Multi-paraphrase Augmentation to Leverage Neural Caption Translation

Johanes Effendi[1], *Sakriani Sakti*[1,2], *Katsuhito Sudoh*[1,2], *Satoshi Nakamura*[1,2]

[1]Nara Institute of Science and Technology, Japan
[2]RIKEN, Center for Advanced Intelligence Project AIP, Japan
{johanes.effendi.ix4,ssakti,sudoh,s-nakamura}@is.naist.jp

Abstract

Paraphrasing has been proven to improve translation quality in machine translation (MT) and has been widely studied alongside with the development of statistical MT (SMT). In this paper, we investigate and utilize neural paraphrasing to improve translation quality in neural MT (NMT), which has not yet been much explored. Our first contribution is to propose a new way of creating a multi-paraphrase corpus through visual description. After that, we also proposed to construct neural paraphrase models which initiate expert models and utilize them to leverage NMT. Here, we diffuse the image information by using image-based paraphrasing without using the image itself. Our proposed image-based multi-paraphrase augmentation strategies showed improvement against a vanilla NMT baseline.

1. Introduction

In general, sentence paraphrasing is a way to restate a concept with different vocabulary, style, and level of detail. As defined by De Beaugrande and Dressler, a paraphrase is an approximate conceptual equivalence among outwardly different material [1]. In many language generation tasks, paraphrasing plays a critical role for enrichment and adding flexibility. In the MT system, paraphrases are often used for multi-reference evaluation [1], pre-editing of source sentences [2, 3, 4] and automatic post-editing [5, 6, 7].

Moreover, since the development of SMT, there have been a lot of approaches for using paraphrasing to elaborate the source language data. Such method have been concluded as a convenient way to handle out-of-vocabulary (OOV) and rare words problem [8]. A study by Madnani and Dorr also showed that by using targeted paraphrases, unfair penalization of translation hypotheses could be avoided [9]. Paraphrasing could also be used to augment the dataset size, which correlates positively with translation result in SMT [4, 10].

However, despite a wide range of existing works of paraphrasing, MT studies usually use a strict definition of paraphrase which accepts only word substitution and reordering. The reason is that we cannot grasp a tangible concept about the idea of the sentence being translated. On the other hand, Hirst argues that paraphrases don't necessarily need to be fully synonymous. It is sufficient for them to be quasi-synonymous, as a mutually replaceable form of truth applicable in some contexts [11]. By taking further this idea, as long as the semantics of the mutual paraphrase sentence can be determined, we actually can widen the paraphrase definition to some extent.

In this research, we treat an image as a symbolic form of sentence idea, regarded as the basis of paraphrasing. We consider two sentences as paraphrase as long as both of them are talking about the same image. This means that the word or phrase insertion and deletion based on the same picture as a concept may now be accepted as one of the paraphrase variations. Slightly different from the usual use case, this definition can be called image-based paraphrasing. Furthermore, as paraphrasing to enable multi-source information in NMT is not much investigated yet, in this study we explore the use of image-based paraphrasing to leverage NMT quality.

Recently, the Second Conference on Machine Translation (WMT17) accelerated a "Multimodal Machine Translation" shared task that aimed to translate the image descriptions into the target language. Most approaches focus on utilizing image features in addition to the information from a single caption of the source language. However, the results from most submitted systems reveal that the additional image features could only slightly contribute to system performance. As pointed out by Calixto et al. [12] the image-text latent representation combination approach has not yielded significant improvement on WMT 2017 Multimodal shared task dataset testing. Here, we attempt to go in another direction in which we diffuse the image information by using image-based paraphrasing without using the image itself. The resulting paraphrase captions are then utilized within a multi-source and multi-expert NMT model.

In summary, the contributions of this work include:

1. Introduce a new way of creating a multi-paraphrase corpus through image captions so-called image-based paraphrasing.

2. Generate multi-paraphrase sentences of the WMT17 Multimodal Translation Task dataset through crowdsourcing, which can be used by the community[1]

3. Develop automatic paraphrase generation in a semi-supervised manner;

[1] The data will be soon available publicly.

4. Utilize multi-expert translation in neural machine translation using our proposed paraphrase; and
5. Improve the baseline used at WMT17 with a 13.2 BLEU score margin, which is close to the top score that used a multimodal model.

2. Multi-paraphrase Generation

2.1. Defining Paraphrase Elementary Operation

To train an NMT model with our image-based multi-paraphrases, firstly we need to build a set of paraphrased source sentences with images as the basis of paraphrasing. However, the process of manually collecting paraphrases is expensive and time-consuming. On the other hand, Resnik et al. (2013) proposed that corpus creation with a crowdsourcing platform provides such advantages as low cost, effectiveness, and reasonable quality [13].

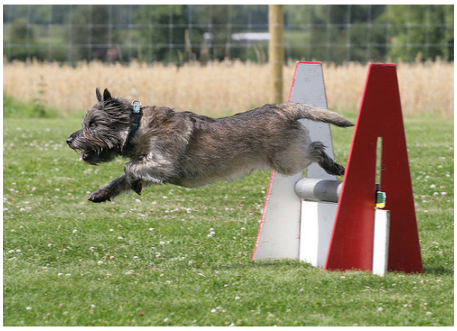

Figure 1: Reference image for captioning and paraphrasing shown in Table 1.

Furthermore, the requirement to have an image and several captions, are similar with an image captioning dataset such as Microsoft Common Object in Context (MSCOCO) dataset [14]. The caption of this dataset can be regarded as paraphrase, such as done by Prakash et al. for their neural paraphrase generation study [15]. They stated that the annotators described the most obvious things in an image and concluded that several captions of an image can be counted as paraphrases. While this may be true, we cannot define what kind of operation has been done from the original sentence to the paraphrase. Consequently, the arbitrary nature of the corpus distribution might cause the paraphrases to become noise to each other.

To prevent this, a set of paraphrase operation which covers all possible paraphrase variations needs to be defined. Bhagat and Hovy categorized the variations of how humans paraphrase [16] and argued that "although the logical definition of paraphrases requires strict semantic equivalence, linguistics accept a broader, approximate, equivalence." Based on this idea, they analyzed paraphrase characteristics in various studies and in corpora and established 25 quasi-paraphrase classes, such as change in tenses, metaphor substitution, and, function-word variations.

Given some quasi-paraphrases have very small frequency in the MTC and MSRP corpora as reported by them, we grouped these into 4 elementary paraphrase operations: deletion, insertion, reordering, and substitution. Then, we constructed a paraphrase corpus based on these four operations. The paraphrase collection was done through a crowdsourcing platform on the partial WMT17 Multimodal Translation Task dataset [17]. After that, we constructed our automatic neural paraphrase model based on partial data to generate the paraphrase sentences of the full WMT17 dataset. The details are described below.

2.2. Crowdsourcing Paraphrases on Partial WMT17 Dataset

The WMT17 Multimodal Translation Task dataset [17] contains a set of images with triplets of captions in English, German, and French. The dataset was created from the Flickr30K Entities dataset of image captions in English [18] that was extended to also contain manually translated German and French captions. The data consists of 29000, 1014, and 1000 triplets respectively for the training, development and testing. An out-of-domain dataset consisting 461 images taken from the MSCOCO dataset [14] was also introduced, which contains ambiguous verbs [19].

We focused on paraphrasing the English sentences which are considered as source language. Table 1 shows an example of a paraphrased image caption based on four elementary operations (deletion, insertion, reordering, and substitution) and Figure 1 shows the reference image. As paraphrasing the whole 29k triplet training dataset (29k training dataset) using crowdsourcing would not be efficient in terms of cost and time, we crowdsourced only 10k triplets of this dataset (10k training dataset), along with the whole development and testing datasets.

We used Crowdflower[2] (now Figure Eight) as the crowdsourcing platform. Each crowdworker was instructed to paraphrase at least two image captions for one session. We limited the task to English speakers, or at least those who spoke English as their second language, to maintain quality. We discarded sentences that were not valid such as randomly inputted character, empty string, or captions that aren't English. The crowdsourcing process took about 3 months and 201 workers participated from 16 countries such as the United States, Philippines, and Malaysia. Each workers created 50.1 quintuplets of paraphrases on average.

2.3. Semi-supervised Paraphrase Generation on Full WMT17 Dataset

Furthermore, to complete the paraphrasing on the full WMT17 dataset, we then used 10k quintuplets of crowdsourced paraphrases and constructed neural paraphrase model using four encoder-decoder long short-term memory (LSTM) models with attention [20] for each paraphrase oper-

[2] http://www.figure-eight.com

Table 1: Image caption and example paraphrases

Operation		Sentence
Image Caption		A little gray dog jumps over a small hurdle.
Paraphrase	**Deletion**	A little gray dog jumps over a hurdle.
	Insertion	A little gray dog jumps over a small hurdle successfully.
	Substitution	A little gray dog pass over a small hurdle.
	Reordering	Over a small hurdle, a little gray dog jumps.

ation. We tuned and tested our automatic neural paraphrase model using these crowdsourced paraphrases of the development and testing datasets, respectively. With these four paraphrasing models, we generated multi-paraphrases on the remaining 19k image captions.

The generated 19k dataset was combined with the original crowdsourced 10k training dataset. Finally, these 29k paraphrased dataset are combined with original dataset resulting 58k-triplet training dataset for each operation. In conclusion, the 29k paraphrased training dataset is working as the regularizer for the original dataset. These are the final data that will be used to train a mixture-of-experts translation model, which is described in the next section. The data will be publicly available to augment the WMT17 dataset.

Based on our empirical observation, using paraphrased data on development and testing dataset will reduce the performance of the overall system. When using paraphrased data on development, the training objective becomes unclear, and the loss returned will not represent the real loss. Given that, we emphasize that the use of paraphrased dataset in translation step was done on training step, in combination with original dataset. In this stage, the paraphrases were acting as regularizer and the means of ensembling, improving robustness of the ensembled model as a whole.

3. Neural Caption Translation

This section describes several approaches on using our proposed multi-paraphrase operations to improve NMT. The score of these approaches will then be compared with WMT baseline and our encoder-decoder LSTM NMT baseline.

3.1. Combining All Data in a Single Model

This method was done by just using the paraphrase as a means for data augmentation in source side, such as reported by Nichols et al. (2010) to leverage SMT system [10]. All paraphrases and its original sentence were combined, and the target sentence was duplicated to the number of multiple paraphrases. This approach was done to measure the baseline performance with augmented data.

3.2. Multi-source Model

We implemented Zoph and Knight (2016) multi-source NMT to incorporate various paraphrase inputs with one output [21]. For this model, the encoded representation and attention were combined by concatenation. They reported that this model has the advantage of information triangulation to reduce ambiguity. In their paper, they used several translation pairs such as $\{$French, German$\}$ to English in which this triplet of language has similar language structure. However, given this advantages, the use of this model to monolingual input has never been investigated.

3.3. Uniform-weighted Ensemble Model

For this uniform weighted ensemble model, we trained NMT models which source sentence has been paraphrased based on each elementary operations and another one that uses original source sentence, resulting five expert NMT models. After that, these five models are ensembled by averaging each output layer probability distribution, so that every model was weighted uniformly. This model is used to compare the performance with mixture-of-experts model listed in the next subsection, where each expert model have different weight.

The training of this translation model consists of two steps. The first step is to train five translation models based on each paraphrase as the source sentence using the 56k dataset (the combination of original and paraphrased source sentences). Five of those models are trained against the same target sentence. Each model is then regarded as an expert model. Each of the expert models operates on subword level, tokenized by Sentence Piece with 3000 vocabulary unit[3].

3.4. Mixture-of-experts Model

Next, we adopted the mixture-of-experts model proposed by Garmash and Monz (2016). Here, instead of linear layer proposed in their study [22], the expert model is implemented into a single LSTM layer hid that receives the concatenated decoder hidden state output h_n.

$$c_t = tanh(LSTM_{hid}([h_0, h_1, ..., h_n]))$$
$$g_{0:i} = softmax(W_{gate} D(c_t) + b_{gate}).$$

A $softmax$ function is then applied to obtain the weights of each expert model's output layer o_n. Assuming W_n is the weight of the output layer from expert n. Then, the aggregated weight W_{agg} is a linear combination function of each of those weights:

$$W_{agg} = g_0 W_0 + g_1 W_1 + ... + g_n W_n.$$

[3] https://github.com/google/sentencepiece

Table 2: Paraphrasing model result in BLEU and METEOR

Operation	BLEU	METEOR
Deletion	53.0	42.2
Insertion	56.1	40.5
Reordering	47.2	42.0
Substitution	59.6	44.8

For this model, a 50% dropout D will be applied on the hidden representation after $tanh$ nonlinearity was applied. The regularized representation was further transformed by the gate layer which has the same output size with the number of expert.

A diagram of mixture-of-experts neural caption translation model using our proposed approach is shown in Fig. 2. First, the source sentence is paraphrased into four different paraphrases used to train each of the expert model. Then, each expert will pass their abstract decoding state into mixture model which will produce weights as many as the number of expert. The resulting weight distribution is the linear combination function between each expert's output probability distribution and gating weight produced by mixture model.

4. Experiments

The purpose of this experiment is to choose the best type of model suitable for our multi-paraphrase, by comparing score between Bahdanau et al. NMT baseline and several popular multi-source NMT.

4.1. Setup

We followed the training, development, and test set-up of WMT17 shared task. All result were scored using *multeval* [23] with lowercased and tokenized sentences. We used BLEU [24] and METEOR [25] as evaluation metrics.

The multi-source NMT has five single-depth encoders with 512 hidden size trained with Adam [26]. The mixture-of-experts model was trained using RMSprop optimizer with 0.0001 learning rate [27]. In every increase of development loss, the learning rate is decayed by half into maximum 5 decays. The results are decoded with beam size of 5.

4.2. Evaluation of Neural Paraphrase Model

We constructed four encoder-decoder LSTM models with attention [20] for each elementary paraphrase operation. Each model has a bidirectional encoder and attentional decoder with one layer, 50% dropout ratio, and 512 hidden layer size. Implementation was done using Chainer framework version 3.0 [28] and ran on GTX Titan X GPU. We used Adam [26] as the optimizer with decaying alpha into half in every development loss increase with maximum of 7 decays for training early stopping. After stopping the training, model with the lowest development was selected and used for decoding.

Table 2 lists the scores of the paraphrases produced with our automatic paraphrasing model. The substitution operation produced the highest BLEU score while the reordering operation producing the lowest BLEU score. This was expected because the reordering operation sometimes includes the changing of the active/passive properties of a sentence. Overall, we believe this score is high enough to paraphrase the remaining 19k WMT dataset.

4.3. Translation Model Results

Table 3 shows the performance of our proposed neural caption translation. All results using our multi-paraphrase outperformed the NMT baseline. There are no improvements gained from combining all data, which is the simplest form of data augmentation. This simple combination of data breaks the relation existed between each paraphrases that mention the same image. Furthermore, we cannot be sure that each source sentence has the same amount of paraphrase. By considering these factors, we utilized multi-source NMT and multi-expert NMT, which yield better BLEU and METEOR score.

This performance increase indicates that each expert model is slightly different between each other, and worked well in uniform-weighted ensemble and mixture-of-experts scenario. This model also performed better than uniform-weighted NMT in three cases. Moreover, the mixture-of-experts model performed better in out-of-domain ambiguous MSCOCO test dataset, implying that overfitting did not occur. This also proves the argument that adding additional knowledge will improve model performance on disambiguating inputs. From applying to these several models, we can conclude that our elementary operation paraphrase is suitable to be used as a means for ensembling.

Table 4 shows the current submission systems in the official WMT17 shared task which submissions consist of one textual model [29] and several multimodal models. Our proposed approach outperformed the baseline in WMT17 with a 13.2 BLEU score margin. Our proposed model, although it is textual, could produce competitive result with other multimodal models. The mixture-of-experts model outperformed several multimodal models such as other WMT submission [30, 31, 32, 33]. Even in the out-of-domain dataset of COCO 2017, the mixture-of-experts model also performed reasonably high with a 28.0 BLEU score. Nevertheless, our score was close to that best score. This proved that the paraphrasing of the source side also helped our model to work with unseen data and prevent overfitting.

4.4. Discussion

To further analyze the contribution between the experts trained on the original data and that trained on paraphrased data, we compared the translation process step-by-step in our proposed approach. This source sentence shown in Table 5 was translated using each baseline model (an expert), resulting five different translation hypotheses. Each expert has been trained with slightly different paraphrased source

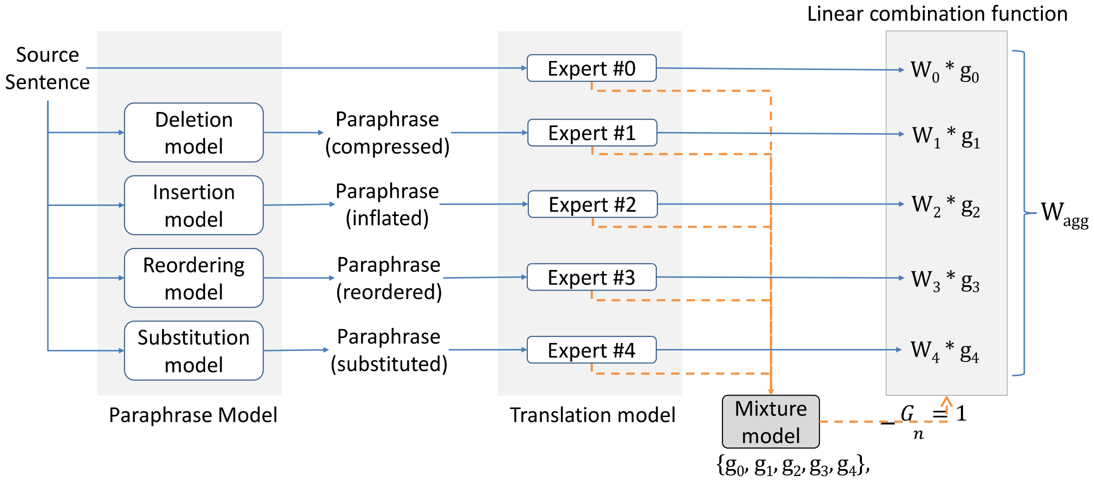

Figure 2: Diagram of proposed mixture-of-experts neural caption translation model

Table 3: The performance of proposed neural caption translation in comparison with the baseline.

Textual Model	**Test 2016**		**Test 2017**		**Test COCO 2017**	
	BLEU	**METEOR**	**BLEU**	**METEOR**	**BLEU**	**METEOR**
Our NMT Baseline	37.7	55.6	30.1	49.7	25.0	44.6
Combine all data	36.7	53.9	29.6	47.7	25.1	43.7
Multi-source NMT	37.6	55.4	30.1	49.4	24.4	44.3
Uniform weighted ensemble	39.6	56.9	31.4	50.7	26.7	46.0
mixture-of-experts ensemble	**40.5**	**57.6**	**32.5**	**51.3**	**28.0**	**46.8**

Table 4: Existing submission systems in official WMT17 shared task.

Textual Model	**Test 2016**		**Test 2017**		**Test COCO 2017**	
	BLEU	**METEOR**	**BLEU**	**METEOR**	**BLEU**	**METEOR**
Official WMT Baseline	32.5	52.5	19.3	41.9	18.7	37.6
Zhang et al. (2017)	-	-	31.9	53.9	28.1	48.5

Multimodal Model	**Test 2016**		**Test 2017**		**Test COCO 2017**	
	BLEU	**METEOR**	**BLEU**	**METEOR**	**BLEU**	**METEOR**
Madhyastha et al. (2017)	-	-	25.0	44.5	21.4	40.7
Calixto et al. (2017)	41.3	59.2	29.8	50.5	26.4	45.8
Ma et al. (2017)	-	-	31.0	50.6	27.4	46.5
Helcl and Libovicky (2017)	36.8	53.1	31.1	51.0	26.6	46.0
Caglayan et al. (2017)	41.0	60.4	33.4	54.0	28.5	48.8

Table 5: Examples of resulting sentences in mixture-of-experts model.

Translation Model	**Type**	**Sentences**	**BLEU-1**
(Data)	**Original**	**two motorcycles drive on a road along the river .**	
Baseline	Original	zwei motorradfahrer fahren auf einer straße entlang .	0.75
Single Paraphrase Model	Deletion	zwei motorräder fahren auf einer straße am fluss .	0.87
	Insertion	zwei motorradfahrer fahren auf einer straße am fluss .	0.84
	Reordering	zwei motorradfahrer fahren auf einer straße am fluss entlang .	0.95
	Substitution	zwei motorradfahrer fahren auf einer straße am flussufer .	0.82
Uniform Weight	Ensemble	zwei motorradfahrer fahren auf einer straße am fluss .	0.84
mixture-of-experts	Ensemble	zwei motorräder fahren auf einer straße am fluss entlang .	**0.97**
(Data)	**Target**	**zwei motorräder fahren auf einer straße dem fluss entlang .**	

sentence. We calculated BLEU-1 scores for each hypothesis against the target, resulting the source-reordered expert model yielded the best result between all experts.

The aim of proposed mixture-of-experts model task is to make sure the best part of each model is kept, and leaving out any noise or error that might occur in each model result. As can be seen from the German result from the mixture-of-experts model compared with the target sentence, the only difference is the word "*am*" in which the correct one should be "*dem*".

In this example, in deletion translation result, the word "*motorräder*" is decoded instead of "*motorradfahrer*". Another example is the phrase "*fluss entlang*" which can only be found in reordering translation result. This goodness on each expert model however, should be kept by the mixture model by distributing right word in every word being decoded. In conclusion, the final result of the ensemble of expert model combines every goodness in each expert model.

Quantitatively, the mixture-of-experts model successfully kept the good feature of best performing 0.87 and 0.95 BLEU-1 score yielded in source-deleted and source-reordered model results respectively, resulting 0.97 BLEU-1 score. This is a significant improvement compared with the BLEU-1 score of the uniform weighted model that was only increased into 0.84.

5. Conclusions and Future Works

A single caption cannot represent all the information of the image to which it refers to. In this study, we elaborated an image by various paraphrase operations. This enables us to incorporate additional knowledge from image to the translation process, without using the image itself, but diffused in a form of paraphrase.

We successfully generated multi-paraphrase sentences of the WMT17 Multimodal Translation Task dataset through crowdsourcing which will be publicly available. We constructed an automatic paraphrase generation model, and used it with the multi-expert approach within NMT.

The results indicate that our proposed paraphrase elementary operations are best to be used for ensembling, especially on multi-expert ensembling settings. The hypothesis of regularizing models by paraphrasing on the source sentence was proven to be effective. In the future, we will further investigate various methods of incorporating visual information into NMT models.

6. Acknowledgement

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

7. References

[1] R. De Beaugrande and W. Dressler, *Introduction to text linguistics*, ser. Longman linguistics library. Longman, 1981. [Online]. Available: https://books.google.co.jp/books?id=mvJsAAAAIAAJ

[2] A. Barreiro, *SPIDER: A System for Paraphrasing in Document Editing and Revision — Applicability in Machine Translation Pre-editing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 365–376. [Online]. Available: https://doi.org/10.1007/978-3-642-19437-5_30

[3] C. Callison-Burch, P. Koehn, and M. Osborne, "Improved statistical machine translation using paraphrases," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL '06)*, Stroudsburg, PA, USA, 2006, pp. 17–24. [Online]. Available: http://dx.doi.org/10.3115/1220835.1220838

[4] W. He, S. Zhao, H. Wang, and T. Liu, "Enriching smt training data via paraphrasing," in *Proceedings of IJCNLP*, 2011.

[5] M. Simard, N. Ueffing, P. Isabelle, and R. Kuhn, "Rule-based translation with statistical phrase-based post-editing," in *Proceedings of the Second Workshop on Statistical Machine Translation (StatMT '07)*, Stroudsburg, PA, USA, 2007, pp. 203–206. [Online]. Available: http://dl.acm.org/citation.cfm?id=1626355.1626383

[6] M. Simard, C. Goutte, and P. Isabelle, "Statistical phrase-based post-editing," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL '07)*, Rochester, NY, USA, 2007, pp. 508–515.

[7] A.-L. Lagarda, V. Alabau, F. Casacuberta, R. Silva, and E. Díaz-de Liaño, "Statistical post-editing of a rule-based machine translation system," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, ser. NAACL-Short '09, Stroudsburg, PA, USA, 2009, pp. 217–220. [Online]. Available: http://dl.acm.org/citation.cfm?id=1620853.1620913

[8] S. Pal, P. Lohar, and S. K. Naskar, "Role of paraphrases in pb-smt," in *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 8404*, ser. CICLing 2014. Berlin, Heidelberg: Springer-Verlag, 2014, pp. 242–253. [Online]. Available: https://doi.org/10.1007/978-3-642-54903-8_21

[9] N. Madnani and B. J. Dorr, "Generating targeted paraphrases for improved translation," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 3, pp. 40:1–40:25, July 2013. [Online]. Available: http://doi.acm.org/10.1145/2483669.2483673

[10] E. Nichols, F. Bond, D. S. Appling, and Y. Matsumoto, “Paraphrasing training data for statistical machine translation,” *Journal of Natural Language Processing*, vol. 17, no. 3, pp. 3_101–3_122, 2010.

[11] G. Hirst, “Paraphrasing paraphrased,” *Invited talk at the ACL International Workshop on Paraphrasing*, 2003.

[12] I. Calixto, Q. Liu, and N. Campbell, “Doubly-attentive decoder for multi-modal neural machine translation,” in *ACL*, 2017.

[13] P. Resnik, O. Buzek, Y. Kronrod, C. Hu, A. J. Quinn, and B. B. Bederson, “Using targeted paraphrasing and monolingual crowdsourcing to improve translation,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 4, no. 3, p. 38, 2013.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[15] A. Prakash, S. A. Hasan, K. Lee, V. Datla, A. Qadir, J. Liu, and O. Farri, “Neural paraphrase generation with stacked residual lstm networks,” in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, December 2016, pp. 2923–2934. [Online]. Available: http://aclweb.org/anthology/C16-1275

[16] R. Bhagat and E. Hovy, “What is a paraphrase?” *Computational Linguistics*, vol. 39, no. 3, pp. 463–472, 2013.

[17] D. Elliott, S. Frank, K. Sima’an, and L. Specia, “Multi30k: Multilingual english-german image descriptions,” in *Proceedings of the 5th Workshop on Vision and Language*, 2016, pp. 70–74.

[18] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” *CoRR*, vol. abs/1505.04870, 2015. [Online]. Available: http://arxiv.org/abs/1505.04870

[19] D. Elliott, S. Frank, L. Barrault, F. Bougares, and L. Specia, “Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description,” in *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, September 2017.

[20] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: http://arxiv.org/abs/1409.0473

[21] B. Zoph and K. Knight, “Multi-source neural translation,” *CoRR*, vol. abs/1601.00710, 2016. [Online]. Available: http://arxiv.org/abs/1601.00710

[22] E. Garmash and C. Monz, “Ensemble learning for multi-source neural machine translation,” in *COLING*, 2016.

[23] A. L. Jonathan Clark, Chris Dyer and N. Smith, “Better hypothesis testing for statistical machine translation: Controlling for optimizer instability,” in *Proceedings of the Association for Computational Lingustics*, 2011.

[24] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu, “Bleu: a method for automatic evaluation of machine translation,” 2002, pp. 311–318.

[25] A. Lavie and A. Agarwal, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” 2005, pp. 65–72.

[26] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.

[27] A. Graves, “Generating sequences with recurrent neural networks,” *CoRR*, vol. abs/1308.0850, 2013. [Online]. Available: http://arxiv.org/abs/1308.0850

[28] S. Tokui, K. Oono, S. Hido, and J. Clayton, “Chainer: a next-generation open source framework for deep learning,” in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*, 2015. [Online]. Available: http://learningsys.org/papers/LearningSys_2015_paper_33.pdf

[29] J. Zhang, M. Utiyama, E. Sumita, G. Neubig, and S. Nakamura, “Nict-naist system for wmt17 multimodal translation task,” in *WMT*, 2017.

[30] P. S. Madhyastha, J. Wang, and L. Specia, “Sheffield multimt: Using object posterior predictions for multimodal machine translation,” in *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 470–476. [Online]. Available: http://www.aclweb.org/anthology/W17-4752

[31] I. Calixto, K. D. Chowdhury, and Q. Liu, “Dcu system report on the wmt 2017 multi-modal machine translation task,” in *Proceedings of the Conference of Machine Translation (WMT)*, vol. 2, 2017.

[32] M. Ma, D. Li, K. Zhao, and L. Huang, “Osu multimodal machine translation system report,” in *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 465–469. [Online]. Available: http://www.aclweb.org/anthology/W17-4751

[33] J. Helcl and J. Libovický, "CUNI system for the WMT17 multimodal translation task," *CoRR*, vol. abs/1707.04550, 2017. [Online]. Available: http://arxiv.org/abs/1707.04550