HỌC BIỂU DIỄN CÂU SỬ DỤNG MÔ HÌNH LSTM TRONG BÀI TOÁN TÌM KIẾM CÂU HỎI

Đinh Khánh Linh*, Trần Quang Huy

Trường Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên, Việt Nam

ARTICLE INFORMATION

Journal: *Vinh University Journal of Science Natural Science, Engineering and Technology*
p-ISSN*: 3030-4563*
e-ISSN: *3030-4180*

Volume: 53
Issue: 3A
****Correspondence:***
dklinh@ictu.edu.vn

Received: *08 May 2024*
Accepted: *27 June 2024*
Published: *20 September 2024*

Citation:
Dinh Khanh Linh, Tran Quang Huy (2024). Sentence representation using LSTM for finding question.
Vinh Uni. J. Sci.
Vol. 53 (3A), pp. 16-22
doi: 10.56824/vujs.2024a063a

TÓM TẮT

Học biểu diễn câu mang đầy đủ ngữ nghĩa của văn bản là thách thức trong các bài toán xử lý ngôn ngữ tự nhiên bởi vì nếu véc tơ biểu diễn ngữ nghĩa của câu tốt thì sẽ làm tăng hiệu năng của các bài toán dự đoán. Trong bài báo này, chúng tôi đề xuất thử nghiệm sử dụng mô hình LSTM với các cách trích rút biểu diễn câu khác nhau và áp dụng vào bài toán tìm câu hỏi tương đồng với mục đích khai thác ngữ nghĩa ẩn của câu. Các phương pháp này tổng hợp biểu diễn câu từ các lớp ẩn của mô hình LSTM. Kết quả chỉ ra rằng kỹ thuật tổng hợp biểu diễn câu dùng kết hợp cả Max Pooling và Mean Pooling cho kết quả cao nhất trên tập dữ liệu SemEval 2017 cho bài toán tìm câu hỏi tương đồng.

Từ khóa: LSTM; học sâu; xử lý ngôn ngữ tự nhiên; hệ thống hỏi đáp; học biểu diễn câu; hệ thống hỏi đáp cộng đồng.

1. Giới thiệu

Tìm câu hỏi tương đồng trong hệ thống hỏi đáp cộng đồng (CQA) là một trong những vấn đề nan giải trong xử lý ngôn ngữ tự nhiên. Nhiều diễn đàn web như Stack Overflow và Qatar Living đang trở nên phổ biến và linh hoạt để cung cấp thông tin cho người dùng [1]. Người dùng có thể đăng câu hỏi và có khả năng nhận được nhiều câu trả lời từ những người khác. Để người dùng có thể tự động nhận được câu trả lời từ những câu trả lời đã có trong cơ sở dữ liệu, bài toán tìm câu hỏi tương đồng đã được đặt ra. Đây là lý do cần thiết để xây dựng một công cụ tự động tìm các câu hỏi liên quan từ các câu hỏi mới.

Bài toán tìm kiếm câu hỏi liên quan được định nghĩa như sau: Cho một câu hỏi mới q và một tập các câu hỏi đã có trong kho dữ liệu $\{q_1, q_2, \dots, q_n\}$. Đầu ra yêu cầu trả về danh sách các câu hỏi tương đồng với q sao cho những câu hỏi liên quan nhất sẽ đứng trước những câu hỏi kém liên quan hơn.

Nghiên cứu [2] đã chỉ ra rằng thách thức lớn nhất của bài toán này là khoảng cách từ vựng. Điều đó có nghĩa là cách sử dụng các từ và cụm từ của câu hỏi thứ nhất khác so với từ và cụm từ của câu hỏi thứ hai mặc dù hai câu có cùng ý nghĩa. Dưới đây là ví dụ về hai câu hỏi được coi là tương đồng với nhau mặc dù cách sử dụng từ ngữ là khác nhau được lấy từ tập dữ liệu SemEval 2017 [3]-[4]:

Câu hỏi 1: Where can I buy good oil for massage?

Câu hỏi 2: Hi there, I can see a lot of massage center here, but I dont which one is better. Can someone help me which massage center is good… and how much will it cost me? Tks.

Hai câu hỏi này cùng một ý hỏi nhưng diễn giải khác nhau. Trong câu hỏi số 2 còn có nhiều nội dung giải thích cho câu hỏi và mang giọng điệu của dạng văn nói, có chứa nhiều từ viết tắt. Một thách thức chính của nhiệm vụ này nằm ở quan hệ ngữ nghĩa phức tạp và linh hoạt được quan sát giữa câu hỏi và câu hỏi đoạn văn. Trong ví dụ trên, câu hỏi 1 chỉ 08 từ, trong khi câu hỏi 2 sử dụng 34 từ để giải thích. Mặt khác, câu hỏi số 2 chứa một nhóm từ bao gồm thông tin không liên quan trực tiếp đến câu hỏi. Ngoài ra, trong khi một câu trả lời hay phải liên quan đến câu hỏi, chúng thường không chia sẻ các đơn vị từ vựng chung. Vấn đề này có thể gây nhầm lẫn cho các hệ thống kết hợp từ đơn giản. Do đó, những thách thức này làm cho các tính năng thủ công ít được mong đợi hơn nhiều so với phương pháp học sâu. Hơn nữa, các hệ thống kết hợp từ cũng cần học cách phân biệt các phần hữu ích với các phần không liên quan và tập trung nhiều hơn vào phần hữu ích.

Bài toán này thường được tiếp cận như một bài toán xếp hạng theo cặp, chiến lược tốt nhất để nắm bắt mối liên hệ giữa các câu hỏi đã có với câu hỏi mới vẫn còn là một vấn đề đang được nghiên cứu. Các phương pháp tiếp cận được thiết lập thường mắc phải điểm yếu sau: Đầu tiên, các nghiên cứu trước đây, chẳng hạn như [5]-[6] sử dụng mạng nơ-ron tích tụ (CNN) hoặc mạng nơ-ron lặp lại (RNN) tương ứng. Tuy nhiên CNN nhấn mạnh sự tương tác cục bộ trong n-gram, trong khi RNN được thiết kế để nắm bắt thông tin tầm xa và quên thông tin cục bộ không quan trọng qua véc tơ ẩn lớp cuối cùng.

Bài báo này đề xuất một phương pháp sử dụng các mô hình học máy thông dụng để giải quyết những điểm yếu trên. Nghiên cứu sẽ bắt đầu với mô hình Long Short-Term Memory (LSTM) cơ bản sử dụng véc tơ ẩn tại lớp cuối cùng để đưa ra biểu diễn câu. Sau đó, biểu diễn câu được tổng hợp bằng cách sử dụng các chiến lược Max Pooling và Mean Pooling qua các lớp ẩn trong mạng LSTM, và cuối cùng mô hình được đánh giá khi kết hợp cả hai đặc trưng Max và Mean Pooling.

2. Tổng quan về vấn đề nghiên cứu

Trong những năm gần đây, nhiều nghiên cứu liên quan đã được đề xuất để giải quyết bài toán tìm câu hỏi tương đồng và đạt được nhiều kết quả khả quan. Cụ thể như sau:

Công việc trước đây về bài toán tìm câu hỏi thường được sử dụng đặc trưng kỹ thuật, các công cụ về ngôn ngữ và tri thức từ bên ngoài. Ví dụ, các tính năng ngữ nghĩa được xây dựng dựa vào Wordnet [7]. Mô hình này ghép các từ liên quan đến ngữ nghĩa dựa trên quan hệ ngữ nghĩa của từ.

Trong hội nghị SemEval 2017, mô hình đứng đầu trong cuộc thi trên tập dữ liệu SemEval sử dụng các đặc trưng kỹ thuật rất phức tạp như thăm dò hàm nhân hoặc trích rút đặc trưng nhân cây từ việc đi phân tích các cây cú pháp [8]. Một nghiên cứu khác khai thác các đặc trưng độ tương tự khác nhau như độ đo Cosine, độ đo Euclidean về khoảng cách từ vựng, cú pháp và ngữ nghĩa [5] để biểu diễn câu học từ mô hình SVM.

Các nghiên cứu trên bài toán tìm câu trả lời [9]-[12] trong hệ thống CQA mang lại hiệu quả tốt hơn với việc sử dụng mạng nơ ron mà không cần phải sử dụng các đặc trưng được trích rút thủ công. Các mô hình này học ra biểu diễn câu, sau đó thực hiện đo độ tương tự của câu hỏi với câu hỏi và câu hỏi với câu trả lời [10].

Nghiên cứu này nhằm thử nghiệm mô hình LSTM cơ bản sử dụng véc tơ ẩn tại lớp cuối cùng để biểu diễn câu. Sau đó, các chiến lược Max Pooling và Mean Pooling sẽ được sử dụng để tổng hợp biểu diễn câu qua các lớp ẩn đó và đánh giá mô hình khi kết hợp cả hai đặc trưng trên.

3. Các mô hình đề xuất

3.1. Mô hình gốc LSTM

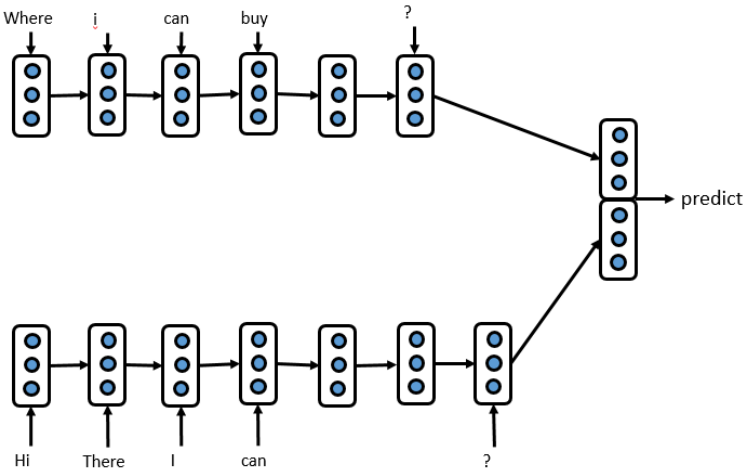
LSTM là một dạng mạng nơ-ron RNN đặc biệt dựa vào dữ liệu dạng chuỗi [13]. LSTM sử dụng một vài véc tơ cổng tại mỗi vị trí để kiểm soát việc truyền thông tin dọc theo trình tự và do đó cải thiện mô hình hóa các phụ thuộc phạm vi dài. Trong khi có các biến thể khác nhau của LSTM. $X = (x_1, x_2,..., x_N)$ được sử dụng để biểu thị một chuỗi đầu vào, trong đó $x_k \in R^L$ $(1 \le k \le N)$. Xác định véc tơ này được sử dụng cùng nhau để tạo ra một chiều d-chiều trạng thái ẩn h_k như sau [11]:

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

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.

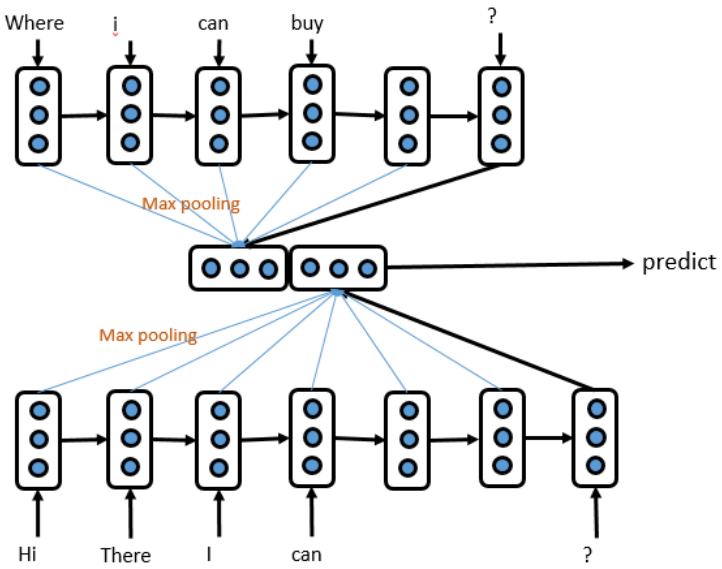
3.2. Các phương pháp biểu diễn câu

Hình 1 mô tả cách lấy biểu diễn câu sử dụng lớp ẩn cuối cùng trong bài toán tìm câu hỏi tương đồng.



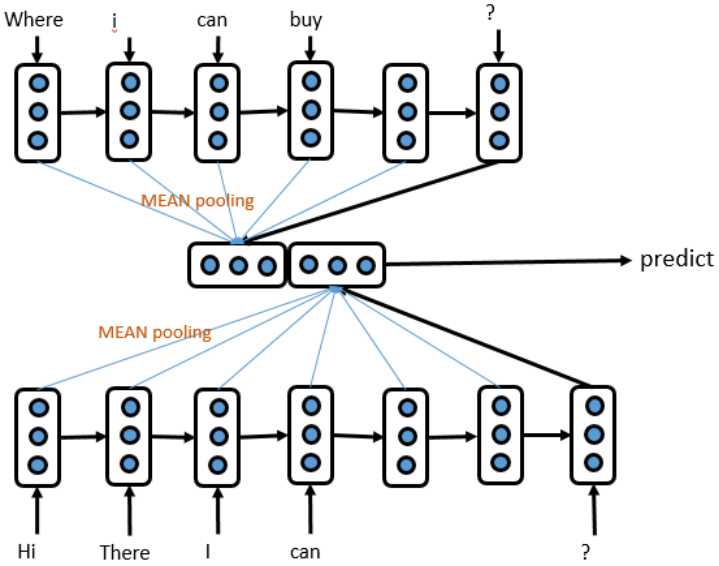
Hình 1: *Mô hình LSTM sử dụng véc tơ ẩn tại lớp cuối cùng dùng để biểu diễn câu*

Hình 2 mô tả phương pháp lấy biểu diễn câu sử dụng phép toán Max Pooling của các lớp ẩn. Max Pooling có nghĩa là thực hiện lấy giá trị lớn nhất của từng thành phần trong các lớp ẩn.



Hình 2: *Mô hình LSTM sử dụng phép toán Max Pooling để lấy biểu diễn câu*

Hình 3 dưới đây mô tả phương pháp lấy biểu diễn câu sử dụng phép toán Mean Pooling của các lớp ẩn. Mean Pooling là thực hiện tính giá trị trung bình của từng thành phần trong các lớp ẩn.



Hình 3: *Mô hình LSTM sử dụng phép toán Mean Pooling để lấy biểu diễn câu.*

Cuối cùng, hai kỹ thuật lấy Mean và Max kết hợp để đưa ra dự đoán câu.

Hàm mất mát là hàm cross entropy [14]:

$$L_{model} = -\frac{1}{S}\sum(y \log \hat{y} + (1-y)\log(1-\hat{y})) + \frac{\gamma}{2S}\|\boldsymbol{W}\|_2^2 \quad (2)$$

Trong đó, S là số lượng cặp câu hỏi trong tập huấn luyện, γ là tham số điều chỉnh của mô hình, **W** là bộ ma trận trọng số của mô hình.

4. Kết quả và thảo luận

4.1. Tập dữ liệu

Tập dữ liệu SemEval 2017 được sử dụng để đánh giá các mô hình đề xuất. Tập dữ liệu này được lấy từ diễn đàn Qatar living [11]. Đây là diễn đàn trao đổi về mọi vấn đề dành cho người nước ngoài sống tại Qatar. Tập dữ liệu được gán nhãn và được chia thành 3 tập: tập huấn luyện, tập phát triển và tập kiểm thử. Bảng 1 thống kê số lượng cặp câu hỏi trong tập dữ liệu.

Bảng 1: *Bảng thống kê cặp câu hỏi trong tập dữ liệu SemEval 2017 [11]*

	Cặp câu hỏi
Tập huấn luyện	3170
Tập phát triển	700
Tập kiểm thử	880

Độ đo MAP và MRR [9] được sử dụng để đánh giá hiệu quả của mô hình đề xuất.

$$\text{MAP} = \frac{1}{|N|}\sum_{j=1}^{|N|}\frac{1}{m_j}\sum_{k=1}^{|m_j|} Precision\left(R_{jk}\right) \quad (3)$$

4.2. Tham số của mô hình

Nghiên cứu sử dụng biểu diễn từ Glove 300 chiều đưa vào mô hình ở lớp đầu vào. Các từ OOV không nằm trong tập từ điển được khởi tạo một cách ngẫu nhiên. Số chiều lớp ẩn trong mô hình LSTM được thiết lập là 400 chiều. Thuật toán tối ưu Adam được sử dụng với tốc độ học được thiết lập là 0,0001, tham số γ được chọn là 0,0001, batch-size là 64, drop-out là 30%. Mô hình được thực thi trên tensorflow và chạy trên google colab. Hiệu năng của mô hình được đánh giá trên tập phát triển và chọn tham số được chọn tốt nhất trên tập phát triển và sau đó được thiết lập tham số thử nghiệm trên tập kiểm thử.

4.3. Kết quả

Kết quả thử nghiệm trên các mô hình được thể hiện ở Bảng 2, cho thấy rằng khi sử dụng kỹ thuật Max và Mean Pooling độ đo Map tăng lên từ 40% lên 40,5%. Điều đó chứng tỏ rằng, khi véc tơ biểu diễn câu được tổng hợp từ các lớp ẩn có khả năng khai thác nhiều thông tin ngữ nghĩa của câu hơn so với sử dụng lớp ẩn cuối cùng. Hơn nữa, khi tổng hợp biểu diễn câu kết hợp cả Mean và Max Pooling thì kết quả MAP tăng lên 41,07%. Như vậy, khi nối hai véc tơ Mean và Max Pooling làm cho việc chứa thông tin tổng hợp câu tốt hơn. Do vậy kết quả dự đoán của mô hình tốt hơn.

Bảng 2: *Kết quả của mô hình đề xuất*

Mô hình	**MAP**
LSTM sử dụng lớp ẩn cuối	40,03
LSTM-Max Pooling	40,50
LSTM-Mean Pooling	40,51
LSTM-Mean+Max Pooling	**41,07**

5. Kết luận

Nghiên cứu đã đề xuất sử dụng mô hình LSTM với các kỹ thuật tổng hợp biểu diễn câu khác nhau cho bài toán tìm câu hỏi tương đồng. Kết quả thực nghiệm cho thấy rằng, việc sử dụng cả hai chiến lược Mean và Max Pooling cũng ảnh hưởng tới kết quả dự đoán cặp câu hỏi tương đồng. Trong tương lai, nghiên cứu sẽ tiến hành thử nghiệm trên các mô hình biLSTM và CNN và kết hợp các mô hình cũng như sử dụng các cơ chế chú ý vào bài toán này.

TÀI LIỆU THAM KHẢO

[1] Guangyou Zhou, Yubo Chen, Daojian Zeng and Jun Zhao, "Towards faster and better retrieval models for question search," In *Proceedings of the 22nd ACM International Conference on Information Knowledge Management*, New York, pp. 2139-2148, 2013. DOI: 10.1145/2505515.2505550

[2] Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. 2015. "Learning continuous word embedding with metadata for question retrieval in community question answering," In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, pp. 250-259, 2015. DOI: 10.3115/v1/P15-1025

[3] Li Cai, Guangyou Zhou, Kang Liu and Jun Zhao ",Learning the latent topics for question retrieval in community QA," In *Proceedings of 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, pp. 273-281, 2011.

[4] Wei Wu, Xu Sun and Houfeng Wang, "Question condensing networks for answer selection in community question answering" In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, pp. 1746-1755, 2018. DOI: 10.18653/v1/P18-1162

[5] Gheibi, O., Weyns, D. and Quin, F., "Applying machine learning in self-adaptive systems: A systematic literature review," *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, *15*(3), 1-37, 2021. DOI: 10.1145/3469440

[6] Moravvej, S. V., Kahaki, M. J. M., Sartakhti, M. S., and Mirzaei, A., "A method based on attention mechanism using bidirectional long-short term memory (BLSTM) for question answering." In *2021 29th Iranian Conference on Electrical Engineering (ICEE)*, pp. 460-464, 2021. IEEE. DOI: 10.1109/ICEE52715.2021.9544258

[7] Dhandapani, A., and Vadivel, V. (2021). "Question answering system over semantic web," *IEEE Access*, pp. 46900-46910, 2021. DOI: 10.1109/ACCESS.2021.3067942

[8] Stephen Robertson, S. Walker, S. Jones, M. M. HancockBeaulieu and M. Gatford, "Okapi at trec 3," In *Overview of the Third Text REtrieval Conference (TREC-3)*, 1995. DOI: 10.6028/NIST.SP.500-225.routing-city

[9] Jiang, Z., Araki, J., Ding, H. and Neubig, G., "How can we know when language models know? on the calibration of language models for question answering," *Transactions of the Association for Computational Linguistics*, 962-977, 2021. DOI: 10.1162/tacl_a_00407

[10] Chauhan, U., and Shah, A., "Topic modeling using latent Dirichlet allocation: A survey," *ACM Computing Surveys (CSUR), 54*(7), 1-35, 2021. DOI: 10.1145/3462478

[11] Preslav Nakov *et al.*, "SemEval-2017 task 3: Community question answering," In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, pp. 27-48, 2017. DOI: 10.18653/v1/S17-2003

[12] Simone Filice, Giovanni Da San Martino and Alessandro Moschitti, "KeLP at SemEval-2017 task 3: Learning pairwise patterns in community question answering," In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, pp. 326-333, 2017. DOI: 10.18653/v1/S17-2053

[13] M. Tan, B. Xiang, and B. Zhou, "LSTM-based Deep Learning Models for non-factoid answer selection," *IBM Watson Core Technologies*, 2015. DOI: 10.48550/arXiv.1511.04108

ABSTRACT

SENTENCE REPRESENTATION USING LSTM FOR FINDING QUESTION

Dinh Khanh Linh, Tran Quang Huy
University of Information and Communication Technology, Thai Nguyen University, Vietnam

Received on 08/5/2024, accepted for publication on 27/6/2024

Learning sentence representation with the full semantics of a document is a challenge in natural language processing problems because if the semantic representation vector of the sentence is suitable, it will increase the performance of finding similar question problems. In this paper, we propose implementing a series of LSTM models with different ways of extracting sentence representations and applying them to question retrieval to exploit the hidden semantics of sentences. These methods give sentence representation from hidden layers of the LSTM model. The results show that the technique using a combination of both Max Pooling and Mean Pooling gives the highest results on the 2017 SemEval dataset for the problem of finding similarity questions.

Keywords: LSTM; Deep Learning; NLP; QA; learning sentence representation; CQA.