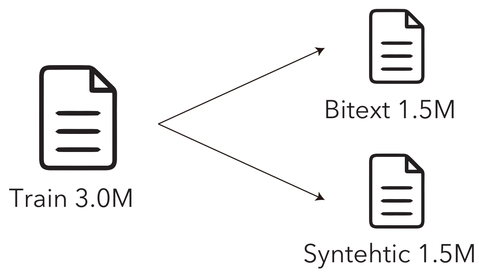
NTT Neural Machine Translation Systems at WAT 2019

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

Scientific Paper Subtask

Corpus Splitting



ASPEC is ordered by sentence alignment scores.
→ Latter sentences are noisy.
→ We split into two parts, use the former one as bitext, the latter one as synthetic data.

Building Synthetic Data

Corpus	En-Ja	Ja-En
w/o synthetic data	44.2	29.9
Back-translation	**45.6**	29.5
Forward-translation	—	**30.1**

Normally, we use back-translation (TRG→SRC) for creating synthetic data.
However, we found that, in our Ja-En case, forward-translation works better than back-translation.

Vocabulary Size

Subword size	En-Ja	Ja-En
4,000	**45.6**	**30.1**
8,000	45.3	29.9
16,000	45.2	29.6
32,000	45.0	29.7

In NMT training, we usually set the vocabulary size to around 32k or larger.
However, we found the smaller vocabulary size may work well in some cases.

Mini-batch Size

Mini-batch size	En-Ja	Ja-En
16 × 4,000 tokens	45.1	29.7
32 × 4,000 tokens	45.3	29.9
64 × 4,000 tokens	45.4	29.8
128 × 4,000 tokens	**45.6**	**30.1**
256 × 4,000 tokens	45.4	29.9

Ref: Scaling Neural Machine Translation, Ott et al., EMNLP 2018

It is known that the Transfomer models tend to work well with a larger mini-batch size.

However, we found that overly large mini-batch degraded the performance.

Ensemble and Right-to-Left Re-ranking

Model type	En-Ja	Ja-En
Single model	45.6	30.1
Ensemble (4)	46.2	30.8
Ensemble (4) + R2L (4)	46.8	**31.2**
Ensemble (6) + R2L (4)	**46.9**	**31.2**

Ensemble and R2L re-ranking are standard techniques for further improving the translation quality.

We used four L2R models and four R2L models for decoding.

Official Results

	Auto Eval		Human Eval			
Lang. pair	BLEU	(Rank)	Pairwise	(Rank)	Adequacy	(Rank)
En-Ja	45.83	(1)	47.75	(1)	**4.50**	**(1)**
Ja-En	30.56	(5)	14.00	(1)	**4.49**	**(2)**

We achieved the best performance for both directions in terms of the pairwise evaluation.
In terms of the adequacy, our En-Ja model ranked first, and the Ja-En model ranked second.

Unconstrained Submission using JParaCrawl

Training data	# sentences	En-Ja	Ja-En
ASPEC	3.0M	45.83	30.56
ASPEC + JParaCrawl	3.0M * 2 + 7.5M	**46.57**	**31.23**

ASPEC is upsampled twice.

JParaCrawl is our new large web crawled data-based parallel corpus.
Adding JParaCrawl to ASPEC improves the performance up to +0.74.
JParaCrawl will be released for the public soon!

http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

Timely Disclosure Subtask

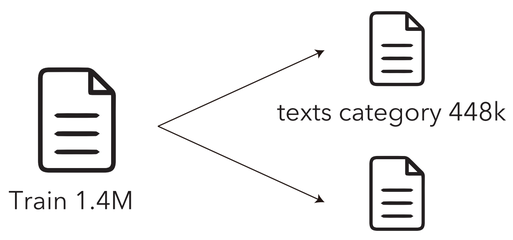
Task

The task focuses on translating the Japanese company's announcements for investors into English.

Difficulties:
- It contains a lot of figures and proper nouns that are critical for readers.

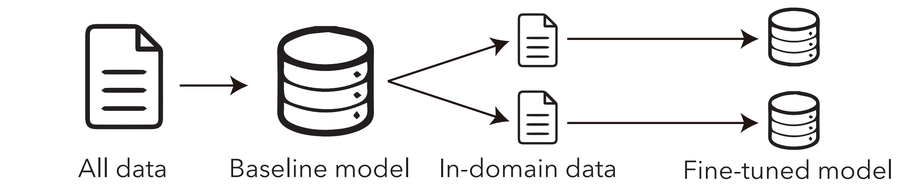
This task is separated into **two subtasks**:
- **Items**: contains subjects, table titles, and bullet points.
- **Texts**: contains full sentences.

Split Training Data



Training data is not separated into sub-categories, so we split it by checking whether the sentence ends with the Japanese period "。" .

Task Specific Fine-tuning



Model type	Texts	Items
Baseline	55.26	54.58
+ Fine-tune	**58.91**	**56.14**

Our baseline model is trained with the whole provided data.
Then fine-tune the model using the in-domain data.
It significantly improves performance..

Ensemble and Right-to-Left Re-ranking

Model type	Texts	Items
Single model	58.91	56.14
Ensemble (4)	60.48	56.93
Ensemble (4) + R2L (4)	**61.19**	**57.34**

We used the ensemble and R2L re-ranking techniques.
We could see additional gains for both subtasks, +2.28 for texts and +1.20 for items compared with the single model.

Official Results

	Auto Eval		Human Eval			
Task	BLEU	(Rank)	Pairwise	(Rank)	Adequacy	(Rank)
Texts	61.19	(1)	55.50	(1)	**4.46**	**(1)**
Items	57.34	(1)	34.00	(2)	**4.47**	**(1)**

* Shown ranks are ordered among only constrained submissions.

Our systems achieved the best performance in terms of the BLEU scores and the adequacy evaluation for both subtasks.