CS11-737 Multilingual NLP

Multilingual Neural Machine Translation Model Architecture

Lei Li
https://lileicc.github.io/course/11737mnlp23fa/
Language Data

Monolingual data
Parallel data

Google-Translate
NLLB

N sentences vs Language index

[Credit: Isaac Caswell, 2022]
Training Multilingual MT Jointly

1 year

Bilingual MT

En → Es
En → Pt
En → Zh

Multilingual Training

En → Es
En → Pt
En → Zh
Many-to-Many Multilingual NMT

**Training**

- En
- De
- Es
- Ja
- Zh

Monolingual

**Testing**

- En
- De
- Es
- Ja
- Zh

no fine-tuning

supervised

zero-shot

unsupervised
Why Multilingual NMT?

• Develop one model to translate between all language pairs.

• Model-side: Languages with rich resource could benefit those with low resource
  ○ Similar languages share tokens

• Serving-side: only one model deployment versus of many deployments. Simpler workload and job management and scheduling.
  ○ Many languages would have much few requests but still need to occupy the servers.
• **Many-to-one:**
  ○ Many source language to a target language
  ○ Usually the target is English

• **One-to-Many:**
  ○ One source language to many target languages
  ○ Usually the source is English

• **Many-to-many:**
  ○ Many source language to many target languages
  ○ Should include non-English pairs (often low-resource or zero-resource setting), very challenging!

• **Which is simpler?**
MNMT Fine-tuning Testing

- **Exotic (Unseen) pair**
  - Both the testing source language and target language appeared in the training, but the source-target pair never appeared in the training
  - Also known as zero-shot MNMT

- **Exotic (Unseen) source**
  - Testing source language never occur in the training

- **Exotic (Unseen) target**
  - Testing target language never occur in the training

- **Exotic (Unseen) full**
  - Neither the source language nor the target language for testing occur in the training
  - Is it even possible? Yes, for the pre-train fine-tuning paradigm.
MNMT with Language Tags
A single model for Multilingual NMT

- Language-specific encoding (@en@car, @de@automobile)
- But hard to learn a joint embedding.
- Challenge:
  - large vocabulary (twice many)
  - how does the model know it is to translate into German or French?

Encoder
- I like singing and dancing

Decoder
- BOS J'adore chanter et danser

• One model can translate between many languages.
• Language Tag is used to indicate the source and target language.
• Vocabulary is built jointly

Vocabulary

• Single joint vocabulary [Johnson 2017]
  ○ combine all corpus together, and apply BPE
• Soft-decoupled encoding [Wang et al 2019]
• Even better: learned vocabulary [Xu 2021], (later in class)
Google’s MNMT

• Training 12 language pairs together

• LSTM-s2s:
  ○ 8 layer LSTM encoder, 1st layer bidirectional
  ○ 8 layer LSTM decoder with attention

Table 4: Large-scale experiments: BLEU scores for single language pair and multilingual models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Single</th>
<th>Multi</th>
<th>Multi</th>
<th>Multi</th>
<th>Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#nodes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>1024</td>
<td>1280</td>
<td>1536</td>
<td>1792</td>
</tr>
<tr>
<td></td>
<td>3B</td>
<td>255M</td>
<td>367M</td>
<td>499M</td>
<td>650M</td>
</tr>
<tr>
<td>En→Ja</td>
<td>23.66</td>
<td>21.10</td>
<td>21.17</td>
<td>21.72</td>
<td>21.70</td>
</tr>
<tr>
<td>En→Ko</td>
<td>19.75</td>
<td>18.41</td>
<td>18.36</td>
<td>18.30</td>
<td>18.28</td>
</tr>
<tr>
<td>Ja→En</td>
<td>23.41</td>
<td>21.62</td>
<td>22.03</td>
<td>22.51</td>
<td>23.18</td>
</tr>
<tr>
<td>Ko→En</td>
<td>25.42</td>
<td>22.87</td>
<td>23.46</td>
<td>24.00</td>
<td>24.67</td>
</tr>
<tr>
<td>En→Es</td>
<td>34.50</td>
<td>34.25</td>
<td>34.40</td>
<td>34.77</td>
<td>34.70</td>
</tr>
<tr>
<td>En→Pt</td>
<td>38.40</td>
<td>37.35</td>
<td>37.42</td>
<td>37.80</td>
<td>37.92</td>
</tr>
<tr>
<td>Es→En</td>
<td>38.00</td>
<td>36.04</td>
<td>36.50</td>
<td>37.26</td>
<td>37.45</td>
</tr>
<tr>
<td>Pt→En</td>
<td>44.40</td>
<td>42.53</td>
<td>42.82</td>
<td>43.64</td>
<td>43.87</td>
</tr>
<tr>
<td>En→De</td>
<td>26.43</td>
<td>23.15</td>
<td>23.77</td>
<td>23.63</td>
<td>24.01</td>
</tr>
<tr>
<td>En→Fr</td>
<td>35.37</td>
<td>34.00</td>
<td>34.19</td>
<td>34.91</td>
<td>34.81</td>
</tr>
<tr>
<td>De→En</td>
<td>31.77</td>
<td>31.17</td>
<td>31.65</td>
<td>32.24</td>
<td>32.32</td>
</tr>
<tr>
<td>Fr→En</td>
<td>36.47</td>
<td>34.40</td>
<td>34.56</td>
<td>35.35</td>
<td>35.52</td>
</tr>
<tr>
<td>ave diff</td>
<td>-</td>
<td>-1.72</td>
<td>-1.43</td>
<td>-0.95</td>
<td>-0.76</td>
</tr>
<tr>
<td>vs single</td>
<td>-</td>
<td>-5.6%</td>
<td>-4.7%</td>
<td>-3.1%</td>
<td>-2.5%</td>
</tr>
</tbody>
</table>

Google’s MNMT Zero-shot

- Bilingual pivot
- Multilingual joint
- What is missing in the table?
  - Multilingual pivot
    - zero-shot
  - no longer zero-shot, since additional Pt-Es pairs are used.

Table 5: Portuguese→Spanish BLEU scores using various models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Zero-shot</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) PBMT bridged</td>
<td>no</td>
<td>28.99</td>
</tr>
<tr>
<td>(b) NMT bridged</td>
<td>no</td>
<td>30.91</td>
</tr>
<tr>
<td>(c) NMT Pt→Es</td>
<td>no</td>
<td>31.50</td>
</tr>
<tr>
<td>(d) Model 1 (Pt→En, En→Es)</td>
<td>yes</td>
<td>21.62</td>
</tr>
<tr>
<td>(e) Model 2 (En↔{Es, Pt})</td>
<td>yes</td>
<td>24.75</td>
</tr>
<tr>
<td>(f) Model 2 + incremental training</td>
<td>no</td>
<td>31.77</td>
</tr>
</tbody>
</table>

Google’s MNMT Zero-shot

- MNMT is worse than pivot on zero-shot directions

Table 6: Spanish→Japanese BLEU scores for explicit and implicit bridging using the 12-language pair large-scale model from Table 4.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT Es→Ja explicitly bridged</td>
<td>18.00</td>
</tr>
<tr>
<td>NMT Es→Ja implicitly bridged</td>
<td>9.14</td>
</tr>
</tbody>
</table>
We investigate the following four language tag strategies:

- **Original**: Hello World! → ¡Hola Mundo
- **T-ENC**: ___es___ Hello World! → ¡Hola Mundo
- **T-DEC**: Hello World! → ___es___ ¡Hola Mundo
- **S-ENC-T-ENC**: ___en___ ___es___ Hello World! → ¡Hola Mundo
- **S-ENC-T-DEC**: ___en___ Hello World! → ___es___ ¡Hola Mundo

Supervised directions: The directions which has been seen together in the training time.

Target Language Tag on Encoder Strategy Gets Best Zero-Shot Performance

Zero-shot directions: The directions between known languages that the model has never seen together at training time.

Mixed Source Language can still be Translated

• {Ja, Ko} -> En

• Japanese: 私は東京大学の学生です。 → I am a student at Tokyo University.

• Korean: 나는 서울대학교 학생입니다. → I am a student at Tokyo University.

• Japanese/Korean: 私は東京大学에서 학생입니다. → I am a student of Tokyo University.
## Mixed Decoder for Target Language

- **En -> {Ja, Ko}**
- Either generate Japanese or Korean

### Table 8: Gradually mixing target languages Ja/Ko.

<table>
<thead>
<tr>
<th>(w_{ko})</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>I must be getting somewhere near the centre of the earth.</td>
</tr>
<tr>
<td>0.40</td>
<td>私は地球の中心近くにどこかを行っているに違いない。</td>
</tr>
<tr>
<td>0.56</td>
<td>私は地球の中心近くのどこかに着いているに違いない。</td>
</tr>
<tr>
<td>0.58</td>
<td>私は地球の中心近くのどこかになっているに違いない。</td>
</tr>
<tr>
<td>0.60</td>
<td>私は地球の中心近くのどこかになっているに違いない。</td>
</tr>
<tr>
<td>0.70</td>
<td>私は地球の中心近くのどこかになっているに違いない。</td>
</tr>
<tr>
<td>0.90</td>
<td>私は地球の中心近くのどこかになっているに違いない。</td>
</tr>
<tr>
<td>1.00</td>
<td>私は地球の中心近くのどこかになっているに違いない。</td>
</tr>
</tbody>
</table>
### Multilingual NMT with mTransformer

- **Model:** Transformer-base (6e6d, 512) $\Rightarrow$ mTransformer
- **Data:** TED-talk, 59 languages, 116 directions

<table>
<thead>
<tr>
<th></th>
<th>Az-En</th>
<th>Be-En</th>
<th>Gl-En</th>
<th>Sk-En</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong># of examples</strong></td>
<td>5.9k</td>
<td>4.5k</td>
<td>10k</td>
<td>61k</td>
<td>20.3k</td>
</tr>
<tr>
<td>Neubig &amp; Hu 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baselines</td>
<td>2.7</td>
<td>2.8</td>
<td>16.2</td>
<td>24</td>
<td>11.42</td>
</tr>
<tr>
<td>many-to-one</td>
<td>11.7</td>
<td>18.3</td>
<td>29.1</td>
<td>28.3</td>
<td>21.85</td>
</tr>
<tr>
<td>Wang et al. 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baselines</td>
<td>11.82</td>
<td>18.71</td>
<td>30.3</td>
<td>28.77</td>
<td>22.4</td>
</tr>
<tr>
<td>many-to-one</td>
<td>11.24</td>
<td>18.28</td>
<td>28.63</td>
<td>26.78</td>
<td>21.23</td>
</tr>
<tr>
<td>many-to-many</td>
<td>12.78</td>
<td>21.73</td>
<td>30.65</td>
<td>29.54</td>
<td>23.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Ar-En</th>
<th>De-En</th>
<th>He-En</th>
<th>It-En</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong># of examples</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baselines</td>
<td>27.84</td>
<td>30.5</td>
<td>34.37</td>
<td>33.64</td>
<td>31.59</td>
</tr>
<tr>
<td>many-to-one</td>
<td>25.93</td>
<td>28.87</td>
<td>30.19</td>
<td>32.42</td>
<td>29.35</td>
</tr>
<tr>
<td>many-to-many</td>
<td><strong>28.32</strong></td>
<td><strong>32.97</strong></td>
<td>33.18</td>
<td><strong>35.14</strong></td>
<td><strong>32.4</strong></td>
</tr>
</tbody>
</table>

Aharoni et al. Massively Multilingual Neural Machine Translation. 2019
Limitation of mTransformer: does not work for Many-to-Many En-X

<table>
<thead>
<tr>
<th></th>
<th>En-Az</th>
<th>En-Be</th>
<th>En-Gl</th>
<th>En-Sk</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td># of examples</td>
<td>5.9k</td>
<td>4.5k</td>
<td>10k</td>
<td>61k</td>
<td>20.3k</td>
</tr>
<tr>
<td>baselines</td>
<td>2.16</td>
<td>2.47</td>
<td>3.26</td>
<td>5.8</td>
<td>3.42</td>
</tr>
<tr>
<td>one-to-many</td>
<td>5.06</td>
<td>10.72</td>
<td>26.59</td>
<td>24.52</td>
<td>16.72</td>
</tr>
<tr>
<td>many-to-many</td>
<td>3.9</td>
<td>7.24</td>
<td>23.78</td>
<td>21.83</td>
<td>14.19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>En-Ar</th>
<th>En-De</th>
<th>En-He</th>
<th>En-It</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td># of examples</td>
<td>213k</td>
<td>167k</td>
<td>211k</td>
<td>203k</td>
<td>198.5k</td>
</tr>
<tr>
<td>baselines</td>
<td>12.95</td>
<td>23.31</td>
<td>23.66</td>
<td>30.33</td>
<td>22.56</td>
</tr>
<tr>
<td>one-to-many</td>
<td>16.67</td>
<td>30.54</td>
<td>27.62</td>
<td>35.89</td>
<td>27.68</td>
</tr>
<tr>
<td>many-to-many</td>
<td>14.25</td>
<td>27.95</td>
<td>24.16</td>
<td>33.26</td>
<td>24.9</td>
</tr>
</tbody>
</table>

Table 3: En→X test BLEU on the TED Talks corpus
• mTransformer
  ○ 6e6d, 1024 -> 8192
  ○ 473m parameters
• 103 Languages (inc. En)
  ○ 64k vocab

<table>
<thead>
<tr>
<th></th>
<th>Ar</th>
<th>Az</th>
<th>Be</th>
<th>De</th>
<th>He</th>
<th>It</th>
<th>Nl</th>
<th>Ro</th>
<th>Sk</th>
<th>Tr</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>baselines</td>
<td>23.34</td>
<td>16.3</td>
<td>21.93</td>
<td>30.18</td>
<td>31.83</td>
<td>36.47</td>
<td>36.12</td>
<td>34.59</td>
<td>25.39</td>
<td>27.13</td>
<td>28.33</td>
</tr>
<tr>
<td>many-to-one</td>
<td>26.04</td>
<td>23.68</td>
<td>25.36</td>
<td>35.05</td>
<td>33.61</td>
<td>36.28</td>
<td>36.33</td>
<td>28.35</td>
<td>29.75</td>
<td>31.01</td>
<td></td>
</tr>
<tr>
<td>many-to-many</td>
<td>22.17</td>
<td>21.45</td>
<td>23.03</td>
<td>37.06</td>
<td>30.71</td>
<td>35.0</td>
<td>36.18</td>
<td>36.57</td>
<td>29.87</td>
<td>27.64</td>
<td>29.97</td>
</tr>
</tbody>
</table>

Table 5: X→En test BLEU on the 103-language corpus

<table>
<thead>
<tr>
<th></th>
<th>Ar</th>
<th>Az</th>
<th>Be</th>
<th>De</th>
<th>He</th>
<th>It</th>
<th>Nl</th>
<th>Ro</th>
<th>Sk</th>
<th>Tr</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>baselines</td>
<td>10.57</td>
<td>8.07</td>
<td>15.3</td>
<td>23.24</td>
<td>19.47</td>
<td>31.42</td>
<td>28.68</td>
<td>27.92</td>
<td>11.08</td>
<td>15.54</td>
<td>19.13</td>
</tr>
<tr>
<td>one-to-many</td>
<td>12.08</td>
<td>9.92</td>
<td>15.6</td>
<td>31.39</td>
<td>20.01</td>
<td>33</td>
<td>31.06</td>
<td>28.43</td>
<td>17.67</td>
<td>17.68</td>
<td>21.68</td>
</tr>
<tr>
<td>many-to-many</td>
<td>10.57</td>
<td>9.84</td>
<td>14.3</td>
<td>28.48</td>
<td>17.91</td>
<td>30.39</td>
<td>29.67</td>
<td>26.23</td>
<td>18.15</td>
<td>15.58</td>
<td>20.11</td>
</tr>
</tbody>
</table>

Table 6: En→X test BLEU on the 103-language corpus

Aharoni et al. Massively Multilingual Neural Machine Translation. 2019
More language trained together, but

<table>
<thead>
<tr>
<th></th>
<th>Ar-En</th>
<th>En-Ar</th>
<th>Fr-En</th>
<th>En-Fr</th>
<th>Ru-En</th>
<th>En-Ru</th>
<th>Uk-En</th>
<th>En-Uk</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>25-to-25</td>
<td>23.43</td>
<td>11.77</td>
<td>38.87</td>
<td>36.79</td>
<td>29.36</td>
<td>23.24</td>
<td>25.81</td>
<td>17.17</td>
<td>25.8</td>
</tr>
<tr>
<td>50-to-50</td>
<td>23.7</td>
<td>11.65</td>
<td>37.81</td>
<td>35.83</td>
<td>29.22</td>
<td>21.95</td>
<td>26.02</td>
<td>15.32</td>
<td>25.18</td>
</tr>
<tr>
<td>75-to-75</td>
<td>22.23</td>
<td>10.69</td>
<td>37.97</td>
<td>34.35</td>
<td>28.55</td>
<td>20.7</td>
<td>25.89</td>
<td>14.59</td>
<td>24.37</td>
</tr>
<tr>
<td>103-to-103</td>
<td>21.16</td>
<td>10.25</td>
<td>35.91</td>
<td>34.42</td>
<td>27.25</td>
<td>19.9</td>
<td>24.53</td>
<td>13.89</td>
<td>23.41</td>
</tr>
</tbody>
</table>
### mTransformer Zero-shot Performance

<table>
<thead>
<tr>
<th></th>
<th>Ar-Fr</th>
<th>Fr-Ar</th>
<th>Ru-Uk</th>
<th>Uk-Ru</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-to-5</td>
<td>1.66</td>
<td>4.49</td>
<td>3.7</td>
<td>3.02</td>
<td>3.21</td>
</tr>
<tr>
<td>25-to-25</td>
<td>1.83</td>
<td>5.52</td>
<td><strong>16.67</strong></td>
<td>4.31</td>
<td>7.08</td>
</tr>
<tr>
<td>50-to-50</td>
<td><strong>4.34</strong></td>
<td>4.72</td>
<td>15.14</td>
<td><strong>20.23</strong></td>
<td><strong>11.1</strong></td>
</tr>
<tr>
<td>75-to-75</td>
<td>1.85</td>
<td>4.26</td>
<td>11.2</td>
<td>15.88</td>
<td>8.3</td>
</tr>
<tr>
<td>103-to-103</td>
<td>2.87</td>
<td>3.05</td>
<td>12.3</td>
<td>18.49</td>
<td>9.17</td>
</tr>
</tbody>
</table>

Table 8: Zero-Shot performance while varying the number of languages involved
Bigger Data

• Data: 25 billion sentence pairs in 103 languages
• Model: mTransformer with 375 million params (larger than Transformer-big)

<table>
<thead>
<tr>
<th></th>
<th>High 25</th>
<th>Med. 52</th>
<th>Low 25</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>En→Any</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilingual</td>
<td>29.34</td>
<td>17.50</td>
<td>11.72</td>
</tr>
<tr>
<td>All→All</td>
<td>28.03</td>
<td>16.91</td>
<td>12.75</td>
</tr>
<tr>
<td><strong>En→Any</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any→En</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilingual</td>
<td>37.61</td>
<td>31.41</td>
<td>21.63</td>
</tr>
<tr>
<td>All→All</td>
<td>33.85</td>
<td>30.25</td>
<td>26.96</td>
</tr>
<tr>
<td>Any→En</td>
<td>36.61</td>
<td>33.66</td>
<td>30.56</td>
</tr>
</tbody>
</table>
Sampling of Data

- sample data prob w.r.t \( p^{\frac{1}{T}} \)

<table>
<thead>
<tr>
<th></th>
<th>High 25</th>
<th>Med. 52</th>
<th>Low 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_V = 1 )</td>
<td>27.81</td>
<td>16.72</td>
<td>12.73</td>
</tr>
<tr>
<td>( T_V = 100 )</td>
<td>27.83</td>
<td>16.86</td>
<td>12.78</td>
</tr>
<tr>
<td>( T_V = 5 )</td>
<td>28.03</td>
<td>16.91</td>
<td>12.75</td>
</tr>
<tr>
<td>( T_V = 1 )</td>
<td>33.82</td>
<td>29.78</td>
<td>26.27</td>
</tr>
<tr>
<td>( T_V = 100 )</td>
<td>33.70</td>
<td>30.15</td>
<td>26.91</td>
</tr>
<tr>
<td>( T_V = 5 )</td>
<td>33.85</td>
<td>30.25</td>
<td>26.96</td>
</tr>
</tbody>
</table>

Data distribution over language pairs

- High Resource → Low Resource
- \{French, German, Spanish, ...\} → \{Yoruba, Sindhi, Hawaiian, ...\}
Bigger Model

- mTransformer:
  - 400m, 1.3B wide (12e12d), 1.3B deep (24e24d)
  - Deep is better than wide!
Limitation

- mTransformer boosts performance on low-resource languages but not high-resource languages.
- Zero-shot directions are not usable yet.

Arivazhagan et al. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. 2019
MT w/ Adapter
Parameter Interference issue for MNMT

- Insufficient model capacity
  - the sharing model capacity has to be split for different translation directions
Multilingual NMT with Serial Adapter

- For each layer, adding language-specific module
  \[ z^\sim = \text{LNT}(z_i) \]
  \[ h = \text{relu}(W z^\sim) \]
  \[ x = Wh + z \]
- Could be used for both domain adaptation and MNMT
- Joint training the whole architecture
Serial Adapter improves Multilingual Translation

• on rich-resource lang.
• But serial-adapter is not plug-and-play
  ○ Joint training mTransformer+SA will be better than training mTransformer, fix, and train adapter.
  ○ Adapter has tight integration with the main architecture.
Counter Interference

- Which adapter will remove noise?

Zhu et al. Counter-Interference Adapter for Multilingual Machine Translation. 2021
Parallel Adapter - CIAT

- Design rationale:
  - process before multilingual interference is introduced in each layer
- Embedding adapter
- Parallel layer adapter
- Training:
  - Pretrain mTransformer on multilingual data
  - Fix mTransformer and train parallel adapters on specific language pairs

Zhu et al. Counter-Interference Adapter for Multilingual Machine Translation. 2021
Comparing MNMT w/ Adapters

- mTransformer could be worse than bilingual Transformer
- Both serial adapter and parallel adapter (CIAT) improves mTransformer
- Parallel even beat bilingual Transformer! Serial adapter does not.

Zhu et al. Counter-Interference Adapter for Multilingual Machine Translation. 2021
Which layer-adapter are more important?

- Upper decoder layer adapter is more important
Embedding Adapter is also important!

- Embedding adapter enhance the word embedding similarity between language pairs

![Graph showing average cosine similarity with and without embedding adapter](image-url)
Benefit of MNMT w/ Adapter

- Improve the performance on MNMT, even beat Bilingual NMT
  - Reducing interference among large languages
  - Boost performance on zero-shot setting
- With a fraction of overhead
  - Bilingual Transformer-big: N x 242m
  - mTransformer: 242m
  - mTransformer+Serial Adapter: 242m + N x 12.6m
  - mTransformer+parallel adapter (CIAT): 242m + N x 12.6~27.3m
- Plug-and-play: CIAT only needs to finetune adapter
Exploiting Model Capacity with Language-specific Subnet
Challenge of Multilingual NMT

- Challenge: Performance degradation for rich-resource
  - caused by Parameter Interference
Language-Specific Sub-network (LaSS)

- Each direction has
  - shared parameters with other directions
  - preserves its language-specific parameters

LaSS overall framework

- For each language pair $s_i \rightarrow t_i$, a sub-network is selected from base model $\theta_0$ indicated by a binary mask $M_{s_i \rightarrow t_i} \in \{0, 1\}^{|	heta|}$
How to find language-specific sub-network: Intuition

- Fine-tuning and pruning
  - Fine-tuning on $s_i \rightarrow t_i$ **amplifies** important weights and **diminishes** the unimportant weights.

How to find language-specific masks

- Start with a vanilla multilingual model $\theta_0$ jointly trained on

$$\left\{ \mathcal{D}_{s_i \rightarrow t_i} \right\}_{i=1}^{N}$$

- For each language pair $s_i \rightarrow t_i$, fine-tuning $\theta_0$ on $\mathcal{D}_{s_i \rightarrow t_i}$, respectively

- Rank the weights in fine-tuned model and prune the lowest $\alpha$ percent to obtain $\mathbf{M}_{s_i \rightarrow t_i}$

Further continue to train $\theta_0$ through structure-aware updating after obtaining $\mathbf{M}_{s_i \rightarrow t_i}$

- Create batch $\mathcal{B}_{s_i \rightarrow t_i}$ full of samples from $s_i \rightarrow t_i$
- Forward and backward with sub-network

$$\theta_{s_i \rightarrow t_i} = \left\{ \theta^j_0 \mid \mathbf{M}^j_{s_i \rightarrow t_i} = 1 \right\}$$
LaSS obtains consistent gains for both Transformer-base and Transformer-big.

LaSS obtains consistent performance gains.

- IWSLT
LaSS obtains large gains in zero-shot translation

- An average of 8.3 BLEU gains on 30 language pairs
- 26.5 BLEU gains for Fr→Zh

Fr → X Results

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Baseline</th>
<th>LaSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fr→Cs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fr→De</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fr→Es</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fr→Ru</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fr→Zh</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Benefits of Language-specific Subnet

- The same number of parameters, no extra parameter
- Improved performance on both rich-resource and zero-shot translation directions.
What do we need for larger scale?
Full Many-to-Many MNMT

- Previous many-to-many MNMT does not work well on non-English pairs.

100 Language Benchmark

- WMT — 13 languages
- WAT — Burmese-English
- IWSLT — 4 languages
- FLORES — Sinhala and Nepali <--> English
- TED — The TED Talks data set (Ye et al., 2018) contains translations between more than 50 languages; most of the pairs do not include English. The evaluation data is n-way parallel and contains thousands of directions.
- Tatoeba — 692 test pairs from mixed domains where sentences are contributed and translated by volunteers online. The evaluation pairs we use from Tatoeba cover 85 different languages.
Data mining for parallel corpus

• CCAigned [El-Kishky et al 2020]
  ○ use LASER encoder to produce sentence embedding
  ○ for every Eng sentence, use vector search engine (e.g. FAISS) to search candidate aligned sentence by comparing sentence embedding
  ○ parallel or comparable web-document pairs in 137 languages aligned with English.

• Use language family as bridge to mine
  ○ non-English pairs

• Total Training Data: 7.5B parallel sentences, corresponding to 2200 directions.
<table>
<thead>
<tr>
<th>Direction</th>
<th>Test Set</th>
<th>Published</th>
<th>BLEU m2m-100</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Without Improvement</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English-Chinese (Li et al., 2019)</td>
<td>WMT’19</td>
<td>38.2</td>
<td>33.2</td>
<td>-5.0</td>
</tr>
<tr>
<td>English-Finnish (Talman et al., 2019)</td>
<td>WMT’17</td>
<td>28.6</td>
<td>28.2</td>
<td>-0.4</td>
</tr>
<tr>
<td>English-Estonian (Pinnis et al., 2018)</td>
<td>WMT’18</td>
<td>24.4</td>
<td>24.1</td>
<td>-0.3</td>
</tr>
<tr>
<td>Chinese-English (Li et al., 2019)</td>
<td>WMT’19</td>
<td>29.1</td>
<td>29.0</td>
<td>-0.1</td>
</tr>
<tr>
<td><strong>With Improvement</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English-French (Edunov et al., 2018)</td>
<td>WMT’14</td>
<td>43.8</td>
<td>43.8</td>
<td>0</td>
</tr>
<tr>
<td>English-Latvian (Pinnis et al., 2017)</td>
<td>WMT’17</td>
<td>20.0</td>
<td>20.5</td>
<td>+0.5</td>
</tr>
<tr>
<td>German-English (Ng et al., 2019)</td>
<td>WMT’19</td>
<td>39.2</td>
<td>40.1</td>
<td>+0.9</td>
</tr>
<tr>
<td>Lithuanian-English (Pinnis et al., 2019)</td>
<td>WMT’19</td>
<td>31.7</td>
<td>32.9</td>
<td>+1.2</td>
</tr>
<tr>
<td>English-Russian (Ng et al., 2019)</td>
<td>WMT’19</td>
<td>31.9</td>
<td>33.3</td>
<td>+1.4</td>
</tr>
<tr>
<td>English-Lithuanian (Pinnis et al., 2019)</td>
<td>WMT’19</td>
<td>19.1</td>
<td>20.7</td>
<td>+1.6</td>
</tr>
<tr>
<td>Finnish-English (Talman et al., 2019)</td>
<td>WMT’17</td>
<td>32.7</td>
<td>34.3</td>
<td>+1.6</td>
</tr>
<tr>
<td>Estonian-English (Pinnis et al., 2018)</td>
<td>WMT’18</td>
<td>30.9</td>
<td>33.4</td>
<td>+2.5</td>
</tr>
<tr>
<td>Latvian-English (Pinnis et al., 2017)</td>
<td>WMT’17</td>
<td>21.9</td>
<td>24.5</td>
<td>+2.6</td>
</tr>
<tr>
<td>Russian-English (Ng et al., 2019)</td>
<td>WMT’19</td>
<td>37.2</td>
<td>40.5</td>
<td>+3.3</td>
</tr>
<tr>
<td>French-English (Edunov et al., 2018)</td>
<td>WMT’14</td>
<td>36.8</td>
<td>40.4</td>
<td>+3.6</td>
</tr>
<tr>
<td>English-German (Ng et al., 2019)</td>
<td>WMT’19</td>
<td>38.1</td>
<td>43.2</td>
<td>+5.1</td>
</tr>
<tr>
<td>English-Turkish (Sennrich et al., 2017)</td>
<td>WMT’17</td>
<td>16.2</td>
<td>23.7</td>
<td>+7.5</td>
</tr>
<tr>
<td>Turkish-English (Sennrich et al., 2017)</td>
<td>WMT’17</td>
<td>20.6</td>
<td>28.2</td>
<td>+7.6</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.0</td>
<td>31.9</td>
<td>+1.9</td>
</tr>
</tbody>
</table>
LegoMT
Lego-MT: Detachable Architecture

Each branch contains a complete encoder-decoder for a language/language group.

7 branches for central languages, and 1 branch for all languages combined.

Hello 你好
Data Flow in Lego-MT

Training Stage

Mix-Flow

Enc-Flow

Dec-Flow

Inference Stage

Hello World!

你好，世界!

三路径

Lego-MT Two-stage Training

- 1\textsuperscript{st} stage: training on many-to-many & one-to-many data

\[
\min L_{mix} + L_{enc}
\]
\[
L_{mix} = - \sum_{x, y \sim D_{\text{multi}}} \log P_{\theta_{mix}}(y \mid x)
\]
\[
L_{enc} = - \sum_{x, y \sim D_{\text{lg}}} \log P_{\theta_{enc}}(y \mid x)
\]

- 2\textsuperscript{nd} stage: training on many-to-one data

\[
L_{dec} = - \sum_{x, y \sim D_{\rightarrow \text{lg}}} \log P_{\theta_{dec}}(y \mid x)
\]

Fix the encoder of mix-flow branch

Multi-centric Data for 433 Languages

• Training Data
  ○ 1.3B sentence pairs collected from OPUS
  ○ 433 languages including 7 central languages

• Testing:
  ○ Flores-101 Devtest, human written translation pairs covering 101 languages.
  ○ 7×85 translation directions

• Evaluation Metric:
  ○ spBLEU, same in Flores-101

Lego-MT Model Configuration

• Model Parameters
  ○ Each Flow: 0.6B parameters
  ○ Total Training Parameters:
    ‣ 9.6B = 1.2B (Mix-Flow) + 0.6 * 7 (Enc-Flow) + 0.6 * 7 (Dec-Flow)
  ○ Inference Parameter:
    ‣ 1.2B (Each branch can be independently loaded during inference)
    ‣ We use Mix-flow for multilingual evaluation

• Training Setting
  ○ Max token 8000
  ○ The training of all centric languages is conducted in random order
  ○ Training duration: 15 days on 32 A100 GPUs.
Lego-MT surpasses plain ChatGPT


![Bar chart comparing Lego-MT, ChatGPT, and ChatGPT 8-Shot in Many-to-English and English-to-Many translation tasks. Lego-MT outperforms both ChatGPT and ChatGPT 8-Shot in both directions.]
Low-resource is recommended to use Mix-Flow.
High-resource is better to use E-Flow + D-Flow.

When to use Mix-flow?

Language Presentation
• Yuan et al. LegoMT: Learning Detachable Models for Massively Multilingual Machine Translation, 2023
• Aharoni et al. Massively Multilingual Neural Machine Translation. 2019
• Arivazhagan et al. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. 2019
• Bapna & Firat, Simple, Scalable Adaptation for Neural Machine Translation, 2019
• Zhu et al. Counter-Interference Adapter for Multilingual Machine Translation. 2021
• Lin et al. Learning Language Specific Sub-network for Multilingual Machine Translation. 2021
• Artetxe et al. Unsupervised Neural Machine Translation. 2018
• Lample et al. Unsupervised Machine Translation Using Monolingual Corpora Only. 2018