Pre-training via Leveraging Assisting Languages for Neural Machine Translation

Haiyue Song¹, Raj Dabre², Zhuoyuan Mao¹, Fei Cheng¹, Sadao Kurohashi¹, Eiichiro Sumita²
¹Kyoto University    ²NICT
Enough parallel data ← Neural Machine Translation

Large Parallel Corpus

Train

L1-L2 NMT model

Parallel data

Seq2Seq with NN (Sutskever+)

NMT by Joint Learning (Bahdanau+)

Time

2014 2015 2016 2018 2019
Lack of large parallel corpora $\rightarrow$ Low performance

Medium

L1-L2 Corpus

Low performance

L1-L2 MT

Parallel data

Seq2Seq with NN (Sutskever+)

NMT by Joint Learning (Bahdanau+)

Time

2014 2015 2016 2018 2019
Lack of large parallel corpus ← Back Translation

- L1, L2 Monolingual Corpus
  - Easier to get

- L1-L2 Corpus

- L1-L2 MT

Parallel data
- Seq2Seq with NN (Sutskever+)
- NMT by Joint Learning (Bahdanau+)

Monolingual data
- Back-Translation (Sennrich+)

Time
- 2014
- 2015
- 2016
- 2018
- 2019

Monolingual data
- Easier to get

Lack of large parallel corpus ← Back Translation
BT: Initial MT $\rightarrow$ Final MT performance

- **L1, L2 Monolingual Corpus**
- **L1-L2 Corpus**
- **L1-L2 MT**

1. **Train**: Initial MT
2. **Translate**: L1-L2 Corpus

**Time**: 2014, 2015, 2016, 2018, 2019

- **Seq2Seq with NN (Sutskever+)**
- **NMT by Joint Learning (Bahdanau+)**
- **Back-Translation (Sennrich+)**
BT: Small parallel corpus $\rightarrow$ Low performance

L1, L2 Monolingual Corpus

Small

L1-L2 Corpus

Initial MT

2. Translate

1. Train

Bottleneck

Seq2Seq with NN (Sutskever+)

NMT by Joint Learning (Bahdanau+)

Back-Translation (Sennrich+)

Time

2014 2015 2016 2018 2019
Low-resource situation ← Pre-train

Pre-trained word embedding (Qi+)

Back-translation (Sennrich+)

NMT by Joint Learning (Bahdanau+)

Seq2Seq with NN (Sutskever+)

Monolingual data

Parallel data

L1-L2 MT

L1-L2 Corpus

L1, L2 Monolingual Corpus

Pre-training

Time

2014 2015 2016 2018 2019

Large

Small
Extreme Low-resource: Lack both parallel and monolingual

Lack of monolingual data?

Parallel data

- Seq2Seq with NN (Sutskever+)
- NMT by Joint Learning (Bahdanau+)
- Back-translation (Sennrich+)

Monolingual data

- Pre-trained word embedding (Qi+)
- MASS (Song+)

Time

- 2014
- 2015
- 2016
- 2018
- 2019
Lack of monolingual data ← Proposed method

Pre-training via Assisting languages

L3 Monolingual Corpus

Small
L1-L2 Corpus

L1-L2 MT

Parallel data
Seq2Seq with NN (Sutskever+)
NMT by Joint Learning (Bahdanau+)
Back-translation (Sennrich+)
Pre-trained word embedding (Qi+)
MASS (Song+)

Monolingual data

Data from assisting languages
Pre-training via Assisting Language (Proposed)

Time
Proposed method: Overview

Goal!

L1-L2
MT model
Proposed method: Overview

Monolingual data

Parallel data

① Pre-train

Pre-trained model

② Fine-tune

L1-L2 MT model

Goal!
Proposed method: Overview

Monolingual data

- L3, L4... Assisting languages
- L1, L2 (Optional)

Parallel data

- L1-L2

1. Pre-train
   Pre-trained model

2. Fine-tune
   L1-L2 MT model

Goal!
Proposed method: Overview

Monolingual data

- L3, L4... Assisting languages
- L1, L2 (Optional)

Mixed data

- Mapping
- Data Selection

Parallel data

- L1-L2

Pre-trained model

① Pre-train

② Fine-tune

L1-L2 MT model

Goal!
Proposed method: Mapping

Goal:

Maximize the cognate sharing
Proposed method: Mapping

Goal:
Maximize the cognate sharing

L3  汉  Map  漢
L1  漢
Proposed method: Mapping

Goal:
Maximize the cognate sharing

Example:
Chinese Hanzi and Japanese Kanji

Chinese Hanzi

Japanese Kanji
Proposed method: Mapping

Goal:
Maximize the cognate sharing

Example:
Chinese Hanzi and Japanese Kanji

Background:
Kanji borrowed from Hanzi
Proposed method: Mapping

Goal:
Maximize the cognate sharing

Example:
Chinese Hanzi and Japanese Kanji

Background:
Kanji borrowed from Hanzi
Over time the written scripts diverged

Chinese Hanzi
Early ages
Simplified (1950s)
Now

Japanese Kanji
漢

漢
Proposed method: Mapping

Goal:
Maximize the cognate sharing

Example:
Chinese Hanzi and Japanese Kanji

Background:
Kanji borrowed from Hanzi
Over time the written scripts diverged

Method:
Map Hanzi to Kanji by a mapping table
(Chu et al., 2012)
Proposed method: Mapping

Goal: Maximize the cognate sharing

Method: Map Hanzi to Kanji
One Hanzi may map to many Kanji
Proposed method: Mapping

Goal:
Maximize the cognate sharing

Method: Map Hanzi to Kanji
One Hanzi may map to many Kanji

Method 1: one-to-one mapping
Proposed method: Mapping

Goal:
Maximize the cognate sharing

Method: Map Hanzi to Kanji
One Hanzi may map to many Kanji

Method 1: one-to-one mapping
Proposed method: Mapping

Goal:
Maximize the cognate sharing

Method:
Map Hanzi to Kanji
One Hanzi may map to many Kanji

Method 1: one-to-one mapping
Method 2: many-to-many mapping

Hanzi    Kanji
机 → [机, 機]
构 → [構, 構]

Chinese word    Japanese word (Synthetic)
机构 → [機構, 机搆, 機構,機搆]
Proposed method: Mapping

Goal:
Maximize the cognate sharing

Method: Map Hanzi to Kanji
One Hanzi may map to many Kanji

Method 1: one-to-one mapping

Method 2: word-to-word mapping
Proposed method: Mapping

Goal:
Maximize the cognate sharing

Method:
Map Hanzi to Kanji
One Hanzi may map to many Kanji

Method 1: one-to-one mapping

Method 2: word-to-word mapping

Hanzi           Kanji
机 → [机, 機]
构 → [構, 構]

Chinese word                Japanese word (Synthetic)
机构 → [機構, 机搆, 機構,机搆]

Japanese LM
Proposed method: Data Selection

Goal:
Reduce difference between train and test
Proposed method: Data Selection

Goal:
Reduce difference between train and test

Method:
Data Selection

Method 1: LM based data selection
Proposed method: Data Selection

Goal:
Reduce difference between train and test

Method:
Data Selection
Method 1: LM based data selection
Method 2: Length based data selection
Proposed method: Data Selection

Goal:
Reduce difference between train and test

Method:
Data Selection

Method 1: LM based data selection

Method 2: Length based data selection
Proposed method: Data Selection

Goal: Reduce difference between train and test

Method: Data Selection

Method 1: LM based data selection

Method 2: Length based data selection
Proposed method: Data Selection

Goal:
Reduce difference between train and test

Method:
Data Selection

Method 1: LM based data selection

Method 2: Length based data selection
Proposed method: Data Selection

Goal:
Reduce difference between train and test

Method:
Data Selection

Method 1: LM based data selection

Method 2: Length based data selection
Pre-train: MASS (Song+, 2019)

Method:
Input: Monolingual sentence with tokens [MASK]ed
Target: [MASK]ed tokens
Experiment settings:

Interested languages:
   Japanese and English
Assisting languages:
   Chinese, French, Arabic and Russian
Experiment settings:

Interested languages:
  Japanese and English
Assisting languages:
  Chinese, French, Arabic and Russian

Dataset:
  Pre-train:
    Ja, En: ASPEC (Nakazawa+, 2016)
    Others: Common Crawl*
  Fine-tune:
    Ja-En: ASPEC (Nakazawa+, 2016)
    No overlap with pre-train data
Data for LM: News commentary*

Data pre-processing:
  Normalization and filtering
  Script mapping for Zh->Ja
  KenLM to train LM

*http://data.statmt.org/ngrams/
*http://data.statmt.org/news-commentary/v14/
Experiment settings:

Interested languages:
Japanese and English

Assisting languages:
Chinese, French, Arabic and Russian

Dataset:
Pre-train:
Ja, En: ASPEC (Nakazawa+, 2016)
Others: Common Crawl*

Fine-tune:
Ja-En: ASPEC (Nakazawa+, 2016)
No overlap with pre-train data

Data for LM: News commentary*

Data pre-processing:
Normalization and filtering
Script mapping for Zh->Ja
KenLM to train LM

Train and evaluate:

- **Tensor2tensor** (Vaswani+, 2018) with ‘transformer_big’ setting
- **Shared vocab** of 64k, using **SentencePiece** (Kuro+, 2018)
- **sacreBLEU**

*http://data.statmt.org/ngrams/
*http://data.statmt.org/news-commentary/v14/
### Results:

<table>
<thead>
<tr>
<th>#</th>
<th>Pre-training</th>
<th>Fine-tuning</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data pre-processing</td>
<td>Zh</td>
<td>Ja</td>
<td>En</td>
<td>Fr</td>
<td>En→Ja 3K</td>
</tr>
<tr>
<td>A1</td>
<td>-</td>
<td>-</td>
<td>2.5</td>
<td>6.0</td>
<td>14.4</td>
<td>22.9</td>
</tr>
<tr>
<td>B1</td>
<td>1-to-1 Zh→Ja mapping + LM</td>
<td>20M</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td><strong>5.3</strong></td>
</tr>
<tr>
<td>B2</td>
<td>LM</td>
<td>-</td>
<td>3.4</td>
<td>9.1</td>
<td>14.9</td>
<td>23.4</td>
</tr>
<tr>
<td>B3</td>
<td>1-to-1 Zh→Ja mapping + LM</td>
<td>20M</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.1</td>
</tr>
</tbody>
</table>

1. **Extreme Low Resource Situation**
   
   Compared with baseline, using **monolingual data from assisting languages helps**. There may be **conflicts between data of different assisting languages**.
## Results:

<table>
<thead>
<tr>
<th>#</th>
<th>Data pre-processing</th>
<th>Pre-training</th>
<th>Fine-tuning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Zh</td>
<td>Ja</td>
</tr>
<tr>
<td>A1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C1</td>
<td>LD</td>
<td>-</td>
<td>1M</td>
</tr>
<tr>
<td>C2</td>
<td>1-to-1 Zh\rightarrow Ja mapping + LD</td>
<td>20M</td>
<td>1M</td>
</tr>
<tr>
<td>C3</td>
<td>LD</td>
<td>-</td>
<td>1M</td>
</tr>
<tr>
<td>C4</td>
<td>1-to-1 Zh\rightarrow Ja mapping + LD</td>
<td>20M</td>
<td>1M</td>
</tr>
</tbody>
</table>

2. **Low Resource Situation**

   Compared with baseline, using monolingual data from assisting languages helps. There may be conflicts between data of different assisting languages.
### Results:

<table>
<thead>
<tr>
<th>#</th>
<th>Data pre-processing</th>
<th>Pre-training</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Flne-tuning</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Zh</td>
<td>Ja</td>
<td>En</td>
<td>Fr</td>
<td>3K</td>
<td>10K</td>
<td>20K</td>
<td>50K</td>
<td>3K</td>
<td>10K</td>
<td>20K</td>
<td>50K</td>
</tr>
<tr>
<td>A1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.5</td>
<td>6.0</td>
<td>14.4</td>
<td>22.9</td>
<td>1.8</td>
<td>4.6</td>
<td>10.9</td>
<td>19.4</td>
</tr>
<tr>
<td>D1</td>
<td>LD</td>
<td>-</td>
<td>15M</td>
<td>15M</td>
<td>-</td>
<td>9.6</td>
<td>17.2</td>
<td>21.5</td>
<td>28.0</td>
<td>8.6</td>
<td>13.5</td>
<td>16.8</td>
<td>20.9</td>
</tr>
<tr>
<td>D2</td>
<td>1-to-1 Zh→Ja mapping + LD</td>
<td>20M</td>
<td>15M</td>
<td>15M</td>
<td>-</td>
<td>9.7</td>
<td>17.1</td>
<td>21.6</td>
<td>27.2</td>
<td>8.3</td>
<td>13.3</td>
<td>16.7</td>
<td>20.6</td>
</tr>
<tr>
<td>D3</td>
<td>LD</td>
<td>-</td>
<td>15M</td>
<td>15M</td>
<td>20M</td>
<td>7.7</td>
<td>15.0</td>
<td>19.8</td>
<td>26.3</td>
<td>6.3</td>
<td>11.7</td>
<td>15.1</td>
<td>20.2</td>
</tr>
<tr>
<td>D4</td>
<td>1-to-1 Zh→Ja mapping + LD</td>
<td>20M</td>
<td>15M</td>
<td>15M</td>
<td>20M</td>
<td>7.7</td>
<td>14.9</td>
<td>19.7</td>
<td>26.1</td>
<td>6.5</td>
<td>11.4</td>
<td>15.4</td>
<td>19.8</td>
</tr>
</tbody>
</table>

#### 3. Rich Resource Situation

Data from assisting languages does not help.
### Results:

<table>
<thead>
<tr>
<th>#</th>
<th>Pre-training</th>
<th>Zh</th>
<th>Ja</th>
<th>En</th>
<th>Fr</th>
<th>Fine-tuning</th>
<th>En→Ja</th>
<th>Ja→En</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data pre-processing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3K</td>
<td>10K</td>
</tr>
<tr>
<td>A1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.5</td>
<td>6.0</td>
<td>14.4</td>
</tr>
<tr>
<td>E1</td>
<td>1-to-1 Zh→Ja mapping</td>
<td>20M</td>
<td>20M</td>
<td>20M</td>
<td>20M</td>
<td>7.0</td>
<td>13.4</td>
<td>19.3</td>
</tr>
<tr>
<td>E2</td>
<td>LM-scoring Zh→Ja mapping</td>
<td>20M</td>
<td>20M</td>
<td>20M</td>
<td>20M</td>
<td>6.3</td>
<td>12.7</td>
<td>18.1</td>
</tr>
</tbody>
</table>

**Mapping:**

1-to-1 Zh→Ja mapping is better than many-to-many mapping

Japanese LM cannot directly apply to Chinese mapped data

Segmentation granularity of Chinese and Japanese data is different
### Results:

<table>
<thead>
<tr>
<th>#</th>
<th>Pre-training</th>
<th>Fine-tuning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data pre-processing</td>
<td>Zh</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3K</td>
</tr>
<tr>
<td>A1</td>
<td>-</td>
<td>2.5</td>
</tr>
<tr>
<td>D1</td>
<td>LD</td>
<td>9.6</td>
</tr>
<tr>
<td>F1</td>
<td>LM-scoring</td>
<td>4.7</td>
</tr>
</tbody>
</table>

### Data Selection:

Sentence length distribution selection is better than LM score method.

Maybe the data used to train the LM is not in-domain.
Results:

<table>
<thead>
<tr>
<th>#</th>
<th>Data pre-processing</th>
<th>Pre-training</th>
<th>Fine-tuning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Zh</td>
<td>Ja</td>
</tr>
<tr>
<td>A1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>F1</td>
<td>LM-scoring</td>
<td>-</td>
<td>20M</td>
</tr>
<tr>
<td>F2</td>
<td>1-to-1 Zh→Ja mapping + LM-scoring</td>
<td>20M</td>
<td>20M</td>
</tr>
<tr>
<td>F3</td>
<td>LM-scoring + Ar20M + Ru20M</td>
<td>-</td>
<td>20M</td>
</tr>
</tbody>
</table>

Different assisting languages:
Similar languages performs better than randomly selected languages
Conclusions:

• Leveraging monolingual data from other languages to improve NMT is possible.

• Script mapping is a good way to improve data similarity thus improve performance.

Future work:

• Explore data selection methods

• Experiments with more challenging language pairs such as Japanese-Russian
Thanks for listening!

Pre-training via Leveraging Assisting Languages for Neural Machine Translation

Haiyue Song\textsuperscript{1}, Raj Dabre\textsuperscript{2}, Zhuoyuan Mao\textsuperscript{1}, Fei Cheng\textsuperscript{1}, Sadao Kurohashi\textsuperscript{1}, Eiichiro Sumita\textsuperscript{2}

\textsuperscript{1}Kyoto University  \textsuperscript{2}NICT