TMU Japanese-Chinese Unsupervised NMT System for WAT 2018 Translation Task

Longtu Zhang and **Yuting Zhao** and **Mamoru Komachi**
Tokyo Metropolitan University
Graduate School of System Design
6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
{zhang-longtu, zhao-yuting}@ed.tmu.ac.jp komachi@tmu.ac.jp

Abstract

This paper describes the unsupervised neural machine translation system of Tokyo Metropolitan University for the WAT 2018 translation task, focusing on Chinese–Japanese translation. Neural machine translation (NMT) has recently achieved impressive performance on some language pairs, although the lack of large parallel corpora poses a major practical problem for its training. In this work, only monolingual data are used to train the NMT system through an unsupervised approach. This system creates synthetic parallel data through back-translation and leverages language models trained on both source and target domains. To enhance the shared information in the bilingual word embeddings further, a decomposed ideograph and stroke dataset for ASPEC Chinese–Japanese Language pairs was also created. BLEU scores of 32.99 for ZH-JA and 26.39 for JA-ZH translation were recorded, respectively (both using stroke data). [1]

1 Introduction

Neural machine translation (NMT) (Bahdanau et al., 2014; Cho et al., 2014; Sutskever et al., 2014) systems have achieved great success in recent years and outperform traditional statistical machine translation (SMT) (Sennrich et al., 2016a; Wu et al., 2016; Zhou et al., 2016) systems. Nevertheless, one of its major challenges has been that it is necessary for NMT models to be trained using large parallel data, meaning that they can fail when the training data is not big enough (Koehn and Knowles, 2017; Isabelle et al., 2017). Unfortunately, the lack of large parallel corpora is a practical problem for the vast majority of language pairs, and these are often non-existent for low-resource languages. On the other hand, monolingual data is much easier to find; many languages with limited parallel data still possess significant amounts of monolingual data.

Lample et al. (2018) have proposed an unsupervised NMT model that is effective on similar language pairs, such as English–French and English–German. In this work, Chinese–Japanese language pair is used because they also share a lot of characters which can be used to replace the need for bilingual dictionaries. New sub-character datasets were also created to enhance the shared information. The byte-pair encodings (BPE) (Sennrich et al., 2016c) vocabularies were shared between the two related languages by jointly trained both monolingual corpora. FastText (Bojanowski et al., 2017) was then used to generate cross-lingual embeddings. Following this, two encoder–decoder language models were trained on noisy data on either monolingual corpora, respectively. For the translation models, back-translation (Sennrich et al., 2016b) was used to handle both direc-

[1] Our team ID for the submission to this shared task (Nakazawa et al., 2018) is TMU.

tions in tandem, from source to target and from target to source (the former generates data to train the later and vice versa). The goal of this back-translation model is to generate a source sentence for each target sentence in the monolingual corpus. The loss is computed based on monolingual data in four-ways: the source and target language models, and the source and target back-translation models. Finally, the model was tested on the translation models only.

The main findings of this paper are summarized as follows:

- The effectiveness of unsupervised NMT is quite promising in Chinese–Japanese language pairs, even if the shared tokens are not as high as 95% (Lample et al., 2018).
- Enhancing the shared information between language pairs will further promote the performance of unsupervised NMT.

2 Data Preparation

Chinese and Japanese are two logographic languages that utilize structuralized strokes to form ideographs and structuralized ideographs to form characters (Japanese also has Kanas that function as phonetic letters). According to UNICODE 10.0 standard, there are 36 strokes ("—", " 丨 ", " 丿 ", "㇏", etc.) composing hundreds of ideographs [2], and further composing 90,000+ of different characters. Table 1 shows examples of Chinese characters and how strokes and ideographs compose different characters.

ASPEC–JC (Japanese Chinese language pairs) parallel corpora (Nakazawa et al., 2016) were used in the experiments. There are 672,315 sentences in training set, and 2,090 and 2,107 sentences in the development and test sets, respectively. Note that although this corpus is bilingual, it was used monolingually in the models for this task. Ideally, a larger monolingual dataset (such as Wikipedia) should be used to obtain better performance.

[2]The number depends on how to define ideographs (usually around 500+); sometimes there are standalone ideographs that can be regard as characters as well.

Character	Semantic ideograph	Phonetic ideograph	Pinyin
驰 run	马 horse	也	chí
池 pool	水(氵) water	也	chí
施 impose	方 direction	也	shī
弛 loosen	弓 bow	也	chí
地 land	土 soil	也	dì
驱 drive	马 horse	区	qū

Table 1: Examples of Chinese characters (Pinyin is the official Romanization of Chinese characters according to its pronunciation.). Note that sometimes a ideograph can also be a character itself (like "马"); some ideographs denote the semantic meaning of the character (semantic ideographs); some denote the pronunciation (phonetic ideographs). Both semantic ideographs and phonetic ideographs can be shared across different characters for similar functions, such that "驰" and "驱" both with "马" have related meanings, while characters with "也" usually pronounce similarly.

Because neither Chinese nor Japanese have natural word boundaries, MeCab (Kudo et al., 2004) was used to pre-tokenize Japanese with the IPADic dictionary, and Jieba to pre-tokenize Chinese with its default dictionary. Then, a BPE sub-word model was trained on concatenated Chinese and Japanese monolingual data with a vocabulary size of 30,000 using fastBPE [3], in order to reduce the vocabulary size and eliminate the presence of unknown words (OOV).

Further, unsupervised NMT models rely heavily on shared information between the source and target data. Therefore, to enhance this information, new ideographs and stroke datasets were created. As opposed to Zhang and Komachi (2018), who utilized three corpora for different language pairs, namely, ASPEC–JC (Japanese Chinese), ASPEC–JE (Japanese English) and Casia2015 [4] (Chinese English) to create decomposed datasets, only ASPEC–JC was chosen in this work in order to focus on the shared information between Chinese and Japanese characters. Another difference is that CHISE was used instead of CNS11643 charset

[3]https://github.com/glample/fastBPE

[4]http://nlp.nju.edu.cn/cwmt-wmt/

32nd Pacific Asia Conference on Language, Information and Computation
The 5th Workshop on Asian Translation
Hong Kong, 1-3 December 2018

Language	Word
JA-character	風 景
JA-ideograph	⿰几虫 ⿰日京
JA-stroke	⿰⿴丿乙⿰日丿⿺⿰日⿴丨フ一⿴日丨一、⿰日⿸⿰⿴⿴丨フ日一一⿰日日、一⿰日⿴丨フ一⿺丨⿴丿、
ZH-character	风 景
ZH-ideograph	⿰几⿻ ⿰日京
ZH-stroke	⿰⿴丿乙⿺丿、⿰日⿸⿰⿴⿴丨フ日一一⿰日日、一⿰日⿴丨フ一⿺丨⿴丿、
EN	landscape

Table 2: Examples of decomposition of a Japanese word "風景" and Chinese word "风景", both meaning "landscape" in English.

for the decomposition information. The CHISE Project [5] provides decomposition mappings for Unicode CJK characters using 12 Ideographic Description Characters, 394 ideographs, and 19 special symbols for "unclear" ideographs. This mapping can help create new datasets. For ideograph datasets, the CHISE mappings were used directly; for stroke dataset, the ideographs and special symbols were manually transcribed to stroke sequences in the CHISE format, and then recursively decomposed characters into strokes. The examples are in Table 2. Similarly, BPE sub-word models were trained by and applied to these stroke and ideograph datasets with a vocabulary size of 30,000.

3 Architecture Description

Three key principles underpin the approach to unsupervised neural machine translation used in this model. The design is largely based on Lample et al. (2018)'s implementation of unsupervised NMT systems.

3.1 Shared BPE Embeddings

Instead of initializing and mapping the bilingual word embeddings based on a bilingual seed dictionary and two monolingual embeddings for unsupervised NMT models (Artetxe et al., 2018), a bilingual embedding is directly trained in two steps: first, the data is segmented using relevant BPE models trained on concatenated monolingual data of character-, ideograph- and stroke-level corpora; second, relying on the shared information between these corpora, word embeddings are trained directly using fastText (Bojanowski et al., 2017). This method is used not only because finding a readily available subword level bilingual embedding is almost impossible, but also because it is found to be efficient enough to encode the shared information directly into one space.

3.2 Encoder–Decoder Language Models

Artetxe et al. (2018) designed a shared encoder for both source and target languages, while Lample et al. (2018) used two different encoders for different languages, where the weights were only shared in last layers. Here, the latter design is followed. Two encoder-decoder models are used as the language models of the source and target languages. The encoders will encode monolingual sentences into latent representations for respective decoders, and the decoders learn to decode the same sentences based on these latent representations. Random blankouts are added to the input sentences as noise to improve the quality of the language model training.

3.3 Back-Translation

The original idea of back-translation (Sennrich et al., 2016b) was to enhance the training of a single NMT model (source–target) using the output of another readily available NMT model (target–source). The difference between the back-translation in the present system and the original one is that two back-translation models are trained together with the two encoder–decoder language models. There is no readily available model, but all models in the architecture learn to encode and generate from scratch.

Specifically, for one translation direction, the forward NMT model translates the source sentences into the target sentences and the backward NMT model translates the target into the

[5] http://www.chise.org/

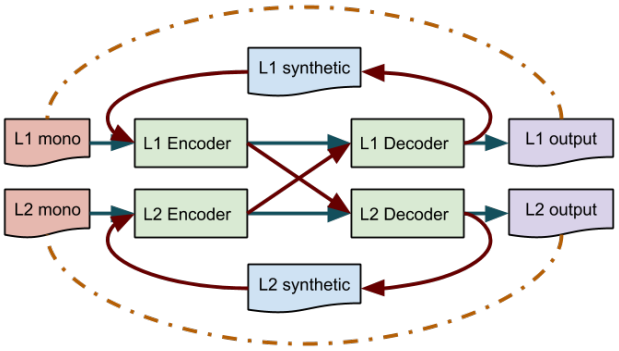

Figure 1: The architecture of unsupervised NMT model. The green arrows indicate the direction of data flow in encoder–decoder language models; the red arrows indicate the direction of data flow in back-translation models. The dotted lines are losses computed from output of the decoders and the original inputs.

source. These models generate sentences separately and then use the resulting translations to train each other. From another perspective, by combining the translation with its original sentence, a pseudo-parallel corpus is created, which is utilized to train the model to reconstruct the original sentence from its translation. More specifically, from the perspective of the two encoders, the models learn to encode both ground truth and synthetic monolingual sentences into latent representations; from the perspective of the two decoders, the models generate good sentences from latent representations, from encoders in both languages.

Figure 1 shows the illustration of the architecture of unsupervised NMT models.

4 Experiments

4.1 Setup

The baseline system in this work was an unsupervised NMT model trained on Chinese–Japanese charcater level data. This was to confirm the effectiveness of the model. Then, two experiments were completed, one for the ideograph model and the other for the stroke model. As discussed in the "Dataset Preparation" section, this is to enhance the shared information between the two languages. Understanding the importance of this in unsupervised NMT models is one of the main goals of this work.

4.2 Training

The system developed in this work was implemented based on Lample et al. (2018) [6]. Transformer (Vaswani et al., 2017) cells were used as the basic units in the encoders and decoders through the PyTorch 0.4.0 toolkit, and the numbers of both the encoder and decoder layers were set to 4. The dimension of the token embeddings and the hidden layers was set to 512. The Adam optimizer (Kingma and Ba, 2015) was used, with a learning rate of 0.0001 and a batch size of 32. A maximum length of 175 tokens per sentence for each type of dataset and a dropout rate of 0.1 was set. It is worth mentioning that the random blank-out rate was set to 0.1 in the last experiment. BLEU scores (Papineni et al., 2002) of the translation in both directions were evaluated at every epoch, and training was stopped when the scores from the last ten epochs did not improve.

5 Results

BLEU scores of 7.01 for translation from ZH-JA and 7.73 for JA-ZH were recorded, respectively, at the time of result submission. However, after bug-fixing and fine-tuning, the best scores increased to 31.99 and 25.87 respectively (both using stroke data).

The results of the baseline systems and the two experiments on sub-character level data are recorded in Table 3. The two sub-character level models outperform the character level baseline model. Moreover, the stroke model performs better than the ideograph model. The translation examples can be found in Table 4.

6 Discussion

6.1 Effectiveness of Unsupervised NMT Model

According to Lample et al. (2018), the source data and target data should share 95% of the tokens in order to make the model effective.

[6] https://github.com/facebookresearch/UnsupervisedMT

Level	Direction	BLEU
Character	JA–ZH	24.18
	ZH–JA	29.79
Ideograph	JA–ZH	25.76
	ZH–JA	32.61
Stroke	JA–ZH	**26.39**
	ZH–JA	**32.99**

Table 3: BLEU scores of 3 unsupervised NMT models on 6 translation directions. The stroke data has the best BLEU scores in both the JA–ZH and ZH–JA translation directions

However, according to the baseline and experiments in this paper, it seems that only 66.89% of shared tokens on character level data are required to generate good translations. Although the BLEU scores of both translation directions is not as good as the most basic supervised NMT model using RNNSearch (Zhang and Komachi, 2018), it is still promising, since the training data used in this work is much smaller than the original setting (Lample et al., 2018).

On the other hand, the testing output produced by the model was closely investigated. In both translation directions, translations were produced which do not use the exact terms in the reference, but instead use synonymous expressions. Several native speakers were asked to judge the grammaticality, fluency, and naturalness of the output translations, and many of the translations were thought to be better than the references. For example in Table 4, the character-level model Chinese translation "中显示" was very close to the reference "所示" semantically, and this translation was consistent in ideograph- and stroke-level models. This might be because of the encoder–decoder language models in the architecture, which successfully grasp the features of the language and express it in the translation. Therefore, if semantic-based metrics (instead of n-gram based metrics, like BLEU) could be introduced to NMT evaluation, the performance of unsupervised NMT could be better reflected in their BLEU scores.

6.2 Shared Information

Zhang and Komachi (2018) proposed that in logographic languages, sub-character decompositions could help supervised NMT models. It is found that sub-character decompositions (ideographs and strokes) are also helpful in unsupervised NMT models. This is largely due to the increase in shared information. Furthermore, since strokes are smaller units than ideographs, and they contain more shared information, the model performance is improved. For example in Table 4, despite the fact that translations produced by ideograph and stroke models were better than that of character model, stroke model was even slightly better than ideograph model. The stroke model translated Japanese "表現" into Chinese "表达", which was considered more precise than ideograph model's "名词". This might be due to the similarity of characters between Chinese and Japanese; and stroke model, as a model of finer granularity of sub-character level, successfully took advantage of this shared information.

Current unsupervised models still perform poorly on distant language pairs, so if the shared information between distant languages can be improved, unsupervised NMTs may be created for more general purpose.

7 Conclusion

The effectiveness of unsupervised NMT models is investigated for another language pair: Chinese–Japanese. The unsupervised NMT system is quite promising for similar languages, even if the monolingual dataset is not large. However, to evaluate its performance more successfully, better semantic-based metrics are required.

Acknowledgments

This work was partially supported by JSPS Grant-in-Aid for Young Scientists (B) Grant Number JP16K16117.

Type	Sentence
Reference–JA	図 3 に 「 会 」 が 固有 表現 で ある か 否 か を 判定 する 2 つ の 例文 を 示し た .
Reference–ZH	图 3 所示 的 是 2 个 关于 判断 “ 会 ” 是否是 固有 表达 的 例句 。
Charcater–JA	図 3 に 示す よう な 2 つ の 判断 について 「 会 」 が 固有 表現 で ある か どう か を 判断 する 例文 を 示す .
Character–ZH	图 3 中 显示 了 判定 “ 会 ” 是 固有 名词 还是 有 2 个 例句 。
Ideograph–JA	図 3 に 示す よう に 2 つ の 判断 「 会 」 が 固有 表現 で ある か どう か について の 例文 を 示す .
Ideograph–ZH	图 3 中 显示 了 判定 “ 会 ” 是否是 固有 名词 的 2 个 例句 。
Stroke–JA	図 3 に 示す の は , 2 つ の 判断 について 「 会 」 が 固有 表現 の 例文 で ある か どう か で ある
Stroke–ZH	图 3 中 显示 了 判定 “ 会 ” 是否是 固有 表达 的 2 个 例句 。
English	Figure 3 showed 2 example sentences of judging whether “会” is an inherent expression.

Table 4: Translation examples from 3 unsupervised NMT models in 6 translation directions. Note that even if the produced translations are not the exact words from the reference sentences, they are synonymous. Furthermore, the stroke model can generate more accurate translations semantically than the ideograph model.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. In *Transactions of the Association for Computational Linguistics (TACL)*, pages 135–146.

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 *The 2014 Conference on Empirical Methods In Natural Language Processing*, pages 1724–1734.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2476–2486.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

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

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5039–5049.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, may.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Win Pa Pa, Isao Goto, Hideya Mino, Katsuhito Sudoh, and Sadao Kurohashi. 2018. Overview of the 5th workshop on asian translation. In *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*, Hong Kong, China, December.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In

Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 311–318.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 371–376.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Lontu Zhang and Mamoru Komachi. 2018. Neural machine translation of logographic language using sub-character level information. In *Proceedings of the Third Conference on Machine Translation*, pages 17–25.

Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. 2016. Deep recurrent models with fast-forward connections for neural machine translation. In *Transactions of the Association for Computational Linguistics (TACL)*, pages 371–383.