國 立 成 功 大 學
人 工 智 慧 科 技 碩 士 學 位 學 程
碩 士 論 文

以聯合語義語音詞嵌入強化中日文神經機器翻譯

Improving Chinese-Japanese Neural Machine Translation with Joint Semantic-Phonetic Word Embedding

研 究 生：王士杰　　Student : Shih-Chieh Wang

指導教授：賀保羅　　Advisor : Paul Horton

Master Degree Program on Artificial Intelligence,
National Cheng Kung University,
Tainan, Taiwan, R.O.C.

Thesis for Master of Science Degree
July, 2021

中華民國一百一十年七月

以聯合語義語音詞嵌入強化中日文神經機器翻譯

王士杰*　　　　　賀保羅†

國立成功大學人工智慧科技碩士學位學程

摘要

中文版簡介。手動換行會自動變成下一段文字區塊。

關鍵字： 關鍵字1、關鍵字2、關鍵字3

*學生
†指導教授

Improving Chinese-Japanese Neural Machine Translation with Joint Semantic-Phonetic Word Embedding

Shih-Chieh Wang* Paul Horton†

Master Degree Program on Artificial Intelligence
National Cheng Kung University

Abstract

Add your abstract here.

Keywords: Keyword1, Keyword2, Keyword3

*Student
†Advisor

Acknowledgements

Add your acknowledgements here.

Shih-Chieh Wang

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Over the past few years, the field of neural machine translation (NMT) between Chinese and Japanese is still an unresolved problem. Recent studies in Chinese-Japanese NMT have used specific methods such as sub-character level features to improve the translation quality. This is due to the lack of parallel corpus and the difference between logogram (a character or symbol that represents a word) and alphabet (a set of letters used when writing in a language) writing systems. This research explores phonetic information as an additional feature for improving the quality of Chinese-Japanese NMT systems.

1.1 Background

NMT is a popular area of natural language processing (NLP), has been proposed by using an end-to-end model which transforms a source sentence into a latent space and decodes it directly into a target sentence [Sutskever et al., 2014,Cho et al., 2014]. The model is called the encoder-decoder model or sequence-to-sequence model, and they are widely used by large technology companies such as Google, Facebook, Microsoft, and DeepL.

1.1.1 Progress of Neural Machine Translation

The progress of NMT and NLP are inseparable. The development of models, tokenization methods, embeddings, and the solutions to less or no parallel data, all involved in the progress of NMT.

In the development of models, recurrent neural network (RNN) was first applied in NMT research [Sutskever et al., 2014,Cho et al., 2014]. After that, [Bahdanau et al., 2014] designed a mechanism called attention, which is based on RNN to address the problem of insufficient information in the latent space between encoder and decoder. The structure of the Transformer was later proposed by [Vaswani et al., 2017], which replace the RNN structure with a full attention mechanism (i.e., self-attention) to achieve better results and was widely used in NMT tasks. This paper will use both attention-based RNN model and Transformer [Bahdanau et al., 2014,Vaswani et al., 2017] as the baseline system in the experiment.

Tokenization is one of the most important parts of any NLP task. It determines how a sentence will be tokenized, and it will generate different meanings to a sentence with different algorithms. Besides word-level and character-level tokenization, several subword-level tokenization algorithms had become the mainstream. For example: Byte-Pair Encoding (BPE) [Sennrich et al., 2016b], Unigram Language Model [Kudo, 2018], WordPiece [Schuster and Nakajima, 2012], and SentencePiece [Kudo and Richardson, 2018]. This paper will utilize BPE, SentencePiece [Sennrich et al., 2016b,Kudo and Richardson, 2018] and two word-level tokenizer (*Jieba*[1] and *Janome*[2]) as tokenization methods.

The concept of embeddings, also known as distributed representations, was first proposed by [Hinton et al., 1986,Bengio et al., 2003], but was difficult to implement due to hardware limitations. With the development of parallel computing and GPU, many embedding implementations have been proposed, such as Word2Vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014], and fastText [Bojanowski et al., 2017]. The contextualized word embedding is another concept that obtains context-dependent word embedding from the whole sentence, meaning that the same word with different position can obtain different embedding through the model. The representative ones are ELMo [Peters et al., 2018] and BERT [Devlin et al., 2019]. This paper will select Word2Vec [Mikolov et al., 2013] as the tool for creating word embeddings because of its simplicity, rapidity, and convenience of analysis.

Several fields have been studied to solve the problems like low-resources and noisy parallel data in NMT tasks. Back-translation [Sennrich et al., 2016a] is a data augmentation method that uses monolingual data of the target language to generate source data and offset the imbalance between encoder and decoder. Parallel corpus filtering was examined for a large number of NMT tasks [Koehn et al., 2018], using pre-filtering rules and scoring functions to retain good sentence pairs can effectively reduce the corpus size and obtained better translation results. This paper will practice corpus filtering to retain quality training data and reduce corpus size to increase experimental efficiency.

1.1.2 Chinese-Japanese Neural Machine Translation

NMT system has gained a lot of improvement in translating between English and other languages by utilizing the techniques described in section 1.1.1. However, the improvement in translating between Chinese and Japanese is limited. The main reasons are the inadequacy of the corpus and the differences in the writing systems of Chinese, Japanese, and Western languages.

[1] https://github.com/fxsjy/jieba
[2] https://mocobeta.github.io/janome

Many studies have focused on improving the Chinese-Japanese (zh-ja) NMT system. In addition to using the methods [Imamura et al., 2018,Chu et al., 2017,Zhang et al., 2020] described in section 1.1.1, many feature engineering techniques have been proposed to utilize the features in Chinese Characters (*Hanzi*) and Japanese *Kanji*. For example, a character-level zh-ja NMT system had been improved by using radicals as character feature information [Zhang and Matsumoto, 2017]. Furthermore, the use of decomposed sub-character level information such as ideographs and strokes of Chinese characters, also improved the results [Zhang and Komachi, 2018].

1.1.3 Phonetic Information

Phonetic information is another feature that had been applied to NMT systems. [Khan and Xu, 2019] had suggested that a phonetic representation usually corresponds to semantically distinct characters or words. [Liu et al., 2019] had pointed out that phonetic information can effectively resist the homophone noises generated by typographical mistakes in Chinese sentences. Both papers had improved the performance of the NMT system between Chinese and other Western languages.

This paper attempts to use *Bopomofo* and *Hiragana* as Chinese and Japanese phonetic information to improve the performance of the zh-ja NMT system. Bopomofo also named *Zhuyin* (注音), is located in the Unicode block in the range U+3100–U+312F. It consists of 37 characters and 4 tone marks to transcribe all possible Chinese characters. Although it is the main component of Mandarin Chinese, it usually does not appear in Chinese sentences. That is, the machine loses some of the phonetic information when reading Chinese sentences. Hiragana (平仮名, ひらがな) is a component of Japanese, along with *Katakana* and Kanji. It consists of 46 base characters and is located in the Unicode block in the range U+3040–U+309F. Compared to Bopomofo, Hiragana is often found in Japanese sentences with Katakana and Kanji, forming mixed writing of Kanji and Kana (仮名交じり文). However, Hiragana disappears after forming Kanji, just like Bopomofo forms Hanzi. Therefore, the machine cannot obtain the phonetic information directly from Japanese sentences.

1.2 Objective

1.3 Related Work

Chapter 2

Method

method.

Chapter 3

Experiment and Result

experiment and result.

Chapter 4

Discussion

discussion.

Chapter 5

Conclusion

conclusion.

REFERENCES

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155.

[Bojanowski et al., 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[Cho et al., 2014] Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[Chu et al., 2017] Chu, C., Dabre, R., and Kurohashi, S. (2017). An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Hinton et al., 1986] Hinton, G. E. et al. (1986). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA.

[Imamura et al., 2018] Imamura, K., Fujita, A., and Sumita, E. (2018). Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 55–63.

[Khan and Xu, 2019] Khan, A. R. and Xu, J. (2019). Diversity by phonetics and its application in neural machine translation. *arXiv preprint arXiv:1911.04292*.

[Koehn et al., 2018] Koehn, P., Khayrallah, H., Heafield, K., and Forcada, M. L. (2018). Findings of the wmt 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739.

[Kudo, 2018] Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

[Kudo and Richardson, 2018] Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

[Liu et al., 2019] Liu, H., Ma, M., Huang, L., Xiong, H., and He, Z. (2019). Robust neural machine translation with joint textual and phonetic embedding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3044–3049, Florence, Italy. Association for Computational Linguistics.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

[Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

[Schuster and Nakajima, 2012] Schuster, M. and Nakajima, K. (2012). Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

[Sennrich et al., 2016a] Sennrich, R., Haddow, B., and Birch, A. (2016a). 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, Berlin, Germany. Association for Computational Linguistics.

[Sennrich et al., 2016b] Sennrich, R., Haddow, B., and Birch, A. (2016b). 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Zhang et al., 2020] Zhang, B., Nagesh, A., and Knight, K. (2020). Parallel corpus filtering via pre-trained language models. *arXiv preprint arXiv:2005.06166*.

[Zhang and Matsumoto, 2017] Zhang, J. and Matsumoto, T. (2017). Improving character-level japanese-chinese neural machine translation with radicals as an additional input feature. In *2017 International Conference on Asian Language Processing (IALP)*, pages 172–175.

[Zhang and Komachi, 2018] Zhang, L. and Komachi, M. (2018). Neural machine translation of logographic language using sub-character level information. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 17–25, Brussels, Belgium. Association for Computational Linguistics.