

NTT Neural Machine Translation Systems at WAT 2017

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



General Settings

Attentional NMT based on Luong et al.
Byte Pair Encoding (vocab: 16k)
2 layers encoder-decoder
- Embed, Hidden, Attention = 512 units

Epochs: 20
SGD (Learning rate = 1.0, decay after 13 epochs)
Minibatch: 128 sentences
Japanese side tokenizer: KyTea
English side tokenizer: Moses tokenizer

Our Implementation

<https://github.com/nttclab-nlp/wat2017>

Try it!

ASPEC (Scientific Paper)

English-to-Japanese

System	Training data	BLEU	Pairwise	Adequacy
Single	3.0M (original)	37.15	-	-
Single	2.0M (original)	37.90	-	-
Single	2.0M (original) + 1.0M synthetic	38.87	-	-
8 Ensemble	2.0M (original)	39.80	72.250	-
8 Ensemble	2.0M (original) + 1.0M synthetic	40.32	75.750	4.41

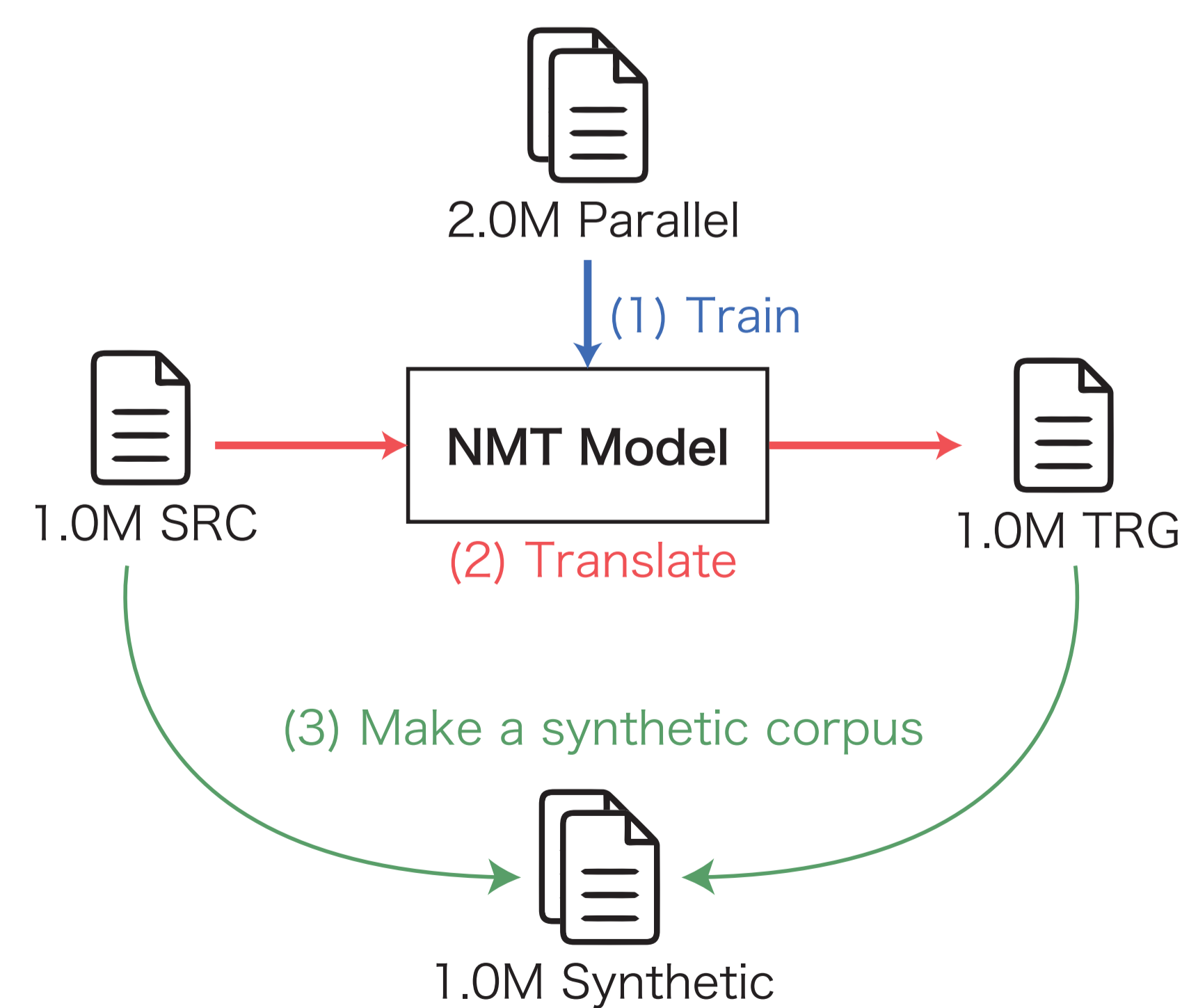
Japanese-to-English

System	Training data	BLEU	Pairwise	Adequacy
Single	3.0M (original)	26.07	-	-
Single	2.0M (original)	27.43	75.000	-
Single	2.0M (original) + 1.0M synthetic	27.62	-	-
8 Ensemble	2.0M (original)	28.36	77.250	4.14
8 Ensemble	2.0M (original) + 1.0M synthetic	28.15	-	-

Noisy part on the corpus

Q Should we use the whole corpus?

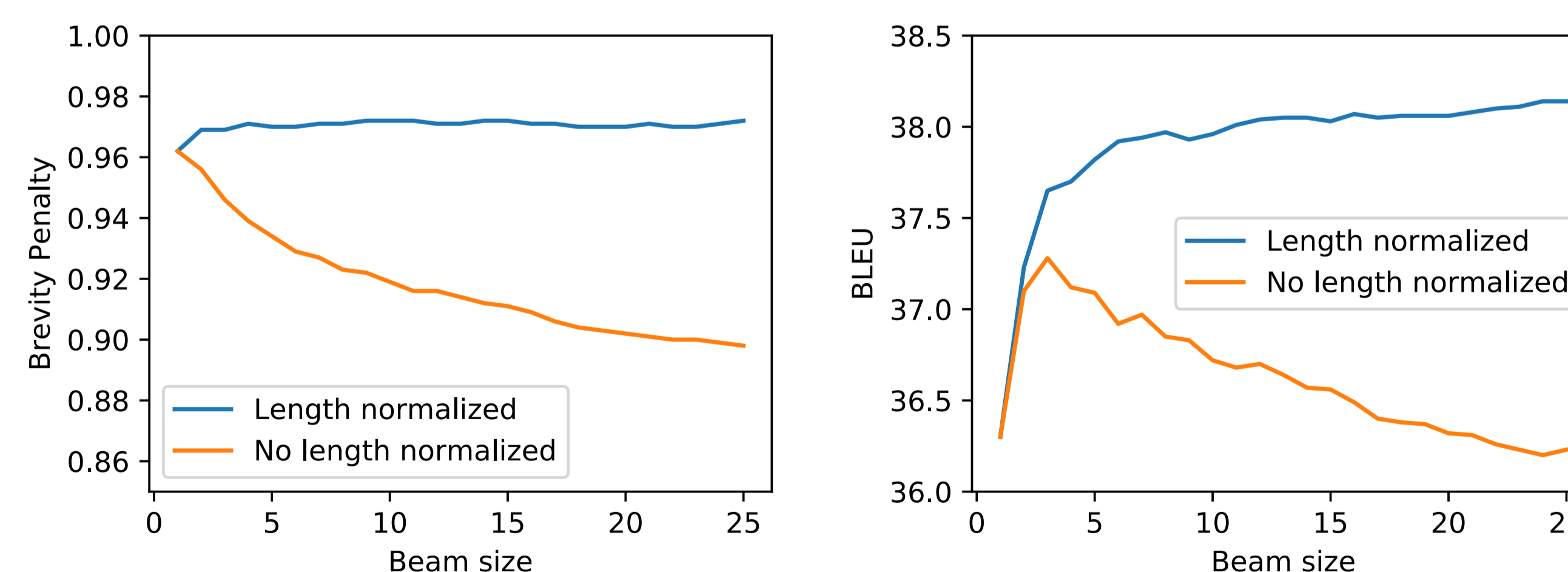
A No! The latter side is noisy.
We only need first 2.0M sentences.
Make synthetic corpus for the latter side.



Length-based score normalization

Beam search with a large beam size tends to select shorter sentences. To rescore the scores, use the following formula proposed by Cromieres et al.

$$\hat{t} = \arg \max_{t \in \mathcal{t}} \left\{ \frac{p(t)}{|t|} \right\}$$



Length-based score normalization works well.

With this normalization, we can use larger beam size with keeping the length of the sentence.

Refs: Improving Neural Machine Translation Models with Monolingual Data, Sennrich et al., ACL 2016
Kyoto University Participation to WAT 2016, Cromieres et al., WAT 2016

JJI Corpus (Newspaper)

Direction	System	BLEU	Pairwise	Adequacy
En→Ja	Single	19.13	14.500	-
	8 Ensemble	20.37	17.750	2.03
Ja→En	Single	19.44	32.000	2.05
	8 Ensemble	20.90	26.750	-

Difficulties on newspaper domain

Direction	System	BLEU	Pairwise
En→Ja	Online-A	11.29	69.750
	RBMT-A	5.31	31.250
	NTT	20.37	17.750

Q We achieved the highest BLEU score, but lower on the human evaluation. Why?

A It's due to the noise on the corpus.

Source	The two leaders initially planned to only send messages, without attending the events.
Target (Original)	韓国の朴槿恵大統領も22日のソウルでの祝賀行事出席を見送る方針を示していた。 (Korean President Park Geun-hye also indicated that she will not be presenting celebratory events in Seoul on the 22nd.)

There are a lot of incorrect aligned sentence pairs.

- NMT model tries to fit these data, leading to drop the human evaluation score.
- We may need to consider how to train a model with noisy parallel corpus like this.