Unsupervised Paraphrasing without Translation

Aurko Roy
Google Research
aurkor@google.com

David Grangier
Google Research
grangier@google.com

Abstract

Paraphrasing exemplifies the ability to abstract semantic content from surface forms. Recent work on automatic paraphrasing is dominated by methods leveraging Machine Translation (MT) as an intermediate step. This contrasts with humans, who can paraphrase without being bilingual. This work proposes to learn paraphrasing models from an unlabeled monolingual corpus only. To that end, we propose a residual variant of vector-quantized variational auto-encoder.

We compare with MT-based approaches on paraphrase identification, generation, and training augmentation. Monolingual paraphrasing outperforms unsupervised translation in all settings. Comparisons with supervised translation are more mixed: monolingual paraphrasing is interesting for identification and augmentation; supervised translation is superior for generation.

1 Introduction

Many methods have been developed to generate paraphrases automatically (Madnani and J. Dorr, 2010). Approaches relying on Machine Translation (MT) have proven popular due to the scarcity of labeled paraphrase pairs (Callison-Burch, 2007; Mallinson et al., 2017; Iyyer et al., 2018). Recent progress in MT with neural methods (Bahdanau et al., 2014; Vaswani et al., 2017) has popularized this latter strategy. Conceptually, translation is appealing since it abstracts semantic content from its linguistic realization. For instance, assigning the same source sentence to multiple translators will result in a rich set of semantically close sentences (Callison-Burch, 2007). At the same time, bilingualism does not seem necessary to humans to generate paraphrases.

This work evaluates if data in two languages is necessary for paraphrasing. We consider three settings: supervised translation (parallel bilingual data is used), unsupervised translation (non-parallel corpora in two languages are used) and monolingual (only unlabeled data in the paraphrasing language is used). Our comparison devises comparable encoder-decoder neural networks for all three settings. While the literature on supervised (Bahdanau et al., 2014; Cho et al., 2014; Vaswani et al., 2017) and unsupervised translation (Lample et al., 2018a; Artetxe et al., 2018; Lample et al., 2018b) offer solutions for the bilingual settings, monolingual neural paraphrase generation has not received the same attention.

We consider discrete and continuous auto-encoders in an unlabeled monolingual setting, and contribute improvements in that context. We introduce a model based on Vector-Quantized Auto-Encoders, VQ-VAE (van den Oord et al., 2017), for generating paraphrases in a purely monolingual setting. Our model introduces residual connections parallel to the quantized bottleneck. This lets us interpolate from classical continuous auto-encoder (Vincent et al., 2010) to VQ-VAE. Compared to VQ-VAE, our architecture offers a better control over the decoder entropy and eases optimization. Compared to continuous auto-encoder, our method permits the generation of diverse, but semantically close sentences from an input sentence.

We compare paraphrasing models over intrinsic and extrinsic metrics. Our intrinsic evaluation evaluates paraphrase identification, and generations. Our extrinsic evaluation reports the impact of training augmentation with paraphrases on text classification. Overall, monolingual approaches can outperform unsupervised translation in all settings. Comparison with supervised translation shows that parallel data provides valuable information for paraphrase generation compared

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6033–6039
Florence, Italy, July 28 - August 2, 2019.

to purely monolingual training.

2 Related Work

Paraphrase Generation Paraphrases express the same content with alternative surface forms. Their automatic generation has been studied for decades: rule-based (McKeown, 1980; Meteer and Shaked, 1988) and data-driven methods (Madnani and J. Dorr, 2010) have been explored. Data-driven approaches have considered different source of training data, including multiple translations of the same text (Barzilay and McKeown, 2001; Pang et al., 2003) or alignments of comparable corpora, such as news from the same period (Dolan et al., 2004; Barzilay and Lee, 2003).

Machine translation later emerged as a dominant method for paraphrase generation. Bannard and Callison-Burch (2005) identify equivalent English phrases mapping to the same non-English phrases from an MT phrase table. Kok and Brockett (2010) performs random walks across multiple phrase tables. Translation-based paraphrasing has recently benefited from neural networks for MT (Bahdanau et al., 2014; Vaswani et al., 2017). Neural MT can generate paraphrase pairs by translating one side of a parallel corpus (Wieting and Gimpel, 2018; Iyyer et al., 2018). Paraphrase generation with pivot/round-trip neural translation has also been used (Mallinson et al., 2017; Yu et al., 2018).

Although less common, monolingual neural sequence models have also been proposed. In supervised settings, Prakash et al. (2016); Gupta et al. (2018) learn sequence-to-sequence models on paraphrase data. In unsupervised settings, Bowman et al. (2016) apply a VAE to paraphrase detection while Li et al. (2017) train a paraphrase generator with adversarial training.

Paraphrase Evaluation Evaluation can be performed by human raters, evaluating both text fluency and semantic similarity. Automatic evaluation is more challenging but necessary for system development and larger scale statistical analysis (Callison-Burch, 2007; Madnani and J. Dorr, 2010). Automatic evaluation and generation are actually linked: if an automated metric would reliably assess the semantic similarity and fluency of a pair of sentences, one would generate by searching the space of sentences to maximize that metric. Automated evaluation can report the overlap with a reference paraphrase, like for translation (Papineni et al., 2002) or summarization (Lin, 2004). BLEU, METEOR and TER metrics have been used (Prakash et al., 2016; Gupta et al., 2018). These metrics do not evaluate whether the generated paraphrase differs from the input sentence and large amount of input copying is not penalized. Galley et al. (2015) compare overlap with multiple references, weighted by quality; while Sun and Zhou (2012) explicitly penalize overlap with the input sentence. Grangier and Auli (2018) alternatively compare systems which have first been calibrated to a reference level of overlap with the input. We follow this strategy and calibrate the generation overlap to match the average overlap observed in paraphrases from humans.

In addition to generation, probabilistic models can be assessed through scoring. For a sentence pair (x, y), the model estimate of $P(y|x)$ can be used to discriminate between paraphrase and non-paraphrase pairs (Dolan and Brockett, 2005). The correlation of model scores with human judgments (Cer et al., 2017) can also be assessed. We report both types of evaluation.

Finally, paraphrasing can also impact downstream tasks, e.g. to generate additional training data by paraphrasing training sentences (Marton et al., 2009; Zhang et al., 2015; Yu et al., 2018). We evaluate this impact for classification tasks.

3 Residual VQ-VAE for Unsupervised Monolingual Paraphrasing

Auto-encoders can be applied to monolingual paraphrasing. Our work combines Transformer networks (Vaswani et al., 2017) and VQ-VAE (van den Oord et al., 2017), building upon recent work in discrete latent models for translation (Kaiser et al., 2018; Roy et al., 2018). VQ-VAEs, as opposed to continuous VAEs, rely on discrete latent variables. This is interesting for paraphrasing as it equips the model with an explicit control over the latent code capacity, allowing the model to group multiple related examples under the same latent assignment, similarly to classical clustering algorithms (Macqueen, 1967). This is conceptually simpler and more effective than rate regularization (Higgins et al., 2016) or denoising objectives (Vincent et al., 2010) for continuous auto-encoders. At the same time, training auto-encoder with discrete bottleneck is difficult (Roy et al., 2018). We address this difficulty with an hybrid model using a continuous residual

connection around the quantization module.

We modify the Transformer encoder (Vaswani et al., 2017) as depicted in Figure 1. Our encoder maps a sentence into a fixed size vector. This is simple and avoids choosing a fixed length compression rate between the input and the latent representation (Kaiser et al., 2018). Our strategy to produce a fixed sized representation from transformer is analogous to the special token employed for sentence classification in (Devlin et al., 2018).

At the first layer, we extend the input sequences with one or more fixed positions which are part of the self-attention stack. At the output layer, the encoder output is restricted to these special positions which constitute the encoder fixed sized-output. As in (Kaiser et al., 2018), this vector is split into multiple heads (sub-vectors of equal dimensions) which each goes through a quantization module. For each head h, the encoder output e_h is quantized as,

$$q_h(e_h) = c_k, \text{ where } k = \underset{i}{\operatorname{argmin}} \|e_h - c_i\|^2$$

where $\{c_i\}_{i=0}^{K}$ denotes the codebook vectors. The codebook is shared across heads and training combines straight-through gradient estimation and exponentiated moving averages (van den Oord et al., 2017). The quantization module is completed with a residual connection, with a learnable weight α, $z_h(e_h) = \alpha e_h + (1 - \alpha) q_h(e_h)$. One can observe that residual vectors and quantized vectors always have similar norms by definition of the VQ module. This is a fundamental difference with classical continuous residual networks, where the network can reduce activation norms of some modules to effectively rely mostly on the residual path. This makes α an important parameter to trade-off continuous and discrete auto-encoding. Our learning encourages the quantized path with a squared penalty α^2.

After residual addition, the multiple heads of the resulting vector are presented as a matrix to which a regular transformer decoder can attend. Models are trained to maximize the likelihood of the training set with Adam optimizer using the learning schedule from (Vaswani et al., 2017).

4 Experiments & Results

We compare neural paraphrasing with and without access to bilingual data. For bilingual settings, we consider supervised and unsupervised translation using round-trip translation (Mallinson et al., 2017; Yu et al., 2018) with German as the pivot language. Supervised translation trains the transformer base model (Vaswani et al., 2017) on the WMT'17 English-German parallel data (Bojar et al., 2017). Unsupervised translation considers a pair of comparable corpora for training, German and English WMT-Newscrawl corpora, and relies on the transformer models from Lample et al. (2018b). Both MT cases train a model from English to German and from German to English to perform round-trip MT. For each model, we also distill the round-trip model into a single artificial English to English model by generating a training set from pivoted data. Distillation relies on the billion word corpus, LM1B (Chelba et al., 2013).

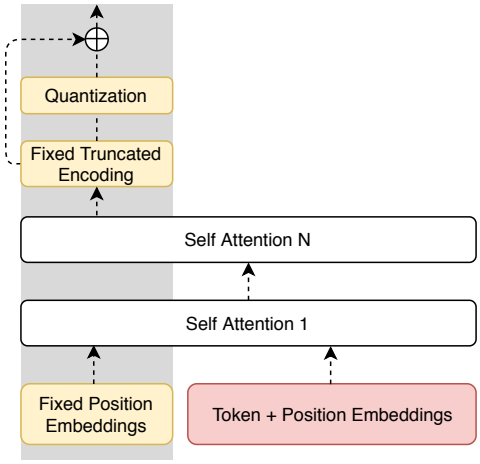

Figure 1: Encoder Architecture

Monolingual Residual VQ-VAE is trained only on LM1B with $K = 2^{16}$, with 2 heads and fixed window of size 16. We also evaluate plain VQ-VAE $\alpha = 0$ to highlight the value of our residual modification. We further compare with a monolingual continuous denoising auto-encoder (DN-AE), with noising from Lample et al. (2018b).

Paraphrase Identification For classification of sentence pairs (x, y) over Microsoft Research Paraphrase Corpus (MRPC) from Dolan and Brockett (2005), we train logistic regression on $P(y|x)$ and $P(x|y)$ from the model, complemented with encoder outputs in fixed context settings. We also perform paraphrase quality regression on Semantic Textual Similarity (STS) from Cer et al. (2017) by training ridge regression on the same features.

Finally, we perform paraphrase ranking on Multiple Translation Chinese (MTC) from Huang et al.

	Parapharase Identification			Generation	
	MRPC	STS	MTC	BLEU	Pref.
Supervised Translation	70.6	46.0	78.6	**8.73**	36.8
+ Distillation	66.5	**60.0**	55.6	7.08	–
Unsupervised Translation	66.0	13.2	65.8	6.59	28.1
+ Distillation	66.9	45.0	52.0	6.45	–
Mono. DN-AE	66.8	46.2	91.6	5.13	–
Mono. VQVAE	66.3	10.6	69.0	3.85	–
+ Residual	**73.3**	59.8	**94.0**	7.26	31.9
+ Distillation	71.3	54.3	88.4	6.88	–

Table 1: Paraphrase Identification & Generation. Identification is evaluated with accuracy on MRPC, Pearson Correlation on STS and ranking on MTC. Generation is evaluated with BLEU and human preferences on MTC.

	SST-2		TREC	
	Acc.	F1	Acc	F1
NB-SVM (trigram)	81.93	83.15	89.77	84.81
Supervised Translation	81.55	82.75	**90.78**	85.44
+ Distillation	81.16	66.59	90.38	**86.05**
Unsupervised Translation	81.87	83.18	88.17	83.42
+ Distillation	81.49	82.78	89.18	84.41
Mono. DN-AE	81.11	82.48	89.37	84.08
Mono. VQ-VAE	81.98	82.95	89.17	83.64
+ Residual	**82.12**	**83.23**	89.98	84.31
+ Distillation	81.60	82.81	89.78	84.31

Table 2: Paraphrasing for Data Augmentation: Accuracy and F1-scores of a Naive Bayes-SVM classifier on sentiment (SST-2) and question (TREC) classification.

(2002). MTC contains English paraphrases collected as translations of the same Chinese sentences from multiple translators (Mallinson et al., 2017). We pair each MTC sentence x with a paraphrase y and 100 randomly chosen non-paraphrases y'. We compare the paraphrase score $P(y|x)$ to the 100 non-paraphrase scores $P(y'|x)$ and report the fraction of comparisons where the paraphrase score is higher.

Table 1 (left) reports that our residual model outperforms alternatives in all identification setting, except for STS, where our Pearson correlation is slightly under supervised translation.

Paraphrases for Data Augmentation We augment the training set of text classification tasks for sentiment analysis on Stanford Sentiment Treebank (SST-2) (Socher et al., 2013) and question classification on Text REtrieval Conference (TREC) (Voorhees and Tice, 2000). In both cases, we double training set size by paraphrasing each sentence and train Support Vector Machines with Naive Bayes features (Wang and Manning, 2012).

In Table 2, augmentation with monolingual models yield the best performance for SST-2 sentiment classification. TREC question classification is better with supervised translation augmentation. Unfortunately, our monolingual training set LM1B does not contain many question sentences. Future work will revisit monolingual training on larger, more diverse resources.

Paraphrase Generation Paraphrase generation are evaluated on MTC. We select the 4 best translators according to MTC documentation and paraphrase pairs with a length ratio under 1.2. Our evaluation prevents trivial copying solutions. We select sampling temperature for all models such that their generation overlap with the input is 20.9 BLEU, the average overlap between humans on MTC. We report BLEU overlap with the target and run a blind human evaluation where raters pick the best generation among supervised translation, unsupervised translation and monolingual.

Table 3 shows examples. Table 1 (right) reports that monolingual paraphrasing compares favorably with unsupervised translation while supervised translation is the best technique. This high-

In:	a worthy substitute
Out:	A worthy replacement.
In:	Local governments will manage the smaller enterprises.
Out:	Local governments will manage smaller companies.
In:	Inchon is 40 kilometers away from the border of North Korea.
Out:	Inchon is 40 km away from the North Korean border.
In:	Executive Chairman of Palestinian Liberation Organization, Yasar Arafat, and other leaders are often critical of aiding countries not fulfilling their promise to provide funds in a timely fashion.
Out:	Yasar Arafat , executive chairman of the Palestinian Liberation Organization and other leaders are often critical of helping countries meet their pledge not to provide funds in a timely fashion.

Table 3: Examples of generated paraphrases from the monolingual residual model (Greedy search).

lights the value of parallel data for paraphrase generation.

5 Discussions

Our experiments highlight the importance of the residual connection for paraphrase identification. From Table 1, we see that a model without the residual connection obtains $66.3\%, 10.6\%$ and 69.0% accuracy on MRPC, STS and MTC. Adding the residual connection improves this to $73.3\%, 59.8\%$ and 94.0% respectively.

The examples in Table 3 show paraphrases generated by the model. The overlap with the input from these examples is high. It is possible to generate sentences with less overlap at higher sampling temperatures, we however observe that this strategy impairs fluency and adequacy. We plan to explore strategies which allow to condition the decoding process on an overlap requirement instead of varying sampling temperatures (Grangier and Auli, 2018).

6 Conclusion

We compared neural paraphrasing with and without access to bilingual data. Bilingual settings considered supervised and unsupervised translation. Monolingual settings considered auto-encoders trained on unlabeled text and introduced continuous residual connections for discrete auto-encoders. This method is advantageous over both discrete and continuous auto-encoders. Overall, we showed that monolingual models can outperform bilingual ones for paraphrase identification and data-augmentation through paraphrasing. We also reported that generation quality from monolingual models can be higher than model based on unsupervised translation but not supervised translation. Access to parallel data is therefore still advantageous for paraphrase generation and our monolingual method can be a helpful resource for languages where such data is not available.

Acknowledgments

We thanks the anonymous reviewers for their suggestions. We thank the authors of the Tensor2tensor library used in our experiments (Vaswani et al., 2018).

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604. Association for Computational Linguistics.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 16–23. Association for Computational Linguistics.

Regina Barzilay and Kathleen R McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th annual meeting on Association for Computational Linguistics*, pages 50–57. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the SIGNLL'16*, pages 10–21.

Christopher Callison-Burch. 2007. *Paraphrasing and translation*. Ph.D. thesis, University of Edinburgh Edinburgh.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *CoRR*, abs/1312.3005.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

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

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350. Association for Computational Linguistics.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. *arXiv preprint arXiv:1506.06863*.

David Grangier and Michael Auli. 2018. Quickedit: Editing text & translations by crossing words out. In *Proc. of NAACL*.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework.

Shudong Huang, David Graff, and George Doddington. 2002. *Multiple-translation Chinese corpus*. Linguistic Data Consortium, University of Pennsylvania.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of NAACL-HLT*, pages 1875–1885.

Łukasz Kaiser, Aurko Roy, Ashish Vaswani, Niki Pamar, Samy Bengio, Jakob Uszkoreit, and Noam Shazeer. 2018. Fast decoding in sequence models using discrete latent variables. *arXiv preprint arXiv:1803.03382*.

Stanley Kok and Chris Brockett. 2010. Hitting the right paraphrases in good time. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 145–153. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2017. Paraphrase generation with deep reinforcement learning. *arXiv preprint arXiv:1711.00279*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

J. B. Macqueen. 1967. Unified techniques for vector quantization and hidden markov modeling using semi-continuous models.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36:341–387.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association*

for Computational Linguistics: Volume 1, Long Papers, volume 1, pages 881–893.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 381–390. Association for Computational Linguistics.

Kathleen R McKeown. 1980. Paraphrasing using given and new information in a question-answer system. *Technical Reports (CIS)*, page 723.

Marie Meteer and Varda Shaked. 1988. Strategies for effective paraphrasing. In *Proceedings of the 12th conference on Computational linguistics-Volume 2*, pages 431–436. Association for Computational Linguistics.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. *CoRR*, abs/1711.00937.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Aaditya Prakash, Sadid A Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual lstm networks. *arXiv preprint arXiv:1610.03098*.

Aurko Roy, Ashish Vaswani, Arvind Neelakantan, and Niki Parmar. 2018. Theory and experiments on vector quantized autoencoders. *arXiv preprint arXiv:1805.11063*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual smt system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 38–42. 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, et al. 2018. Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*.

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

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207. ACM.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.

John Wieting and Kevin Gimpel. 2018. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 451–462.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.