Preliminary Experiments toward NMT on E-commerce Product Titles

Aizhan Imankulova⋆† Koji Murakami⋆

⋆Rakuten Institute of Technology, Rakuten Inc., †Tokyo Metropolitan University
{ar-aizhan.imankulova, koji.murakami}@rakuten.com

1 Introduction

E-commerce products sales are dramatically rising around the world, so is selling e-commerce products abroad. However, to successfully sell e-commerce products abroad, overcoming the language barrier becomes one of the important steps. Therefore, machine translation could be a solution in translating a great amount of e-commerce products texts. Along with the development of neural machine translation (NMT) [1, 7, 9], the accuracy of machine translation has been increasing, especially in academia, which often uses data with controlled language.

Although our goal is to explore the possibility of using NMT on e-commerce product titles, e-commerce products data are different from those used in academia. Below are the differences between Rakuten Ichiba Japanese-English parallel e-commerce product and academia-wise data: 1) E-commerce product data were created by individual stores that have translated and registered text about their products. However, some Japanese sentences were translated using machine translation tools without any proofreading. This resulted in low-quality, noisy parallel data with mistranslated proper names and erroneous grammar which required transcreation. 2) The amount of created parallel data in Rakuten Ichiba is small in comparison to the existing Japanese monolingual e-commerce product data. 3) E-commerce product data are divided into many categories and each category has a different set of vocabulary. This kind of parallel data include a wide range of products, proper names, and descriptions which leads to the data sparseness problem with too many word types, especially in NMT. All of these factors adversely affect the performance of a NMT.

We compare several Japanese-English corpora which are used in academia (NTCIR, ASPEC, Tanaka) to train NMT systems and Rakuten Ichiba parallel e-commerce product corpus (Rakuten), which is a concatenation of all available Rakuten Ichiba parallel data from different categories. Table 1 shows the ratio of the number of word types to the number of total tokens in the Japanese-English parallel corpora. The higher the ratio, the more difficult to train an NMT model using given parallel corpus.

Table 1: Ratio for Japanese-English parallel corpora.

Corpus	# of sent	Ratio for Ja	Ratio for En
NTCIR	1,169,201	0.002	0.005
ASPEC	3,008,500	0.004	0.011
Tanaka	148,835	0.018	0.016
Rakuten	1,228,207	0.023	0.025

Under these restricted conditions, it is necessary to consider how NMT should be applied to translate e-commerce product titles. Therefore, we investigate the following questions:

- Unknown words: how to handle low-frequency words in e-commerce product data;
- Granularity: if one general NMT model could be enough for all e-commerce categories or if we should train an NMT model for each category;
- Data selection: select data with better quality to train an NMT model for e-commerce products;
- Data augmentation: how to use monolingual data in order to further improve NMT for e-commerce product titles.

2 Experimental Setup

2.1 Translation Model

For all experiments, we translate from Japanese into English.

We used an open source OpenNMT toolkit[1] described by Klein et al. [6] for experiments. We used recommended methods by Denkowski and Neubig [3] such as Byte Pair Encoding (BPE) [10] and annealing Adam optimization [5]. Adam has a maximum step size of 0.0002. A bi-directional encoder and decoder with a single LSTM layer have 1,024 units and for word representations we used 512 units.

For evaluation we used BLEU [8] and RIT scoring system [11]. The latter uses external vocabulary such as product information registered in a US E-commerce company Rakuten.com[2] to calculate how

[1] http://opennmt.net
[2] https://www.rakuten.com

Table 2: Data statistics for 12 category.

Dataset	# of sentence pairs			Word frequency for Japanese			Word frequency for English		
	Train	Dev	Test	# of types	1	2	# of types	1	2
Breadmaker	500	91	100	1,076	423	171	1,091	535	146
Microwave	1,065	199	200	1,825	704	257	1,948	905	271
Pendant	140,390	1,000	1,000	72,193	39,409	9,811	60,392	35,238	7,595
Rice cooker	3,049	600	600	3,480	1,461	594	4,370	2,246	787
Shampoo	32,368	1,000	1,000	14,731	5,434	2,089	11,901	4,640	1,621
Tools	177,372	2,890	2,900	85,089	44,190	11,586	71,337	40,423	9,078
Shoes	414,316	2,000	2,000	192,080	87,571	30,077	185,211	100,383	28,996
Skirt	52,007	1,000	1,000	48,544	29,030	6,263	36,955	22,793	4,134
Socks	34,356	1,000	1,000	25,748	11,707	3,958	20,698	9,787	3,035
Tops	550,156	2,000	2,000	264,140	148,776	35,502	239,347	158,349	27,388
Clothes	1,050,835	6,000	6,000	427,737	231,182	60,239	411,161	259,819	53,307
Rakuten	1,228,207	8,890	8,900	479,034	259,430	67,345	457,342	289,244	59,113

Table 3: BLEU and average RIT scores of Rakuten word-level and BPE-level *NMT models* on Rakuten test set.

NMT model	BLEU	RIT score
Rakuten all Word	53.77	6.83
Rakuten all BPE	**60.67**	**6.90**

many words have appeared in the equivalent categories. Therefore, the sentence-level RIT score (RIT score) is a weighted sum of precision of all correct n-grams (N=1,2,3) in the external vocabulary. It was calculated as: RIT score $= 4 * uni + 9 * bi + 3 * tri$.

We tokenized English sentences using NLTK script and removed non-letter characters. For Japanese sentences, we removed meta-tags and used MeCab 0.996 with the mecab-ipadic-NEologd[3] dictionary for word segmentation. We eliminated the sentence pairs exceeding 50-word length, with length ratio bigger than 3 and duplicated pairs. For data selection, we used RIT score.

2.2 Dataset

We experimented with Japanese-English translation using in-house e-commerce product titles, which are spread across different categories with different size. We chose 9 separate categories to experiment with: Breadmaker, Microwave, Pendant, Rice cooker, Shampoo, Shoes, Skirt, Socks, and Tops. In our case, data of the Rakuten e-commerce product titles have a tree structure: each leaf represents one category; nodes represent combined categories (Table 2, Tools and Clothes), and root contains data from all listed categories (Table 2, Rakuten). Tools contain data from Breadmaker, Microwave, Pendant, Rice cooker, and Shampoo. Clothes contain data from Shoes, Skirt, Socks, and Tops. These make up for the 12 datasets we experimented on. For development and test sets, we randomly sampled sentences from each dataset that have RIT score ≥ 3 (for Breadmaker, which has little data, we set RIT score ≥ 2.5) in order to calculate BLEU scores on more reliable data. Table 2 shows the data statistics after preprocessing for 12 datasets, which we used in our experiments.

[3] https://github.com/neologd/mecab-ipadic-neologd

3 NMT on E-commerce Product Titles

3.1 Handling Unknown Words

E-commerce product data have many low-frequency words which lead to the problem of large vocabularies for NMT. Table 2 shows the number of word types of 1 and 2 frequency. Up to 66% of word types occur only once in training corpora. Recently, NMT systems trained on sub-words are widely used to deal with the data sparseness problem.

We examined the impact of sub-words [10] on e-commerce NMT model and compared the results with the output of NMT model trained on word-level. We trained BPE models for each language using Rakuten data, setting BPE merge operations to 16K for each language. Then we tokenized the data from each category using pre-trained BPE model. For the word-level NMT model, we limited the vocabulary to the top 50K source words and 50K target words by frequency. We set others as an unknown token $<unk>$.

As shown in Table 3, the Rakuten BPE-level model display better performance than the Rakuten word-level model on BLEU score (+6.8 BLEU) and on average RIT score (+0.07 RIT score). For that reason, we decided to train all models on BPE-level[4].

3.2 Granularity

In this section, we investigate how granular a translation model should be to effectively translate data from each category. Table 2 shows that the size of training data are too small for some categories, especially for Breadmaker, Microwave and Rice cooker.

[4] From now on, we do not include the word "BPE" to the names of models for the sake of simplicity.

Table 4: BLEU scores of *NMT models* on 12 test data. **Bold**: best results; Underlined : second best results.

Test data	*Pendant all*	*Rice cooker all*	*Shampoo all*	*Tops all*	*Tools all*	*Clothes all*	*Rakuten all*
Breadmaker	7.41	5.93	4.73	6.61	32.66	10.04	**45.31**
Microwave	4.15	1.24	3.03	6.68	17.95	12.04	**43.58**
Pendant	**61.72**	0.00	0.63	7.91	59.79	12.83	46.19
Rice cooker	4.91	33.80	3.71	1.76	**47.59**	7.47	30.99
Shampoo	7.43	4.12	**63.76**	4.40	55.58	6.53	28.18
Tools	19.99	10.60	19.11	4.73	**58.73**	8.96	37.21
Shoes	5.08	0.15	0.43	16.45	5.83	62.15	**64.27**
Skirt	5.92	0.00	0.00	31.89	6.63	50.39	**51.59**
Socks	5.84	0.63	0.88	17.23	5.65	54.12	**56.66**
Tops	7.63	0.00	0.30	**62.40**	6.68	54.53	52.47
Clothes	5.99	0.22	0.48	32.62	6.16	**60.34**	56.58
Rakuten	10.14	3.04	4.14	25.12	17.36	44.62	**60.57**

Table 5: Number of sentence pairs of selected and augmented training data.

Training data	Selected	Augmented
Tools	132,192	175,687
Clothes	807,307	1,034,577
Rakuten	939,494	1,207,445

Table 6: Comparison of average RIT scores. orig: the original English sentences; *all*: output of NMT models trained on all data; *sel*: output of NMT models trained on the selected data.

Additional data	# of sent	orig	*all*	*sel*
Tools	45,155	4.12	4.29	**5.48**
Clothes	242,989	4.56	4.72	**5.17**
Rakuten	288,677	4.68	4.86	**5.36**

Rakuten Ichiba contains around 30K categories [2]. It would be nearly impossible to create a translation model for each category; however, in case of an insufficient volume of domain-specific data, adding generic content may help to improve the quality of NMT. Therefore, we concatenate similar datasets to 1) increase the size of training data and 2) to decrease the amount of created NMT models to translate data from each category. To investigate its effect, we trained 4 fine-granular models for different categories with different size of training data from Table 2: *Rice cooker all, Shampoo all, Pendant all* and *Tops all*. We also trained medium-granular *Tools all* and *Clothes all* models, and a coarse-granular *Rakuten all* model using all training data. Then we compared the ability of each model to translate the data from each category.

Table 4 shows the BLEU scores of each NMT model on test data of each category. All models demonstrate the best results on their own in-domain data (test data with the same name), except fine-granular *Rice cooker all* model, which was trained on very small data, and fail on translating out-of-domain data. On the other hand, medium-granular *Tools all* and *Clothes all* models show the best and the second best results on in-domain and sub-in-domain (which training data was included in medium-granular training data) datasets. A coarse-granular *Rakuten all* model outperforms *Tools all* on 2 (Breadmaker and Microwave) and *Clothes all* on 3 (Shoes, Skirt and Socks) sub-in-domain datasets. From this point onwards, we experimented with medium-granular and coarse-granular models only.

Table 7: BLEU scores of *NMT models* trained on the selected and augmented data.

Test data	*Tools*		*Clothes*		*Rakuten*	
	sel	*aug*	*sel*	*aug*	*sel*	*aug*
Tools	**57.32**	36.04	1.95	10.98	36.47	37.24
Clothes	2.60	7.47	**60.79**	57.06	57.59	57.12
Rakuten	9.41	22.85	29.70	47.97	**60.98**	60.93

Table 8: Average RIT scores of *NMT models* on in–domain test data.

Test data	orig	*all*	*sel*	*aug*
Tools	6.12	6.17	6.32	**6.73**
Clothes	6.87	6.90	6.89	**7.01**
Rakuten	6.88	6.90	6.97	**7.00**

3.3 Data Selection

The quality of the training data plays an important role in training NMT systems. Therefore, selecting high-quality data from a noisy parallel corpus [4] is considered to be one of the solutions. In this section, we applied data selection from training data and its contribution to the quality of translation for Tools, Clothes, and Rakuten dataset from Table 2. For that purpose, we sampled from these training datasets only sentence pairs with RIT score ≥ 3 and trained *Tools sel*, *Clothes sel* and *Rakuten sel* models using the selected data. The sizes of the selected training sets are shown in Table 5 (Selected). Development and test sets are the same as in Table 2.

The results are shown in Table 7 (columns: *sel*). Compared to the results of NMT models trained on all data (*all*) from Table 4, NMT models trained on

Table 9: Example from Rakuten test data translated by *NMT models*.

Model	Model output
Source sentence	楽天 大 感謝祭 ！ 小花 リボンシフォンシャツブラウス
Original target sentence	rakuten great thanksgiving ! florets ribon chiffon shirt blouse
Rakuten all Word	rakuten great thanksgiving ! flower <*unk*>
Rakuten all	rakuten great thanksgiving ! florets ribon chiffon shirt blouse
Rakuten sel	rakuten great thanksgiving ! floret ribon chiffon shirt blouse
Rakuten aug	rakuten great thanksgiving ! florets ribbon chiffon shirt blouse

the selected data (*sel*) performed slightly better on an in-domain case and much worse on out-of-domain cases, except for *Rakuten sel* model.

3.4 Data Augmentation

Rakuten Ichiba has a great amount of Japanese monolingual data. In this section, we investigate how to effectively use Japanese monolingual data to further improve the quality of NMT models for e-commerce product titles. For that purpose, we used *all* and *sel* models to translate in-domain Japanese sentences from Tools, Clothes, and Rakuten training data (Table 2) which have RIT score < 3. Table 6 shows the size of obtained pseudo-parallel data and the average RIT scores for the original English sentences (orig) and for outputs of *all* and *sel* models. In all cases, outputs of *sel* models are better than that of *all* and original target sentences. Furthermore, we selected sentences from the pseudo-parallel data to use as additional data to the selected training data from Table 5 (Selected). Sentences with the highest RIT score among the outputs of *orig*, *all*, and *sel* were kept, and sentences with the highest RIT score < 3 were eliminated. The sizes of the obtained augmented training data are shown in Table 5 (Augmented). Finally, we trained NMT models using the created augmented training data (*aug*).

BLEU scores of *aug* models are shown in Table 7 (columns: *aug*). Compared to the results of *sel* models, *aug* models are worse for in-domain datasets, however, they demonstrate better results for out-of-domain datasets.

4 Discussion

Table 8 illustrates the calculated average RIT scores for in-domain data, where we can see that all NMT outputs are better than the original target sentences, which we could not evaluate using BLEU. Also, in contrast to BLEU results in the previous section (Section 3.4), we can conclude that *aug* models outperform all other models. We assume that the reasons of such discrepancy between these scores are: 1) NMT models are trying to recreate original data, so they do not correlate with RIT score at some parts; 2) RIT score cares more about how many words, which appear in the equivalent categories on Rakuten.com, are contained in each sentence, while BLEU uses the original data (noisy pair) as the reference; 3) BLEU evaluates from the originally incorrect translation title, so the "correct" NMT outputs are considered "wrong".

Table 9 shows an example of original and translated sentences from Rakuten test set. The original target translation of the Japanese word "リボン" is "ribon", which is the erroneous translation. *Rakuten all Word* model output <*unk*> word translating "リボンシフォンシャツブラウス" (Section 3.1). *Rakuten all* and *Rakuten sel* models, which were trained on BPE-level, translated all words, however, output an erroneous translation such as "ribon". Finally, *Rakuten aug* correctly translated it to "ribbon".

5 Conclusion and Future Work

In this research, we have explored the possibility of using NMT on e-commerce product titles and demonstrated the effectiveness of using BPE, data selection, and data augmentation methods.

As part of future work, we are planning to improve the evaluation, the data selection, and the data augmentation methods. We would like to also adapt domain adaptation techniques.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] A. Cevahir and K. Murakami. Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant. In *COLING 2016: Technical Papers*, pp. 525–535, 2016.

[3] M. Denkowski and G. Neubig. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pp. 18–27, 2017.

[4] A. Imankulova, T. Sato, and M. Komachi. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *WAT2017*, pp. 70–78, 2017.

[5] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[6] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. Opennmt: Open-source toolkit for neural machine translation. In *Proc. ACL*, pp. 67–72, 2017.

[7] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, pp. 1412–1421, 2015.

[8] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318, 2002.

[9] R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data. In *ACL*, pp. 86–96, 2016.

[10] R. Sennrich, Barry Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *ACL*, pp. 1715–1725, 2016.

[11] 村上浩司, 須藤清, 新里圭司. 参照文を用いない暫定的な翻訳評価と翻訳辞書作成ツールの開発. 言語処理学会第 23 回年次大会 (NLP2017), pp. 1188–1191, 2017.