PAPER

Leveraging Neural Caption Translation with Visually Grounded Paraphrase Augmentation

Johanes EFFENDI[†a)], ***Nonmember***, **Sakriani SAKTI**[†,††], ***Member***, **Katsuhito SUDOH**[†,††], ***Nonmember***, ***and*** **Satoshi NAKAMURA**[†,††], ***Member***

SUMMARY

Since a concept can be represented by different vocabularies, styles, and levels of detail, a translation task resembles a many-to-many mapping task from a distribution of sentences in the source language into a distribution of sentences in the target language. This viewpoint, however, is not fully implemented in current neural machine translation (NMT), which is one-to-one sentence mapping. In this study, we represent the distribution itself as multiple paraphrase sentences, which will enrich the model context understanding and trigger it to produce numerous hypotheses. We use a visually grounded paraphrase (VGP),which uses images as a constraint of the concept in paraphrasing, to guarantee that the created paraphrases are within the intended distribution. In this way, our method can also be considered as incorporating image information into NMT without using the image itself. We implement this idea by crowdsourcing a paraphrasing corpus that realizes VGP and construct neural paraphrasing that behaves as expert models in a NMT. Our experimental results reveal that our proposed VGP augmentation strategies showed improvement against a vanilla NMT baseline.

key words: *visually grounded paraphrase, data augmentation, neural machine translation*

1. Introduction

Sequence-to-sequence problems are usually solved by assuming that the model is mapping from the source sequence distribution to the target sequence distribution. The representation of these distributions can be varied depending on the task. In question-answering tasks, the source and target distributions can be represented by questions and answers. Similarly, in machine translation (MT), the source and target distributions are represented by source and target sentences. Since these distributions are the concept's representations, their sentences can be replaced with other equivalent ones.

However, common approaches in machine translation (MT) assumes that this task is simplified by a one-to-one sentence mapping problem. In this study, we propose to represent this mapping task by a many-to-many approach, specifically by depicting the distribution with multiple paraphrase sentences. Although all of these sentences are valid, the model's understanding can be enriched and triggered to produce multiple valid hypotheses. From the linguistic perspective, this definition also resembles Hirst's (2003) idea, which defines paraphrasing as describing a situation by many different way of expressions [1].

Furthermore, to represent the source and target distributions in multiple paraphrase sentences, a semantic boundary needs to be defined. Unconstrained paraphrasing on a sentence might lead to an inappropriate shifts in semantic meaning that diverged from the original idea of the paraphrased sentence. For this reason, we need a boundary within the distribution to ensure that there is a common ground exists among several paraphrases.

In this study, we propose a new variant of paraphrases called visually grounded paraphrases (VGP), which we define VGP as a set of paraphrases that describe about the same image. Different phrases and wording between two paraphrase sentences can also be associated with an object or an action shown in the image, which work as a pivot [2]. This association implies that multiple paraphrases can be produced by operating with these associated phrases or words. Additionally, from the amount of information contained, an image has more information than a single caption that represents it. Therefore, to sufficiently represent an image, multiple sentences are preferable for representing the information contained in it than just a single sentence.

Grounding visual description and its image representation enables a wide variety of functionality, especially in multimodal settings. VGP can be used in image captioning [3] and visual question answering [4] to improve the understanding of both textual and visual information. In this study, we also showed that VGP is useful for neural machine translation as an augmentation method, in which we use several variations of VGP as multi-source input for a neural MT (NMT). In all of its use cases, VGP improves the relations between textual and visual data. In addition, it can also represent visual data in a textual representation.

We evaluated our proposed VGP concept using a caption translation task to measure its effectiveness in a practical task and compared it with other approaches. This task hosted by the Second Conference of Machine Translation (WMT17) seeks to translate a given image description into the target language. Most of multimodal 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 show that the additional image features only slightly contribute to system performance. As pointed out by Calixto et al. [5], the image-text latent repre-

Manuscript received March 6, 2019.
Manuscript revised August 1, 2019.

[†]The authors are with the Augmented Human Communication Lab., Nara Institute of Science and Technology, Ikoma-shi, 630-0192 Japan.

[†]The author are with the RIKEN, Center for Advanced Intelligence Project AIP, Ikoma-shi, 630-0192 Japan.

a) E-mail: johanes.effendi.ix4@is.naist.jp

DOI: 10.1587/trans.E0.??.1

Table 1 Image caption and example paraphrases

Operation		Sentence
Image Caption		Two white dogs are running on the grass.
Paraphrase	**Deletion**	Two white dogs are running.
	Insertion	Two white dogs wearing collars are running on the grass.
	Substitution	On the grass, two white dogs are running.
	Reordering	Two white dogs are racing on the grass.

sentation combination approach has not yielded significant improvement on WMT 2017 Multimodal shared task dataset testing.

However, compared to the previous attempts, our approach can be considered as diffusing the image information with visually-grounded paraphrases without using the image itself. The resulting paraphrase captions are then utilized within a multi-source and multi-expert NMT model. Another advantage of our textual approach is that it uses fewer computational resources than the multimodal approach, and still maintains comparable performance.

The following are the contributions of this work:

1. Proposed the concept of visually grounded paraphrases that introduce a new way of creating a paraphrase corpus through image captions;
2. Generated VGP sentences of the WMT17 Multimodal Translation Task dataset through crowdsourcing;
3. Developed automatic paraphrase generation in a semi-supervised manner;
4. Utilized multi-expert translation in neural machine translation using our proposed paraphrases;
5. Improved the baseline used at WMT17 with a 13.2 BLEU score margin, which is near the top score that used a multimodal model.

2. Visually Grounded Paraphrases Generation

2.1 Defining Visually Grounded Paraphrases

Although the logical definition of a paraphrase is generally as simple as a “semantic equivalence", linguists accept a broader, approximate equivalence that allows more relaxed paraphrasing action constrained by an idea. Hirst (2003) argued that paraphrases aren’t fully synonymous because they have some pragmatic differences [1]. Such hard-to-define definitions causes paraphrasing-related technologies to use the safest available operations to avoid producing non-valid paraphrases.

For example, by analyzing the nature of the operations found in paraphrase applications in MT, we found that only non-intrusive paraphrase operations like phrase substitution or phrase reordering were used. Simply adding or flexibly rephrasing information is impossible because the only source of information that can be used as the inspiration for the paraphrasing is the source sentence. Nevertheless, such common usages of paraphrases in MT fail to demonstrate these wide and flexible definitions of paraphrases. From a bigger perspective, we can say that this issue happens because the sentence being paraphrased doesn’t have a tangible representation of ideas. Further intrusive operation on this original sentence, might completely change its meaning due to a lack of grounded representation of the idea.

Therefore, we proposed a visually grounded paraphrase (VGP)[†], to implement a wider definition of paraphrasing and completely utilize its benefits. We also implement paraphrases for each visual representation as a set of captions being paraphrased by a set of paraphrase operations. This idea covers every possible paraphrase that is being made and enables the tracking of the operation that has been done on the source sentence. The grouping of possible paraphrase operations is described in the next subsection.

2.2 Paraphrase Elementary Operations

Fig. 1 Reference image for captioning and paraphrasing shown in Table 1 [6].

To train a paraphrasing model with our VGP, first, we need to build a set of paraphrased source sentences with images as the basis for paraphrasing. This requirement resembles image captioning datasets such as the Microsoft Common Object in Context (MSCOCO) dataset [7]. The captions of this dataset can be regarded as paraphrase sets, such as those done by Prakash et al. for their neural paraphrase generation study [8]. They reported how their annotators described the most obvious things in an image and concluded that several captions of an image can be counted as paraphrases. Although this result may be true, we cannot

[†]Not to be confused with Chu et al.’s (2018) work which uses the same term that refers to a synonymous set of clustered phrases with the help of object grounding in an image [2]. In our work, we focus on generating paraphrase sentences.

define what operations have been done from the original sentence to the paraphrase. Consequently, the arbitrary nature of the corpus distribution might cause the paraphrases to be regarded as noise by each other.

To prevent this situation, a set of paraphrase operations covering every possible paraphrase variation needs to be defined. Bhagat and Hovy categorized the variations of human paraphrases [9] 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 operations, including tense changes, metaphor substitution, and function-word variations.

Some quasi-paraphrase operations have very small frequency in the MTC [10] and MSRP corpora [11], as reported by them. On the other hand, creating 25 kinds of paraphrases from one original sentence in corpus creation might be too difficult and too fine-grained. Then, we grouped them into four elementary paraphrase operations: deletion, insertion, reordering, and substitution. These groups cover the possible quasi-paraphrases defined, while reducing the number of paraphrases needs to be made from one sentence. Moreover, we call our paraphrase as *elementary* operation because it's possible to utilize more than one operations to create a new paraphrases, which has multiple elementary operations. For this study, we focused on constructing a paraphrase corpus based on these four operations.

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 [12]. The paraphrase collection was done through a crowdsourcing platform on part of the WMT17 Multimodal Translation Task dataset [6]. After that, we constructed our automatic neural paraphrase model based on partial data to generate paraphrase sentences of the full WMT17 dataset. The details are described below.

2.3 Crowdsourcing Paraphrases on Partial WMT17 Dataset

The WMT17 Multimodal Translation Task dataset [6] 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 [13] that was extended to include manually translated German and French captions. The data respectively 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 [7] was also introduced, which contains ambiguous verbs [14].

We focused on paraphrasing English sentences, which are considered the source language. Table 1 shows an example of a paraphrased image caption based on the four elementary operations (deletion, insertion, reordering, and substitution), and Fig. 1 shows the reference image. Since paraphrasing the entire 29k triplet training dataset (29k training dataset) with crowdsourcing is inefficient in terms of cost and time, we crowdsourced only 10k triplets of this dataset (10k training dataset) along with the entire development and testing datasets.

We used Figure Eight[†] as the crowdsourcing platform. Each crowdworker was instructed to paraphrase at least two image captions per session, based on the image to which it refers. As shown in Fig. 2, the crowdworker are always be able to see the reference image and the original caption. With this way, the created paraphrases are always refers to the original caption, and its definition doesn't stray away because an image is always provided as reference.

We limited the task to English speakers or those who spoke English as their second language to maintain quality. We discarded such invalid sentences as those with randomly inputted characters, empty strings, or captions that aren't in English. The crowdsourcing process took about three months with 201 participants from 16 countries including the United States, Philippines, and Malaysia. Each worker created an average of 50.1 paraphrase quintuplets.

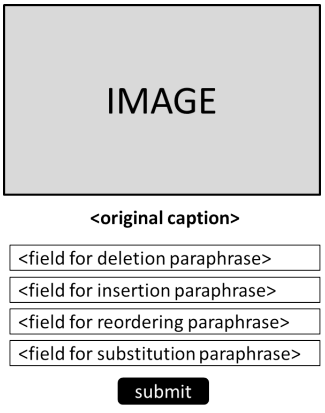

Fig. 2 Form layout during the crowdsourcing.

2.4 Corpus Analysis Method

In this subsection, we show the statistics of the corpus to show how each operation is implemented in the crowdsourcing process. There are two kinds of analysis: word-based and part-of-speech-based. First, we created the following word-based measurements to show the statistics of our corpus based on each operation:

Ratio of target sentence vs. source sentence: By calculating this ratio, we identified the compression or inflation rate of each sub-operation corpus with which we can evaluate the effectiveness of each operation.

Average number of deleted words per sentence: This metric shows how many words were deleted from the source

[†]http://www.figure-eight.com

sentence on average.

Average number of inserted words per sentence: This metric shows how many words were inserted to the target sentence on average.

Average reordered word distance: This metric shows how a word was reordered in the sentence by comparing its position distance between the source and target sentences.

Average number of substituted words per sentence: This metric shows how many words were substituted between the source and target sentences on average. For reordering and substitution operation analysis, we used the word alignment between the source and target language sentences estimated using Fast Align [15].

Next, we performed a part-of-speech-based analysis by comparing the source and target sentences with their part-of-speech (POS) tags predicted by the Stanford POS Tagger [16]. For this analysis, we show the top three POS tags for every operation. This will give an overview on how the crowdworker prefer which kinds of words to be deleted, inserted, reordered, or substituted.

Then, in order to measure the semantic consistency of the paraphrases, we calculate the average semantic embedding distance from original source caption to its paraphrased counterpart. We used Bidirectional Encoder Representations from Transformers (BERT) [17] to generate sentence embedding. As the baseline, we compare with the distance of source sentence and random sentence in dataset. We use mean squared error (MSE) and cosine distance between two sentence embedding vectors as distance measurement.

2.5 Corpus Analysis

(1) Word-based operation characteristics

To calculate the deletion operation effect on the source sentences, we compared the ratio of the number of words in the target sentence to the number in the source sentence and found that the words in the target sentences decreased by 25.61%, where 3.20 words were deleted per sentence on average (Table 2). We did the same calculation in the insertion sub-corpus and found that the number of words in the target sentences increased by 25.24%, where an average of 2.89 words were inserted per sentence.

Table 2 Sub-corpus characteristics

Parameter	# word
Average number of deleted words per sentence	3.197
Average number of inserted words per sentence	2.877
Average distance of reordered word	5.348
Average number of word substitutions per sentence	1.663

To measure the reordering elementary operation for the sub-corpus, we calculated the shift distance of a word in the source and target sentences and found that those in the latter shifted on average by as many as 5.35 words. The distance calculated in the reordering happened when the source sentence was paraphrased into its passive form. Another reordering approach switched the order of such sentence information as time, place, or tool.

For the substitution sub-corpus, we measured how many words were replaced by checking whether the source and target words matched in an alignment. If they are different, then we counted that as one word substitution. As seen in Table 2, we found an average of 1.66 word substitutions per sentence in the substitution sub-corpus. This means that at most 1 or 2 words were substituted in a sentence.

(2) POS-based operation characteristics

We counted the types of words that were most deleted and inserted. For reordering and substitution, we counted a pair of source and target word types to identify which word type was usually preferred by the crowdworkers.

Table 3 Top Three Most Operated POS Tags

Deletion	Insertion	Reordering	Substitution
NN	NN	DT DT	NN NN
IN	JJ	NN NN	NNS NNS
JJ	IN	IN IN	VBG VBG

Table 3 shows that nouns (NN), adjectives (JJ), and conjunctions (IN) were usually deleted or inserted. This correlates with how most sentences are deleted or inserted: by adding or removing time, place, or tool information, such as *"on a dark night"* or *"on a beautiful house"*.

For reordering the sub-corpus, we found that no word type was changed by the reordering, showing that reordering just changes a sentence's order or its structure without making major alterations to the words or the semantics. This also happens in substitutions where a word is usually replaced by another word with the same meaning. Such target words are usually hypernyms or synonyms of the source word.

(3) Embedding-based operation characteristics

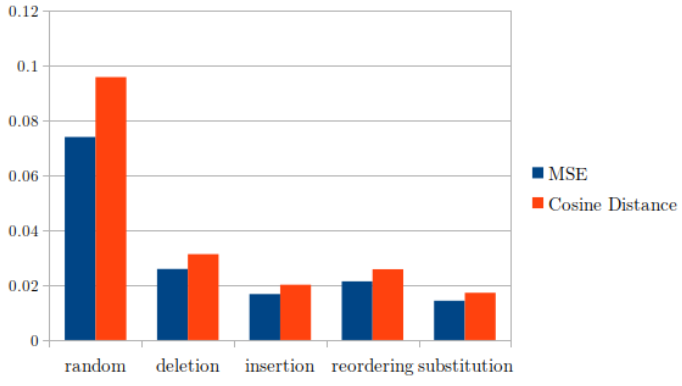

Fig. 3 Semantic embedding distance from original source sentence

In Fig. 3 we show the average distance between original source sentence embedding and the paraphrases (see column:

deletion, insertion, reordering, substitution) which are relatively similar. Compared with the distance to the random sentence (see column: random), our paraphrased sentence has much lower semantic embedding distance. This shows that the paraphrase operation maintained the consistency of the source sentence semantic meaning.

2.6 Semi-supervised Paraphrase Generation on Full WMT17 Dataset

To complete the paraphrasing on the full WMT17 dataset, we used 10k quintuplets of crowdsourced paraphrases and constructed a neural paraphrase models using four encoder-decoder, long short-term memory (LSTM) models with attention [18] for each paraphrase operation. We tuned and tested our automatic neural paraphrase model using these crowdsourced paraphrases of the development and testing datasets. With these four paraphrasing models, we generated VGPs on the remaining 19k image captions.

3. Improving MT with Visually Grounded Paraphrase

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

3.1 Two Scenarios of Data Usage

Given the created VGP corpus, there are two usage possibilities. First, the simplest form of data augmentation is to concatenate all the paraphrases with the original sentence. However, this simple approach suffers from disadvantages, because the relationships between the paraphrases and their original sentences are lost in the concatenation process with other quintuplets. Therefore, we also investigated the use of both multi-source and ensemble NMTs to preserve this relationship between paraphrases.

3.2 Proposed Approaches

In this section, we listed our proposed approaches to improve MT using our created VGP dataset.

3.2.1 Combining All Data in a Single Model

This method was done by just using the paraphrases as a means for data augmentation on the source side, such as reported by Nichols et al. (2010), to leverage SMT systems [19]. All of the paraphrases and their original sentence were combined, and the target sentence was duplicated by the number of multiple paraphrases. This approach measured the baseline performance with augmented data.

3.2.2 Multi-source Model

We implemented a multi-source NMT model proposed by Zoph and Knight (2016) to incorporate various paraphrase inputs with one output [20]. For their model, the encoded representation and attention were combined by concatenation. Zoph and Knight reported that this model has an advantage of information triangulation to reduce ambiguity. In their paper, they used several translation pairs such as {French, German} with English for which this language triplet shares a similar language structure. However, given these advantages, using this model as monolingual input has never been investigated.

3.2.3 Uniform-weighted Ensemble Model

For the uniform-weighted ensemble model, we trained NMT models whose source sentences were paraphrased based on the four elementary operations and another that uses the original source sentence to create five expert NMT models. Next, these five models were ensembled by averaging each output layer probability distribution, so that every model was weighted uniformly. We used this model to compare the performance with the mixture-of-experts model described in the next subsection, where each expert model has 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 a source sentence using the 56k dataset (the combination of the original and paraphrased source sentences). Five of these models are trained against the same target sentence. Each model is then regarded as an expert. Each of the expert models operates on the sub-word level, tokenized by Sentencepiece with 3000 vocabulary units[†].

3.2.4 Mixture-of-experts Model

Next we adopted the mixture-of-experts model proposed by Garmash and Monz (2016). Here, instead of using the linear layers proposed in their study [21], we implemented an expert model into a single LSTM layer hid that receives concatenated decoder hidden state output h_n:

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

We then applied a $softmax$ function 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:

$$W_{agg} = g_0W_0 + g_1W_1 + ... + g_nW_n.$$

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 whose output size matches the number of experts.

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

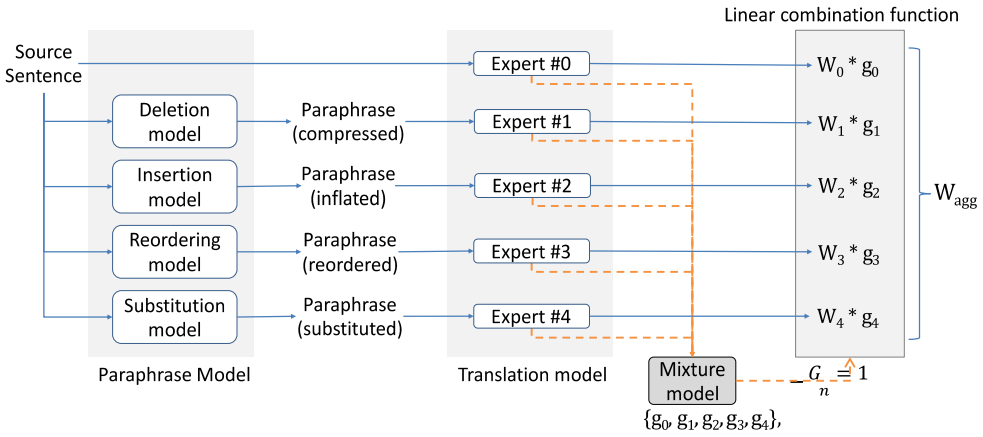

Fig. 4 Diagram of proposed mixture-of-experts neural caption translation model

A diagram of a mixture-of-experts neural caption translation model using our proposed approach is shown in Fig. 4. First, the source sentence is paraphrased into four different paraphrases used to train each of the expert model. Then, each expert passes its abstract decoding state into mixture model that will produce weights as the number of experts. The resulting weight distribution is a linear combination function between each expert's output probability distribution and the gating weight produced by the mixture model.

4. Experiments

The purpose of this experiment is to choose the approach that is the most suitable for our VGP by comparing the scores between Bahdanau et al.'s NMT baseline and several popular multi-source NMTs.

4.1 Setup and Dataset Composition

We combined the generated 19k dataset described in Section 2.6 with the original crowdsourced 10k training dataset. These 29k paraphrased datasets are combined with the original dataset, creating a 58k-triplet training dataset for each operation. The 29k paraphrased training dataset functions as a regularizer for the original dataset. These final data will be used to train a mixture-of-experts translation model, which will be described in the next section. The data will be publicly available to augment the WMT17 dataset.

Based on our empirical observations, using paraphrased data on the development and test 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 actual loss. Given that, we emphasize that using the paraphrased dataset in the translation step was done in the training steps in combination with the original dataset. In this stage, the paraphrases acted as a regularizer of the source sentences and a way of ensembling, improving the ensembled model's robustness as a whole.

Furthermore, for the experiment, we followed the training, development, and test set-up of the WMT17 shared task supplemented using our augmented training data. All results were scored using *multeval* [22] with lowercased and tokenized sentences. We used BLEU [23] and METEOR [24] as evaluation metrics.

Table 4 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

4.2 Model Specification and Implementation Details

We used single depth bidirectional LSTM with the size of 512 for all of our encoders in paraphrasing and translation model. For the decoder in all model, we used unidirectional LSTM with linear attention over encoded representation [18]. For neural paraphrasing, we trained four encoder-decoder model with attention for each elementary paraphrase operation. Note that although the architecture of the neural paraphrasing model is similar with NMT, the input and output are monolingual. On the other hand, in NMT, to receive five sentences as input, the multi-source NMT has five encoder and one decoder as a single model. At the same time, the multi-expert has five NMT model with one encoder and one decoder, combined with the combination layer.

The paraphrasing and translation model are optimized using Adam optimizer [25] with 1e-3 learning rate, while the combination layer of the multi-expert model are trained with RMSprop optimizer [26] with 1e-4 learning rate. We used unweighted softmax cross entropy loss for all of our models.

Table 5 Performance of proposed neural caption translation in comparison with 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
Combining 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 6 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 7 Examples of resulting sentences in mixture-of-experts model.

Translation Model	Type	Sentences	BLEU1
(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 .**	

In every increase of development loss, the learning rate is decayed by half with the exception of the first five epochs. We set a maximum of 7 decays for paraphrase models, and 5 decays in translation models for training early stopping. After the training stops, model with the lowest development loss was selected was selected and used for decoding with the beam size of five. All implementation was done with Chainer deep learning framework version 3.0 [27].

4.3 Evaluation of Neural Paraphrase Model

Table 4 lists the scores of the paraphrases produced with our automatic paraphrasing model. The substitution operation produced the highest BLEU score, and the reordering operation produced the lowest BLEU score. We expected this result because the reordering operation sometimes includes the changes of the active/passive properties of a sentence. Overall, we believe this score is high enough to paraphrase the remaining 19k WMT dataset.

4.4 Translation Model Results

Table 5 shows the performance of our proposed neural caption translation. All of the results using our VGP outperformed the NMT baseline. No improvements were gained from combining all of the data, which is the simplest form of data augmentation. This simple combination of data severs the relation that existed between each paraphrase that mentioned the same image. Furthermore, we cannot be sure that each source sentence has the same amount of paraphrases. By considering these factors, we utilized multi-source NMT and multi-expert NMT, which yielded better BLEU and METEOR scores.

This performance increase indicates that each expert model is slightly different between each other, and worked well in the uniform-weighted ensemble and mixture-of-experts scenario. This model also outperformed the uniform-weighted NMT in three cases. Moreover, the mixture-of-experts model performed better in out-of-domain ambiguous MSCOCO test dataset, suggesting that overfitting did not occur with such data augmentation. By applying to these several models, we can conclude that our elementary operation paraphrase is suitable to be used as a means for ensembling.

Table 6 shows the current submission systems in the official WMT17 shared task whose submissions consist of one textual model [28] and several multimodal models. Our proposed approach outperformed the baseline in WMT17 by a 13.2 BLEU score margin. Our proposed model, even though it is textual, produced competitive results with other multimodal models. The mixture-of-experts model outperformed several multimodal models, including another WMT submission [29]–[32]. Even in the out-of-domain dataset of COCO 2017, the mixture-of-experts model also performed reasonably well with a 28.0 BLEU score. Nevertheless, our score was almost the best score, proving that paraphrasing the source side also helped our model work with unseen data and prevented overfitting.

4.5 Discussion

To further analyze the contribution of the experts trained on the original data and on the paraphrased data, we compared the translation process step-by-step in our proposed approach. The source sentence shown in Table 7 was translated using each baseline model (an expert), resulting five different translation hypotheses. Each expert has been trained with slightly different paraphrased source sentence. We calculated the BLEU1 scores for each hypothesis against the target, resulting the source-reordered expert model yielded the best result between all experts.

The aim of the proposed mixture-of-experts model task is to ensure that the best part of each model is kept, as well as removing any noise or error that might occur in each model result. As seen from the comparison of the German result from the mixture-of-experts model with the target sentence, the only difference is the word "*am*" in which the correct one should be "*dem*".

In this example, in the deletion translation result, the word "*motorräder*" is decoded instead of "*motorradfahrer*". Another example is the phrase "*fluss entlang*" which is only found in the reordering translation result. Such quality of each expert model however, must be kept by the mixture model by distributing the correct word for every word being decoded. In conclusion, the final result of the ensemble of expert model combines the correct phrases in each expert model.

Quantitatively, the mixture-of-experts model successfully retained the good features of the best performing 0.87 and 0.95 BLEU1 scores that were yielded in the source-deleted and source-reordered model results, resulting in a 0.97 BLEU1 score. This is a significant improvement compared with the BLEU1 score of the uniform weighted model that only increased to 0.84.

5. Conclusions and Future Works

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

We successfully generated visually grounded paraphrase (VGP) sentences of the WMT17 Multimodal Translation Task dataset by crowdsourcing and our evaluation shows the effectiveness of our corpus creation method. We used this corpus to construct an automatic paraphrase generation model, and employed it within various multi-source and multi-expert approaches in NMT. In the future, we are interested on how the combinations of the elementary operations can also be employed in this various NMT methods.

From the usage perspective, our results indicate that our proposed paraphrase elementary operations are optimal for ensembling, especially with multi-expert ensembling settings. We proved our hypothesis that regularizing models by paraphrasing the source sentences is effective. In the future, we will further investigate various methods of incorporating visual information into NMT models. Furthermore, we would also investigate a large-scale paraphrase augmentation to enable ensembling in self-attentional NMT approach such as the Transformer model [33].

6. Acknowledgement

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

References

[1] G. Hirst, "Paraphrasing paraphrased," Invited talk at the ACL International Workshop on Paraphrasing, 2003.
[2] C. Chu, M. Otani, and Y. Nakashima, "iparaphrasing: Extracting visually grounded paraphrases via an image," Proceedings of the 27th International Conference on Computational Linguistics, pp.3479–3492, Association for Computational Linguistics, 2018.
[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3156–3164, 2015.
[4] Q. Wu, D. Teney, P. Wang, C. Shen, A.R. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," Computer Vision and Image Understanding, vol.163, pp.21–40, 2017.
[5] I. Calixto, Q. Liu, and N. Campbell, "Doubly-attentive decoder for multi-modal neural machine translation," ACL, 2017.
[6] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual english-german image descriptions," Proceedings of the 5th Workshop on Vision and Language, pp.70–74, 2016.
[7] 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," European conference on computer vision, pp.740–755, Springer, 2014.
[8] A. Prakash, S.A. Hasan, K. Lee, V. Datla, A. Qadir, J. Liu, and O. Farri, "Neural paraphrase generation with stacked residual lstm networks," Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, pp.2923–2934, The COLING 2016 Organizing Committee, December 2016.
[9] R. Bhagat and E. Hovy, "What is a paraphrase?," Computational Linguistics, vol.39, no.3, pp.463–472, 2013.
[10] G.D. Shudong Huang, David Graff, Multiple-Translation Chinese Corpus, Linguistic Data Consortium, University of Pennsylvania, 2002.
[11] M. Research, "Microsoft research paraphrase corpus," mar 2005.

[12] 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.
[13] 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.
[14] 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," Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, September 2017.
[15] C. Dyer, V. Chahuneau, and N.A. Smith, "A simple, fast, and effective reparameterization of IBM model 2," Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL '13), Atlanta, GA, USA, 2013.
[16] K. Toutanova, D. Klein, C.D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL '03), Stroudsburg, PA, USA, pp.173–180, 2003.
[17] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," CoRR, vol.abs/1409.0473, 2014.
[19] 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.
[20] B. Zoph and K. Knight, "Multi-source neural translation," CoRR, vol.abs/1601.00710, 2016.
[21] E. Garmash and C. Monz, "Ensemble learning for multi-source neural machine translation," COLING, 2016.
[22] A.L. Jonathan Clark, Chris Dyer and N. Smith, "Better hypothesis testing for statistical machine translation: Controlling for optimizer instability," Proceedings of the Association for Computational Lingustics, 2011.
[23] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu, "Bleu: a method for automatic evaluation of machine translation," pp.311–318, 2002.
[24] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," pp.65–72, 2005.
[25] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," CoRR, vol.abs/1412.6980, 2014.
[26] A. Graves, "Generating sequences with recurrent neural networks," CoRR, vol.abs/1308.0850, 2013.
[27] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a next-generation open source framework for deep learning," Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS), 2015.
[28] J. Zhang, M. Utiyama, E. Sumita, G. Neubig, and S. Nakamura, "Nict-naist system for wmt17 multimodal translation task," WMT, 2017.
[29] P.S. Madhyastha, J. Wang, and L. Specia, "Sheffield multimt: Using object posterior predictions for multimodal machine translation," Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, Copenhagen, Denmark, pp.470–476, Association for Computational Linguistics, September 2017.
[30] I. Calixto, K.D. Chowdhury, and Q. Liu, "Dcu system report on the wmt 2017 multi-modal machine translation task," Proceedings of the Conference of Machine Translation (WMT), 2017.
[31] M. Ma, D. Li, K. Zhao, and L. Huang, "Osu multimodal machine translation system report," Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, Copenhagen, Denmark, pp.465–469, Association for Computational Linguistics, September 2017.
[32] J. Helcl and J. Libovický, "CUNI system for the WMT17 multimodal translation task," CoRR, vol.abs/1707.04550, 2017.
[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30, ed. I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, pp.5998–6008, Curran Associates, Inc., 2017.

Johanes Effendi received his B.S. degree in Computer Science (cum laude) from Universitas Indonesia, Indonesia in 2015 and his M.Eng. degree in 2018 from Nara Institute of Science and Technology (NAIST), Japan. He is now a Ph.D. Student at Augmented Human Communication Laboratory, NAIST, Japan. He is a recipient of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) scholarship. His research interests include neural machine translation, paraphrasing, and natural language processing.

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005-2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003-2007), A-STAR and U-STAR (2006-2011). In 2009-2011, she served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia. In 2011-2017, she was an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She served also as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015-2016, under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation". Currently, she is a research associate professor at NAIST, as well as a research scientist at RIKEN, Center for Advanced Intelligent Project AIP, Japan. She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. She is also the officer of ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL) and a board Member of Spoken Language Technologies for Under-Resourced Languages (SLTU). Her research interests include statistical pattern recognition, graphical modeling framework, deep learning, multilingual speech recognition & synthesis, spoken language translation, affective dialog system, and cognitive communication.

Katsuhito Sudoh is an associate professor of Nara Institute of Science and Technology and a visiting researcher at RIKEN, Center for Advanced Intelligent Project. He received a bachelor's degree in engineering in 2000, and a master's and Ph.D. degree in informatics in 2002 and 2015, respectively, from Kyoto University. He was in NTT Communication Science Laboratories from 2002 to 2017. He currently works on machine translation and natural language processing. He is a member of the Association for Computational Linguistics (ACL), the Association of Natural Language Processing (ANLP), the Information Processing Society of Japan (IPSJ) and the Acoustical Society of Japan (ASJ)

Satoshi Nakamura is Professor of Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan, Project Leader of Tourism Information Analytics Team of RIKEN, Center for Advanced Intelligence Project AIP, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampolli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.