HW-TSC's Participation in the WAT 2020 Indic Languages Multilingual Task

Zhengzhe Yu[1], **Zhanglin Wu**[1], **Xiaoyu Chen**[1], **Daimeng Wei**[1], **Hengchao Shang**[1], **Jiaxin Guo**[1], **Zongyao Li**[1], **Minghan Wang**[1], **Liangyou Li**[2], **Lizhi Lei**[1], **Hao Yang**[1], **Ying Qin**[1],
[1]Huawei Translation Service Center, Beijing, China
[2]Huawei Noah's Ark Lab, Hong Kong, China
{yuzhengzhe,wuzhanglin2,chenxiaoyu35,weidaimeng,
shanghengchao,guojiaxin1,lizongyao,wangminghan,
liliangyou,leilizhi,yanghao30,qinying}@huawei.com

Abstract

This paper describes our work in the WAT 2020 Indic Multilingual Translation Task. We participated in all 7 language pairs (En↔Bn/Hi/Gu/Ml/Mr/Ta/Te) in both directions under the constrained condition—using only the officially provided data. Using transformer as a baseline, our Multi→En and En→Multi translation systems achieve the best performances. Detailed data filtering and data domain selection are the keys to performance enhancement in our experiment, with an average improvement of 2.6 BLEU scores for each language pair in the En→Multi system and an average improvement of 4.6 BLEU scores regarding the Multi→En. In addition, we employed language independent adapter to further improve the system performances. Our submission obtains competitive results in the final evaluation.

1 Introduction

This paper describes our work in the WAT 2020 Indic Multilingual Translation Task (Nakazawa et al., 2020). Our team (Team ID: HW-TSC) participated in all seven language pairs (En↔Bn/Hi/Gu/Ml/Mr/Ta/Te) by training Multi→En and En→Multi multilingual translation models. Based on previous works, we mainly focus on exploiting fine-grained data filtering and domain data selection techniques to enhance system performance. Multistep filtering is conducted to sort out the high-quality subset for training. Several other strategies, including Back-Translation (Edunov et al., 2018), Tagged Back-Translation (Caswell et al., 2019), Joint Training (Zhang et al., 2018), Fine-Tuning (Sun et al., 2019), Ensemble and Adapter Fine-Tuning (Bapna et al., 2019) are employed and tested in our experiments. sacreBLEU (Post, 2018) is used to evaluation the system performance.

2 Data

This section describes the size and source of the dataset as well as our data filtering techniques.

2.1 Data Source

We use PM India and CVIT-PIB datasets for the training of the 7 language pairs. PM India (dataset size: 255k) is a high-quality alignment corpus already being filtered while CVIT-PIB (dataset size: 478k) contains mainly multilingual hybrid data that requires alignment. Table 1 shows the data distribution of 7 language pairs. Apart from the two multilingual datasets, 700k monolingual data provided by the organizer is also used in our experiments.

2.2 Data Pre-processing

Our data pre-processing procedures include:

- Convert full-width text to half-width text;
- De-duplicate the data;
- Remove text which the source or target side is empty;
- Perform language identification (Joulin et al., 2016b,a) on the dataset and remove texts with undesired tags;
- Employ multilingual sentencepiece model (SPM) with regularization (Kudo and Richardson, 2018; Kudo, 2018) for all language pairs;
- Filter the corpora with fast-align (Dyer et al., 2013);
- Delete extra-long sentences with more than 100 sub-tokens.

It should be noted that we trained the hybrid SPM in conjunction with English and 7 other indic languages. In order to ensure that each language

Proceedings of the 7th Workshop on Asian Translation, pages 92–97

has equivalent vocabulary size, we averaged the training data for each language when training SPM, namely, over-sampling low resource languages. In consideration of the small dataset size, we did not perform strict data cleansing strategy at the beginning but merely observed poor alignment results regarding the CVIT-PIB dataset compared with the PM India dataset. So we further use Fast-align on the dataset to improve the data quality, although a quite large amount of data was removed during this process.

2.3 Data Selection

During the experiment, we observed that the system trained only with the PM India dataset performed better than the system trained jointly with PM India and CVIT-PIB datasets. We believe the reason is that the domain of the PM India dataset is much more align with that of the test set. So we further filtered the CVIT-PIB dataset to select the "in-domain" data. Inspired by curriculum learning ideas (Wang et al., 2019), we exploited the Static Data Selection strategy. We regarded the PM India dataset and the dev set as "in-domain" and tried to sort out "in-domain" data in the CVIT-PIB dataset with a trained classifier. First we use PM India dataset combine CVIT-PIB dataset to train a base model. Then we sampled a fixed number of sentences (e.g. 30k) from the source side (EN) of the PM India dataset plus the dev sets and labeled them as IN-domain. Then we sampled the same amount of sentence from the CVIT-PIB dataset and labeled then as OUT-domain. We trained a Fasttext (Bojanowski et al., 2017) classifier on the sampled dataset to score sentences in the CVIT-PIB with the classification probability of $P(y = InDomain|x)$ to retrieve the top-k bi-text pairs. Where k is set to 5k in our experiment. Not that even the probability score is lower than 0.5, we still kept the sentence pairs as long as their ranks are within the top-k. Then we used the "in-domain" CVIT-PIB data and PM India data to fine-tune the base model we trained and observed better performances.

From the experiment, we find that data selection is quite effective compared to using entire CVIT-PIB dataset on both En→Multi and Multi→En.

3 System Overview

This section describes our system used in the WAT 2020 Indic Multilingual Translation Task. The following introduced strategies are tested sequentially and our experimental results regarding each strategy is listed in each part.

Language	PM India	CVIT-PIB	Mono
En-Bn	26K	48K	114k
En-Gu	44K	29K	121k
En-Hi	52K	195K	155k
En-Ml	29K	31K	80k
En-Mr	31K	80K	116k
En-Ta	35K	87K	87k
En-Te	35K	5K	109k
Total	255K	478K	

Table 1: Data source of Indic Multilingual Translation Task

3.1 Model

Transformer (Vaswani et al., 2017a) has been widely used for machine translation in recent years, which has achieved good performance even with the most primitive architecture without much modifications. Therefore, we choose to start from Transformer-Deep and consider it as a baseline, which is the model with deeper encoder version proposed in (Sun et al., 2019), with 35 encoder layers and 3 docoder layers, 512 hidden size and 4096 batch size. We used the Adam optimizer (Kingma and Ba, 2014) with $\beta 1 = 0.9$, $\beta 2 = 0.98$. We used the same warmup and decay strategy for learning rate as (Vaswani et al., 2017b), with 4,000 warmup steps. During training, we employed label smoothing value of 0.1 (Szegedy et al., 2016). For evaluation, we used beam search with a beam size of 4 and length penalty $\alpha = 0.6$ (Wu et al., 2016). Our models are implemented with THUMT (Zhang et al., 2017), and trained on a platform with 8 V100 GPUs. We train models for 100k steps and average the last 6 checkpoints for evaluation.

3.2 Multilingual Strategy

For this Indic Multilingual Translation Task, we exploited different multilingual training strategies regarding multilingual training on the basis of transformer. We trained the hybrid SPM model in conjunction with English and 7 indic languages as the shared word segmentation system for all language pairs. We kept the vocabulary within 30k, which included all tokens of all 8 languages (En/Bn/Hi/Gu/Ml/Mr/Ta/Te). Two mainstream methods about multilingual training are available: two models with En→Multi and Multi→En separately and a mono Multi→Multi model. We tested

	En→Bn	Bn→En	En→Gu	Gu→En	En→Hi	Hi→En	En→Ml	Ml→En	En→Mr	Mr→En	En→Ta	Ta→En	En→Te	Te→En
En→Multi / Multi→En	**15.61**	**16.97**	**10.6**	**18.40**	**19.03**	**18.60**	**3.55**	**12.58**	**8.03**	**15.42**	4.59	**12.56**	3.63	**16.97**
Multi→Multi	14.66	16.31	10.25	15.61	18.36	18.68	3.13	11.37	7.97	13.59	4.59	10.11	3.63	16.31

Table 2: BLEU score of Multilingual strategy for En→Multi, Multi→En and Multi→Multi. for line En→Multi / Multi→En, En→XX inferenced by En→Multi model and XX→En inferenced by Multi→En model

both methods by training three models with the PM India dataset. The results listed in Table 2 shows that Multi→Multi performs worse than the other strategy and thus we only consider the separate En→Multi and Multi→En models in the following experiments. We believe that a Multi→Multi model contains too many languages pairs (14 in this case) so conflicts and confusions may occur among language pairs in different directions. Regarding our En→Multi model, we added tags "2XX" (XX indicates the target language, e.g. 2bn) at the beginning of the source sentence for each bilingual sentence pair, a strategy used in (Johnson et al., 2017). Then we mixed all data for training.

Due to the limitations of the multilingual translation model, once the model is trained, other further training methods (fine-tuning, etc.) might be difficult to improve the performance of the model, so we will introduce the fine-tuning method we use below to improve each language pair without affecting the performance of others.

3.3 Data Augmentation

Our experiment demonstrates that simply combining all bilingual data altogether does not produce gains to model quality, as described in the previous section as well in Table 3 and Table 4 that adding the whole CVIT-PIB dataset negatively influenced the model performance with respect to most of the language pairs. Two strategies regarding data augmentation are leveraged:

- Data filtering: To address the poor-quality CVIT-PIB data, as we introduced in the previous section, we used fast-align to further filter the dataset despite a significant reduction of the training data size. This strategy works as we can see from Table 3 and Table 4 that the BLEU scores of several languages achieve increases of more than 0.5 points.

- Domain transfer: Static Data Selection is leveraged to filter "in-domain" data. As we introduced in the previous section, we regarded the domain of PM India dataset more align with the test set and CVIT-PIB more like "out-of-domain" data. We use the techniques described before to select more "in-domain" data in the CVIT-PIB dataset and combined the filtered CVIT-PIB data and PM India data to fine-tuning the models. Another key issue constraining the system performance is the imbalanced data sizes for each language. In the ideal setting, the amount of data in each language is supposed to be equal. En-Hi is regarded as a high-resource language in this experiment as the size of its training data far exceeding that of other language in this task. Therefore, we over-sampled the training data of other low-resource language data (Arivazhagan et al., 2019) to ensure their training data size are balanced. This strategy led to a huge improvement in BLEU scores, as shown in Table 3 and Table 4: an average improvement of 2.6 BLEU scores for En→Multi model and an average improvement of 4.6 BLEU scores for Multi→En model.

3.4 Ensemble

We trained four models in each direction with different seeds and ensembed these models. Ensemble also contributed to the increase of BLEU scores in our experiment. Particularly, we observed an 2.7 improvement of BLEU with regard to En→Gu.

3.5 Language independence Adapter Fine-tuning

Previous works demonstrate that fine-tuning a model with in-domain data could effectively improve the performance of model. However, due to the limitations of the multilingual translation model, once the model is trained, when fine tuning one of the language pairs, the performance of others will go worse. Thanks to the finding of Adapter (Bapna et al., 2019), we are able to fine-tune each language pair without impacting the performance of others. In the experiment, we set the adapter size to 128 and fine-tuned the model on the dev set for each language pair in En→Multi with 3,000 tokens per batch for one epoch, successfully achieving 1.02 of BLEU improvements on En→Ta and 1.8 of BLEU improvements on En→Te. However, we do not gain any improvement for other language pairs.

	En→Bn	**En→Gu**	**En→Hi**	**En→Ml**	**En→Mr**	**En→Ta**	**En→Te**	Avg
PM India Data	15.61	10.6	19.03	3.55	8.03	4.59	3.63	
+ CVIT-PIB Data	7.27 (-8.34)	11.07 (+0.74)	15.98 (-3.05)	2.87 (-0.68)	8.95 (+0.92)	4.39 (-0.2)	2.05 (-1.58)	-1.74
+ Fast-align	10.71 (+3.44)	10.74 (+0.33)	16.63 (+0.65)	3.23 (+0.36)	9.39 (+0.44)	4.17 (-0.22)	3.50 (+1.45)	+0.92
+ Domain Transfer	18.30 (+7.59)	11.9 (+1.16)	22.12 (+5.49)	4.10 (+0.87)	10.54 (+1.15)	5.64 (+1.47)	4.22 (+0.72)	+2.64
+ Ensemble	18.47 (+0.17)	14.64 (+2.74)	23.13 (+1.01)	4.57 (+0.47)	11.32 (+0.78)	6.17 (+0.53)	4.64 (+0.42)	+0.87
+ Adapter Fine-tuning	-	-	-	-	-	7.19 (+1.02)	6.49 (+1.85)	+1.44
2020 Submission	**19.64**	**14.66**	**24.48**	4.60	**11.52**	**7.21**	**6.93**	
Official Baseline	15.03	9.73	13.96	6.32	8.84	4.33	5.20	

Table 3: The experimental result of En→Multi

	Bn→En	**Gu→En**	**Hi→En**	**Ml→En**	**Mr→En**	**Ta→En**	**Te→En**	Avg
PM India Data	16.97	18.40	18.60	12.58	15.42	12.56	16.97	
+ CVIT-PIB Data	14.98 (-1.99)	19.39 (+0.99)	18.97 (+0.37)	14.59 (+2.01)	17.13 (+1.71)	14.17 (+1.61)	11.74 (-5.23)	-0.08
+ Fast-align	15.89 (+0.91)	21.26 (+1.87)	22.70 (+3.73)	14.26 (-0.33)	18.61 (+1.48)	14.58 (+0.41)	11.63 (-0.11)	+1.14
+ Domain Transfer	21.52 (+5.63)	27.33 (+6.07)	26.96 (+4.26)	18.90 (+4.64)	22.88 (+4.27)	16.12 (+1.54)	17.32 (+5.69)	+4.58
+ Ensemble	22.99 (+1.47)	29.91 (+2.58)	28.26 (+1.3)	20.63 (+1.73)	23.84 (+0.96)	19.98 (+3.84)	18.74 (+1.42)	+1.90
2020 Submission	**23.38**	**30.26**	**28.51**	**20.87**	**24.05**	**20.16**	**19.03**	
Official Baseline	21.80	24.48	25.68	15.46	21.15	18.37	15.44	

Table 4: The experimental result of Multi→En

Due to time restriction and heavy workload, we did not fine-tune the Multi→En model.

One should noticed that whether En→Multi or Multi→en are multilingual translation models, fine-tuning cannot be used usually, because the improvement of a one language pair and will hurt others' performance. Through adapter fine-tuning, we can guarantee that fine-tuning one language pair does not affect the quality of other language pairs in the model.

4 Result

This section presents the experimental results for each direction of all three language pairs in Table 3 and Table 4, where the contribution of strategies introduced in previous sections are listed in each row. In this competition, among the 14 directions of the 7 Indic language pairs (En↔Bn/Hi/Gu/Ml/Mr/Ta/Te), our submission ranks the first place in 13 language directions while En→Hi even achieve an improvement of 10.5 points in term of BLEU when comparing with the baseline.

5 Analysis

Here are several findings worthy of sharing during our experiments:

- In this experiment, we also used Back Translation and Tagged Back-Translation, but only saw undesired results. The performance of most language pairs became even worse and the BLEU scores of some languages even reduced more than 10 points. We think data domain may be responsible for the BLEU reduction, similar to the situation when we adding CVIT-PIB data for training but only gained worse results. Therefore we give up Back-Translation in our experiment.

- Both En→Multi and Multi→En models are better than the Multi→Multi model, which is the same as the result of mainstream viewpoint. Although many believe that the reason is insufficient model capacity of a Multi→Multi model, we think another possible reason is language confusion. This experiment contains 14 language pairs (two-way) while the data size is under one million. So the transformer capability is certainly enough. But since there are 14 languages in one model, there may be conflicts and confusion between language pairs in different directions, especially when they come from the same language family. Off-target (Zhang et al., 2020) could be a key issue, which we will further investigate in our future work.

- We find that data selection plays a more important role in our experiment when comparing with other training strategies. We observe that the domain of a small dataset is usually too narrow so the introduction of other data source will cause a great shift on domain, thereby affecting the performance of models on dev/test sets. For example, En→Bn can reach a BLEU score of 15.61 with only PM India data, but only 7.27 after adding the CVIT-PIB data. So

we refer to the idea of Static Data Selection in the curriculum learning and ensured little domain-shifting while training data size increases, thus the system performance enhances.

6 Conclusion

This paper presents the submissions by HW-TSC on the WAT 2020 Indic Multilingual Translation Task. We perform experiments with a series of pre-processing and training strategies. The effectiveness of each strategy is demonstrated. Our submission finally achieves competitive result in the evaluation.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Ankur Bapna, Naveen Arivazhagan, and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. *arXiv preprint arXiv:1906.06442*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 66–75.

Taku Kudo and John Richardson. 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, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, Suzhou, China. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Meng Sun, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Baidu neural machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 374–381.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, \Lukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wei Wang, Isaac Caswell, and Ciprian Chelba. 2019. Dynamically composing domain-data selection with clean-data selection by" co-curricular learning" for neural machine translation. *arXiv preprint arXiv:1906.01130*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huan-Bo Luan, and Yang Liu. 2017. THUMT: an open source toolkit for neural machine translation. *CoRR*, abs/1706.06415.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. *arXiv preprint arXiv:1803.00353*.