Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder

Thanh-Le Ha, Jan Niehues and Alexander Waibel
Outline

- Introduction
- Multilingual Neural Machine Translation
  - Related works
  - Our proposed approach
- Experimental results
- Conclusion & Future Work
Attention Neural Machine Translation

Transcribed Source Sentence (spoken German):

Ich bin nach Hause gegangen <EoS>

Transcribed Translated Sentence (English):

I went home <EoS>

Diagram illustrating the process of attention in neural machine translation with an English and a German sentence.
Multilingual NMT: Prospective Benefits

Automatic Language Transfer:
- Can be applied to under-resource scenarios

Number of parameters grows linearly with the number of languages
Multilingual NMT: Challenges

- Attention is language-specific
  ⇒ An encoder-attention-decoder triple for each language pair

- Multilingual (and Multimodal) models [Luong 2016]
  - Without attention: attention is modality-(and language-)specific

- One-to-Many NMT [Dong 2015]
  - Single encoder, several pairs of attention-decoder for each target language
Multilingual NMT: Challenges

- Attention is language-specific
  ⇒ An encoder-attention-decoder triple for each language pair

- Want a shared attention or decoder NMT?
  ⇒ We must modify the architecture

- Many-to-One NMT [Zoph&Knight 2016]
  - Many encoders, need addition layers to combine their outputs before feeding to the attention.

- Multilingual (Many-to-Many) NMT [Firat 2016 papers]
  - Multi-way: Multiple encoders and decoders
  - With shared attention (they must change their architecture)
Multilingual NMT: Our motivations

- NMT should learn **common semantic space** of all languages
  - “work”, “working” and “worked”
  - “car” and “automobile”
  - “player”-English, “joueur”-French, “Spieler”-German

[From Socher 2012]
Multilingual NMT: Our approach

- NMT should learn common semantic space of all languages

- Our multilingual NMT system should:
  - Learn language-independent source and target sentence representations
  - Have a shared language-dependent word embeddings

=> A simple preprocessing step: Language-specific Coding
Multilingual NMT: Our approach

Language-specific Coding

- Append a language code to the words belonging to that language:
  - (excuse me | excusez moi) (En-Fr)
    \[ \Rightarrow (EN\textunderscore excuse EN\textunderscore me | FR\textunderscore excusez FR\textunderscore moi) \]
  - (entschuldigen Sie | excusez moi) (De-Fr)
    \[ \Rightarrow (DE\textunderscore entschuldigen DE\textunderscore Sie | FR\textunderscore excusez FR\textunderscore moi) \]
Multilingual NMT: Our approach

- Able to feature attention mechanism for multilingual NMT
- Everything (encoder, attention, decoder) is shared (universal)
- Do not need to change the NMT architecture
  - Language-specific coding is a preprocessing step
  - Can use any NMT framework with any translation unit
Experiments

- Training, validation and testing data
  - TED talks from WIT3
  - WMT’16 parallel and monolingual data

- Framework: Nematus [Sennrich 2016]
  - Sub-word with BPE on joint corpus
  - Vocabularies’ size: 40K, sentence-length cut-off at 50
  - One 1024-cell GRU layer, one 1000D embeddings for encoder and decoder
  - Adadelta, mini-batch size: 80. grad norm: 0.1
  - Dropout at every layer

- Experiments on different scenarios:
  - Under-resource (simulated): En-De TED
  - Large-scale, real task: IWSLT’16: En-De WMT tuning on TED
Goal:
- Translating En to De
- Using multilingual corpora:
  - En-De: TED 196K
  - Fr-De: TED 165K
- Two kinds of configurations: Mix-source & Multi-source
Experiments: Mix-source Multilingual NMT

Language-specific coded English

EN_excuse EN_me
EN_see EN_ya EN_soon

DE_entschuldigen DE_Sie
DE_bis DE_bald

Language-specific coded German

Multilinguality in Neural Machine Translation

Institute for Anthropomatics and Robotics
Experiments: Multi-source Multilingual NMT

Language-specific coded English
EN_excuse EN_me
EN_see EN_ya ENSoon

Language-specific coded French
FR_excusez FR_moi
FR_merci FR_moi

Language-specific coded German
DE_entschuldigen DE_Sie
DE_bis DE_bald

Language-specific coded German
DE_entschuldigen DE_Sie
DE_danke DE_tön

Multi-specific Coding
## Experiments: Under-resource scenario

<table>
<thead>
<tr>
<th>System</th>
<th>tst2013</th>
<th>tst2014</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>ΔBLEU</td>
</tr>
<tr>
<td>Baseline (En =&gt; De)</td>
<td>24.35</td>
<td>-</td>
</tr>
<tr>
<td>Mix-source (En,De =&gt; De,De)</td>
<td>26.99</td>
<td>+2.64</td>
</tr>
<tr>
<td>Multi-source (En,Fr =&gt; De,De)</td>
<td>26.64</td>
<td>+2.21</td>
</tr>
</tbody>
</table>

Both **Mix-source** and **Multi-source** improve the translation significantly.
## Experiments: Under-resource scenario

<table>
<thead>
<tr>
<th>System</th>
<th>tst2013</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>ΔBLEU</td>
</tr>
<tr>
<td>Baseline (En =&gt; De)</td>
<td>24.35</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mix-source (En,De =&gt; De,De)</td>
<td>26.99</td>
<td>+2.64</td>
<td>22.71</td>
</tr>
<tr>
<td>Multi-source (En,Fr =&gt; De,De)</td>
<td>26.64</td>
<td>+2.21</td>
<td>22.21</td>
</tr>
<tr>
<td>Baseline 2 (En =&gt; De) x2</td>
<td>24.58</td>
<td>+0.23</td>
<td>20.55</td>
</tr>
</tbody>
</table>

- Both **Mix-source** and **Multi-source** improve the translation significantly
- Because we have larger data (double the baseline)?
  - Baseline 2: Double the corpus
Experiments: Under-resource scenario

<table>
<thead>
<tr>
<th>System</th>
<th>tst2013</th>
<th>tst2014</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>ΔBLEU</td>
</tr>
<tr>
<td>Baseline (En =&gt; De)</td>
<td>24.35</td>
<td>-</td>
</tr>
<tr>
<td>Mix-source (En,De =&gt; De,De)</td>
<td>26.99</td>
<td>+2.64</td>
</tr>
<tr>
<td>Multi-source (En,Fr =&gt; De,De)</td>
<td>26.64</td>
<td>+2.21</td>
</tr>
<tr>
<td>Mix-source 2 (En,De =&gt; De,De)</td>
<td>27.18</td>
<td>+2.83</td>
</tr>
</tbody>
</table>

- **Multi-source** performs worse than **Mix-source**. Why?
  - **Smaller training data?**
    - Mix-source: 392K, Multi-source: 361K
    - Mix-source 2: De part of En-Fr: 361K < Mix-source:392K
  - **Having more data in other languages confuses NMT?**
  - Need more analyses (more source languages, more language types)
Experiments: Multi-source Visualization

- Take the source word embeddings (1000 dims) to visualize
- Using t-SNE [Maaten 2008] to project to 2-dim points

En&Fr Word Embeddings topic “human”
Experiments: Multi-source Visualization

- Take the source word embeddings (1000 dims) to visualize
- Using t-SNE [Maaten 2008] to project to 2-dim points

En&Fr Word Embeddings topic “computer”
Experiments: Large-scale, real task

Translate En-De for the real task of IWSLT16:
- Baseline: WMT data + BackTranslation

Train Mix-source configuration on:
- 1) WMT parallel data (En-De) + sampled additional mono data (De-De)
- 2) WMT parallel data (En-De) + mono part of that parallel data (De-De)

Adapt on TED En-De (continue training):
- Also Mix-source on TED

<table>
<thead>
<tr>
<th>System</th>
<th>tst2013</th>
<th>tst2014</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>ΔBLEU</td>
</tr>
<tr>
<td>Baseline (En =&gt; De)</td>
<td>25.74</td>
<td>-</td>
</tr>
<tr>
<td>1) Sampled Mix-source (En,De =&gt; De,De)</td>
<td>27.74</td>
<td>+2.00</td>
</tr>
<tr>
<td>2) Mono Mix-source (En,De =&gt; De,De)</td>
<td>28.89</td>
<td>+3.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System</th>
<th>tst2013</th>
<th>tst2014</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>ΔBLEU</td>
</tr>
<tr>
<td>Baseline (En =&gt; De)</td>
<td>25.74</td>
<td>-</td>
</tr>
<tr>
<td>1) Sampled Mix-source (En,De =&gt; De,De)</td>
<td>27.74</td>
<td>+2.00</td>
</tr>
<tr>
<td>2) Mono Mix-source (En,De =&gt; De,De)</td>
<td>28.89</td>
<td>+3.15</td>
</tr>
</tbody>
</table>
Conclusion & Future work

Conclusion

- We proposed a simple but elegant approach for multilingual NMT
  - Allows to use attention seamlessly
  - A preprocessing step, no need to change an NMT architecture
- Improve significantly in under-resource scenarios
- Provide natural, effective way to leverage monolingual data in NMT

Future work

- More languages and the impact
- Apply multilingual NMT in zero-resource scenario
Multilinguality in Neural Machine Translation