Hệ thống Hỏi-Đáp song ngữ Anh-Việt với hướng tiếp cận BERT đa ngôn ngữ

*Tóm tắt*—Hỏi đáp là một nhiệm vụ đầy thách thức đối với máy móc, vì nó đòi hỏi cả hiểu biết về ngôn ngữ tự nhiên và kiến thức về thế giới. Đây cũng là một trong những vấn đề trọng tâm của xử lý ngôn ngữ tự nhiên nói riêng và lĩnh vực trí tuệ nhân tạo nói chung, vì hỏi đáp là một ví dụ tốt cho khả năng của máy móc trong việc hiểu lý luận phức tạp bằng ngôn ngữ của con người và là một thử thách để có được một trí tuệ nhân tạo mạnh mẽ. Nó cũng có nhiều ứng dụng trong thế giới thực, từ cải thiện tìm kiếm đến xây dựng trợ lý ảo. Ngày càng có nhiều các kỹ thuật tân tiến cho nhiệm vụ hỏi đáp và gặt hái nhiều thành công. Tuy nhiên, các kỹ thuật này chủ yếu dựa trên ngôn ngữ tiếng Anh, khi áp dụng nó trong nhiệm vụ hỏi đáp cho tiếng Việt thì kết quả mang lại không cao. Ngoài ra còn một điểm gây khó khăn nữa trong các kỹ thuật này là kích thước to lớn của mô hình, với một tài nguyên tính toán hạn chế thì khó có thể triển khai được. Bài báo của chúng tôi đề xuất cách xây dựng hệ thống hỏi đáp song ngữ Anh-Việt, dựa trên hướng tiếp cận BERT đa ngôn ngữ (multilingual BERT - mBERT), vì đây là một mô hình đã thể hiện sự hiệu quả của mình trong việc giải quyết các tác vụ đa ngôn ngữ. Phương pháp của chúng tôi đã cải tiến hiệu suất hỏi đáp với ngôn ngữ Anh-Việt, và đạt được kết quả cạnh tranh so với phương pháp tiếp cận học sâu khác.

Từ khóa—hỏi đáp, song ngữ Anh-Việt, BERT, đa ngôn ngữ

# giới thiệu

Hỏi đáp từ lâu đã là một nhiệm vụ thách thức trong tác vụ hiểu ngôn ngữ tự nhiên và xử lý ngôn ngữ tự nhiên. Những thành phần chính trong hỏi đáp yêu cầu khả năng hiểu câu hỏi và ngữ cảnh chứa câu hỏi được tạo ra từ ngữ cảnh đó. Hỏi đáp là vô cùng khó bởi vì bản chất động của ngôn ngữ tự nhiên. Điều này đã dẫn đến việc áp dụng các phương pháp hướng dữ liệu trong trả lời câu hỏi. Ý tưởng là cho phép dữ liệu, thay vì các phương thức, thực hiện hầu hết các công việc trong công việc trả lời câu hỏi. Điều này thực hiện được là nhờ một số lượng lớn các kho lưu trữ văn bản có sẵn. Từ trước đến nay đã có nhiều kỹ thuật được sử dụng cho hệ thống hỏi đáp, tiêu biểu trong số đó có thể kể đến là hướng tiếp cận dựa trên quy tắc (rule-based), hướng tiếp cận thống kê và hướng tiếp cận học sâu [4].

Hướng tiếp cận dựa trên quy tắc [4] là một trong những phương pháp ban đầu đáng chú ý nhất được sử dụng cho những hệ thống hỏi đáp. Các hệ thống này sử dụng những quy tắc được tạo ra từ ngữ nghĩa ngữ pháp để xác định câu trả lời chính xác cho một câu hỏi. Các quy tắc này thường được làm thủ công và sử dụng kinh nghiệm, dựa trên các gợi ý từ vựng và ngữ nghĩa trên ngữ cảnh. Những quy tắc này gồm các mẫu được xác định trước để phân loại câu hỏi dựa trên loại câu trả lời. Các quy tắc ngữ pháp đại diện cho ngữ cảnh dưới dạng cây quyết định và điều này được sử dụng để tìm ra đường đi dẫn đến câu trả lời đúng [4]. Một nhược điểm lớn của hệ thống hỏi đáp dựa trên quy tắc là các quy tắc kinh nghiệm cần được tạo thủ công.

Với sự phát triển nhanh chóng của tài liệu văn bản có sẵn trên mạng, tầm quan trọng của các hướng tiếp cận thống kê [4] đối với hỏi đáp tăng lên. Cách tiếp cận này dựa vào việc dự đoán câu trả lời dựa trên dữ liệu. Vì phương pháp này khó có khả năng giải quyết sự không đồng nhất của dữ liệu và không có ngôn ngữ truy vấn có cấu trúc nên chúng đã được điều chỉnh cho phù hợp với các giai đoạn khác nhau của hỏi đáp [4]. Những phương pháp tiếp cận thống kê yêu cầu hình thành giả thuyết trước khi tiến hành xây dựng mô hình. Giả thuyết này đóng vai trò tạo ra mô hình.

Với sự xuất hiện của mạng nơ-ron cho các hệ thống hỏi đáp đã mở ra rất nhiều khả năng. Các kỹ thuật học máy thông thường bị hạn chế về khả năng xử lý dữ liệu tự nhiên ở dạng thô của chúng. Phương pháp học sâu là một nhánh phát triển từ phương pháp học máy, được đặc trưng bởi cách sử dụng mạng nơ-ron. Học sâu [4] là phương pháp học cho phép một máy được cung cấp dữ liệu thô và tự động khám phá các biểu diễn cần thiết để phát hiện, dự đoán hoặc phân lớp. Có nhiều cấp độ biểu diễn, thu được bằng cách tạo các mô-đun đơn giản nhưng phi tuyến tính mà mỗi mô-đun chuyển đổi biểu diễn ở một cấp (bắt đầu với đầu vào thô) thành biểu diễn ở cấp độ cao hơn. Với thành phần của đủ các loại biến đổi như vậy, các hàm rất phức tạp có thể được học, dẫn tới hiệu suất của phương pháp học sâu đã trở nên vượt trội so với các phương pháp khác. Vì vậy, chúng tôi đã sử dụng phương pháp học sâu để xây dựng hệ thống Hỏi-Đáp của mình.

BERT là một mô hình đào tạo trước, sử dụng phương pháp học sâu và là một trong những mô hình phổ biến cho tác vụ xử lý ngôn ngữ tự nhiên. Được giới thiệu vào tháng 10/2018, BERT là viết tắt của cụm từ Bidirectional Encoder Representations from Transformers. BERT được thiết kế để đào tạo trước các biểu diễn hai chiều sâu từ văn bản không được gắn nhãn bằng cách điều chỉnh chung trên cả ngữ cảnh bên trái và bên phải trong tất cả các lớp. Cấu trúc mô hình của BERT là bộ mã hóa Biến áp hai chiều nhiều lớp phân phối hoàn toàn bằng cơ chế lặp lại và tích chập và chỉ dựa vào cơ chế chú ý, do đó giảm đáng kể thời gian đào tạo [3]. BERT được huấn luyện cho đơn ngôn ngữ, và mBERT là biến thể được huấn luyện cho đa ngôn ngữ. mBERT ánh xạ chuỗi văn bản vào một không gian đa ngôn ngữ được dùng chung bằng cách cùng tiền huấn luyện nhiều ngôn ngữ [3]. Điều này cho phép ta chuyển các bộ mã hóa đa ngôn ngữ qua các ngôn ngữ cụ thể.

Việc tinh chỉnh tác vụ xuôi dòng trong ngôn ngữ nguồn có thể không yêu cầu mBERT giữ lại các đặc điểm cấu trúc hoặc cách mã hóa cú pháp. Việc khuyến khích mBERT tìm hiểu mối tương quan giữa cấu trúc cú pháp và nhãn đích có thể mang lại lợi ích cho việc chuyển giao đa ngôn ngữ. Nhóm tác giả Ahmad [2] đề xuất tăng cường mBERT với cú pháp ngôn ngữ chung trong khi tinh chỉnh các nhiệm vụ phía sau. Họ sử dụng mạng chú ý đồ thị (graph attention network - GAT) để tìm hiểu các biểu diễn có cấu trúc của các chuỗi đầu vào được tích hợp vào cơ chế tự chú ý.

Hiện nay, mBERT đã được học với 104 ngôn ngữ trên thế giới với khoảng 119 nghìn từ vựng (trong bản tiếng Anh - BERTBASE chỉ có khoảng 30 nghìn từ), nhưng trong thực tế thì ít khi nào ta dùng nhiều đến như vậy. Kích thước của bộ từ vựng sẽ gây ảnh hưởng lớn đến số lượng tham số. Giảm kích thước của mô hình là một bước quan trọng để có thể triển khai được trên thực tế. Nhóm Abdaoui [1] đã đề xuất cách làm giảm kích thước mô hình bằng cách làm giảm kích thước của bộ từ vựng. Trong trường hợp mô hình đa ngôn ngữ, hầu hết các tham số đều nằm trong lớp nhúng nên việc giảm kích thước từ vựng sẽ có tác động quan trọng đến tổng số tham số. Do đó phương pháp được đề xuất là tạo ra một mô hình nhỏ hơn chuyên dùng để xử lý các ngôn ngữ mục tiêu của hệ thống. Với mục tiêu là không ảnh hưởng nhiều đến độ chính xác chung hoặc thậm chí gia tăng khả năng suy luận của mô hình tại mỗi ngôn ngữ mục tiêu và khắc phục được phần nào các hạn chế về mặt tài nguyên phần cứng.

Tuy nhiên, các mô hình này đa phần có hiệu suất tốt cho tiếng Anh, còn với tiếng Việt thì vẫn còn hạn chế. Chúng tôi muốn cải tiến hiệu suất của mô hình với tiếng Việt, từ đó sử dụng mô hình để tạo ra hệ thống Hỏi-Đáp dùng cho tiếng Việt đạt kết quả tốt hơn. Hệ thống của chúng tôi có khả năng trả lời câu hỏi trên đoạn ngữ cảnh cho trước. Vì mang tính song ngữ Anh-Việt, ngữ cảnh và câu hỏi không nhất thiết phải cùng một ngôn ngữ. Ngữ cảnh tiếng Anh mà câu hỏi tiếng Việt thì hệ thống vẫn đưa ra được câu trả lời, và ngược lại cho ngữ cảnh tiếng Việt và câu hỏi tiếng Anh. Câu trả lời sẽ được rút trích ra từ ngữ cảnh, nên ngôn ngữ sẽ trùng với ngôn ngữ của đoạn văn ngữ cảnh.

Trong bài báo này, chúng tôi sẽ trình bày phương pháp xây dựng hệ thống Hỏi-Đáp song ngữ Anh-Việt. Những đóng góp của chúng tôi được mô tả như sau:

* Chúng tôi xây dựng tập dữ liệu Anh-Việt để huấn luyện mô hình, vì tập dữ liệu dùng cho tiếng Việt rất hạn chế nên chúng tôi phải tiến hành xây dựng thêm.
* Chúng tôi tiến hành thu gọn bộ từ vựng mô hình mBERT để phù hợp với tài nguyên tính toán hiện có.
* Chúng tôi tiến hành tinh chỉnh mô hình mBERT để hiệu suất thu được với tiếng Việt cao hơn.
* Sau cùng, chúng tôi sẽ dùng mô hình thu được để xây dựng hệ thống Hỏi-Đáp.

Kết quả ban đầu thu được khá khả quan, mô hình mBERT với bộ từ vựng song ngữ Anh-Việt (mBERT-en-vi) đạt được kết quả tương ứng với các chỉ số EM và F1 là 67.70% và 80.80% trên tập dữ liệu tiếng Anh, 44.97% và 66.28% trên tập dữ liệu tiếng Việt.

# nghiên cứu tổng quan

# phương pháp đề xuất

# kết quả

Tại phần này, chúng tôi sẽ trình bày về kết quả thu được từ phương pháp chúng tôi đề xuất. Chúng tôi sử dụng tập dữ liệu MLQA [5] để đánh giá kết quả của mình. MLQA là một tập dữ liệu benchmark đánh giá hỏi đáp rút trích song song đa chiều trong bảy ngôn ngữ: tiếng Anh, tiếng Ả Rập, tiếng Đức, tiếng Việt, tiếng Tây Ban Nha, tiếng Trung giản thể và tiếng Hindi. Kho ngữ liệu kết quả của tập dữ liệu có từ 5,000 đến 6,000 ví dụ trong mỗi ngôn ngữ, và hơn 12,000 ví dụ trong tiếng Anh. Mỗi ví dụ có một tương đương được căn chỉnh trong nhiều ngôn ngữ khác (luôn bao gồm cả tiếng Anh), phần lớn được căn chỉnh 4 hướng. Tổng hợp, có hơn 46,000 chú thích hỏi đáp. Để đánh giá mô hình của mình, chúng tôi sử dụng hai chỉ số hiệu suất tiêu chuẩn của SQuAD [6] là điểm Exact Match (EM) và điểm F1.

Hiện tại, chúng tôi đã thực hiện một số thí nghiệm và nhận thấy rằng, mô hình chúng tôi có hiệu suất tốt hơn khi kiểm thử trên tập dữ liệu MLQA so với những mô hình trong bài báo của nhóm tác giả Lewis [5]. Cụ thể hơn, mô hình mBERT với bộ từ vựng song ngữ Anh-Việt (mBERT-en-vi) đạt được kết quả tương ứng với các chỉ số EM và F1 là 67.70% và 80.80% trên tập dữ liệu tiếng Anh, 44.97% và 66.28% trên tập dữ liệu tiếng Việt, 53.72% và 66.62% trên tập dữ liệu phối trộn giữa cả hai ngôn ngữ. Kết quả thu được thành tựu đáng mong đợi, nhưng vẫn cần nhiều thí nghiệm hơn nữa trong tương lai.

# kết luận

Trong bài báo này, chúng tôi đã trình bày công việc nghiên cứu, hiện thực cùng kết quả của mô hình Hỏi-Đáp song ngữ Anh-Việt. Chúng tôi đã đưa ra phương pháp để xây dựng một Hệ thống Hỏi-Đáp song ngữ. Do một số khó khăn trong việc xây dựng hệ thống như tìm kiếm dữ liệu huấn luyện, đặc biệt là dữ liệu tiếng Việt, cũng như các hạn chế về mặt tài nguyên tính toán đã gây ra nhiều khó khăn trong quá trình tìm hiểu cũng như kiểm nghiệm các phương án khả thi, đây là một thách thức trong việc xây dựng một mô hình đa ngôn ngữ. Chúng tôi nhận thấy rằng, sử dụng cách tiếp cận bằng multilingual BERT là một phương pháp khả thi để đạt được mục tiêu đã đề ra. Kết quả của chúng tôi thu được với mô hình Anh-Việt cũng có hiệu suất tốt hơn so với mô hình gốc.

##### tài liệu tham khảo

1. Amine Abdaoui, Camille Pradel, and Grégoire Sigel. Load what you need: Smaller versions of mutililingual BERT. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online, November 2020. Association for Computational Linguistics.
2. Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554, Online, August 2021. Association for Computational Linguistics.
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171– 4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
4. K. S. D. Ishwari, A. K. R. R. Aneeze, S. Sudheesan, H. J. D. A. Karunaratne, Anupiya Nugaliyadde, and Yashas Mallawarachchi. Advances in natural language question answering: A review. *CoRR*, abs/1904.05276, 2019.
5. Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online, July 2020. Association for Computational Linguistics.
6. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.