Multilingual and Multitask Learning in seq2seq Models

CMSC 470
Marine Carpuat
Multilingual Machine Translation
Neural MT only helps in high-resource settings

Ongoing research

- Learn from other sources of supervision than pairs (E,F)
  - Monolingual text
  - Multiple languages

- Incorporate linguistic knowledge
  - As additional embeddings
  - As prior on network structure or parameters
  - To make better use of training data

[Koehn & Knowles 2017]
Multilingual Translation

• Goal: support translation between any N languages

• Naïve approach: build on translation system for each language pair and translation direction
  • Results in $N^2$ models
  • Impractical computation time
  • Some language pairs have more training data than others

• Can we train a single model instead?
The Google Multilingual NMT System
[Johnson et al. 2017]
The Google Multilingual NMT System
[Johnson et al. 2017]

• Shared encoder, shared decoder for all languages

• Train on sentence pairs in all languages

• Add token to the input to mark target language

<2es> Hello, how are you? → Hola, ¿cómo estás?
A standard encoder-decoder LSTM architecture, updated to enable parallelization/multi-GPU training.
Pros and Cons?

**Advantages**
- Translation for low resource languages benefits from data for high resource languages
- Enables “zero shot” translation
  - Translation between language pairs which have not been seen (as a pair) during training
- Can handle code-switched input
  - Sequences that contain more than one language

**Drawbacks/Issues**
- Requires a single shared vocabulary for all languages
  - BPE, wordpiece
- Model size
- Opaque
- No direct control on output language
  - Bias toward high-resource languages?
How well does this work?

Evaluation Set Up

WMT
  Train
    English↔French(Fr)
    English↔German(De)
  Test: newstest2014+15

Google production
  English↔Japanese(Ja)
  English↔Korean(Ko)
  English↔Spanish(Es)
  English↔Portuguese(Pt)

BLEU evaluation
### BLEU scores in the “many to one” condition

<table>
<thead>
<tr>
<th>Model</th>
<th>Single</th>
<th>Multi</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT German→English (oversampling)</td>
<td>30.43</td>
<td>30.59</td>
<td>+0.16</td>
</tr>
<tr>
<td>WMT French→English (oversampling)</td>
<td>35.50</td>
<td>35.73</td>
<td>+0.23</td>
</tr>
<tr>
<td>WMT German→English (no oversampling)</td>
<td>30.43</td>
<td>30.54</td>
<td>+0.11</td>
</tr>
<tr>
<td>WMT French→English (no oversampling)</td>
<td>35.50</td>
<td>36.77</td>
<td>+1.27</td>
</tr>
<tr>
<td>Prod Japanese→English</td>
<td>23.41</td>
<td>23.87</td>
<td>+0.46</td>
</tr>
<tr>
<td>Prod Korean→English</td>
<td>25.42</td>
<td>25.47</td>
<td>+0.05</td>
</tr>
<tr>
<td>Prod Spanish→English</td>
<td>38.00</td>
<td>38.73</td>
<td>+0.73</td>
</tr>
<tr>
<td>Prod Portuguese→English</td>
<td>44.40</td>
<td>45.19</td>
<td>+0.79</td>
</tr>
</tbody>
</table>

**Note:** Single language pair baseline vs. Multilingual model.
### BLEU scores in the “one to many” condition

<table>
<thead>
<tr>
<th>Model</th>
<th>Single</th>
<th>Multi</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT English→German (oversampling)</td>
<td>24.67</td>
<td>24.97</td>
<td>+0.30</td>
</tr>
<tr>
<td>WMT English→French (oversampling)</td>
<td>38.95</td>
<td>36.84</td>
<td>-2.11</td>
</tr>
<tr>
<td>WMT English→German (no oversampling)</td>
<td>24.67</td>
<td>22.61</td>
<td>-2.06</td>
</tr>
<tr>
<td>WMT English→French (no oversampling)</td>
<td>38.95</td>
<td>38.16</td>
<td>-0.79</td>
</tr>
<tr>
<td>Prod English→Japanese</td>
<td>23.66</td>
<td>23.73</td>
<td>+0.07</td>
</tr>
<tr>
<td>Prod English→Korean</td>
<td>19.75</td>
<td>19.58</td>
<td>-0.17</td>
</tr>
<tr>
<td>Prod English→Spanish</td>
<td>34.50</td>
<td>35.40</td>
<td>+0.90</td>
</tr>
<tr>
<td>Prod English→Portuguese</td>
<td>38.40</td>
<td>38.63</td>
<td>+0.23</td>
</tr>
</tbody>
</table>

**Note:** Single language pair baseline vs Multilingual model.
BLEU scores in the “many to many” condition

<table>
<thead>
<tr>
<th>Model</th>
<th>Single</th>
<th>Multi</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT English→German (oversampling)</td>
<td>24.67</td>
<td>24.49</td>
<td>-0.18</td>
</tr>
<tr>
<td>WMT English→French (oversampling)</td>
<td>38.95</td>
<td>36.23</td>
<td>-2.72</td>
</tr>
<tr>
<td>WMT German→English (oversampling)</td>
<td>30.43</td>
<td>29.84</td>
<td>-0.59</td>
</tr>
<tr>
<td>WMT French→English (oversampling)</td>
<td>35.50</td>
<td>34.89</td>
<td>-0.61</td>
</tr>
<tr>
<td>WMT English→German (no oversampling)</td>
<td>24.67</td>
<td>21.92</td>
<td>-2.75</td>
</tr>
<tr>
<td>WMT English→French (no oversampling)</td>
<td>38.95</td>
<td>37.45</td>
<td>-1.50</td>
</tr>
<tr>
<td>WMT German→English (no oversampling)</td>
<td>30.43</td>
<td>29.22</td>
<td>-1.21</td>
</tr>
<tr>
<td>WMT French→English (no oversampling)</td>
<td>35.50</td>
<td>35.93</td>
<td>+0.43</td>
</tr>
<tr>
<td>Prod English→Japanese</td>
<td>23.66</td>
<td>23.12</td>
<td>-0.54</td>
</tr>
<tr>
<td>Prod English→Korean</td>
<td>19.75</td>
<td>19.73</td>
<td>-0.02</td>
</tr>
<tr>
<td>Prod Japanese→English</td>
<td>23.41</td>
<td>22.86</td>
<td>-0.55</td>
</tr>
<tr>
<td>Prod Korean→English</td>
<td>25.42</td>
<td>24.76</td>
<td>-0.66</td>
</tr>
<tr>
<td>Prod English→Spanish</td>
<td>34.50</td>
<td>34.69</td>
<td>+0.19</td>
</tr>
<tr>
<td>Prod English→Portuguese</td>
<td>38.40</td>
<td>37.25</td>
<td>-1.15</td>
</tr>
<tr>
<td>Prod Spanish→English</td>
<td>38.00</td>
<td>37.65</td>
<td>-0.35</td>
</tr>
<tr>
<td>Prod Portuguese→English</td>
<td>44.40</td>
<td>44.02</td>
<td>-0.38</td>
</tr>
</tbody>
</table>
### Impact of model size in “many to many” condition

<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>#nodes</td>
<td>1024</td>
<td>1024</td>
<td>1280</td>
<td>1536</td>
<td>1792</td>
</tr>
<tr>
<td>#params</td>
<td>3B</td>
<td>255M</td>
<td>367M</td>
<td>499M</td>
<td>650M</td>
</tr>
<tr>
<td>Prod English→Korean</td>
<td>19.75</td>
<td>18.41</td>
<td>18.36</td>
<td>18.30</td>
<td>18.28</td>
</tr>
<tr>
<td>Prod Japanese→English</td>
<td>23.41</td>
<td>21.62</td>
<td>22.03</td>
<td>22.51</td>
<td>23.18</td>
</tr>
<tr>
<td>Prod Korean→English</td>
<td>25.42</td>
<td>22.87</td>
<td>23.46</td>
<td>24.00</td>
<td>24.67</td>
</tr>
<tr>
<td>Prod English→Spanish</td>
<td>34.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prod English→Portuguese</td>
<td>38.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prod Spanish→English</td>
<td>38.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prod Portuguese→English</td>
<td>44.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prod English→German</td>
<td>26.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prod English→French</td>
<td>35.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prod German→English</td>
<td>31.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prod French→English</td>
<td>36.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ave diff vs single</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Findings so far: multilingual model
- can improve translation quality (BLEU) for low resource language pairs
- reduce training costs compared to training one model per language pair, at no (or little) loss in translation quality
Follow up work: evaluating multilingual models at scale

- 25+ billion sentence pairs
- from 100+ languages to and from English
- with 50+ billion parameters

• Comparing against strong bilingual baselines

Data distribution over language pairs

https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html
Follow up work: evaluating multilingual models at scale

- The multilingual model improves BLEU by 5 points (on average) for low-resource language pairs.
- With multilingual and bilingual models of the same capacity (i.e. number of parameters)!
- Suggests that the multilingual model is able to transfer knowledge from high-resource to low-resource languages.

Translation quality comparison of a single massively multilingual model against bilingual baselines that are trained for each one of the 103 language pairs.
Analysis: representations in multilingual model cluster by language family [Kudugunta et al. 2019]
Multilingual Machine Translation

Summary

• A simple idea:
  • Shared model for all language pairs
  • Add a token to input to identify output language

• Improves BLEU for low-resource language pairs

• But open questions remain
  • How to train massive models efficiently?
  • What properties are transferred from one language to another?
  • Are there unwanted effects on translation output? Bias toward high-resource languages / dominant language families?
Multitask Models for Controlling MT Output Style

Case Study I: formality
Style Matters for Translation

Is it more "Hey Dude" or "Dear Sir"?
Improve translation accuracy by telling us the tone of the content.
New Task: Formality-Sensitive Machine Translation (FSMT)

Comment ça va? \( f \), Source \( f \) \( \Rightarrow \) FSMT \( \theta \) \( \Rightarrow \) Translation-1 \( e_1 \) How are you doing?

or

Desired formality level \( \ell \) \( \Rightarrow \) Translation-2 \( e_2 \) What's up?

How to train?

\[ f, \ell_1, e_1 \] Ideal training data doesn’t occur naturally!
Formality in MT Corpora

delegates are kindly requested to bring their copies of documents to meetings.

in these centers, the children were fed, medically treated and rehabilitated on both a physical and mental level.

there can be no turning back the clock.

I just wanted to introduce myself.

-yeah, bro, up top.
Formality Transfer (FT)

Given a large parallel formal-informal corpus (e.g., Grammarly’s Yahoo Answers Formality Corpus) these are sequence-to-sequence tasks

[Rao and Tetreault, 2018]
Formality Sensitive MT as Multitask Formality Transfer + MT

How are you doing?  
What's up?  
Comment ça va?  
To formal or informal?

Bî-FT +FSMT

Formal-Target  
Informal-Target

EN  
FR

EN  
EN

EN  
EN
Multitask Formality Transfer + MT

• Model: shared encoder, shared decoder as in multilingual NMT [Johnson et al. 2017]

• Training objective:

\[ \mathcal{L}_{MT} + \mathcal{L}_{FT} \]

\[ \mathcal{L}_{MT} = \sum_{(X,Y)} \log P(Y|X; \theta) \quad \text{MT pairs} \]

\[ \mathcal{L}_{FT} = \sum_{(Y_{\ell}, Y_{\ell})} \log P(Y_{\ell} | Y_{\ell}, \ell; \theta) \quad \text{FT pairs} \]
## Formality Transfer MT
### Human Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Formality Difference</th>
<th>Meaning Preservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiTask</td>
<td>0.35</td>
<td>2.95</td>
</tr>
<tr>
<td>Phrase-based MT + formality reranking</td>
<td>0.05</td>
<td>2.97</td>
</tr>
</tbody>
</table>

| Niu & Carpuat 2017                      |

- 300 samples per model
- 3 judgments per sample
- Protocol based on Rao & Tetreault
Multitask model makes more formality changes than re-ranking baseline

<table>
<thead>
<tr>
<th></th>
<th>Reference</th>
<th>Formal</th>
<th>MultiTask</th>
<th>You need to be quiet and answer the question, Chief Toohey.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Refrain from the commentary and respond to the question, Chief Toohey.</td>
<td>Baseline</td>
<td>Please refrain from comment and just answer the question, Tooheys’s boss.</td>
<td>Please refrain from comment and just answer the question, Tooheys’s boss.</td>
</tr>
<tr>
<td>Informal</td>
<td>MultiTask</td>
<td>Baseline</td>
<td>Please refrain from comment and answer my question, Tooheys’s boss.</td>
<td>Please refrain from comment and answer my question, Tooheys’s boss.</td>
</tr>
</tbody>
</table>
Multitask model introduces more meaning errors than re-ranking baseline

<table>
<thead>
<tr>
<th>Reference</th>
<th>Try to file any additional motions as soon as you can.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formal</td>
<td>You should try to introduce the <strong>sharks</strong> as soon as you can.</td>
</tr>
<tr>
<td>MultiTask</td>
<td>Try to introduce any additional requests as soon as you can.</td>
</tr>
<tr>
<td>Baseline</td>
<td>Try to introduce any additional requests as soon as you can.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Informal</th>
<th>Try to introduce <strong>sharks</strong> as soon as you can.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiTask</td>
<td>Try to introduce any additional requests as soon as you can.</td>
</tr>
<tr>
<td>Baseline</td>
<td>Try to introduce any additional requests as soon as you can.</td>
</tr>
</tbody>
</table>

Meaning errors can be addressed by introducing additional synthetic supervision
[Niu, PhD thesis 2019]
Controlling Machine Translation formality via multitask learning

- A multitask formality transfer + MT model
- Can produce distinct formal/informal translations of same input
- Introduces more formality rewrites, while roughly preserving meaning, esp. with synthetic supervision

Details:

[github.com/xingniu/multitask-ft-fsmt](https://github.com/xingniu/multitask-ft-fsmt)
Multitask Models for Controlling MT Output Style

Case Study II: Complexity
Our goal: control the complexity of MT output

To make machine translation output accessible to broader audiences

Es:  
El museo Mauritshuis abre una exposición dedicada a los autorretratos del siglo XVII.

En (grade 8):  
The Mauritshuis museum is staging an exhibition focused solely on 17th century self-portraits.

En (grade 3):  
The Mauritshuis museum is going to show self-portraits.
Our goal: control the complexity of MT output

- El museo Mauritshuis abre una exposición dedicada a los autorretratos del siglo XVII.
- Desired output reading grade level [2-10]
- Complexity Controlled MT
- The Mauritshuis museum is going to show self-portraits.
Summary

What you should know

• Multitask sequence-to-sequence models
  • How they are defined and trained (loss function)

• A simple yet powerful approach that can be applied to many translation and related sequence-to-sequence tasks
  • Can help improve performance by sharing data from multiple tasks
  • Has been applied to multilingual MT, style controlled MT, among other tasks

Also in discussing recent research papers, we illustrated:

• Pros and cons of automatic vs. manual evaluation
• Experiment design and result interpretation