Breaking Down Multilingual Machine Translation

Ting-Rui Chiang¹  Yi-Pei Chen²  Yi-Ting Yeh¹  Graham Neubig¹  Carnegie Mellon University¹, The University of Tokyo²
Background: Multilingual Training for Machine Translation

encoder | decoder  
---|---
Belarusian | Arabic | German | Azerbaijani | English

Many-to-one

↑ has more improvement than ↓

encoder | decoder  
---|---
English | Azerbaijani | Belarusian | Arabic | German

We were wondering why?
Observation

- Azerbaijani
  - Belarusian
    - Arabic
  - German

encoder \rightarrow \text{decoder} \rightarrow \text{English}

- many-to-one

affects the number of modalities exposed to the encoder/decoder

- English
  - encoder \rightarrow \text{decoder}

- one-to-many

- Azerbaijani
  - Belarusian
    - Arabic
  - German
Investigation

- How does multilingual training affect the encoder/decoder?
  - i.e. How useful are the parameters learned from multilingual training?
Experiment - Step 1: Train a Multilingual Model

encoder

decoder

parameters

Azerbaijani
Belarusian
Arabic
German

train

Azerbaijani
Belarusian
Arabic
German
Experiment - Step 2: Initialize Several Bilingual Models

- Encoder
- Decoder

parameters

English → Encoder → Decoder → Arabic

English → Encoder → Decoder → Arabic

English → Encoder → Decoder → Arabic

English → Encoder → Decoder → Arabic
Experiment - Step 2: Initialize Several Bilingual Models

Load both encoder and decoder parameters.
Experiment - Step 2: Initialize Several Bilingual Models

Load both

Load encoder
Experiment - Step 2: Initialize Several Bilingual Models

Load both parameters

Load encoder parameters

Load decoder parameters
Experiment - Step 2: Initialize Several Bilingual Models

- **Load both**
  - English
  - Encoder
  - Decoder
  - Arabic

- **Load encoder**
  - English
  - Encoder
  - Decoder
  - Arabic

- **Load decoder**
  - English
  - Encoder
  - Decoder
  - Arabic

- **From scratch**
  - English
  - Encoder
  - Decoder
  - Arabic
Experiment - Step 3: Train with Bilingual Data

- Load both
- Load encoder
- Load decoder
- From scratch
Experiment - Final Step: Compare their performance

We can infer how multilingual training benefits the encoder/decoder.
Low-resource: Multilingual training benefits both the encoder and the decoder.
High-resource: Multilingual training only benefits encoder.
Investigating Parameter Sharing

1. Identify important attention heads for languages.
2. Compute the coherence of important heads.
Investigating Parameter Sharing
### Improvement by Training with Related Languages

<table>
<thead>
<tr>
<th>Model</th>
<th>az</th>
<th>be</th>
<th>gl</th>
<th>sk</th>
<th>ar</th>
<th>de</th>
<th>he</th>
<th>it</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-All (Aharoni et al., 2019)</td>
<td>5.1</td>
<td>10.7</td>
<td>26.6</td>
<td>24.5</td>
<td>16.7</td>
<td>30.5</td>
<td>27.6</td>
<td>35.9</td>
</tr>
<tr>
<td>Bilingual Baseline</td>
<td>1.3</td>
<td>1.9</td>
<td>3.9</td>
<td>13.1</td>
<td>15.6</td>
<td>27.1</td>
<td>25.4</td>
<td>32.0</td>
</tr>
<tr>
<td>All-All</td>
<td>3.1</td>
<td>6.2</td>
<td>20.5</td>
<td>18.4</td>
<td>12.7</td>
<td>24.5</td>
<td>21.1</td>
<td>30.5</td>
</tr>
<tr>
<td>All-All w/ f.t. on related clusters</td>
<td>7.9</td>
<td>12.8</td>
<td>27.5</td>
<td>24.9</td>
<td>-</td>
<td>30.2</td>
<td>27.0</td>
<td>35.4</td>
</tr>
<tr>
<td>All-All w/ f.t. on random groups</td>
<td>6.9</td>
<td>13.3</td>
<td>22.5</td>
<td>24.3</td>
<td>-</td>
<td>-</td>
<td>27.5</td>
<td>35.2</td>
</tr>
<tr>
<td>En-All</td>
<td>4.9</td>
<td>9.00</td>
<td>24.2</td>
<td>21.9</td>
<td>15.1</td>
<td>27.9</td>
<td>24.1</td>
<td>33.3</td>
</tr>
<tr>
<td>En-All w/ f.t. on related clusters</td>
<td>7.9</td>
<td>13.9</td>
<td>21.0</td>
<td><strong>26.2</strong></td>
<td>16.7</td>
<td>30.4</td>
<td>27.1</td>
<td>35.4</td>
</tr>
<tr>
<td>En-All w/ f.t. on random groups</td>
<td>7.0</td>
<td>13.1</td>
<td>23.1</td>
<td>24.7</td>
<td>-</td>
<td>-</td>
<td>27.6</td>
<td>35.2</td>
</tr>
<tr>
<td>Load En-All w/ f.t. on closest</td>
<td>7.8</td>
<td><strong>15.2</strong></td>
<td><strong>28.6</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

We found that multilingual training is more useful for the encoder.

We proposed a purely data-driven way to identify related languages.

Our experiments can serve as analysis tools for future research.