A REVIEW OF DEEP LEARNING FOR FINDING THE BEST ANSWER IN COMMUNITY QUESTION ANSWERING SYSTEM

Ha Thi Thanh[1*], Mong Thi Minh Huong[2], Ho Thi Tuyen[1], Luong Thi Minh Hue[1]
[1]TNU - University of Information and Communication Technology
[2]TNU - University of Technology

ARTICLE INFO

Received: 13/4/2021
Revised: 12/8/2021
Published: 18/8/2021

KEYWORDS

CQA
Deep Learning
Selection Answer
Attention Mechanism
Finding Best Answer

ABSTRACT

Answer selection (also called finding the best answer) is a major problem in community question answering system. When a question is posted on the forum, users can answer the question. The purpose of answer selection problem is to sort the answers according to the level of relevance to the question. The best answers will be preceded by less relevant answers. In recent years, many deep learning models have been proposed in many natural language processing problems, including the answer selection. However, these proposed models are performed on different data sets. Therefore, the aim of this paper is to survey and describe thoroughly some deep learning models applying problem of finding the best answer and analyzing some challenges on the data sets for this task in community question answering system.

TỔNG HỢP MỘT SỐ PHƯƠNG PHÁP HỌC SÂU ÁP DỤNG VÀO BÀI TOÁN LỰA CHỌN CÂU TRẢ LỜI TRONG HỆ THỐNG HỎI ĐÁP CỘNG ĐỒNG

Hà Thị Thanh[1], Mông Thị Minh Hường[2], Hồ Thị Tuyến[1], Lương Minh Huế[1]
[1]Trường Đại học Công nghệ Thông tin và Truyền thông – ĐH Thái Nguyên
[2]Trường Đại học Kỹ thuật Công nghiệp – ĐH Thái Nguyên

THÔNG TIN BÀI BÁO

Ngày nhận bài: 13/4/2021
Ngày hoàn thiện: 12/8/2021
Ngày đăng: 18/8/2021

TỪ KHÓA

CQA
Học sâu
Lựa chọn câu trả lời
Cơ chế sự chú ý
Hệ thống hỏi đáp cộng đồng

TÓM TẮT

Bài toán tìm câu trả lời (còn gọi là bài toán lựa chọn câu trả lời hay tìm câu trả lời tốt nhất) là một bài toán chính trong hệ thống hỏi đáp. Khi một câu hỏi được đăng lên forum sẽ có nhiều người tham gia trả lời câu hỏi. Bài toán lựa chọn câu trả lời với mục đích thực hiện sắp xếp các câu trả lời theo mức độ liên quan tới câu hỏi. Những câu trả lời nào đúng nhất sẽ được đứng trước các câu trả lời kém liên quan hơn. Trong những năm gần đây, rất nhiều mô hình học sâu được đề xuất sử dụng vào nhiều bài toán xử lý ngôn ngữ tự nhiên (NLP) trong đó có bài toán lựa chọn câu trả lời trong hệ thống hỏi đáp nói chung và trong hệ thống hỏi đáp cộng đồng (CQA) nói riêng. Hơn nữa, các mô hình được đề xuất lại thực hiện trên các tập dữ liệu khác nhau. Vì vậy, trong bài báo này, chúng tôi tiến hành tổng hợp và trình bày một số mô hình học sâu điển hình khi áp dụng vào bài toán tìm câu trả lời đúng trong hệ thống hỏi đáp và phân tích một số thách thức trên các tập dữ liệu cho bài toán trên hệ thống hỏi đáp.

DOI: https://doi.org/10.34238/tnu-jst.4321

[*] Corresponding author. *Email: htthanh@ictu.edu.vn*

1. Giới thiệu

Hệ thống hỏi đáp cộng đồng (ví dụ như các website nổi tiếng như Stack Overflow (https://stackoverflow.com/) and Qatar Living (https://www.qatarliving.com/forum) được biết đến với kho dữ liệu lớn lên tới hàng triệu cặp câu hỏi và các câu trả lời do người dùng trong cộng đồng tạo ra. Kho dữ liệu này qua thời gian trở thành kho dữ liệu chứa tri thức rất quý giá được nhiều người dùng sử dụng để tìm kiếm thông tin. Các nghiên cứu gần đây thực hiện trên các bài toán như lựa chọn câu trả lời, tìm câu hỏi liên quan hay phân lớp câu trả lời [1], [2]. Trong đó, bài toán lựa chọn câu trả lời là bài toán quan trọng và là bài toán chính của hệ thống hỏi đáp. Một người dùng có thể đăng câu hỏi và mong muốn nhận được các câu trả lời từ người dùng khác. Trong một số forum, nhiều câu hỏi có thể có hàng trăm câu trả lời (ví dụ như hệ thống Yahoo!answer). Do đó người dùng rất mất thời gian khi đọc tất cả câu trả lời đó và đánh giá từng câu trả lời một. Hơn nữa, những câu hỏi có nội dung đặc thù đặc biệt người bình thường không phải chuyên gia trong lĩnh vực đó khó có thể phân biệt được câu trả lời đúng hay sai. Vì những lý do này, việc xây dựng một công cụ tự động đánh giá câu trả lời tốt hay không tốt là một công việc rất cần thiết.

Ví dụ 1:
Subject: Nationalities banned in Qatar
Question: Hello! Can you help me, is there anyone knows the list of nationalities who are banned and cannot apply employment visa in Qatar?
Answer(good): Pakistanis are facing severe problems. There is no ban on Visa but it is very hard, near impossible to get.
Answer(bad): I just want to know because, our company will move in Qatar and we will be hiring some staff like Pakistani, Egyptian, Jordan etc...Im currently here in the UAE, and Bangladeshi, Syrian and Egyptian are not issuing employment visa now, I just want to know is Qatar also are banning this nationals?
Answer(bad): Hi are you suspecting your nationality.:)
Answer(bad):@khalli &Swift Unlock- are you twin?
*Answer(good):*Yup Pakistanis are banned becoz Qatar people are brilliant enough to do that... I wonder why USA and Uk doesn't have that kind of brains :).
Answer(good): There is no offical ban on any nationality as that would cause diplomatic problems, however most of the nationalities you have mentioned are near impossible to get work visas for.
Answer(bad): What work do you want to start here?
Answer(bad): NOTA- are correct answer nor the Quota- its Wasta.
Answer(bad): Zafir, There were 7 sets of twins of Khalli till yesterday.

Hình 1. *Ví dụ về câu hỏi và các câu trả lời trong tập dữ liệu Semeval 2017*

Bài toán lựa chọn câu trả lời được phát biểu như sau: Cho một câu hỏi q và các câu trả lời ứng viên $a_1, a_2, ..., a_n$. Chúng ta cần phải xác định xem các câu trả lời đó câu nào đúng. Đây là một bài toán rất quan trọng và được nhiều nhà nghiên cứu quan tâm [1], [3]-[5]. Với bài toán này thách thức lớn nhất là vấn đề khoảng cách từ vựng. Khoảng cách từ vựng là sự sai khác giữa từ vựng của câu hỏi và câu trả lời. Ngoài sự khác nhau về từ vựng trong câu hỏi và câu trả lời, độ dài của câu hỏi và câu trả lời cũng lệch nhau. Câu hỏi và câu trả lời lại chứa nhiều câu. Một lý do dẫn tới thách thức về khoảng cách từ vựng nữa là ngôn ngữ dùng trong các forum ở dạng văn nói. Nhiều câu hỏi và câu trả lời chứa nhiều thông tin dư thừa, không đề cập trực tiếp tới nội dung chính của câu hỏi và câu trả lời như lời chào hỏi, biểu tượng cảm xúc, từ viết tắt, viết sai chính tả. Những nguyên nhân này gây khó khăn cho mô hình dự đoán câu trả lời đúng. Hình 1 là ví dụ về cặp câu hỏi và câu trả lời minh họa các thách thức này trong tập dữ liệu SemEval 2017. Trong ví dụ 1, câu hỏi chứa phần dư thừa không liên quan tới nội dung chính như là "Hello, Can you help me". Hơn nữa, trong ví dụ còn chứa nhiều biểu tượng cảm xúc như ':0', ':)', ':P', các từ viết tắt. Trong ví dụ thứ hai, câu hỏi chứa nhiều câu và nhiều ý hỏi.

Các nghiên cứu gần đây sử dụng phương pháp tiếp cận dựa vào mạng học sâu và cơ chế sự chú ý để giải quyết bài toán tìm câu trả lời đúng mà không cần sử dụng các kỹ thuật trích rút đặc trưng đặc biệt hoặc sử dụng thêm nguồn tri thức bên ngoài [2], [6]. Các phương pháp này hướng tới việc tìm ra những từ mang thông tin quan trọng của câu hỏi và câu trả lời.

Trong những năm gần đây, nhiều nghiên cứu đã chỉ ra rằng, cơ chế sự chú ý mang lại thành tựu to lớn trong các bài toán NLP như dịch máy, suy diễn ngôn ngữ, đọc hiểu và hỏi đáp [4]. Hơn nữa, thông qua việc học trọng số sự chú ý của các từ và cụm từ trong câu thì trọng số của cụm từ dư thừa và nhiễu thường có trọng số nhỏ. Điều này dẫn tới mức độ ảnh hưởng của những phần này tới toàn bộ ngữ nghĩa của câu không còn đáng kể. Do đó, ngữ nghĩa của câu chỉ tập trung vào những từ và cụm từ quan trọng mà liên quan trực tiếp tới nội dung của câu hỏi và câu trả lời. Vì vậy, mạng học sâu dựa vào cơ chế sự chú ý là sự lựa chọn phù hợp với dữ liệu văn bản trong hệ thống hỏi đáp cộng đồng.

Trong khi rất nhiều nghiên cứu đã công nhận hiệu quả của các mô hình mạng học sâu trong bài toán lựa chọn câu trả lời nhưng chưa có đánh giá tổng hợp cụ thể nào về các mô hình học sâu ứng dụng trong bài toán này [6]-[8]. Trong bài báo này, chúng tôi tiến hành tổng hợp và phân nhóm một số mô hình điển hình đã đề xuất giải quyết bài toán lựa chọn câu trả lời. Đồng thời chúng tôi chọn ra một số mô hình học sâu điển hình để trình bày cụ thể cách sử dụng các mô hình này vào bài toán lựa chọn câu trả lời. Qua đó, chúng tôi đề xuất các hướng nghiên cứu trong tương lai.

2. Các phương pháp

Bài toán lựa chọn câu trả lời là bài toán cốt lõi và được nghiên cứu nhiều nhất trong hệ thống hỏi đáp cộng đồng. Quá trình nghiên cứu về bài toán này có thể gồm 3 giai đoạn: Giai đoạn sử dụng các đặc trưng của từ vựng, giai đoạn tiếp theo sử dụng đặc trưng kỹ thuật và giai đoạn thứ 3 là giai đoạn sử dụng mạng nơron học sâu và cơ chế sự chú ý.

Trong giai đoạn đầu các nghiên cứu sử dụng sự trùng lặp giữa câu hỏi và câu trả lời. Trong phương pháp này, câu trả lời tốt nhất được lựa chọn dựa vào so sánh từ trùng nhau giữa câu hỏi và câu trả lời. Phương pháp túi từ Bag-of-word và túi n-gram (Bag-of-Ngram) [5] được sử dụng phổ biến trong giai đoạn đầu. Ngoài ra một số phương pháp cũng sử dụng đặc trưng về trọng số của túi từ. Tuy nhiên, những phương pháp này được chỉ ra là không hợp lý. Điểm yếu nhất của những phương pháp này đó là không sử dụng đặc trưng ngữ nghĩa và đặc trưng ngôn ngữ của câu. Để khắc phục nhược điểm này một số nghiên cứu sử dụng mạng ngữ nghĩa Wordnet để giải quyết thách thức về ngữ nghĩa. Tuy nhiên, phương pháp này có hạn chế về ngôn ngữ vì một số từ không có trong nguồn từ vựng Wordnet [6].

Trong giai đoạn thứ hai, các nghiên cứu cố gắng đưa các đặc trưng kỹ thuật sử dụng cấu trúc cú pháp và ngữ nghĩa của câu. Cây phụ thuộc được sử dụng để biểu diễn câu hỏi và các câu trả lời ứng viên, đồng thời tích hợp thông tin ngữ nghĩa như sử dụng thực thể có tên vào biểu diễn này. Nghiên cứu khác gần đây lại sử dụng cây phụ thuộc và thuật toán khoảng cách sửa cây trong bài toán lựa chọn câu trả lời [7]. Ngoài ra các đặc trưng này được sử dụng đưa vào mô hình học sâu như CNN, mô hình RNN [7]. Trong cuộc thi SemEval CQA 2017 [2], các đội đứng đầu khai thác rất nhiều đặc trưng như cây phụ thuộc, độ tương tự và nhiều đặc trưng đặc biệt khác.

Giai đoạn thứ 3 là giai đoạn phát triển nhất khi giải quyết bài toán lựa chọn câu trả lời trong hệ thống hỏi đáp vì hiệu suất của mô hình được cải thiện lớn hơn hẳn những giai đoạn trước. Giai đoạn này gọi là giai đoạn bùng nổ về số lượng các nghiên cứu về AI cùng với mô hình học sâu mạng nơron mà nó loại bỏ việc sử dụng các đặc trưng kỹ thuật được trích rút thủ công. Với số lượng nghiên cứu lớn trên các bài toán về QA, các nhà nghiên cứu đã chia thành 5 nhóm chính: Nhóm dựa trên Siamese, nhóm dựa vào cơ chế sự chú ý, nhóm dựa vào so sánh tổng hợp, nhóm dùng mô hình ngôn ngữ và nhóm gồm các kiến trúc đặc biệt cho bài toán hỏi đáp.

2.1. Các mô hình dựa vào kiến trúc Siamese

Những mô hình dựa vào mạng Siamese là những mô hình theo cấu trúc mạng Siamese. Những mô hình này sẽ xử lý câu hỏi và câu trả lời một cách độc lập và học ra biểu diễn của chúng. Trong quá trình xử lý thông tin của câu khác không ảnh hưởng đến quá trình này của mỗi câu [3]. Yu và cộng sự [8] là mô hình đầu tiên sử dụng mạng nơron vào giải quyết bài toán lựa chọn câu trả lời. Mô hình này sử dụng mạng CNN và hồi quy logistic vào việc lựa chọn câu trả lời liên quan nhất với câu hỏi. Feng và cộng sự sử dụng mô hình của Yu với việc kết hợp sử dụng mạng nơron sâu với lớp kết nối đầy đủ (fully-connected). Trong mô hình này các lớp ẩn khác nhau, các phép toán tích chập, pooling với các hàm kích hoạt khác nhau được sử dụng để thăm dò ảnh hưởng của các yếu tố này. Tuy nhiên, các mô hình này được tính toán một cách độc lập và đánh giá riêng biệt. He và cộng sự [2] đã đề xuất mô hình kết hợp nhiều khía cạnh của mô hình hóa độ tương tự câu vào một mô hình duy nhất và cuối cùng đưa ra véctơ biểu diễn cho từng câu.

Các mô hình học sâu được nghiên cứu và sử dụng rộng rãi trong các bài toán này. Yu và cộng sự [8] đã đề xuất mô hình Convolutional Bigram để phân lớp câu trả lời ứng viên thành lớp câu hỏi đúng và câu hỏi sai. Tan và cộng sự [9] đã sử dụng mô hình attentive-biLSTM để tính trọng số sự chú ý, sau đó tổng hợp ngữ nghĩa dựa vào độ liên quan của các đoạn trong câu trả lời với câu hỏi. Madabushi và cộng sự [10] đã cung cấp giải pháp cho bước tiền xử lý thay vì cải tiến mô hình. Trong mô hình này các thực thể được gán tên trong các câu trả lời ứng viên được chuyển thành những từ đặc biệt giúp cho mô hình tìm kiếm câu trả lời phù hợp một cách dễ dàng nhất. Quá trình này cũng được ứng dụng vào mô hình của Rao và cộng sự [2] và nghiên cứu này cũng đã xác nhận hiệu quả của quá trình này [2].

2.2. Mô hình mạng nơron dựa vào cơ chế sự chú ý ứng dụng vào bài toán lựa chọn câu trả lời

Không giống như mô hình siamese, mô hình dựa vào cơ chế sự chú ý sử dụng sự tương tác ngữ cảnh giữa các câu để đạt được thông tin tương tác giữa câu hỏi và câu trả lời. Cơ chế sự chú ý đầu tiên được sử dụng trong dịch máy, sau đó được áp dụng sang các bài toán khác của NLP như hỏi đáp và lựa chọn câu trả lời [3]. Cơ chế sự chú ý của Bahdanau được sử dụng trên mạng RNN đã vượt qua được hiệu năng của bài toán lựa chọn câu trả lời vào thời điểm đó. He và cộng sự [11] cũng đã sử dụng cơ chế sự chú ý này kết hợp với mạng CNN. Mô hình này chứng minh rằng khi cơ chế sự chú ý này kết hợp với CNN cho kết quả tốt hơn so với khi kết hợp với mạng RNN. Sau thành công của cơ chế sự chú ý, Tan [9] đã đề xuất để gióng các từ liên quan của câu hỏi với câu trả lời. Do câu hỏi và câu trả lời có nhiều nhiễu nên làm cho thông tin quan trọng của chúng bị phân tán, điều đó gây khó khăn cho việc dự đoán câu trả lời đúng. Cũng có những nghiên cứu tận dụng thông tin bổ sung để bù đắp sự mất cân bằng giữa câu hỏi và câu trả lời như sử dụng mô hình người dùng, sử dụng mô hình chủ đề, sử dụng tri thức bên ngoài từ đồ thị tri thức để làm giàu học biểu diễn của câu hỏi.

2.3. Các mô hình dựa trên so sánh - tổng hợp

Mô hình dựa vào cơ chế so sánh - tổng hợp cũng tập trung vào tương tác ngữ cảnh giữa các câu như mô hình sự chú ý nhưng mức độ tương tác nhiều hơn. Những mô hình này ban đầu thường là so sánh ở mức từ để đạt được nhiều thông tin, sau đó tích hợp thông tin so sánh ở mức từ với véctơ biểu diễn ở mức câu [3]. Trong mô hình của He và cộng sự là mô hình đầu tiên sử dụng cơ chế so sánh - tổng hợp để cải tiến chất lượng của bài toán lựa chọn câu trả lời. Thay vì sử dụng biểu diễn câu đầu vào sang dạng biểu diễn một véctơ và tính độ tương tự của hai câu, tác giả đã thực hiện tương tác giữa các cặp từ với nhau để học biểu diễn của các câu đầu vào qua việc tổng hợp các giá trị này. Một nghiên cứu của Bian [1] đã bổ sung thêm một kỹ thuật sự chú ý động vào mô hình so sánh - tổng hợp. Kỹ thuật mới này giúp lọc nhiễu trong ma trận sự chú ý, đồng thời giúp khai thác ngữ nghĩa tốt hơn ở cấp độ từ và làm cho mô hình học ra biểu diễn câu tốt hơn. Mô hình Shen đề xuất một lớp liên trọng số và cố thiết lập trọng số của mỗi từ.

3. Tập dữ liệu

Trong phần này chúng tôi trình bày một số tập dữ liệu được sử dụng để đánh giá các mô hình đề xuất trong các nghiên cứu gần đây. Bảng 1 dưới đây thống kê một số tập dữ liệu được dùng để đánh giá các mô hình trong bài toán của hệ thống hỏi đáp.

Bảng 1. *Bảng thống kê một số tập dữ liệu sử dụng trong các bài toán của hệ thống hỏi đáp cộng đồng*

	Train	DeV	Test	Tổng
Yahoo!answer				87.390 câu hỏi và 414.446 câu trả lời
Trec- QA	1229	80	100	1409 cặp câu hỏi – câu trả lời
Quora				404.289 cặp câu hỏi
SemEval 2017	267	50	88	405 câu hỏi gốc và 4050 câu trả lời

Yahoo!webscope: Dữ liệu được thu thập từ trang hỏi đáp Yahoo!answer với đa dạng các thể loại. Đây là tập dữ liệu rất giàu thông tin chưa được gán nhãn bao gồm 87.390 câu hỏi và 314.446 câu trả lời. Tập dữ liệu này chứa rất nhiều thông tin hữu ích cho việc nghiên cứu trên các bài toán của CQA như chủ đề câu hỏi, nội dung câu hỏi, mô tả chi tiết của câu hỏi, câu trả lời tốt nhất do người hỏi chọn và các câu trả lời khác cho câu hỏi đó. Các thông tin khác liên quan tới người hỏi, thời gian hỏi và trả lời, ngày bình chọn cho câu trả lời.

Trec-QA: Tập TREC-QA bao gồm 1409 cặp câu hỏi - câu trả lời được chia thành 1229, 80 và 100 cặp câu tương ứng với ba tập: Tập huấn luyện, tập phát triển và tập kiểm thử. Tập này chứa các cặp câu hỏi factoid và câu trả lời của nó. Câu hỏi factoid là câu hỏi ngắn gọn và thường chứa từ để hỏi như *what, where, when, who*. Trong tập này mỗi câu hỏi chỉ có một câu trả lời và được gán nhãn POS, NER và phân tích câu phụ thuộc.

Quora: Đây là tập dữ liệu được công bố trong cuộc thi Kaggle (https://www.kaggle.com/c/quora-question-pairs/data). Tập dữ liệu này được thu thập từ trang hỏi đáp Quora.com về các lĩnh vực trong cuộc sống hay công việc hàng ngày. Nó bao gồm các câu hỏi được gán nhãn duplicate và non-duplicate phục vụ cho bài toán tìm câu hỏi tương đồng. Trong 404351 cặp câu hỏi có 149306 cặp câu có nhãn positive và 255,045 cặp câu có nhãn negative.

SemEval: Tập này được thu thập từ forum hỏi đáp chia sẻ mọi thứ liên quan tới công việc ở Qatar (https://www.qatarliving.com/forum). Chủ đề ở đây cũng rất phong phú và đa dạng với nhiều lĩnh vực. Đây là tập dữ liệu được công bố trong Workshop đánh giá về mặt ngữ nghĩa (http://alt.qcri.org/semeval2017/task3/index.php?id=data-and-tools). Từ khía cạnh ngôn ngữ, tập dữ liệu này rất có giá trị và thách thức. Tập dữ liệu này chứa nhiều đặc trưng của văn bản web như URLs, biểu tượng cảm xúc, địa chỉ email, lỗi sai chính tả, kí hiệu viết tắt. Forum sử dụng ngôn ngữ tiếng Anh và là nơi trao đổi, cung cấp mọi thông tin về Qatar cho mọi người mới sống và có ý định tới sống ở đây. Do không phải là người bản ngữ dùng tiếng Anh nên câu có nhiều lỗi về mặt ngữ pháp, nhiều từ không phổ biến hoặc những từ không tồn tại.

Workshop Semeval được tổ chức hàng năm với sự tham gia của nhiều đội tuyển. Tập dữ liệu cụ thể công bố đến năm 2017. Tập dữ liệu này cũng được chia làm ba tập: train, dev và test chứa các câu hỏi và các câu trả lời của nó. Với mỗi câu hỏi gốc có 10 câu hỏi liên quan (được lấy qua máy tìm kiếm) được gán ba nhãn: Perfect match, Relevant và Irrelevant. Với mỗi câu hỏi gốc có 10 câu trả lời được gán ba nhãn Good, Bad, Potentially useful. Mỗi câu hỏi liên quan lại có 10 câu trả lời cũng được gán ba nhãn như trên.

Khác biệt lớn nhất giữa tập Trec-QA và các tập dữ liệu còn lại đó là về đặt trưng ngôn ngữ. Tập dữ liệu TREC-QA là tập dữ liệu với ngôn ngữ tiếng Anh chuẩn. Các câu hỏi chủ yếu là câu hỏi factoid và các câu hỏi thường ngắn gọn không mô tả được hết những thách thức của hệ thống hỏi đáp cộng đồng. Trong khi đó, tập dữ liệu khác như Yahoo!answer, Quora, SemEval ngôn ngữ dùng là ngôn ngữ nói. Đặc biệt hơn, tập SemEval đôi khi người dùng còn dùng ngôn ngữ khác không phải tiếng Anh. Ngoài ra các tập dữ liệu như Yahoo!answer và Quora lại không chia thành các tập huấn luyện, tập phát triển và kiểm thử chuẩn. Vì mỗi bài báo lại chia tập dữ liệu thử nghiệm khác nhau nên các phương pháp được đề xuất khó so sánh với nhau. Khác biệt thứ hai là

các câu hỏi trong tập CQA chứa nhiều câu hỏi mở với nhiều lĩnh vực khác nhau, còn tập TREC-QA chứa nhiều các câu hỏi factoid có nội dung ngắn gọn và rõ ràng. Khác biệt thứ 3 giữa tập dữ liệu CQA và QA là các tập CQA thường có lượng dữ liệu lớn hơn nhiều so với TREC-QA. Khác biệt cuối cùng đó là trong các tập dữ liệu CQA, tập dữ liệu SemEval có sẵn công cụ đánh giá chuẩn và được công khai, trong khi các tập dữ liệu khác kịch bản đánh giá không được thống nhất. Hơn nữa, vì tập dữ liệu Semeval này chứa nhiều miền dữ liệu nên khi sử dụng vào các mô hình có thể dễ dàng cho việc điều chỉnh và chuyển đổi miền sử dụng.

Khó khăn trong nghiên cứu các bài toán trên hệ thống CQA là không có tập dữ liệu chuẩn để so sánh các phương pháp với nhau. Các bảng 2 và bảng 3 là các thống kê kết quả của một số mô hình đã được đề xuất và thực hiện trên các tập dữ liệu trên một nghiên cứu tổng hợp trong bài báo. Nhiều nhà nghiên cứu sử dụng tập dữ liệu được lấy từ Yahoo!answer nhưng các tập dữ liệu huấn luyện, tập phát triển và tập kiểm thử lại khác nhau, không cố định và không công bố công khai. Trong khi nhiều tác giả lại công bố nghiên cứu của mình trên tập TREC-QA nhưng tập dữ liệu chỉ chứa các câu hỏi factoid. Trong khi câu hỏi trên CQA là những câu hỏi phức tạp và dài, nhiễu. Vì vậy, khó khăn của việc nghiên cứu trên bài toán lựa chọn câu trả lời là không có tập dữ liệu chuẩn để thử nghiệm đánh giá chung cho các mô hình được đề xuất. Mỗi mô hình lại phù hợp với từng tập dữ liệu riêng có đặc trưng ngôn ngữ riêng.

Bảng 2. *Bảng kết quả MAP và MRR của một số mô hình học sâu trên tập dữ liệu TrecQA*

Mô hình	MAP	MRR
Bigram+Word count+CNN	71,13	78,46
Embedding+CNN+Max pooling	71,06	79,98
QA-LSTM	68,19	76,52
QA-LSTM/CNN	70,61	81,04
QA-LSTM attention	68,96	78,49
QA-LSTM/CNN attention	72,79	82,40

Bảng 3. *Bảng kết quả P@1 của một số mô hình trên tập Yahoo!answer*

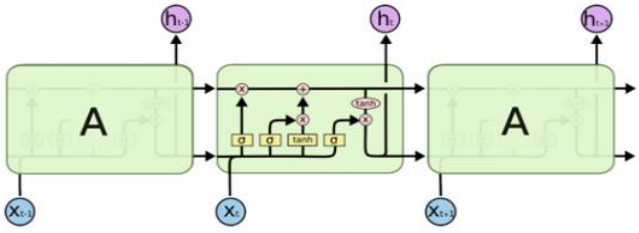
Mô hình	P@1
OKapi BM25	35,6
TransLM	48,5
BOW embeddings	66,8
CNN_MLP	68,5

4. Một số mô hình điển hình

Trong phần này, chúng tôi lựa chọn ra một số mô hình đại diện cho các nhóm mô hình được trình bày tại mục 2 để mô tả kĩ hơn kiến trúc của các mô hình này khi áp dụng vào bài toán lựa chọn câu trả lời.

4.1. Mô hình LSTM

Mô hình LSTM được đề xuất bởi Hochreiter và Schmidhuber vào năm 1997 để khắc phục nhược điểm của mô hình RNN. Mô hình LSTM như hình 2.



Hình 2. *Mô hình LSTM [9]*

Mạng LSTM (Long Short-Term Memory) bao gồm nhiều tế bào LSTM liên kết với nhau thay vì chỉ tương tác với nhau qua đơn vị tầng ẩn như mạng RNN. LSTM bao gồm trạng thái tế bào

giống như băng truyền chạy xuyên suốt các nút mạng. Do đó, các thông tin được truyền đi dễ dàng thông suốt. LSTM có khả năng bỏ đi hoặc thêm các thông tin cho trạng thái tế bào thông qua các nhóm gọi là cổng. Cổng là nơi sàng lọc thông tin đi qua nó thông qua phép toán *sigmoid* và phép nhân. Các phương trình lan truyền trong mạng LSTM như sau:

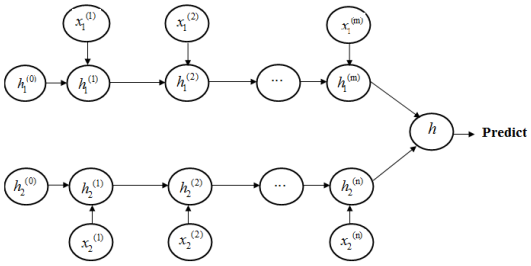
$$\begin{aligned} i_k &= \sigma\left(\mathrm{W}^i x_k + V^i h_{k-1} + b^i\right), \\ f_k &= \sigma\left(\mathrm{W}^f x_k + V^f h_{k-1} + b^f\right), \\ o_k &= \sigma\left(\mathrm{W}^o x_k + V^o h_{k-1} + b^f\right), \\ c_k &= f_k \odot c_{k-1} + i_k \odot \tanh\left(\mathrm{W}^c x_k + V^c h_{k-1} + b^c\right) \\ h_k &= o_k \odot \tanh\left(c_k\right) \end{aligned} \quad (1)$$

Trong đó: **i, f, o** là cổng vào, cổng quên và cổng ra tương ứng, ma trận **W**, **V** và **b** là ma trận học từ mô hình.

Véctơ $\mathbf{c}_k$ là bộ nhớ trong của đơn vị. Nó là sự kết hợp của bộ nhớ trước đó và đầu vào mới. Chúng ta có thể chọn bỏ qua hoàn toàn bộ nhớ cũ (cổng quên bằng 0) hoặc bỏ qua hoàn toàn trạng thái mới được tính toán (cổng đầu vào bằng 0), hoặc một giá trị ở giữa hai thái cực này.

Mạng bộ nhớ ngắn hạn hướng dài hạn đã chứng tỏ khả năng khắc phục hạn chế vấn đề phụ thuộc dài của mình qua nhiều thử nghiệm thực tế, giải quyết một số bài toán trong học máy nói chung và trong xử lý ngôn ngữ tự nhiên nói riêng.

Mô hình LSTM được ứng dụng vào bài toán lựa chọn câu trả lời như sau: Cho câu hỏi và câu trả lời đi qua hai đường LSTM như hình 3. Sau đó véctơ ẩn cuối cùng $h_1^{(m)}$ và $h_2^{(n)}$ được nối lại và đi qua hàm sofmax để dự đoán. Bài toán lựa chọn câu trả lời được đưa về bài toán phân lớp nhị phân.



Hình 3. *Mô hình siamese sử dụng LSTM cho bài toán lựa chọn câu trả lời*

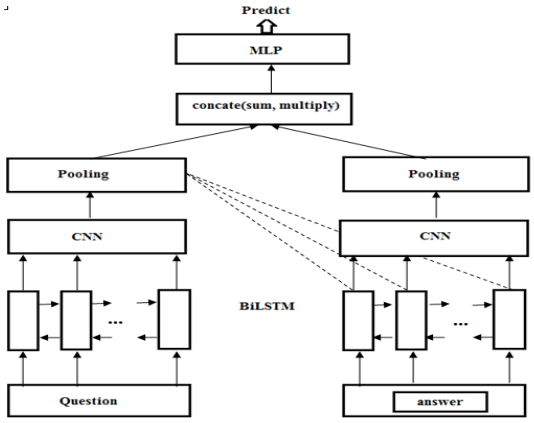
4.2. Mô hình LSTM/CNN attention

Trong mô hình này, đầu ra của hai câu hỏi sau khi đưa qua mô hình LSTM và CNN sẽ được sử dụng để tính ma trận trọng số sự chú ý từ với từ. Sau đó biểu diễn từ của câu thứ hai sẽ được cập nhật lại qua trọng số sự chú ý. Cuối cùng, phép toán tổng hợp lớn nhất (max pooling) được sử dụng để thu thập các đặc trưng quan trọng trước khi đưa vào lớp dự đoán. Mô hình này (hình 4) gần giống với mô hình của Tan và cộng sự [9]. Trong đó, công thức tính trọng số chú ý như sau:

$$m_{a,q}\left(t\right) = \mathrm{W}_{am} h_a(\mathrm{t}) + \mathrm{W}_{qm} o_q \quad (2)$$

$$s_{a,q}(t) \propto exp\left(\mathrm{w}_{ms}^T \tanh\left(m_{a,q}\left(t\right)\right)\right) \quad (3)$$

$$\tilde{h}_a\left(t\right) = h_a\left(t\right) s_{a,q}\left(t\right) \quad (4)$$



Hình 4. *Mô hình LSTM/CNN attention cho bài toán lựa chọn câu trả lời*

4.3. Mô hình tổng hợp so sánh

Mô hình match-LSTM làm mô hình được lựa chọn để mô tả về phương pháp tổng hợp so sánh áp dụng vào bài toán lựa chọn câu trả lời. Mô hình này được đề xuất cho bài toán suy diễn ngôn ngữ. Sau đó mô hình được áp dụng vào bài toán lựa chọn câu tra lời [12]. Mô hình bao gồm 5 lớp:

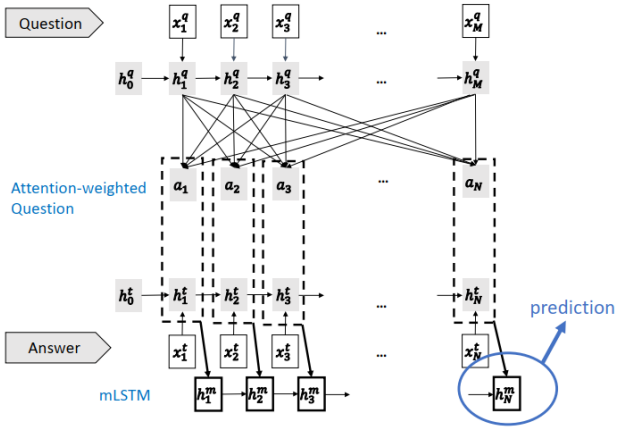
- Lớp biểu diễn từ: Mục đích của lớp này là học biểu diễn mỗi từ trong câu sang không gian có số chiều cố định sử dụng mô hình Glove.
- Lớp biểu diễn theo ngữ cảnh: Câu hỏi và câu trả lời đưa qua hai đường LSTM để cập nhật biểu diễn từ trong câu theo ngữ cảnh.
- Lớp matching: Trong mô hình so sánh từng từ cập nhật theo ngữ cảnh của câu trả lời với các từ trong câu hỏi qua việc tính trọng số và véctơ sự chú ý theo công thức sau [12]:

$$e_{kj} = \mathrm{w}^{e} \cdot \tanh\left(\mathrm{W}^{q} h_{j}^{q} + \mathrm{W}^{t} h_{k}^{t} + \mathrm{W}^{m} h_{k-1}^{m}\right) \qquad (5)$$

$$\alpha_{kj} = \frac{exp\left(e_{kj}\right)}{\sum_{j'=1}^{M} exp\left(e_{kj'}\right)} \qquad (6)$$

$$a_{k} = \sum_{j=1}^{M} a_{kj} h_{j'}^{q} \qquad (7)$$

- Tiếp theo là lớp tổng hợp. Lớp này làm nhiệm vụ tổng hợp so sánh ở bước trên qua đường mLSTM sang không gian véctơ với số chiều cố đinh.
- Cuối cùng là lớp dự đoán. Mô hình sử dụng biểu diễn của lớp ẩn cuối cùng của bước trên trong mô hình mLSTM dùng để dự đoán bằng hàm softmax.



Hình 5. *Mô hình match-LSTM* [12]

5. Thảo luận và hướng phát triển

Mục đích của bài báo nhằm tổng hợp một số kiến trúc về mô hình học sâu áp dụng vào bài toán lựa chọn câu trả lời trong hệ thống hỏi đáp bao gồm các kiến trúc Siamese, kiến trúc học sâu với cơ chế chú ý và kiến trúc so sánh tổng hợp. Qua ba kiến trúc này, chúng tôi trình bày 3 mô hình học sâu tương ứng để làm rõ cách áp dụng vào bài toán lựa chọn câu trả lời.

Như trình bày ở phần 3 về dữ liệu thử nghiệm trên bài toán CQA, mỗi tập dữ liệu có những đặc trưng ngôn ngữ riêng. Các nhóm mô hình đề xuất để giải quyết bài toán này cũng được áp dụng trên tập dữ liệu khác nhau. Do đó khó có thể đánh giá một cách đầy đủ và toàn diện các mô hình trên. Từ các phân tích trên, chúng tôi đề xuất hướng nghiên cứu bài toán trong tương lai:

- Xây dựng tập dữ liệu chuẩn đủ lớn mang đầy đủ thách thức của bài toán tìm câu trả lời đúng trong hệ thống hỏi đáp cộng đồng.
- Cài đặt thử nghiệm và đánh giá toàn diện và đầy đủ các mô hình học sâu điển hình trên các tập dữ liệu khác nhau; từ đó thấy được ưu nhược điểm của từng mô hình trên.
- Các mô hình đề xuất chỉ được đánh giá trên tập dữ liệu tiếng Anh mà chưa có đánh giá trên tập dữ liệu tiếng Việt.

Lời cám ơn

Chúng tôi xin cảm ơn đề tài có mã số T2021-07-03 đã hỗ trợ một phần kinh phí để chúng tôi thực hiện công việc này.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] W. Bian, S. Li, Z. Yang, G. Chen, and Z. Lin, "A Compare-Aggregate Model with Dynamic-Clip Attention for Answer Selection," *CIKM*, New York – NY - USA, 2017, pp. 1987-1990.

[2] H. He, J.Wieting, K. Gimpel, J. Rao, and J. Lin, "Attention- based multi-perspective convolutional neural networks for textual similarity measurement," *The Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval- 2016)*, San Diego - California, 2016, pp. 1103-1108.

[3] T. M. Lai, T. Bui, and S. Li, "A Review on Deep Learning Techniques Applied to Answer Selection," *COLING*, Santa Fe - New Mexico - USA, 2018, pp. 2132-2144.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, Minneapolis - Minnesota - USA, 2019, pp. 4171-4186.

[5] S. Wan, M. Dras, R. Dale, and C. Paris, "Using dependency-based features to take the para-farce out of paraphrase," *The Proceedings of the Australasian Language Technology Workshop 2006*, Sydney - Australia, 2006, pp. 131-138.

[6] Yi, Liang and Wang, JianXiang and Lan, Man, "ECNU: Using Multiple Sources of CQA-based Information for Answers Selection and Response Inference", *The proceedings of the 9th International Workshop on Semantic Evaluation SemEval*, Denver, Colorado, 2015, pp.236--241.

[7] M. Wang and C. D. Manning, "Probabilistic Tree-Edit Models with Structured Latent Variables for Textual Entailment and Question Answering," *The COLING*, Beijing - China, 2010, pp. 1164-1172.

[8] L. Yu, & K. M. Hermann, P. Blunsom, and S. Pulman, "Deep Learning for Answer Sentence Selection", 2014. [Online]. Available: https://arxiv.org/abs/1412.1632. [Accessed May 2021].

[9] M. Tan, B. Xiang, and B. Zhou, "LSTM-based Deep Learning Models for non-factoid answer selection," 2015. [Online]. Available: https://arxiv.org/abs/1511.04108. [Accessed May 2021].

[10] H. T. Madabushi, M. Lee, and J. Barnden, "Integrating Question Classification and Deep Learning for improved Answer Selection," *COLING 2018*, Santa Fe - New Mexico - USA, 2018, pp. 3283-3294

[11] H. He, K. Gimpel, and J. Lin, "Multi-perspective sentence similarity modeling with convolutional neural networks," *EMNLP*, Lisbon - Portugal, 2015, pp. 1576-1586.

[12] T. T. Ha, A. Takasu, T. C. Nguyen, K. H. Nguyen, V. N. Nguyen, K. A. Nguyen, and S. G. Tran, "Supervised attention for answer selection in community question answering," *IJAI*, vol 9, no. 2, pp. 203-11, 2020.