The University of Sydney's Machine Translation System for WMT19

Liang Ding **Dacheng Tao**
UBTECH Sydney AI Center, School of Computer Science, FEIT
University of Sydney, Australia
ldin3097@uni.sydney.edu.au, dacheng.tao@sydney.edu.au

Abstract

This paper describes the University of Sydney's submission of the WMT 2019 shared news translation task. We participated in the Finnish→English direction and got the best BLEU(33.0) score among all the participants. Our system is based on the self-attentional Transformer networks, into which we integrated the most recent effective strategies from academic research (*e.g.*, BPE, back translation, multi-features data selection, data augmentation, greedy model ensemble, reranking, ConMBR system combination, and post-processing). Furthermore, we propose a novel augmentation method **Cycle Translation** and a data mixture strategy $\boldsymbol{Big/Small}$ **parallel construction** to entirely exploit the synthetic corpus. Extensive experiments show that adding the above techniques can make continuous improvements of the BLEU scores, and the best result outperforms the baseline (Transformer ensemble model trained with the original parallel corpus) by approximately 5.3 BLEU score, achieving the state-of-the-art performance.

1 Introduction

Neural machine translation (NMT), as a succinct end-to-end paradigm, has resulted in massive leap in state-of-the-art performances for many language pairs (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2017; Wu et al., 2016; Vaswani et al., 2017). Among these encoder-decoder networks, the Transformer (Vaswani et al., 2017), which solely uses along attention mechanism and eschews the recurrent or convolutional networks, leads to state-of-the-art translation quality and fast convergence speed (Ahmed et al., 2017). Although many Transformer-based variants are proposed (*e.g.*, DynamicConv (Wu et al., 2019), sparse-transformer (Child et al., 2019)), our preliminary experiments show that their performances are unstable compared to the traditional Transformer. Traditional Transformer therefore was employed as our baseline system. In this paper, we summarize the USYD NMT systems for the WMT 2019 Finnish→English (FI→EN) translation task.

#	**cycle translated sample sentence pair**
1	*She stuck to her principles even when some suggest that in an environment often considered devoid of such thing there are little point.*
2	*She insists on her own principles, even if some people think that it doesn't make sense in an environment that is often considered to be absent.*

Table 1: Example of difference between original sentence (line 1) and cycle translated result (line 2). Pretrained BERT model using all available English corpora show that the $\mathcal{L}oss$ decreased from 6.98 to 1.52.

As the limitation of time and computation resources, we only participated in one challenging task FI→EN, which lags behind other language pairs in translation performance (Bojar et al., 2018). We introduce our system with three parts.

First, at data level, we find that the data quality of both parallel and monolingual is unbalanced (*i.e.*, contains a large number of low quality sentences). Thus, we apply several features to select the data after pre-processing, for example, language models, alignment scores etc. Meanwhile, in order to fully utilize monolingual corpus, not only back translation (Sennrich et al., 2015) is adopted to back translate the high quality monolingual sentences with target-to-source(T2S) model, we also propose **Cycle Translation** to improve the low-quality sentences, in turn resulting in corresponding high-quality back translation results. Note that unlike text style transfer task (Shen et al., 2017; Fu et al., 2018; Prabhumoye et al., 2018) which transfers text to specific style (*e.g.*, political

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 175–182
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

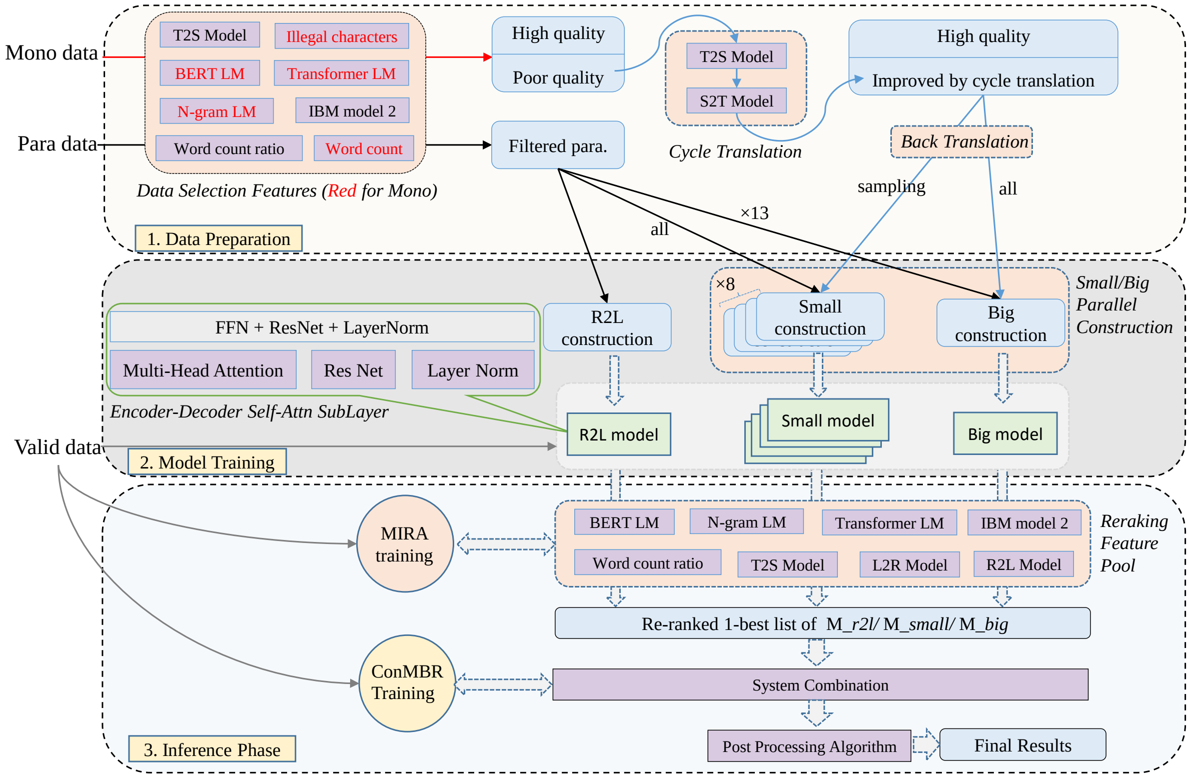

Figure 1: The schematic structure of the three main stages of the USYD-NMT. They are data preparation stage, model training stage and inference phrase. For brevity, here Mono, Para, and Valid represent the monolingual, parallel and validation data respectively.

slant, gender), we aim to improve the fluency of sentences, for instance, through cycle translation, low quality sentence in Table 1 becomes more fluent in terms of language model score. The top diagram of Figure 1 depicts data preparation process concretely.

As to model training in the middle part of Figure 1, we empirically introduced **Big/Small parallel construction** strategy to construct training data for different models. The intuition is all the data are advantageous and can be fully exploited by different models, thus we train 8 Transformer_base models ($\mathcal{M}_{small} \times 8$) by using different small scale corpus constructed by small parallel construction method and a Transformer_big model ($\mathcal{M}_{big} \times 1$) based on the big parallel construction method. In the meantime, a right-to-left model ($\mathcal{M}_{r2l}$) is trained.

In addition, in inference phrase, we comprehensively consider the ensemble strategies at model level, sentence level and word level. For model level ensemble, while brutal ensemble top-N or last-M models may improve translation performance, it is difficult to obtain the optimal result. Hence we employ Greedy Model Selection based Ensembling (GMSE) (Partalas et al., 2008; Deng et al., 2018). For sentence level ensemble, we keep top n-best for multi-features reranking. And for word aspect, we adopt the confusion network decoding (Bangalore et al., 2001; Matusov et al., 2006; Sim et al., 2007) with using the consensus network minimum Bayes risk (MBR) criterion (Sim et al., 2007). After combination, a post-processing algorithm is employed to correct inconsistent number and years between the source and target sentences. The bottom part of Figure 1 shows the inference process.

Our omnivorous model achieved the best BLEU (Papineni et al., 2002) scores among submitted systems, demonstrating the effectiveness of the proposed approach. Theoretically, our approach is not specific to the Finnish→English language pair, *i.e.*, it is universal and effective for any language pairs. The remainder of this article is organized as follows: Section 2 will describe each component of the system. In Section 3, we introduce the data preparing details. Then, the experimental results are showed in Section 4. Finally, we conclude in Section 5.

model_parameters	$\mathcal{M}$_small	$\mathcal{M}$_big
num_stack	6	6
hidden_size	512	1024
FFN_size	2048	4096
num_heads	8	16
p_dropout	0.1	0.3

Table 2: Model differences between base and big.

2 Approach

2.1 Neural Machine Translation Models

Given a source sentence $X = x_1, ..., x_{T'}$, NMT model factors the distribution over target sentence $Y = y_1, ..., y_T$ into a conditional probabilities:

$$p(Y|X;\theta) = \prod_{t=1}^{T+1} p(y_t|y_{0:t-1}, x_{1:T'};\theta) \qquad (1)$$

where the conditional probabilities are parameterized by neural networks.

The NMT model consists of two units: an encoder and a decoder. The encoder is assumed that it can adequately represent the source sentence. Then, the decoder can recursively predict each target word. Parameters of encoder, decoder and attention mechanism are trained to maximize the likelihood with a cross-entropy loss applied:

$$\begin{aligned}\mathcal{L}_{ML} &= \log p(Y|X;\theta) \\ &= \sum_{t=1}^{T+1} \log p(y_t|y_{0:t-1}, x_{1:T'};\theta)\end{aligned} \qquad (2)$$

Concretely, an self-attentional encoder-decoder architecture (Vaswani et al., 2017) was selected to capture the causal structure. For training with different size of corpus, we employ the Transformer_base ($\mathcal{M}$_**base**) and Transformer_big ($\mathcal{M}$_**big**) in our structure, see Table 2.

2.2 Data Selection Features

Inspired by (Bei et al., 2018), where their system shows data selection can obtain substantial gains, we deliberately design criteria for parallel and monolingual corpus. Both of them employ rule-based features, count features, language model features. And for parallel data, word alignment-based features, T2S translation model score features are applied. The feature types are described in Table 3. Our BERT language model used here is trained from scratch by the open-source tool[1] with target side data.

Category	Features
NMT Features	T2S score (Sennrich et al., 2016)
LM Features	BERT LM (Devlin et al., 2018)
	Transformer LM (Bei et al., 2018)
	N-gram LM (Stolcke, 2002)
Alignment Features	IBM model 2 (Dyer et al., 2013)
Rule-based features	Illegal characters (Bei et al., 2018)
Count Features	Word count
	Word count ratio

Table 3: Features for data selection.

According to our observations, by using above multiple data selection filters, issues like misalignment, translation error, illegal characters, over translation and under translation in terms of length could be significantly reduced.

2.3 Cycle Translation for Low-quality Data

Although the data selection procedure has preserved relatively high quality monolingual data, there are still a large batch of data is incomplete or grammatically incorrect. To address this problem, we proposed Cycle Translation (denoted as $\mathcal{CT}(\cdot)$, as Figure 2) to improve the mono-lingual data that below the quality-threshold (According to our empirical ablation study in section 4, the latter 50% will be cycle translated in our submitted system).

2.4 Back Translation for monolingual corpus

Back-translation (Sennrich et al., 2015; Bojar et al., 2018), translating the large scale monolingual corpus to generate synthetic parallel data by Target-to-Source pretrained model, has been widely utilized to improve the translation quality since adding the synthetic data into parallel data can enhance the in-domain information over the original corpus distributions, allowing the translation model to be more robust and deterministic.

2.5 Greedy Model Selection Based Ensemble

Model ensemble is a typical boosting technique, which refers to combining multiple models to reduce stochastic differences in the output that may not be avoided at a single run. Also normally, ensemble model outperforms the the best single one.

[1] https://github.com/huggingface/pytorch-pretrained-BERT

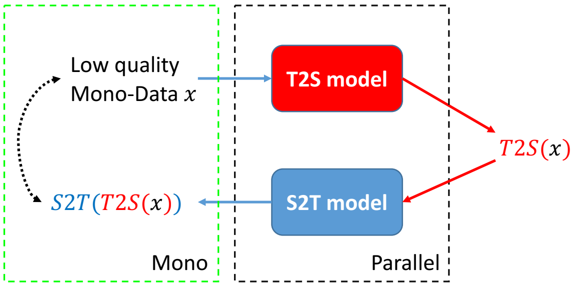

Figure 2: The Cycle Translation process, into which we feed the low quality monolingual data x, and then correspondingly obtain the improved data $\mathcal{CT}(x)$ (denoted as $S2T(T2S(x))$ in figure). Note that models marked in red and green represent the T2S and S2T model trained by $\mathcal{M}_{small}$ with the processed given parallel corpus, the red arrows indicate the data flows of the opposite language type of the inputs. The dotted double-headed arrow between the input x and the final output $\mathcal{CT}(x)$ means that they share the semantics but differs in fluency.

In neural machine translation, we generally ensemble several checkpoints saved during a single model training. However, our preliminary experiments show that both top-N or last-M ensembling approaches could only bring very insignificant improvements but consume a lot of GPU resources.

To overcome this issue, we adopt greedy model selection based ensembling(GMSE), which technically follows the instruction of (Deng et al., 2018).

2.6 Reranking n-best Hypotheses

As the NMT decoding being generally from left to right, this leads to label bias problem (Lafferty et al., 2001). To alleviate this problem, we rerank the n-best hypotheses through training a k-best batch MIRA ranker (Cherry and Foster, 2012) with multiple features on validation set. The feature pool we integrated include left-to-right (L2R) translation model, (right-to-left) R2L translation model, (target-to-source) T2S translation model, language model, IBM model 2 alignment score, and word count ratio. After multi-feature reranking, the best hypothesis of each model ($\mathcal{M}_{big} \times 1$, $\mathcal{M}_{small} \times 8$ and R2L model) was retained for system combination.

2.6.1 Left-to-right NMT model

The L2R feature refers to the original translation model that could generate the n-best list. During reranking training, we keep the original perplexity score evaluated by this L2R model as L2R feature.

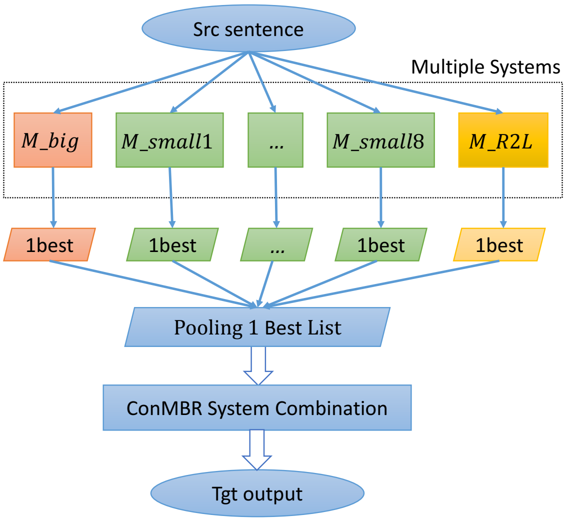

Figure 3: The System Combination process, into which we feed each system/model with the source sentence x, in turn obtain corresponding 1-best result $\mathcal{M}_{big}(x)$, $\mathcal{M}_{small1}(x)$, ... ,$\mathcal{M}_{small2}(x)$,$\mathcal{M}_{R2L}(x)$ (Note that the 1-best result here of each system was already reranked). After pooling all system results, we can perform the ConMBR system combination decoding and obtain the final target side results.

2.6.2 Right-to-Left NMT Model

The R2L NMT model using the same training data but with inverted target sentences (*i.e.*, reverse target side characters "a b c d"→"d c b a"). Then, inverting the hypothesis in the n-best list such that each sequence can be given a perplexity score by R2L model.

2.6.3 Target-to-Source NMT Model

The T2S model was initially trained for back-translation, we can employ this model to assess the translation adequacy as well by adding the T2S feature to reranking feature pool.

2.6.4 Language Model

Besides above features, we employ language models as an auxiliary feature to give the fluent sentences better scores such that the results are easier to understand by human.

2.6.5 Word Count Ratio

To alleviate over-translation or under-translation in terms of length, we set the optimal ratio of $\mathcal{L}_{fi} : \mathcal{L}_{en}$ to 0.76 according to the corpus-based statistics. We use the deviation between the ratio of each sentence pair and this optimal ratio as the score.

src	*Siltalan edellinen kausi liigassa oli 2006-07*
pred	*Siltala's previous season in the league was 2006 at 07*
+post	*Siltala's previous season in the league was 2006-07*

Table 4: Example of the effectiveness of post-processing in handling inconsistent number translation.

Data	Sentences
filtered parallel corpus	5,831,606
reconstructed mono	82,773,126
filtered synthetic parallel	75,940,978
small construction($\times 8$)	11,663,212
big construction	151,751,856

Table 5: Data statistics after data preparation

2.7 System Combination

As is shown in Figure 3, in order to take full advantages of different models($\mathcal{M}_{big}\times 1$, $\mathcal{M}_{small}\times 8$ and R2L model), we adopted word-level combination where confusion network was built. Concretely, our method follows Consensus Network Minimum Bayes Risk (ConMBR) (Sim et al., 2007), which can be modeled as

$$E_{ConMBR} = \text{argmin}_{E'}\mathcal{L}(E', E_{con}) \qquad (3)$$

where E_{con} was obtained as backbone through performing consensus network decoding.

2.8 Post-processing

In addition to general post-processing strategies (*i.e.*, de-BPE, de-tokenization and de-truecase [2]), we also employed a post-processing algorithm (Wang et al., 2018) for inconsistent number, date translation, for example, "*2006-07*" might be segmented as "*2006 -@@ 07*" by BPE, resulting in the wrong translation "*2006 at 07*". Our post-processing algorithm will search for the best matching number string from the source sentence to replace these types of errors, see Table 4.

3 Data Preparation

We used all available parallel corpus [3] for Finnish→English except the "Wiki Headlines" due to the large number of incomplete sentences, and for monolingual target side English data, we selected all besides the "Common Crawl" and "News Discussions". The criteria is inspired by (Marie et al., 2018), who won the first place in this direction at WMT18. Table 5 shows the final corpus statistics. More details are as follows:

Parallel Data: We use the criteria in section 2.2, the overall criteria are following:

- Remove duplicate sentence pairs.
- Remove sentence pairs containing illegal characters.
- Retain sentence pairs between 3 and 80 in length.
- Remove sentence pairs that are too far from the best ratio($\mathcal{L}_{fi}$: $\mathcal{L}_{en}$=0.76)
- Remove pairs containing influent English sentences according to a series of LM features.
- Remove inadequate translation sentence pairs according to $\mathcal{M}_{T2S}$ score.
- Remove sentence pairs with poor alignment quality according to IBM model 2.

After data selection, there are approximately 5.8M parallel sentences.

Monolingual Data: For our Finnish→English system, back translation was performed for monolingual English data. Before back-translation, we filter them according to the aforementioned criteria in section 2.2 and concurrently, the scores of each sentence is obtained. After monolingual selection, there are 82M sentences remained, which is still a gigantic scale. We ***cycle translate*** the last 25%, 50% and 75% of it in terms of the LM scores to empirically identify the optimal threshold and improve the fluency of monolingual corpora. In doing so, all monolingual corpus is kept at relatively high quality.

Synthetic Parallel Data: The synthetic parallel data also needs to be filtered by alignment score and word count ratio to alleviate poor translation. Further filtration retains 75M synthetic data.

On the other hand, previous works have shown that the maximum gain can be obtained by mixing

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

[3] both parallel and monolingual corpus can be obtained from: http://www.statmt.org/wmt19/translation-task.html

#	Models	news-test18	news-test19	Δ_{ave}
1	Baseline(original_parallel + ensemble)	21.8	27.3	–
2	$\mathcal{M}_{small}$(selected_parallel)	22.6	27.9	+0.70
3	`+synthetic`	23.9	28.8	
4	`+GMSE`	24.2	29.2	
5	`+reranking`	24.6	29.5	
6	`+post processing`	24.8	29.6	+2.65
7	Cycle translation + B/S construction	25.3	30.9	+3.55
8	`+GMSE`	25.9	31.7	
9	`+reranking`	26.3	32.4	
10	`+system combination`	26.6	32.8	
11	`+post processing`	**26.7**	**33.0**	**+5.30**

Table 6: FI→EN Results on newstest2018 and newstest2019. The submitted system is the last one.

#	$\mathcal{CT}$ **Ratio**	**Val.**	Δ
1	[0%]	22.62	-
2	[25%]	23.18	+0.56
3	[50%]	**23.70**	**+1.08**
4	[75%]	23.07	+0.45

Table 7: Different experimental settings that employed different cycle translation thresholds. Val. denotes that the results are reported on validation set.

the sampled synthetic and original corpus in a ratio of 1:1 (Sennrich et al., 2015, 2016). The size of the synthetic corpus is generally larger than the parallel corpus, thus partial sampling is required to satisfy the 1-1 ratio. However, such sampling leads to waste of enormous synthetic data. To address this issue, we argue that a better construction strategy can be introduced to make full use of the synthetic corpus, subsequently leading to better translation quality.

Small Parallel Construction: We randomly sampled approximate 5.8M corpus from the shuffled synthetic data for 8 times and mix them with parallel data respectively.

Big Parallel Construction: The aim of big construction is to fully utilize the synthetic data. To achieve this, we repeated the parallel corpus 13 times and then mixed it with all synthetic corpora.

4 Experiments

The metric we employed is detokenized case-sensitive BLEU score. `news-test2018` is utilized as validation set and test set is officially released `news-test2019`. Training set, validation set and test set are processed consistently. Both Finnish and English sentences are performed tokenization and truecasing with Moses scripts (Koehn et al., 2007). In order to limit the size of vocabulary of NMT models, we adopted byte pair encoding (BPE) (Sennrich et al., 2016) with 50k operations for each side. All the model we trained are optimized with Adam (Kingma and Ba, 2014). Larger beam size may worsen translation quality (Koehn and Knowles, 2017), thus we set beam_size=10 for each model. All models were trained on 4 `NVIDIA V100` GPUs.

In order to find the optimal threshold in cycle translation procedure, we first report our experimental results on validation data set with different thresholds, which ranges from [0%, 25%, 50%, 75%]. Intuitively, the quality improvement of monolingual sentences afforded by cycle translation could bring better synthetic parallel data, subsequently leading to more accurate translation model. Thus, this ablation experiment was trained with synthetic parallel corpus only with different cycle translation ratios on Transformer_base model. As is shown in Table 7, when cycle translation threshold is 50%, the model could achieve the relatively best performance. We therefore set the cycle translation ratio to 50% in our following main experiment.

Our main experiment is shown in Table 6, our baseline system is developed with the $\mathcal{M}_{small}$ configuration using the original parallel corpus and last-20 ensemble strategy. Unsurprisingly, the baseline system relatively performs the worst in Table 6. The $\mathcal{M}_{small}$ configuration trained with selected parallel data improves BLEU by

+0.7 points. According to *exp.*[3-6], adding these components can lead to continuous improvements. Notably, with Cycle Translation and Big/Small parallel construction strategy, our system could obtains +3.55 significant improvement. And *exp.*[8-11] show that with performing GMSE, multi-features reranking, ConMBR system combination and post-processing, our system further improved the BLEU score from 30.9 to 33.0 on the official data set `news-test2019`, which substantially outperforms the baseline by 5.3 BLEU score.

5 Conclusion and Future Work

This paper presents the University of Sydney's NMT systems for WMT2019 Finnish→English news translation task. We leveraged multi-dimensional strategies to improve translation quality in three levels: 1) At data level, in addition to using various data selection criteria, we proposed cycle translation to improve monolingual sentence fluency. 2) For model training, we trained multiple models with R2L corpus and big/small parallel construction corpus respectively. 3) As for inference, we prove the effectiveness of multi-features rescoring, ConMBR system combination and post-processing. We find that cycle translation and B/S construction approach bring the most significant improvement for our system.

In future work, we will apply the beam+noise method (Edunov et al., 2018) to generate robust synthetic data during back translation, we assume that this method combined with our proposed cycle translation strategy can bring greater improvement. Also, we would like to investigate hyper-parameter optimization for neural machine translation to avoid empirical settings.

Acknowledgments

This research was supported by Australian Research Council Projects FL-170100117, DP-180103424 and IH-180100002. We would also thank anonymous reviewers for their comments.

References

Karim Ahmed, Nitish Shirish Keskar, and Richard Socher. 2017. Weighted transformer network for machine translation. *arXiv preprint arXiv:1711.02132*.

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.

B Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01*. IEEE.

Chao Bei, Hao Zong, Yiming Wang, Baoyong Fan, Shiqi Li, and Conghu Yuan. 2018. An empirical study of machine translation for the shared task of WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of NAACL 2012*. Association for Computational Linguistics.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers.

Yongchao Deng, Shanbo Cheng, Jun Lu, Kai Song, Jingang Wang, Shenglan Wu, Liang Yao, Guchun Zhang, Haibo Zhang, Pei Zhang, Changfeng Zhu, and Boxing Chen. 2018. Alibaba's neural machine translation systems for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of EMNLP 2019*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of AAAI 2018*.

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

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of EMNLP 2013*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv: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 ACL 2007*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*. Morgan Kaufmann Publishers Inc.

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*. Association for Computational Linguistics.

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment. In *Proceedings of EACL 2006*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*. Association for Computational Linguistics.

Ioannis Partalas, Grigorios Tsoumakas, and Ioannis P Vlahavas. 2008. Focused ensemble selection: A diversity-based method for greedy ensemble selection. In *ECAI*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of NIPS 2017*.

Khe Chai Sim, William J Byrne, Mark JF Gales, Hichem Sahbi, and Philip C Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. IEEE.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS 2014*.

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

Qiang Wang, Bei Li, Jiqiang Liu, Bojian Jiang, Zheyang Zhang, Yinqiao Li, Ye Lin, Tong Xiao, and Jingbo Zhu. 2018. The NiuTrans machine translation system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *Proceedings of ICLR 2019*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. In *Proceedings of NIPS 2016*.