Utilizing BERT for Question Retrieval in Vietnamese E-commerce Sites

Thi-Thanh Ha
HaNoi Uni. of Science and Technology, VietNam
ThaiNguyen Uni. of Information and Communication Technology, VietNam
htthanh@ictu.edu.vn

Van-Nha Nguyen
HaNoi Uni. of Science and Technology
nha282@gmail.com

Kiem-Hieu Nguyen
HaNoi Uni. of Science and Technology
hieunk@soict.hust.vn

Kim-Anh Nguyen
HaNoi Uni. of Science and Technology
anhnk@soict.hust.vn

Tien-Thanh Nguyen
HaNoi Uni. of Science and Technology
20144052@student.hust.edu.vn

Abstract

Question retrieval is an important task in question answering. This task is considered to be challenging due to the lexical gap issue, i.e., similar questions could be expressed in different words or phrases. Although there are numerous researches conducted on question retrieval task in English, the corresponding problem in Vietnamese hasn't been studied much. In this investigation, we highlight our efforts on question retrieval in Vietnamese e-commerce sites majorly in two directions: (1) Building a Vietnamese dataset for question retrieval in e-commerce domain. (2) Conducting experiments using recent deep learning techniques including BERT-based classifiers. Our results provide practical examples of effectively employing these models on Vietnamese e-commerce data. Particularly, we demonstrate that a BERT model trained on e-commerce texts yields significant improvement on question retrieval over BERT trained on general-domain texts.

1 Introduction

Community-based Question Answering (CQA) systems[1][2] have become an increasingly popular online platform. Community websites, where users can post their own questions or answers to other users' questions, provide frameworks for people with dissimilar backgrounds to share their knowledge and experiences. When a user posts a new question on a community website, it usually takes a while for other users to respond. Moreover, in a certain period of time, the number of questions and answers stored in a database gradually becomes enormous and challenging to handle, which means that the possibility of finding duplicated questions increases. As a result, it is time-consuming to retrieve good answers to a given question in an archive of question-answer pairs. In order to reduce latency, CQA systems should automatically find questions which are similar to a given new question. It is hoped that the answers of these related questions could be useful for the new question.

The problem of question retrieval is defined as follows: Given a query question and a set of existing questions, return the most similar questions to the query. Question retrieval has been extensively investigated with the purpose of answering new questions using previous answers in databases [Zhou et al.2013, Zhou et al.2015]. Previous studies delved into the lexical gap challenge in which query question might contain words and phrases different from its similar questions. Figure 1 is a typical pair of similar questions in our Vietnamese dataset.

In order to deal with lexical gap challenge, previous research applied soft alignment technique originated from machine translation or implicitly disambiguated word meaning using topic models [Cai et al.2011]. A huge number of research methods in recent years have focused on end-to-end approaches based on deep neural networks without depending on feature engineering or external knowledge bases [Wu et al.2018, Tay et al.2017]. These approaches leverage pre-trained embeddings and specific-purpose network structures aiming at

[1] https://stackoverflow.com/
[2] https://www.qatarliving.com/

Question 1: Làm ơn chỉ giùm tôi cách tắt phím slide to unclock trên samsung s9 plus
(Can you please show me how to turn off slide to unlock button on samsung s9 plus)
Question 2: Cách tắt màn hình slide to unclock chỉ để màn hình kiểu vuốt để mở khóa máy ss j7 pro
(how to turn off slide to unlock screen on ss j7 pro)

Figure 1: An example of similar question pair

representing syntactic and semantic information in questions. Until recently, BERT, a pre-trained language model, achieves state-of-the-art performance in many natural language processing (NLP) tasks [Devlin et al.2018]. However, to our knowledge, BERT has not been applied to Vietnamese question retrieval.

In the scope of this paper, we advocate: (1) A public CQA Vietnamese dataset in E-commerce domain for question retrieval problem. (2) Experimentation with various deep learning models on this dataset. (3) Empirical findings on tuning and visualizing attention of these models. (4) A pre-trained BERT embedding model for Vietnamese E-commerce texts.

2 Related Work

Over the recent years, numerous methods have been proposed to deal with community question answering tasks and achieved state-of-the-art results.

Traditional methods attempt to deal with CQA problems by transforming the text in questions into Bag-of-Words (BoW) representation with tf-idf weighting scheme, such as BM25 [Robertson et al.1995]. Count-based language models [Cao et al.2009] have also been considered as a popular method to model questions as sequences instead of bags of words. Nonetheless, such models might not be useful when there are a vast number of possible sequences. A sentence should have an exact pattern, such as string or word sequence, matching to a particular part of another sentence. Another popular model based on semantic similarity is Latent Dirichlet Allocation (LDA) [Blei et al.2002], which is a probabilistic model applied in representing questions through a set of latent topics. The learned topic distribution is then applied to retrieve similar historical questions. In another direction, various methods have been developed based on machine translation techniques, such as the monolingual phrase-based translation model, to measure question similarity [Jeon et al.2005] or question-answer similarity.

Top performing systems in SemEval 2017 Task 3 challenge [Nakov et al.2017] use sophisticated feature engineering such as exploiting kernel functions or extracting tree kernel features from parse trees. For instance, the best-performance system [Filice et al.2017], uses similarity features like cosine distance or Euclidean distance and lexical, syntactic, semantic and distributed representations to learn an SVM classifier.

Recent studies in question retrieval and answer selection [Severyn and Moschitti2015, Tan et al.2015] in CQA highlight the effectiveness of neural network models over time-consuming handcrafted feature engineering. These methods learn distributed vector representation of texts and measure question-question or question-answer similarity for question retrieval or answer selection, respectively [Bonadiman et al.2017, Severyn and Moschitti2015].

BERT (Bidirectional Encoder Representations from Transformers) was proposed in [Devlin et al.2018] as a kind of pre-trained transformer network [Vaswani et al.2017], which was applied to various NLP tasks with state-of-the-art performance, including sentence classification, question answering, and sentence pair regression. Several prior studies substantiate that BERT could perform well in many cases [Liu et al.2019, Hao et al.2019]. Particularly, [Liu et al.2019] illustrated that the performance of BERT can be further improved by some small adjustments in the pre-training process. Besides, [Hao et al.2019] focused on the interpretation of self-attention, which is one of the most fundamental components of BERT.

Prior researches were generally conducted on English datasets. In this paper, we explore how well recent deep learning models, especially pre-trained BERT, could possibly perform on Vietnamese. At the same time, we visualize some attention layers to illustrate the effectiveness of BERT models on Viet-

namese.

3 BERT for Vietnamese Question Retrieval

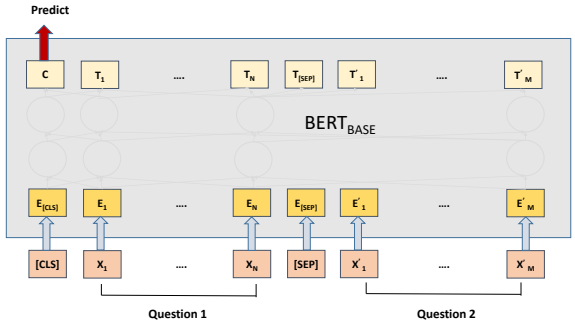


Figure 2: BERT for question retrieval [Devlin el al.2018]

3.1 BERT

BERT is a Bidirectional Encoder Representations achieved from Transformers [Devlin el al.2018] that generates a sentence representation by jointly learning two tasks: masked language modeling and next-sentence prediction. BERT models can be fine-tuned well on both sentence level as well as word level tasks.

BERT has a deep architecture, which has 12 layers of 768 hidden size and 12 self-attention heads. This model begins from the word embeddings layer. In 12 layers, multi-headed attention is calculated using word representations of the previous layer to generate a new intermediate representation. As a result, a token will have 12 intermediate representations with the same size.

In the masked language modeling task, 15% of the tokens are chosen at random to obtain bi-directional pre-trained language model. To avoid mistmatching between pre-training and fine-tuning, in those 15% tokens, a token is replaced with [MASK] 80% of times, 10% of times it is replaced by another random token, and the rest 10% of times it is unchanged.

In the next-sentence prediction task, given a pair of sentences, the aim of this task is to predict whether the second sentence is the true next sentence of the first one.

3.2 BERT for Vietnamese Question Retrieval

In this paper, we apply Multilingual BERT-BASE model (Figure 2), which is considered to be effective on small datasets. It is proved to be good at the ability of cross-lingual generalization by a multilingual representation without being explicitly trained.

Our experiments consist of two parts: Pre-training BERT on unlabeled 1.1M texts of Vietnamese E-commerce (see table 2); and fine-tuning for question retrieval problem on a labeled E-commerce dataset. The parameters of all the layers of our model are fine-tuned at once. A special classification token ([CLS]) and separation token ([SEP]) are added as inputs of our model as followed: $Bert - Input(q_1, q_2) = [CLS]q_1[SEP]q_2[SEP]$, where q_1, q_2 are two questions. The final hidden state corresponding to [CLS] token is applied as an aggregate sequence representation for classification tasks. *Softmax* activation in the last layer is used to predict the label of the considering question.

4 Dataset

We collected questions from users in QA section of The gioi Di dong - an e-commerce website on mobiles, laptops and other electronic devices[3]. An ElasticSearch engine was built from the corpus. We selected a random subset as original questions. Each question was put into ElasticSearch as query. Thereafter, for the first 10 returned questions, human annotators were asked to assess their equivalence to the original question. To increase the difficulty of the task, we removed original questions that could be easily handled by ElasticSearch (i.e questions that have little lexical gap challenge).

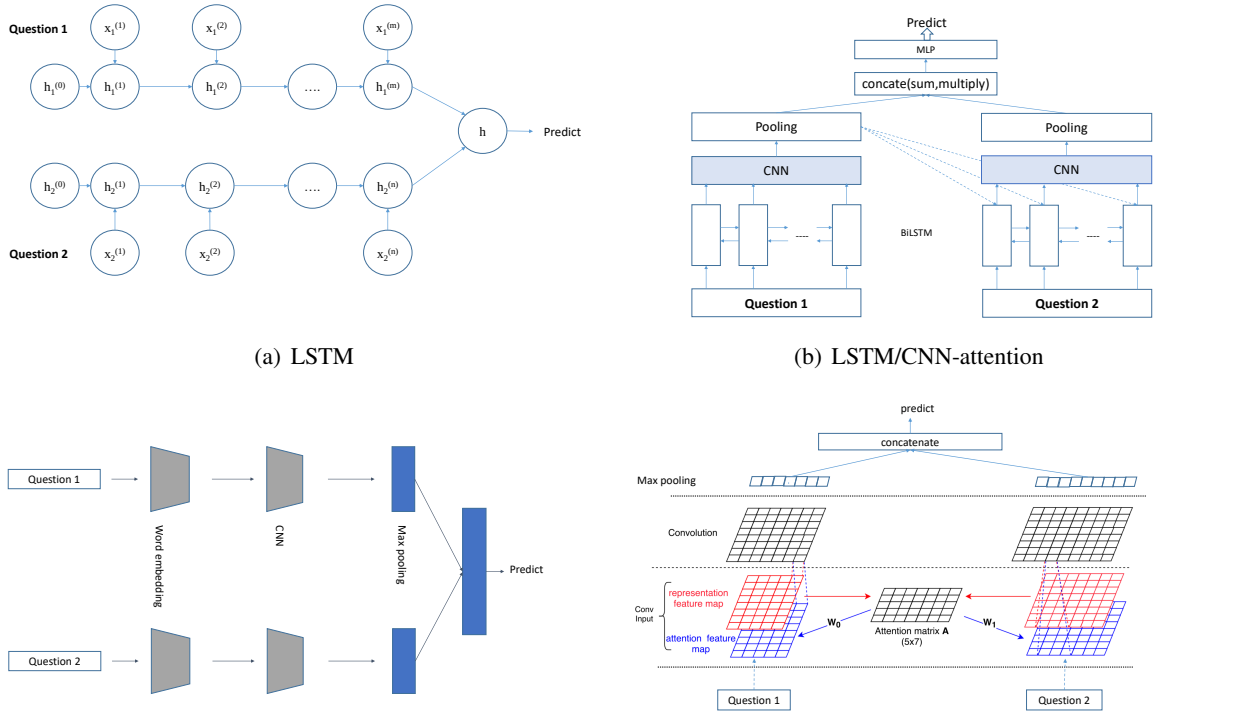
We divided annotated data into three separated sets: training, development, and test (Table 1). In average, 30% of questions were annotated as relevant to the original question.

We also use the large corpus for pre-trained embeddings (Table 2).

5 Experiments and discussions

Our models were implemented using Tensorflow and all experiments were conducted on GPU Nvidia Tesla p100 16Gb. We used Mean Average Precision

[3]https://www.thegioididong.com/hoi-dap



(a) LSTM

(b) LSTM/CNN-attention

(c) CNN

(d) ABCNN

Figure 3: Baseline deep learning models in question retrieval

	Pairs of questions
Train	5,996
Dev	847
Test	1,068
Average length (syllable)	27
Vocabulary (syllable)	5,821

Table 1: Statistics of Thegioididong dataset.

Corpus size	1.1M
Vocabulary size (syllable)	151,735
Average length (syllable)	31

Table 2: Statistics of unlabeled corpus crawled from The gioi Di dong.

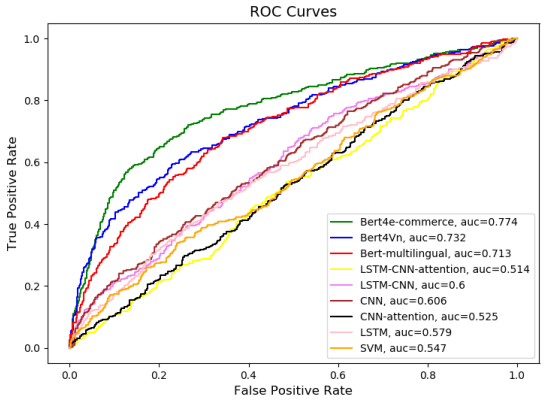


Figure 4: The ROC curves of prediction models.

(MAP) for evaluation. Hyper-parameters were tuned on the development set.

Table 3 presents detailed experimental results on Thegioididong. The results are divided into three parts: vanilla neural networks with LSTM/CNN encoder; BERT pre-trained on different corpora; and baseline bag-of-word models. In all models except PhoBERT, we used syllables as unit input. In PhoBERT, we used its built-in module for word segmentation[4].

Figure 4 illustrates the accuracy of nine models.

[4]https://github.com/VinAIResearch/PhoBERT

Models	MAP
LSTM	52.60
CNN	53.10
ABCNN	51.52
LSTM attention	55.50
BERT-multilingual	61.06
BERT4Vn	63.75
PhoBERT	65.50
BERT4ecommerce	**70.50**
ElasticSearch	52,00
SVM	49.75

Table 3: MAP score of models on Vietnamese dataset.

In general, both Table 3 and Figure 4 show that deep learning approach is better than baseline models; and there was a substantial rise of BERT models, especially when pre-trained on domain data.

5.1 LSTM/CNN Networks

Figure 3 shows the architecture of our models.

- **LSTM**: Both questions are encoded by a shared-weight bi-directional LSTM. The representation of each question is concatenation of the last hidden units of each direction. The representations of two questions are concatenated and is fed into an MLP for prediction.
- **CNN**: Bi-directional LSTM building-block is replaced by CNN.
- **ABCNN** [Yin et al.2015]: This model employs an attention feature matrix to influence convolution. Attention matrix is generated by matching units of the first question representation feature map with units of the second question representation feature map. It can be viewed as a new feature map of two questions to put into next layer.
- **LSTM/CNN-attention**: In this model, outputs from all words of both questions are passed through a word-wise dot product to create a word-by-word attention alike matrix. Updated hidden vector of both question from attention serve as inputs of CNN structure. A global max pooling is then applied to collect important features before prediction. This model is close to LSTM siamese networks as in [Tan et al.2016].

We pre-train syllable embeddings using word2vec on the unlabeled e-commerce corpus. Embedding layers were initialized by pre-train vectors. Adam [Kingma and Ba2014] is used as optimization function. Hyper-parameters used in each experiment are shown in Table 4.

As shown in Table 3, simple concatenation of output from LSTM/CNN and using MLP for prediction slightly outperform baseline models. Learning attention weights as in ABCNN even hurts the performance. In LSTM/CNN-attention, directly calculating word-by-word attention using dot product results in significant improvement.

5.2 Pre-training and Fine-tuning Bert

BERT experiments are performed using Multiligual BERT-BASE model[5]. We first pre-trained BERT on unlabeled E-commerce Vietnamese with maximum length of 200, batch size of 32, and learning rate of $2e^{-5}$ with 20000 steps. We call this model BERT4ecommerce. After pre-training, our model was fine-tuned on question retrieval using Thegioididong dataset.

We also compare our in-domain pre-trained model with other general-domain pre-train BERTs:

- **BERT-multilingual** [Pires et al.2019]: 110K wordpiece vocab, pre-trained on Vietnamese Wikipedia corpus
- **BERT4Vn**[6]: Pre-trained on 500M words of Vietnamese news.

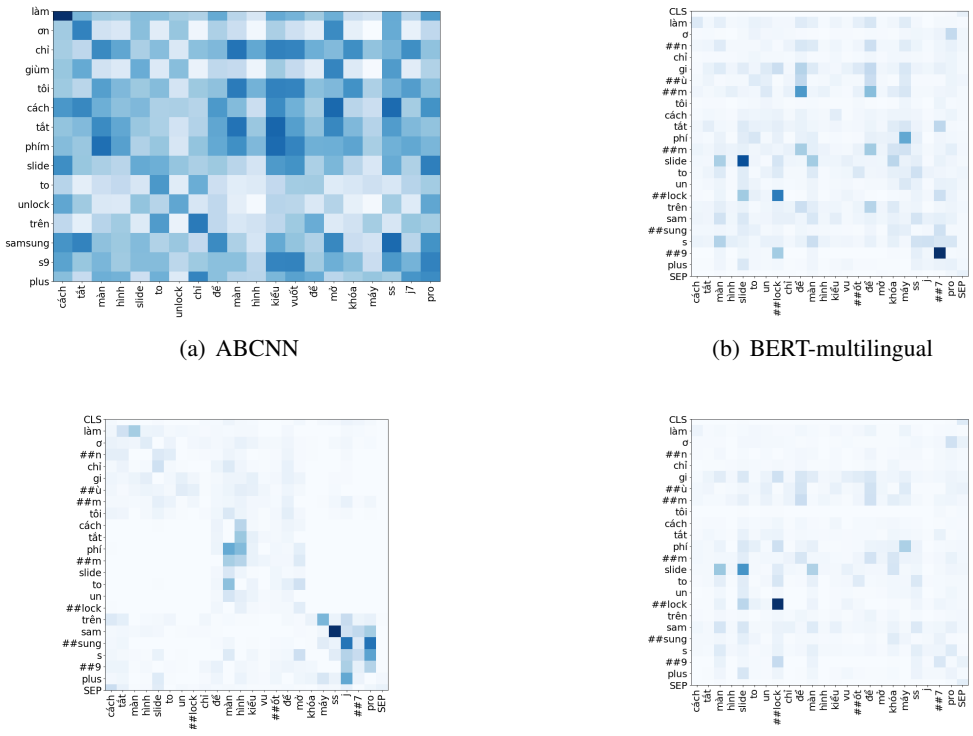
As both shown in Table 3 and Figure 4, significant improvement was obtained by using BERT. Especially, Bert4E-commerce achieved the highest performance (70.50% in MAP, 77.4% in AUC). These experiments advance the idea that when source domain used in pre-training model and target domain are the same, it could have good impact on the final result. E-commerce vocabulary consists of a wide range of words used for technological devices such as Iphone, Samsung S9, "mua-tra-gop" (pay by installments) and so on. Moreover, E-commerce data or social data in general has no guarantee in spelling, grammar and word usage. For instance, numerous spelling mistakes and abbreviations such as

[5]https://github.com/google-research/bert

[6]https://github.com/lampts/bert4vn

	Emb-size	Hid-size/filter-size	L-rate	P_{drop}	Batch size	epochs	Params (x10^5)
LSTM	300	300	0.0001	0.2	64	25	21
LSTM/CNN-attention	300	300	0.0001	0.2	64	25	27
CNN	300	3	0.003	0.5	64	25	33
ABCNN	300	3	0.001	0.2	32	25	34

Table 4: The hyper-parameters set of LSTM/CNN models



(a) ABCNN (b) BERT-multilingual

(c) BERT4Vn (d) BERT4ecommerce

Figure 5: Visualization of BERT and ABCNN

"thoong bao" (notification), "mk" (password) "ss" (Samsung), "f" (keyboard) were found in our dataset. Thus, retraining word embedding on E-commerce domain is required and much more effective than using pre-trained model on news source data such as Wiki and news in this situation.

5.3 Word-based BERT

So far, all our models were based on syllables. In this section, we use a word-based BERT model and apply it to segmented questions. We chose PhoBERT [Nguyen and Nguyen2020], a pre-trained model on 3B segmented texts from Wikipedia and news.

Results show that PhoBERT performs better than BERT-multilingual and BERT4Vn which indicates that word segmentation is helpful for question retrieval in in-domain social texts. Nevertheless, without word segmentation, BERT pre-trained on in-domain texts still outperforms PhoBERT in a large margin. This result is encouraging as word segmentation in in-domain texts suffers from unknown words and spelling mistakes that could propagate errors to downstream tasks.

5.4 Attention visualization

It is argued in [Wiegreffe and Pinter2019] that attention can be use to explain model prediction. In this section, we visualize attention of BERT4 and

	max-length	learning rate	steps reach max
BERT-multilingual	200	$2e^{-5}$	650
BERT4Vn	200	$2e^{-5}$	1600
PhoBERT	200	$2e^{-5}$	1000
BERT4ecommerce	200	$2e^{-5}$	900

Table 5: The hyper-parameters set of fine-turning BERT models

ABCNN to point out that self attention of BERT could learn semantic relationship in questions better than some commonly known attention mechanism such as ABCNN. An attention matrix of Bert was extracted from the first attention layer.

Figure 5 visualizes word-by-word attention between query question (Y-axis) and candidate question (X-axis). This visualization presents alignment weights between two questions, where darker color correlates with larger value.

The attention distribution of BERT is sparser than that of ABCNN. This helps to strengthen interaction between important words such as 'slide' with 'màn hình' (screen), 'lock', 'tắt phím' with 'khóa máy' as seen in the example. The research in [Cui et al.2019] shows that sparse attention matrix achieved from BERT leads to a more interpretable representation of inputs.

6 Conclusion

We carried out a range of experiments with LSTM, LSTM attention, CNN, ABCNN and fine-tuning BERT for question retrieval on a Vietnamese dataset. In particular, our BERT model pre-trained on an ecommerce corpus could be useful for related research.

We hope our work can give a boost to applications related to CQA on Vietnamese Ecommerce data. In the future, we are going to investigate the effect of word segmentation to question answering in ecommerce domain.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2002. Latent dirichlet allocation. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 601–608. MIT Press.

Daniele Bonadiman, Antonio E. Uva, and Alessandro Moschitti. 2017. Multitask learning with deep neural networks for community question answering. *CoRR*, abs/1702.03706.

Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao. 2011. Learning the latent topics for question retrieval in community QA. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 273–281, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Xin Cao, Gao Cong, Bin Cui, Christian Søndergaard Jensen, and Ce Zhang. 2009. 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, page 265–274, New York, NY, USA. Association for Computing Machinery.

Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. 2019. Fine-tune BERT with sparse self-attention mechanism. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3548–3553, Hong Kong, China, November. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Simone Filice, Giovanni Da San Martino, and Alessandro Moschitti. 2017. 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)*, pages 326–333, Vancouver, Canada, August. Association for Computational Linguistics.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019. Visualizing and understanding the effectiveness of bert. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and an-

swer archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, page 84–90, New York, NY, USA. Association for Computing Machinery.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48, Vancouver, Canada, August. Association for Computational Linguistics.

Dat Quoc Nguyen and A. Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *ArXiv*, abs/2003.00744.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, January.

Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 373–382, New York, NY, USA. Association for Computing Machinery.

Ming Tan, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *CoRR*, abs/1511.04108.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 464–473, Berlin, Germany, August. Association for Computational Linguistics.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2017. Enabling efficient question answer retrieval via hyperbolic neural networks. *CoRR*, abs/1707.07847.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November. Association for Computational Linguistics.

Wei Wu, Xu Sun, and Houfeng Wang. 2018. 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)*, pages 1746–1755, Melbourne, Australia, July. Association for Computational Linguistics.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2015. ABCNN: attention-based convolutional neural network for modeling sentence pairs. *CoRR*, abs/1512.05193.

Guangyou Zhou, Yubo Chen, Daojian Zeng, and Jun Zhao. 2013. Towards faster and better retrieval models for question search. In *Proceedings of the 22nd ACM International Conference on Information Knowledge Management*, CIKM13, page 2139–2148, New York, NY, USA. Association for Computing Machinery.

Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. 2015. Learning continuous word embedding with metadata for question retrieval in community question answering. pages 250–259, 01.