Dynamics of Multilingual Translation

Sunit Bhattacharya

September 25, 2020
Introduction

Experiments
  Experiments with Czech
  Experiments with Czech+German
  Experiments with Czech+German+French

Sentinel Attention Activation Ratio

Future Work
Introduction
Outline

Introduction

Experiments
  - Experiments with Czech
  - Experiments with Czech+German
  - Experiments with Czech+German+French

Sentinel Attention Activation Ratio

Future Work
Introduction

• NMT is generally modelled as sequence-to-sequence learning problem

Figure 1: Architecture of NMT system with attention

• The architecture has been successfully extended for multilingual translation and image captioning tasks.
Interpreting attention weights

“If you can’t explain it simply, you don’t understand it well enough.”

- Albert Einstein

• A clear interpretation of how exactly neural networks do what they do and how they do it is often unclear.

• Using the attention mechanism as a tool for understanding model behavior has been proposed and implemented (Mareček and Rosa (2018), Pham et al. (2019)).

• There is however a debate pertaining to the usefulness of attention weights as a measure of interpretability.
  
  • Some feel that attention cannot be used to understand the basis for prediction for models (Jain and Wallace (2019), Serrano and Smith (2019)).
  
  • Others (Vashishth et al. (2019), Vig and Belinkov (2019)) have shown that attention weights are interpretable and are capable of capturing linguistic notions and giving ‘human-interpretable descriptions of model behavior’.

Introduction  Experiments  Sentinel Attention Activation Ratio  Future Work
Introduction

Experiments
   Experiments with Czech
   Experiments with Czech+German
   Experiments with Czech+German+French

Sentinel Attention Activation Ratio

Future Work
Architecture

Figure 2: High-level network overview
• All experiments were done using NeuralMonkey and using data from the Multi30k dataset.
• The architecture is based on models successfully used for multimodal tasks (Libovický and Helcl, 2017).
• Three kinds of experiments were done:
  • Mono-encoder experiments: 1 encoder - 1 decoder
  • Bi-encoder experiments: 2 encoders - 1 decoder
  • Tri-encoder experiments: 3 encoders - 1 decoder
• Models with different combinations of French, Czech and German for the encoder and English for the decoder were trained.
• Analysis was done of the “forced decoding” model behavior when scoring a given expected output.
• For the sake of analysis, a criteria where the BLEU performance of the validation set does not improve in 300 training steps was chosen as a possible early stopping mechanism.
Learning curves

**Figure 3:** Learning curve of BLEU for DE→EN.

**Figure 4:** Learning curve of BLEU for FR→EN.

**Figure 5:** Learning curve of BLEU for CZ→EN.
Experiments
Experiments with Czech
Outline

Introduction

Experiments
- Experiments with Czech
- Experiments with Czech+German
- Experiments with Czech+German+French

Sentinel Attention Activation Ratio

Future Work
Figure 6: Attention energy distribution for CZ→EN.
Experiments
Experiments with Czech+German
Outline

Introduction

Experiments
  - Experiments with Czech
  - Experiments with Czech+German
  - Experiments with Czech+German+French

Sentinel Attention Activation Ratio

Future Work
Figure 7: Learning curve of BLEU for CZ+DE→EN.
Czech+German→English attention distribution

Figure 8: Attention energy distribution for CZ+DE→EN.
Experiments

Experiments with Czech+German+French
Introduction

Experiments
  - Experiments with Czech
  - Experiments with Czech+German
  - Experiments with Czech+German+French

Sentinel Attention Activation Ratio

Future Work
Figure 9: Learning curve of BLEU for CZ+DE+FR→EN.
Figure 10: Attention energy distribution for CZ+DE+FR→EN.
Sentinel Attention Activation Ratio
Outline

Introduction

Experiments
- Experiments with Czech
- Experiments with Czech+German
- Experiments with Czech+German+French

Sentinel Attention Activation Ratio

Future Work
A metric in the form of sentinel attention activation ratio (SAAR) was used to understand how much the decoder was relied upon by the model to make its final predictions.

For a particular sentence $S_i$, SAAR was calculated as:

$$S_i = \frac{A_s}{A_t}$$

where $A_s$ was the number of words whose prediction was based on the decoder during the entire training and $A_t$ represents the total count of attention units activated during training.

For each model, the corresponding SAAR for all sentences in the validation set was calculated followed by calculating their correlation with sentence length.
Sentinel Attention Activation Ratio

<table>
<thead>
<tr>
<th>cz_en</th>
<th>fr_en</th>
<th>de_en</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.393</td>
<td>0.010</td>
<td>-0.175</td>
</tr>
</tbody>
</table>

**Table 1:** Correlation between SAAR and sentence length of monolingual models.

<table>
<thead>
<tr>
<th>cz_de_en</th>
<th>cz_fr_en</th>
<th>de_fr_en</th>
<th>3_en</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.1145</td>
<td>-0.242</td>
<td>-0.362</td>
<td>-0.126</td>
</tr>
</tbody>
</table>

**Table 2:** Correlation between SAAR and sentence length of multilingual models.

The correlation values indicate that SAAR decreases with increasing sentence length.
Future Work
Outline

Introduction

Experiments
  Experiments with Czech
  Experiments with Czech+German
  Experiments with Czech+German+French

Sentinel Attention Activation Ratio

Future Work
• Eye-tracking study (October-November) to observe how human attention (in the form of eye movement) behaves during translation.

• Compare computational models of attention shift for translation with human attention patterns.

• Investigate the nature of 'representations' learnt by multilingual models.
Summary

1. Using a setup that employs the hierarchical attention combination mechanism can be useful for doing model analysis.
2. The model seems to pay greater importance to features from the source language when the target sentence is shorter.
3. The model exhibits a number of flips in how it spreads its attention throughout the training.


