Domain Fine-tuning Narrows the Gap: HwTscSU's Submissions on WAT 2022 Shared Tasks

Yilun Liu[1], **Zhen Zhang**[2], **Shimin Tao**[1], **Junhui Li**[2], **Hao Yang**[1]
[1]2012 Lab, Huawei
[2]School of Computer Science and Technology, Soochow University, Suzhou, China
{liuyilun3,taoshimin,yanghao30}@huawei.com
zzhang99@stu.suda.edu.cn, lijunhui@suda.edu.cn

Abstract

In this paper we describe our submission to the shared tasks of the 9th Workshop on Asian Translation (WAT 2022) on NICT–SAP Task under the team name "HwTscSU". The tasks involve translations from 5 languages into English and vice-versa in two domains: IT domain and Wikinews domain. The purpose is to determine the feasibility of multilingualism, domain adaptation or document-level knowledge given very little to none clean parallel corpora for training. Our approach for all translation tasks mainly focus on pre-training NMT models on general datasets and fine-tuning them on domain-specific datasets. Due to the scarcity of parallel corpora, we collect and clean the OPUS dataset, including three IT domain corpora, i.e., GNOME, KDE4, and Ubuntu. We then train Transformer models on the collected datasets and fine-tune them on corresponding dev sets. The BLEU scores are greatly improved in comparison with other systems.

1 Introduction

Explorations on machine translation have come far since the era of neural machine translation (NMT) (Kalchbrenner and Blunsom, 2013). Owing to the incorporation of novel structures such as CNN (Gehring et al., 2017) and Transformer (Vaswani et al., 2017), modern NMT models are able to compete with human translation.

However, the performance of neural machine translation is often highly relevant to the size of available datasets. When the training datasets are small in quantity, performances of NMT models are often poor, especially for low-resource languages. Considering that it is often helpful, in such low-resource scenarios, to leverage monolingual or bilingual corpora from multiple languages and domains to boost translation quality, we collected a large amount of web-crawled datasets for training models in the task.

The Workshop on Asian Translation[1] (Nakazawa et al., 2022) is an open machine translation evaluation campaign focusing on Asian languages. WAT gathers and shares the resources and knowledge of Asian language translation to understand the problems to be solved for the practical tasks of machine translation technologies among all Asian countries. Among those tasks, we participated on NICT-SAP tasks which evaluate Hindi/Thai/Malay/Indonesian/Vietnamese ↔ English translation in two domains: IT domain (Software Documentation) and Wikinews domain (ALT). IT domain and Wikinews are two extremely low-resource domains for machine translation, especially when concerning languages such as Hindi, Thai, Malay, Indonesian and Vietnamese. Often, in these domains, there is no clean bilingual parallel corpus at all (the IT domain), or the size of available corpora is extremely small (the Wikinews).

Both two corpora contain a lot of technical terms. Moreover, some technical terms are domain-specific and do not exist in general dictionaries. Therefore, we focus on domain adaptation for translations of both IT and Wikinews domains.

In this paper, we describe our simple approach involving Transformer pre-training and fine-tuning. We first collected and cleaned rich sentence pairs from public dataset. Following Berling Lab (Park and Lee, 2021), for both NICT-SAP IT domain and ALT domain tasks we first collected public dataset from OPUS (Tiedemann, 2012) such as but not limited to GNOME, KDE4 and Ubuntu. Then we chose G-Transformer (Bao et al., 2021) as our model and pre-train the baseline with these datasets. Finally, as fine-tuning on domain-specific dataset greatly boosts translation performance in WMT evaluation (Barrault et al., 2020; Akhbardeh et al., 2021), we fine-tuned the pre-trained models on corresponding dev set officially provided for

[1] https://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2022/index.html

Proceedings of the 9th Workshop on Asian Translation, pages 59–63
October 17, 2022.

high performance for all sub-tasks. Our method obtained the new state-of-the-art results on IT-domain tasks. We ranked first place on all NICT-SAP IT domain tasks, especially achieving 10.74 improvement for English to Malay. On ALT domain tasks, we ranked first in one out of eight sub-tasks.

2 Task Description

2.1 NICT-SAP Shared Task

The NICT-SAP shared task is to translate texts between English and other five languages, that is, Hindi (Hi), Thai (Th), Malay (Ms), Indonesian (Id), and Vietnamese (Vi) in extremely low-resource conditions. The task contains two domains: IT domain and ALT domain.

The data in the Asian Language Translation (ALT) domain (Thu et al., 2016) consists of translations obtained from WikiNews which is a multilingual parallel corpus. The training, development, and test sets are provided by the WAT organizers. We filter translations that are longer than 512 tokens, resulting in fewer than 20K training sentences in all languages.

The data in the IT domain consist of translations of software documents. However, there is no clean corpus from the IT domain for training. Different from ALT domain, the WAT organizers only provide the development and test sets (Buschbeck and Exel, 2020). In this case, we collected and cleaned parallel corpora available through OPUS for training, where the domain is not fully identical with the domains of the provided dev/test sets.

The dataset sizes of two given corpora are shown in Table 1.

2.2 Evaluation Metric

We report the performance in BLEU (Papineni et al., 2002) and RIBES (Isozaki et al., 2010), which are official evaluation metrics.

3 Our Approaches

For our submissions we focus on training G-Transformer (Bao et al., 2021) on OPUS dataset from scratch and fine-tuning on dev sets. G-Transformer is developed on FairSeq (Ott et al., 2019) for document-level translation, and also supports Transformer-based sentence-level translation.

3.1 Data crawling and preprocessing

For all tasks, we pre-trained the sentence-level Transformer models on web-crawled dataset as baselines. Since the WAT organizers do not provide the training dataset for IT domain, we collect it from public dataset including GNOME, KDE4, Ubuntu,Tateoba, Tanzil, QED (Abdelali et al., 2014), tico-19, OpenSubtitles, ELRC. We download all the dataset from OPUS site and filter translations that are longer than 512 tokens. Table 2 shows the statistics of the data obtained from the site. Note that, the data obtained from GNOME, KDE4 and Ubuntu are all in the IT domain, while others are not.

3.2 Model configuration

For the NMT system, we use G-Transformer (Bao et al., 2021) to train Transformer (Vaswani et al., 2017) architecture models. We use Transformer-base as our basic model setting, which has 6 layers in both the encoder and decoder, respectively. For each layer, it consists of a multi-head attention sublayer with 8 heads. We set the max sequence length as 512 for both source and target sides. We use an effective batch size of 8192 tokens. We chose Adam (Kingma and Ba, 2015) as our optimizer, with parameters settings $\beta_1 = 0.9$, $\beta_2 = 0.98$, and warm-up steps 4000. The learning rate is set to be $5e^{-4}$ for NMT pre-training and domain fine-tuning. We set the data type to the floating point 16 for fast computation. Following Berling Lab (Park and Lee, 2021), we change the hidden layer size from 512 to 1024 and the feed forward networks from 2048 to 4096 for better performances. In both pre-training and fine-tuning, we save the checkpoints every epoch and set the early-stop patience as 10 by evaluating the loss on the dev set. Each model was trained on 2 V100 (32GB).

In preprocessing, we use Google sentencepiece library[2] to train separate SentencePiece models (Kudo, 2018) for each source-side and target-side language. Then following Berling Lab (Park and Lee, 2021), we set vocabulary size to 8,000 for English, Malaysian and Vietnamese and to 16,000 for Hindi, Indonesian and Thai. We set a character coverage to 0.995. Specifically, we only use IT domain datasets (Ubuntu, GNOME, KDE4) to train SentencePiece models.

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

Domain	Set	En-Hi	En-Th	En-Ms	En-Id	En-Vi
ALT	Train	18,088	18,088	18,088	18,087	18,088
	Test	1,018	1,018	1,018	1,018	1,018
	Dev	1,000	1,000	1,000	1,000	1,000
IT	Train	-	-	-	-	-
	Test	2,073	2,050	2,050	2,037	2,000
	Dev	2,016	2,048	2,050	2,023	2,003

Table 1: Data sizes (number of sentence pairs) for the NICT-SAP domain task provided officially after filtering.

Pair	GNOME	KDE4	Ubuntu	ELRC	TANZIL	Opensubtitles	tico-19	QED	Tatoeba
En-Hi	145,706	97,227	11,309	245	187,080	93,016	3,071	11,314	10,900
En-Th	78	70,634	3,785	236	93,540	3,281,533	-	264,677	1,162
En-Ms	299,601	87,122	120,016	1,697	122,483	1,928,345	3,071	79,697	-
En-Id	47,234	14,782	96,456	2,679	393,552	926,8181	3,071	274,581	9,967
En-Vi	149	42,782	5,056	4,273	-	3,505,276	-	338,024	5,693

Table 2: Statistics (number of sentence pairs) of parallel corpora from OPUS. The data from GNOME, KDE4, Ubuntu are IT domain.

Tasks	Pre-training	Fine-tuning
En→Hi	13.05	41.85
Hi→En	14.79	40.42
En→Th	15.81	40.44
Th→En	7.92	31.95
En→Ms	31.35	56.75
Ms→En	27.97	45.65
En→Id	42.77	59.36
Id→En	37.02	58.20
En→Vi	13.09	10.68
Vi→En	25.40	50.90

Table 3: BLEU's comparison of pre-training and fine-tuning in IT domain tasks.

4 Result

4.1 Pre-training and Fine-tuning

We pre-train the Transformer with the clean data shown in Table 2. Then we fine-tune on corresponding dev set for each sub-task. Table 3 shows the comparison of their performances in IT domain. Note that the BLEU scores are obtained by the Mosesdecoder[3] scripts rather than official results because the official would evaluate Thai language using character level BLEU. Except En→Vi, domain fine-tuning could get better performance.

4.2 NICT-SAP IT Domain Translation Task

We submitted the fine-tuned models which show the best performance. Table 4 shows the overall results on NICT-SAP IT domain. For multilingual translation, it is popular to fine-tune mBART (Liu et al., 2020) which is pre-trained on large-scale monolingual corpora in many languages. However, we simply pre-trained the models from scratch and used relatively small corpus from OPUS. Domain fine-tuning makes a huge improvement in performance and we rank first in all sub-tasks in IT domain, as shown in Table 4. After submitting the translations, we noticed that the improvement was partially due to the overlaps between the dev set and test set.

Tasks	BLEU	RIBES	Rank
En→Hi	41.70	0.74	1
Hi→En	40.20	0.73	1
En→Th	70.10	0.89	1
Th→En	31.80	0.71	1
En→Ms	56.70	0.88	1
Ms→En	45.50	0.82	1
En→Id	58.80	0.78	1
Id→En	57.20	0.78	1
En→Vi	32.70	0.68	1
Vi→En	61.50	0.84	1

Table 4: Official BLEU/RIBES scores for NICT-SAP IT domain tasks on leader-board. The rank is sorted by BLEU score.

[3] https://github.com/moses-smt/mosesdecoder

Tasks	BLEU	RIBES	Rank
En→Hi	20.30	0.74	7
Hi→En	21.30	0.76	3
En→Th	49.70	0.79	3
Th→En	16.10	0.75	3
En→Ms	43.10	0.91	3
Ms→En	38.90	0.89	3
En→Id	42.40	0.91	1
Id→En	40.00	0.89	3

Table 5: Official BLEU/RIBES scores for NICT-SAP ALT domain tasks on leader-board. The rank is sorted by BLEU score.

4.3 NICT-SAP ALT Domain Translation Task

Table 5 shows official results on NICT-SAP ALT domain. We fine-tune the pre-trained models showed in Table 3 on corresponding dev set. Although the models are not pre-trained with in-domain corpus, the performances are better than other Transformer-base models. However, there is still a gap between our models and other models which are fine-tuned from mBART (Liu et al., 2020).

4.4 Fine-tuning on Document-Level Dataset

As G-Transformer (Bao et al., 2021) is designed for document-level translation, finally we try to fine-tune the pre-trained models on the dev sets at the document-level through G-Transformer. However, fine-tuning the document-level translation model on dev sets does not achieve good performance. For example, the dev set for En↔Ms contains 210 documents. And the performance changes from 31.35 to 29.14 in BLEU and 29.97 to 33.42 on the two tasks, respectively, when moving from sentence-level fine-tuning to document-level fine-tuning. Therefore, the document-level fine-tuning is less effective than the sentence-level fine-tuning. We attribute it to two reasons. First, the number of document in dev sets is too small to properly train the new added document-level parameters. Second, with small fine-tuning set, the model is not well adopted to accept long sequences as inputs.

5 Conclusion

In this paper, we have described our translation models to the NICT-SAP translation tasks on NICT-SAP track. We first pre-train our models from scratch on the datasets from OPUS. Then we fine-tune the models on corresponding dev sets. Experimental results have shown that our model ranked firsts place for NICT-SAP IT domain tasks and achieved good performance for NICT-SAP ALT domain tasks.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The amara corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1044–1054. European Language Resources Association.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3442–3455. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55. Association for Computational Linguistics.

Bianka Buschbeck and Miriam Exel. 2020. A parallel evaluation data set of software documentation with document structure annotation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 160–169. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the*

34th International Conference on Machine Learning, pages 1243–1252. PMLR.

Hideki Isozaki, Tsutomu 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*, pages 944–952. Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Taku Kudo. 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. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation (WAT2022)*. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53. Association for Computational Linguistics.

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 of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Heesoo Park and Dongjun Lee. 2021. Bering lab's submissions on WAT 2021 shared task. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 141–145. Association for Computational Linguistics.

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew M. Finch, and Eiichiro Sumita. 2016. Introducing the asian language treebank (ALT). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2214–2218. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 6000–6010.