Context-aware Neural Machine Translation with Mini-batch Embedding

Makoto Morishita, Jun Suzuki, Tomoharu Iwata, Masaaki Nagata

Context-aware NMT

NTT

- Conventional NMT translates each sentence independently
- Human evaluation
  - Single sentence NMT < Document-level human translation

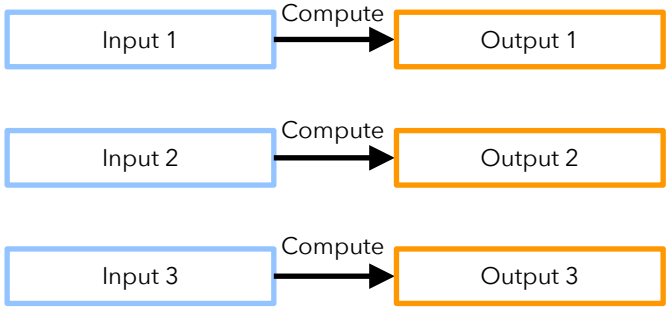
[Läubli et al., EMNLP 2018]

- Current context-aware NMT methods
  - Consider only a few previous sentences
  - Require a large modification to the NMT model

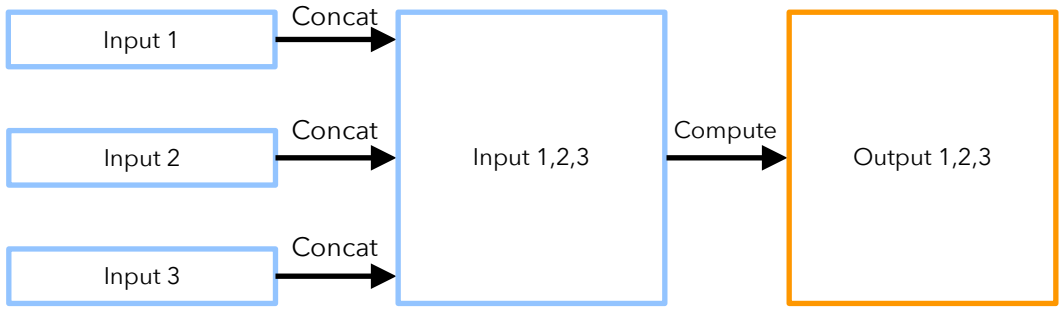
Proposed Method

Mini-batch

- NMT normally uses a mini-batch for training/decoding
- Several inputs are concatenated and computed together
- Requires few numbers of matrix computations
  - Results in faster computation on GPUs



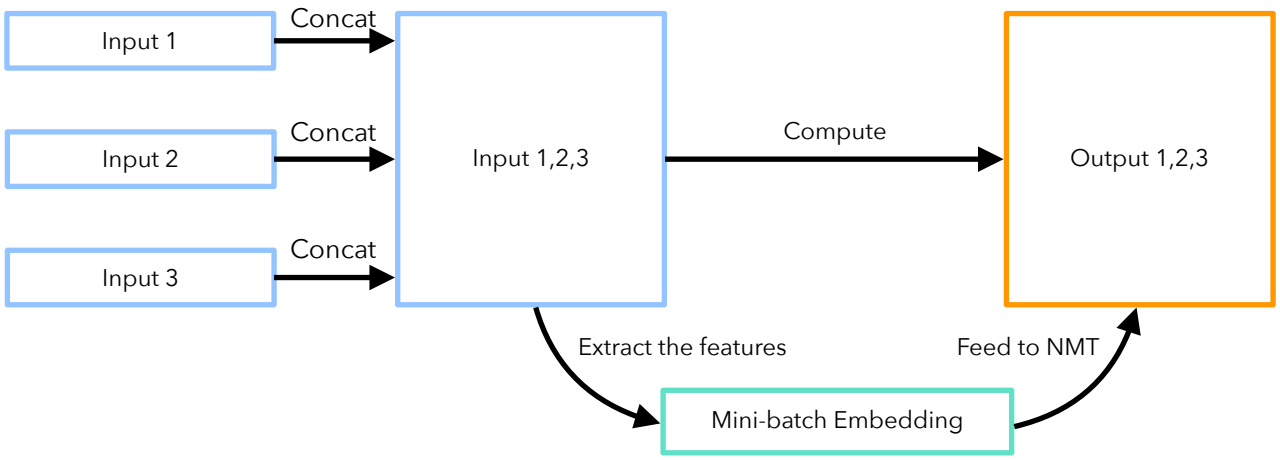
Without mini-batch



With mini-batch

Mini-batch Embedding

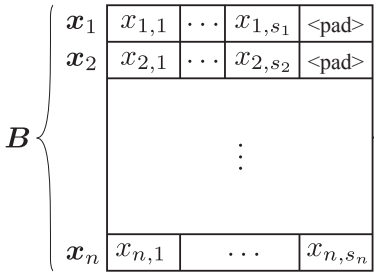
- We propose a mini-batch embedding
  - Include features of the mini-batch
- Document-level mini-batching
  - Sentences from the same document will be in the same mini-batch
  - Mini-batch embedding will include features of the document



How to Make a Mini-batch Embedding

Input sentence Output sentence

Mini-batch $\boldsymbol{B} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$



How to Make a Mini-batch Embedding

Input sentence Output sentence

Mini-batch $\boldsymbol{B} = \underbrace{\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}}_{\text{From the same document}}$

B

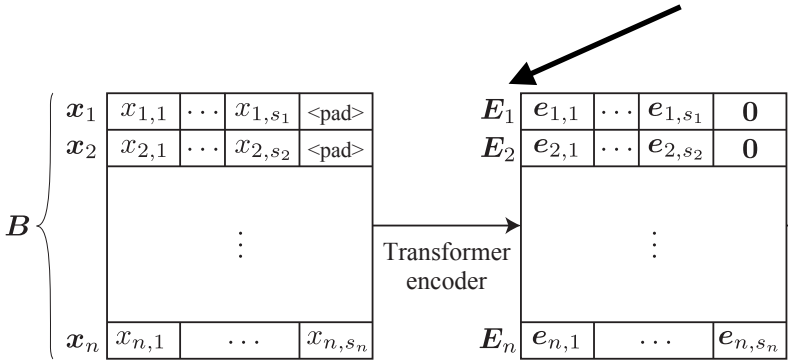
$\boldsymbol{x}_1$	$x_{1,1}$	$\cdots$	x_{1,s_1}	<pad>
$\boldsymbol{x}_2$	$x_{2,1}$	$\cdots$	x_{2,s_2}	<pad>
		$\vdots$		
$\boldsymbol{x}_n$	$x_{n,1}$	$\cdots$	x_{n,s_n}	

How to Make a Mini-batch Embedding

Input sentence Output sentence

Mini-batch $\boldsymbol{B} = \underbrace{\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}}_{\text{From the same document}}$

Contextualized embeddings $\boldsymbol{E}_i = (\boldsymbol{e}_{i,1}, \ldots, \boldsymbol{e}_{i,s_i})$



Uses single-layer Transformer encoder

How to Make a Mini-batch Embedding

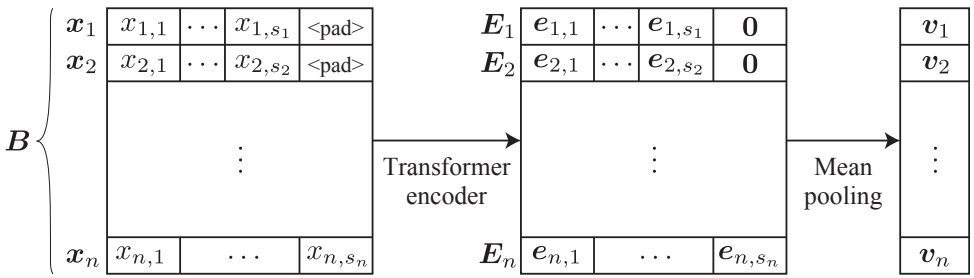
Input sentence Output sentence

Mini-batch $\boldsymbol{B} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$

From the same document

Sentence embeddings $\boldsymbol{v}_i$

Contextualized embeddings $\boldsymbol{E}_i = (\boldsymbol{e}_{i,1}, \ldots, \boldsymbol{e}_{i,s_i})$



Uses single-layer Transformer encoder

How to Make a Mini-batch Embedding

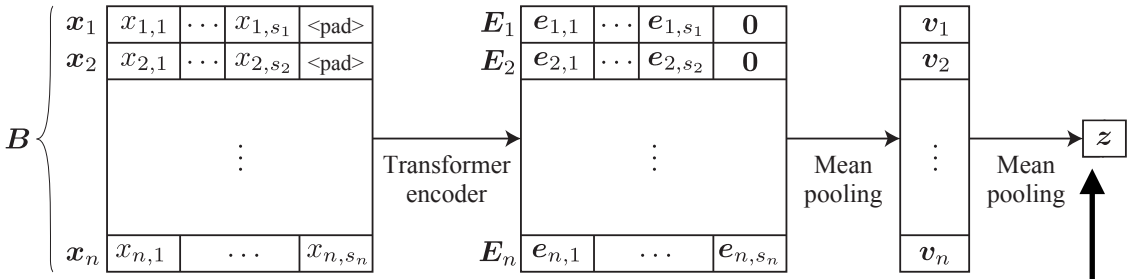
Input sentence Output sentence

Mini-batch $\boldsymbol{B} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$

From the same document

Sentence embeddings $\boldsymbol{v}_i$

Contextualized embeddings $\boldsymbol{E}_i = (\boldsymbol{e}_{i,1}, \ldots, \boldsymbol{e}_{i,s_i})$



Uses single-layer Transformer encoder

Mini-batch embedding $\boldsymbol{z}$

How to Make a Mini-batch Embedding

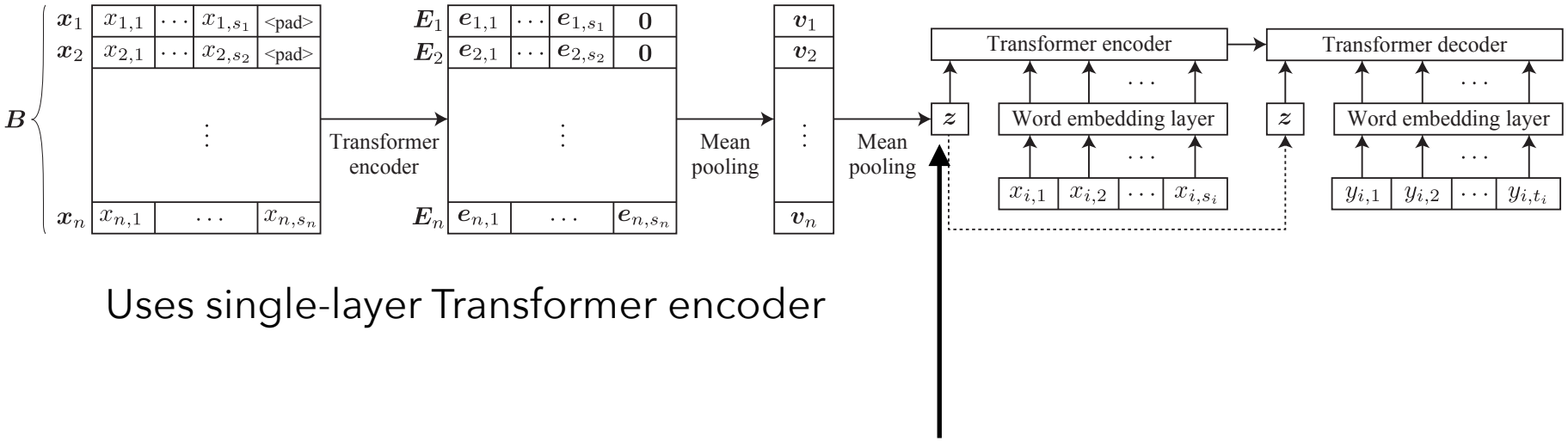
Input sentence Output sentence

Mini-batch $\boldsymbol{B} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$

From the same document

Sentence embeddings $\boldsymbol{v}_i$

Contextualized embeddings $\boldsymbol{E}_i = (\boldsymbol{e}_{i,1}, \ldots, \boldsymbol{e}_{i,s_i})$



Uses single-layer Transformer encoder

Mini-batch embedding $\boldsymbol{z}$

Added as a tag to the downstream task (NMT)

Experiments

Experimental Settings

NTT

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

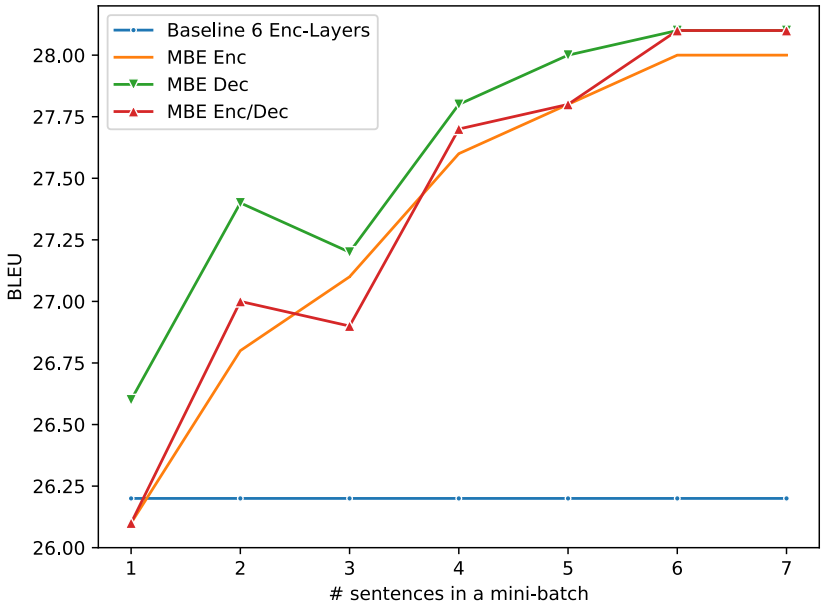
Experimental Results

NTT

Model	ASPEC	WMT	IWSLT
Single-sentence NMT	26.2	18.4	12.0
Add mini-batch embedding to Enc	28.0 (+1.8)	19.9 (+1.5)	12.2 (+0.2)
Add mini-batch embedding to Dec	**28.1 (+1.9)**	19.9 (+1.5)	**13.8 (+1.8)**
Add mini-batch embedding to Enc/Dec	**28.1 (+1.9)**	**20.0 (+1.6)**	13.4 (+1.4)

- Our model surpassed the single-sentence baseline
- Feeding the mini-batch embedding to the decoder-side works better
- We also compared with previous context-aware baselines
  - Please refer to the paper for more results

Effect of Mini-batch Size



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

Conclusion

- We proposed the mini-batch embedding and applied it to the context-aware NMT.
  - A simple, fast, and effective context-aware NMT method that considers a whole document context.
- Future work
  - Apply the mini-batch embedding to other tasks

END