Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

Author: Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, Jeffrey Dean

Presented by: Kejia Jiang
Introduction

• A **single** Neural Machine Translation (NMT) model to translate between multiple languages.

• **Simplicity**
  
  Requires no change to the traditional NMT model architecture.

• **Low-resource language improvements**
  
  Language pairs with little available data and language pairs with abundant data are mixed together.

• **Zero-shot translation**
  
  Translates between arbitrary languages, including unseen language pairs during the training process.
Related work

• The multilingual model architecture is identical to Google’s Neural Machine Translation (GNMT) system (Wu et al., 2016)
  
  Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (Wu et al., 2016)

• GNMT model consists of a deep LSTM network with 8 encoder and 8 decoder layers using residual connections and attention connections.
  • Accurate
  • Fast
  • Robustness to rare words
GNMT Deep Stacked LSTMs

Encoder LSTMs

GPU8

8 layers

GPU3

GPU2

GPU1

Decoder LSTMs

Softmax

y1 → y2 → ... → </s>

Dection LSTMs

GPU8

GPU3

GPU2

GPU1

\[ x_3 \rightarrow x_2 \rightarrow ... \rightarrow </s> \]

Attention
GNMT attention module

• Context $a_i$ for the current time step is computed according to the following formulas:

$$s_t = \text{AttentionFunction}(y_{i-1}, x_t) \quad \forall t, \quad 1 \leq t \leq M$$

$$p_t = \exp(s_t) / \sum_{t=1}^{M} \exp(s_t) \quad \forall t, \quad 1 \leq t \leq M$$

$$a_i = \sum_{t=1}^{M} p_t . x_t$$

• Here the $\text{AttentionFunction}$ is a feed forward network with one hidden layer.
GNMT: Residual Connections
GNMT Residual Connections

\[ c_t^i, m_t^i = \text{LSTM}_i(c_{t-1}^i, m_{t-1}^i, x_t^{i-1}; W^i) \]

\[ x_t^i = m_t^i \]

\[ c_{t+1}^{i+1}, m_{t+1}^{i+1} = \text{LSTM}_{i+1}(c_{t-1}^{i+1}, m_{t-1}^{i+1}, x_t^i; W^{i+1}) \]

• With residual connections between LSTM_i and LSTM_{i+1}, the above equations become:

\[ c_t^i, m_t^i = \text{LSTM}_i(c_{t-1}^i, m_{t-1}^i, x_t^{i-1}; W^i) \]

\[ x_t^i = m_t^i + x_t^{i-1} \]

\[ c_{t+1}^{i+1}, m_{t+1}^{i+1} = \text{LSTM}_{i+1}(c_{t-1}^{i+1}, m_{t-1}^{i+1}, x_t^i; W^{i+1}) \]
To address the translation of out-of-vocabulary (OOV) words, GNMT applies sub-word units to do segmentation.

Example:

Word: Jet makers feud over seat width with big orders at stake.

Wordpieces: _Jet_makers_feud_over_seat_width_with_big_orders_at_stake.

This method provides a good balance between the flexibility of “character”-delimited models and the efficiency of “word”-delimited models.
GNMT with zero-shot translation

• Based on the GNMT, the system adds an artificial token at the beginning of the input sentence to indicate the target language the model should translate to.

• Example: En→Es

  Instead of:

  How are you? -> ¿Cómo estás?

  put <2es> at the beginning:

  <2es> How are you? -> ¿Cómo estás?
Zero-shot translation

- The system uses implicit bridging to deal with the problem. No explicit parallel training data has been seen.
  - Although the source and target languages should be seen individually during the training at some point.

<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>
To improve zero-shot translation quality

- Incrementally training the multilingual model on the additional parallel data for the zero-shot directions.

- Zero-shot:
  - En↔{Be,Ru,Uk}
- From-scratch:
  - En↔{Be,Ru,Uk}
  - Ru↔{Be, Uk}
- Incremental:
  - Zero-shot
  - + From-scratch

<table>
<thead>
<tr>
<th></th>
<th>Zero-Shot</th>
<th>From-Scratch</th>
<th>Incremental</th>
</tr>
</thead>
<tbody>
<tr>
<td>En→Be</td>
<td>16.85</td>
<td>17.03</td>
<td>16.99</td>
</tr>
<tr>
<td>En→Ru</td>
<td>22.21</td>
<td>22.03</td>
<td>21.92</td>
</tr>
<tr>
<td>En→Uk</td>
<td>18.16</td>
<td>17.75</td>
<td>18.27</td>
</tr>
<tr>
<td>Be→En</td>
<td>25.44</td>
<td>24.72</td>
<td>25.54</td>
</tr>
<tr>
<td>Ru→En</td>
<td>28.36</td>
<td>27.90</td>
<td>28.46</td>
</tr>
<tr>
<td>Uk→En</td>
<td>28.60</td>
<td>28.51</td>
<td>28.58</td>
</tr>
<tr>
<td>Be→Ru</td>
<td>56.53</td>
<td>82.50</td>
<td>78.63</td>
</tr>
<tr>
<td>Ru→Be</td>
<td>58.75</td>
<td>72.06</td>
<td>70.01</td>
</tr>
<tr>
<td>Ru→Uk</td>
<td>21.92</td>
<td>25.75</td>
<td>25.34</td>
</tr>
<tr>
<td>Uk→Ru</td>
<td>16.73</td>
<td>30.53</td>
<td>29.92</td>
</tr>
</tbody>
</table>
Mixed language

• Can a multilingual model successfully handle multi-language input (code-switching) in the middle of a sentence?
• Yes! Because the individual characters/wordpieces are present in the shared vocabulary.

  • **Japanese:** 私は東京大学の学生です。 → I am a student at Tokyo University.
  • **Korean:** 나는 도쿄 대학의 학생입니다. → I am a student at Tokyo University.
  • **Japanese/Korean:** 我是東京大学学生. → I am a student of Tokyo University.
Mixed language (2)

• What happens when a multilingual model is triggered with a linear mix of two target language tokens?

• Example:
  Using a multilingual En→{Ja, Ko} model, feed a linear combination \((1-w)<2ja>+w<2ko>\) of the embedding vectors for “<2ja>” and “<2ko>”, \(0 <= w <= 1\).
  Result : with \(w = 0.5\), the model switches languages mid-sentence.
\( w_{ko} \)  I must be getting somewhere near the centre of the earth.

<table>
<thead>
<tr>
<th>Score</th>
<th>Korean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>私は地球の中心の近くにどこかに行っているに違いない。</td>
</tr>
<tr>
<td>0.40</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>0.56</td>
<td>私は地球の中心の近くのどこかになっているに違いない。</td>
</tr>
<tr>
<td>0.58</td>
<td>私は지구의 중심에서가까이에어던가에도착하고있어야한다。</td>
</tr>
<tr>
<td>0.60</td>
<td>나는지구의센터의가까이에어던가에도착하고있어야한다。</td>
</tr>
<tr>
<td>0.70</td>
<td>나는지구의중심근처에어던가에도착해야합니다。</td>
</tr>
<tr>
<td>0.90</td>
<td>나는어던가지구의중심근처에도착해야합니다。</td>
</tr>
<tr>
<td>1.00</td>
<td>나는어던가지구의중심근처에도착해야합니다。</td>
</tr>
</tbody>
</table>
Conclusion

• Use a single model where all parameters are shared, which improves the translation quality of low resource languages in the mix.

• Zero-shot translation without explicit bridging is possible.

• To improve the zero-shot translation quality:
  Incrementally training the multilingual model on the additional parallel data for the zero-shot directions.

• Mix languages on the source or target side can yield interesting but reliable translation results.
Thank you!