NTT Neural Machine Translation Systems at WAT 2017

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

Participated tasks

Scientific paper (ASPEC)

- Japanese-to-English, English-to-Japanese
- Training data: 3.0M

Newspaper (JIJI)

- Japanese-to-English, English-to-Japanese
- Training data: 200k

Overview of our systems

Innovative R&D by NTT

General settings

- Attentional NMT
- Byte Pair Encoding
- 2 layers encoder-decoder
  - Embed, Hidden, Attention = 512 Units
- SGD (Learning rate = 1.0, decay after 13 epochs)

Effective approaches to attention- based neural machine translation, Luong et al., EMNLP 2015

Neural Machine Translation of Rare Words with Subword Units, Sennrich et al., ACL 2015

NTT

Overview of our systems

Features

- Synthetic corpus for noisy sentences
- Length-based score normalization
- Model ensembling

Noisy part on ASPEC

- ASPEC has been collected by aligning sentences automatically
  - It is sorted by alignment scores
  - Latter sentence pairs are often noisy

- We separated ASPEC as follows:
  - First 2.0M as clean
  - Latter 1.0M as noisy

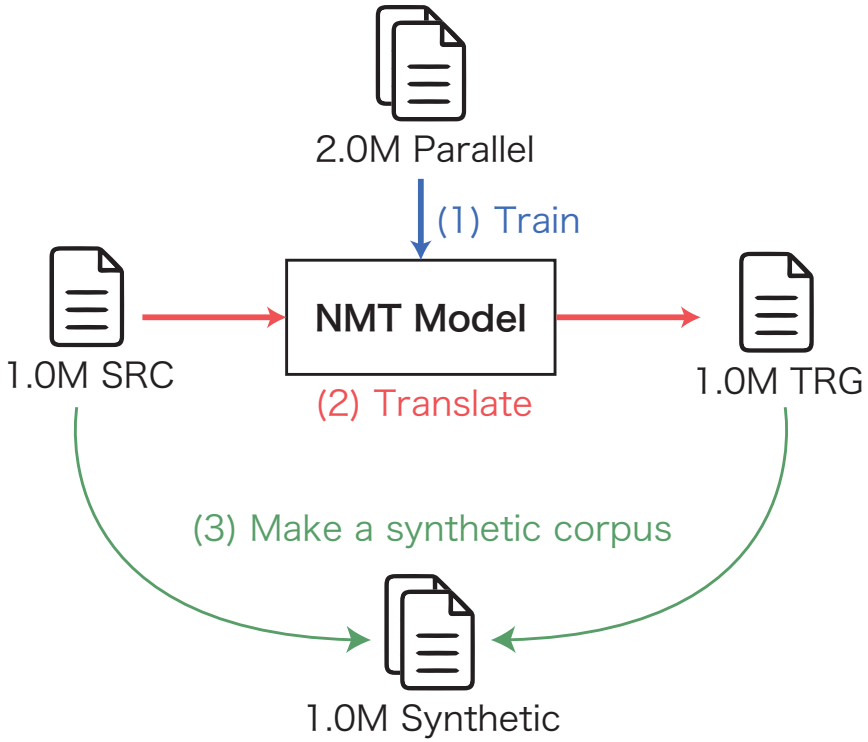
How to use noisy part on ASPEC

- Previously
  - SMT: use it for training LM
  - NMT: throw away

- We use these noisy part by making synthetic corpus

Improving Neural Machine Translation Models with Monolingual Data, Sennrich et al., ACL 2016

How to make a synthetic corpus



1. Train NMT model with reliable parallel corpus.
2. Translate unreliable part of the corpus.
3. Make a synthetic corpus and re-train NMT model.

Example of synthetic corpus

Source	The search procedure utilizes a nonlinear least squares method coupled with the method of steepest descent.
Target (Original)	また，具体的な探索の手順を示した。 (We also show the specific search procedure.)
Target (Synthetic)	探索手順は最急降下法と結合した非線形最小二乗法を用いた。 (The search procedure utilizes a nonlinear least squares method coupled with the method of steepest descent.)

Synthetic sentence seems to be better than original sentence.

Length-based score normalization

- Beam search with a large beam size tends to select shorter sentences.
  - We need to normalize the scores
  - Length-based score normalization

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

- Divide the score by the length
  - Simple but effective method
  - Proposed by Cromieres et al., WAT 2016

Kyoto University Participation to WAT 2016, Cromieres et al., WAT 2016

Experimental Results and Analysis

Experimental results on ASPEC (En-Ja)

System	Training data	BLEU	Pairwise	Adequacy
Single	3.0M (original)	37.15	–	–
Single	2.0M (original)	**37.90**	–	–

Q Should we use the whole corpus?

A No! The latter side is noisy.
We only need first 2.0M sentences.

Experimental results on ASPEC (En-Ja)

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**	–	–

How should we use the noisy part of the corpus?

Make a synthetic corpus!

Experimental results on ASPEC (En-Ja)

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**

Q Should we ensemble the models?

A Yes!

Experimental results on ASPEC (En-Ja)

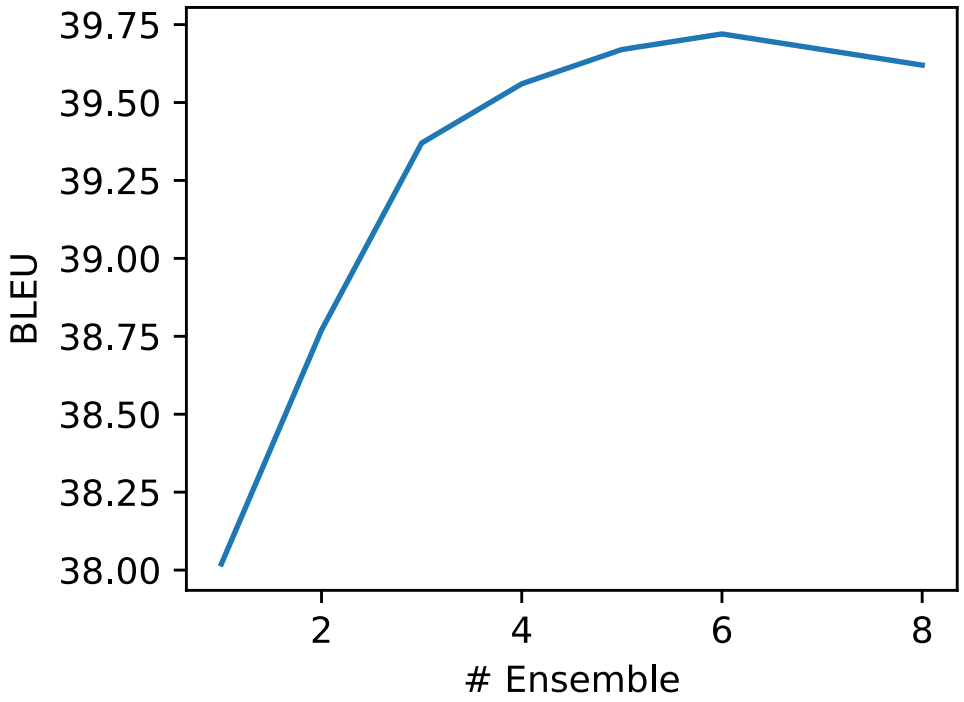
System	Training data	BLEU	Pairwise	Adequacy
Single	3.0M (original)	37.15	–	–
Single	2.0M (original)	37.90	–	–
Single	2.0M (original) + 1.0M synthetic	38.87	–	–
8 Ensemble	2.0M (original)	39.80	72.250	
8 Ensemble	2.0M (original) + 1.0M synthetic	**40.32**	**75.750**	**4.41**

Q Should we ensemble the models?

A Yes!

1st Place!

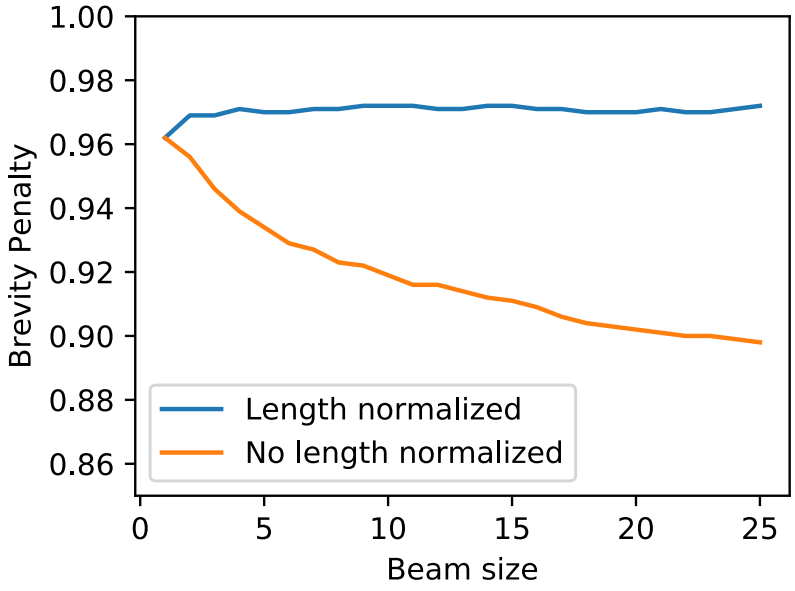
Ensembling

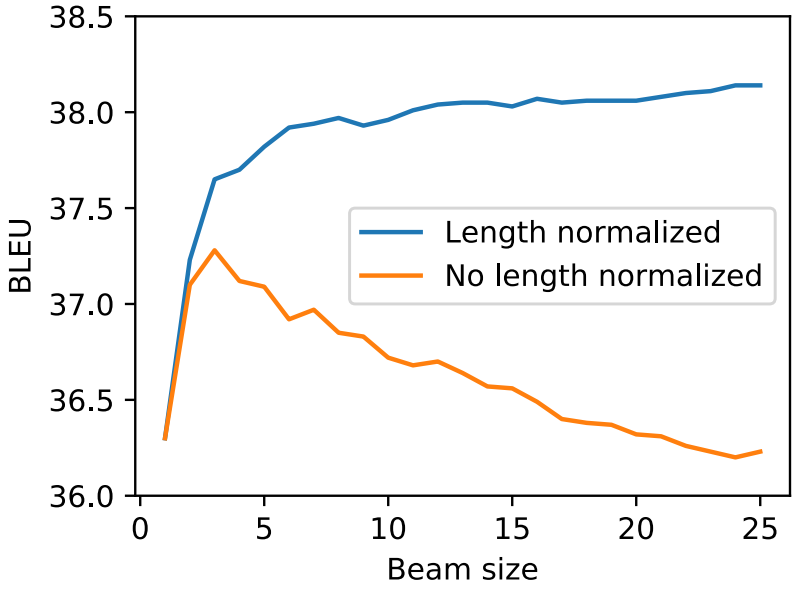


Model ensembling improves the BLEU score.

- But the impact gradually decreased.

Length-based score normalization





Q Should we use length-based score normalization?

A Yes!

Experimental results on ASPEC (Ja-En)

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	–	–

Experimental results on ASPEC (Ja-En)

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	–	–

1st Place!

Experimental results on JIJI

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	–

Model ensembling gains BLEU scores.

- But lower on the Ja-En pairwise evaluation.

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

Actually, we lose to 2 commercial systems.

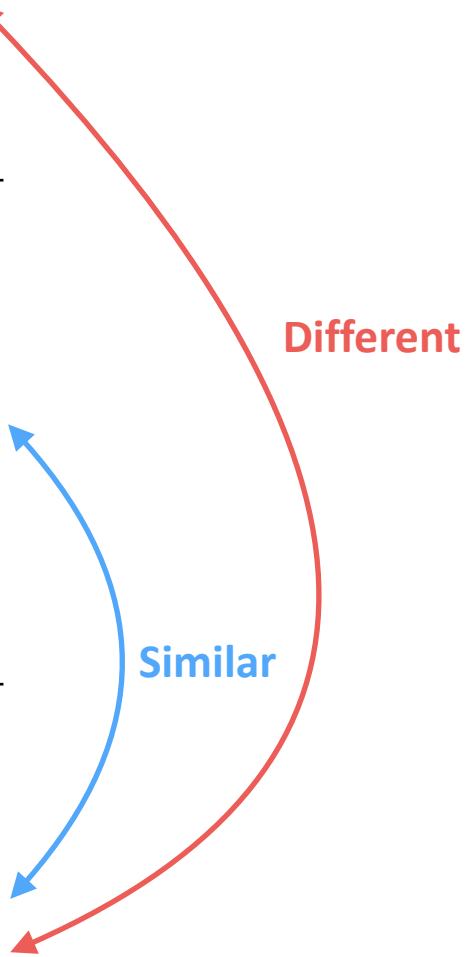
- But we won w.r.t. BLEU score.
- Why?

Difficulties on newspaper domain

The answer is **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.)

Source	NKSJ aims to rebuild itself through the merger, as it projects a group net loss of about 100 billion yen for the business year ending March 31, due to insurance payouts related to the massive flooding in Thailand last autumn.
Target (Original)	ＮＫＳＪは、タイの洪水被害に伴う保険金支払いがかさみ、１２年３月期は連結純損益が約１０００億円の赤字と２期連続で赤字の見通し。 (NKSJ projects a group net loss of about 100 billion yen for the business year 2012 ending March 31, due to insurance payouts related to the massive flooding in Thailand.)
Translation	ＮＫＳＪは昨年秋、タイ洪水に伴う保険金支払いに伴う保険金の支払いなどで純損益が１００００億円程度の赤字になるとの見通しを示している。 (NKSJ projects a group net loss of about 100 billion yen due to insurance payouts related to the massive flooding in Thailand last autumn.)



Noise on the corpus

- There are a lot of sentence pairs like this.
- Our model achieves higher BLEU score.
  - But low score on human evaluation.
- JIJI corpus is one of the most difficult corpus to train.

Our implementation is now available!

Innovative R&D by NTT

https://github.com/nttcslab-nlp/wat2017

You can easily re-build our NMT model.

NTT

Conclusion

- Synthetic corpus and length-based score normalization work effectively.

- Future work
  - Think how to train the system with noisy parallel corpus.

- Our implementation: https://github.com/nttcslab-nlp/wat2017

END