# Context-aware Neural Machine Translation with Mini-batch Embedding

Makoto Morishita[1, 2], Jun Suzuki[2], Tomoharu Iwata[1], Masaaki Nagata[1]

[1]NTT Communication Science Laboratories, [2]Tohoku University

## Abstract

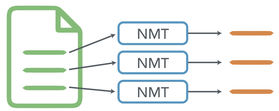
Document-level context is essential for a better translation, though current context-aware NMT models are usually complexed and/or slow.

→ We proposed a mini-batch embedding that contains the features of a mini-batch (document).

→ We feed that embedding into an NMT model to achieve a simple and fast context-aware machine translation.

## Neural Machine Translation

### Context-aware NMT




Single-sentence NMT

Input/Output: Single sentence
+ Fast
+ Simple
- NMT cannnot consider the document-level context.


Context-aware NMT

Input/Output: Document
+ NMT considers the context.
- Slow
- Complexed model.

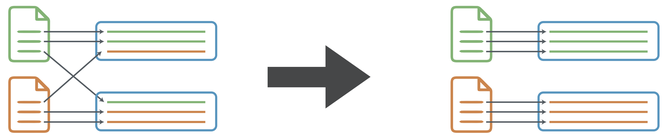
### Mini-batch


Without mini-batch | With mini-batch

NMT normally uses a mini-batch for training/decoding.
Several inputs are concatenated and computed together.
Requires few numbers of matrix computations.

## Context-aware NMT with Mini-batch Embedding

### Document-level Mini-batching



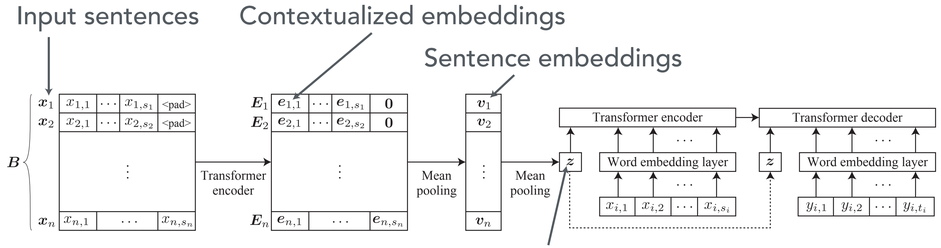
Normal mini-batching

→ Randomly chosen sentences across the document will be in a mini-batch.

Document-level mini-batching

→ Sentences from the same document will be in the same mini-batch

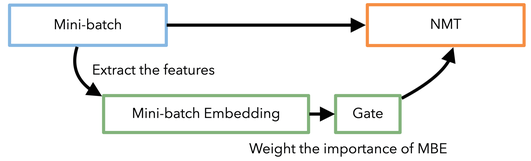
### Mini-batch Embedding




Mini-batch embedding (MBE)
Added as a tag to the downstream task (NMT)

We propose a mini-batch embedding (MBE)
→ includes features of sentences in a mini-batch
→ with the document-level mini-batching, MBE inlcudes document-level features

### Mini-batch Embedding Gate




We add a gate to weight the importance of MBE
→ If the model failed to extract the features of the mini-batch, the weight would be small

## Experiments

### Experimental Settings

Language pair: English-to-Japanese
Training data: JParaCrawl (10M sentences)
Test data:
- ASPEC (Scientific paper)
- WMT (News)
- IWSLT (TED Talk)

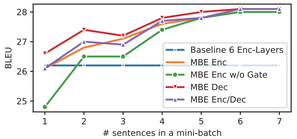
NMT model: Transformer (big)

### Translation Performance

| Model | ASPEC | WMT | IWSLT |
|---|---|---|---|
| Single-sentence NMT | 26.2 | 18.4 | 12.0 |
| 2-to-1 | 27.0 (+0.8) | 19.2 (+0.8) | 12.9 (+0.9) |
| DocRepair | 27.9 (+1.7) | 19.3 (+0.9) | 12.3 (+0.3) |
| MBE Enc | 28.0 (+1.8) | 19.9 (+1.5) | 12.2 (+0.2) |
| MBE Enc w/o Gate | 28.0 (+1.8) | 19.4 (+1.0) | 13.0 (+1.0) |
| MBE Dec | **28.1 (+1.9)** | 19.9 (+1.5) | **13.8 (+1.8)** |
| MBE Enc/Dec | **28.1 (+1.9)** | **20.0 (+1.6)** | 13.4 (+1.4) |

Our model surpassed the single-sentence baseline and the current context-aware NMT methods.
→ Feeding the MBE to the decoder-side works better

### Effect of Decoding Batch-size




As we increase the mini-batch size, our model achieves better performance.
→ The large mini-batch size makes the MBE more informative