

# A Bag of Useful Tricks For Neural Machine Translation: Embedding Layer Initialization and Large Batch Size

Masato Neishi<sup>\*1</sup>, Jin Sakuma<sup>\*1</sup>, Satoshi Tohda<sup>\*1</sup>, Shonosuke Ishiwatari<sup>1</sup>, Naoki Yoshinaga<sup>2</sup>, Masashi Toyoda<sup>2</sup>  
 {neishi, tohda, jsakuma, ishiwatari, ynaga, toyoda}@tkl.iis.u-tokyo.ac.jp  
<sup>1</sup>The University of Tokyo <sup>2</sup>Institute of Industrial Science (IIS), the University of Tokyo <sup>\*</sup>contributed equally

## System Overview of Team UT-IIS

### Task:

ASPEC English - Japanese

### Approach:

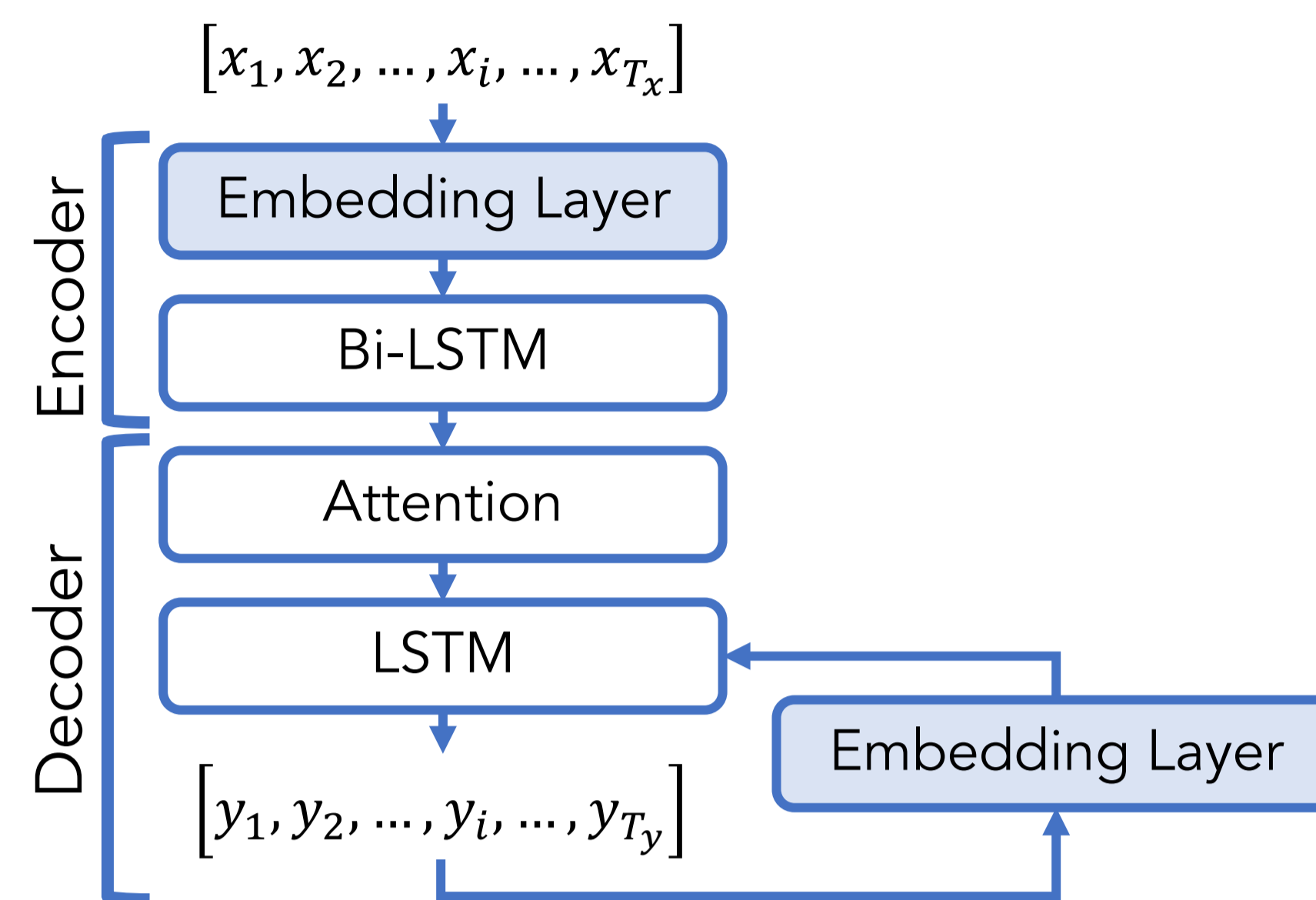
Seq2seq model with attention [Bahdanau+, 2015]

**+ Model independent tricks**

### Result:

Evaluation Metric	Score
BLEU (KyTea)	38.93
Human Evaluation	68.000

### Seq2seq with Attention:



### Tricks:

#### Training Phase

- Adam Optimization [Kingma and Ba, 2015]
- Subword Translation (SentencePiece)
- Embedding Layer Initialization
- Large Batch Size

#### Prediction Phase

- Ensemble of 8 Models
- Beam Search (width=256)

## Embedding Layer Initialization

### Background:

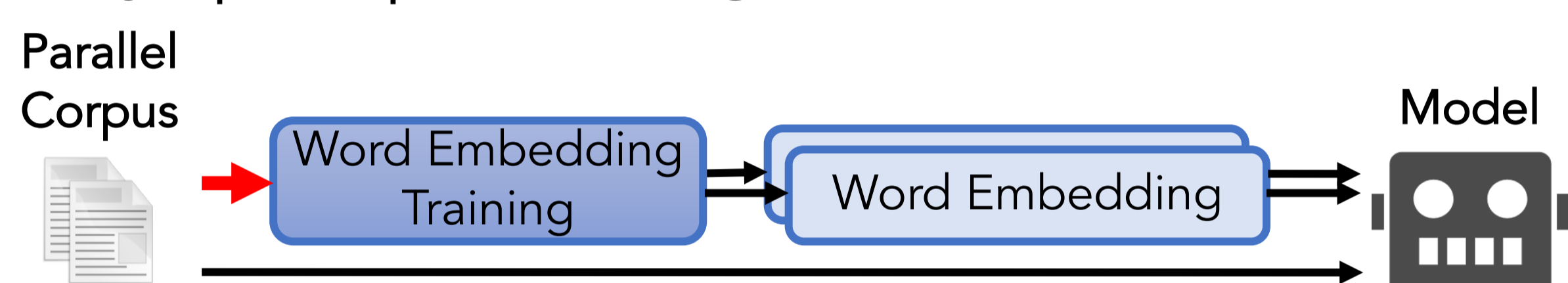
Ramachandran+ (2017) obtained a significant BLEU gain by using LSTM to pretrain layers of a network on a **large external corpus** (one week on 32 GPUs).

Is external data necessary?

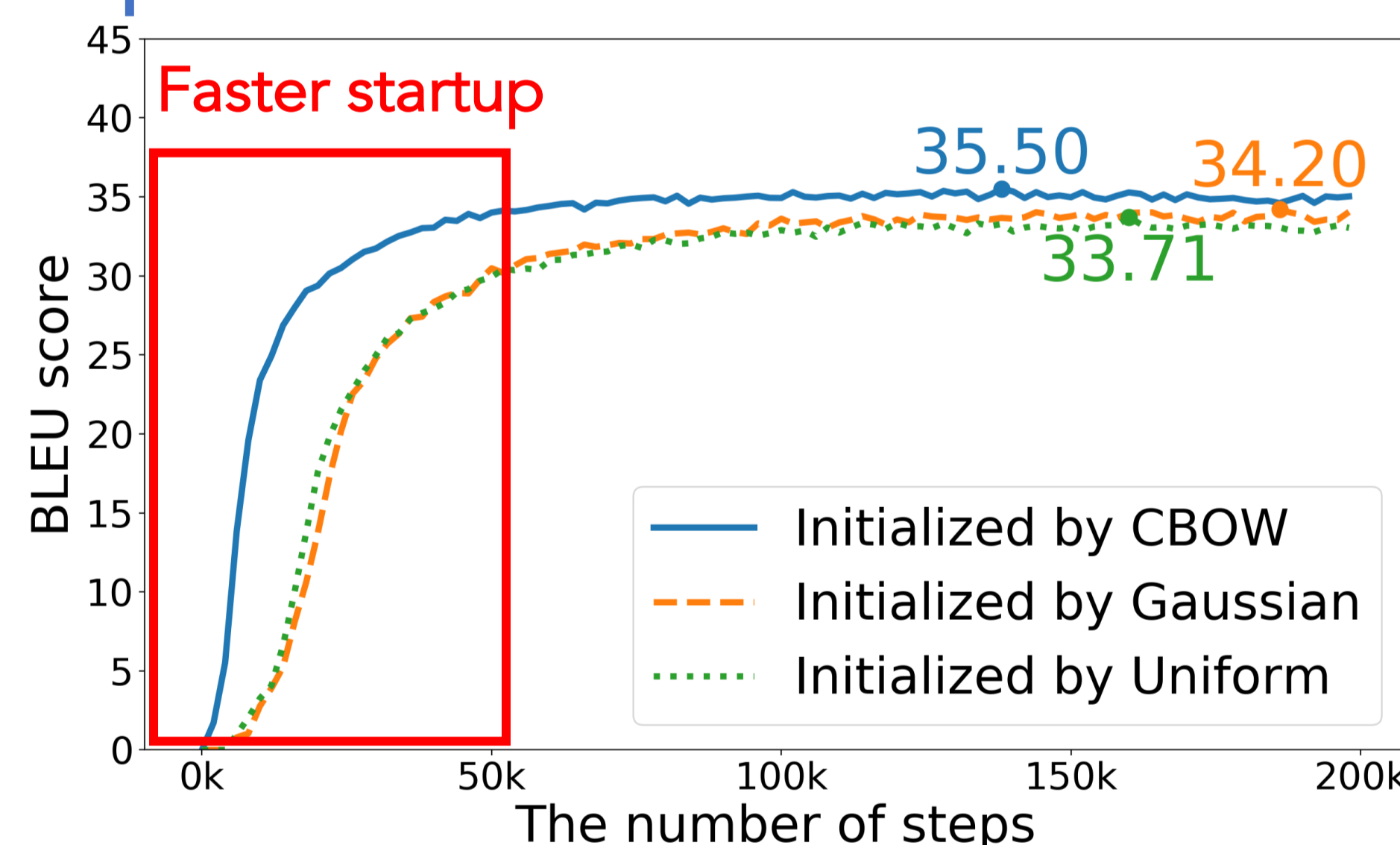
### Proposal:

Pretrain word embeddings from **the training data only**

- No additional resources
- Very quick pretraining (less than 30 min on a CPU)



### Experiments:



Method	BLEU	Diff.
Random (Gaussian)	34.20	-
<b>CBOw</b>	<b>35.50</b>	<b>+1.30</b>
Skip-gram	34.44	+0.24
SI-Skip-gram	34.44	+0.24
GloVe	34.58	+0.38

- **+1.30 BLEU** (34.20 w/ Gaussian → 35.50 w/ CBOw)
- CBOw performed best among initialization methods

## Large Batch Size

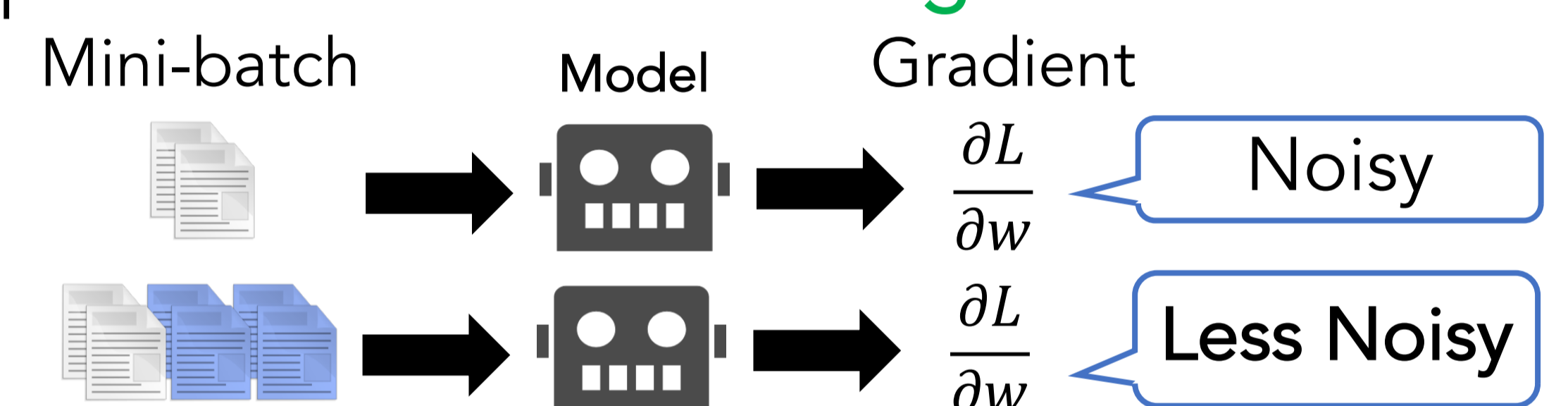
### Background:

Morishita+ (2017) tested larger batch sizes, **up to 64**, for improvements in (mini-batch) training NMT.

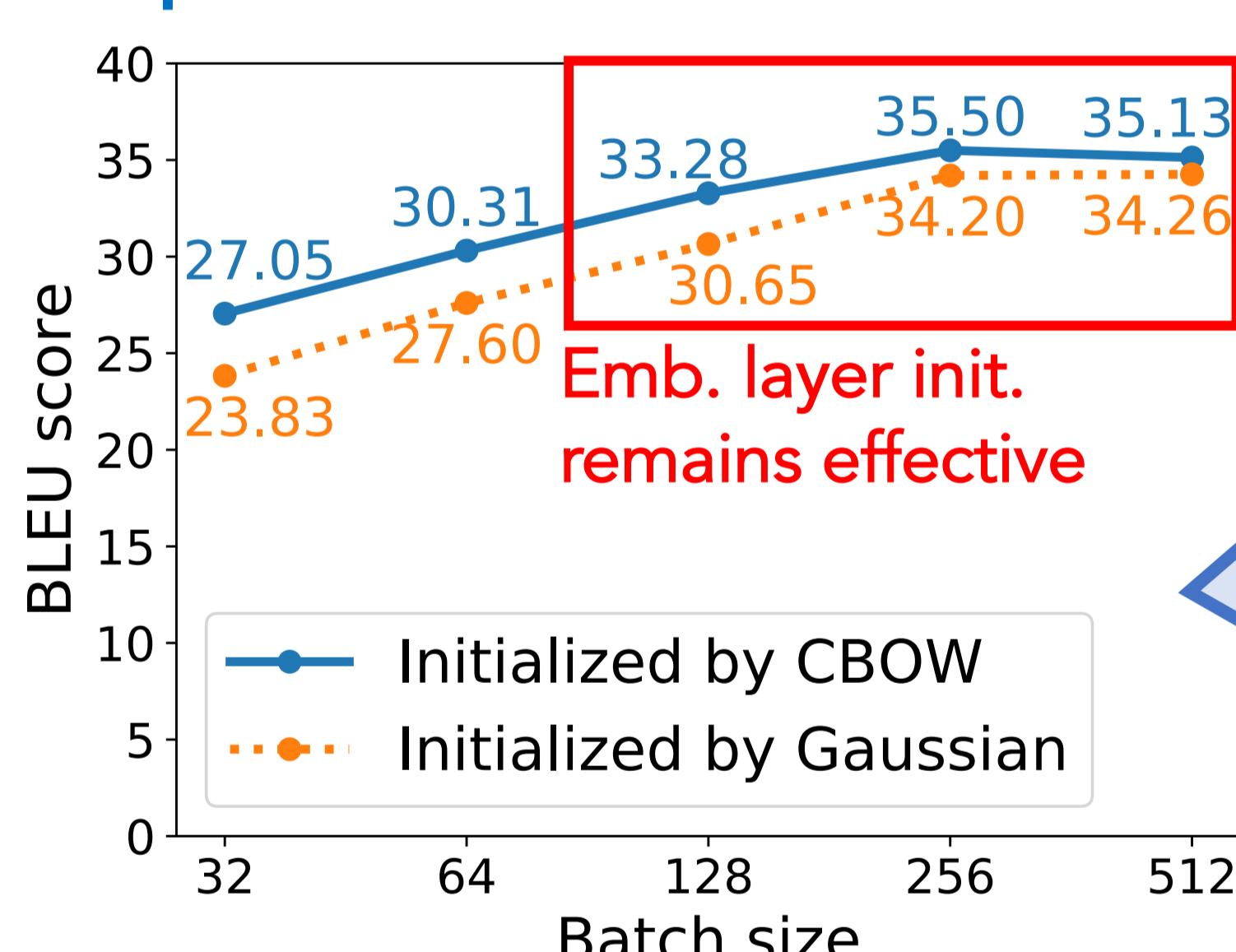
How about even larger batch sizes?

### Proposal:

Test whether translation quality will continue to improve with batch sizes **larger than 64**



### Experiments:



### Tradeoff

- Pros:
- Better optimal
- Cons:
- More memory needed
  - Slower update

- **+5.19 BLEU** (30.31 @64 → 35.50 @256)
- Effect of large batch size saturates at 256

## Overall Result

Tricks	BLEU	Gain
Baseline (existing tricks)	23.83	-
+ Embedding Layer Initialization	27.05	+3.22
+ Larger Batch Size	35.50	+11.67
+ Ensemble of 8 Models	38.00	+14.17
+ Beam Search (width=256)	38.93	+15.10

Training Tricks  
Prediction Tricks

- **+3.43 BLEU** by prediction tricks
- **+15.10 BLEU** against vanilla seq2seq
- The tricks are simple and applicable to **any NMT model**

Code available: <https://github.com/nem6ishi/wat17>

## Translation Examples

Source	Doping induced a noticeable change at the lower boundary of the three-dimensional orderly vortex phase.
Reference	ドーピングにより三次元規則的渦糸相の下界に著しい変化が生じた。
No tricks	ドーピングドーピングは三次元の秩序渦相の低下で注目可能な変化を誘導する。
W/ tricks	ドーピングは三次元秩序渦相の下部境界で顕著な変化を誘起した。
Source	The outline of the 23rd white paper which is used to be issued yearly from Ministry of Posts and Telecommunications is specified.
Reference	郵政省が毎年発表し、今回23回目当たる標記白書の概要を述べた。
No tricks	st上から行われているまでに使用されるまでに使用されている23rd白色紙の概要を特定した。
W/ tricks	郵政省から発行されている第23回白書の概要を述べた。
Source	It has been already entering into the ubiquitous society, and the diffusion of the portable telephone is over 70% of the total population.
Reference	すでに、ユビキタス社会の入り口にあり、携帯電話は総人口の70%以上の普及率である。
No tricks	既にユビキタス社会に入院し、携帯電話の普及は全人口の70%以上である。
W/ tricks	既にユビキタス社会に入り、携帯電話の普及率は全人口の70%以上である。