NICT's Supervised Neural Machine Translation Systems for the WMT19 News Translation Task

Raj Dabre* and **Kehai Chen*** and **Benjamin Marie*** and **Rui Wang*** and **Atsushi Fujita** and **Masao Utiyama** and **Eiichiro Sumita**
National Institute of Information and Communications Technology, Kyoto, Japan
{raj.dabre,khchen,bmarie,wangrui}@nict.go.jp
{atsushi.fujita,mutiyama,eiichiro.sumita}@nict.go.jp

Abstract

In this paper, we describe our supervised neural machine translation (NMT) systems that we developed for the news translation task for Kazakh↔English, Gujarati↔English, Chinese↔English, and English→Finnish translation directions. We focused on leveraging multilingual transfer learning and back-translation for the extremely low-resource language pairs: Kazakh↔English and Gujarati↔English translation. For the Chinese↔English translation, we used the provided parallel data augmented with a large quantity of back-translated monolingual data to train state-of-the-art NMT systems. We then employed techniques that have been proven to be most effective, such as back-translation, fine-tuning, and model ensembling, to generate the primary submissions of Chinese↔English. For English→Finnish, our submission from WMT18 remains a strong baseline despite the increase in parallel corpora for this year's task.

1 Introduction

Neural machine translation (NMT) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015) has enabled end-to-end training of a translation system without needing to deal with word alignments, translation rules, and complicated decoding algorithms, which are the characteristics of phrase-based statistical machine translation (PBSMT) (Koehn et al., 2007). NMT performs well in resource-rich scenarios but badly in resource-poor ones (Zoph et al., 2016). With the aid of multilingualism, transfer learning, and monolingual corpora, researchers have shown that the translation quality in a low-resource scenario can be significantly boosted (Zoph et al., 2016; Firat et al., 2016; Sennrich et al., 2016a). Furthermore, unsupervised NMT (Lample et al., 2018) has enabled translation in a scenario where only monolingual corpora are available.

In this paper, we describe all the systems for Kazakh↔English, Gujarati↔English, Chinese↔English, and English→Finnish, that we developed and submitted for WMT 2019 under the team name "NICT." In particular our observations can be summarized as follows:

Kazakh→English translation heavily benefits from the existence of Russian as a pivot language in the form of a Russian–Kazakh corpus which can be used to generate a pseudo-parallel Kazakh–English corpus from the Russian–English corpus.

Gujarati→English translation can be drastically improved by training a robust Hindi→English model and fine tuning it on the Gujarati–English corpus.

Chinese↔English translation can benefit from back-translation, model ensembling, and fine-tuning based on the development data.

English→Finnish translation generated by our WMT18's NMT system (Marie et al., 2018) remains a strong baseline despite the availability of larger bilingual corpora for training this year.

Noisy parallel corpora for back-translation leads to poor quality pseudo-parallel data which leads to poor translations.

Kindly refer to the overview paper (Bojar et al., 2019) for additional details about the tasks, comparisons to other submissions, human analyses and insights.

2 The Transformer NMT Model

The Transformer (Vaswani et al., 2017) is the current state-of-the-art model for NMT. It is a

*equal contribution

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 168–174
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

sequence-to-sequence neural model that consists of two components: the *encoder* and the *decoder*. The encoder converts the input word sequence into a sequence of vectors. The decoder, on the other hand, produces the target word sequence by predicting the words using a combination of the previously predicted word and relevant parts of the input sequence representations. The reader is encouraged to read the original paper (Vaswani et al., 2017) for a deeper understanding.

3 Kazakh↔English Task

3.1 Use of Pseudo-Parallel Data

In this paper, we rely on a highly reliable data-augmentation technique known as back-translation (Sennrich et al., 2016a). This technique relies on a L2→L1 model to translate an L2 monolingual corpus, thereby yielding a large L1–L2 pseudo-parallel corpus for L1→L2 translation. The final L1→L2 translation quality depends on the quality of the pseudo-parallel corpus which in turn depends on L2→L1 translation quality. For a low-resource L1–L2 pair, this approach is rather infeasible.[1] However, the existence of a pivot-language, L3, can prove beneficial. In this situation, we can assume large L3–L1 and L3–L2 corpora. Using a robust L3→L1 model, we can translate the L3 side of the L3–L2 corpus to obtain a high quality L1–L2 pseudo-parallel corpus (Firat et al., 2016).

In our participation, we regard Russian as the helping language, L3.

3.2 Datasets

We used the official Kazakh–English, Kazakh–Russian, and Russian–English datasets provided by WMT. All three datasets belong to the news domain. After filtering the Kazakh–English parallel corpus using the "clean-corpus.perl" script in Moses (Koehn et al., 2007),[2] we obtained 98,602 (noisy) sentence pairs.

We filtered the Kazakh–Russian corpus of 5,063,666 lines according to the scores provided with the corpus files. The real-valued scores ranged from 0 to a maximum value of 11. Since higher scores meant better pairs, we filtered the corpora using the thresholds 1, 2, 3, 4, and 5 and trained NMT models on the filtered corpora. We found out that a threshold of a score of at least 1 gave a corpus of 2,905,538 lines and performs the best on a development set.[3] Using scores of 2, 3, and 4 gave slightly lower BLEU scores on the development set and thus we decided to use as large a corpus as possible.

We used 4,596,000 lines[4] of Russian sentences, randomly selected from the 12,061,155 sentences Russian–English corpus, for back-translation. No other type of pre-processing was performed.

3.3 Systems

We used the tensor2tensor[5] version 1.6 implementation of the Transformer (Vaswani et al., 2017) model. We used the default hyper-parameters in tensor2tensor for all our models with the exception of the number of training iterations. Unless mentioned otherwise we used the Transformer "base" model hyper-parameter settings with a $2^{15} = 32,768$ sub-word vocabulary which was learned using tensor2tensor's internal tokenization and sub-word segmentation mechanism. We learned separate sub-word vocabularies for the source and target languages.

During training, a model checkpoint was saved every 1000 iterations. All models were trained till convergence on the WMT19's official development set BLEU score. We averaged the last N model checkpoints and used it for decoding the test sets. N is 20 for Kazakh↔English. The choice of N depended on the number of iterations for convergence which in turn depended on the size and quality of the data used to train models. We chose the beam size and length penalty by tuning on the development set. We did not ensemble multiple models although it could possibly improve the translation quality even further.

We first trained Russian→Kazakh and Russian→English models for back-translation purposes. The Russian→Kazakh model was trained for 300,000 iterations on one GPU with a batch size of 2048 words and the

[1] We had initially experimented with the large Kazakh and English monolingual corpora for back-translation but observed no benefits.

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

[3] We chose a set of 2,000 sentences, not included in the training set, to monitor convergence.

[4] Due to lack of time, we were unable to back-translate all Russian sentences before the task deadline. After the deadline we experimented with back-translating all Russian sentences but did not observe any appreciable improvements in translation quality.

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

Task	BLEU	BLEU cased	IGNORE BLEU (11b)	IGNORE BLEU-cased (11b)	IGNORE BLEU-cased-norm	TER	BEER 2.0	CharactTER	Rank
Kazakh→English	28.1	26.2	28.1	26.2	26.2	0.670	0.555	0.701	3/9
English→Kazakh	6.4	6.4	6.4	6.4	7.8	0.926	0.418	0.841	8/9
Gujarati→English	18.6	17.2	18.6	17.2	17.3	0.733	0.508	0.705	5/10
English→Gujarati	10.5	10.5	10.5	10.5	10.6	0.856	0.448	0.785	6/8

Table 1: Results for Kazakh↔English and Gujarati↔English tasks. These scores are simply copied from the official runs list.

Russian→English for 100,000 iterations[6] on two GPUs with a batch size of 4096 words. We used the Russian→English model to translate the Russian side of the Russian–Kazakh corpus into English. On the other hand, we used the Russian→Kazakh model to translate the Russian side of the Russian–English corpus into Kazakh. We used greedy decoding (to save time) with a length penalty of 1.0 in both cases.

Both Kazakh→English and English→Kazakh models were trained only on the pseudo-parallel data, using two GPUs with a batch size of 4096 words, till the convergence of BLEU on the development set. As a result, the Kazakh→English model was trained for 200,000 iterations, whereas the English→Kazakh model was trained for 220,000 iterations. For both translation directions, decoding was done using a beam of size 10 and length penalty of 0.8 (determined by tuning on the development set).

3.4 Results

Refer to rows 1 and 2 of Table 1 for the various automatic evaluation scores. For Kazakh→English our submitted system achieved a cased BLEU score of 26.2 placing our system at 3rd rank out of 9 primary systems. On the other hand, our English→Kazakh performed poorly with its system achieving a BLEU score of 6.4 placing it at 8th out of 9 primary systems.

Initially, we had experimented with back-translating English monolingual corpora to Kazakh using models trained on the Kazakh–English parallel corpora. However, this led to a BLEU score of less than 15. After repeated experimentation we realized that the Kazakh–English parallel corpus was of extremely poor quality and hence decided to experiment with Russian as a pivot language. We trained a multilingual English–Russian–Kazakh model and pivot translation (Firat et al., 2016) gave a BLEU of around 18 which motivated us to exploit the Russian–Kazakh data. The main lesson we learned was: always exploit a pivot language whenever possible instead of relying on a parallel corpus of bad quality. Note once again that our submissions did not involve the use of the Kazakh–English corpus provided by the organizers.

[6]Given that the Russian–English corpus contains over 12M sentence pairs, training for more iterations could give better results.

4 Gujarati↔English Task

4.1 Fine-Tuning for Transfer Learning

In addition to the approaches in Section 3.1, we also use fine-tuning for transfer learning. Zoph et al. (2016) proposed to train a robust L3→L1 parent model using a large L3–L1 parallel corpus and then fine-tune it on a small L2–L1 corpus to obtain a robust L2→L1 child model. The underlying assumption is that the pre-trained L3→L1 model contains prior probabilities for translation into L1. The prior information is divided into two parts: language modeling information (strong prior) and cross-lingual information (weak or strong depending on the relationship between L3 and L2). Dabre et al. (2017) have shown that linguistically similar L3 and L2 allow for better transfer learning. As such, we transliterate L3 to L2 before pre-training a parent model. This could help in faster convergence, ensure cognate overlap, and potentially lead to a better translation quality.

In this participation, we used Hindi as the helping language, L3.

4.2 Datasets

We used the official Gujarati–English and Hindi–English datasets provided by WMT. The Gujarati–English corpus contains 28,683 sentence pairs belonging to the news and Wiki domains. We also used the ILCI Gujarati–English corpus (Jha, 2010) of 44,777 sentence pairs belonging to the tourism and health domains. In total the size

of the Gujarati–English parallel corpus is 73,460 sentence pairs. The Hindi–English corpus of 1,492,827 sentence pairs contains sentence pairs belonging to multiple domains.

We used around 2,700,919 lines of Gujarati monolingual corpora (of which approximately 244,919 lines were from the news domain) for back-translation.[7] We mapped the script on the Hindi side of the Hindi–English corpus to Gujarati using the Indic languages toolkit.[8] No other type of pre-processing was performed.

We had initially experimented with a large English monolingual corpus for back-translation but observed no benefits.

4.3 Systems

Most training details, including the size of subword vocabulary, are same as those in Section 3.3. The only exception is the number of checkpoints we averaged before decoding which is 10 instead of 20. This is because Gujarati↔English models converged rather quickly and hence were not trained for a long period of time.

We first trained a bi-directional Gujarati↔English model[9] using the parallel corpora mentioned above, for 60,000 iterations on one GPU with a batch size of 2048 words. We then used this model to translate Gujarati monolingual data into English using greedy decoding with a length penalty of 1.0. We also pre-trained a Hindi→English model where the scripts on the Hindi side was mapped to those in Gujarati. This model was trained for 90,000 iterations on one GPU with a batch size of 4096 words.

We then trained a Gujarati→English model by fine-tuning the Hindi→English model on the Gujarati→English data for an additional 15,000 iterations[10] on one GPU with a batch size of 4096 words. We also trained a English→Gujarati model using the pseudo-parallel corpus by training for 60,000 iterations[11] on one GPU with a batch size of 2048 words. For both cases, decoding was done using a beam of size 10 and length penalty of 0.8.

4.4 Results

Refer to rows 3 and 4 of Table 1 for the various automatic evaluation scores. For Gujarati→English our submitted system run achieved a cased BLEU score of 17.2 placing our system at 5th position out of 10 primary systems. On the other hand, our English→Gujarati performed poorly with its system run achieving a BLEU score of 10.6 placing it at 6th position out of 8 primary systems.

Similar to our experience in Kazakh↔English, using the NMT models trained using Gujarati–English parallel corpora for back-translation, led to poor translation quality. Our Gujarati→English system achieved less than 10 BLEU when relying on a naive back-translation approach. As such, we decided to rely on transfer learning by fine-tuning a Hindi–English model on the Gujarati–English corpus. In WMT19, Hindi was the only language linguistically similar to Gujarati and hence we did not explore other resource-rich language pairs. Other participants used Czech–English for transfer learning and achieved similar success. On the other hand, only the pseudo English–Gujarati corpus was available for developing the English→Gujarati system. Due to lack of time, we did not try using our transfer learning based Gujarati→English model for back-translation. Given that our submitted Gujarati→English system is over 8 BLEU points higher than the naive back-translation based system, we expect that English→Gujarati has a huge potential for improvement.

As in the case of Kazakh↔English, we noted that it is extremely beneficial to leverage a helping language, such as Hindi, for improving translation quality.

5 Chinese↔English Tasks

5.1 Datasets

The training data for the Chinese↔English (ZH↔EN) translation tasks consists of two parts: 1) we selected the first 10 million lines of the News Crawl 2016 English corpus according to our last year's finding (Marie et al., 2018), 2) the corresponding synthetic data was generated through back-translation (Sennrich et al., 2016a). We applied tokenizer and truecaser of Moses (Koehn

[7] During back-translation, some parts of the monolingual corpus remained untranslated due to out-of-memory errors caused by very long input sentences.

[8] https://github.com/anoopkunchukuttan/indic_nlp_library

[9] We chose a bi-directional model because we observed higher BLEU scores on the development set compared to a unidirectional model.

[10] Fine-tuning converges quickly.

[11] Given the size of the pseudo-parallel corpus we expected to train for much longer but observed convergence rather quickly. It is likely that our generated corpus was quite noisy and hence the models had limited learning potential.

Task	System	BLEU	BLEU cased	IGNORE BLEU (11b)	IGNORE BLEU-cased (11b)	IGNORE BLEU-cased-norm	TER	BEER 2.0	CharactTER
ZH→EN	Single model	24.1	23.3	24.1	23.3	23.5	0.667	0.574	0.643
	+back-translation	26.6	25.3	26.6	25.3	25.5	0.652	0.585	0.632
	+fine-tuning	28.7	27.5	28.7	27.5	27.7	0.621	0.599	0.613
	+ensemble five models	32.3	31.0	32.3	31.0	31.3	0.599	0.615	0.569
EN→ZH	Single model	30.3	30.3	0.4	0.4	2.2	0.999	0.304	0.839
	+back-translation	31.8	31.8	0.6	0.6	2.6	0.999	0.315	0.765
	+fine-tuning	33.1	33.1	0.0	0.0	2.3	1.000	0.319	0.747
	+ensemble five models	34.5	34.5	0.7	0.7	2.6	0.999	0.326	0.734

Table 2: Results for ZH↔EN translation task. "Single model" denotes that it was trained by only using the first 10M lines of the News Crawl-2016 English corpus as training data. These scores are simply copied from the official runs list.

et al., 2007) to the English sentences. Jieba[12] was used to tokenize the Chinese sentence. For cleaning, we filtered out sentences longer than 80 tokens in the training data.

5.2 Systems

We used Marian toolkit (Junczys-Dowmunt et al., 2018)[13] to build competitive NMT systems based on the Transformer (Vaswani et al., 2017) architecture. We used the byte pair encoding (BPE) algorithm (Sennrich et al., 2016b) for obtaining the sub-word vocabulary whose size was set to 50,000. The number of dimensions of all input and output layers was set to 512, and that of the inner feed-forward neural network layer was set to 2048. The number of attention heads in each encoder and decoder layer was set to eight. During training, the value of label smoothing was set to 0.1, and the attention dropout and residual dropout were set to 0.1. The Adam optimizer (Kingma and Ba, 2014) was used to tune the parameters of the model. The learning rate was varied under a warm-up strategy with warm-up steps of 16,000. All NMT models for ZH↔EN tasks were consistently trained on four P100 GPUs. We validated the model with an interval of 5,000 batches on the development set and selected the best model according to BLEU (Papineni et al., 2002) score on the newsdev2018 data set.

We performed the following training run independently for five times to obtain the models for ensembling. First, an initial model was trained on the provided parallel data and used to generate pseudo-parallel data through back-translation. A new model was then trained from scratch on the mixture of the original parallel data and the pseudo-parallel data. The new model was further fine-tuned on the concatenation of newsdev2017 and newstest2017 data sets for 20 epochs. Finally, we decoded the newstest2019 test set with an ensemble of the five fine-tuned models to generate the primary submissions for the ZH↔EN task.

5.3 Results

Table 2 shows the results of ZH↔EN tasks. It is obvious that the back-translation, fine-tuning, and ensemble methods are greatly effective for the ZH↔EN tasks. In particular, the ensemble gave more improvements on the ZH→EN task over the "Single model+back-translation+fine-tuning" model than the EN→ZH task. In addition, these three methods can incrementally improve translation performance of the Transformer NMT.

6 English→Finnish Task

For the translation direction English→Finnish, we used the exactly same NMT models and system used to generate our last year's submission (Marie et al., 2018). We did not exploit the new larger parallel data provided for this year. For this year, we only submitted the output produced by the ensemble of our three NMT models. Our system was ranked third for the task according to BLEU-cased, at 23.2 BLEU points, which is 4.2 BLEU points below the best system submitted to the task.

7 Conclusion

In this paper, we have described our primary systems whose translations we have submitted to WMT2019. In general, we found that back-translation, fine-tuning, and ensembling are the most effective means of maximizing the translation quality for all language pairs. In addition to this, we have observed that leveraging a helping language, such as Russian for Kazakh↔English

[12] https://github.com/fxsjy/jieba
[13] https://marian-nmt.github.io

translation and Hindi for Gujarati→English translation, can lead to large benefits as compared to using only parallel corpora and back-translation.

Acknowledgments

We thank the organizers for providing the datasets and the reviewers for their valuable suggestions in improving this paper. This work was conducted under the program "Research and Development of Enhanced Multilingual and Multipurpose Speech Translation Systems" of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Christof Monz, Mathias Müller, and Matt Post. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*, Florence, Italy. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, USA.

Girish Nath Jha. 2010. The TDIL program and the Indian Language Corpora Initiative (ILCI). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 982–985, Valletta, Malta. European Language Resources Association (ELRA).

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej 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.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations*.

Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2018. NICT's neural and statistical machine translation systems for the WMT18 news translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 449–455, Belgium, Brussels.

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, Philadelphia, USA.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistic*, pages 1715–1725, Berlin, Germany.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 3104–3112, Montréal, Canada.

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 30th Advances in Neural Information Processing Systems*, pages 5998–6008.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, USA.