Japanese-English Machine Translation of Recipe Texts

Takayuki Sato
Tokyo Metropolitan University
Tokyo, Japan
sato-takayuki@ed.tmu.ac.jp

Jun Harashima
Cookpad Inc.
Tokyo, Japan
jun-harashima@cookpad.com

Mamoru Komachi
Tokyo Metropolitan University
Tokyo, Japan
komachi@tmu.ac.jp

Abstract

Concomitant with the globalization of food culture, demand for the recipes of specialty dishes has been increasing. The recent growth in recipe sharing websites and food blogs has resulted in numerous recipe texts being available for diverse foods in various languages. However, little work has been done on machine translation of recipe texts. In this paper, we address the task of translating recipes and investigate the advantages and disadvantages of traditional phrase-based statistical machine translation and more recent neural machine translation. Specifically, we translate Japanese recipes into English, analyze errors in the translated recipes, and discuss available room for improvements.

1 Introduction

In recent years, an increasing amount of recipe data has become available on the web. For example, as of September 2016, more than 2.45 million recipes are available on cookpad, 1 million on Yummly, and 0.3 million on Allrecipes, to name a few. These recipes are from all over the world, and are written in various languages, including English and Japanese. However, language barriers may prevent the users from discovering recipes of local specialities.

Many researchers have focused on various tasks such as recipe analysis (Maeta et al., 2015), information retrieval (Yasukawa et al., 2014), summarization (Yamakata et al., 2013), and recommendation (Forbes and Zhu, 2011). However, to date, little work has been done on machine translation of recipe texts. In particular, Japanese foods are gaining popularity because they are considered healthy. We believe that many people would be able to use cooking recipes currently available only in Japanese if those Japanese recipes were translated into other languages.

In this study, we translated recipes via machine translation and investigated the advantages and disadvantages of machine translation in the recipe domain. First, we translated Japanese recipes into English using phrase-based statistical machine translation (PBSMT) and neural machine translation (NMT). Then, we classified translation errors into several categories in accordance with Multidimensional Quality Metrics (MQM) (Burchardt and Lommel, 2014). Finally, we analyzed the classified errors and discussed how to mitigate them.

2 Recipe Parallel Corpus

As described in the previous section, we focused on translating Japanese recipe texts into English, because almost all of the recipe texts on cookpad, which is one of the largest recipe sharing services in the world, are written in Japanese. We used a Japanese-English parallel corpus provided by Cookpad Inc. that includes 16,283 recipes. Each recipe mainly consists of a title, ingredients, and steps. Examples of a title, an ingredient, and a step are shown in Table 1.[1] Unlike general parallel corpora, a translation pair of a step does not always consist of one parallel sentence. Examples of step texts in Table 1 show the case where there are two sentences in the translation pair.

[1] In this paper, we use the abbreviation of the cases: NOM (nominative), ACC (accusative), and TOP (topic marker).

Proceedings of the 3rd Workshop on Asian Translation,
pages 58–67, Osaka, Japan, December 11-17 2016.

Table 1: Examples of title, ingredient and step.

Title	簡単 シンプル ! ふわふわ 卵 の オムライス easy simple ! fluffy egg of omurice Easy and Simple Fluffy Omurice
Ingredient	ご飯 (冷や ご飯 で も 可) rice (cold rice also available) Rice (or cold rice)
Step	ケチャップ と ソース を 混ぜ 合わ せ ます . 味見 し ながら 比率 は 調節 し て ください . ketchup and sauce ACC mix . taste while ratio TOP adjust please . Mix the ketchup and Japanese Worcestershire-style sauce. Taste and adjust the ratio.

Table 2: Number of sentences and words in each field.

	Language	Title	Ingredient	Step	Total
sentence		16,170	131,938	124,771	272,879
word	Japanese	115,336	322,529	1,830,209	2,268,074
	English	100,796	361,931	1,932,636	2,395,363

These translation pairs were collected through the following two processes: translation and modification. First, a Japanese native speaker fluent in English translated Japanese recipes into English. Then, two native English speakers checked the translation and modified it as necessary. Note that the participants in these two processes were familiar with cooking.

We adopted the following three preprocessing procedures to this corpus in order to easily handle it. First, each Japanese text and its English translation in steps were split into sentences by a period. We used sentences that met the following conditions in our experiments: (1) the number of the split sentences in Japanese is the same as that in English or (2) there are exactly one Japanese and two English sentences. In the sentences in English that met the second condition, the first period was changed into ‘, and’ to join two English sentences. This preprocessing excluded 25,654 texts where there were 59,282 Japanese step sentences and 57,016 English step sentences. Second, we excluded sentence pairs where the longer sentence is more than two times longer than the other. This process is necessary because some English sentences were translated as simple expressions, and hence the ratio of the length of the sentence pairs was sometimes large. An example is shown below.

(1) 関西 の お 店 の 味 ! 我が家 の お好み焼き .
kansai-style restaurant taste ! my own home okonomiyaki .
kansai-style okonomiyaki .

Third, sentences that contain more than 40 words were excluded from our experiments. Table 2 shows the number of sentences and words in each field after preprocessing. The size of the Japanese vocabulary was 23,519, while that of the English vocabulary was 17,307.

After prepossessing, we randomly chose 100 recipes as a development set (1,706 sentences) and 100 recipes as a test set (1,647 sentences). The former was used to tune our translation models, while the latter was used to analyze translation errors and to evaluate the translation models.

3 Machine Translation Methods

We used two methods in our experiments: PBSMT and NMT. The former has been widely accepted as one of the bases of machine translation systems that we generally use, whereas the latter has been gaining great attention in research community because of its fluency and simplicity.

PBSMT obtains a language model and a translation model (phrase table) from a parallel corpus and translates sentences based on these models (Koehn et al., 2003). The method achieves good performance on any language pair consisting of languages whose word orders are similar to each other, as in the case of English and French. Conversely, it performs poorly when the word orders of the languages differ, as in the case of English and Japanese. In addition, PBSMT often generates ungrammatical sentences because it does not consider syntactic information.

NMT embeds each source word into a d-dimensional vector and generates a target sentence from the vectors (Sutskever et al., 2014). Even though the method does not use any syntactic information, it can generate grammatical sentences. However, due to the execution time it requires, NMT generally limits the size of the vocabulary for a target language. Therefore, compared with PBSMT, which can handle many phrases in the target language, NMT has a disadvantage in that it cannot generate low frequent words. The method also has the disadvantage that it often generates target words that do not correspond to any words in the source sentences (Tu et al., 2016).

The setting for each method in this study was as follows. We used the parallel corpus described in Section 2 as our corpus, Moses (ver.2.1.1) (Koehn et al., 2007) as the PBSMT method, and conducted Japanese word segmentation using MeCab (Kudo et al., 2004) with IPADIC (ver.2.7.0) as the dictionary. Word alignment was obtained by running Giza++. The language model was learned with the English side of the recipe corpus using KenLM (Heafield, 2011) with 5-gram. Other resources in English were not used for training the language model because the style of recipe texts is different from general corpus in that it contains many noun phrases in title and ingredient, and many imperatives in step. The size of the phrase table was approximately 3 million pairs, and we used the development set to tune the weights for all features by minimum error rate training (MERT) (Och, 2003). We used the default parameter 6 for the distortion limit.

We reimplemented the NMT model in accordance with (Bahdanau et al., 2015). Note that we used long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) instead of gated recurrent unit (Cho et al., 2014) for each recurrent neural network (RNN) unit of the model. The model had 512-dimensional word embeddings and 512-dimensional hidden units with one layer LSTM. We set the vocabulary size of the model to $30,000$, and we did not perform any unknown word processing during training. Adagrad (Duchi et al., 2011) was used with the initial learning rate of 0.01 as an optimization method. The initial values for word embeddings on both sides were obtained by training word2vec[2] with default setting because better results were shown in our preliminary experiments. The initial word embeddings on the source side were learned with a raw Japanese recipe corpus (Harashima et al., 2016) consisting of approximately 13 million step sentences. Conversely, initial word embeddings on the target side were learned with approximately $120,000$ English step sentences included in the parallel corpus. Title sentences were not used for learning because they were often written with free expression largely different depending on each recipe. Ingredient sentences were also not used because most of them consisted of a few words. The batch size was set to 64 and the number of epochs was set to 10. We selected the model that gave the highest BLEU score in the development set for testing. Beam search for decoding in NMT was not carried out. When testing, the output length was set up to 40 words.

Each output was evaluated via two metrics: bilingual evaluation understudy (BLEU) (Papineni et al., 2002) score and rank-based intuitive bilingual evaluation score (RIBES) (Isozaki et al., 2010). BLEU is more sensitive to word agreement than RIBES, whereas RIBES is more sensitive to word order evaluation. We set two hyper-parameters for RIBES: α was 0.25 and β was 0.10.

4 Error Classification of Recipe Translation

We conducted blackbox analysis on the outputs of PBSMT and NMT. Blackbox analysis is a type of analysis that does not take into account how translation output is obtained. The error classification used for this analysis is based on the MQM ANNOTATION DECISION TREE (Burchardt and Lommel, 2014) because it makes the classification of each error more consistent. The method classifies each error by following a decision tree where each node asks a Yes/No question. If the question is answered with a 'Yes', the corresponding text span is classified as the error specified by the tree node. When a text span is classified as an error at higher priority, that part is not classified as other errors. The same process continues until the lowest priority error is checked.

The error classification defined in MQM is roughly divided into two parts: Accuracy and Fluency. Accuracy addresses the extent to which the target text accurately renders the meaning of the source text. It is usually called 'Adequacy' in the literature. Fluency relates to the monolingual qualities of the target

[2] https://radimrehurek.com/gensim/models/word2vec.html

text. In this section, we explain the fine classification of accuracy and fluency metrics in detail and describe how to classify and analyze the errors.

4.1 Accuracy

In terms of accuracy, MQM defines (1) Omission, (2) Untranslated, (3) Addition, (4) Terminology, (5) Mistranslation and (6) General. In this study, we adapted the MQM ANNOTATION DECISION TREE to recipe translation to classify each error. We modified the original MQM ANNOTATION DECISION TREE in three different ways. First, we divided mistranslation errors into substitution and word order and considered them independently. This is because the tendency of substitution and word order is so different in a distant language pair such as Japanese and English that the difference should be reflected. Second, we defined a new order to classify each error, in which substitution and word order are given the highest priority. This makes the classification of substitution easy, especially for NMT, which sometimes outputs completely different target words from source sentences. Third, we excluded terminology because terminology-related errors do not occur when only a single domain, such as food recipes, is considered. Therefore, in this study, the following fine classification was applied: (1) Substitution, (2) Word order, (3) Omission, (4) Untranslated, (5) Addition, (6) General. Here, we explain the definition of each error classification with examples.

Accuracy (Substitution) The target content does not represent the source content owing to inappropriate words. In the following example, 'Heat' is used for '割る' (break).

(2) 卵 を 割る .
egg ACC break .
Heat an egg .

Accuracy (Word order) The target content does not represent the source content owing to inappropriate word positions. In the following example, 'from step 1' should be placed after 'into a bowl .'.

(3) 1 の 器 に レタス を 入れる .
1 from bowl into lettuce ACC add .
Add the lettuce from step 1 into a bowl .

Accuracy (Omission) Content in the source sentence is missing from the translation. In the following example, the translation does not contain a word for 'はちみつ' (honey).

(4) はちみつ 生地 は 1 次 発酵 まで 済ませる .
honey dough ACC first fermenatation until finish .
Make the dough until the first rising .

Accuracy (Untranslated) Source words have been left untranslated. From the following example, it can be seen that there is an untranslated word '狭い' (narrow).

(5) 長 さ を 整え , 幅 の 狭い ほう で カット する .
length ACC adjust , width NOM narrow with cut .
Adjust the length , and cut the 狭い into it .

Accuracy (Addition) The translation includes words or phrases that are not present in the source sentence. In the following example, 'red' and 'into a pot' should not have been added.

(6) ソース を 加える .
sauce ACC add .
Add the red sauce into a pot .

Accuracy (General) It is applied when the translation error is difficult to classify into a certain category in terms of accuracy. The number of errors is counted by phrases in the source that are not represented in the target. In the following example, there are four errors.

(7) 出来上がっ た 時に 倒れ ない ため です .
finished when fall not for .
It will be hard to cover the cake .

4.2 Fluency

In terms of fluency, MQM defines (1) Word order, (2) Word form, (3) Function words, (4) Grammar general, (5) Fluency general. Here, we explain the definition of each error classes with examples.

Grammar (Word order) The word order is incorrect. In the following example, the position of the word 'place' is incorrect and is considered to be before the word 'parts'. The number of errors equals the number of content words at wrong positions.

(8) Parts of the face , place on a baking sheet .

Grammar (Word form) The wrong form of a word is used. The following example includes one error because 'uses' is incorrect.

(9) I uses the dough for step 4 .

Grammar (Function Words) This is a misuse of function words such as preposition, particle, and pronoun. From the following example, it can be seen that the function word 'to' is unnecessary.

(10) It 's finished to .

Grammar (General) In addition to the errors identified above, there are other grammatical errors such as insertion and omission of unnecessary content words. In the following example, there is not a verb.

(11) The honey dough for the first rising .

Fluency (General) Even when a sentence is grammatically correct, it may have some issues in terms of fluency. The sentence used as the example of this category is unintelligible because of the phrase 'from the cake'. For each unintelligible phrase, we count content words in it as errors (in this case, 'cake' and 'future.').

(12) I was going to be taken from the cake in the future .

5 Results and Discussion

We translated Japanese sentences in the corpus described in Section 2 into English sentences following the procedure described in Section 3. We then evaluated the outputs with automatic evaluation metrics, BLEU and RIBES. Finally, we discussed the results for each type of sentence, title, ingredient, and step. The outputs were also analyzed following the error classification procedure outlined in Section 4. Note that all the recipes in the test set were used for the automatic evaluation and 25 recipes randomly chosen from the test set were used for error analysis.

5.1 Automatic Evaluation

The results obtained following automatic evaluation by BLEU and RIBES are shown in Table 3.

Title is represented with a relatively large vocabulary and free expression largely different depending on each recipe. In other words, it includes low frequent expressions. The percentage of the number of sentences for which title accounts is very low compared with ingredient and step as shown in Table 2. Hence, the translation of title is more difficult than that of ingredient and step owing to data sparsity.

Table 3: Automatic evaluation BLEU/RIBES results.

Method	Title	Ingredient	Step	Total
PBSMT	**22.15 / 61.85**	**56.10 / 90.03**	25.37 / 74.98	**28.09 / 81.72**
NMT	19.68 / 61.49	55.75 / 89.70	**25.68 / 77.84**	28.01 / **82.79**

Table 4: Number of accuracy errors in 25 recipes.

Method	Substitution	Word order	Omission	Untranslated	Addition	General	Total
PBSMT	49 (11.0%)	98 (21.9%)	139 (31.1%)	23 (5.1%)	95 (21.3%)	43 (9.6%)	447
NMT	102 (19.2%)	20 (3.8%)	176 (33.1%)	0 (0.0%)	114 (21.5%)	119 (22.4%)	531

PBSMT shows better performance for title translation than NMT both in BLEU and RIBES, because it is possible for PBSMT to partially translate title using the phrase table created from infrequent expressions. On the other hand, some NMT outputs are very short and do not include any word that corresponds to any source words. It resulted in poor performance of BLEU.

Ingredient has very short sentences, with an average length of 3.0 words. In addition, there are not many translation candidates for each ingredient. Consequently, the BLEU and RIBES scores for both methods are very high. Although the margin between PBSMT and NMT is small, PBSMT exhibited better performance in both metrics. Translating the names of ingredients was similar to translation using a dictionary, at which PBSMT is better.

In the translation of step, NMT shows better performance than PBSMT in BLEU and RIBES. When several nouns are enumerated, the reordering distance tends to be long because the target sentence is usually written in imperative form. However, it appears that NMT does not have any difficulty in translating such sentences. This is because NMT is good at modeling long dependencies owing to the use of RNN. There is also a case where omission occurs in a source sentence and zero-anaphora and/or coreference resolution will be required to generate the omitted word in a target sentence. It appears difficult for both methods to output a word for the omitted word but NMT tended to estimate more words than PBSMT.

Finally, let us look at the results for RIBES. It is possible that RIBES is a metric that can be higher for NMT than for PBSMT. NMT tends to output shorter sentences than the references. Conversely, PBSMT does not output sentences that are as short as those of NMT because it ensures that all the source phrases are translated. However, the default parameter of RIBES optimized for patent translation (Isozaki et al., 2010) does not significantly penalize omission errors that frequently occur in NMT. Instead, it penalizes substitution errors and word order errors, which are abundant in PBSMT. This suggests that we need to investigate a better evaluation metric for assessing the quality of NMT.

5.2 Error Analysis

5.2.1 Accuracy

The number of accuracy errors is shown in Table 4. Compared with NMT, PBSMT has many errors related to the word order. In general, PBSMT exhibits poor results against syntactically different language pairs because reordering words is difficult in such cases. As the sentence length becomes longer, word order errors increase, because reordering words becomes more difficult. The majority of the corpus used in this study comprised short sentences, especially for title and ingredient. Ingredient sentences are very short and title sentences are relatively short. The average length of step sentences is also not so long, and is 14.0 words in Japanese and 15.0 words in English. However, many steps are written in imperative order form in English. Consequently, even when the sentence length is short, inevitably a word order error occurs because word reordering frequently occurs in the case of long distances. The example below is a part of a sentence in which some ingredients are enumerated; thus, PBSMT has difficulty in reordering word positions.

(13) 4 の 鍋 に 1 の ブリ & 3 の 大根 & しいたけ & 生姜 を
4 from pan to 1 from amberjack and 3 from daikon radish and shiitake mashrooms and ginger ACC
入れ ,
add ,
PBSMT: Amberjack and daikon radish and shiitake mushrooms , and add the ginger from step 1 to the pan from step 3

This error is frequently seen because the names of ingredients often appear in steps. It appears that the solution in order for PBSMT to handle these errors requires a translation model with a syntactic rule such as a constituent structure or dependency structure.

On the other hand, NMT has many more errors in terms of substitution than with PBSMT. In substitution, there were errors in which the meanings of the source word and the target word were not similar at all. For example, 'sweet potato' was output as the translated word for 'キャベツ' (cabbage). To solve this problem, the use of lexicon probability obtained from a phrase table or a dictionary is considered promising for the NMT model (Arthur et al., 2016).

There were many omission errors and addition errors in both PBSMT and NMT. In particular, omission errors account for a large percentage in both methods. The following example shows that omission errors or addition errors occur in either, or both methods.

(14) ホーム ベーカリー の 生地 作り コース で 生地 を 作る .
bread maker of dough setting with dough ACC make.
PBSMT: Make the dough in the bread maker to make the dough.
NMT: Make the dough using the dough setting.
Reference: Use the bread dough function on the bread maker to make the bread dough.

In terms of omission or addition errors, PBSMT and NMT output errors occur in the same sentences although the error positions are different. In the example above, omission of '生地 作り コース' (dough setting) and addition of 'to make' and 'the dough' are seen in the PBSMT output. On the other hand, NMT omits the translation of 'ホーム ベーカリー の' (on the bread maker). Thus, it appears that sentences in which machine translation output errors occur in both methods are somewhat similar.

Addition is seen in a sentence where an object in Japanese is omitted. Recipe steps in Japanese tend to omit words that have already appeared in the same recipe. In the translation of such sentences, some words should be inserted in the target sentence. An example is given below.

(15) 紙 に 包ん で ,
paper in wrap ,
NMT: Wrap the cake in the cake paper,
Reference: Wrap the cakes in parchment paper,

This sentence does not contain the source word that corresponds to 'the cake', but the word exists in the reference. NMT succeeded in generating 'the cake' in this example. However, in general, performing zero-anaphora resolution for inter-sentential arguments is difficult. NMT is more promising than PBSMT in terms of modeling of long dependency to estimate omitted arguments. It appears important to take into account the ingredients used or the order in which actions are completed in the flow of the recipe.

Although 'untranslated' is considered an error that occurs only in PBSMT, the ratio proves to be very low. The corpus used in this study did not have a large vocabulary; therefore, the words that appeared in the training dataset include almost all of the words in the test set. Therefore, untranslated errors rarely occurred in this dataset.

5.2.2 Fluency

The number of fluency errors is shown in Table 5. Word order errors appear to have occurred for the same reason as word order errors that adversely affect accuracy.

Few word form errors were seen in both methods. There was little ambiguity in tense, because title and ingredient are mostly noun phrases, and most of the steps are written in imperative form. In addition, disagreement between subject and verb or that of tense rarely occurred, because most of the subjects

Table 5: Number of fluency errors in 25 recipes.

Method	Grammar				Fluency	Total
	Word order	Word form	Function words	General	General	
PBSMT	18 (14.0%)	2 (1.6%)	24 (18.6%)	73 (56.9%)	12 (9.3%)	129
NMT	4 (4.8%)	1 (1.2%)	6 (7.2%)	17 (20.5%)	55 (66.3%)	83

corresponded to ingredients, which are expressed in third person singular.

More function word errors were seen in PBSMT than in NMT. The main class of word error encountered was the addition of an unnecessary function word. The reason for this appears to be the noise in the phrase extraction process when creating a phrase table. Output consisting of phrases with noise can be avoided by taking syntactic constraints into account. In the following example, 'in' is an inappropriate word:

(16) PBSMT: Remove the sinew from the chicken tenders and fold in lightly .

The errors in grammar in general were mainly errors related to a content word. In particular, omission and addition of a noun and a verb are observed in many outputs. This appears to have the same cause as function word errors. The following example shows the omission of a verb:

(17) PBSMT: Basic chiffon cake milk to make the dough .

The output of NMT has many unintelligible sentences that are classified under fluency general. NMT outputs a few grammar-related errors, such as word order, function word, and grammar general. Repetition of the same word and phrase were commonly seen in NMT but never in PBSMT.

(18) NMT: leave to steam for about 2 hours , and open the pot , and open the pot .

6 Related Work

In machine translation in the recipe domain, solving zero-anaphora analysis problems appears to be essential because some of step sentences have an order relationship in which reference is made to words that have previously appeared, especially ingredients with zero pronouns. In other words, better translation performance can be obtained if ingredients in the flow of the recipe are correctly detected. Mori et al. (2014) annotated a role label for each ingredient in a monolingual recipe corpus to model the recipe flow. If the information is appropriately adapted to the machine translation process well, some problems encountered by the machine translation systems in the recipe domain can be solved.

Bentivogli et al. (2016) conducted error analysis of PBSMT and NMT with the English-German language pair. THe authors were the first to work on error analysis of NMT and also with PBSMT and tree-based statistical machine translation in which they analyzed errors in several ways. The automatic evaluation metrics used in their study were BLEU and two types of modified translation error rate (TER) (Snover et al., 2006): Human-targeted TER and Multi-reference TER. For analysis of linguistic errors, three error categories were used: morphology errors, lexical errors and word order errors. In terms of word order errors, they also conducted fine-grained word order error analysis in which they took part-of-speech tagging and dependency parsing into account.

Ishiwatari et al. (2016) used the same recipe corpus as we used for domain adaptation of SMT without a sentence-aligned parallel corpus. In their research, the MT system was trained only with an out-domain corpus that consisted of words related to Japanese history and the temples of shrines in Kyoto. Then, they adapted the MT system to a recipe corpus in which there were many words that did not appear in the out-domain corpus, using count-based vectors to translate unknown words. Although their method performed well in the translation of the out-domain corpus, it did not focus on recipe translation itself.

7 Conclusion and Future Work

In this paper, we proposed a new task of translating cooking recipes. We translated Japanese recipes into English using PBSMT and NMT and evaluated the outputs with BLEU and RIBES. Further, we discussed the tendency observed by studying the outputs. Each of three parts comprising a recipe (title, ingredient, and step) had its own characteristics. Title proved difficult to translate owing to a relatively large vocabulary despite its limited length. Good performance was achieved in the translation of ingredient because it is very simply written compared with title and step. In translating step, PBSMT and NMT exhibited different tendencies. Many word order errors were found in PBSMT outputs corresponding to step, resulting in a lower score for RIBES in PBSMT than in NMT.

Error analysis of the outputs was also conducted with the error classification expanded from the MQM ANNOTATION DECISION TREE. The results of the error analysis showed that the tendency of each type of errors differs according to the translation method applied. Compared with that of NMT, the output of PBSMT contained many grammatical errors. On the other hand, NMT had more substitution errors than PBSMT. NMT also tended to output target words that differ in meaning form the original source word. In addition, although the outputs of NMT were usually grammatically correct, some of them were unintelligible. Many omission errors and addition errors were found in both methods.

As our future work, we plan to tackle on the machine translation of recipe texts, taking into account the ingredients used and the order in which actions are completed in the flow of the recipe. It may be possible to solve omission errors in either or both sides using the information. To achieve that, we also need to perform machine translation without sentence-alignment, but with the whole document.

Acknowledgement

This research was (partly) supported by Grant-in-Aid for Research on Priority Areas, Tokyo Metropolitan University, “ Research on social big data. ”

References

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating Discrete Translation Lexicons into Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1557–1667.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *5th International Conference on Learning Representations (ICLR)*.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 257–267.

Aljoscha Burchardt and Arle Lommel. 2014. Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality. Technical report, QTLaunchPad.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. In *Journal of Machine Learning Research 12*, pages 2121–2159.

Peter Forbes and Mu Zhu. 2011. Content-boosted Matrix Factorization for Recommender Systems: Experiments with Recipe Recommendation. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys)*, pages 261–264.

Jun Harashima, Michiaki Ariga, Kenta Murata, and Masayuki Ioki. 2016. A Large-Scale Recipe and Meal Data Collection as Infrastructure for Food Research. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 2455–2459.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 187–197.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. LONG SHORT-TERM MEMORY. In *Neural Computation 9*, pages 1735–1780.

Shonosuke Ishiwatari, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2016. Instant Translation Model Adaptation by Translating Unseen Words in Continuous Vector Space. In *The 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*.

Hideki Isozaki, Tsutom Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic Evaluation of Translation Quality for Distant Language Pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 944–952.

Philipp Koehn, Franz Josef Och, and Daniel Maruc. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, and Richard Zens. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 230–237.

Hirokuni Maeta, Tetsuro Sasada, and Shinsuke Mori. 2015. A Framework for Procedural Text Understanding. In *Proceedings of the 14th International Conference on Parsing Technologies (IWPT)*, pages 50–60.

Shinsuke Mori, Hirokuni Maeta, Yoko Yamakata, and Tetsuro Sasada. 2014. Flow Graph Corpus from Recipe Texts. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 2370–2377.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 138–145.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *In Advances in Neural Information Processing Systems 27 (NIPS)*, pages 3104–3112.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling Coverage for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180.

Yoko Yamakata, Shinji Imahori, Yuichi Sugiyama, Shinsuke Mori, and Katsumi Tanaka. 2013. Feature Extraction and Summarization of Recipes using Flow Graph. In *Proceedings of the 5th International Conference on Social Informatics (SocInfo)*, pages 241–254.

Michiko Yasukawa, Fernando Diaz, Gregory Druck, and Nobu Tsukada. 2014. Overview of the NTCIR-11 Cooking Recipe Search Task. In *Proceedings of the 11th NTCIR Conference (NTCIR-11)*, pages 483–496.