NAIST-IS-MT1651208

Master's Thesis

Neural Paraphrasing Framework to Leverage Machine Translation Quality

Johanes Effendi The

September 11, 2018

Department of Information Science
Graduate School of Information Science
Nara Institute of Science and Technology

A Master's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
MASTER of ENGINEERING

Johanes Effendi The

Thesis Committee:

Professor Satoshi Nakamura	(Supervisor)
Professor Yuji Matsumoto	(Co-supervisor)
Associate Professor Sakriani Sakti	(Co-supervisor)
Associate Professor Katsuhito Sudoh	(Co-supervisor)

Neural Paraphrasing Framework to Leverage Machine Translation Quality*

Johanes Effendi The

Abstract

Machine translation (MT) is a field that breaks language barrier by translating a sentence in source language into other sentence in target language. Several generation of MT has been developed since rule-based machine translation until neural machine translation (NMT). One of the challenges that is still relevant until now is to introduce additional knowledge, following human nature that translates something with prior knowledge as opposed to system. In this thesis, we investigate the paraphrasing as a means to introduce additional knowledge into NMT. Paraphrasing has been proven to improve translation quality in machine translation and has been widely studied alongside with the development of statistical machine translation (SMT). In this thesis, we investigate and utilize neural paraphrasing to improve translation quality in NMT, which has not yet been much explored. Our contribution are separated based on the occurrence of paraphrasing in either on both source and target side of a machine translation model.

Paraphrasing on the source side of NMT will augment the data, that can be used to widen the model context achieved through multi-source information. This information can also disambiguate uncertain input as it is known that MT is prone to ambiguity. For this source side, we develop a multi-paraphrase corpus semi-supervisedly by crowdsourcing the WMT 2017 Multimodal Translation Task. After

*Master's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1651208, September 11, 2018.

that, we use the corpus to do multi-expert neural caption translation. This proposed model is the alternative from the trend of using visual information as additional context which could only slightly contribute to performance based on recent results. Our proposed approach outperformed the baseline of WMT 2017 with 13.2 BLEU score margin, which is close to the top score that used a multimodal model.

On the other hand, paraphrasing with a robust model on the target side aims to improve translation quality by the means of automatic post-editing (APE). In this task, we develop a multi-source neural APE model which receives translation hypothesis and the source translation. We also introduce several decoding strategy to take advantages of the availability of translation hypothesis. Our proposed method successfully improved the baseline on NICT-QE/APE dataset by 1.8 TER margin.

Keywords:

paraphrasing, machine translation, automatic post-editing, visual description, image caption translation

Acknowledgements

I want to express my gratitude to Professor Satoshi Nakamura for welcoming me to his lab and supervised me for these 2 years. He introduced me to the scientific research environment in Japan through internship and recommended me for the MEXT-IPGP scholarship program, that makes me able to continue my study here in NAIST. His great insight and leadership are becoming my inspiration on how I strive to become a good researcher.

I would like to thank Research Associate Professor Sakriani Sakti for tirelessly supervising me through my research here in Japan. From her I learned a lot of research and academic skills such as the proper way to do research, academic writing, giving a good presentation, time management and attention to detail. Her working style and enthusiasm always inspires me, and I am grateful that I can still learn many things from her in my doctoral course.

I also want to thank Associate Professor Katsuhito Sudoh for always giving me helpful comments and suggestions from MT perspective, in every group meeting, paper submission, and this thesis. I am grateful to have supervisors from two differently related fields so that I can have various constructive perspectives on every concern.

I would also like to thank the thesis committee, Professor Yuji Matsumoto for his insightful comments in NAIST Seminar and especially for this thesis. Next, I would like to extend my gratitude for all faculty I have worked with in NAIST, for their support in every aspect of my research life.

I would like to thank the people in Augmented Human Communication Lab.,

especially for Speech Processing group, being the biggest and most diverse group in our lab, for becoming a great companion and delivering friendly environment during my master course. My special thanks to our lab Secretary, Manami Matsuda for always cheerfully helping me to solve a wide range of problems that occur during my stay as a foreign student in Japan.

I am particularly grateful for the assistance given by Senior Researcher Atsushi Fujita during my internship at the National Institute of Information and Communications Technology (NICT). With his guidance and support, I can improve my research skills and get invaluable best practices for research in MT.

Thank you also to Hashimoto-sensei, Inoue-sensei, and Motoi-sensei, for your help in teaching me Japanese. Thank you Yanagita-san and Kano-san for becoming my first friends here in Japan, and also helping me during the beginning of my life in Japan.

Finally, my family and friends that are always giving me support and encouragement, whose warmth can be felt from thousands of kilometers apart.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The idea to augment human-to-human communication through breaking language barrier has been pursued by many people since the philosopher and mathematician Rene Descartes proposed the idea of universal language in 17th century. In this society where communication doesn't have any barrier, knowledge sharing and intercultural cooperation can be smoothly realized. Machine translation (MT) is a field to fulfill this task to translate information between source and target language.

1.1. The Development of Machine Translation

1.1.1 From Rule-based into Neural-based Machine Translation

The first MT system was publicly demonstrated in 1951 by the MT research team of Georgetown University. This influential demonstration were then encourages the research of MT in other countries such as Japan, where the National Institute of Advanced Industrial Science and Technology (AIST) tested an English-Japanese MT system *Yamato*. This system was able to reached 90 point score on the textbook of 1st grade junior high school, showing the potential of realizing a robust translation

system.

Since then, a lot of MT system was developed by relying on a huge linguistic rules and dictionaries. This first approach to MT is called rule-based machine translation (RBMT). The most popular RBMT system is SYSTRAN[1], which is used widely until 2007. Another approach to MT are translation using interlingua [2, 3], where an intermediary universal language is being used for translating between language pair. Building this rule-based model and its variance requires extensive efforts on defining specific translation rules and dictionary explicitly. This problem has become the main problems for RBMT as solving ambiguity and adapting domain has proven to be difficult.

Furthermore, the first approach on using corpus to develop machine translation is example-based machine translation (EBMT) [4, 5, 6, 7]. EBMT receives a database of examples as input, and derive rules from this examples. There must be some close matches to the source text in the database in order to make this approach to be successful. On the other hand, the increase in computational power has contributed to the development of statistical machine translation (SMT) which recognizes language pattern from even bigger bilingual and monolingual data. Compared with RBMT, SMT has been proven to be effective by being able to recognize a language pattern automatically, and taking less time to built. In order to combine the advantages of RBMT and SMT, there has been some approach to statistically post-edit the result of RBMT [8, 9].

The rapid development of deep learning technologies also applied into MT domain, since it was done by Sutskever et al. that introduced encoder-decoder architecture to do neural machine translation (NMT) [10]. In International Workshop on Spoken Language Translation (IWSLT) 2015 MT shared task, NMT outperformed SMT systems on English-German pair by significant margin [11]. Bentivogli et al. reported that NMT works better on diverse lexical condition, makes less morphology and lexical errors, and better language modeling. Some problems still remains, such

[1]https://platform.systran.net/reference/translation

as difficulty in handling long sequences [12], wrong reordering on some phrase due to diverging alignment model, and prone to ambiguous phrases or terms.

1.1.2 Towards Better Context Understanding in MT

Up until now, building an effective and universal MT is still an open problem that needs to be solved. One of the critical problem remaining in MT is to improve the MT system ability to cope with the ambiguity of natural language. Several studies has shown that MT result quality can be improved by better word or phrase disambiguation [13, 14, 15]. One of the way to solve this problem is by introducing external knowledge, which improves MT system ability to understand context and generalizing across domains.

Many studies have attempted to incorporate additional knowledge to improve translation quality. The most straightforward approach to this is by augmenting lexicons through some data or computations that happened outside of the MT model itself. Zhang et al. introduces the ground truth translational equivalents as additional knowledge in an NMT system [16]. Arthur et al. incorporated discrete translation lexicons into an NMT system, and improved the performance on low-frequency words successfully and promoted faster convergence [17]. In semantic space, Shi et al. used knowledge-based semantic embedding to better translate domain specific electric business data which has a lot of specific terms [18].

There's also been some approach on introducing external knowledge through external corpus such as introduced first in SMT as monolingual corpus to increase the quality of language modelling. Waschle et al. introduced monolingual corpus as translation memory in SMT [19], which is also studied intensively in NMT [20, 21]. This external corpus is not always monolingual, but can be in form of symbolic translation rules in form of extensible markup language (XML), as introduced by Chatterjee et al. [22].

1.2. Introducing External Knowledge into MT using Paraphrasing

As explained in the previous section, introducing external knowledge is one of the approaches to realize a regularized and better context understanding in MT. These approaches, such as monolingual corpus, are introduced with the expectation that the model will find similar kind of sentences with the same meaning, which enables the better creation of more regularized representation, compared with the model that is only trained on one sentence for each meaning or sense. Monolingual corpus approach is indeed effective because it has the same text modality with the MT system, and introduces several senses similarity through introducing similar sentences with the same idea. However, we argue that this approach is not controllable as it is difficult to control the distribution of sentences in the monolingual collection of uncontrolled sentences.

We propose that the source sentence itself should be written in different linguistic variations such as vocabulary, style, and level of detail. We hypothesize that this approach will improves model comprehension of the sentence semantic, by seeing multiple paraphrased sentence examples describing a piece of information. Another advantages on using this approach is because the additional knowledge modalities is the same with the translation source and target modalities, which both are textual. Therefore, there is no need to do any effort to adapt the representation in a cross-modality manner.

1.2.1 Paraphrase Definition in MT

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, paraphrase is an approximate conceptual equivalence among outwardly different material [23].

Paraphrase definition in MT existing works usually used strict definition by emphasizing semantic equivalence such as substitution and reordering. For example, a

sentence pair below can be considered paraphrases because both of them have the same idea, regardless of the word substitution and phrase reordering that happened:

> A motorcycle is parked on the side of the road.
> On the side of the road, a motorbike is being parked.

The strict paraphrase definition in this idea is needed, because there is no tangible form of the sentence idea itself. One should rely on this strict definition in order to maintain the semantic consistency.

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 [24]. Furthermore, by taking further this idea, as long as the semantic of the mutual paraphrase sentence can be determined, we can widen the paraphrase definition to some extent.

In this research, we use image as a symbolic form of sentence idea, regarded as the basis of paraphrasing. We regard two sentences as paraphrase as long as both of it are talking about the same image.

Figure 1.1: Image used as the basis of paraphrases

> A motorcycle is parked on the side of the road.
> On the side of the road, a motorbike is being parked near the street sign.

For example, given an image of Fig. 1.1, the above sentence pair can be considered paraphrase because both of the sentence are describing about the same image as the

representation of the idea. This mean that the word or phrase insertion and deletion based on the same image as a concept may now be acceptable as one of paraphrase variations. Slightly different from the usual use case, this definition can be called image-based paraphrasing.

Another sentence rewriting task that aim to improve the sentence condition can also be seen as a monolingual translation between an erroneous source sentence and its fixed representation such as:

> It 's tell me the weather forecast for tomorrow.
> I want you to tell me the weather forecast for tomorrow.

In this example, the first sentence is being edited into the second sentence, with the target to improve the sentence quality. As both of the sentence is describing about the same semantics, we may allow correction as one of the operation in our paraphrasing task.

In summary, this study covers wide paraphrase definition not only substitution and reordering but also covers deletion, insertion, and correction. Although it might be quite different with common paraphrase definition usually being used in MT study, it enables us to investigate more diverse sentence rewriting problems within the same framework.

1.2.2 Paraphrase Application in MT

In many language generation task, paraphrasing plays a critical role for enrichment and added flexibility. For example, in question answering task, paraphrasing is needed to increase the similarity of the questions and retrieved documents [25]. In abstractive summarization, some paraphrasing is needed to transform the filtered sentences into a coherent shorter sentences. Cohn and Lapata leverages paraphrasing to enrich their abstractive summarization system and developed their abstractive summarization corpus [26].

In MT system, the paraphrase is often used for multi-reference evaluation [27]

since there might be a lot of valid translated sentences. Since the development of SMT, there has been a lot of approach of using paraphrasing to paraphrase the source language data. This approach has been concluded as a convenient way to handle out-of-vocabulary (OOV) and rare words problem [28]. Madnani and Dorr also showed that by using targeted paraphrases, unfair penalization of translation hypotheses can be avoided [29].

Paraphrasing is also used to augment the dataset size, which correlates positively with translation result in SMT [30, 31]. Some sentence pre-editing also can be made to make the source sentences easier to be translated [32, 33]. As opposed with doing paraphrasing in source side of MT, paraphrasing can also be done in target side of MT to leverage translation quality. One of the way is to make the translation becomes more natural by mono-lingually translate the translation hypothesis into better post-edited version, which task is called automatic post-editing [8, 9, 34].

With this recent studies in mind, we conclude that although paraphrasing has been intensively studied in SMT, we found that paraphrasing to enable multi-source information in NMT is not much investigated yet. Therefore, here we explore the use of paraphrases to make a mixture-of-expert and multi-source NMT to leverages multi-paraphrases input. Based on the occurrence of the paraphrasing, we separate our contributions into two main tasks: paraphrasing on source and target side as shown in Fig. 1.2.

1.2.3 Paraphrasing on source side: Visual Description Translation

Paraphrasing on source side is done on the source sentence such as shown on Fig. 1.2 (b). The objective of this paraphrasing is to elaborate the context contained in the original source sentence. Therefore, there will be smaller chance for ambiguity in the model, increasing the ability of the model to better grasp context.

To prove this hypothesis, we chose visual description translation task for the source side paraphrasing. In this task, a caption describing an image is being trans-

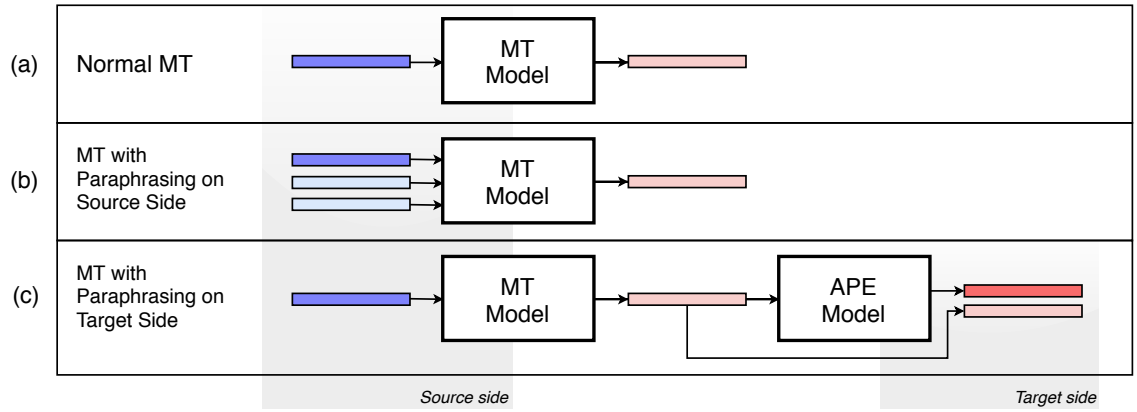

Figure 1.2: Paraphrasing on source and target side of NMT

lated from one language to another. Several straightforward approaches were utilizing the image latent representation directly into some NMT components such as most approaches have done in WMT 2017 Multimodal Translation shared task. However, to the best of our knowledge, there hasn't been any significant improvement yet on using that approach, especially given the size of the small dataset.

For this task, we take on different approach by paraphrasing the source caption, given the image itself, which in the previous section we described as image-based paraphrasing. Textual information is easier to be processed in NMT model as opposed to image latent representation. The additional information that could be helpful to disambiguate or widen the model context could be taken from the paraphrase itself.

1.2.4 Paraphrasing on target side: Automatic Post-editing

MT system in real-life situation has been used in several industries to translate documents from many language pairs. The result of this system however, still suffer from systematic errors such as unnatural choice of words, wrong sentence structure, and so on. The task of fixing this translation hypothesis is named post-editing. Usually post-editing is done by human editor with the principle of doing minimum

edit possible in order to make the sentence grammatically and semantically plausible with respect to the source sentence.

From the both paraphrasing and translation perspective, APE task is regarded as a paraphrasing task in the form of monolingual translation from MT translation result into a newly post-edited. The paraphrasing objective itself is to improve the sentence structure, while keeping the semantic plausibility. In order to achieve that, a further high level understanding of the target language modeling needs to be achieved, compared with the MT task.

As shown in Fig. 1.2 (c), we propose a neural APE model that fixed the translation result from SMT system. Represented by the same red color in the figure, the APE model can be seen as a paraphrasing model, which fixes systematic error found in MT hypothesis. We compare that by taking source information as the translation context, the post-editing task can produce better result.

1.3. Thesis objective and contribution

As discussed before, incorporating external knowledge has been a challenge in any generation of MT. From many ways incorporating this information, in this thesis, we focus on using paraphrase as this form of additional information. Taking a look at section 1.2.3 and 1.2.4, this thesis proposes approach to utilize paraphrasing advantages on both source and target sides of an NMT system.

We summarize this thesis objective into three main contributions as follows:

- Developing multi-paraphrase of WMT 2017 multimodal translation shared task dataset in a semi-supervised manner by partially crowdsourcing. This dataset were then used to train a neural paraphrasing model.

- Introducing multi-expert visual description translation through multi-paraphrase as multi source input. Our proposed approach improved the baseline of WMT 2017 by 13.2 BLEU scores, comparable with multimodal model.

- Improving MT result by means of automatic post-editing on NICT-QE/APE dataset by using multi-source model that paraphrase the translation hypothesis in regard to source sentence.

In conclusion, this thesis aims to investigate and utilize the use of paraphrasing to leverage the quality of MT system. We discuss the background knowledge of NMT on Chapter 2, in which our proposed model was based on. Chapter 3 elaborates our paraphrase incorporation strategies using our proposed model on both tasks. After that, we report and discuss the result of our experiment on visual description translation in Chapter 4. Furthermore, in Chapter 5, we report and analyze the result of our proposed method on automatic post-editing task. Finally, in Chapter 6, we conclude our contribution in this thesis and elaborate some possible advancement in the future works.

Chapter 2

Neural Machine Translation

This chapter discuss about NMT technology which is used in this thesis.

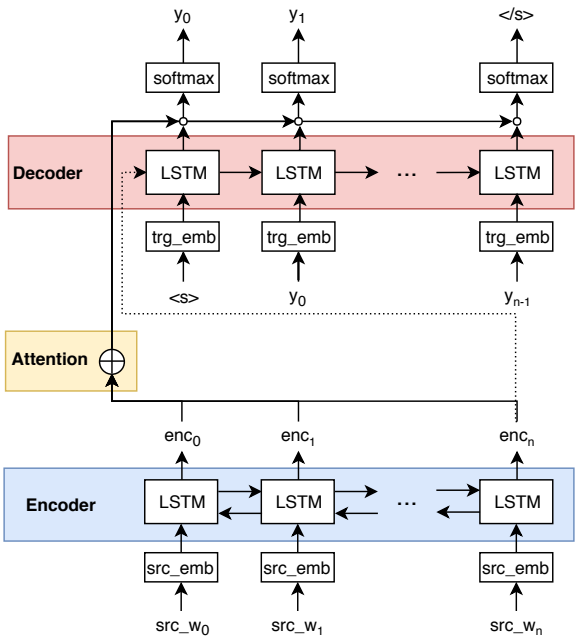

Figure 2.1: Encoder Decoder LSTM

2.1. LSTM Encoder Decoder with Attention

There are several modifications done to neural network cell in order to be able to process sequential data better. A recurrent neural network (RNN) cell was done to enable the use of previous element, compared with normal perceptron that can doesn't regard time order into decision process.

However, RNN cell still suffers from a problem called vanishing gradient when handling long sequences. The main reason is because RNN doesn't have memory to remember what is happening in several previous time-steps in the sequence. To tackle the problem, a long short term memory (LSTM) cell was developed. LSTM cell has the ability to add or remove information to the cell, controlled by the gates architecture [35]. This enables LSTM to handle long sequences, especially a long term dependencies in a sequences.

Another problem that rises from processing a sequence is the variability of the sequence length. A sequential data such as sentence can consists by many number of words, requiring a certain mechanism to encode this variable length sequence into one latent representation. To address this length variability on both input and output, Sutskever et al. developed an encoder-decoder architecture to encode a variable length sequence, and decode it into another target sequence [10]. Improved with attention mechanism [36], this approach has been proven to be successful to outperform SMT as the previous standard of MT.

In Fig. 2.1, step-by-step computation of each encoding and decoding step are shown in an MT task. The source sentence in source language consisting of several words $src\_w_{0:n}$ are embedded using src_emb embedding layer for source language, which output are then encoded by a bidirectional LSTM. The encoded representation enc_n which represents the sentence are then used to reset the LSTM decoder.

The decoder are fed with embedded target language token of sentence beginning $< s >$, which the hidden representation are then combined with attention before inputted into output layer. In output layer, the probability distribution of the next word are then normalized by a softmax function. The selected word with the max-

imum probability are then used to initialize the decoder for the next step. The decoding stop when the decoder produces an end-of-sentence token $< /s >$.

2.2. Beam Reranking Approach for NMT Decoding

In the decoding process, the word with the highest softmax probability in the output layer is chosen as the decoded word. After that, the next word decoding use this selected word for the basis of the next translation. From the beam search problem perspective, the way of only using the word with the highest probability is similar with greedy search. Better translation candidate might be produced if the beam decoding process also consider the word with second or third highest probability score. As the decoding process work from left to right, better word or sequence of word candidate can be seen the sub-optimal word probability is also considered.

If we consider the top N word with the highest probability, in every decoding step t there will be N^t translation candidate, in which the decoding will not be efficient in terms of memory. In contrast, the NMT decoding is defining beam size as the size of the pool, on which each translation candidate are listed. With the beam size of N, for each time step, there will be fixed number of N translation candidate in the pool. If a translation candidate has reached the end-of-sentence token, the size of the beam size N will be reduced by one. The decoding will be stopped when the beam size is zero. After that, the candidates with the highest score will be selected.

There are various modification to improve the beam search. One of them is by reranking with language model. The reranking process is done by getting the language model score for each translation candidate and combine the score with model probability. This scoring can be done in all steps of decoding, or just on the last step to help the algorithm choose the final candidate.

2.3. Unsupervised Text Tokenizer to Discover Subword Units

The word-by-word generation of target sentence in decoding process are aimed to maximize the likelihood of the output layer distribution compared to a discrete translation target word distribution. The output layer itself, has a fixed size, where each cell represents a word in target language vocabulary. This methodology poses a problem when the model translates a rare word, resulting an unknown token $< unk >$ which reduces the evaluation score. To tackle this problem, the subword NMT has been developed based on the intuition that various word classes are translatable using smaller units than words

For example, there exists German word "spielt" in training data, and word "spielen" in testing data. Both of these word has the same meaning (to play), but with different ending based on the subject of the sentence. If the system is using word-by-word tokenization, the system will not recognize the word "spielen", and return the token $< unk >$ which later will reduce the evaluation quality.

Let's assume there are other word "tanzen" in training data, which has the same ending "en" with the word "spielen" in testing data. If each word in training data are tokenized into subword such as "spiel" + "t" and "tanz" + "en", then the system can recognize the word "spielen" in testing data, just by combining subwords "spiel" and "en".

The first approaches on tokenizing this text automatically is by using Byte Pair Encoding (BPE) to discover subword unit [37]. The BPE itself is a data compression technique uses frequency to do sequential byte replacement [38]. The implementation for BPE are further improved by Google Sentence Piece which is reversibly convertible by preserving space as a special meta symbol. For example, the phrase `Hello_World.` will be tokenized as `[Hello] [_Wor] [ld] [.]`.

Chapter 3

Leveraging Machine Translation Quality using Neural Paraphrasing Framework

In this chapter, we discuss our proposed architecture to leverage MT quality using neural paraphrasing on both source and target side of NMT.

3.1. Determining integration point

To deal with multiple inputs, there are various integration strategies can be done to combine these inputs feature. This combination can be seen as combining several systems in which each systems has several potential integration point. The point where the integration possibly conducted can be made clear if we divide the flow of information in an NMT system described in Chapter 2 into four phases. We describe these four phases and each challenges in this list as follows:

1. **Integration in feature space**
   Integration in the feature space involves doing feature combination. This is the simplest form of doing data augmentation. However, given the sequential

nature of the source sentence, if we take words as feature, it is difficult to determine the importance of each features. This is also not the direction that we want, because the reason of using paraphrasing itself is to augment the data and widen the model search space.

2. **Integration in encoding phase**
   The purpose of the encoding in NMT is to represent a sequence of words with variable length into a single vector representation. Integrating multiple inputs in encoding phase poses a new challenge on how to develop encoder that could receive variable length but decode at once. Given the reordering that happens in the input, the solution's implementation of this problem involves a re-aligning operation. This re-aligning operation is an operation that cannot be done in batch, because each sentence has different alignment with its counterpart. Therefore, this solution is not suitable for neural network because batching is needed to improve the speed and stabilizes the training process.

3. **Integration in decoding phase**
   The combination in decoding phase involves the combination of attention from each input. By this method, the model will decode with the help of several attentions as prior knowledge. This approach is more suitable for several input that has different word reordering. The initialization of the decoder can also become the integration point, using combined encoded representation.

4. **Integration in result space / ensembling**
   Integrating in result space consists of two possible approaches: expert ensembling and hypothesis reranking. Expert ensembling approach is using a learned weight combination to combine the output of each NMT model, given the decoder hidden state. Garmash and Monz reported that this method is effective for small sized dataset [39]. On the other hand, hypothesis reranking is using external model to rerank several hypothesis that can possibly produced by several model.

Based on these listed potential integration point, we choose to integrate our model in decoding phase and result phase. The integration in decoding phase is implemented as a modified multi-source NMT [40], and the integration on result space is implemented in a form of multi-expert ensembling [39].

3.2. Paraphrase Incorporation Strategies

This section is describing our general framework which is used on two cases such as written in Section 3.3 and Section 3.4.

3.2.1 Various multi-source combination strategies

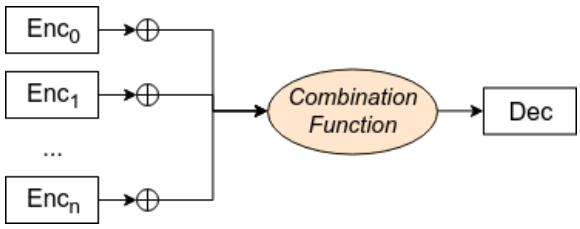

Figure 3.1: Multi-source NMT

This model is a modification of Zoph and Knight's multi source neural translation [40]. They claimed that this model is taking advantage of information triangulation to reduce ambiguity. In their paper, they reported several translation pair such as {French, German} to English in which this triplets of language has similar language structure. However, given this advantages, the use of this model to monolingual input has never been investigated. Furthermore, different set of combination function might perform better given the different language characteristics.

For visual description translation task, we modify this existing model to be able to receive more than two input. On the other hand, for APE task, we hypothesize that the use of source information alongside with translation hypothesis will give better

context to the APE model itself. For both task, we also use various combination functions to investigate which combination function is suitable for each task.

We implement the joint attention mechanism by using various combination function such as:

1. **Linear layer concatenation**

   This method concatenates all attentions $att_{0:n}$ using a single linear layer that receives the concatenation of N number of hidden representation and output one combined attention att_{agg} after a $tanh$ non-linearity.

   $$att_{agg} = tanh(W.[att_0, att_1, ..., att_n] + b)$$

2. **Summation**

   The combination of attention were done by summing the available attentions.

   $$att_{agg} = \sum_{i=0}^{n} att_i$$

3. **Weighted summation**

   For this summation strategy, each of the attention were weighted before being summed. The weights $W_{0:n}$ itself is predicted using a linear layer from decoder hidden state dec_{hid} to the number of attentions with $softmax$ non-linearity.

   $$W_{0:n} = softmax(W.dec_{hid} + b)$$
   $$att_{agg} = \sum_{i=0}^{n} W_i.att_i$$

4. **Max or average pooling**

   This combination were done by using a max-pooling or average with the window size of N, run through the second order axis of the attention list as shown on Fig. 3.2.

These combination functions can be applied to combine encoded representation (f_{enc}) and attention (f_{att}).

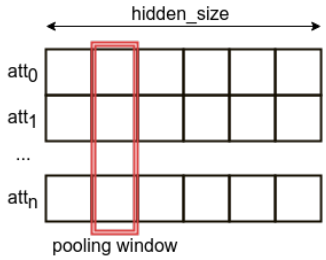

Figure 3.2: Pooling on concatenated attention

3.2.2 Multi-expert ensembling

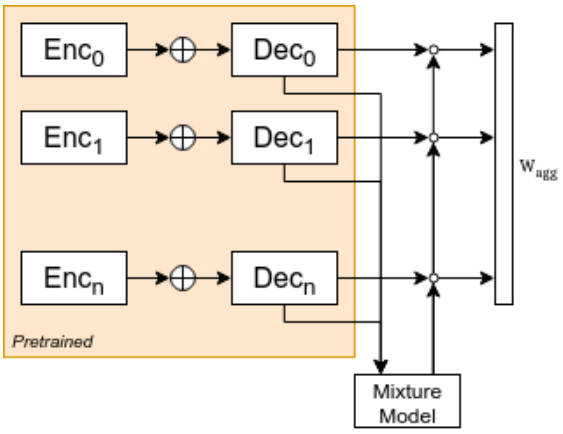

Figure 3.3: Multi-expert NMT

An NMT model decoder determines which word should be chosen for the next output by choosing the highest probability word in the output layer. Thus, the weighting of the output layer probability distributions becomes a crucial step in determining the quality of the output. Common practice in realizing this idea is to linearly interpolate them together by putting uniform weights to each model output probability distributions. However, this might not be ideal because each of the ensembled models might not contribute uniformly depending on the input it is given.

To give proper weighting for each model output probability distributions, some learning named mixture-of-experts needs to be done to identify systematic errors

created with each model[41]. Garmash and Monz showed that by letting the expert model to learn from decoder hidden state output as an abstract decoding state which contains context and prediction information for the next word, the expert model's weight distribution can perform better than those of the models with uniform weight distribution [39]. In addition, generalization can be done by creating a very large neural network with a sparsely-gated mixture-of-experts layer [42].

Commonly, the mixture-of-expert method is used with several NMT models that has been initialized with different random seeds, or several NMT models that have different source language, but with the same target language. In this study, we used the method on the several NMT models that have been trained on paraphrased source sentences which has the same meanings, and translated into the same target language.

Determining the weight for each output layer probability distribution were done as follows:

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

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 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:

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

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.

3.3. The Use of Paraphrasing for Visual Description Translation

In this sub-section, we described our approach on incorporating additional knowledge by the way of paraphrasing on the source side of NMT, such as shown in Fig. 3.4. Such approach is implemented in a task named visual description translation.

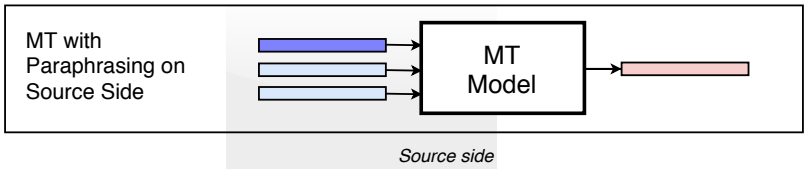

Figure 3.4: Leveraging NMT with paraphrasing on source side

Visual description research is useful for many applications such as image searching with natural language query, image description for visually impaired users, or understanding slides in a lecture. But, most of the activities were done focusing only on a monolingual English setting [43]. Recently, with the availability of multilingual image caption dataset such as Multi30k, there has been a lot of attempt in expanding the research to generate visual descriptions in various languages using MT, such as shown in Fig. 3.5. WMT 2017 "Multimodal Machine Translation" is one of the shared tasks that facilitate these activities.

Most approaches focus on utilizing image features in addition to the information from a single caption of the source language. Using NMT approaches, combining multi-modalities in NMT become straightforward because of the intermediary hidden vector representation is assumed to be compatible with each other. However, as pointed by Calixto et al. [44] and looking at the results of WMT 2017 Multimodal shared task, we argue that this image-text latent representation combination approach has not yield significant improvement on WMT 2017 Multimodal shared task dataset testing.

The main source of sentence translation is the source sentence itself. The transla-

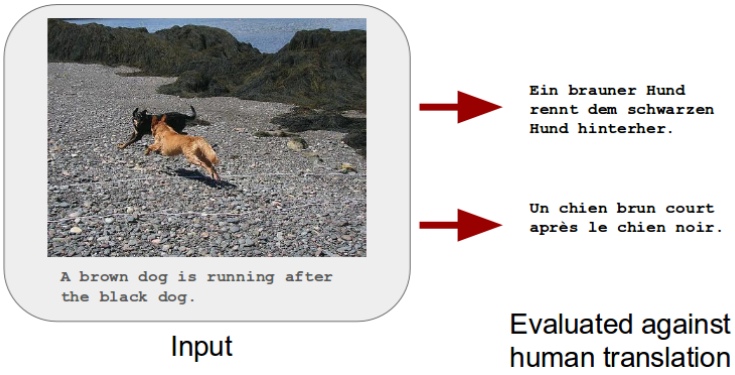

Figure 3.5: Visual Description Translation workflow

tion model doesn't always need the image information, especially when the sentence itself is already clear and non-ambiguous. Therefore, we hypothesized that the image itself is not critical, but the source sentence itself should be improved according to the image. One of the way to improve the source sentence is by using paraphrasing as it is successfully proven to improve MT quality described in previous section.

Moreover, we also argue that a single caption might not be able to represent all the information presented on an image. Therefore, we attempt another direction in which we use additional information by way of multi-paraphrase captions of the source language. In this attempt, we use image as the basis of paraphrasing. The resulting paraphrase captions are then utilized within a multi-source and multi-expert NMT model.

3.3.1 Related works

Many studies have attempted to incorporate additional knowledge in multi-source manner to improve translation quality. Multi-source neural translation is one method to enrich the model knowledge and reduce ambiguity. [45] stated that multi-source translation is expected to have better word sense disambiguation, better word reordering, and reduce the use of anaphora resolution. Several multi-source NMT system have been developed following this preliminary research on SMT [46, 47].

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. Most approaches use visual information as additional context by directly embedding image features and combining them with the source-language description within a multimodal system. The best system in this shared task was developed by modulating each target embedding with global or convolutional image features [48].

However, the results from most submitted systems reveal that the additional image features could only slightly contribute to system performance [48, 49]. There were even negative results in the study by [50], where they found that their textual model performs better than their multimodal model. They suggested that their multimodal model could eventually outperform the textual one if there is sufficient amount of data [50]. Under this condition, it seems that there is no certain method yet on how to incorporate image features directly into a translation system.

In this study, we go in another direction in which we use various visual descriptions by way of paraphrasing to describe the image. The resulting paraphrase captions are then utilized within a multi-expert NMT model. This enables us to incorporate additional knowledge from image to the translation process, without using the image itself. Paraphrasing has been shown to be effective to improve coverage of image aspects and translation quality in MT model [33].

3.3.2 Proposed approach

For this task, we use both the multi-source and multi-expert ensembling proposed model with five input following our paraphrase elementary operations: original, deletion, insertion, reordering, substitution.

In multi-source proposed model as shown in Fig 3.6, there are five bidirectional LSTM encoder, in which the latent output is combined with various combination function. The combined latent representation and attentions are then used for decoding the visual description in target language.

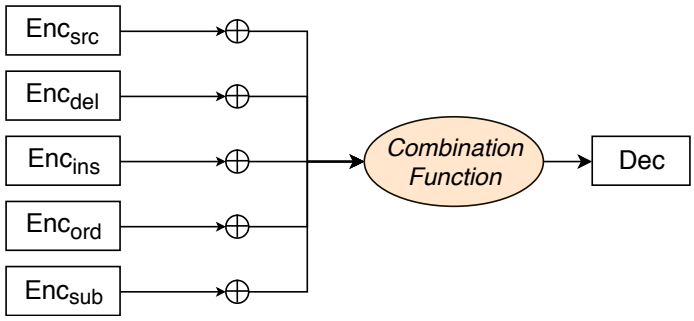

Figure 3.6: Visual description translation proposed multi-source model

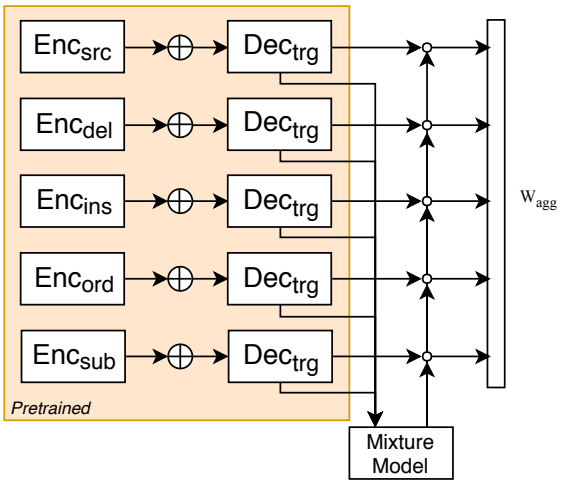

Figure 3.7: Visual description translation proposed multi-expert model

On the contrary, in proposed multi-expert model, there will be five encoder-decoder LSTM NMT model which has been pretrained to each paraphrased source sentence (Fig. 3.7). The mixture model will take the hidden state of each decoder, and use it to output weight for each output layer. By this approach, we hypothesize that each paraphrased source will trigger the NMT model to also produce slightly different results. These slightly different results are then combined by the mixture model.

3.4. The Use of Paraphrasing for Automatic Post-editing

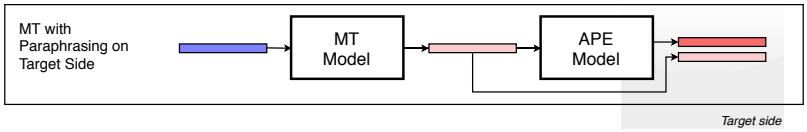

Figure 3.8: Leveraging NMT with paraphrasing on target side

In this paraphrasing on the target side, the APE task itself is the task of paraphrasing. The objective of this task is to fix some systematic errors produced in MT step.

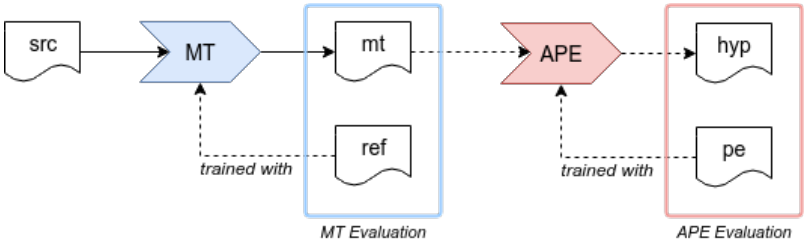

Figure 3.9: Automatic Post-editing workflow

As seen in Fig. 3.9, MT model and APE model are trained on different dataset. MT model is trained on a pair of source (src) and reference (ref) bilingual pair of sentences resulting a translation hypothesis (mt). Meanwhile, APE model should receive the translation hypothesis of the previous MT model (mt), and trained it with human post-edited version of that hypothesis (pe). The evaluation score of an MT process is calculated from $\{mt, ref\}$ pair, while APE evaluation should be calculated from $\{hyp, pe\}$ pair.

3.4.1 Related works

The development of APE system in earlier days was started by Dugast et al. [51]. that developed a statistical post-editing (SPE) system for a rule-based machine translation (RBMT) system. An RBMT system is defined by a set of translation rules. However, these static rules are of course cannot cover all translation rules perfectly, and therefore the result of it has a systematic error. Therefore, the task of APE system is to learn how to repair this systematic error automatically. This concept was further generalized for other MT system implementation such as statistical machine translation (SMT). On the next development, some system that does APE to SMT result also developed such as the one developed by Simard et al. [52].

Furthermore, a research by Bechara et al. [53] introduced Contextual SPE, where source information is introduced to SPE model by word alignment. This strategy is proved to improved the result of SPE system since it can help the system to understand the context of given MT translation result.

As mentioned in Section 2.1, the use of LSTM Encoder-Decoder has become the popular way of doing machine translation and some of its auxiliary component. This recent improvement has also been introduced to APE task by Pal et al. by developing neural network based automatic post-editing (NNAPE). However, post-editing task needs the model to understand the target language in more fine-grained way. In order to reach this model convergence, a neural network model need a large amount of data. Junczys-Dowmunt and Grundkiewicz introduced a NNAPE model consisting of log-linear combinations of several model pair, trained using a large amount of artificial monolingual dataset [54].

This studies mostly relied on sequence-to-sequence model either from the source (src) to target (pe) or mt-hypothesis (mt) to target (pe). It has been proven that the source information is useful for post-editing task, especially if it is used together in a model. A recent research uses multi-source NMT to combine weighted attentions between the source and mt-hypothesis attention. We argue that this study's attention combination process is arbitrary, and there should be more various attention

combination strategies being studied.

This recent studies also just focused on WMT 2016-2017 Automatic Post Editing shared task dataset which only investigates post-editing in English-German language pair which is quite similar in terms of language structure. This is not the case in a language pair like Japanese-English where both languages has different language structure, which makes the combination of attentions is more difficult.

3.4.2 Proposed approach

For this task, we use both the multi-source and multi-expert ensembling proposed model with two multilingual inputs which consist of source sentence (src) and MT hypothesis (mt).

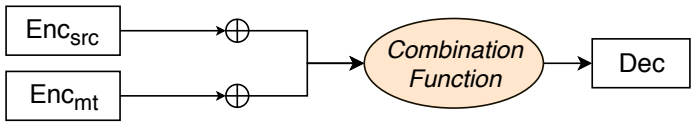

Figure 3.10: Automatic post-editing proposed multi-source model

As shown in Fig. 3.10, combination function will combine each of the encoded representations and attentions from src and mt encoder. By this approach, we also investigate which combination function is useful for combining two different kind of attention of Japanese-English translation.

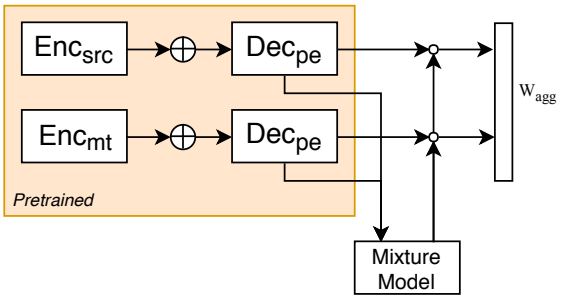

Figure 3.11: Automatic post-editing proposed multi-expert model

In previous research Juncsys-Dowmunt and Grundkiewicz used log-linear combinations to join several $src - pe$ and $mt - pe$ NMT model into one APE model [54]. However, they did it on English-German dataset, which has similar language structure. In this thesis, we want to compare which APE approach is more suitable for far-distanced language pair such as English-Japanese: multi-source NMT or multi-expert NMT.

Chapter 4

Experiment on Visual Description Translation

This section discuss about the result and evaluation from the task defined in Section 3.3.

4.1. Defining Paraphrase Elementary Operation

The definition of paraphrase accepted by the linguist are broader than its strict semantic equivalence in its logical definition counterparts. This relaxed definition are created due to the nature of paraphrasing itself. For example, there are a lot of paraphrase variations that can be made from a sentence. This huge amount of variations needs to be stated in order to define what kind of variations specifically done on a sentence.

Bhagat and Hovy's research on paraphrase definition and the classification of quasi-paraphrases [1] specified 25 quasi-paraphrases based on lexical perspectives (Table 4.1). Each quasi-paraphrase class has its own way of implementing the semantic equivalence standards of a paraphrase. By this classification, they also proved that although approximate equivalence is hard to characterize, paraphrasing opera-

tion is not a completely unstructured phenomenon.

Table 4.1: Bhagat and Hovy's Study of 25 Quasi Paraphrases in MTC and MSR Corpora [1]

#	**Category**	**% MTC**	**% MSR**
1	Synonym substitution	37	19
2	Antonym substitution	0	0
3	Converse substitution	1	0
4	Change of voice	1	1
5	Change of person	0	1
6	Pronoun/Co-referent substitution	1	1
7	Repetition/Ellipsis	4	4
8	Function word variations	37	30
9	Actor/Action Substitution	0	0
10	Verb/Semantic-role noun substitution	1	0
11	Manipulator/Device substitution	0	0
12	General/Specific substitution	4	3
13	Metaphor substitution	0	1
14	Part/Whole substitution	0	0
15	Verb/Noun conversion	2	3
16	Verb/Adjective conversion	1	0
17	Verb/Adverb conversion	0	0
18	Noun/Adjective conversion	0	0
19	Verb-preposition/Noun substitution	0	0
20	Change of tense	4	1
21	Change of aspect	1	0
22	Change of modality	1	0
23	Semantic implication	1	4
24	Approximate numerical equivalences	0	2
25	External knowledge	6	32

As shown in Table 4.1, many quasi-paraphrases have very small frequency in the Multiple Translation Corpus (MTC) and Microsoft Research Paraphrase (MSRP) corpora. We argue that these 25 classes is too fine-grained and can actually be grouped into fewer classes to reduce ambiguity and makes further usage of this paraphrase group simpler. We grouped them into these following four operations: deletion, insertion, reordering, and substitution. We slightly modified Snover et al.'s Translation Edit Rate (TER) strict definition of sentence building components into these four elementary paraphrase operations [55].

4.2. Experimental Settings

4.2.1 Baseline Model

The baseline given by WMT was produced with Moses SMT toolkit [56] for 2016 baseline. For 2017 baseline, it was produced by text-only NMT system built with Nematus toolkit [57] with hyperparameters were kept as default. The NMT model was using batch size of 40, Adam optimizer, and BPE compression algorithm to segment word [37]. We also evaluate one of our baseline NMT implementation to be compared with other models. This baseline NMT implementation will only be trained on original data to enable the comparison of translation without paraphrasing.

4.2.2 Proposed Model

The experiment in this task is divided into two step. Firstly, we train four encoder-decoder LSTM with attention using crowdsourced paraphrased dataset in order to generate the remaining paraphrase in semi-supervised manner. Each model has bidirectional encoder and dot-product attentional decoder with single depth, 50% dropout ratio, and 512 hidden layer size. Implementation was done using Chainer framework version 4.0 [58] and ran on GTX Titan X GPU. For optimizer, we use Adam [59] with decaying alpha into half every development loss increase with maximum 7 decay for early stopping. After stopping training, model with the lowest development loss will be selected and be used for decoding with beam size of 10.

Second, for translation step, we tokenize both the source and target of the paraphrased dataset using Google Sentence Piece [1] trained on 3000 unit vocabulary.

In this step, we use the multi-source and multi-expert NMT for each pair of original, deletion, insertion, reordering, and substitution to German target language. In multi-source NMT proposed model, this five input will require five encoders and one decoder, while in multi-expert NMT proposed model, five different sequence-to-sequence encoder-decoder LSTM model are used for each elementary operation.

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

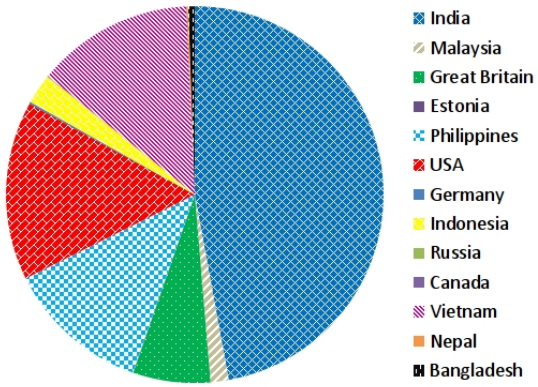

Figure 4.1: Country distribution of crowdworkers

We use 256 hidden layer size for each model, with single depth for each bidirectional-encoder and decoder. We did regularization using dropout with 0.2 ratio for embedding layer, and 0.3 ratio for other layer.

4.3. Dataset Construction

4.3.1 Crowdsourcing multi-paraphrase corpus based on elementary operations

We conducted a preliminary study to know how to build a multi-paraphrase corpus based on our proposed elementary operations. We collected 10k images from Visual Genome [60], which consists of such diverse objects as animals, people, and vehicles, and asked the crowdworkers to caption these images and create paraphrases of the captions into four different paraphrases based on each operation. The corpus consists of 5-pairs of paraphrases with 50 k sentences.

We used Crowdflower[2] as a crowdsourcing platform. For one session, each crowdworker had to caption and paraphrase at least two images and also be available for

[2]CrowdFlower – https://www.crowdflower.com/

additional sessions of image annotation. We limited this task to English-speaking countries or those countries where English is the second language. We also monitored the crowdworker results. If the resulting sentences were not valid paraphrases, we discarded them from the data.

We successfully gathered more than 200 workers from 13 countries for this corpus creation. Most workers came from India, the United States, and Philippines. The remaining countries and distribution are shown in Fig. 4.1. Each worker created an average of 52.56 images took an average of 7 minutes 54 seconds to caption an image and write its four paraphrases.

However, we discovered that our preliminary study dataset have shorter caption compared with WMT 2017 dataset, This is because we let the crowdworker to create their own caption, in which we found that unconstrained captioning by crowdsourcing will cause the result tends to be shorter. We also found that the vocabulary in this crowdsourced corpus is quite different from the one used in WMT 2017 dataset. We calculated about 50.6% sentences in WMT 2017 training dataset has a word which does not belong to this crowdsourced dataset.

After that, we measured each operation characteristics with quantitative metrics. 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 number of words in the target sentences decreased by 19.53%, where an average of 2.23 words was deleted per sentence (Table 4.2). We did the same calculation in the insertion sub-corpus and found that the number of words in the target sentences increased by 25.15%, where an average of 2.93 words were inserted per sentence.

Table 4.2: Operation characteristics

Parameter	**# word**
Avg. deleted words per sentence	2.2286
Avg. inserted words per sentence	2.9311
Avg. distance of reordered word	4.4252
Avg. word substitutions per sentence	1.6770

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 4.42 words. The distance calculated in the reordering happened when the source sentence was paraphrased into its passive form, or to exchange the order of such sentence information as time, place, and tool. As seen in Table 4.2, we found an average of 1.68 word substitutions per sentence in the substitution sub-corpus. This means that at most 1 or 2 words were substituted in a sentence.

Furthermore, 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. This evaluation identifies which word type was usually preferred by the crowdworkers.

Table 4.3: Top 5 Most Operated POS Tags

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

Table 4.3 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. 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 is changed by reordering. This shows that reordering just changes the order or the structure of the sentence without making major alterations to the words or the semantics. This also happens in substitution 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.

By reflecting from this preliminary study, we are paraphrasing the WMT 2017 dataset by asking the crowdworker to paraphrase the English source caption from

Figure 4.2: Reference image for captioning and paraphrasing shown in Table 4.4.

WMT dataset, instead of letting them do captioning by themselves. We keep the data collection settings such as elementary operations definition, countries, platform, and worker criteria as it has been found effective in this preliminary study.

4.3.2 Paraphrasing the WMT 2017 multimodal translation shared task dataset

The WMT17 Multimodal Translation Task dataset [43] 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 caption in English [61] 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. All captions are preprocessed by lowercasing, punctuation normalization, and tokenization. An out-of-domain dataset consisting 461 images taken from the Microsoft Common Objects in Context (MSCOCO) dataset [62] was also introduced, the captions of which contain ambiguous verbs [63].

We focused on paraphrasing the English sentences which is considered source language. Table 4.4 shows example of a paraphrased image caption based on four elementary operations (deletion, insertion, reordering, and substitution) and Fig. 4.2 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

Table 4.4: 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.

and time, we crowdsourced only 10k triplets of this dataset (10k training dataset), along with the whole development and testing datasets.

We used Crowdflower[3] (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. 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.

Finally, to complete the paraphrasing on the full WMT17 dataset, we then used 10k quintuplets of crowdsourced paraphrases and constructed four encoder-decoder long short-term memory (LSTM) models with attention [36] for each paraphrase operation. We tuned and tested our automatic neural paraphrasing 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.

4.3.3 Corpus usage

The resulting 19k generated paraphrased dataset was combined with the original crowdsourced 10k training dataset. This 29k sentences was further combined with the original 29k training dataset resulting 58k-triplet training dataset for each operation.

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

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 multi-source and mixture-of-expert translation model, which is described in the next section.

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 emphasized that the use of paraphrased dataset in translation step was done on training stage. In this stage, the paraphrase were acting as regularizer and the means of ensembling, improving robustness of the ensembled model as a whole.

4.4. Experiment result and analysis

4.4.1 Result

Table 4.5: 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

Table 4.5 lists the scores of the paraphrase 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.

Table 4.6 shows the performance of our proposed neural caption translation. Both of the proposed models outperforms the NMT baseline. The proposed multi-source

Table 4.6: The performance of proposed neural caption translation in comparison with the baseline

Model Name	dec_reset	dec_type	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	sum	sum	37.0	55.1	30.2	49.5	25.1	45.0
	concat	sum	36.6	54.8	29.8	49.2	24.2	44.1
		gated_sum	37.2	54.9	30.4	49.7	26.1	44.9
		concat	37.6	55.4	30.1	49.4	24.4	44.3
		max_pooling	37.1	54.8	30.0	49.2	24.3	44.0
		avg_pooling	37.0	55.0	30.8	49.6	25.0	44.3
Uniform weighted ensemble	-	-	39.6	56.9	31.4	50.7	26.7	46.0
Mixture-of-expert ensemble	-	-	**40.5**	**57.6**	**32.5**	**51.3**	**28.0**	**46.8**

Table 4.7: 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

NMT shows slight improvement, although the improvement from proposed mixture-of-expert NMT was higher.

For the mixture-of-expert model, this performance increase indicates that the each expert model is slightly different between each other, and worked well in uniform-weighted ensemble and mixture-of-expert scenario. The mixture-of-expert also performed better than uniform-weighted NMT in three cases. The results also indicate that the mixture model successfully distribute weight to each expert translation model sequentially, compared with regarding all the weight uniformly.

Another interesting result is when we compare our proposed model result with the result of combining all available paraphrased data into one single data as shown in "Combine all data". This baseline shows that what we are doing is not simply data

Figure 4.3: Reference image for translation shown in Table 4.8.

augmentation, because this result proved that just by combining all of the data, some performance reduction were happened. As elaborated in the Chapter 1, we want to control the distribution of the similar sentence and its source sentence, as opposed to just using an external corpus, or using an arbitrary kind of paraphrase.

Table 4.7 shows the current submission systems in the official WMT17 shared task. Our proposed model outperformed the baseline in WMT17 with a 13.2 BLEU score margin. Although it may not be fair to compare directly with other multmodal models since our proposed model was developed only on textual modalities, it is surprising to see that our proposed model actually could produce competitive result with other multimodal models. Our proposed mixture-of-expert model outperformed several multimodal models such as those developed by [64], [65], [66], and [50].

4.4.2 Discussion

To further analyze the contribution between the experts trained on the original data and that trained on paraphrased data, we compare the paraphrasing and translation process step-by-step in our proposed model. Table 4.8 lists the resulting output sentences. "Baseline" model result means the result of one expert in our ensemble model.

This source sentence shown in Table 4.8 was translated using each baseline model (an expert), resulting five different translation hypotheses. We calculated BLEU+1

Table 4.8: Examples of resulting sentences.

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
[Multi-paraphrase] Uniform Weight	Ensemble	zwei motorradfahrer fahren auf einer straße am fluss .	0.84
[Multi-paraphrase] Mixture-of-expert	Ensemble	zwei motorräder fahren auf einer straße am fluss entlang .	0.97
(Data)	**Target**	**zwei motorräder fahren auf einer straße dem fluss entlang .**	

scores for each hypothesis against the target, resulting the source-deleted and source-inserted expert model yielded the best result.

The aim of proposed mixture-of-expert model task is to make sure the best part of each model are 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-expert 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, removing the word "two" in the source sentence implies some changes in the word "*motorradfahrer*" which is decoded in other model, with the exception of deletion translation model which resulting the word "*motorräder*". Another example is the word "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. Here we can see that 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.

Figure 4.4: Reference image for translation shown in Table 4.9.

Additionally, in Table 4.9 and Fig. 4.4 we also provide examples when our proposed model failed to improve the translation result, compared with our baseline model. In this example, all the expert model except the original produced significantly shorter sentences compared with the target sentence. This resulted a shorter sentence result when all of this expert was combined with either ensembling method.

The reordering expert model, which is expected to be more robust on reordered target sentence returns a reordered sentence hypothesis. This reordered translation hypothesis has the highest BLEU+1 score compared with other translation hypothesis except the original. But actually, if we see the target sentence, there isn't much reordering happened in the alignment of source sentence and target sentence. Therefore, reordering hypothesis doesn't help much here. The expected behavior, when reordering can propose "an metallseilen" instead of "metall an einem sell" in the original translation expert, doesn't happened.

Another expected insertion is, the word "das" in the original expert translation hypothesis, is decoded before the word "an metall", not before the phrase "einen langen rosafarbenen rock". But in this case, insertion translation expert model simply doesn't improve the result. The overall BLEU+1 score of mixture-of-expert model is higher than the uniform weight, but still lower than the baseline model.

Looking at some of the result errors brought us to the hypothesis that these errors can be prevented by developing a prior guess on which operation is more useful even

Table 4.9: Examples of resulting sentences.

Translation Model	Type	Sentences	BLEU+1
(Data)	**Original**	**a little girl climbing metal rope cables wearing a long pink skirt and black t-shirt .**	
Baseline	Original	ein kleines mädchen klettert metall an einem seil , das einen langen rosafarbenen rock und einem schwarzen t-shirt klettert .	0.9
Single Paraphrase Model	Deletion	ein kleines mädchen klettert metall seilen und einem schwarzen t-shirt klettert .	0.47
	Insertion	ein kleines mädchen klettert mit einem langen rosafarbenen rock an einem seil hoch .	0.61
	Reordering	ein kleines mädchen in einem langen pinkfarbenen rock und schwarzem t-shirt klettert metall an einem seil .	0.76
	Substitution	ein kleines mädchen klettert an einem seil seil seilen und einem schwarzen t-shirt klettert .	0.63
[Multi-paraphrase] Uniform Weight	Ensemble	ein kleines mädchen klettert an einem seil seil und einem schwarzen t-shirt hoch .	0.59
[Multi-paraphrase] Mixture-of-expert	Ensemble	ein kleines mädchen klettert metall mit einem langen rosafarbenen rock und einem schwarzen t-shirt .	0.67
(Data)	**Target**	**ein kleines mädchen , das an metallseilen hochklettert und einen langen rosafarbenen rock und ein schwarzes t-shirt trägt .**	

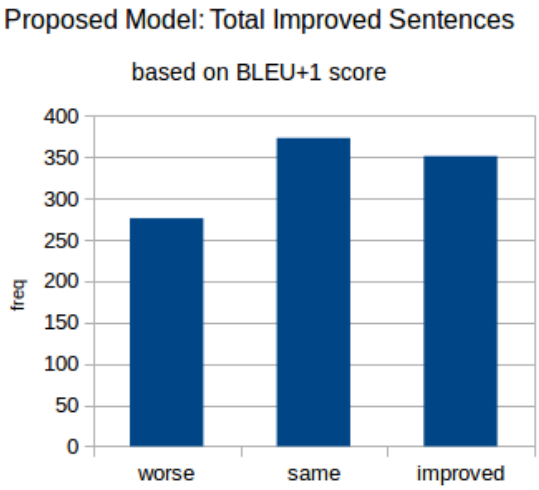

Figure 4.5: BLEU+1 differences between best proposed model result and baseline

before the translation or decoding starts. This will help the model to put more emphasis on which operation is more useful, given the source sentence. Another interesting approach is also to introduce variational error, where the system is also trained not only with the true example, but also with the false example.

Finally, in Figure 4.5 we showed the number of translation hypothesis that were worse, same, or improved, compared with the baseline translation hypothesis. There are 276, 373, and 351 sentences which were worse, same, and improved, consecutively. Therefore, we showed that by using our proposed paraphrasing strategies on the source side, can statistically improved 35% sentences on the test set.

4.5. Conclusion

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

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 models, and used it with the multi-expert approach within NMT. The use of image information is just for references for crowdworkers for paraphrasing, and it's not used in the neural models both for paraphrasing and translation.

The results indicate that our textual model with the mixture-of-experts NMT performed better than the uniform-weighted NMT model. Our best proposed system outperformed the baseline in WMT17 with a 13.2 BLEU score margin, which is close to the top score that used a multimodal model. The hypothesis of regularizing models by paraphrasing on the source sentence was proven to be effective. In the future, we will further investigates various methods of incorporating visual information into NMT models.

Chapter 5

Experiment on Neural Automatic Post-editing

This section discuss about the result and evaluation from the task defined in Section 3.4.

5.1. Experimental Settings

5.1.1 Baseline

There are two baselines for this task. To know whether the APE model successfully made improvement, we compare the *mt* directly with the *pe* as a baseline. The second baseline is taken from Fujita et al. proposed model of using SMT trained using identical pairs of sentences on the target side of their in-house daily life conversation corpus [67].

5.1.2 Proposed Model

Similar with the previous task, we propose two kinds of model for this task: multi-source NMT and multi-expert NMT. All models are trained using 512 hidden size,

0.0001 initial learning rate, 0.2 dropout ratio for embedding layer, and 0.5 dropout ratio for recurrent layer. The hypothesis were then decoded using beam strategy with size of 5. Other parameter are the same with as mentioned in Section 4.2.

The multi-source NMT proposed model consists of 2 encoders for {*src*,*mt*} and one decoder for {*pe*}. On the other hand, the proposed multi-expert NMT are consists of two encoder-decoder LSTM model of *src* to *pe* and *mt* to *pe* pairs.

5.1.3 Proposed Decoding Strategies

To take advantages of the hypothesis in the triplet of APE dataset, we propose a reranking and filtering algorithm using a 3-gram language model trained on our English part of dataset using KenLM [68]. The strategies are listed as follows:

- ***Rerank*** approach used sorted the hypothesis of beam decoding process using the normalized score obtained from language model.
- ***LMfilter*** approach compare final hypothesis with unmodified *mt* sentence from dataset. If the final hypothesis has lower language model score than the original *mt* sentence score, *mt* sentence will be returned.
- ***Rerank + LMfilter*** is the combination of both *Rerank* and *LMfilter* decoding strategy

5.2. Dataset

5.2.1 NICT QE-APE dataset

NICT QE-APE dataset was developed in order to support the research on quality estimation and automatic post-editing of machine translation results in involving Asian languages. This dataset composed by the triplets of translation sources (*src*), machine translation system hypothesis (*mt*), and post-editing target (*pe*). The total

number of triplets in this multi-lingual dataset is 8459 triplets for each language pair: Japanese to English, Japanese to Chinese, and Japanese to Korean.

The creation of this dataset involved a human evaluator that grade the quality of a hyp resulted from an in-house statistical machine translation system that translated a src sentences about travel and hospital utterances. After that, the annotator then do post-editing by performing minimal edit by keeping grammatical and semantical information quality in mind.

In this dataset, Fujita et al. also did some post-editing attempt using SMT system and reached best performance by introducing identical pair of sentences in the target side of their corpus. We were using this experiment result as our baseline for this Neural Automatic Post-editing task.

5.2.2 Generation of artificial dataset for pre-training

Training a comprehensive model only with NICT-QE/APE dataset is not enough, we need more data to make the model really grasp the idea of post-editing. However, to produce a large-scale dataset using human annotation is slow and expensive. Therefore, we propose a generation of an artificial dataset from a pair of MT dataset.

Our artificial dataset is composed from the concatenation of several datasets as follows:

- Basic Travel Expression Corpus [69]
- Tanaka Corpus [70]
- REIJI Corpus
- in-house medical domain corpus

This concatenated dataset now only consists of $\{src, trg\}$ pair. As described in Section 1.2.4, an APE model needs to be trained on triplets of *src*, *mt*, *pe*. To artificially generate the triplet, we train an SMT model using this data, and use this model to translate the *src* into *trg* resulting translation hypothesis *hyp*. As shown

in Fig. 5.1, the resulting *hyp* are then regarded as *mt*, and the target sentences are used as *pe*.

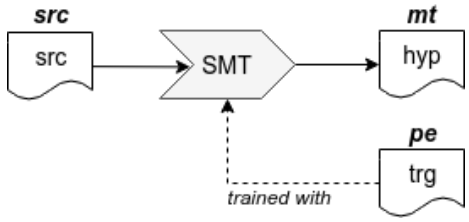

Figure 5.1: Creating artificial triplets using SMT

We also follow the intelligent selection algorithm based on cross-entropy differences as proposed by Moore and Lewis [71]. Using this method, we sort the sentences in artificial dataset based on its cross-entropy similarity with the NICT-QE/APE dataset. After that, we take 50k most similar sentences, excluding the sentences which has less than 3 words. This 50k most relevant sentences are then combined with NICT-QE/APE dataset for finetuning.

5.2.3 Corpus usage

We use 1.1M artificial dataset described in previous section for pre-training our proposed model. After that, the model is finetuned with the combination of 50k relevant artificial dataset and 8k NICT-QE/APE dataset. The reason why we don't use only the NICT-QE/APE dataset is because this dataset is too small and prone to make our proposed model goes overfit.

5.3. Experiment result and analysis

5.3.1 Result

Table 5.1 shows the result of our proposed method in BLEU [72] and TER [55]. Different with BLEU score, lower TER score indicates better result. Result No. 3

until No. 8 shows the result of our proposed multi-source model. On the other hand, our proposed multi-expert model didn't perform as good as the multi-source model.

The best combination function in multi-source NMT for this task was found using both concatenation for f_{enc} and f_{att} by reaching 44.7 BLEU score and 40.1 TER score. It is also shown that in the same model, the proposed decoding strategy LMFilter can increase 0.5 BLEU score from the normal decoding approach. Different with the reported combination function in [40], here we found that in APE task, concatenation is more effective for f_{enc} compared with summation.

Table 5.1: APE Result in BLEU and TER

No	Model	f_{enc}	f_{att}	Normal		Rerank		LMFilter		LMFilter+Rerank	
				BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
1	**Baseline mt to pe**	-	-	43.7	42.4	-	-	-	-	-	-
2	**Fujita et al. (2017)**	-	-	44.0	41.9	-	-	-	-	-	-
3	**Multi-source**	sum	sum	44.4	40.2	44.0	40.7	44.5	40.9	44.0	41.3
4		**concat**	sum	43.7	40.7	43.5	41.1	44.1	41.4	43.7	41.8
5			gated_sum	43.7	40.8	43.7	41.1	43.8	41.5	43.6	41.7
6			concat	44.2	**40.1**	44.3	40.3	**44.7**	40.9	44.6	40.9
7			max_pooling	43.9	40.8	43.8	41.0	43.8	41.7	43.8	41.4
8			avg_pooling	43.3	41.1	43.4	41.3	43.7	41.4	43.6	41.6
9	**Multi-expert**	-	-	41.4	43.4	41.1	44.2	41.4	43.4	41.1	44.2

Table 5.2: APE Example 1

Type	Sentence
src	あなた の 意見 が いい と 思い ます
mt	i think your opinion .
hyp	i think your ideas are good .
pe	i support your opinion .

This example in table 5.2 shows the case when our proposed model produced post-edit hypothesis that has similar meaning with the post-edit target. Compared with the *mt* sentences, the translation result has been significantly improved.

On the other hand, in table 5.3, sometimes the translation hypothesis mt is already adequate, but the model was paraphrasing it into other sentence. In this example, either *mt*, *hyp*, and *pe* has the same idea, and grammatically plausible.

Table 5.3: APE Example 2

Type	Sentence
src	それでは 説明書 を 見 ながら 詳しく 説明 し ます 。
mt	then i 'll explain it in detail over a look at the manual .
hyp	i will explain it in detail over the instructions .
pe	then i will explain in detail while looking at the information sheet .

Table 5.4: APE Example 3

Type	Sentence
src	あなた と 医師 の 間 に ある カーテン は 開け ます か 。
mt	can i open the curtains in between you and doctor ?
hyp	do you open the curtains in between you and doctor ?
pe	can i open the curtains in between you and doctor ?

Table 5.4 shows the case when our proposed model failed to improve or keep the MT translation hypothesis. In this example, the *mt* sentence is already good, but the model change two words "can I" into "do you", which change the sentence meaning completely. Although if we just consider this part, the word "do you" might be a good sentence starter such as "do you want me to open the curtains in between you and doctor?"

5.3.2 TER distribution comparison

In Fig. 5.2, we visualize the post-editing operation in terms of TER score. The expected behaviour of an APE model is that it represent the post-edit distribution from *mt* to *pe* as shown in yellow-colored bar (*ideal*) in contrast with the result with our APE model which is shown in blue-colored bar (*result_normal*) and red-colored bar (*result_lmfilter*) for both result No. 6 which used normal and LMFilter decoding strategy shown in Table 5.1 consecutively. We separated the x-axis into four groups to categorize how extensive the operation has been and should been done.

One of the expected behaviour is when the *mt* sentence is already good enough, the APE model should not modify the sentence. From the 0.0 and 0.01-0.50 TER range category in the bar, we can see that the proposed model should leave more sentence as is than it should. The LM filter, which decides not to modify *mt* sentence

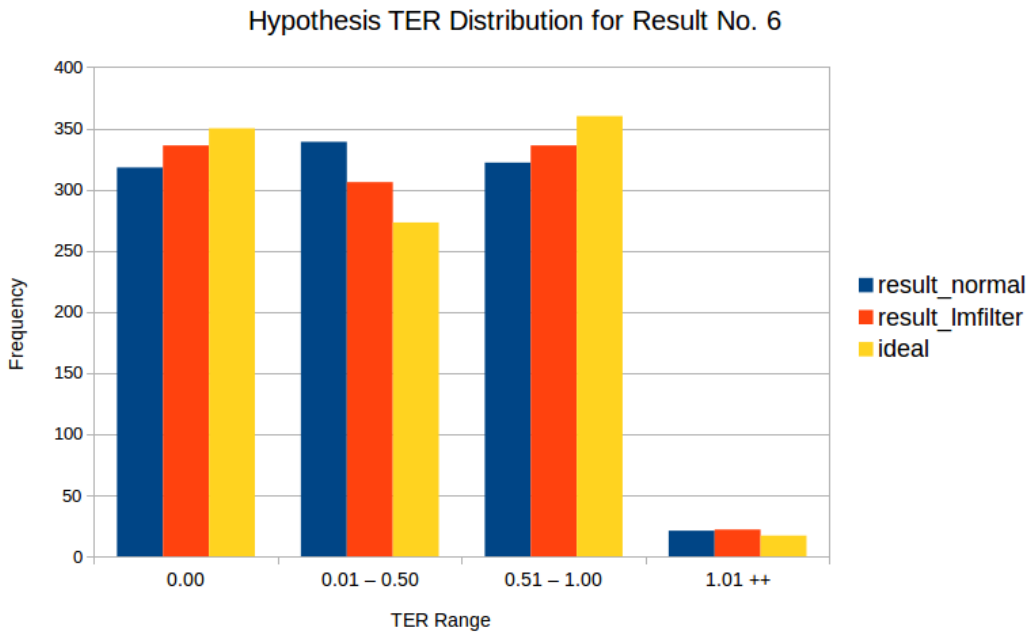

Figure 5.2: TER distribution between model result and ideal result

when it's already good, works as expected, shown by smaller TER frequency differences compared with normal decoding. We suggest that some more tuning needs to be done to discourage editing the model when it is already good.

5.4. Conclusion

We have developed sequence-to-sequence neural APE model that incorporate source information. There was some +0.7 BLEU and −1.8 TER points improvement when using both of the attentions from *src* and *mt* are concatenated. Analysis from TER showed that the model APE operation distribution is quite similar with the ideal expected distribution. The only problem is that the model are modifying more than it should be. Some quality estimation might be needed in order to help the model decide how extensive the operation should be.

Chapter 6

Conclusions and Future Directions

6.1. Conclusions

This thesis alleviates paraphrasing on both source side and target side to leverage machine translation quality. We proposed a novel perspective through general framework to utilize neural paraphrasing to leverage machine translation quality. This approach is divided on two main contribution area such as source and target side of MT, where the paraphrasing occurs.

In the source side, we showcased the use of paraphrasing to improve the result of visual description translation on which it shows that a single caption cannot represents all the information of the image to which it refers to. We developed multi-paraphrase of WMT 2017 dataset in a semi-supervised manner by partially crowdsourcing. Our proposed multi-expert model influenced by paraphrasing has been proven to be effective by 13.2 BLEU improvement over the baseline.

From the error analysis we found that the expert model already showed specialization in terms of paraphrase operation it trained to. However, sometimes this paraphrase operation is not what the model needs in a particular case. Some methods to analyze or guess what kind operation is needed to improve the model is needed to emphasize the elementary operations each translation case is suitable to. In conclu-

sion, we reported a novel perspective on how paraphrasing can be useful to leverage neural machine translation quality by proposing our elementary operation inspired by the basic operations that builds a sentence.

We also investigated and leveraged the use of paraphrasing on target side of NMT to improve translation quality in way of Automatic post-editing. Our proposed method that paraphrase the translation hypothesis in regard to source sentence has been proven to be effective by improving 0.7 BLEU and 1.8 TER points on NICT-QE/APE dataset. Looking at the TER distribution comparison shown in section 5.3.2, we can see that the LMFilter decoding strategy, as the simplest implementation of quality estimation, can improve post-editing result that still has distribution differences with the ideal distribution.

To conclude, we investigated post-editing approach using various combination functions on which we found effective way to post-edit, especially in distant language group pair such as Japanese-English which is under-explored. From both source of NMT, we proposed

6.2. Future Directions

There are still a lot of room for improvement in terms of incorporating additional knowledge to MT, given its advantages. For we have proved that the basic operation itself is useful for ensembling, it is also possible to do this in target side through some ways.

For the visual description task, we have showed a method of ensembling on several NMT model using our proposed elementary paraphrase operations to incorporate textual knowledge. Our proposed paraphrasing operation relies on using image as the idea representation of the sentences being paraphrased. Consequently, any form of idea representation such as semantic form or speech, can be used to further extend this work into other modalities.

Another interesting future work for this task is also to develop a prior hypothesis

on which elementary operation is suitable for each translation pair with respect to the source sentence, given the fact that each elementary operations contributes differently on each data. Further improvement of this task can also involve variational loss, where the model is trained not only with the true example, but also with the false example.

On APE task, we saw that the model quality can be more improved if the model can understand or estimate the quality of the translation hypothesis. By assuming that a good quality estimator is available, it can guide the training process of the model as these both post-editing and quality estimation task is tightly related. Our proposed model works in word-level, so to influence the training in every time step, granularity level such as word might be effective.

Appendices

Appendix A

Data Details

A.1. WMT 2017 Multimodal Translation Shared Task Dataset

WMT 2017 Multimodal Translation Shared Task Dataset (WMT Dataset) is an extension of Multi30k dataset [43]. The Multi30k dataset was developed to stimulate multilingual multimodal research. Given the advantages of multimodal systems such as image captioning, the development of this task on other language has not received attention yet. The development of this multilingual multimodal data opens up the possibility for researchers in machine translation to work in this domain.

The Multi30k dataset get its images from Flickr30k image caption dataset [73] that contains 31.014 images sourced from online photo-sharing websites. In this dataset, each images has five English descriptions, collected using crowdsourcing. The Multi30k dataset therefore extends the Flickr30k dataset with German captions. There are two types of German caption in this dataset: translated and independent. The translated caption were translated directly from the English caption, and independent caption are the German caption that were created directly from the image itself, without looking at the English caption.

The dataset that we used is the translated dataset, which is also used in WMT

2017 Multimodal Translation Shared Task. It contains 29k caption for training, 1k caption for development. On the other hand, this WMT Shared Task has new dataset each year it is conducted. Therefore, there are dataset from 2016 and 2017 [1]. The dataset for the year 2017 consists of two kinds: in-domain dataset and out-domain dataset. The in-domain dataset are taken from the same Flickr images, while out-domain dataset images were taken from MSCOCO dataset [62]. This out-domain dataset caption were consisted of ambiguous sentences, in order to benchmark the performance of the model while being in the out-of-domain situation.

A.2. NICT-QE/APE Dataset

The Quality Estimation (QE) and Automatic Post-editing (APE) task were commonly done on MT outputs. This MT output quality augmentation is especially needed for translation language pair which is quite different such as in Asian language. However, the dataset for this task, especially in Asian language pair such as Japanese, Chinese, and Korean with English, is less studied. Therefore, the creation of NICT-QE/APE dataset is aimed to facilitate this [67].

This dataset were designed to improve translation in its most practical domain: travel and hospital. Most of the sentence in this dataset is spoken language produced from speech translation service managed by National Institute of Information and Communications Technology (NICT), Japan. They created three language pairs by considering the largest speakers of the language who visited Japan as reported by Japan National Tourism Organization (JNTO). The pairs are Japanese-English, Japanese-Chinese, and Japanese-Korean.

This dataset consisted of triplets per row: translation source (*src*), machine translation system hypothesis (*mt*), and post-editing target (*pe*). The total number of triplets in this dataset is 8459 triplets for each language pair. Aside from the sentence triplet, this dataset also provides score for sentence level QE and word-level

[1]When this thesis is being written, the WMT has not released the full 2018 dataset yet.

QE, which can be used as the label in QE task.

For APE task, the creation of this dataset involves human annotator which is doing manual post-editing under this following guidance, as quoted from their paper:

1. Refer only to *src* and *hyp* basically.
2. Make each *hyp* grammatical and semantically appropriate with respect to its *src*
3. Perform minimal edits

This study provided preliminary result of APE intended as a baseline for future uses of this dataset. They proposed various baseline of various data combinations between the main dataset and their internal dataset as pseudo training data and bitext back-off. The reported highest score was 44.0 BLEU and 41.9 TER score for Japanese-English language pair. It was trained using bitext back-off, where an SMT system was trained using the data combination of the main dataset and {*pe*,*pe*} pair of the pseudo dataset.

References

[1] Rahul Bhagat and Eduard Hovy. What is a paraphrase? *Computational Linguistics*, 39(3):463–472, 2013.

[2] Teruko Mitamura, Eric Nyberg, and Jaime Carbonell. An efficient interlingua translation system for multi-lingual document production. In *Proceedings of Machine Translation Summit III*, pages 2–4, 1991.

[3] Arturo Trujillo. Interlingua machine translation. In *Translation Engines: Techniques for Machine Translation*, pages 167–201. Springer, 1999.

[4] M. Nagao, J. Tsujii, and J. Nakamura. Machine translation from japanese into english. *Proceedings of the IEEE*, 74(7):993–1012, July 1986.

[5] Eiichiro Sumita, Hitoshi Iida, and Hideo Kohyama. Translating with examples: a new approach to machine translation. In *The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, number 3, pages 203–212, 1990.

[6] Harold Somers. Review article: Example-based machine translation. *Machine Translation*, 14(2):113–157, June 1999.

[7] Satoshi Sato. Ctm: An example-based translation aid system. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 4*, COLING '92, pages 1259–1263, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[8] 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*, NAACL-Short '09, pages 217–220, Stroudsburg, PA, USA, 2009.

[9] Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation (StatMT '07)*, pages 203–206, Stroudsburg, PA, USA, 2007.

[10] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.

[11] Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus phrase-based machine translation quality: a case study. *CoRR*, abs/1608.04631, 2016.

[12] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39. Association for Computational Linguistics, 2017.

[13] Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word sense disambiguation improves statistical machine translation. In *ACL*, 2007.

[14] Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 61–72, 2007.

[15] Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. Improving word sense disambiguation in neural machine translation with sense embeddings. In

Proceedings of the Second Conference on Machine Translation, pages 11–19. Association for Computational Linguistics, 2017.

[16] Jiacheng Zhang, Yang Liu, Huanbo Luan, Jingfang Xu, and Maosong Sun. Prior knowledge integration for neural machine translation using posterior regularization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1514–1523. Association for Computational Linguistics, 2017.

[17] Philip Arthur, Graham Neubig, and Satoshi Nakamura. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567. Association for Computational Linguistics, 2016.

[18] Chen Shi, Shujie Liu, Shuo Ren, Shi Feng, Mu Li, Ming Zhou, Xu Sun, and Houfeng Wang. Knowledge-based semantic embedding for machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2245–2254. Association for Computational Linguistics, 2016.

[19] Katharina Wäschle and Stefan Riezler. Integrating a large, monolingual corpus as translation memory into statistical machine translation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, 2015.

[20] Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535, 2015.

[21] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96. Association for Computational Linguistics, 2016.

[22] Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. Guiding neural machine translation decoding with external knowledge, 01 2017.

[23] R. De Beaugrande and W.. V. Dressler. Introduction to text linguistics. *Longman, NY*, pages 23–38, 1981.

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

[25] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[26] Trevor Cohn and Mirella Lapata. An abstractive approach to sentence compression. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):41, 2013.

[27] Liang Zhou, Chin-Yew Lin, and Eduard Hovy. Re-evaluating machine translation results with paraphrase support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 77–84, Stroudsburg, PA, USA, 2006.

[28] Santanu Pal, Pintu Lohar, and Sudip Kumar Naskar. Role of paraphrases in pb-smt. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 8404*, CICLing 2014, pages 242–253, Berlin, Heidelberg, 2014. Springer-Verlag.

[29] Nitin Madnani and Bonnie J. Dorr. Generating targeted paraphrases for improved translation. *ACM Trans. Intell. Syst. Technol.*, 4(3):40:1–40:25, July 2013.

[30] Eric Nichols, Francis Bond, D Scott Appling, and Yuji Matsumoto. Paraphrasing training data for statistical machine translation. *Journal of Natural Language Processing*, 17(3):3_101–3_122, 2010.

[31] Wei He, Shiqi Zhao, Haifeng Wang, and Ting Liu. Enriching smt training data via paraphrasing. In *Proceedings of IJCNLP*, 2011.

[32] Anabela Barreiro. *SPIDER: A System for Paraphrasing in Document Editing and Revision — Applicability in Machine Translation Pre-editing*, pages 365–376. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[33] Chris Callison-Burch, Philipp Koehn, and Miles 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)*, pages 17–24, Stroudsburg, PA, USA, 2006.

[34] Michel Simard, Cyril Goutte, and Pierre 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)*, pages 508–515, Rochester, NY, USA, 2007.

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

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

[37] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, 2015.

[38] Philip Gage. A new algorithm for data compression. *The C Users Journal*, pages 23–38, 1994.

[39] Ekaterina Garmash and Christof Monz. Ensemble learning for multi-source neural machine translation. In *COLING*, 2016.

[40] Barret Zoph and Kevin Knight. Multi-source neural translation. *CoRR*, abs/1601.00710, 2016.

[41] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Comput.*, 3(1):79–87, March 1991.

[42] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *CoRR*, abs/1701.06538, 2017.

[43] 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*, pages 70–74, 2016.

[44] Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *ACL*, 2017.

[45] Franz Josef Och and Hermann Ney. Statistical multi-source translation. In *In MT Summit 2001*, pages 253–258, 2001.

[46] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *HLT-NAACL*, pages 866–875. The Association for Computational Linguistics, 2016.

[47] Barret Zoph and Kevin Knight. Multi source neural translation. *arXiv preprint arXiv:1601.00710*, 2016.

[48] 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*, abs/1707.04481, 2017.

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

[50] Jindrich Helcl and Jindrich Libovický. CUNI system for the WMT17 multimodal translation task. *CoRR*, abs/1707.04550, 2017.

[51] Loïc Dugast, Jean Senellart, and Philipp Koehn. Statistical post-editing on systran's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[52] Michel Simard, Cyril Goutte, and Pierre Isabelle. Statistical phrase-based post-editing. In *Proceedings of NAACL*, 2007.

[53] Hanna Béchara, Yanjun Ma, and Josef Genabith. Statistical post-editing for a statistical mt system, 2011.

[54] Marcin Junczys-Dowmunt and Roman Grundkiewicz. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. *CoRR*, abs/1605.04800, 2016.

[55] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200, 2006.

[56] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration*

Sessions, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[57] Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain, April 2017. Association for Computational Linguistics.

[58] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin 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.

[59] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[60] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[61] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870, 2015.

[62] 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*, pages 740–755. Springer, 2014.

[63] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia 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.

[64] Pranava Swaroop Madhyastha, Josiah Wang, and Lucia 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*, pages 470–476, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[65] Iacer Calixto, Koel Dutta Chowdhury, and Qun Liu. Dcu system report on the wmt 2017 multi-modal machine translation task. In *Proceedings of the Conference of Machine Translation (WMT)*, volume 2, 2017.

[66] Mingbo Ma, Dapeng Li, Kai Zhao, and Liang Huang. Osu multimodal machine translation system report. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 465–469, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[67] Eiichiro Sumita Atsushi Fujita. Japanese to english/chinese/korean datasets for translation quality estimation and automatic post-editing. In *Proceedings of WAT 2017*, Taipei, Taiwan, November 2017.

[68] Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *In Proc. of the Sixth Workshop on Statistical Machine Translation*, 2011.

[69] Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *LREC*, pages 147–152, 2002.

[70] Yasuhito Tanaka. Compilation of a multilingual parallel corpus. *Proceedings of PACLING 2001*, pages 265–268, 2001.

[71] Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[72] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. Bleu: a method for automatic evaluation of machine translation. pages 311–318, 2002.

[73] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 2641–2649, Washington, DC, USA, 2015. IEEE Computer Society.