để chuyển câu hỏi về dạng các vector rồi đưa vào bộ phân lớp. Sau khi câu hỏi được phân lớp và gán nhãn, hệ thống sẽ đưa ra câu hỏi với loại nhãn tương ứng.

Xây dựng bộ dữ liệu dựa theo chuẩn TREC

Việc lựa chọn các lớp để phân loại câu hỏi cũng đã được đề xuất bởi nhiều nghiên cứu khác nhau. Tuy nhiên, cách phân loại được Li và Roth [5] đề xuất được sử dụng hầu hết trong các nghiên cứu những năm gần đây. Tập dữ liệu này được xuất bản lần đầu tiên tại Đại học Illinois Urbana-Champaign được gọi là tập dữ liệu UIUC. Ngoài ra, tập dữ liệu thường được biết với tên gọi khác là tập dữ liệu TREC vì nó được biết đến rộng rãi từ hội nghị Truy xuất Văn bản (TREC). Dữ liệu TREC là dữ liệu tiếng Anh, bao gồm 6000 câu hỏi được chia làm thành 2 tập nhỏ: tập huấn luyện gồm 5500 câu và tập thử nghiệm gồm 500 câu. Các câu hỏi trong tập dữ liệu đều được gán nhãn thành 6 lớp thô (Coarse) và 50 lớp hạt mịn (Fine-Grained) [2].

Bảng 1. Nhãn phân loại câu hỏi theo TREC [3]

Lớp thô	**Lớp mịn**
ABBR	abbreviation, expansion
DESC	definition, descrption, manner, reason
ENTY	animal, body, color, creation, currency, disease, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUM	description, group, individual, title
LOC	city, country, mountain, other, state
NUM	code, count, date, distance, money, order, other, percent, percent, period, speed, temperature, size, weight

Chúng tôi tiến hành huấn luyện và kiểm tra trên hai bộ dữ liệu đều được xây dựng từ chuẩn TREC: bộ câu hỏi TREC tiếng Việt và bộ câu hỏi do chúng tôi xây dựng nhằm kiểm tra tính khái quát hoá khả năng ứng dụng của hệ thống.

Bộ câu hỏi TREC tiếng Việt

Bộ câu hỏi TREC tiếng Việt được xây dựng bởi nhóm tác giả Dang Hai Tran và cộng sự [3] dựa theo bộ dữ liệu TREC tiếng Anh. Sau đó, tác giả chuyển đổi sang ngôn ngữ tiếng Việt. Bộ câu hỏi được tiến hành thực hiện bởi 5 sinh viên có TOEFL lớn hơn 500. Các sinh viên trên đã chuyển đổi ngôn ngữ và kiểm tra chéo nhau nhằm nâng cao độ chính xác dịch thuật. Trong quá trình chuyển đổi họ cũng tuân theo các quy tắc sau:

- Quy tắc 1: Nội dung câu hỏi tiếng Việt đúng với nhãn tiếng Anh của nó trước khi được chuyển đổi.
- Quy tắc 2: Tên thực thể được thay đổi nhưng vẫn tránh thay đổi ngữ nghĩa.
- Quy tắc 3: Một câu hỏi tiếng Anh có thể dịch thành nhiều câu hỏi tiếng Việt khác cấu trúc ngữ pháp nhưng có cùng ý nghĩa và vẫn đảm bảo được chất lượng của câu hỏi[2].

Do đó, độ tin cậy của bộ dữ liệu được đảm bảo. Tuy nhiên, bộ dữ liệu vẫn còn tồn tại những câu hỏi mang nhược điểm sau:

- Câu hỏi mang nhiều đặc trưng văn hóa, phong tục của phương Tây. Ví dụ: "Gandy dancer là gì?", "Philebus like nghĩa là gì?". Nếu không là người phương Tây hoặc dữ liệu tiếng nước ngoài thì không thể hiểu được khái niệm "gandy dancer", "philebus-like".
- Câu hỏi đơn thuần chỉ là dịch nghĩa từ vựng tiếng Anh sang tiếng Việt. Ví dụ: "Antidisestablishmentarianism có nghĩa_là gì?","Nest-ce pas nghĩa_là gì?". Từ "antidisestablishmentarianism" và "nest-ce pas" lần lượt là một từ tiếng Anh và tiếng Pháp cần được dịch nghĩa.
- Câu hỏi mang tính tối nghĩa. Ví dụ: "Hình thái đại diện cái gì trên đảo Phục Sinh?", "Một giấc_mơ ướt là gì?". Hai cụm từ "hình thái đại diện" và "giấc mơ ướt" được dịch từ tiếng Anh sang tiếng Việt làm câu hỏi tối nghĩa. Người đọc hay hệ thống khó tìm ra được câu trả lời.

Xây dựng bộ câu hỏi tiếng Việt mới

Nhằm khắc phục những nhược điểm của bộ dữ liệu trên, chúng tôi đã xây dựng ra một bộ câu hỏi hoàn toàn xuất phát từ tiếng Việt. Dữ liệu ban đầu gồm 1500 câu hỏi về mọi lĩnh vực được chúng tôi tổng hợp từ chương trình truyền hình "Ai là triệu phú" và "Đường lên đỉnh Olympia", tuy nhiên còn mang nhiều lỗi ngữ pháp. Sau khi chỉnh sửa dữ liệu bao gồm 1,416 câu, trong đó tập huấn luyện có 944 câu và tập thử nghiệm là 472 câu. Bộ câu hỏi cũng được gắn nhãn theo chuẩn TREC. Do sự khác nhau về ngôn ngữ cũng như sự nhập nhằng trong định nghĩa gán nhãn của TREC nên chúng tôi đã xây dựng hướng dẫn gán nhãn để người gán nhãn có thể tham khảo khi gán nhãn cho bộ dữ liệu.

Bảng 2. Nhãn phân loại bộ dữ liệu UIT-OQA theo chuẩn TREC[5]

Lớp	**Định nghĩa**	**Lớp**	**Định nghĩa**
ABBREVIATION	Viết tắt	description	Mô tả của vật
abb	Từ viết tắt	reason	Nguyên nhân, lý do
exp	Nghĩa từ viết tắt	HUMAN	Con người
ENTITY	Thực thể	title	Chức vụ, họ của một người
animal	Các loài động vật, linh vật	group	Nhóm người hoặc tổ chức: công ty, dân tộc, triều đại, ban nhạc, v.v…
body	Các bộ phận cơ thể của người, động vật, thực vật	ind	Tên người
color	Màu sắc	description	Mô tả về người
creative	Những thực thể liên quan đến nghệ thuật: sách, phim, báo, thơ, kịch, v.v…	LOCATION	Địa điểm
currency	Tên các loại tiền tệ	city	Thành phố
dis.med	Các loại bệnh và phương thuốc	country	Đất nước
event	Các sự kiện: lễ hội, giải thưởng, trận chiến lịch sử, v.v…	mountain	Núi
food	Thức ăn, nước uống.	other	Các địa điểm khác: Sông, chùa, cầu, huyện, xã, v.v…
instrument	Các nhạc cụ	state	Tỉnh, bang
lang	Ngôn ngữ	NUMERIC	Số
letter	Chữ cái	code	Mật mã, số điện thoại
other	Các thực thể khác	count	Đếm số lượng
plant	Các loài thực vạt	date	Ngày, tháng, năm
product	Sản phẩm	distance	Khoảng cách, độ dài
religion	Tôn giáo	money	Tiền

sport	Thể thao	order	Thứ hạng
substance	Yếu tố và các chất	other	Các loại số khác
symbol	Kí tự	period	Thời gian kéo dài: giai đoạn, thời kì, độ tuổi, v.v…
technique	Kĩ thuật và phương pháp	percent	Phần trăm
term	Khái niệm tương đương	speed	Tốc độ
vehicle	Phương tiện giao thông: xe, máy bay, v.v…	temp	Nhiệt độ
word	Từ có tính chất đặc biệt	size	Kích thước chung, âm lượng
DESCRIPTION	Mô tả và tóm tắt khái niệm	weight	Cân nặng
definition	Định nghĩa của vật		

Trong 6 loại nhãn thô, chúng tôi nhận thấy rằng, nhãn câu hỏi loại ENTY chiếm số lượng câu nhiều nhất với 442 câu. Nhãn chiếm số lượng thấp nhất là ABBR với 16 câu. Trong lớp mịn của bộ dữ liệu, nhóm nhãn ind của HUM chiếm tỉ lệ cao nhất với 363 câu. Nhóm nhãn chiếm tỉ lệ thấp nhất có số lượng câu hỏi là 1 gồm currency, letter, code, oder, temp. Bảng 3 và bảng 4 mô tả tỉ lệ phần trăm loại câu hỏi được gắn nhãn theo lớp thô và lớp mịn.

Bảng 3. Tỉ lệ phần trăm loại câu hỏi được gắn theo lớp thô

Lớp thô	Phần trăm trong bộ dữ liệu (%)
ABBR	1
DESC	2
ENTY	31
HUM	28
LOC	24
NUM	14

Bảng 4. Tỉ lệ phần trăm loại câu hỏi được gắn theo lớp mịn

Lớp	**Phần trăm trong bộ dữ liệu (%)**	**Lớp**	**Phần trăm trong bộ dữ liệu (%)**
abb	0.78	vehicle	0.21
exp	0.35	word	0.21
animal	2.9	description	0.85
body	1.13	reason	0.78
color	0.56	title	0.28
creative	5.01	group	2.4
currency	0.07	ind	25.6
dis.med	0.14	city	2.33
event	0.98	country	7.56
food	0.21	mountain	0.49
instrument	0.56	other	10.66
lang	0.28	state	3.1
letter	0.07	code	0.07
other	6.36	count	3.88
plant	0.92	date	8.55
product	0.21	distance	0.07
religion	0.14	order	0.21
sport	0.21	other	0.21
substance	5.65	period	0.49
symbol	0.49	temp	0.07
technique	0.28	temp	0.07
term	3.53	temp	0.07

Bộ câu hỏi của chúng tôi gồm có 1,416 câu. Vì vậy bộ dữ liệu chỉ tồn tại 42 nhãn mịn. Ngoài ra, bộ dữ liệu vẫn tồn tại những câu hỏi được gán nhãn với độ tin cậy thấp chiếm 10.6% (150 câu/1,416 câu) bộ dữ liệu. Những câu hỏi này không được tin cậy vì tồn tại sự nhập nhằng giữa hai nhóm term thuộc ENTY (ENTY_term) và other thuộc ENTY (ENTY_other) hoặc manner thuộc DESC (DESC_manner) và technique thuộc ENTY (ENTY_technique).

Rút trích các đặc trưng

Hệ thống sử dụng bốn đặc trưng chính: Từ, từ khóa, túi từ và quan hệ phụ thuộc. Mỗi đặc trưng sẽ được sử dụng tại các thành phần khác nhau của mô hình. Bảng 5 trình bày các thành phần mô hình sử dụng đặc trưng.

Bảng 5. Tóm tắt các đặc trưng được xử dụng

Đặc trưng	Thành phần mô hình sử dụng đặc trưng
Từ	Tiền xử lý
Từ khóa	Rút trích đặc trưng
Túi từ	Rút trích đặc trưng
Quan hệ phụ thuộc	Rút trích đặc trưng

Từ

Các câu hỏi sẽ được tách từ bằng công cụ vnTokenizer 4.1.1 của nhóm tác giả Hong Phuong Le và cộng sự [6]. vnTokenizer là một bộ tách từ tự động các văn bản Tiếng Việt, với độ chính xác cao từ 96%-98%.

Từ khóa

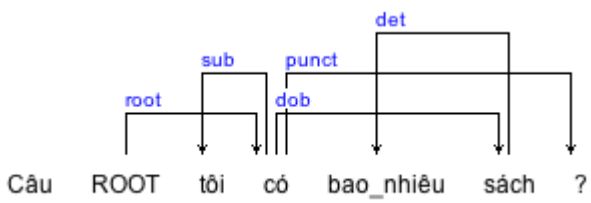
Từ khóa là những từ quan trọng được rút trích trong câu hỏi của bộ dữ liệu huấn luyện. Sự xuất hiện của các từ khóa trong câu hỏi được ghi nhận và tạo thành vector đặc trưng riêng của mỗi câu. Đối với bộ dữ liệu huấn luyện, sau khi được tách từ, chúng tôi đem đi rút trích chỉ lấy những từ khóa có số lượng nhất định. Để đảm bảo những từ khóa chúng tôi lấy có độ tin cậy cao, chúng tôi sẽ làm tuần tự theo các bước. Đầu tiên, mỗi từ khoá sau khi tính tần số và tần số nghịch, chúng tôi sắp xếp chúng theo thứ tự tần số giảm dần. Tiếp theo chúng tôi lấy kết quả đó loại bỏ tất cả các từ khoá có tần số nghịch thấp. Cuối cùng, các từ khóa cần lấy chính là những từ trên cùng của danh sách đang xét [3].

Túi từ

Thông qua túi từ, mỗi câu hỏi được chuyển về thành dạng các vector đặc trưng cho bộ phân lớp. Bộ phân lớp sẽ gắn nhãn cho từng câu hỏi dựa trên các vector đặc trưng. Trước khi chuyển về vector, câu hỏi sẽ được loại bỏ đi các stop-word bằng cách loại bỏ hết các từ có tần số nghịch thấp. Vì những từ có tần số nghịch thấp không mang tính đặc trưng cho câu hỏi chúng xuất hiện [3].

Quan hệ phụ thuộc

Chúng tôi sử dụng đặc trưng quan hệ phụ thuộc giữa hai từ trong câu. Trong nghiên cứu này, chúng tôi sử dụng bộ phân tích cú pháp phụ thuộc của nhóm tác giả Kiet Nguyen và Ngan Nguyen [7, 8]. Hình 2 minh họa về quan hệ phụ thuộc cho câu hỏi "tôi có bao_nhiêu sách?".



Hình 2. Minh họa quan hệ phụ thuộc

Nhờ vào bộ phân tích trên, hệ thống sẽ rút trích đươc các cặp từ có quan hệ với nhau. Chúng tôi hy vọng hệ thống sẽ cho ra kết quả phân tích tốt hơn.

IV. THÍ NGHIỆM

Chúng tôi thực hiện ba thí nghiệm để kiểm tra độ chính xác cũng như tính ứng dụng của hệ thống trên các bộ dữ liệu khác nhau. Hai thí nghiệm đầu tiên, chúng tôi thực hiện trên bộ dữ liệu TREC tiếng Việt. Thí nghiệm thứ nhất chúng tôi chỉ sử dụng đặc trưng túi từ để rút trích dữ liệu. Thí nghiệm hai, chúng tôi thêm đặc trưng quan hệ phụ thuộc vào hệ thống. Thí nghiệm cuối cùng, hệ thống chạy trên bộ dữ liệu UIT-OQA với phương pháp kết hợp hai đặc trưng túi từ và quan hệ phụ thuộc. Chúng tôi sử dụng phương pháp phân loại SVM của thư viện máy học Weka [9] để tiến hành thử nghiệm phân lớp các bộ dữ liệu sau khi đã rút trích các đặc trưng.

Đặc trưng túi từ trên dữ liệu TREC tiếng Việt

Bảng 6. Kết quả phân tích lớp thô trên từng số lượng câu hỏi riêng biệt với đặc trưng túi từ

Số lượng từ khóa	1000	2000	3000	4000	5000	6000	7000	**7238**
Độ chính xác	19.8%	19.8%	20.0%	20.2%	20.4%	20%	20.2%	**87.6%**

Bảng 6 trình bày kết quả hệ thống sau khi áp dụng rút trích đặc trưng túi từ trên bộ dữ liệu của dữ liệu TREC. Chúng tôi thu được kết quả cao nhất là 7238 từ khóa với độ chính xác là 87.6%. Sau đó, chúng tôi đem kết quả trên đi phân lớp mịn hơn và cho ra kết quả là 47.2%. Tuy nhiên, ta có thể thấy kết quả này cao bất thường so với các bước nhảy trước. Do đó hệ thống có khả năng đã bị overfitting.

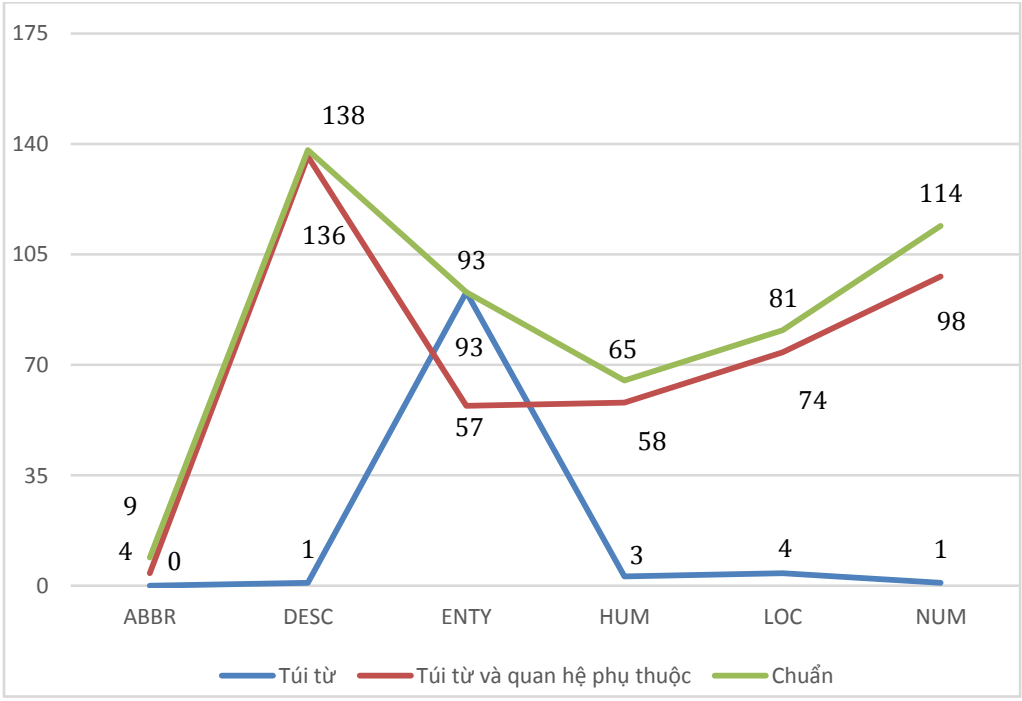
Đặc trưng túi từ và quan hệ phụ thuộc trên dữ liệu TREC tiếng Việt

Bảng 7. Kết quả phân tích lớp thô với đặc trưng Túi từ với quan hệ phụ thuộc trên bộ dữ liệu TREC tiếng Việt

Số lượng từ khóa	1000	2000	3000	4000	**5000**	6000	7000	8000
Độ chính xác	82.4%	84.4%	84.4%	84.6%	**85.4%**	85.2%	84.4%	84.2%

Bảng 7 trình bày kết quả phân tích lớp thô khi hệ thống sử dụng đặc trưng túi từ kết hợp với quan hệ phụ thuộc. Kết quả cho thấy, độ chính xác ở mỗi lần thay đổi bước nhảy có sự thay đổi ổn định và cao nhất là ở mức 5000 từ khóa, độ chính xác 85.4%. Khi chúng tôi lấy kết quả ấy đem chạy với phân lớp mịn hơn cho ra kết quả 70.2%. Sau đó, chúng tôi tiến hành so sánh kết quả phân tích lớp thô giữa phương pháp đặc trưng túi từ và phương pháp sử dụng túi từ kết hợp quan hệ phụ thuộc cho ra kết quả như bảng 8.

Bảng 8. So sánh kết quả phân tích câu hỏi gán nhãn chính xác theo lớp thô giữa phương pháp đặc trưng túi từ và phương pháp sử dụng túi từ kết hợp quan hệ phụ thuộc trên bộ dữ liệu TREC tiếng Việt tại mốc keyword 5000



Từ những kết quả phân lớp, chúng tôi nhận thấy rằng khi thêm đặc trưng quan hệ phụ thuộc, độ chính xác tăng trên 60% so với chỉ dùng một đặc trưng túi từ. Nếu hệ thống chỉ dùng riêng lẻ phương pháp túi từ, hệ thống chỉ nhận được nhãn ENTY chính xác còn các nhãn còn lại nhãn đúng trả về rất thấp. Xảy ra điều trên vì số lượng câu hỏi nhãn ENTY trong bộ dữ liệu huấn luyện chiếm tỷ lệ cao. Vì thế, số lượng từ khóa được xem là đặc trưng của nhãn ENTY chiếm phần lớn, làm cho hệ thống gán nhãn sai. Khi thêm đặc trưng quan hệ phụ thuộc, các từ khóa thêm mối quan hệ ràng buộc nhau nên kết quả trả về có độ chính xác cao hơn.

C. Đặc trưng túi từ và quan hệ phụ thuộc trên bộ dữ liệu UIT-OQA

Vì kết quả phương pháp rút trích đặc trưng túi từ và quan hệ phụ thuộc cho ra kết quả khá tốt nên chúng tôi sử dụng phương pháp ấy cho bộ dữ liệu thử nghiệm mới. Bảng 9 là kết quả phân tích lớp thô khi hệ thống áp dụng đặc trưng túi từ và quan hệ phụ thuộc.

Bảng 9. Kết quả phân tích lớp thô với đặc trưng túi từ với quan hệ phụ thuộc trên bộ dữ liệu mới

Số lượng keyword	1000	2000	3000	4000	5000	6000	6345
Độ chính xác	74.41%	74.41%	76.54%	76.75%	75.26%	72.1%	71.2%

Đối với cách làm tương tự các thí nghiệm trên, chúng tôi thu được kết quả lớp thô cao nhất ở số lượng 4000 từ khóa là 76.75% và ở lớp mịn là 54.37%.

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã thu được hai kết quả nhất định. Thứ nhất, chúng tôi xây dựng được bộ dữ liệu câu hỏi UIT-OQS cho tiếng Việt gồm 1,416 câu hỏi trải khắp 6 loại nhãn phân lớp thô và 42 loại nhãn phân lớp mịn. Độ tin cậy của bộ dữ liệu lên đến 90%. Với ngôn từ dễ hiểu và gần gũi với văn hóa Việt Nam thì khả năng ứng dụng của bộ dữ liệu là không thể phủ nhận. Thứ hai, chúng tôi đã kết hợp được đặc trưng quan hệ phụ thuộc vào hệ thống. Dù kết quả thử nghiệm trên bộ dữ liệu TREC tiếng Việt thấp hơn so với nghiên cứu trước đó nhưng độ chính xác tương đối cao. Ngoài ra, kết quả từ các thí nghiệm còn cho ta thấy được sự ảnh hưởng rõ rệt của quan hệ phụ thuộc đối với hệ thống, độ chính xác tăng khoảng 60% so với chỉ dùng túi từ.

Để phát triển hệ thống, chúng tôi sẽ cải thiện thêm đặc trưng túi từ. Chúng tôi nhận ra rằng cùng với một bộ dữ liệu và đặc trưng nhưng kết quả thu được thấp hơn nghiên cứu VnQCS[2]. Ngoài ra, chúng tôi sẽ xây dựng những mẫu dựa trên quan hệ phụ thuộc dành cho từng loại câu hỏi, nhằm mang lại độ chính xác cao hơn cho hệ thống. Trong tương lai, chúng tôi tiếp tục xây dựng bộ câu hỏi lên mức 5,000 câu. Bộ dữ liệu này sẽ hoàn toàn dựa trên các câu hỏi mang tính thực tế trong tiếng Việt, không tiếp cận theo phương pháp dịch thuật từ Anh.

TÀI LIỆU THAM KHẢO

[1] Amit Mishra, Sanjay Kumar Jain, "A survey on question answering systems with classification", Journal of King Saud University: Computer and Information Sciences, Volume 28, Number 3, 2016, , 2014.

[2] Babak Loni, "A Survey of State-of-the-Art Methods on Question Classification", Delft University of Technology, Technical report, 2011.

[3] Dang Hai Tran, Cuong Xuan Chu, Son Bao Pham, Minh Le Nguyen, "Learning Based Approaches for Vietnamese Question Classification Using Keywords Extraction from the Web", Proceeding IJCNLP 2013, At Nagoya, Japan, 2013.

[4] Li Xin, Huang Xuan-Jing, Wu Li-de "Question Classification using Multiple Classifiers", In Proceedings of the 5th Workshop on Asian Language Resources and First Symposium on Asian Language Resources Network, 2005.

[5] Xin Li, Dan Roth, "Learning Question Classifier", In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02), 2002.

[6] Kiet V. Nguyen, Ngan Luu-Thuy Nguyen, "Error Analysis for Vietnamese Dependency Parsing", 2015 Seventh International Conference on Knowledge and Systems Engineering (KSE 2015), Oct. 2015.

[7] Kiet V. Nguyen, Ngan Luu-Thuy Nguyen, "Vietnamese Dependency Parsing with Supertag Features", 2016 Eighth International Conference on Knowledge and Systems Engineering (KSE 2016), Oct. 2016.

[8] Hong Phuong Le, Thi Minh Huyen Nguyen, Azim Roussanaly, Tuong Vinh Ho, "A Hybrid Approach to Word Segmentation of Vietnamese Texts", Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, 2007.

[9] Mark Hall, Eibe Frank, Geoffey Holmes, Bernhard Pfahringer, Peter Reutemann, Lan H.Witten, "The WEKA Data Mining Software: An Update", ACM SIGKDD Explorations Newsletter 11(1):10-18, 2009.

VIETNAMESE QUESTION CLASSIFICATION FOR OPEN-DOMAIN QUESTION ANSWERING SYSTEM

Le Thi Thanh Thuy, Nguyen Van Kiet, Nguyen Luu Thuy Ngan

ABSTRACT: *Question classification is an important component of question answering systems, specially open-domain question answering system. Question classification helps to determine objects which are needed to find their answers and the knowledge around these answers, so the accuracy of the question classification has a big impact on the quality of Open-domain question answering system. We propose to use combinative methods: bag-of-word, keywords and dependency relations. We implemented this method by two corpus: the Vietnamese corpus TREC and other corpus what we build. Experimental results showed that the accuracy of the question classification systemis 85.4% in Coarse class and 70.2% in Fine-Gained class. The system also built a corpus called UIT-OQA. The corpus included 1416 questions for studies of question classification and question answering questions in Vietnamese.*