A BAG OF USEFUL TRICKS FOR PRACTICAL NEURAL MACHINE TRANSLATION

M. Neishi*, J. Sakuma*, S. Tohda*, S. Ishiwatari (The University of Tokyo) *Contributed Equally

N. Yoshinaga, M. Toyoda (IIS, the University of Tokyo)

CONTENT

Asian
Universities
Alliance

Overview 01

About Paper

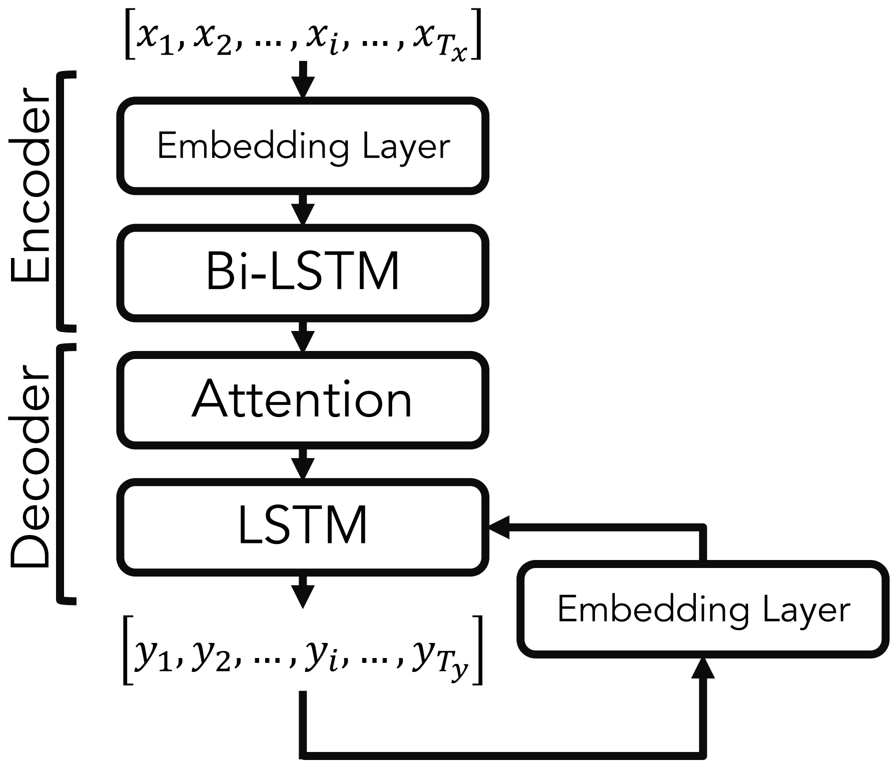
Original Paper

- A system description paper for The 4th Workshop on Asian Translation (WAT 2017)

Summary

- Proposed novel tricks for Neural Machine Translation (NMT)
  - Model-independent
  - Easy to apply
- Apply all the possible tricks to a vanilla NMT system
- Outperformed best score of WAT 2016

System Overview



Task:

ASPEC En-Ja Translation

Model:

Seq2seq model with attention [Bahdanau+, 2015]

+ Model Independent Tricks

Approaches

- Trick used when:
  - Training the model
    - Adam Optimization [Kingma and Ba, 2015]
    - Sub-word Translation (SentencePiece)
    - **Embedding Layer Initialization**
    - **Large Batch Size**

Novel Tricks

  - Prediction
    - Exhaustive Ensemble Search
    - Beam Search

Asian Universities Alliance

02 Proposed Tricks

Novel Tricks for a Better Optimum

– **Embedding Layer Initialization:**

Good initialization should lead to fast convergence to a good local optimum

– **Large Batch Size:**

Tested improvements for sizes up to **512 sents**

Novel Tricks for a Better Optimum

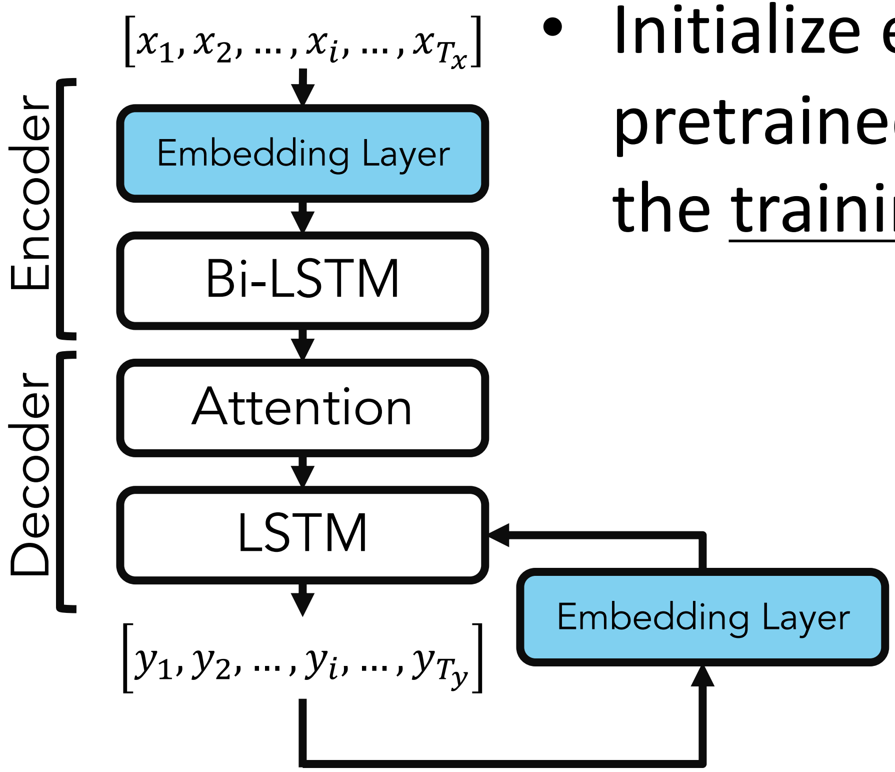
– **Embedding Layer Initialization:**

Good initialization should lead to fast convergence to a good local optimum

– **Large Batch Size:**

Tested improvements for sizes up to **512 sents**

Embedding Layer Initialization



- Initialize embedding layers with pretrained embeddings induced from the training data

Pretraining on a large external corpus [Ramachandran+ 2017]

Easy to apply:
- No additional resources
- Very quick pretraining

Novel Tricks for a Better Optimum

- **Embedding Layer Initialization:**

  Good initialization should lead to fast convergence to a good local optimum

- **Large Batch Size:**

  Tested improvements for sizes up to **512 sents**

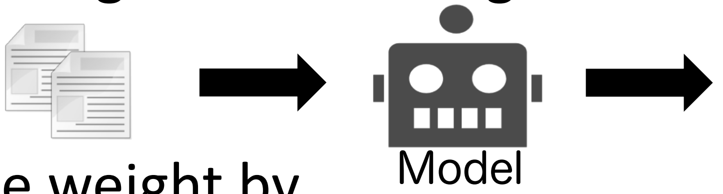
Small Batch makes Update Noisy

In a step of SGD (and its variance):

1. Take small portion of data (batch)

32 sents

2. Compute gradient of weights on batch



$$\frac{\partial L}{\partial w}$$

Noisy gradient

3. Update weight by

$$w \leftarrow w - \frac{\partial L}{\partial w}$$

Noisy update

Small Batch makes Update Noisy

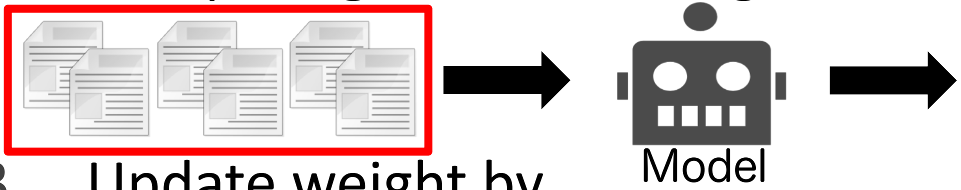
~64 sentences
[Morishita+ 2017]

In a step of SGD (and its variance):

1. Take small portion of data (batch)

32~512
sents

2. Compute gradient of weights on batch



$$\frac{\partial L}{\partial w}$$

Less noisy
gradient

3. Update weight by

$$w \leftarrow w - \frac{\partial L}{\partial w}$$

Less noisy
update

Asian
Universities
Alliance

Experiments 03

Experiments

1. **Effect of Initialization Methods:**
   Will the proposed method speed up convergence and improve translation quality?

2. **Effect of Large Batch Size:**
   Will large batch sizes (32 to 512) improve translation quality?

Experiment Setup

- Training
  - 200k steps (save checkpoint at every 2k)
  - Checkpoint with highest BLEU score (in dev) is used in evaluation
- Evaluation
  - KyTea segmentation to compute the BLEU score
  - Greedy search for experiments

Experiments

1. **Effect of Initialization Methods:**
   Will the proposed method speed up convergence and improve translation quality?

2. **Effect of Large Batch Size:**
   Will large batch sizes (32 to 512) improve translation quality?

Effect of Initialization Methods

– Purpose:
  Investigate the effect of embedding layer initialization using CBOW embeddings

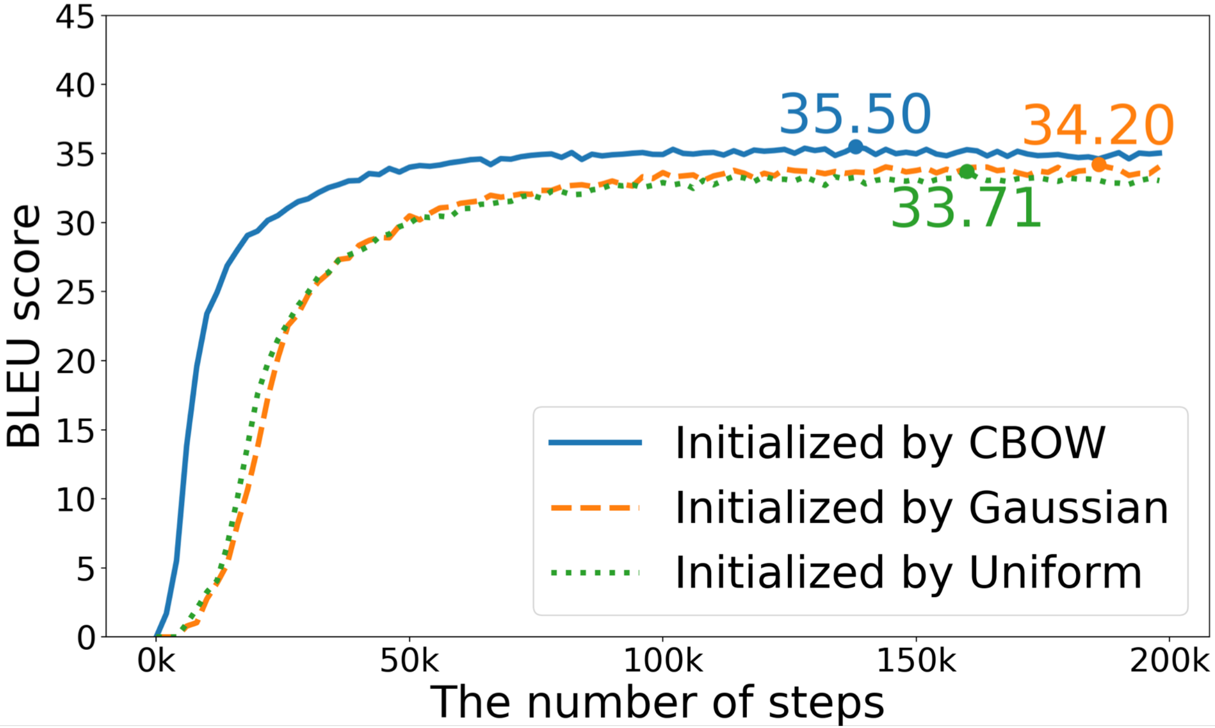
– Compare:
  - CBOW embeddings
  - Random initialization (Gaussian Distribution)
  - Random initialization (Uniform Distribution)

Best performance among:

CBOW	[Mikolov+ 2013]
Skip-gram	[Mikolov+ 2013]
GloVe	[Pennington+ 2014]
SI-Skip-gram	[Bojanowski+ 2017]

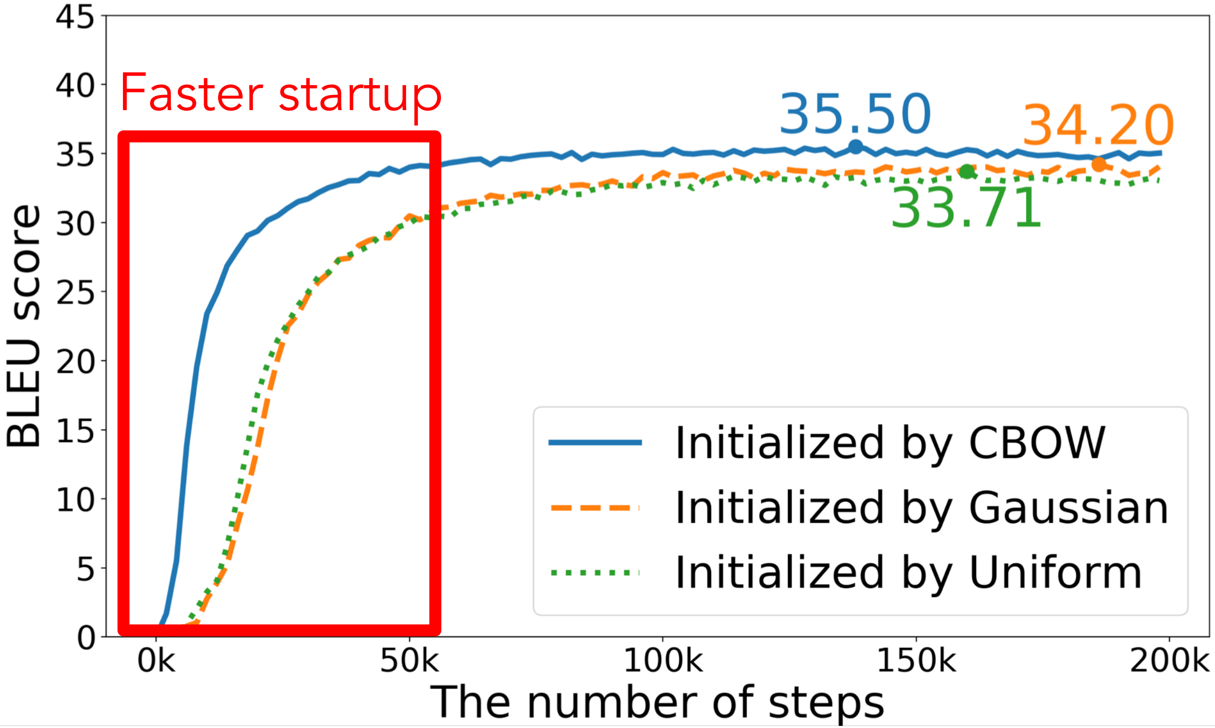
Effect of Initialization Methods: Results

Training curves for different initialization methods

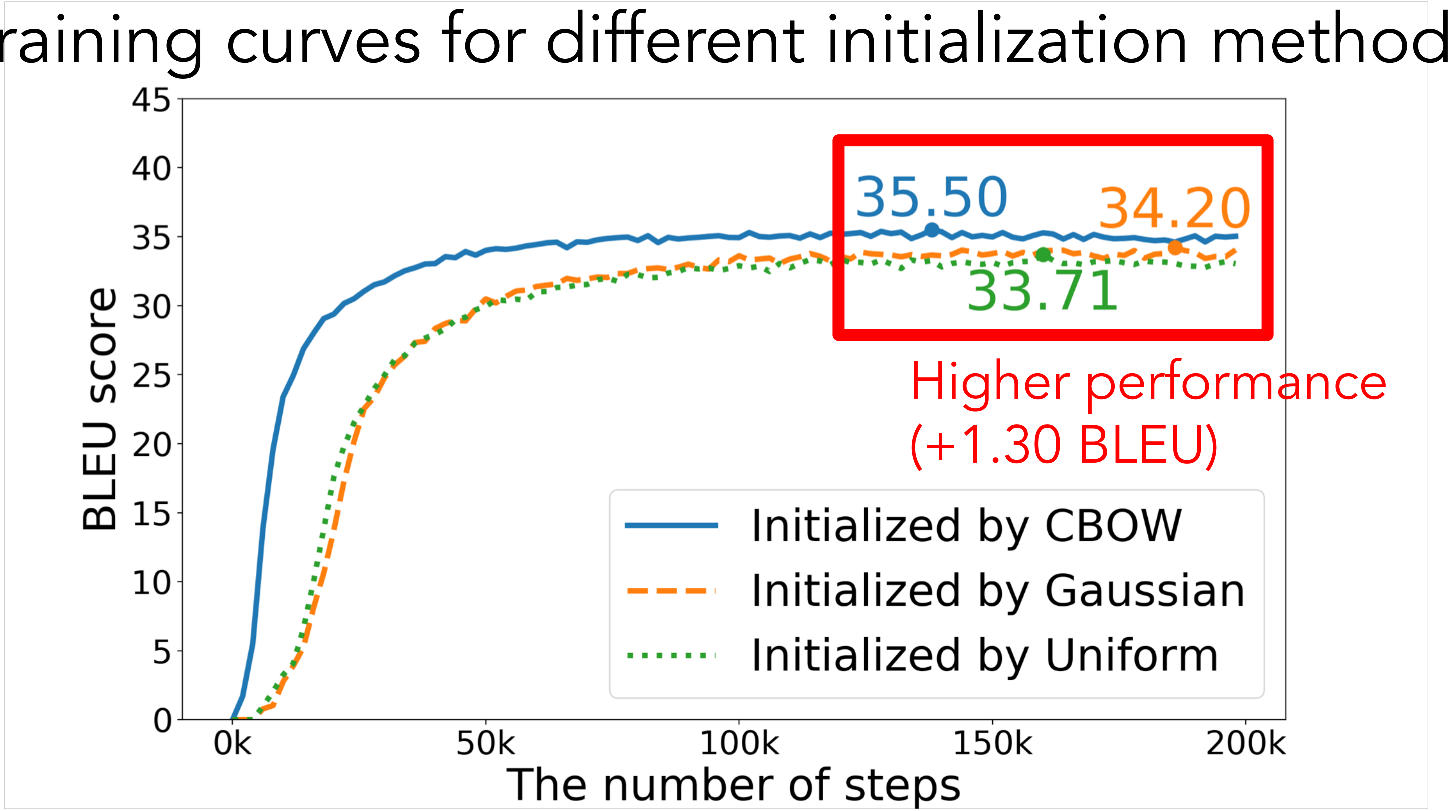


Effect of Initialization Methods: Results

Training curves for different initialization methods



Effect of Initialization Methods: Results



Experiments

1. **Effect of Initialization Methods:**
   Will the proposed method speed up convergence and improve translation quality?

2. **Effect of Large Batch Size:**
   Will large batch sizes (32 to 512) improve translation quality?

Effect of Large Batch Size

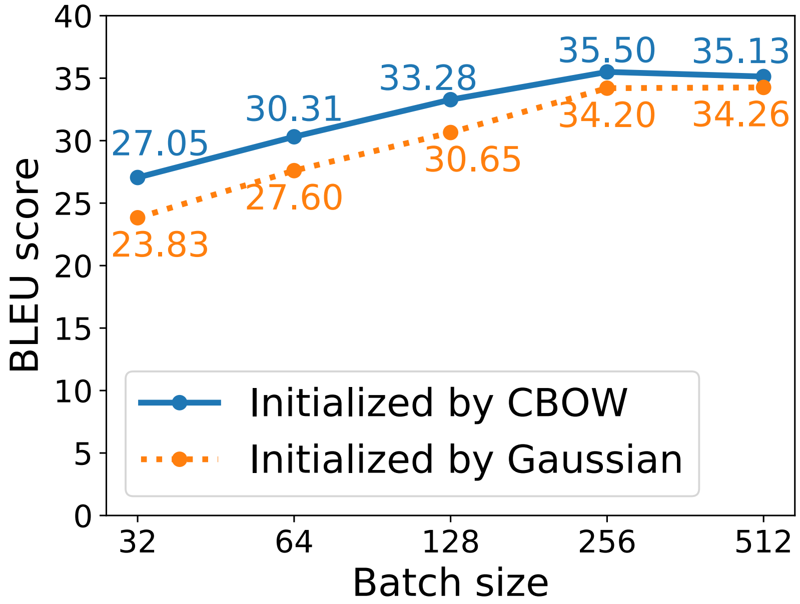
- Purpose:

  Investigate the effect of large batch size

- Compare:
  - Batch sizes: 32, 64,128, 256, 512
  - Initialization methods: CBOW, Gaussian

Effect of Large Batch Size: Results

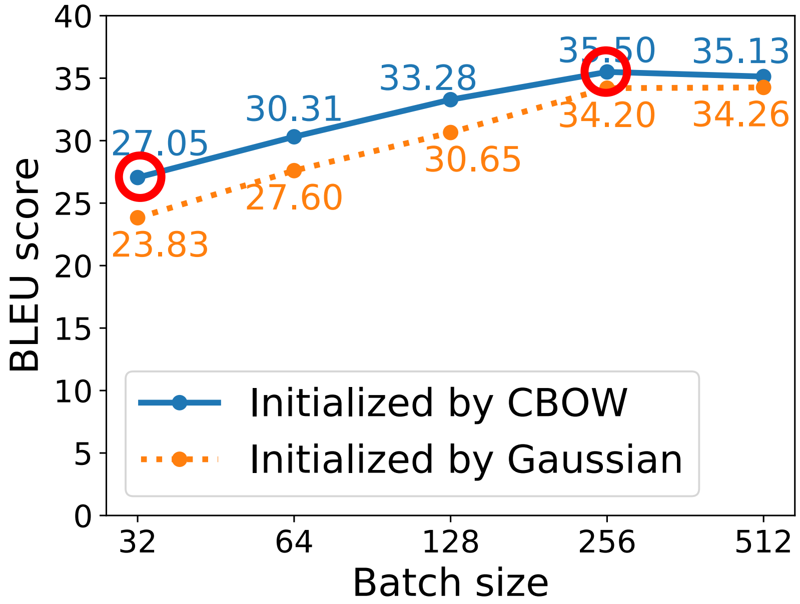
Performance at highest BLEU for each model



Larger batch size leads to higher BLEU score until 256

Effect of Large Batch Size: Results

Performance at highest BLEU for each model

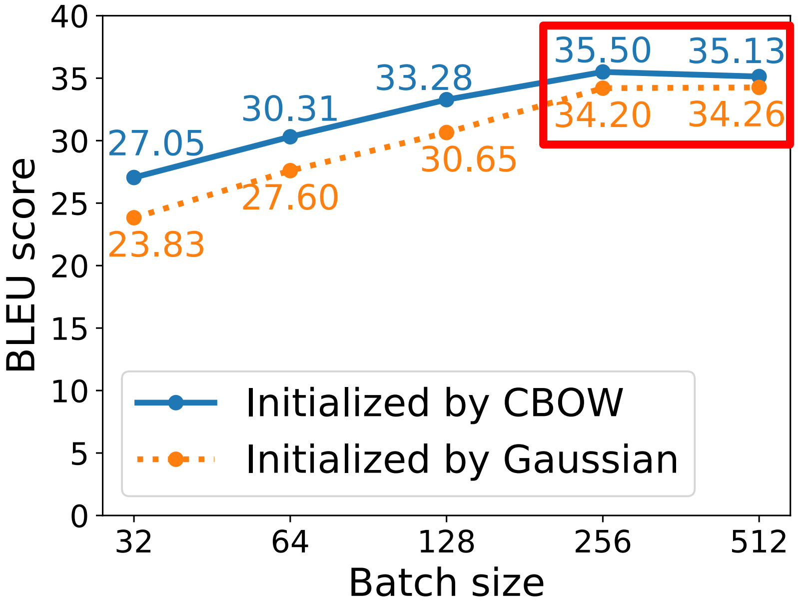


+8.45 BLEU

Larger batch size leads to higher BLEU score until 256

Effect of Large Batch Size: Results

Performance at highest BLEU for each model



Saturates at 256

Larger batch size leads to higher BLEU score until 256

Tradeoff of Large Batch Size

- Pros:
  - Better translation performance
- Cons:
  - Higher memory consumption
    - Titan X/Xp (12GB RAM) not enough for batch size 512
  - Slower convergence
    - Training of 512 batch size takes 7 days
      (c.f. batch size 256: 3 days)

- Rule of thumb: 256 performs well and trains in an acceptable time

BLEU Gains by Two Tricks

Batch Size	Initialization	BLEU Score	Gain
32	Gaussian	23.83	-
32	CBOW	27.05	+4.86
256	Gaussian	34.20	+10.37
256	CBOW	35.50	+11.67

By combining these two tricks, we gained **+11.67** BLEU score

Asian Universities Alliance

04

Model with All Tricks

Prediction Tricks

To further improve translation quality, we implemented these techniques for prediction:

- **Exhaustive Ensemble Search:**
  Search all combinations of models for the best performance when combined
- **Beam Search:**
  Keep multiple hypothesis sentences to get the best prediction on the model ensemble

Summary of Approaches

- Impact of tricks on the BLEU score

Tricks	BLEU (dev)	Gain
Baseline (existing tricks)	23.83	-
+ Embedding Layer Initialization	27.05	+3.22
+ Large Batch Size	35.50	+11.67
+ Exhaustive Ensemble Search	38.00	+14.17
+ Beam Search (width=256)	39.03	+15.20

Tricks have an additive effect on translation quality

Summary of Approaches

- Impact of tricks on the BLEU score

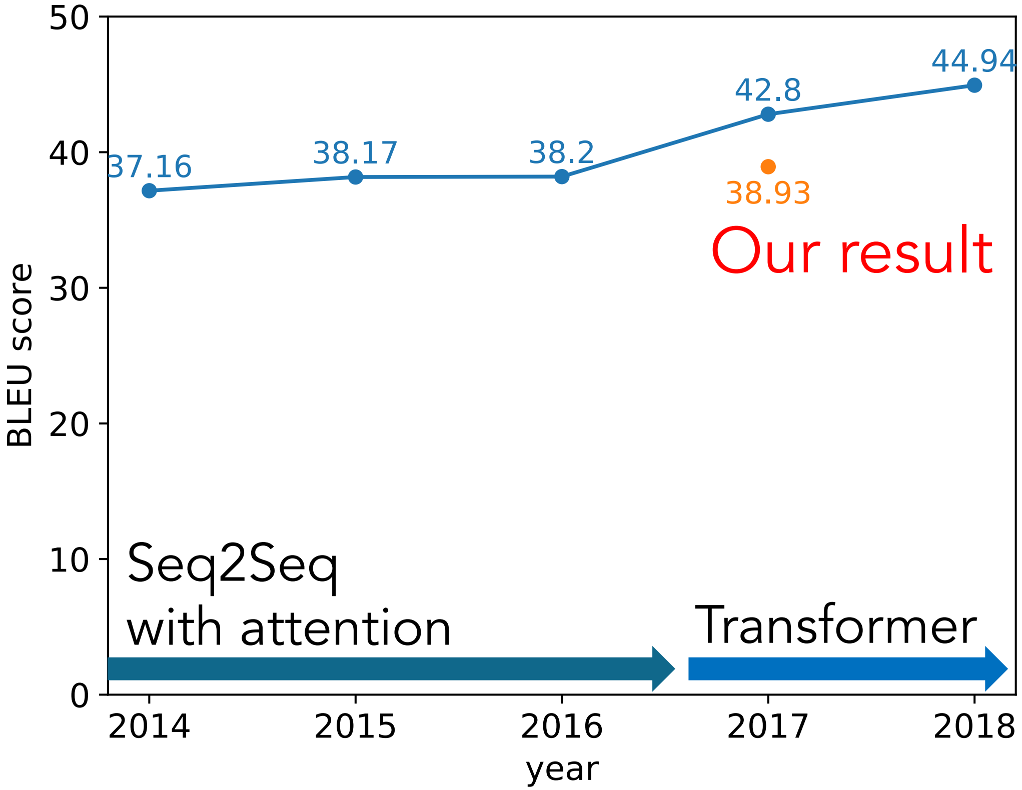
Tricks	BLEU (dev)	Gain
Ba		
+ I		.22
+ I		.67
+ I		.17
+ Beam Search (width=256)	39.03	+15.20

System Performance

BLEU (KyTea) (Test)	38.93
Human Evaluation	68.000

Tricks have an additive effect on translation quality

Transition of best score in WAT



05 Conclusion

Conclusion

- Demonstrated improvements with:
  - Training the model
    - Adam Optimization
    - Sub-word Translation
    - Embedding Layer Initialization
    - Large Batch Size

Novel tricks: leads to a better local optimum

  - Prediction
    - Exhaustive Ensemble Search
    - Beam Search

Improves upon proposed tricks