# Enhancing Neural Machine Translation with Image-based Paraphrase Augmentation

Johanes Effendi[1] Sakriani Sakti[1,2] Katsuhito Sudoh[1,2] Satoshi Nakamura[1,2]
[1]Nara Institute of Science and Technology [2]RIKEN AIP
{johanes.effendi.ix4, ssakti, sudoh, s-nakamura}@is.naist.jp

## 1 Introduction

As defined by De Beaugrande and Dressler, a paraphrase is an approximate conceptual equivalence among outwardly different material [4]. In this research, we treat an image as a representation of sentence idea, which can be regarded as the basis of paraphrasing. Furthermore, as paraphrasing to enable multi-source information in NMT is not much investigated yet, in this study we explore the use of this 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 generate image descriptions in a target language. The results from most submitted systems reveal that the additional image features could only slightly contribute to system performance although the model itself uses a lot of resource.

Here, we attempt to go in another direction in which we incorporate the image information by using image-based paraphrasing without using the image itself. Then, we applied this proposed way of paraphrasing to build a paraphrase corpus in which we developed a neural paraphrasing model. Furthermore, this initiates multi-expert neural machine translation (NMT) model that translates an image caption without using the image itself, but instead using a series of visual descriptions describing the image.

In summary, the contributions of this work include:

1. Introduce a novel way of image-based paraphrasing.
2. Generate multiple paraphrase sentences of the WMT17 Multimodal Translation Task dataset through crowdsourcing.
3. Utilize multi-expert translation in neural machine translation using our proposed paraphrase; and
4. 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 Proposed Method

By having five caption paraphrases and their translation, our proposed 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 consisting of the original both 29k dataset and paraphrased dataset. Five of those models are trained against the same target sentence. Each model is then regarded as an expert model.

#### 2.0.1 Uniform-weighted Ensemble Model

For this uniform weighted ensemble model, we ensembled five expert models by averaging each output layer probability distribution so every model was weighted uniformly.

#### 2.0.2 Mixture-of-experts Model

Next, we adopted the mixture-of-experts model to determine the weight for each output layer probability distribution:

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

Here, the expert model is implemented into a single LSTM layer *hid* that receives the concatenated decoder hidden state output $h_n$. A $softmax$ function

Table 1: 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 |
| 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 2: 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 |
| Caglayan et al. (2017) | 41.0 | 60.4 | 33.4 | 54.0 | 28.5 | 48.8 |

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:

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

The resulting weight distribution in the output layer is the linear combination function between each expert's output probability distribution and gating weight produced by mixture model.

# 3 Corpus Creation

## 3.1 Image-based paraphrase definition

Our proposed paraphrase needs a set of sentences with an image as its representation of idea. The caption of this dataset can be regarded as paraphrase, such as done by Prakash et al. for their neural paraphrase generation study [7] using MSCOCO dataset [6]. While this is sufficient, we cannot define what kind of operation has been done from the original sentence to the paraphrase. Therefore, this might cause the paraphrase to become noise to each other.

To prevent this, we need to define a set of paraphrase operation which covers all possible paraphrase variations. Based on Bhagat and Hovy's idea of quasi-paraphrases [2], we constructed a paraphrase corpus based on various elementary operations which is grouped from their quasi-paraphrases into 4 elementary paraphrase operations such as: deletion, insertion, reordering, and substitution, and used the WMT17 Multimodal Translation Task dataset [5]. The collection was done through a crowdsourcing platform on the partial WMT17 dataset. After that, we constructed our automatic neural paraphrasing model based on partial data to generate the paraphrase sentences of the full WMT17 dataset.

## 3.2 Partially crowdsource the WMT17 dataset

The WMT Multimodal Translation task data consists of 29000, 1014, and 1000 triplets respectively for the training, development and testing [5]. An out-of-domain dataset consisting 461 images was also introduced, which contains ambiguous verbs.

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. For the remaining dataset, we paraphrased it using a neural paraphrase model trained using the crowdsourced dataset.

# 4 Experiments

## 4.1 Setup

We constructed four encoder-decoder LSTM models with attention [1] 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.

### 4.2 Translation Model Results

Table 1 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.

The mixture-of-expert model performed better than uniform-weighted NMT in three cases. 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 2 shows the current submission systems in the official WMT17 shared task which submissions consist of one textual model [8] and a multimodal model. 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 except [3].

## 5 Conclusions

In this study, we elaborated an image by various paraphrase operations which enables us to incorporate additional knowledge from image to the translation process. Furthermore, we successfully generated multi-paraphrase sentences of the WMT17 Multimodal Translation Task dataset through crowdsourcing. 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 multi-expert ensembling settings.

## 6 Acknowledgement

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

## References

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

[2] Rahul Bhagat and Eduard Hovy. What is a paraphrase? *Computational Linguistics*, Vol. 39, No. 3, pp. 463–472, 2013.

[3] Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. LIUM-CVC submissions for WMT17 multimodal translation task. *CoRR*, Vol. abs/1707.04481, , 2017.

[4] R. De Beaugrande and W.U. Dressler. *Introduction to text linguistics*. Longman linguistics library. Longman, 1981.

[5] 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*, pp. 70–74, 2016.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

[7] Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Neural paraphrase generation with stacked residual lstm networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2923–2934, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[8] Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Graham Neubig, and Satoshi Nakamura. Nictnaist system for wmt17 multimodal translation task. In *WMT*, 2017.