Paraphrase Generation with High-quality Data Using T-CVAE

Zhiming Xu
Department of Computer Science
and Technology, Nanjing University,
Nanjing, China
zhimingxu@smail.nju.edu.cn

Shujian Huang
National Key Laboratory for
Novel Software Technology,
Nanjing University, Nanjing, China
huangsj@nju.edu.cn

Abstract

Paraphrase generation has been drawing increasing attention from the research community during the last few years. Capable paraphrase generation models can greatly benefit various downstream tasks, such as question answering, and information retrieval. Many neural networks based on seq2seq have achieved decent performances on several commonly used datasets. However, they ignored certain limitations appearing in those datasets, which can possibly degraded performances or even altered the objective of generating paraphrases. In this paper, we carefully study the drawback underneath one dataset, propose a simple and effective way to reorganize it, and show this can improve performance by a large margin with Transformer-based conditional variational antoencoder.

1 Introduction

Paraphrases refer to the restatements of *original* texts in a different form, often with modified words, phrases and orders. It can be very useful in many closely related tasks within natural language generation, such as abstract summarization (Chen and Bansal, 2018) and chat-bot (Yan et al., 2016), as well as in other non-generative tasks like question answering (Fader et al., 2014) and relation extraction. Those works show that paraphrase is not only important as a specific task, but also proves to help improve performances in other downstream tasks.

Existent approaches to generating paraphrases could be categorized as follows: rule-based ones (Zhao et al., 2009; Hassan et al., 2007), variational autoencoder ones (Gupta et al., 2017), and reinforcement learning ones (Yang et al., 2019; Qian et al., 2019). The latter two kinds of models often include encoder/decoder architectures, which is usually implemented with sequential models, such as LSTM (Hochreiter and Schmidhuber, 1997) in (Yang et al., 2019). Apart from those, other models also seek to use an architecture similar to that of machine translation with Transformer (Vaswani et al., 2017) like (Wang et al., 2019).

In this paper, we first analyze certain limitations in a prevailing dataset, namely MSCOCO (Lin et al., 2014) found via our observation that previous state-of-the-art model (Gupta et al., 2017) trained on these data actually learn more about language model than paraphrase in some cases (it makes up sentences based on training data instead of generating a paraphrase). Then we propose a practical metric to measure paraphrase and regroup original data accordingly, and propose a novel framework for paraphrase generation based on conditional variational autoencoder (CVAE) that solely exploits the Transformer model (Vaswani et al., 2017), namely T-CVAE (Wang and Wan, 2019). Since the individual attention heads in Transformer imitates behavior related to the syntactic and semantic structure of the sentence (Vaswani et al., 2017, 2018) which is critical to paraphrase generation.

Our main contributions include:

- We point out severe flaws in MSCOCO dataset, and overcome it with simple and practical regrouping.
- We propose a novel and concise framework for paraphrase generation that produces quality paraphrases of their source sentences compared to previous state-of-the-art ones.

2 Dataset Analysis

A commonly used dataset in training paraphrase generation, i.e., MSCOCO (Lin et al., 2014) was originally derived from the image caption task which aimed to provide a descriptive caption for a given image. In the original dataset, an image is often annotated with five captions (we will call them a *group* of captions thereon). Previous works usually assume that the semantic meanings of captions in a group are equivalent. Therefore, each sentence taken from it is thought to be a paraphrase of the others. However, we find that this assumption is not necessarily true. Since different captions in a group might describe the same image in different ways. Suppose we would like to caption an image of a desk, one might say "a laptop sits on a brown desk", another might say "a pile of books lies beneath a laptop". They are both genuine captions, but they fails to conforms to the assumption that they are semantically equivalent since either contains information the other ignores. More such examples are shown in Table 1.

To clearly demonstrate that the captions within a group can be significantly discrepant in their meanings, we apply two different methods of mainstream sentence embedding to them, specifically, BERT (Devlin et al., 2018) and InferSent (Conneau et al., 2017) and project the resulting vectors to 2D space with Principle Components Analysis (PCA). Without loss of generality, we use these models without fine-tuning to avoid introducing biases underneath this dataset. We randomly sample 6 groups of captions from MSCOCO and plot their embeddings after compressed to 2D vectors by BERT[1] in Figure 1, and by InferSent[2] in Figure 2, respectively.

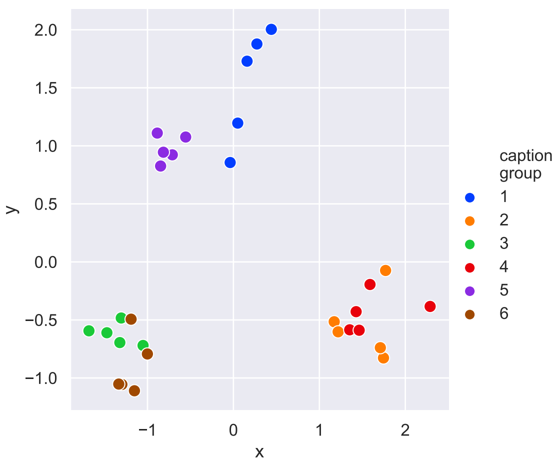

Figure 1: Embedding vectors obtained from InferSent. Each color represents a group of captions.

Although compressing high dimensional vectors to 2D might yield great losses in information, we can clearly observe that even captions in the same group (shown in the same color) lie together, they also tend to mix with other groups. This means that some captions are more mutually semantically *equivalent* than the others despite in the same group. Therefore, assuming them to be paraphrases is not necessarily true. A dataset which merely takes captions uniformly from the same image actually consists of considerable non-paraphrase components that are irrelevant, or even harmful to this task. To better capture how similar/disrep a group of captions can be, we study a simple yet expressive measure, cosine similarities between caption pairs' embedding vectors and plot their distribution. We use InferSent to plot three different kinds of sentence pairs' cosine similarity scores in Figure 3: random sentence pairs, caption pairs from MSCOCO, and human-annotated paraphrase pairs from Quora Question Pairs[3].

We can see that while an caption pairs in a group are more similar than random ones, they are still significantly different compared to true paraphrase pairs. Consequently, we will regroup MSCOCO dataset below with a threshold of .8.

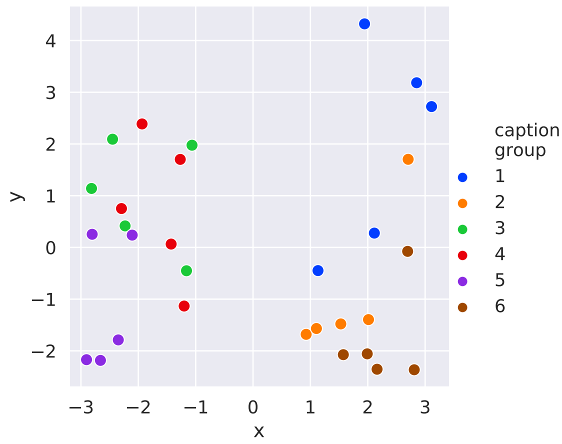

Figure 2: Embedding vectors obtained from BERT. Each color represents a group of captions.

3 Regroup

Based on the discussion above, we will regroup original groups of captions and compose a "more" paraphrase dataset. There are 168930 and 5085 images in MSCOCO's train and validation set[4], respectively. Since each image is usually accompa-

[1] https://github.com/dmlc/gluon-nlp/tree/master/scripts/bert
[2] https://github.com/facebookresearch/InferSent
[3] https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs
[4] COCO 2015 Image Captioning Task

Caption 1	Caption 2
a street sign modified to read stop bush	a vandalized stop sign and a red beetle on the road
the two people are walking down the beach	two teenagers at a white sanded beach with surfboards
a woman walks by a couple of shop windows	two bicycles and a woman walking in front of a shop

Table 1: Some semantically different captions.

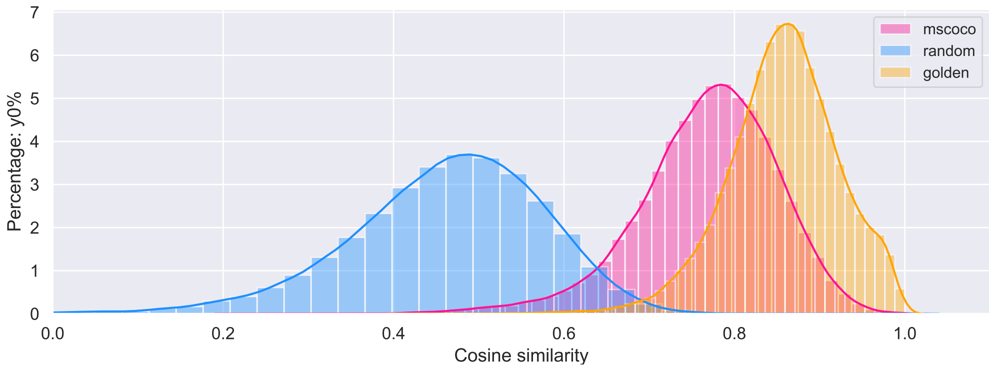

Figure 3: Cosine similarity distribution from InferSent.

nied with 5 captions, denote the number of images as N, we can construct around $N \cdot \binom{5}{2}$ paraphrase pairs. However, a pair might not be semantically equivalent as discussed above. Therefore, We traverse all those possible pairs, while only keeping those that have a cosine similarity exceeding .8, which result in 648334 and 19251 pairs derived from original train and validation set, respectively. It could be the case that some sentences frequently appear as the first while others as the second, so that the model would only be able to learn paraphrase in a fixed direction. However, during training time we will address this by randomly peek either sentence as the *source* to generate the other, i.e., *target*. We will call the newly partitioned more paraphrase version of MSCOCO as COCO-P.

4 Model

In this section, we present our model for generating paraphrase, which is very similar to the T-CVAE (Wang and Wan, 2019) model for story completion. The main difference is that in story completion task, they encode five sentences with one of them masked and aim to predict the masked sentence. While adopted to paraphrase generation task, we encode a pair of sentences, the *source* and its masked *target*, and try to generate the latter with encoded information of *source* and a latent variable z. The overall architecture is shown in Figure 4

5 Experiments

We will describe our implementation and compare with recent state-of-the-art models in this section.

5.1 Baselines

We first compare our model's performances on MSCOCO dataset with recent models (Gupta et al., 2017; Li et al., 2018; Huang et al., 2019; Yang et al., 2019) and show the improvements brought by Transformer as both encoder and decoder. Besides, we will show on the more paraphrasing version of MSCOCO, COCO-P, the performance can be even better.

- VAE-SVG-EQ (Gupta et al., 2017): This model is the former state-of-the-art in paraphrase generation. It main components are variational autoencoder which uses LSTM (Hochreiter and Schmidhuber, 1997) for both encoding and generating.

- GAP (Yang et al., 2019): This model uses a generator-discriminator paradigm resembling GAN (Goodfellow et al., 2014). It also adds one more hidden representation for constructing latent variable z.

5.2 Implementation

Since our model resembles the aforementioned story completion one, our code is based on their official implementation[5]. We use GloVe word

[5]https://github.com/sodawater/T-CVAE/

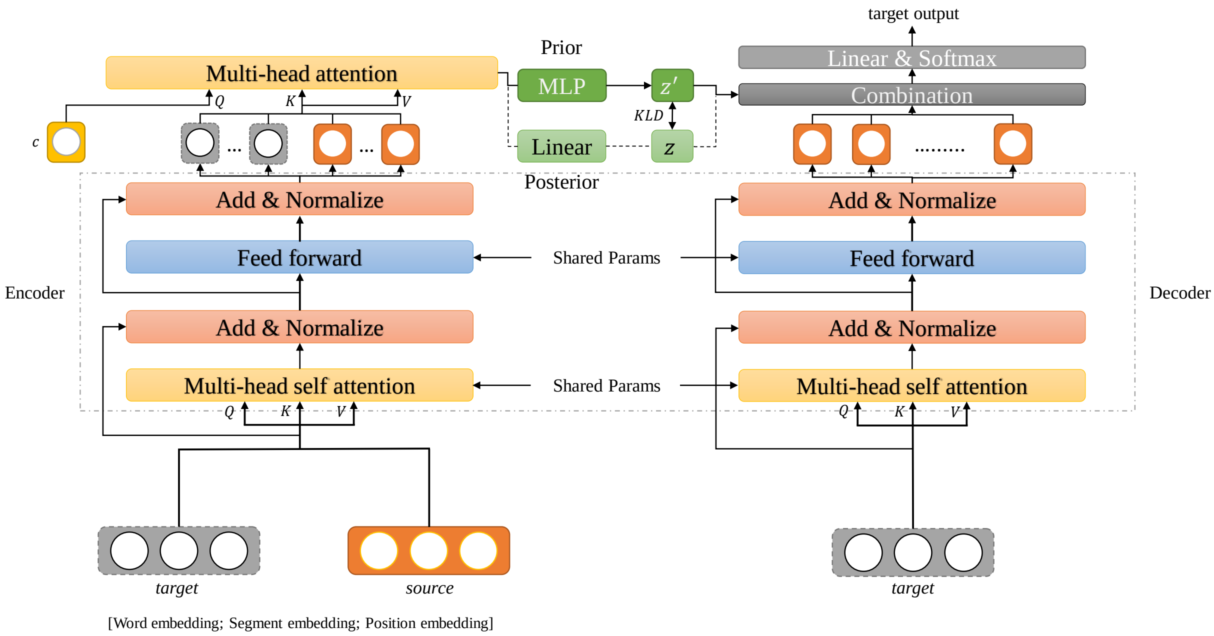

Figure 4: Architecture of our T-CVAE model. Both prior net and the posterior net are built upon the encoder, and the posterior net takes an extra input target represented by dashed box. During training, latent variable z fed to the combination layer is calculated by the posterior (connected with dashed lines); during inference, the prior net replaces the posterior net and generates the latent variable z' (connected by solid lines). The reparametrization trick is used to obtain samples of latent variable either from z in training or z' in inferring.

MODEL	BLEU↑	METEOR↑	ROUGE-L↑
VAE-SVG-EQ (Gupta et al., 2017)	41.7	31.0	-
GAP (Yang et al., 2019)	45.6	36.17	-
T-CVAE (ours) trained and evaluated on MSCOCO	42.7	-	37.2
T-CVAE (ours) trained on COCO-P, evaluated on MSCOCO	43.2	-	38.7
T-CVAE (ours) trained on MSCOCO, evaluated on COCO-P	45.6	-	43.1
T-CVAE (ours) trained and evaluated on COCO-P	**48.3**	-	**45.8**

Table 2: Performances of our model on MSCOCO/COCO-P against other models.

embedding (Pennington et al., 2014), specifically, the Common Crawl 300d vectors with 840B tokens[6]. The vocabulary is built on the most frequent $20,000$ words from training data. We set the number of self-attention layers in Transformer to 2 with a hidden size of 256. For the latent random variable z, we set its dimension to 64. Besides, we use a batch size of 128, a fixed learning rate of 1.0×10^{-4}, and clip gradient to $[-3, 3]$. A dropout of .15 is also applied to each Transform layer for regularization.

5.3 Evaluation

We present the results on original MSCOCO and reconstructed COCO-P dataset in Table 2. It can be seen that our proposed model along can improve performances compared to previous state-of-the-art (Gupta et al., 2017) with a large margin. Besides, training on the COCO-P dataset we construct above will further improve performances even tested on (not that paraphrased) MSCOCO. Evaluation on COCO-P shows even better performances, which proves the benefit brought by our regrouped dataset.

[6]http://nlp.stanford.edu/data/glove.840B.300d.zip

6 Conclusions

We investigated critical shortcomings in a widely used paraphrase dataset, MSCOCO and overcome it with simple and practical regrouping. Besides, We proposed a novel and concise framework that improves on the current state-of-the-art with our regrouped dataset.

In the future, we will use T-CVAE with variational attention (Bahuleyan et al., 2018) to investigate its potential benefit for increasing diversity.

References

Hareesh Bahuleyan, Lili Mou, Olga Vechtomova, and Pascal Poupart. 2018. Variational attention for sequence-to-sequence models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1672–1682, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1156–1165, New York, NY, USA. ACM.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2017. A deep generative framework for paraphrase generation.

Samer Hassan, Andras Csomai, Carmen Banea, Ravi Sinha, and Rada Mihalcea. 2007. UNT: SubFinder: Combining knowledge sources for automatic lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 410–413, Prague, Czech Republic. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Shaohan Huang, Yu Wu, Furu Wei, and Zhongzhi Luan. 2019. Dictionary-guided editing networks for paraphrase generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6546–6553.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *Lecture Notes in Computer Science*, page 740–755.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Lihua Qian, Lin Qiu, Weinan Zhang, Xin Jiang, and Yong Yu. 2019. Exploring diverse expressions for paraphrase generation. 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 3171–3180, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Su Wang, Rahul Gupta, Nancy Chang, and Jason Baldridge. 2019. A task in a suit and a tie: Paraphrase generation with semantic augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:7176–7183.

Tianming Wang and Xiaojun Wan. 2019. T-cvae: Transformer-based conditioned variational autoencoder for story completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5233–5239. International Joint Conferences on Artificial Intelligence Organization.

Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhou, Zhoujun Li, and Jianshe Zhou. 2016. DocChat: An information retrieval approach for chatbot engines using unstructured documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 516–525, Berlin, Germany. Association for Computational Linguistics.

Qian Yang, Dinghan Shen, Yong Cheng, Wenlin Wang, Guoyin Wang, Lawrence Carin, et al. 2019. An end-to-end generative architecture for paraphrase generation. 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 3123–3133.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 834–842, Suntec, Singapore. Association for Computational Linguistics.