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

Masato Neishi\textsuperscript{1}, Jin Sakuma\textsuperscript{1}, Satoshi Tohda\textsuperscript{1}, Shonosuke Ishiwatari\textsuperscript{1}, Naoki Yoshinaga\textsuperscript{2}, Masashi Toyoda\textsuperscript{2}

\textsuperscript{1}The University of Tokyo \textsuperscript{2}Institute of Industrial Science (IIS), the University of Tokyo \textsuperscript{*contributed equally

System Overview of Team UT-IIS

**Task:**
ASHPEC English - Japanese

**Approach:**
Seq2seq model with attention

+B + Larger Batch Size

**Proposal:**

- Beam
- Larger Batch Size

**Seq2seq with Attention:**

\[ x_1, x_2, \ldots, x_T \]

\[ y_1, y_2, \ldots, y_T \]

**Evaluation Metric**

<table>
<thead>
<tr>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU (KyTea) 38.93</td>
</tr>
<tr>
<td>Human Evaluation 68.000</td>
</tr>
</tbody>
</table>

**Tricks:**

- **Training Phase**
  - Adam Optimization \([\text{Kingma and Ba, 2015}]
  - Subword Translation (SentencePiece)

- **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:**

- Faster startup

**Method**

<table>
<thead>
<tr>
<th>BLEU</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random (Gaussian) 34.20</td>
<td>-</td>
</tr>
<tr>
<td>CBOW 35.50</td>
<td>+1.30</td>
</tr>
<tr>
<td>Skip-gram 34.44</td>
<td>+0.24</td>
</tr>
<tr>
<td>SL-Skip-gram 34.44</td>
<td>+0.24</td>
</tr>
<tr>
<td>GloVe 34.58</td>
<td>+0.38</td>
</tr>
</tbody>
</table>

**Large Batch Size**

**Background:**

Morishita+ (2017) obtained 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:**

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

**Overall Result**

**Source**

Doping induced a noticeable change at the lower boundary of the three-dimensional ordered vortex phase.

**Tricks**

- 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

**Translation Examples**

**Source**

2015

**Reference**

奪うという形で三次元規則の導入による変化が生じた。

**Source**

線形移動に3次元の構造がとられた。こうした変化を説明できる。

**Source**

The outline of the 23rd white paper which is used to be issued yearly from Ministry of Posts and Telecommunications is specified.

**Reference**

2017年12月20日発表の第24回白書まとめを依頼した。

**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、ユビキタス社会の一端に入り、携帯電話の総保有数は約70%を越えるためである。