NTT's Neural Machine Translation Systems for WMT 2018

Makoto Morishita, Jun Suzuki, Masaaki Nagata
NTT Communication Science Laboratories, NTT Corporation, Japan

Innovative R&D by NTT

Abstract

Parallel and synthetic corpora filtering are essential.
→ The Transformer model might be largely affected by noisy data.

The Transformer model might need more data than the RNN-based model.
→ The Transformer has too many parameters to train with a small corpus.

1. Transformer

Fast

It does not rely on RNNs.
More GPU efficient architecture.

Training on Multi-GPUs

Almost no overhead.
RNN with 1 GPU → 31 days
Transformer with 8 GPUs → 3 days

Better Accuracy

It surpasses the RNN when trained with a large corpus and well-tuned hyperparameters.

Ref: Attention Is All Youu Need, Vaswani et al., NIPS 2017

2. Parallel Corpus Filtering

qe-clean toolkit

A toolkit for automatically cleaning parallel sentences.
It uses a language model and a word alignment model to evaluate how clean the parallel sentence is.

Lanugage Model

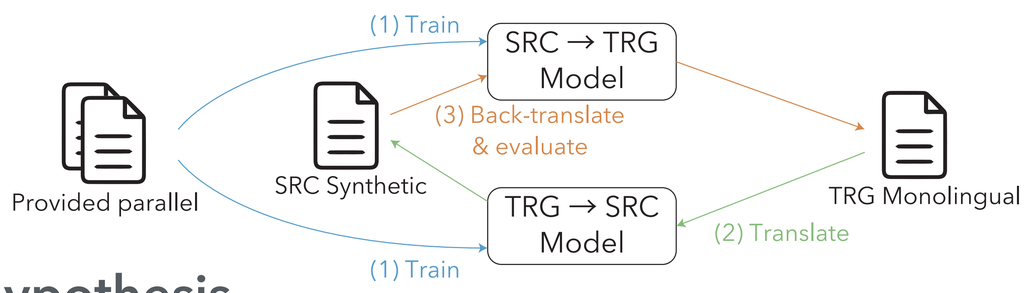
To evaluate how grammatical the sentence is.

Word Alignment Model

To evaluate how correctly the sentence has been translated.

Ref: The CMU-Avenue French-English Translation System, Denkowski et al., WMT 2012

3. Synthetic Corpus Filtering



Hypothesis

The correct synthetic sentence can be back-translated to the original sentence.

Method

Check the back-translation BLEU+1 scores and filter out the low scored sentence pairs.

Ref: Improving Low-Resource Neural Machine Translation with Filtered Pseudo-Parallel Corpus, Imankulova et al., WAT 2017

4. Right-to-Left Re-Ranking

Problem

NMT uses previous predictions as context information.
It lacks reliability when decoding the latter side of the sentence.

Method

Rescore n-best translations of Left-to-Right model by Right-to-Left model, which generates a sentence from the end to the beginning of the sentence.

Ref: Agreement on Target-Bidirectional LSTMs for Sequence-to-Sequence Learning, Liu et al., AAAI 2016

Experimental Results

Parallel Corpus Filtering

Parallel corpus filtering works well.
We could see up to 11.3 points improvement.

Synthetic Corpus Filtering

Synthetic corpus filtering also helps.
We could see up to 3.5 points improvement.
Average back translation BLEU+1 score is largely improved by filtering.

	En-De	De-En
Unfiltered	44.02	53.96
Filtered	80.12	80.81

Average back translation BLEU+1 scores of the synthetic corpus

	Settings	En-De Sentences	En-De Transformer	En-De RNN	De-En Sentences	De-En Transformer	De-En RNN
(1)	Europarl + News Commentary + Rapid	3.10M	32.5	30.4	3.10M	31.0	31.0
(2)	(1) + unfiltered Common Crawl + ParaCrawl	38.66M	26.6	—	38.66M	32.7	—
(3)	(1) + filtered Common Crawl + ParaCrawl	7.11M	37.9	**39.6**	7.11M	37.5	**39.6**
(4)	(3) + unfiltered synthetic corpus	45.05M	41.5	—	32.97M	46.4	—
(5)	(3) + filtered synthetic corpus	14.22M	45.0	39.8	14.22M	46.3	43.7
(6)	(5) + R2L re-ranking (submission)	14.22M	**46.5**	—	14.22M	**46.8**	—

RNN vs. Transformer

RNN sometimes surpass the Transformer.
This might be due to the corpus size is too small to train the Transformer.
The Transformer has more than twice parameters than the RNN has.
This makes difficult to train with the smaller corpus.