NICT's Submission To WAT 2020: How Effective Are Simple Many-To-Many Neural Machine Translation Models?

Raj Dabre[1] **Abhisek Chakrabarty**[2]
National Institute of Information and Communications Technology, Kyoto, Japan
[1]raj.dabre@nict.go.jp [2]abhisek.chakra@nict.go.jp

Abstract

In this paper we describe our team's (NICT-5) Neural Machine Translation (NMT) models whose translations were submitted to shared tasks of the 7^{th} Workshop on Asian Translation. We participated in the Indic language multilingual sub-task as well as the NICT-SAP multilingual multi-domain sub-task. We focused on naive many-to-many NMT models which gave reasonable translation quality despite their simplicity. Our observations are twofold: (a.) Many-to-many models suffer from a lack of consistency where the translation quality for some language pairs is very good but for some others it is terrible when compared against one-to-many and many-to-one baselines. (b.) Oversampling smaller corpora does not necessarily give the best translation quality for the language pair associated with that pair.

1 Introduction

Neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2014) is an end-to-end machine translation (MT) modeling approach that is known to give state-of-the-art translations for a variety of language pairs. Although, it is known to work particularly well for language pairs with an abundance of parallel corpora it tends to perform rather poorly for language pairs that lack large parallel corpora. Fortunately, multilingual neural machine translation (MNMT) (Dabre et al., 2020) methods can be applied in order to significantly improve the translation quality for such language pairs. The underlying reason for improvement is that sharing parameters among several language pairs enables transfer learning which is proven to improve translation quality regardless of the language pair.

For the 7^{th} Workshop on Asian Translation (WAT, 2020), our team (NICT-5) decided to focus on the Indic languages task and the NICT-SAP task. Both tasks showcase resource poor Indic and South-East Asian Languages and thus multilingual NMT solutions can be applied to great effect in these tasks. It is common to train one-to-many or many-to-one NMT models for multilingual tasks but many-to-many models are often not showcased. This sparked out curiosity and we decided to investigate how well a many-to-many model would perform in the case of the two sub-tasks we chose.

The many-to-many models we trained used the Transformer (Vaswani et al., 2017) architecture using the simple black-box token-prepending technique (Johnson et al., 2017). In essence we simply concatenated the individual parallel corpora while prepending an artificial token such as $2xx$ where xx indicates the target language. Typically, the smallest corpus in the multilingual dataset is oversampled to match the size of the largest one but we tried settings with and without oversampling. Furthermore, following (Chu et al., 2017) we additionally prepended source sentences with tokens such as $2dom$ where dom indicates the domain of the corpus. We only did this when we knew that the test (and train) sets would involve multiple domains. An evaluation of our models showed that their performance is not consistent because they sometimes outperform the one-to-many and many-to-one models and sometimes underperform them. Furthermore the performance also depends on the language pair. As a secondary observation, we noticed that when parallel corpora sizes are not too different, oversampling smaller corpora negatively affects the final translation quality. We hope that our many-to-many models will serve as baselines which can be significantly improved upon in the future. Although there are many-to-one and one-to-many models that may be better, many-to-many models have zero-shot translation (Johnson et al., 2017) capabilities and thus should be focused on in the interest of a one-for-all NMT model.

Proceedings of the 7th Workshop on Asian Translation, pages 98–102

2 Related Work

Our work in this paper focuses on neural machine translation and multilingualism.

Neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2014) is now the de-facto machine translation paradigm that is used in research as well as engineering applications. While the initial architectures used recurrent neural networks, the more recent architectures use self-attention and feed-forward networks (Vaswani et al., 2017) which enable faster training and decoding. The main advantage of NMT is that the translation models are small (can fit on low-memory and low-computation devices) and the training approach is end-to-end (rather than modular). Large translation models and modular (multiple components) design were the key features of the predecessor of NMT aka statistical machine translation (SMT) (Koehn et al., 2007) which involved error compounding as the input sentences were processed by multiple components that were prone to making mistakes.

Another advantage of NMT is that its inner working is non-symbolic which enables it to incorporate multiple languages without any need to modify the basic architecture. While we use the multilingual NMT approach proposed by (Johnson et al., 2017) for multilingualism and the derivative multi-domain NMT approach by (Chu et al., 2017) we refer readers to (Dabre et al., 2020) and (Chu and Wang, 2018) for overviews on multilingualism and domain-adaptation, respectively. We do not describe the multilingual or multi-domain NMT modeling techniques in this paper as they are the same as described in Johnson et al. (2017) and Chu et al. (2017).

Specific to WAT, multilingual multi-domain approaches have been shown to improve translation quality for low-resource languages (Banerjee et al., 2018; Dabre et al., 2018; Philip et al., 2018).

3 Experiments

In this section we describe the tasks, datasets, implementation details, evaluation methodology and actual models trained.

3.1 Tasks

We participated in the NICT-SAP and Indic multilingual tasks. Our team name is "NICT-5". The NICT-SAP task involves two domains: Wikinews and Software Documentation (loosely speaking a part of the IT domain). The languages involved are Thai (Th), Hindi (Hi), Malay (Ms), Indonesian (Id) and English (En). The Indic task involves mixed domain corpora for evaluation (various articles composed by Indian Prime Minister) and involves the languages Hindi (Hi), Marathi (Mr), Tamil (Ta), Telugu (Te), Gujarati (Gu), Malayalam (Ml), Bengali (Bg) and English (En). For both tasks, the objective was to train a single multilingual and multi-domain NMT model. The desired models could be one-to-many, many-to-one or many-to-many. English is either the source or the target language for both tasks.

Split	Domain	Language			
		Hi	Id	Ms	Th
Train	ALT	18,088			
	IT	254,242	158,472	506,739	74,497
Dev	ALT	1,000			
	IT	2,016	2,023	2,050	2,049
Test	ALT	1,018			
	IT	2,073	2,037	2,050	2,050

Table 1: The NICT-SAP task corpora splits. The corpora belong to two domains: wikinews (ALT) and software documentation (IT). The Wikinews corpora are N-way parallel.

3.2 Datasets

We used some corpora from the many listed in the official task descriptions[1][2]. In particular the following parallel corpora were used:

NICT-SAP Task: We used parallel corpora from the Asian Language Treebank (ALT) (Thu et al., 2016), KDE, GNOME and Ubuntu. The last three corpora were taken from OPUS[3]. Where the ALT corpus is for the ALT domain test set domain[4], the other three are for the IT or Software Documentation domain[5] (Buschbeck and Exel, 2020). Detailed statistics for corpora can be found in Table 1.

Indic Task: We used the filtered PM India dataset provided by organizers[6] and the CVIT-PIB

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/NICT-SAP-Task/

[2] http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/

[3] http://opus.nlpl.eu/

[4] http://lotus.kuee.kyoto-u.ac.jp/WAT/NICT-SAP-Task/altsplits-sap-nict.zip

[5] Software Domain Evaluation Splits

[6] http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/cvit-pmindia-mono-bi.zip

	Language						
Split	**Bn**	**Gu**	**Hi**	**Ml**	**Mr**	**Ta**	**Te**
Train	74,593	73,504	247,926	61,678	112,429	122,337	41,741
Dev	2,000	2,000	2,000	2,000	2,000	2,000	2,000
Test	3,522	4,463	3,169	2,886	3,760	3,637	3,049

Table 2: The Indic task corpora splits. The training corpora statistics are the result of combining the PIB and PMI corpora. While the number of development set sentences are the same, they are not N-way parallel as in the case of the Wikinews corpora.

dataset [7]. Detailed statistics for corpora can be found in Table 2.

With the exception of character splitting Thai, we do not perform any explicit pre-processing of any of the corpora used.

3.3 Implementation and Models Trained

We performed necessary modifications to the tensor2tensor v1.14[8] implementation of the transformer model. Tensor2tensor has an internal subword segmentation and we chose the option to train separate subword vocabularies of size 32,000. As we only train many-to-many models, our vocabularies are multilingual. We modified the original code to enable oversampling of smaller corpora during data pre-processing. We also modified the code to prepend the source sentences with a token $2xx$ to indicate the target language to be generated where xx is one of $en, bg, hi, mr, ml, ta, te, gu, th, ms, id$ as applicable. Additionally for the NICT-SAP task, we prepend the source sentences with a token like $2it$ or $2alt$ to distinguish between the IT and Wikinews domains.

As for the models trained, we trained transformer big models on single Tesla V100 GPUs using the hyperparameter settings corresponding to "transformer_big_single_gpu". Some important hyperparameters are: 6-layer encoder and decoder models with 16 attention heads, 1024-4096 hidden-filter sizes. We trained the models till convergence on development set BLEU score (Papineni et al., 2002). The development set BLEU score is the average of the BLEU scores of individual language pairs pairs. Evaluation on development set is done every 1000 batches (of 2048 tokens) and training stops when the BLEU score does not improve for 10 consecutive evaluations. Before evaluation, the model parameters are saved as a checkpoint and the last 10 checkpoints are averaged to give a single model which is then decoded using a beam of size 4 and a length penalty of 0.6[9].

We trained a total of 4 models, 2 models per task; one with oversampling the smaller corpora to match the size of the largest corpora and one without.

4 Results

Tables 3 and 4 contain results for the NICT-SAP and Indic tasks for translation to and from English. We primarily report BLEU scores for our translations and mark scores that are either better than or not better than (check captions) the organizers translations. We used the same data as the organizers did. The organizers trained one-to-many or many-to-one models whereas we only trained many-to-many-models. For other scores such as RIBES, JPO adequacy and AM-FM scores kindly check the workshop overview paper (Nakazawa et al., 2020) or the evaluation page[10] as applicable.

4.1 NICT-SAP results

From table 3, it is clear that our many-to-many models outperform the organizer's one-to-many or many-to-one models. Upon further investigation it seems that our models are better for the IT domain translation. Furthermore, models with and without oversampling of smaller corpora do not exhibit significant difference in performance for the IT domain (in most cases) but in the case of the Wikinews domain, models without oversampling tend to be significantly better (in most cases). Apart from Thai and Hindi to English translation, all other translation directions show reasonable BLEU scores indicating that multilingual NMT models

[7] http://preon.iiit.ac.in/~jerin/resources/datasets/pib_v0.2.tar

[8] https://github.com/tensorflow/tensor2tensor/

[9] We recommend tuning these two decoding hyperparameters on the development set in order to determine optimal values.

[10] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

Os	Domain	Translation Direction							
		En→Hi	**Hi→En**	**En→Id**	**Id→En**	**En→Ms**	**Ms→En**	**En→Th**	**Th→En**
Y	**ALT**	24.23	12.37	32.88	17.39*	36.77	18.03*	42.13*	10.78
N		22.74*	12.89	32.94	20.64	37.17	22.02	44.12*	12.51
Y	**IT**	14.03	16.89	32.52	25.95	34.62	26.33	28.24*	10.00
N		14.02	15.46	31.73	26.07	33.83	26.71	34.49*	10.34

Table 3: Results of our single multi-domain (ALT and IT) many-to-many NMT models for the NICT-SAP task. We indicate in the column labelled "Os" whether oversampling of the smaller corpora was done. Scores marked with "*" are **those that do not beat** organizers baselines.

Os	Translation Direction						
	En→Bn	**En→Gu**	**En→Hi**	**En→Ml**	**En→Mr**	**En→Ta**	**En→Te**
Y	9.69	8.00	12.68	4.76	7.29	3.94	3.54
N	7.97	10.13+	15.65+	5.00	8.97+	4.43+	4.73
	Bn→En	**Gu→En**	**Hi→En**	**Ml→En**	**Mr→En**	**Ta→En**	**Te→En**
Y	15.30	16.69	17.50	11.44	14.34	12.09	10.08
N	16.14	20.93	21.25	13.50	17.57	14.66	11.81

Table 4: Results of our single many-to-many NMT models for the Indic task. We indicate in the column labelled "Os" whether oversampling of the smaller corpora was done. Scores marked with "+" are **those that beat** organizers baselines.

are a reasonable solution for the involved language pairs. As for the reasons why Thai to English translation quality is poor (below or around 15 BLEU), it is clear that the parallel corpus for that pair is the smallest which has a strong negative impact.

4.2 Indic results

Table 4 presents results that are rather disappointing. Most of our translations were unable to beat those of the organizers'. However, there might be a simple explanation for this. Note that the Indic task involves almost twice as many translation directions as the NICT-SAP task. We used big transformer models for both tasks and so, the problem is not representation capacity but rather a lack of it. This lack of capacity likely comes from our naive multilingual solution coupled with rather small parallel corpora for each language pair. Future work will focus on expanding the amount of parallel corpora via popular techniques such as backtranslation (Sennrich et al., 2016). Another observation, just as in several cases of the ALT domain translations of the NICT-SAP task, we see that using oversampling is detrimental to translation quality.

5 Conclusion

In this paper we have described our submissions to the NICT-SAP and Indic multilingual tasks in WAT 2020. In general our many-to-many models have mixed performance where they sometimes outperform one-to-many or many-to-one models (organzier's models) and sometimes do not. Furthermore we observed that our models trained without oversampling smaller corpora tended to perform better than their counterparts that used oversampling of smaller corpora. This is especially true when parallel corpora for those pairs/domains contain approximately 100,000 sentences or fewer. This shows that investigation into improved data balancing methods might be necessary rather than relying on naive approaches we used in this paper. In the future we hope to improve upon our results by leveraging sophisticated methods involving better corpora balancing and monolingual corpora through back-translation.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv e-prints*, page arXiv:1409.0473.

Tamali Banerjee, Anoop Kunchukuttan, and Pushpak Bhattacharya. 2018. Multilingual Indian language translation system at WAT 2018: Many-to-one phrase-based SMT. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

Bianka Buschbeck and Miriam Exel. 2020. A parallel evaluation data set of software documentation with document structure annotation.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 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, Vancouver, Canada. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Anoop Kunchukuttan, Atsushi Fujita, and Eiichiro Sumita. 2018. NICT's participation in WAT 2018: Approaches using multilingualism and recurrently stacked layers. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jerin Philip, Vinay P. Namboodiri, and C.V. Jawahar. 2018. Cvit-mt systems for wat-2018. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew 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)*, pages 1574–1578, Portorož, Slovenia. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.