Crossing Variational Autoencoders for Answer Retrieval

Wenhao Yu[†], **Lingfei Wu**[‡], **Qingkai Zeng**[†], **Yu Deng**[‡], **Shu Tao**[‡], **Meng Jiang**[†]
†University of Notre Dame, Notre Dame, IN, USA
‡IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA
{wyu1, qzeng, mjiang2}@nd.edu, {wuli, dengy, shutao}@us.ibm.com

arXiv:2005.02557v1 [cs.CL] 6 May 2020

Abstract

Answer retrieval is to find the most aligned answer from a large set of candidates given a question. Learning vector representations of questions/answers is the key factor. Question-answer alignment and question/answer semantics are two important signals for learning the representations. Existing methods learned semantic representations with dual encoders or dual variational auto-encoders. The semantic information was learned from language models or question-to-question (answer-to-answer) generative processes. However, the alignment and semantics were too separate to capture the *aligned semantics* between question and answer. In this work, we propose to *cross* variational auto-encoders by generating questions with aligned answers and generating answers with aligned questions. Experiments show that our method outperforms the state-of-the-art answer retrieval method on SQuAD.

1 Introduction

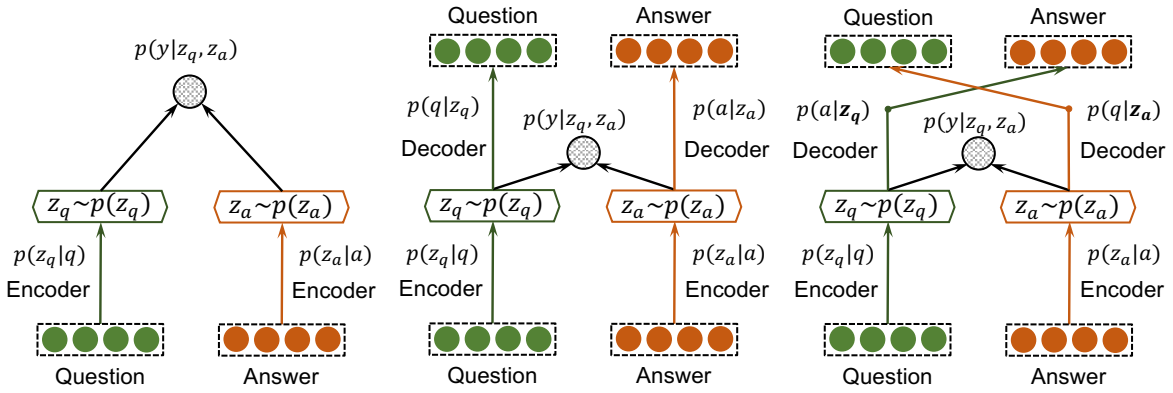
Answer retrieval is to find the most aligned answer from a large set of candidates given a question (Ahmad et al., 2019; Abbasiyantaeb and Momtazi, 2020). It has been paid increasing attention by the NLP and information retrieval community (Yoon et al., 2019; Chang et al., 2020). Sentence-level answer retrieval approaches rely on learning vector representations (i.e., embeddings) of questions and answers from pairs of question-answer texts. The question-answer alignment and question/answer semantics are expected to be preserved in the representations. In other words, the question/answer embeddings must reflect their semantics in the texts of being aligned as pairs.

One popular scheme "Dual-Encoders" (also known as "Siamese network" (Triantafillou et al., 2017; Das et al., 2016)) has two separate encoders to generate question and answer embeddings and a predictor to match two embedding vectors (Cer et al., 2018; Yang et al., 2019). Unfortunately, it has been shown difficult to train deep encoders with the weak signal of matching prediction (Bowman et al., 2015). Then there has been growing interests in developing deep generative models such as variational auto-encoders (VAEs) and generative adversial networks (GANs) for learning text embeddings (Xu et al., 2017; Xie and Ma, 2019). As shown in Figure 1(b), the scheme of "Dual-VAEs" has two VAEs, one for question and the other for answer (Shen et al., 2018). It used the tasks of generating reasonable question and answer texts from latent spaces for preserving semantics into the latent representations.

Table 1: The answer at the bottom of this table was aligned to 17 different questions at the sentence level.

Question (1): What three stadiums did the NFL decide between for the game?
Question (2): What three cities did the NFL consider for the game of Super Bowl 50?
...
Question (17): How many sites did the NFL narrow down Super Bowl 50's location to?
Answer: The league eventually narrowed the bids to three sites: New Orleans Mercedes-Benz Superdome, Miami Sun Life Stadium, and the San Francisco Bay Area's Levi's Stadium.

Although Dual-VAEs was trained jointly on question-to-question and answer-to-answer reconstruction, the question and answer embeddings can only preserve *isolated* semantics of themselves. In the model, the Q-A alignment and Q/A semantics were too separate to capture the *aligned semantics* (as we mentioned at the end of the first paragraph) between question and answer. Learning the alignment with the weak Q-A matching signal, though now based on generatable embeddings, can lead to confusing results, when (1) dif-



(a) Dual-Encoders (Yang et al., 2019) (b) Dual-VAEs (Shen et al., 2018) (c) Dual-CrossVAEs (**Ours**)

Figure 1: (a)–(b) The Q-A alignment and Q/A semantics were learned too separately to capture the aligned semantics between question and answer. (c) We propose to cross VAEs by generating questions with aligned answers and generating answers with aligned questions.

ferent questions have similar answers and (2) similar questions have different answers. Table 1 shows an examples in SQuAD: 17 different questions share the same sentence-level answer.

Our idea is that if aligned semantics were preserved, the embeddings of a question would be able to generate its answer, and the embeddings of an answer would be able to generate the corresponding question. In this work, we propose to *cross* variational auto-encoders, shown in Figure 1(c), by reconstructing answers from question embeddings and reconstructing questions from answer embeddings. Note that compared with Dual-VAEs, the encoders do not change but decoders work across the question and answer semantics.

Experiments show that our method improves MRR and R@1 over the state-of-the-art method by 1.06% and 2.44% on SQuAD, respectively. On a subset of the data where any answer has at least 10 different aligned questions, our method improves MRR and R@1 by 1.46% and 3.65%, respectively.

2 Related Work

Answer retrieval (AR) is defined as the answer of a candidate question is obtained by finding the most similar answer between multiple candidate answers (Abbasiyantaeb and Momtazi, 2020). While another popular task on SQuAD dataset is machine reading comprehension (MRC), which is introduced to ask the machine to answer questions based on one given context (Liu et al., 2019). In this section, we review existing work related to answer retrieval and variational autoencoders.

Answer Retrieval. It has been widely studied with information retrieval techniques and has received increasing attention in the recent years by considering deep neural network approaches. Recent works have proposed different deep neural models in text-based QA which compares two segments of texts and produces a similarity score. Document-level retrieval (Chen et al., 2017; Wu et al., 2018; Seo et al., 2018, 2019) has been studied on many public datasets including including SQuAD (Rajpurkar et al., 2016), MsMarco (Nguyen et al., 2016) and NQ (Kwiatkowski et al., 2019) etc. ReQA proposed to investigate sentence-level retrieval and provided strong baselines over a reproducible construction of a retrieval evaluation set from the SQuAD data (Ahmad et al., 2019). We also focus on sentence-level answer retrieval.

Variational Autoencoders. VAE consists of encoder and generator networks which encode a data example to a latent representation and generate samples from the latent space, respectively (Kingma and Welling, 2013). Recent advances in neural variational inference have manifested deep latent-variable models for natural language processing tasks (Bowman et al., 2016; Kingma et al., 2016; Hu et al., 2017a,b; Miao et al., 2016). The general idea is to map the sentence into a continuous latent variable, or code, via an inference network (encoder), and then use the generative network (decoder) to reconstruct the input sentence conditioned on samples from the latent code (via its posterior distribution). Recent work in cross-modal generation adopted cross alignment VAEs to jointly learn rep-

resentative features from multiple modalities (Liu et al., 2017; Shen et al., 2017; Schonfeld et al., 2019). DeConv-LVM (Shen et al., 2018) and VAR-Siamese (Deudon, 2018) are most relevant to us, both of which adopt Dual-VAEs models (see Figure 1(b)) for two text sequence matching task. In our work, we propose a Cross-VAEs for questions and answers alignment to enhance QA matching performance.

3 Proposed Method

Problem Definition. Suppose we have a question set $\mathcal{Q}$ and an answer set $\mathcal{A}$. Each question and answer have only one sentence. Each question $q \in \mathcal{Q}$ and answer $a \in \mathcal{A}$ can be represented as (q, a, y), where y is a binary variable indicating whether q and a are aligned. Therefore, the solution of sentence-level retrieval task could be considered as a matching problem. Given a question q and a list of answer candidates $C(q) \subset \mathcal{A}$, our goal is to predict $p(y|q, a)$ of each input question q with each answer candidate $a \in C(q)$.

3.1 Crossing Variational Autoencder

Learning cross-domain constructions under generative assumption is essentially learning the conditional distribution $p(q|z_a)$ and $p(a|z_q)$ where two continuous latent variables $z_q, z_a \in \mathbb{R}^{d_z}$ are independently sampled from $p(z_q)$ and $p(z_a)$:

$$p(q|a) = \mathbb{E}_{z_a \sim p(z_a|a)}[p(q|z_a)], \quad (1)$$

$$p(a|q) = \mathbb{E}_{z_q \sim p(z_q|q)}[p(a|z_q)]. \quad (2)$$

The question-answer pair matching can be represented as the conditional distribution $p(y|z_q, z_a)$ from latent variables $p(q|z_a)$ and $p(a|z_q)$:

$$p(y|q, a) = \mathbb{E}_{z_q \sim p(z_q|q), z_a \sim p(z_a|a)}[p(y|z_q, z_a)], \quad (3)$$

Objectives. We denote E_q and E_a as question and answer encoders that infer the latent variable z_q and z_a from a given question answer pair (q, a, y), and D_q and D_a as two different decoders that generate corresponding question and answer q and a from latent variables z_a and z_q. Then, we have cross construction loss:

$$\begin{aligned}\mathcal{L}_{cross}(\theta_E,\theta_D) \\ =& y \cdot \mathbb{E}_{q\sim Q}[-\log p_D(q|a, E(a))] \\ &+ y \cdot \mathbb{E}_{a\sim A}[-\log p_D(a|q, E(q))].\end{aligned} \quad (4)$$

Variational Autoencoder (Kingma and Welling, 2013) imposes KL-divergence regularizer to align both posteriors $p_E(z_q|q)$ and $p_E(z_a|a)$:

$$\begin{aligned}\mathcal{L}_{KL}(\theta_E) =& y \cdot \mathbb{E}_{q\sim Q}[D_{KL}(p_E(z_q|q)||p(z_q))] \\ &+ y \cdot \mathbb{E}_{a\sim A}[D_{KL}(p_E(z_a|a)||p(z_a))],\end{aligned} \quad (5)$$

where θ_E, θ_D are all parameters to be optimized. Besides, we have question answer matching loss from $f_\phi(y|q, a)$ as:

$$\begin{aligned}\mathcal{L}_{matching}(\phi_f) = -\big[& y \cdot \log p_{f_\phi}(y|z_q, z_a) \\ &+ (1-y) \cdot \log(1 - p_{f_\phi}(y|z_q, z_a))\big],\end{aligned} \quad (6)$$

where f is a matching function and ϕ_f are parameters to be optimized. Finally, we obtain the overall object function to be minimized:

$$\mathcal{J} = \alpha \cdot \mathcal{L}_{cross} + \beta \cdot \mathcal{L}_{KL} + \gamma \cdot \mathcal{L}_{matching}, \quad (7)$$

where α, β and γ are introduced as hyper-parameters to control the importance of each task.

3.2 Model Implementation

Dual Encoders. We use Gated Recurrent Unit (GRU) as encoders to learn contextual words embeddings (Cho et al., 2014). Question and answer embeddings are reduced by weighted sum through multiple hops self-attention (Lin et al., 2017) of GRU units and then fed into two linear transition to obtain mean and standard deviation as $\mathcal{N}(z_q; \mu_q, diag(\sigma_q^2))$ and $\mathcal{N}(z_a; \mu_a, diag(\sigma_a^2))$.

Dual Decoders. We adopt another Gated Recurrent Unit (GRU) for generating token sequence conditioned on the latent variables z_q and z_a.

Question Answer Matching. We adopt cosine similarity with l_2 normalization to measure the matching probability of a question answer pair.

4 Experiment

4.1 Dataset

Our experiments were conducted on SQuAD 1.1 (Rajpurkar et al., 2016). It has over 100,000 questions composed to be answerable by text from Wikipedia documents. Each question has one corresponding answer sentence extracted from the Wikipedia document. Since the test set is not publicly available, we partition the dataset into 79,554 (training) / 7,801 (dev) / 10,539 (test) objects.

4.2 Baselines

InferSent (Conneau et al., 2017)**.** It is not explicitly designed for answer retrieval, but it produces results on semantic tasks without requiring additional fine tuning.

Table 2: Performance of answer retrieval on SQuAD.

Method	SQuAD		
	MRR	R@1	R@5
InferSent	36.90	27.91	46.92
SenBERT	38.01	27.34	49.59
BERT$_{\text{QA}}$	48.07	40.63	57.45
QA-Lite	50.29	40.69	61.38
USE-QA	61.23	53.16	69.93
Dual-GRUs	61.06	54.70	68.25
Dual-VAEs	61.48	55.01	68.49
Cross-VAEs	**62.29**	**55.60**	**70.05**

Table 3: Performance of answer retrieval on a subset of SQuAD in which any answer has more than 8 questions. Our method outperforms baselines much more. SSE indicates the sum of squared distances/errors between two different questions aligned to same answer.

Method	SQuAD Subset			
	MRR	R@1	R@5	SSE
BERT$_{\text{QA}}$	37.90	30.81	45.24	0.23
USE-QA	47.06	40.90	53.44	0.14
Cross-VAEs	**48.52**	**44.55**	**53.52**	**0.09**

USE-QA (Yang et al., 2019) **.** It is based on Universal Sentence Encoder (Cer et al., 2018), but trained with multilingual QA retrieval and two other tasks: translation ranking and natural language inference. The training corpus contains over a billion question answer pairs from popular online forums and QA websites (e.g, Reddit).

QA-Lite. Like USE-QA, this model is also trained over online forum data based on transformer. The main differences are reduction in width and depth of model layers, and sub-word vocabulary size.

BERT$_{\text{QA}}$ (Devlin et al., 2019) **.** BERT$_{QA}$ first concatenates the question and answer into a text sequence $[[CLS], Q, [SEP], A, [SEP]]$, then passes through a 12-layers BERT and takes the $[CLS]$ vector as input to a binary classifier.

SenBERT (Reimers and Gurevych, 2019) **.** It consists of twin structured BERT-like encoders to represent question and answer sentence, and then applies a similarity measure at the top layer.

4.3 Experimental Settings

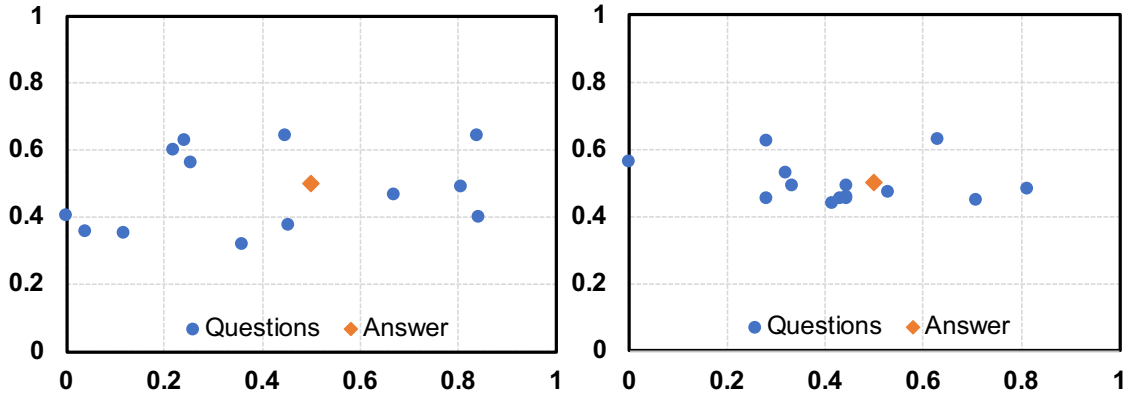
Implementation details. We initialize each word with a 768-dim BERT token embedding vector. If a word is not in the vocabulary, we use the average vector of its sub-word embedding vectors in the vocabulary. The number of hidden units in GRU encoder are all set as 768. All decoders are multi-layer perceptions (MLP) with one 768 units hidden layer. The latent embedding size is 512. The model is trained for 100 epochs by SGD using Adam optimizer (Kingma and Ba, 2014). For the KL-divergence, we use an KL cost annealing scheme (Bowman et al., 2016), which serves the purpose of letting the VAE learn useful representations before they are smoothed out. We increase the weight β of the KL-divergence by a rate of $2/epochs$ per epoch until it reaches 1. We set learning rate as 1e-5, and implemented on Pytorch.

Competitive Methods. We compare our proposed method cross variational autoencoder (Cross-VAEs) with dual-encoder model and dual variational autoencoder (Dual-VAEs). For fair comparisons, we all use GRU as encoder and decoder, and keep all other hyperparameters the same.

Evaluation Metrics. The models are evaluated on retrieving and ranking answers to questions using three metrics, mean reciprocal rank (MRR) and recall at K (R@K). R@K is the percentage of correct answers in topK out of all the relevant answers. MRR represents the average of the reciprocal ranks of results for a set of queries.

Comparing performance with baselines. As shown in Table 2, two BERT based models do not perform well, which indicates fune tuning BERT may not be a good choice for answer retrieval task due to unrelated pre-training tasks (e.g, masked language model). In contrast, using BERT token embedding can perform better in our retrieval task. Our proposed method outperforms all baseline methods. Comparing with USE-QA, our method improves MRR and R@1 by +1.06% and +2.44% on SQuAD, respectively. In addition, Dual variational autoencoder (Dual-VAEs) does not make much improvement on question answering retrieval task because it can only preserve isolated semantics of themselves. Our proposed crossing variational autoencoder (Cross-VAEs) could outperform dual-encoder model and dual variational autoencoder model, which improves MRR and R@1 by +1.23%/+0.81% and +0.90%/+0.59%, respectively.

Analyzing performance on sub-dataset. We extract a subset of SQuAD, in which any answer has at least eight different questions. As shown in Table 3, our proposed cross variational au-



(a) USE-QA (b) CrossVAEs

Question (1): What halftime performer previously headlined Super Bowl XLVIII? **Mismatched Answer:** Coincidentally, both teams were coached by John Fox in their last Super Bowl appearance prior to Super Bowl 50.	**Question (2):** Which Super Bowl halftime show did Beyonće headline? **Mismatched Answer:** On December 3, the league confirmed that the show would be headlined by the British rock group Coldplay.
Correct Answer of Question (1) and (2): The Super Bowl 50 halftime show was headlined by the British rock group Cold-play with special guest performers Beyonće and Bruno Mars, who headlined the Super Bowl XLVII and Super Bowl XLVIII halftime shows.	

(c) Two questions were incorrectly matched by USE-QA, but correctly matched by CrossVAEs.

Figure 2: A case of 14 different questions aligned to the same answer. We use SVD to reduce embedding dimensions to 2, and then project them on the X-Y coordinate axis. The scale of X-Y axis is relative with no practical significance. We observe that our method makes questions that share the same answer to be closer with each other.

toencoder (Cross-VAEs) could outperform baseline methods on the subset. Our method improves MRR and R@1 by +1.46% and +3.65% over USE-QA. Cross-VAEs significantly improve the performance when an answer has multiple aligned questions. Additionally, SSE of our method is smaller than that of USE-QA. Therefore, the questions of the same answer are closer in the latent space.

4.4 Case Study

Figures 2(a) and 2(b) visualize embeddings of 14 questions of the same answer. We observe that crossing variational autoencoders (CrossVAE) can better capture the aligned semantics between questions and answers, making latent representations of questions and answers more prominent. Figure 2(c) demonstrates two of example questions and corresponding answers produced by USE-QA and CrossVAEs. We observe that CrossVAEs can better distinguish similar answers even though they all share several same words with the question.

5 Conclusion

Given a candidate question, answer retrieval aims to find the most similar answer text between candidate answer texts. In this paper, We proposed to cross variational autoencoders by generating questions with aligned answers and generating answers with aligned questions. Experiments show that our method improves MRR and R@1 over the best baseline by 1.06% and 2.44% on SQuAD.

Acknowledgements

We thank Drs. Nicholas Fuller, Sinem Guven, and Ruchi Mahindru for their constructive comments and suggestions. This project was partially supported by National Science Foundation (NSF) IIS-1849816 and Notre Dame Global Gateway Faculty Research Award.

References

Zahra Abbasiyantaeb and Saeedeh Momtazi. 2020. Text-based question answering from information retrieval and deep neural network perspectives: A survey. *arXiv preprint arXiv:2002.06612*.

Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. 2019. Reqa: An evaluation for end-to-end answer retrieval models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Wei-Cheng Chang, Felix X Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. *In Proceedings of 8th International Conference for Learning Representation (ICLR)*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Arpita Das, Harish Yenala, Manoj Chinnakotla, and Manish Shrivastava. 2016. Together we stand: Siamese networks for similar question retrieval. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Michel Deudon. 2018. Learning semantic similarity in a continuous space. In *Advances in neural information processing systems (NeurIPS)*, pages 986–997.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017a. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P Xing. 2017b. On unifying deep generative models. *In Proceedings of 5th International Conference for Learning Representation (ICLR)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *In Proceedings of 2nd International Conference for Learning Representation (ICLR)*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *In Proceedings of 1st International Conference for Learning Representation (ICLR)*.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems (NeurIPS)*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *In Proceedings of 5th International Conference for Learning Representation (ICLR)*.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems (NeurIPS)*.

Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural machine reading comprehension: Methods and trends. *Applied Sciences*.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*.

Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Minjoon Seo, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2018. Phrase-indexed question answering: A new challenge for scalable document comprehension. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-time open-domain question answering with dense-sparse phrase index. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Dinghan Shen, Yizhe Zhang, Ricardo Henao, Qinliang Su, and Lawrence Carin. 2018. Deconvolutional latent-variable model for text sequence matching. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems (NeurIPS)*.

Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. 2017. Few-shot learning through an information retrieval lens. In *Advances in neural information processing systems (NeurIPS)*.

Lingfei Wu, Ian EH Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J Witbrock. 2018. Word mover's embedding: From word2vec to document embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhongbin Xie and Shuai Ma. 2019. Dual-view variational autoencoders for semi-supervised text matching. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press.

Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. 2017. Variational autoencoder for semi-supervised text classification. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2019. A compare-aggregate model with latent clustering for answer selection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.