USING MULTI-LAYER LSTMS FOR QUESTION RETRIEVAL

Luong Thi Minh Hue
TNU - University of Information and Communication Technology

ARTICLE INFO

Received: 01/4/2022
Revised: 26/5/2022
Published: 27/5/2022

KEYWORDS

LSTM
NLP
Deep learning
CQA
Multi-layerLSTM

ABSTRACT

Question retrieval is one of the important problems in the Community Question Answering system. The biggest challenge of this problem is the lexical gap between the words and phrases of the first and second question. Although there are many studies applied to this problem, the exploitation of multi-layer LSTM model has not been tested on this problem. In this paper, we exploit a multi-layer LSTM model applied to the problem of finding similar questions for the purpose of exploiting hidden semantics of sentences. The multi-layer LSTM model is capable of synthesizing semantics by multiple layers and exploits hidden semantics through many layers. Our model learned the semantics of sentences and improved the performance of finding question. The results show that the model with 3 layers gives the best results compared to the original LSTM model and other multi-layer models on the 2017 semeval dataset for the problem of finding similar questions.

SỬ DỤNG MÔ HÌNH LSTM NHIỀU TẦNG VÀO BÀI TOÁN TÌM KIẾM CÂU HỎI

Lương Thị Minh Huế
Trường Đại học Công nghệ Thông tin và Truyền thông – ĐH Thái Nguyên

THÔNG TIN BÀI BÁO

Ngày nhận bài: 01/4/2022
Ngày hoàn thiện: 26/5/2022
Ngày đăng: 27/5/2022

TỪ KHÓA

LSTM
Học sâu
Xử lý ngôn ngữ tự nhiên
Hỏi đáp cộng đồng
Mô hình đa tầng

TÓM TẮT

Tìm câu hỏi tương đồng là một trong những bài toán quan trọng trong hệ thống hỏi đáp. Thách thức lớn nhất của bài toán này là thách thức về khoảng cách từ vựng giữa các từ trong câu hỏi thứ nhất và câu hỏi thứ hai. Mặc dù có nhiều nghiên cứu đề xuất các mô hình, tuy nhiên việc khai thác mô hình LSTM nhiều lớp chưa được thử nghiệm trên bài toán này. Trong bài báo này, chúng tôi khai thác mô hình LSTM nhiều tầng áp dụng vào bài toán tìm câu hỏi tương đồng với mục đích khai thác ngữ nghĩa ẩn của câu. Mô hình LSTM nhiều tầng có khả năng tổng hợp ngữ nghĩa qua nhiều lớp. Nó khai thác ngữ nghĩa ẩn qua nhiều tầng, từ đó giúp cho mô hình hiểu được ngữ nghĩa của câu. Kết quả chỉ ra rằng mô hình 3 tầng cho kết quả tốt nhất so với mô hình gốc LSTM và các mô hình nhiều tầng khác trên tập dữ liệu semeval 2017 cho bài toán tìm câu hỏi tương đồng.

DOI: https://doi.org/10.34238/tnu-jst.5799

Email: lmhue@ictu.edu.vn

1. Giới thiệu

Hệ thống hỏi đáp dựa trên cộng đồng (CQA) đã trở thành một nền tảng trực tuyến ngày càng phổ biến. Các forum, nơi người dùng có thể đăng câu hỏi hoặc câu trả lời các câu hỏi của người dùng khác đã đăng lên, cung cấp cho người dùng nơi mà họ có thể chia sẻ kiến thức và kinh nghiệm của mình. Khi một người dùng đăng một câu hỏi mới lên hệ thống thì người dùng sẽ phải chờ một thời gian trễ nào đó để nhận câu trả lời từ người dùng khác. Hơn nữa, forum sau một thời gian hoạt động, lượng câu hỏi và câu trả lời sẽ được tích lũy trong kho dữ liệu là rất lớn. Điều đó có nghĩa là khả năng người dùng hỏi lại những câu hỏi lặp lại là rất lớn. Một lý do khác, khi lượng câu hỏi và câu trả lời lớn thì việc tìm câu trả lời cho câu hỏi trong kho dữ liệu có sẵn rất mất thời gian. Vì vậy, bài toán tìm kiếm câu hỏi tương đồng với câu hỏi mới với mục đích tận dụng câu trả lời đã có của những câu hỏi tương đồng với câu truy vấn [1], [2]. Hệ thống CQA hướng tới tìm câu trả lời một cách tự động từ câu trả lời của những câu hỏi đã có.

Bài toán tìm kiếm câu hỏi tương đồng được định nghĩa như sau: Cho một câu hỏi truy vấn q và một tập các câu hỏi đã có trong hệ thống $\{q_1, q_2, \ldots, q_n\}$, 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.

Các nghiên cứu trước [3] 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 [4], [5]:

Câu hỏi 1: which is a good bank as per your experience in Doha

Câu hỏi 2: Hi guys, I need to open a new bank account. Which is the best bank in Qatar? I assume all of them wil roughly be the same, but stll which has a slight edge (money transfer, benefits etc) Thanks!!

Hai câu hỏi này cùng một ý hỏi nhưng diễn đạt 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. Để giải quyết thách thức này, các nghiên cứu trước đó sử dụng các kỹ thuật như kỹ thuật gióng mềm trong dịch máy [3]. Các nghiên cứu khác sử dụng các mô hình học sâu sử dụng các đặc trưng kỹ thuật và tri thức bên ngoài [4]-[6]. Các nghiên cứu này khai thác các đặc trưng về ngữ nghĩa và cú pháp trong câu sử dụng mô hình LSTM. Tuy nhiên, mô hình LSTM chưa được thử nghiệm trên mô hình nhiều tầng LTSM. Vì vậy, trong bài báo này chúng tôi thử nghiệm mô hình LSTM nhiều tầng để học ra ngữ nghĩa của câu. Bài báo tập trung vào trình bày kinh nghiệm thử nghiệm mô hình LSTM nhiều tầng trên bài toán tìm câu hỏi tương đồng.

Phần tiếp theo của bài báo chúng tôi trình bày: (2) Các công việc liên quan, (3) Mô hình LSTM, (4) Các thử nghiệm và thảo luận, (5) Kết luận và công việc trong tương lai.

2. Các công việc liên quan

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:

Các phương pháp truyền thống giải quyết các bài toán CQA bằng cách biểu diễn câu hỏi sang túi từ (Bag of word) sử dụng trọng số tf.idf như mô hình BM25 [6]. Mô hình ngôn ngữ dựa vào danh mục câu hỏi [7] với mục đích cải tiến chất lượng tìm kiếm câu hỏi và câu trả lời cũng được xem xét như là một phương pháp phổ biến để mô hình hóa câu hỏi qua trình tự các túi từ. Tuy nhiên, các mô hình như vậy không thực sự hiệu quả khi câu dài. Một câu nên được trích ra nhiều phần và thực hiện so khớp với phần cụ thể của câu khác. Một mô hình cũng được sử dụng phổ biến khác, đó là mô hình LDA [8]. Đây là mô hình xác suất với mục đích học ra biểu diễn của câu qua một tập các chủ đề ẩn. Phân phối chủ đề do mô hình học ra được ứng dụng vào tìm câu hỏi tương đồng. Một hướng nghiên cứu khác là các mô hình dịch máy được sử dụng như mô hình

dịch máy dựa vào cụm từ [9]. Các mô hình này được dùng để tính độ tương đồng của câu hỏi với câu hỏi và câu hỏi với câu trả lời.

Trong hội nghị Semeval 2017, mô hình đạt kết quả cao nhất trên tập dữ liệu Semeval sử dụng các đặc trưng kỹ thuật rất phức tạp [9] như thăm dò hàm nhân hoặc trích rút đặc trưng nhân cây từ việc đi phân tích các cây cú pháp. Một nghiên cứu khác đã công bố trên tập dữ liệu này khai thác các đặc trưng độ tương tự khác nhau như độ đo cosin, độ đo Euclidean về khoảng cách từ vựng, cú pháp và ngữ nghĩa [5] để biểu diễn câu học từ mô hình SVM.

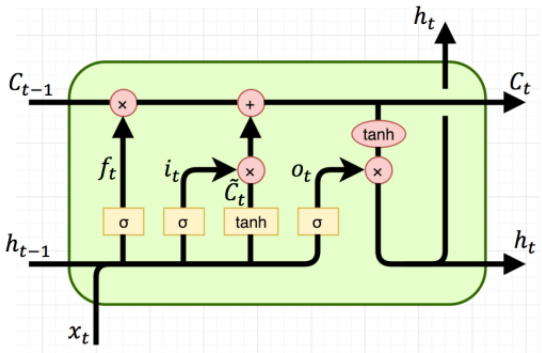
Các nghiên cứu gần đây trên bài toán lựa chọn câu hỏi và câu trả lời [10] trong hệ thống CQA mang lại hiệu quả tốt sử dụng mạng nơ ron mà không cần phải sử dụng các đặc trưng được trích rút thủ công. Các mô hình này học ra biểu diễn câu, sau đó thực hiện đo độ tương tự của câu hỏi với câu hỏi và câu hỏi với câu trả lời [7].

Các nghiên cứu sử dụng mô hình LSTM trên bài toán này tập trung vào khai thác ngữ nghĩa của câu [9] hoặc đề xuất các mô hình đặc biệt cho bài toán này. Tuy nhiên, các việc sử dụng mô hình LSTM nhiều tầng cho bài toán này hiện nay chưa được thăm dò tính hiệu quả. Trong bài báo này, chúng tôi đề xuất sử dụng mô hình LSTM nhiều tầng để thăm dò tính hiệu quả của mô hình LSTM.

3. Mô hình LSTM nhiều tầng

3.1. Mô hình LSTM

Mô hình LSTM (Long Short-Term Memory) được đề xuất đầu tiên vào năm 1997 [11]. Mô hình LSTM như Hình 1. Phần này chúng tôi giới thiệu mô hình LSTM là cơ sở của mô hình nhiều tầng được đề xuất bên dưới.



Hình 1. *Mô hình LSTM* [11]

Mô hình LSTM 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 công thức [11] trong mạng LSTM như sau:

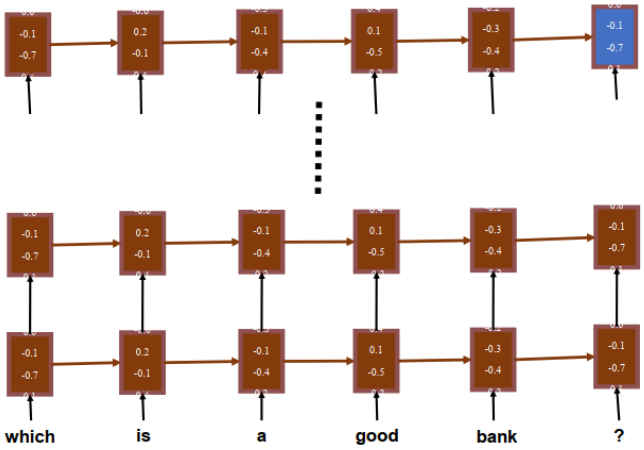
$$\begin{aligned} i_t &= \sigma\left(W^i x_t + V^i h_{t-1} + b^i\right), \\ f_t &= \sigma\left(W^f x_t + V^f h_{t-1} + b^f\right), \\ o_t &= \sigma\left(W^o x_t + V^o h_{t-1} + b^f\right), \\ c_t &= f_t \odot c_{t-1} + i_t \odot tanh(W^c x_t + V^c h_{t-1} + b^c) \\ h_t &= o_t \odot tanh(c_t) \end{aligned} \quad (1)$$

Trong đó: i, f, o là cổng vào, cổng quên và cổng ra tương ứng; h_t là các véc tơ ẩn tại mỗi bước thứ *t*; c_t là một băng chuyền ở trên mô hình LSTM, thông tin nào cần quan trọng và dùng ở sau sẽ được gửi vào và dùng khi cần. Do vậy, mô hình LSTM có thể mang thông tin từ đi xa (long term

memory), mô hình LSTM có chứa cả thông tin ngắn và dài; ma trận **W**, **V** và **b** là ma trận học từ mô hình.

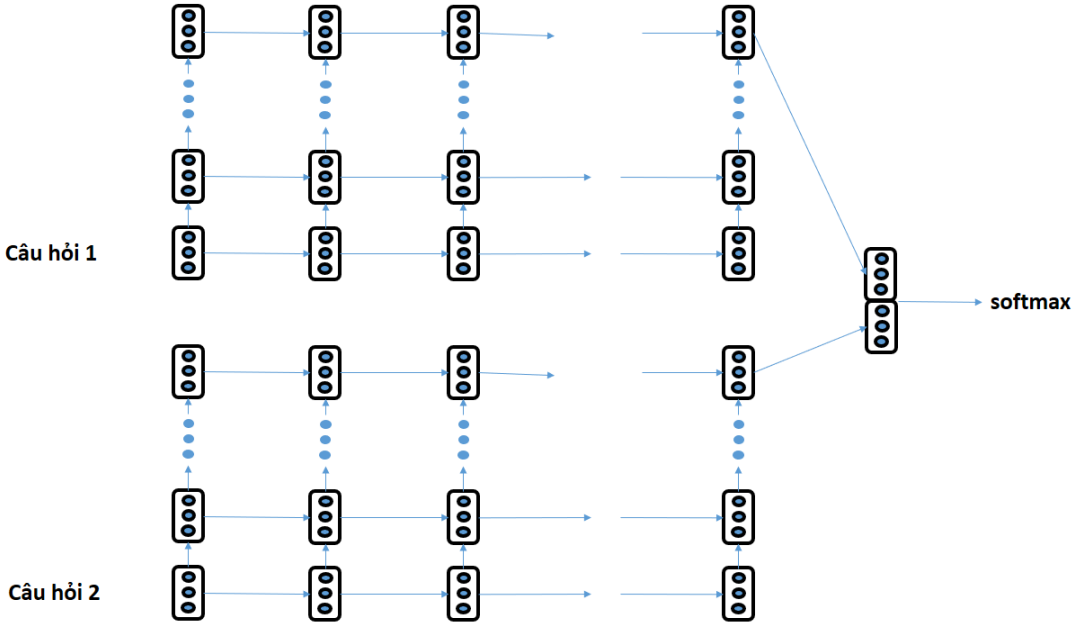
3.2. Mô hình LSTM nhiều tầng

Mô hình LSTM nhiều tầng được chúng tôi đề xuất áp dụng vào bài toán tìm kiếm câu hỏi tương đồng. Hình 2 mô tả mô hình LSTM nhiều tầng áp.



Hình 2. *Mô hình LSTM nhiều tầng*

Để đi dự đoán cặp câu hỏi chúng tôi mô tả dùng hai đường LSTM nhiều tầng học ra hai biểu diễn h_1 và h_2 tương ứng với hai câu hỏi. Cuối cùng hai véc tơ biểu diễn nối lại với nhau $h = [h_1h_2]$ và cho đi qua hàm *sofmax* để dự đoán như hình 3 sau:



Hình 3. *Mô hình LSTM sử dụng cho bài toán tìm câu hỏi tương đồng*

Hàm mất mát là hàm cross entropy [1]:

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

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

Để đánh giá mô hình đề xuất, chúng tôi sử dụng tập dữ liệu Semeval 2017. Tập dữ liệu này được lấy từ forum về cuộc sống ở Qatar (https://www.qatarliving.com/) [9] và được gán nhãn và bao gồm 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 tiếng Anh - Semeval 2017.

Bảng 1. *Bảng thống kê cặp câu hỏi trong tập dữ liệu Semeval 2017* [9]

	Semeval 2017
Tập huấn luyện	3170
Tập phát triển	700
Tập kiểm thử	880

Chúng tôi sử dụng độ đo MAP (mean Average Precision) [9] để đánh giá hiệu quả của mô hình đề xuất.

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

Trong đó, N là số câu trong tập kiểm thử, R_{jk} là tập kết quả tìm kiếm được xếp hạng từ kết quả tốt nhất cho tới khi tìm thấy câu hỏi thứ *k*, m_j là số câu hỏi đúng của câu hỏi thứ j trong số N câu hỏi.

4.2. Tham số của mô hình

Chúng tôi sử dụng biểu diễn từ Word2vec 300 chiều đưa vào mô hình ở lớp đầu vào. Các từ không nằm trong tập từ điển được khởi tạo một cách ngẫu nhiên. Số chiều lớp ẩn trong mô hình LSTM 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 thiết lập là 0,0001; batch-size là 64, drop-out là 30%. Mô hình được thực thi trên tensorflow và chạy trên GPU Nvidia Tesla p100 16Gb. Chúng tôi đánh giá hiệu năng của mô hình trên tập phát triển và chọn tham số tốt nhất trên tập phát triển để thiết lập tham số thử nghiệm trên tập kiểm thử.

4.3. Kết quả

Bảng 2 biểu diễn kết quả thử nghiệm trên các mô hình:

Bảng 2. *Kết quả của mô hình đề xuất*

Mô hình	**MAP**
LSTM	40,03
LSTM 2 tầng	41,00
LSTM 3 tầng	**41,43**
LSTM 4 tầng	40,23
LSTM 5 tầng	39,38

Trước hết chúng tôi thử nghiệm trên mô hình LSTM gốc để đi dự đoán cặp câu hỏi. Mô hình LSTM gốc cho kết quả 40,03% trên độ đo MAP. Sau đó, chúng tôi thử nghiệm mô hình đề xuất LSTM nhiều tầng. Kết quả chỉ ra rằng, khi chồng 2 tầng LSTM kết quả tăng lên 1% so với mô hình LSTM ban đầu. Khi thử nghiệm trêm mô hình LSTM 3 tầng, kết quả đạt giá trị cao nhất là 41,23%. Sau đó, khi thử nghiệm trên số tầng nhiều hơn là 4 và 5 tầng kết quả giảm dần. Do đó chúng tôi lựa chọn thử nghiệm trêm mô hình LSTM 3 tầng. Điều đó chứng tỏ rằng, khi khai thác ngữ nghĩa trên nhiều mức độ thì ảnh hưởng tới kết quả của bài toán và có khả năng khai thác ngữ nghĩa của câu tốt hơn. Điều này cũng được chứng minh trong bài báo khi sử dụng mô hình LSTM nhiều tầng trong bài toán dịch máy [12].

5. Kết luận và công việc trong tương lai

Trong bài báo này chúng tôi đã đề xuất sử dụng mô hình LSTM nhiều tầng cho bài toán tìm câu hỏi tương đồng. Qua thực nghiệm, chúng tôi thấy rằng, việc sử dụng nhiều tầng LSTM cũng ảnh hưởng tới kết quả dự đoán cặp câu hỏi tương đồng. Trong tương lai, chúng tôi sẽ tiến hành khảo sát các phương pháp biểu diễn câu khác nhau thay vì sử dụng véc tơ ẩn tại lớp cuối dùng để đi dự đoán.

Lời cảm ơn

Chúng tôi xin cảm ơn đề tài có mã số T2022-07-04 đã hỗ trợ một phần kinh phí để chúng tôi thực hiện công việc này.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] G. Zhou, Y. Chen, D. Zeng, and J. Zhao, "Towards faster and better retrieval models for question search," In *Proceedings of the 22nd ACM International Conference on Information Knowledge Management*, CIKM13, New York, NY, USA. Association for Computing Machinery, 2013, pp. 2139-2148.

[2] G. Zhou, T. He, J. Zhao, and P. Hu, "Learning continuous word embedding with metadata for question retrieval in community question answering," *CIKM13*, vol. 01, pp. 250-259, 2015.

[3] L. Cai, G. Zhou, K. Liu, and J. 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, November. Asian Federation of Natural Language Processing, 2011, pp. 273-281.

[4] W. Wu, X. Sun, and H. 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, July. Association for Computational Linguistics, 2018, pp. 1746-1755.

[5] Y. Tay, A. T. Luu, and S. C. Hui, "Enabling efficient question answer retrieval via hyperbolic neural networks," *CoRR*, pp. 265-274, 2017, doi: abs/1707.07847.

[6] S. 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)*, January, 1995.

[7] X. Cao, G. Cong, B. Cui, C. S. Jensen, and C. Zhang, "The use of categorization information in language models for question retrieval," In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, New York, NY, USA. Association for Computing Machinery, 2019, pp. 265-274.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pp. 601-608. MIT Press, 2002.

[9] P. Nakov, D. Hoogeveen, L. Màrquez, A. Moschitti, H. Mubarak, T. Baldwin, and K. Verspoor, "SemEval-2017 task 3: Community question answering," In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, August. Association for Computational Linguistics, 2017, pp. 27-48.

[10] S. Filice, G. Da San Martino, and A. 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, August. Association for Computational Linguistics, 2017, pp. 326-333.

[11] 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].

[12] D. Britz, A. Goldie, M.-T. Luong, and Q. Le, "Massive Exploration of Neural Machine Translation Architectures," In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics, 2017, pp. 1442-1451.