The University of Maryland's Chinese-English Neural Machine Translation Systems at WMT18

Weijia Xu and **Marine Carpuat**
Department of Computer Science
University of Maryland
College Park, MD 20742, USA
weijia@cs.umd.edu, marine@cs.umd.edu

Abstract

This paper describes the University of Maryland's submission to the WMT 2018 Chinese↔English news translation tasks. Our systems are BPE-based self-attentional Transformer networks with parallel and backtranslated monolingual training data. Using ensembling and reranking, we improve over the Transformer baseline by +1.4 BLEU for Chinese→English and +3.97 BLEU for English→Chinese on *newstest2017*. Our best systems reach BLEU scores of 24.4 for Chinese→English and 39.0 for English→Chinese on *newstest2018*.

1 Introduction

While machine translation between Chinese and English has long been considered a challenging task, with performance lagging behind other language pairs (Bojar et al., 2017), neural architectures have helped achieve large improvements. A new state-of-the-art on Chinese→English news translation was recently obtained (Hassan et al., 2018) using a deep Transformer model in combination with many other techniques including Dual Learning (He et al., 2016), joint training of source-to-target and target-to-source models, and Deliberation Networks (Xia et al., 2017). The resulting high quality translation comes at the cost of large models and complex training pipelines, which make such models difficult to train and deploy with constrained resources.

In this shared task, our goal is to evaluate the performance of systems inspired by Hassan et al. (2018) but with fewer and smaller components, which require less time and memory at training and decoding time. Our systems are based on a multi-layer encoder-decoder architecture with attention mechanism. We experiment with different network architectures, including single-layer RNN, deep Stacked RNN as used in Zhou et al. (2016), and self-attentional Transformer networks (Vaswani et al., 2017). The best results are obtained with deep Transformer models.

Our best systems reach BLEU scores of 24.4 for Chinese→English and 39.0 for English→Chinese on *newstest2018*. Using a combination of backtranslation (Section 2.2), ensembling, and reranking (Section 2.3) we improve over the base Transformer models by +1.4 BLEU (Chinese→English) and +3.97 BLEU (English→Chinese) on *newstest2017*. We describe each component of the system (Section 2), and its contribution for each language pair (Section 4). We show that the impact of backtranslation and reranking is not symmetric in the two translation directions, and that, compared to oracle scores, the reranker leaves much room for improvement.

2 Approach

2.1 Neural Machine Translation Models

Currently, state-of-the-art Neural Machine Translation (NMT) (Bahdanau et al., 2014) is generally based on a sequence-to-sequence encoder-decoder model with attention mechanism, which represent the conditional probability $p(y|x)$ of a target sentence y given a source input x.

This model comprises two components: an encoder Θ_{enc} and a decoder Θ_{dec}. The encoder encodes an input sentence x into a sequence or set of continuous representations, while the decoder predicts the conditional probability distribution of the target words given the encoder's output states. Θ_{enc} and Θ_{dec} are trained to maximize the likelihood of a parallel training data comprised of N pairs of source and target sentences:

$$\mathcal{L}(\Theta) = \sum_{n=1}^{N} \sum_{t=1}^{T} \log p(y_t^{(n)} | h_{t-1}^{(n)}, Attn; \Theta_{dec}) \tag{1}$$

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 539–544
Brussels, Belgium, October 31 - November 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-6431

where

$$Attn = f^{attn}(f^{enc}(x^{(n)}; \Theta enc), h^{(n)}_{t-1}) \quad (2)$$

$h^{(n)}_{t-1}$ denotes the decoder's hidden states conditioned on $y^{(n)}_{<t}$, the target words preceding step t. The attention model f^{attn} computes a weighted sum over the encoder's outputs $f^{enc}(x^{(n)}; \Theta enc)$ where the weights are determined by the "similarity" between each of the encoder's outputs and the decoder's hidden state $h^{(n)}_{t-1}$.

State-of-the-art NMT encoders and decoders include Stacked RNNs (Zhou et al., 2016), convolutional sequence-to-sequence models (ConvS2S) (Gehring et al., 2017), and Transformer models (Vaswani et al., 2017). The ConvS2S and Transformer models differ from RNNs in that they replace the recurrent processing in RNNs with convolutional representation and self-attention respectively, which enable the parallelization of the computation and make the encoded representation less sensitive to the sequence length.

ConvS2S uses stacked convolutional representation to model the dependencies between nearby words on lower layers, while longer-distance dependencies are modeled through upper layers. In contrast, the Transformer model captures source context via self-attention, which allows to attend to any source word regardless of position, and therefore has the potential to model long-distance dependencies more directly.

In addition, the Transformer uses multi-head attention, which lets the model attend to information from different representation subspaces at different positions. The attention function can be interpreted as mapping a query and a set of key-value pairs into an output – the output is generally computed as a weighted sum of the values, and the weights are computed by a function of the query and the corresponding key. Instead of computing a single attention pass, multi-head attention consists of several stacked attention layers in which the same attention function is applied to different transformations of the query, keys and values. And then the output vectors from the above attention layers are concatenated together and linearly transformed, resulting in the final output.

The Transformer model has achieved significant improvements over RNN-based encoder-decoders on several NMT tasks (Vaswani et al., 2017), while RNNs outperform ConvS2S (Hieber et al., 2017). We therefore only experiment with the Transformer and RNN architectures.

2.2 Backtranslating Monolingual Data

We leverage the monolingual data provided in the shared task using backtranslation (Sennrich et al., 2016a). For each language pair, we select monolingual corpora from the target language based on their similarity to the parallel corpus as measured by cross-entropy difference (Moore and Lewis, 2010). Following the setup from Hassan et al. (2018), we backtranslate the monolingual data using a single Transformer model, and then use a mixture of parallel and backtranslated monolingual data with a proportion of 2:1 for training a new Transformer model.

2.3 Reranking n-best Hypotheses

In order to improve the translation quality, we rerank the n-best results using features extracted from different NMT models (Cherry and Foster, 2012; Neubig et al., 2015; Hassan et al., 2018).

Right-to-left NMT Model Sequence-to-sequence models generate sequences on a token-by-token basis, and suffer from the exposure bias problem (Bengio et al., 2015). Exposure bias refers to the problem that models are trained using contexts from human generated references while tested using model-generated contexts, and thus at test time previous mistakes may be amplified and lead to subsequent errors. In order to address this issue, we train a right-to-left (R2L) NMT model using the same training data but with inverted target data. Then for each hypothesis from the n-best list, we invert the hypothesis sequence and use the perplexity score given by the right-to-left NMT model as a reranking feature.

Target-to-source NMT Model In order to improve the translation quality in terms of adequacy, we also use features from target-to-source (T2S) NMT models for reranking. We use the perplexity score given the translation as input and the source sentence as reference. The score represents the conditional probability of the source sentence given the translation, which can be viewed as an adequacy score. Since we participate in both Chinese→English and English→Chinese tasks, we can just use the models trained in the opposite direction for reranking.

Reranking Model First we generate n-best translation hypotheses for each source sentence. Then we get the perplexity scores for each hypothesis with L2R, R2L, and T2S models. The scores are treated as features which we use to train a k-best batch MIRA ranker (Cherry and Foster, 2012) to find out the optimal weights for reranking.

3 Data and Preprocessing

Parallel Data We use all the parallel data available for the shared tasks. The training data for both tasks consists of about 15.8M sentence pairs from the UN Parallel Corpus, 9M sentence pairs from the CWMT Corpora, 332K sentence pairs from the News Commentary Corpus. In addition to the criteria used in Hassan et al. (2018) to filter the parallel data, we add a criterion of bad sentences according to the alignment score given by the `fast-align` toolkit[1]. The overall criteria are the following:

- Duplicate sentence pairs are removed.
- Sentences with characters of other languages are removed.
- Chinese sentences without Chinese characters are removed.
- The length of each sentence must be between 3 and 50.
- The length ratio of sentence pairs must not exceed 1.6.
- Bad sentence pairs according to the alignment score are removed.

Table 1 shows the data statistics after filtering, tokenization, truecasing, and BPE.

Monolingual Data We further augment the training data with backtranslated monolingual data. For Chinese→English systems, we select 8M sentences from "News Crawl: articles from 2017" that are most similar to the bilingual data using cross-entropy difference (Moore and Lewis, 2010). For English→Chinese systems, we select 8M sentences from the XMU Corpus based on the same criteria.

Tuning and Testing Data The official newsdev2017 is used as the validation set, and newstest2017 is used as the test set.

Preprocessing All corpora are processed consistently. We tokenize the English sentences and perform truecasing with the Moses scripts (Koehn et al., 2007). Chinese sentences are segmented with the Jieba segmenter[2]. We segment English and Chinese tokens into subwords via Byte-pair Encoding (BPE) (Sennrich et al., 2016b). We train the BPE models for English and Chinese separately, and use 32K subwords for each side.

4 Experiments

4.1 Baseline systems

The baseline system is a bidirectional RNN with attention mechanism as used in Bahdanau et al. (2014). Our systems are built on Sockeye (Hieber et al., 2017). We use word embedding size of 1024 and hidden layer size of 1024. We filter out sentences with length larger than 50. We use Adam optimizer with initial learning rate of 0.0002. We adopt layer normalization (Ba et al., 2016) and label smoothing (Szegedy et al., 2016). We tie the output weight matrix with the target embeddings (Press and Wolf, 2017). The beam size is set to 10.

The deep RNN is based on Stacked RNNs with attention (Zhou et al., 2016). We use the same system settings as the baseline but set the number of stack layers to 4.

The Transformer network (Section 2.1) is a 6-layer Transformer model with model size of 1024, feed-forward network size of 4096, and 16 attention heads. We adopt label smoothing and weight tying, and set the beam size to 10.

Table 2 shows the total number of parameters for each model and the BLEU scores on Chinese→English and English→Chinese newstest2017. Results show that the Transformer outperforms RNNs in both directions, although it is not a controlled comparison since the Transformer has 1.6 times as many parameters as the deep RNN model. Based on this strong performance, we select the Transformer as the base model for further improvements.

4.2 Results on Chinese→English Translation

Table 3 shows the results for the Chinese→English translation task. We report cased BLEU computed on detokenized output with the `multi-bleu-detok.pl` script. The baseline, deep RNN, and Transformer models are trained on the 17.6M bilingual data. We backtranslate

[1] https://github.com/clab/fast_align

[2] https://github.com/fxsjy/jieba

	train		valid		test	
Sentences	17577153	17577153	2002	2002	2001	2001
Tokens	392490201	433127957	72494	69775	68360	64012
Types	49475	32102	4593	9911	4913	9171
OOVs	–	–	104	32	121	25

Table 1: Data sizes for Chinese/English training (train), validation (valid) and test sets respectively. All statistics are computed after filtering, tokenization, truecasing, and BPE. The *Types* column shows the number of distinct tokens in each data set. The *OOVs* column shows the number of distinct out-of-vocabulary tokens.

System	Size	C→E	E→C
baseline	108.77M	20.99	30.45
deep RNN	165.46M	21.65	31.63
Transformer	259.94M	24.00	34.50

Table 2: BLEU scores for baseline models on Chinese→English and English→Chinese newstest2017. The *Size* column shows the total number of parameters.

System	BLEU
baseline	20.99
deep RNN	21.65
Transformer	24.00
+synthetic	24.12
+ensemble	24.76
+reranking (L2R, T2S)	25.20
+reranking (L2R, T2S, R2L)	25.37
+beam size from 10 to 30	**25.41**

Table 3: Chinese → English Results on newstest2017. The submitted system is the last one.

System	BLEU
baseline	30.45
deep RNN	31.63
Transformer	34.50
+synthetic	36.69
+ensemble	38.28
+reranking (L2R, T2S)	38.19
+reranking (L2R, T2S, R2L)	38.42
+beam size from 10 to 30	**38.47**

Table 4: English → Chinese Results on newstest2017. The submitted system is the last one.

the selected 8M monolingual data using the English→Chinese Transformer model. Training the Transformer model on the mixed parallel/synthetic data improves the model by +0.1 BLEU. We further train 3 independent Transformer models with different random seeds, and gain +0.6 BLEU score by ensembing. Finally, by rescoring the n-best lists with L2R, R2L, and T2S models, we gain +0.6 BLEU score. Increasing the beam size from 10 to 30 also brings improvements when reranking. We submit the last system and get 24.4 BLEU score on the official test set.

4.3 Results on English→Chinese Translation

Table 4 shows the results for the English→Chinese translation task. We report character-based BLEU calculated with the Moses `multi-bleu-detok.pl` script. Similar to the Chinese→English systems, the baseline systems are trained on the parallel data. Augmenting the training data with the backtranslated monolingual data improves BLEU by +2.2 points. The ensemble model improves over the single best system by +1.6 BLEU. Rescoring with L2R, R2L, and T2S models brings an improvement of +0.1 BLEU. We further increase the beam size from 10 to 30 to gain more from reranking. Our submitted system outperforms the best system in WMT17 (Wang et al., 2017) by +2.1 BLEU on newstest2017 and obtains a BLEU score of 39.0 on the official test set.

We note that the components added to the baseline Transformer model have an asymmetric impact in the two translation directions. While backtranslation improves the results by +2.2 BLEU for the English→Chinese task, it doesn't help as much for Chinese→English (+0.1). In contrast, rescoring with L2R, R2L, and T2S models brings more improvements for Chinese→English (+0.6) than the other (+0.2). One possible explanation is that in a parallel corpus sentences originally written in language A and sentences translated from language B to A may have different styles due to translationese effects (Volansky et al., 2015).

While the original language is not known for all training documents, it seems reasonable to assume that the majority of documents are translated from English into Chinese: the UN corpus is known to comprise primarily original English documents

Beam	C→E		E→C	
	reranker	oracle	reranker	oracle
10	25.37	28.72	38.42	42.84
30	**25.41**	30.44	**38.47**	44.78
100	25.40	**33.05**	38.38	**47.17**

Table 5: A comparison of BLEU scores when using the reranker trained with L2R, R2L, and T2S features versus the oracle, with varying beam sizes.

(Tolochinsky et al., 2018). For other training data sources beyond UN, a bilingual Chinese-English speaker manually inspected a random sample of 100 sentence pairs, and estimated that 87% sentences were originally written in English. This might explain why rescoring with the T2S models helps more in the Chinese→English direction than in the other, and why the English→Chinese systems benefit more from backtranslated data which introduces some (machine) translated Chinese to complement the translation direction observed in the parallel training data.

4.4 Experiments on Reranking

To estimate an upper-bound for reranking methods, we build an oracle that returns the translation in the n-best list that gets the highest BLEU score.

Table 5 shows the comparison of BLEU scores when using the reranker trained with L2R, R2L, and T2S features versus the oracle. Increasing the beam size from 30 to 100 doesn't improve the results when using the reranker, but improves the oracle scores. This is consistent with prior findings that beam search only improves translation quality for narrow beams and deteriorates with larger beams (Koehn and Knowles, 2017), but differs in that we rerank the n-best lists instead of adopting the 1-best results from beam search. The results also show that better translations according to BLEU exist in the n-best lists with larger beam size, but are ranked low by the models.

In addition, we find that the oracle scores are always higher than the reranker scores, and the gap increases with beam size. When comparing the MSR's best system results (28.46 BLEU achieved by Combo-4 in Hassan et al. (2018) with the oracle, we find that the oracle score is still higher by 4-5 BLEU. The results show that there is room for improvement by introducing more useful rescoring features and warrant further investigation.

5 Conclusion

This paper presents the University of Maryland's NMT systems for WMT 2018 Chinese↔English news translation tasks. Our experiments confirm the benefits of using Transformer networks over RNN-based architectures. We report performance gains from incorporating monolingual data, using ensemble models and reranking with target-to-source and right-to-left models, although the impact of these techniques depends on the translation direction. By comparing the oracle and reranking results, we find that there is potential for further improvement with more useful rescoring features.

Acknowledgments

Thanks to Xing Niu for his initial scripts, and to the anonymous reviewers for their valuable comments. This work is partially funded by an AWS Machine Learning Research Award.

References

Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *CoRR*, abs/1607.06450.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1171–1179. Curran Associates, Inc.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *CoRR*, abs/1803.05567.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 820–828. Curran Associates, Inc.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden.

Graham Neubig, Makoto Morishita, and Satoshi Nakamura. 2015. Neural reranking improves subjective quality of machine translation: Naist at wat2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 35–41.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163.

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.

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 Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Elad Tolochinsky, Ohad Mosafi, Ella Rabinovich, and Shuly Wintner. 2018. The UN parallel corpus annotated for translation direction. *CoRR*, abs/1805.07697.

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.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. 2017. Sogou neural machine translation systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 410–415.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. 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 1784–1794. Curran Associates, Inc.

Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. 2016. Deep recurrent models with fast-forward connections for neural machine translation. *Transactions of the Association for Computational Linguistics*, 4:371–383.