ĐẠI HỌC HẠ LONG
Học để thành công
TẠP CHÍ KHOA HỌC ĐẠI HỌC HẠ LONG
Scientific Journal of Ha Long University
http://uhl.edu.vn/

TẬP DỮ LIỆU TIẾNG VIỆT CHO BÀI TOÁN TÌM CÂU HỎI TƯƠNG ĐỒNG

Hà Thị Thanh[1*], Nguyễn Thị Oanh[1]

[1]Trường Đại học Công nghệ Thông tin và Truyền thông, Đại học Thái Nguyên

** Email: htthanh@ictu.edu.vn*

Ngày nhận bài: 05/8/2022 *Ngày nhận bài sửa sau phản biện: 10/11/2022* *Ngày chấp nhận đăng: 14/11/2022*

TÓM TẮT

Bài toán tìm kiếm câu hỏi tương đồng là bài toán phổ biến và quan trọng trong xử lý ngôn ngữ tự nhiên. Tuy nhiên, có rất ít nghiên cứu về bài toán này trên tập dữ liệu tiếng Việt. Nguyên nhân của hiện tượng trên là do chưa có tập dữ liệu tiếng Việt chuẩn cho bài toán tìm kiếm câu hỏi. Trong bài báo này, chúng tôi trình bày một phương pháp xây dựng tập dữ liệu tiếng Việt cho bài toán tìm kiếm câu hỏi tương đồng. Chúng tôi xây dựng được 7911 cặp câu hỏi được gán nhãn. Đồng thời, tập dữ liệu này cũng được thử nghiệm đánh giá trên một số mô hình học máy cơ bản.

Từ khóa: *elastic search, máy tìm kiếm, tập dữ liệu.*

VIETNAMESE DATASET FOR THE FINDING SIMILAR QUESTION PROBLEM

ABSTRACT

Finding similar questions is a common problem in natural language processing. However, little research has been conducted on the question retrieval problem for Vietnamese. The reason for this is that there is no standard Vietnamese dataset for the finding question problem. In this paper, we created a method to build a Vietnamese dataset for the problem of finding similar questions. As a result, we built 7911 pairs of labeled questions. This dataset was evaluated on some basic machine learning models.

Keywords: *dataset, elastic search, search engine.*

1. GIỚI THIỆU

Bài toán tìm câu hỏi tương đồng là bài toán trung gian hỗ trợ cho các hệ thống hỏi đáp tự động tìm kiếm câu trả lời cho câu hỏi mới. Bài toán tìm kiếm câu hỏi tương đồng tuy không phải là lĩnh vực nghiên cứu mới nhưng nó được sử dụng nhiều trong các hệ thống hỏi đáp. Bài toán này còn có tên gọi khác là bài toán tìm kiếm câu hỏi hay phát hiện câu hỏi trùng lặp.

Bài toán tìm câu hỏi tương đồng được định nghĩa như sau: Cho một câu hỏi truy vấn (câu hỏi mới) q và các câu *hỏi* $q_1, q_2, \ldots, q_n$ trong kho dữ liệu của hệ thống hỏi đáp. Đầu ra trả về danh sách xếp hạng các câu hỏi sao cho những câu tương đồng nhất với câu hỏi truy vấn ở trên và câu không tương đồng nhất ở cuối của danh sách. Bài toán tìm câu hỏi tương đồng về bản chất là một bước trung gian trong hệ thống hỏi đáp. Trong hội nghị Semeval 2017 (Nakov và cs., 2017), để giải quyết bài toán tìm câu trả lời tốt nhất cho câu hỏi mới, đầu tiên hệ thống sẽ thực hiện tìm các câu hỏi tương đồng với câu hỏi mới, sau đó, một câu trả lời tốt nhất được chọn trong

số các câu trả lời của các câu hỏi tương đồng. Trong các nghiên cứu của Zhou (Chan và cs., 2012; Yin và cs., 2016) sử dụng bài toán tìm kiếm câu hỏi tương đồng với mục đích tìm câu trả lời cho câu hỏi mới từ các câu trả lời có trong cơ sở dữ liệu.

Ví dụ về cặp câu hỏi tương đồng:

Câu hỏi 1: Làm ơn chỉ giùm tôi cách tắt phím slide to unclock trên Samsung S9 Plus

Câu hỏi 2: Cách tắt màn hình slide to unclock chỉ để màn hình kiểu vuốt để mở khóa máy Samsung J7 Pro

Để đánh giá các mô hình tìm kiếm, các nguồn dữ liệu của các cặp câu hỏi được thu thập hoặc do con người tạo ra. Việc xây dựng tập dữ liệu chuẩn đóng vai trò quan trọng trong việc đánh giá các mô hình cho các bài toán trong xử lý ngôn ngữ tự nhiên, đặc biệt là cho ngôn ngữ tài nguyên thấp như tiếng Việt.

Các tập dữ liệu tiếng Anh phổ biến như: Yahoo!webscope (Chan và cs., 2012), tập Trec-QA (Wang và cs., 2007), tập Quora (Sharma và cs., 2019), SemEval (Nakov và cs., 2015, 2016, 2017).

Do sự phát triển nhanh chóng của các bộ hỏi đáp trên tiếng Anh, các mô hình học máy khác nhau được đề xuất và thực hiện trên các bộ dữ liệu này như các mô hình SVM, LSTM, CNN được tổng hợp trên tài liệu của Nakov (Nakov và cs., 2017) trên bộ dữ liệu Semeval. Mô hình LSTM cho hỏi đáp trên bộ Yahoo!answer (Chan và cs., 2012). Mấy năm gần đây, mô hình BERT (Devlin và cs., 2019) được áp dụng vào các bài toán hỏi đáp, cụ thể là trên bài toán tìm câu hỏi tương đồng trong các nghiên cứu (Sakata và cs., 2019; Yang và cs., 2019) cho kết quả vượt trội so với các phương pháp trước đó.

Với tập dữ liệu tiếng Việt, hầu như chưa có nghiên cứu về bài toán tìm câu hỏi tương đồng trên tập dữ liệu này. Hơn nữa, cần yêu cầu tập dữ liệu đủ lớn để các mô hình học sâu có thể chạy ổn định. Để đẩy mạnh nghiên cứu về bài toán tìm câu hỏi tương đồng trên ngôn ngữ tiếng Việt, chúng tôi đề xuất phương pháp xây dựng bộ dữ liệu cho bài toán này. Kết quả nghiên cứu đã đóng góp bộ dữ liệu gồm 7911 cặp câu hỏi được gán nhãn. Tiếp theo, chúng tôi thực hiện đánh giá một số mô hình học máy cơ bản trên tập dữ liệu này.

2. MỘT SỐ TẬP DỮ LIỆU TIẾNG ANH

Trong phần này chúng tôi miêu tả một vài tập dữ liệu tiếng Anh cho hệ thống hỏi đáp, trong đó có bài toán tìm câu hỏi tương đồng:

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 87390 câu hỏi và 314446 câu trả lời. Tập dữ liệu này chứa rất nhiều thông tin hữu ích cho việc nghiên cứu các bài toán trên hệ thống hỏi đáp, ví dụ như các thông tin về chủ đề câu hỏi, nội dung câu hỏi, mô tả chi tiết (giải thích) 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 (Chan và cs., 2012). Tập này chứa các cặp câu hỏi dạng factoid và một 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 từ loại POS, thực thể có tên NER và phân tích câu phụ thuộc.

Quora: Đây là tập dữ liệu được công bố trong cuộc thi Kaggle. Tập dữ liệu này được thu thập từ trang hỏi đáp Quora.com bao gồm các lĩnh vực trong cuộc sống hay công việc hàng ngày. Tập dữ liệu này chứa các câu hỏi được gán nhãn duplicate (1) và non-duplicate (0) phục vụ cho bài toán tìm câu hỏi tương đồng. Trong 404351 cặp câu hỏi có 149306 cặp câu có nhãn 1 và 255045 cặp câu có nhãn 0.

SemEval: Tập này được thu thập từ forum hỏi đáp chia sẻ mọi thứ liên quan tới công việc và cuộc sống ở Qatar. 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 ngữ nghĩa (Nakov và

cs., 2015, 2016, 2017). Từ khía cạnh ngôn ngữ, tập dữ liệu này rất có giá trị và thách thức. Tập dữ liệu này chứa lượng lớn 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 không tồn tại. Workshop 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 thành ba tập là tập huấn luyện, tập phát triển và tập kiểm thử chứa các câu hỏi và các câu trả lời của nó. Với mỗi câu hỏi gốc có 10 câu hỏi liên quan (được đưa qua máy tìm kiếm) và được gán ba nhãn: Perfect match, Relevant và Irrelevant. Với mỗi câu hỏi liên quan có 10 câu trả lời được gán ba nhãn Good, Bad và Potentially useful. Mỗi câu hỏi liên quan lại có 10 câu trả lời cũng được gán ba nhãn như trên.

Cho đến nay, chưa có bất kỳ bộ dữ liệu nào về dữ liệu tiếng Việt để phục vụ cho nghiên cứu về bài toán tìm câu hỏi tương đồng. Như đã đề cập ở trên, các bộ dữ liệu là tiêu chuẩn để đánh giá các mô hình học máy và được sử dụng khuyến khích các nhà nghiên cứu khám phá các mô hình hiểu ngôn ngữ cho tiếng Việt. Vì vậy, việc xây dựng dữ liệu tiếng Việt là động lực chính để chúng tôi xây dựng bộ dữ liệu mới cho bài toán tìm kiếm câu hỏi tương đồng.

3. PHƯƠNG PHÁP XÂY DỰNG TẬP DỮ LIỆU

Để xây dựng tập dữ liệu tiếng Việt, chúng tôi thực hiện qua các bước như sau:

– Bước 1: Chúng tôi chọn nguồn thu thập dữ liệu. Chúng tôi tiến hành chọn website chứa các dữ liệu là câu hỏi của người dùng. Chúng tôi chọn website của Thế giới di động trong mục hỏi đáp của người dùng về các nội dung liên quan tới mua bán các thiết bị điện tử như điện thoại, máy tính. Qua bước này chúng tôi thu thập được bộ câu hỏi không có nhãn có kích thước 1.1Mb dữ liệu.

– Bước 2: Chúng tôi sử dụng máy tìm kiếm Elasticsearch (Kuc & Rogozinski, 2013) tiến hành chọn và gán nhãn dữ liệu như sau: Đầu tiên, tập con các câu hỏi được chọn và dùng làm câu hỏi gốc. Mỗi câu hỏi này sẽ được đưa vào máy tìm kiếm coi như là câu truy vấn. Sau đó, từng câu hỏi từ tập câu hỏi gốc trên được đưa vào máy tìm kiếm. Kết quả trả về một danh sách các câu hỏi liên quan tới câu truy vấn. Mười câu hỏi đầu tiên trong danh sách kết quả được chọn để tiến hành gán nhãn.

– Bước 3: Gán nhãn. Cứ mỗi câu hỏi gốc có 10 cặp câu hỏi tương ứng với các nhãn là 1 và 0 được gán bởi con người. Một cặp câu hỏi được chọn nhãn là 1 nếu phần trả lời của câu hỏi thứ nhất có thể dùng để trả lời một phần hoặc toàn bộ cho câu hỏi thứ hai và ngược lại. Công việc gán nhãn được thực hiện bởi bốn thành viên trong nhóm nghiên cứu. Sau đó, các thành viên trong nhóm sẽ tiến hành kiểm tra chéo các kết quả gán nhãn. Kết thúc giai đoạn gán nhãn chéo, chúng tôi tiến hành thống kê kết quả gán nhãn. Kết quả gán nhãn trùng nhau khoảng 80 – 85%. Những câu gán nhãn không giống nhau được tiến hành rà soát lại và thống nhất kết quả gán nhãn cuối cùng.

– Bước 4: Cuối cùng, các câu hỏi gốc mà không có câu hỏi nào tương đồng cũng bị loại khỏi tập dữ liệu. Để làm tăng độ khó của tập dữ liệu, các cặp câu hỏi dễ (là những câu dễ dàng tìm được qua máy tìm kiếm, thường có ít thách thức về khoảng cách từ vựng) cũng được rà soát lại và đưa ra quyết định có bị loại khỏi tập dữ liệu hay không.

Sau khi có tập dữ liệu, các cặp câu được gán nhãn, tập dữ liệu này được chia thành 3 tập: tập huấn luyện, tập phát triển và tập kiểm thử. Tập dữ liệu thu được có 30% dữ liệu có nhãn 1 có liên quan tới câu hỏi gốc, còn lại 70% là cặp câu có nhãn 0.

Trong quá trình làm dữ liệu, máy tìm kiếm Elastic (Kuc & Rogozinski, 2013) phiên bản 6.6.1 được sử dụng. Đây là máy tìm kiếm được xây dựng trên thư viện Lucence. Máy tìm kiếm Elastic tìm kiếm và trả lại kết quả là danh sách các câu hỏi liên quan với câu hỏi

gốc theo độ đo (Kuc & Rogozinski, 2013) được tính như sau:

$$\text{score}(q,d) = queryNorm(q) * \text{coord}(q,d) * \sum\left(tf(t) * idf(t)^2 * \text{t.getBoost}() * \text{norm}(t,d)\right),$$

trong đó: t là từ trong văn bản d, $score(q,d)$ là độ đo mức độ liên quan của văn bản d với truy vấn q, $queryNorm(q)$ là hệ số chuẩn hóa truy vấn để các truy vấn này có thể so sánh được với các truy vấn khác, $coord(q,d)$ là hệ số ngang hàng, thông thường những văn bản chứa nhiều từ trong truy vấn q sẽ có điểm số cao hơn, $t.getBoost()$ là hệ số tăng cường truy vấn, $norm(t,d)$ chuẩn hóa trường độ dài.

Đồng thời trong quá trình thu thập dữ liệu, bộ dữ liệu không gán nhãn được giữ lại để huấn luyện các mô hình ngôn ngữ. Tập này dùng để học biểu diễn của từ trong giai đoạn tiền huấn luyện từ nhúng như Word2vec hoặc Bert. Bảng 1 dưới đây là thống kê tập dữ liệu sau khi gán nhãn:

Bảng 1. Bảng thống kê tập dữ liệu tiếng Việt

	Số lượng cặp câu hỏi	Số lượng câu hỏi gốc
Tập train	5996	615
Tập dev	847	86
Tập test	1068	110

4. PHÂN TÍCH TẬP DỮ LIỆU

Để hiểu rõ hơn về tập dữ liệu, chúng tôi phân tích tập dữ liệu được gán nhãn theo các khía cạnh phân tích dựa trên độ dài (độ dài câu hỏi theo syllabus, theo từ và câu). Việc làm này có ý nghĩa rất quan trọng trong việc lựa chọn các mô hình học máy phù hợp. Bảng 2 trình bày số liệu thống kê chi tiết như sau:

Bảng 2. Một số kết quả thống kê trên tập dữ liệu tiếng Việt

	Số lượng
Số lượng cặp câu có nhãn 1	5177
Số lượng cặp câu nhãn 0	2734
Độ dài trung bình câu theo syllable	27
Số câu hỏi có 1 câu	5294
Số câu hỏi có từ 2 câu trở lên	2539
Số từ theo từ điển theo syllable	5821
Số từ trong từ điển theo tách từ tiếng Việt	6337

Tiếp theo chúng tôi cũng thực hiện một số thống kê trên tập dữ liệu không gán nhãn như Bảng 3 dưới đây:

Bảng 3. Bảng thống kê tập dữ liệu không có nhãn tiếng Việt

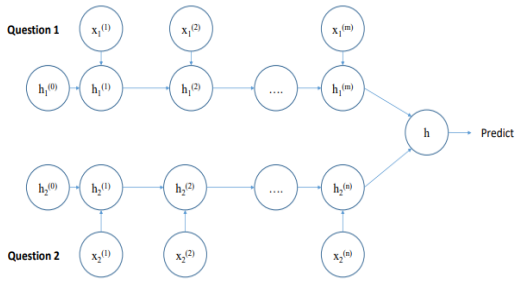
	Số lượng
Kích thước tập dữ liệu không có nhãn	1.1Mb
Kích thước từ điển theo syllable	151 735
Độ dài trung bình câu hỏi theo syllable	31

Đồng thời, chúng tôi tiến hành đánh giá tập dữ liệu mới với một số mô hình cơ bản sau:

– Elastic Search: Là kết quả đánh giá trên máy tìm kiếm trong quá trình xây dựng tập dữ liệu được mô tả ở mục 3.

– Mô hình SVM (Trần Cao Đệ & Phạm Nguyên Khang, 2012): Cặp câu hỏi được đưa vào mô hình SVM sử dụng biểu diễn câu dùng trọng số tf.idf.

– Mô hình LSTM (Hình 1): Cặp câu hỏi được mã hóa bởi hai mô hình LSTM và hai mô hình được sử dụng cùng bộ tham số. Lớp ẩn cuối cùng của LSTM được sử dụng làm biểu diễn của câu hỏi. Cuối cùng, hai biểu diễn của hai câu hỏi được nối lại và cho qua lớp MLP để dự đoán.



Hình 1. Mô hình LSTM cho bài toán tìm câu hỏi tương đồng trên tập dữ liệu tiếng Việt

Kết quả đánh giá các mô hình học máy điển hình trên tập dữ liệu mới như Bảng 4 dưới đây:

Bảng 4. Bảng thống kê tập dữ liệu không có nhãn tiếng Việt trên độ đo MAP

Mô hình	MAP
Elastic search	52.00
SVM	49.75
LSTM	52.60

Trong đó, mô hình SVM cho kết quả thấp hơn so với mô hình thực hiện trên máy tìm kiếm Elastic nhưng mô hình LSTM cho kết quả tốt hơn so với mô hình Elastic và SVM. Điều này chứng tỏ rằng, mô hình học sâu có thể hoạt động tốt trên tập dữ liệu mới.

5. KẾT LUẬN

Trong bài báo này, chúng tôi trình bày phương pháp xây dựng tập dữ liệu tiếng Việt cho bài toán tìm câu hỏi tương đồng. Tập dữ liệu này được phân tích và chứng tỏ đủ lớn có thể đáp ứng thử nghiệm trên các mô hình học sâu. Trong thời gian tới, chúng tôi tiếp tục khai thác các mô hình học sâu với các cơ chế chú ý trên tập dữ liệu này.

TÀI LIỆU THAM KHẢO

Chan, W., Zhou, X., Wang, W., & Chua, T.-S. (2012). Community Answer Summarization for Multi-Sentence Question with Group L1 Regularization. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 582–591.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *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)*, 4171–4186.

Kuc, R., & Rogozinski, M. (2013). *Mastering ElasticSearch*. Packt Publishing.

Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., & Verspoor, K. (2017). SemEval-2017 Task 3: Community Question Answering. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 27–48.

Nakov, P., Màrquez, L., Magdy, W., Moschitti, A., Glass, J., & Randeree, B. (2015). SemEval-2015 Task 3: Answer Selection in Community Question Answering. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 269–281.

Nakov, P., Màrquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, A. A., Glass, J., & Randeree, B. (2016). SemEval-2016 Task 3: Community Question Answering. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 525–545.

Sakata, W., Shibata, T., Tanaka, R., & Kurohashi, S. (2019). FAQ Retrieval using Query-Question Similarity and BERT-Based Query-Answer Relevance. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1113–1116.

Sharma, L., Graesser, L., Nangia, N., & Evci, U. (2019). Natural Language Understanding with the Quora Question Pairs Dataset. *ArXiv*.

Trần Cao Đệ & Phạm Nguyên Khang. (2012). Phân loại văn bản với máy học vector hỗ trợ và cây quyết định. *Tạp chí Khoa học Đại học Cần Thơ*, *2012:21a*, 52–63.

Wang, M., Smith, N. A., & Mitamura, T. (2007). What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 22–32.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, *32*.

Yin, Y., Wei, F., Dong, L., Xu, K., Zhang, M., & Zhou, M. (2016). Unsupervised word and dependency path embeddings for aspect term extraction. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2979–2985.