Kyoto University MT System

Description for IWSLT 2017

Raj Dabre\textsuperscript{1}, Fabien Cromieres\textsuperscript{2}, Sadao Kurohashi\textsuperscript{1}

\textsuperscript{1}Graduate School of Informatics, Kyoto University, Kyoto, Japan
\textsuperscript{2}Japan Science and Technology Agency, Saitama, Japan

14-12-2017 IWSLT 2017
QR code to access Slides
Flow of This Talk

● Overview
  ○ Multilingual Task
  ○ AIAYN

● Our approaches
  ○ Using NMT
  ○ Using SMT (for internal evaluation)

● Experimental Settings

● Results and Observations

● Conclusion
Multilingual Task

- 5 languages
  - German, Dutch, Romanian, Italian and English
  - 3 Germanic and 2 Romance
- Objective: One multilingual model for all 5 languages (20 directions)
- Non zero-shot setting
  - Use all data (20 parallel corpora)
- Zero-shot setting
  - All data except for German-Dutch, Dutch-German, Romanian-Italian and Italian-Romanian (16 parallel corpora)
Preferred Paradigm: Non-Recurrent NMT

● Why NMT?
  ○ Easier to develop end-to-end multilingual models with parameter sharing (Johnson et al., 2016)
  ○ NMT as a black box is good enough

● Why Non-Recurrent?
  ○ Faster to train (multilingual model training takes time as it is)
  ○ Known to perform better than recurrent models (Vaswani et al., 2017)
Attention is All You Need (AIAYN)

- Faster training
  - Feed Forward Layers
  - Positional Encoding
  - Residual connections
  - Batch Normalization
- Better attention mechanism
  - Multi-head
  - Self and cross
- Adam with decay

Taken from Vaswani et al. 2017
Encoder

- $H_{Memes}$
- $H_{are}$
- $H_{life}$

Feed Forward Layer

Self Attention Layer

$E_{Memes}$
$E_{are}$
$E_{life}$

N such layers
Residual connections
Normalization
Attention

Scaled Dot-Product Attention

Multi-Head Attention

Taken from Vaswani et al. 2017
Our Approach: MLNMT Using Artificial Tokens
Multilingual Phrase based SMT

- Hacky Approach
- Only works for non-zero shot conditions
- **Technique:** For each language pair append “#tgt” to each source token
- **Example:**
  - Original: “I am a boy” --> “Watashi wa otokonoko desu”
  - Modified: “I#ja am#ja a#ja boy#ja” --> “Watashi wa otokonoko desu”
- **Outcome:** Single phrase table with multiple language directions
- **Working:** Token “#tgt” helps match phrase pairs for exactly one language pair
Experimental Settings

● Corpora
  ○ 20 way corpora provided by organizers (~200K sentences per direction)
  ○ dev2010 and tst2010 for internal evaluation
  ○ tst2017 for official evaluation

● Generic Preprocessing
  ○ XML to Moses format
  ○ Tokenization (using Moses tokenizer)
  ○ Truecasing (using Moses truecaser)

● Specific Preprocessing For NMT:
  ○ Prepending the "<2xx>" token to source sentences for all corpora

● Specific Preprocessing For PBSMT:
  ○ Appending "#xx" token to all source word tokens for all corpora
  ○ Byte Pair Encoding
    ■ Not needed for NMT: AIAYN has in built sub-word encoder
PBSMT SETTINGS

- Moses toolkit for training, tuning and testing
- Sub-word vocabulary size: 32000
- Language model: 7-gram KenLM
- Default settings for alignment and phrase extraction, tuning and testing.
NMT SETTINGS

- Google’s implementation of AIAYN
  - [https://github.com/tensorflow/tensor2tensor](https://github.com/tensorflow/tensor2tensor)
- Sub-word vocabulary size of 32000 (managed by EMS)
- Embedding and output layer sizes: 512
- Feed forward hidden layer size: 2018
- Adam optimizer with weight decay (Noam LR Decay)
  - 16000 of learning rate warmup before decay
- Beam search decoding:
  - Beam width of size 4
  - Alpha of 0.6 (for decoded sequence length penalty)
- Iterations: 400000 (~10 epochs)
- Data parallelism: 5 GPUs (3-4 days for convergence)
**Internal Evaluation (tst2010)**

- NMT is inherently superior to PBSMT
- But needs 3-4 times longer training time
- PBSMT does not really allow for languages to interact
  - No parameters are shared in reality
  - Phrase table sharing is more of a hack

<table>
<thead>
<tr>
<th>L1/L2</th>
<th>de</th>
<th>en</th>
<th>it</th>
<th>nl</th>
<th>ro</th>
</tr>
</thead>
<tbody>
<tr>
<td>de</td>
<td>-</td>
<td>29.63</td>
<td>17.57</td>
<td>23.51</td>
<td>14.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>34.98</td>
<td>21.37</td>
<td>23.69</td>
<td>18.96</td>
</tr>
<tr>
<td>en</td>
<td>21.70</td>
<td>-</td>
<td>24.04</td>
<td>27.25</td>
<td>21.38</td>
</tr>
<tr>
<td></td>
<td>27.81</td>
<td>29.07</td>
<td>30.91</td>
<td>26.65</td>
<td></td>
</tr>
<tr>
<td>it</td>
<td>15.88</td>
<td>28.89</td>
<td>-</td>
<td>18.48</td>
<td>19.46</td>
</tr>
<tr>
<td></td>
<td>21.37</td>
<td>34.58</td>
<td></td>
<td>21.83</td>
<td>20.72</td>
</tr>
<tr>
<td>nl</td>
<td>21.57</td>
<td>34.79</td>
<td>18.84</td>
<td>-</td>
<td>15.99</td>
</tr>
<tr>
<td></td>
<td>24.45</td>
<td>38.86</td>
<td>23.02</td>
<td></td>
<td>20.68</td>
</tr>
<tr>
<td>ro</td>
<td>15.96</td>
<td>31.10</td>
<td>22.65</td>
<td>18.57</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>21.81</td>
<td>37.10</td>
<td>24.07</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Upper score is SMT
Lower score is NMT
Official Evaluation: tst2017

- **Surprise:** Zero-shot results are almost as good as non-zero shot results
- **Analysis:** Extracted 5-lingual corpora from the 20 parallel corpora
- **Observation:** 150k sentences are 5 lingual
  - 60% of corpus
- **Conclusion:** Missing parallel sentences between Italian and Romanian and Dutch and German are remedied by indirect translations from other languages

### Table

#### Non Zero Shot

<table>
<thead>
<tr>
<th>L1/L2</th>
<th>de</th>
<th>en</th>
<th>it</th>
<th>nl</th>
<th>ro</th>
</tr>
</thead>
<tbody>
<tr>
<td>de</td>
<td>-</td>
<td>26.45</td>
<td>17.54</td>
<td>19.64</td>
<td>16.27</td>
</tr>
<tr>
<td>en</td>
<td>23.25</td>
<td>-</td>
<td>30.79</td>
<td>28.80</td>
<td>24.66</td>
</tr>
<tr>
<td>it</td>
<td>19.10</td>
<td>34.73</td>
<td>-</td>
<td>22.32</td>
<td>20.60</td>
</tr>
<tr>
<td>nl</td>
<td>20.27</td>
<td>30.49</td>
<td>19.86</td>
<td>-</td>
<td>17.65</td>
</tr>
<tr>
<td>ro</td>
<td>17.94</td>
<td>29.58</td>
<td>21.89</td>
<td>20.24</td>
<td>-</td>
</tr>
</tbody>
</table>

#### Zero Shot

<table>
<thead>
<tr>
<th>L1/L2</th>
<th>de</th>
<th>en</th>
<th>it</th>
<th>nl</th>
<th>ro</th>
</tr>
</thead>
<tbody>
<tr>
<td>de</td>
<td>-</td>
<td>27.08</td>
<td>17.67</td>
<td><strong>20.31</strong></td>
<td>16.08</td>
</tr>
<tr>
<td>en</td>
<td>23.63</td>
<td>-</td>
<td>30.99</td>
<td>30.18</td>
<td>24.49</td>
</tr>
<tr>
<td>it</td>
<td>19.20</td>
<td>35.28</td>
<td>-</td>
<td>22.76</td>
<td><strong>20.37</strong></td>
</tr>
<tr>
<td>nl</td>
<td><strong>19.68</strong></td>
<td>30.63</td>
<td>20.74</td>
<td>-</td>
<td>17.74</td>
</tr>
<tr>
<td>ro</td>
<td>18.40</td>
<td>30.23</td>
<td><strong>21.85</strong></td>
<td>20.47</td>
<td>-</td>
</tr>
</tbody>
</table>
How does MLNMT stack against bilingual models?

- **Dutch-German**
  - Bilingual: 19.5
  - Non zero shot: **20.27**
  - Zero shot: 19.68
- **Romanian-Italian**
  - Bilingual: **23.14**
  - Non zero shot: 21.89
  - Zero shot: 21.85
- More or less comparable performance
- Bilingual models required a few hours of training on 5 GPUs
Conclusions and Future Work

- Set foundations for low resource multilingual NMT baselines
- AIAYN is fast and effective
  - Better than PBSMT setting we tried
- Zero-shot performance is almost as good or better than non zero-shot performance
  - Suspicion: Setting is not truly zero shot
- Future work
  - Train more robust models (dropout, annealing, checkpoint averaging)
  - Try out stricter zero-shot conditions
  - Better training methods for related languages (European)
  - Modifications for AIAYN for multilinguality
Thank You for Listening