Attention
- Encoder hidden states capture contextual source word identity

- Decoder hidden states are now mostly responsible for selecting what to attend to

- Doesn’t take a complex hidden state to walk monotonically through a sentence and spit out word-by-word translations
Neural MT
Results: WMT English-French

- 12M sentence pairs

Classic PBMT system: \(\sim 33\) BLEU, uses additional target-language data

  PBMT + rerank w/LSTMs: \(36.5\) BLEU (long line of work here; Devlin+ 2014)

Sutskever+ (2014) seq2seq single: \(30.6\) BLEU (input reversed)

Sutskever+ (2014) seq2seq ensemble: \(34.8\) BLEU

Luong+ (2015) seq2seq ensemble with attention and rare word handling: \(37.5\) BLEU

- But English-French is a really easy language pair and there’s tons of data for it! Does this approach work for anything harder?
Results: WMT English-German

- 4.5M sentence pairs

Classic phrase-based system: 20.7 BLEU

Luong+ (2014) seq2seq: 14 BLEU

Luong+ (2015) seq2seq ensemble with rare word handling: 23.0 BLEU

- Not nearly as good in absolute BLEU, but BLEU scores aren’t really comparable across languages

- French, Spanish = easiest
  German, Czech = harder
  Japanese, Russian = hard (grammatically different, lots of morphology...)
### MT Examples

<table>
<thead>
<tr>
<th>src</th>
<th>In einem Interview sagte Bloom jedoch, dass er und Kerr sich noch immer lieben.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ref</td>
<td>However, in an interview, Bloom has said that he and <em>Kerr</em> still love each other.</td>
</tr>
<tr>
<td>best</td>
<td>In an interview, however, Bloom said that he and <em>Kerr</em> still love.</td>
</tr>
<tr>
<td>base</td>
<td>However, in an interview, Bloom said that he and <em>Tina</em> were still <em>&lt;unk&gt;</em>.</td>
</tr>
</tbody>
</table>

- best = with attention, base = no attention
- NMT systems can hallucinate words, especially when not using attention — phrase-based doesn’t do this

Luong et al. (2015)
<table>
<thead>
<tr>
<th>.src</th>
<th>Wegen der von Berlin und der Europäischen Zentralbank verhängten strengen Sparpolitik in Verbindung mit der Zwangsjacke, in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Währung genötigt wird, sind viele Menschen der Ansicht, das Projekt Europa sei zu weit gegangen</th>
</tr>
</thead>
<tbody>
<tr>
<td>ref</td>
<td>The <em>austerity imposed by Berlin and the European Central Bank, coupled with the straitjacket</em> imposed on national economies through adherence to the common currency, has led many people to think Project Europe has gone too far.</td>
</tr>
<tr>
<td>best</td>
<td>Because of the strict <em>austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket</em> in which the respective national economy is forced to adhere to the common currency, many people believe that the European project has gone too far.</td>
</tr>
<tr>
<td>base</td>
<td>Because of the pressure <em>imposed by the European Central Bank and the Federal Central Bank with the strict austerity</em> imposed on the national economy in the face of the single currency, many people believe that the European project has gone too far.</td>
</tr>
</tbody>
</table>

- best = with attention, base = no attention

Luong et al. (2015)
Handling Rare Words

- Words are a difficult unit to work with: copying can be cumbersome, word vocabularies get very large
- Character-level models don’t work well
- Compromise solution: use thousands of “word pieces” (which may be full words but may also be parts of words)

  Input: _the _eco tax _port i co _in _Po nt - de - Bu is...

  Output: _le _port ique _éco taxe _de _Pont - de - Bui s

- Can achieve transliteration with this, subword structure makes some translations easier to achieve

Sennrich et al. (2016)
Byte Pair Encoding (BPE)

- Start with every individual byte (basically character) as its own symbol

```python
for i in range(num_merges):
    pairs = get_stats(vocab)
    best = max(pairs, key=pairs.get)
    vocab = merge_vocab(best, vocab)
```

- Count bigram character cooccurrences
- Merge the most frequent pair of adjacent characters

- Doing 8k merges => vocabulary of around 8000 word pieces. Includes many whole words

- Most SOTA NMT systems use this on both source + target

Sennrich et al. (2016)
### Byte Pair Encoding (BPE)

<table>
<thead>
<tr>
<th>Original:</th>
<th>furiously</th>
<th>Original:</th>
<th>tricycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) BPE:</td>
<td>_fur</td>
<td>iously</td>
<td>(b) BPE:</td>
</tr>
<tr>
<td>Unigram LM:</td>
<td>_fur</td>
<td>ious</td>
<td>ly</td>
</tr>
<tr>
<td>(c) BPE:</td>
<td>_Comple</td>
<td>t</td>
<td>ely</td>
</tr>
<tr>
<td>Unigram LM:</td>
<td>_Complete</td>
<td>ly</td>
<td>_pre</td>
</tr>
</tbody>
</table>

- BPE produces less linguistically plausible units than another technique based on a unigram language model

Bostrom and Durrett (2020)
Google’s NMT System

- 8-layer LSTM encoder-decoder with attention, word piece vocabulary of 8k-32k

Wu et al. (2016)
Google’s NMT System

English-French:
Google’s phrase-based system: 37.0 BLEU
Luong+ (2015) seq2seq ensemble with rare word handling: 37.5 BLEU
Google’s 32k word pieces: 38.95 BLEU

English-German:
Google’s phrase-based system: 20.7 BLEU
Luong+ (2015) seq2seq ensemble with rare word handling: 23.0 BLEU
Google’s 32k word pieces: 24.2 BLEU

Wu et al. (2016)
Similar to human-level performance on English-Spanish
# Google’s NMT System

<table>
<thead>
<tr>
<th>Source</th>
<th>She was spotted three days later by a dog walker trapped in the quarry</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBMT</td>
<td>Elle a été repéré trois jours plus tard par un promeneur de chien piégé dans la carrière</td>
</tr>
<tr>
<td>GNMT</td>
<td>Elle a été repérée trois jours plus tard par un traîneau à chiens piégé dans la carrière.</td>
</tr>
<tr>
<td>Human</td>
<td>Elle a été repérée trois jours plus tard par une personne qui promenait son chien coincée dans la carrière</td>
</tr>
</tbody>
</table>

Gender is correct in GNMT but not in PBMT

“sled”  “walker”

Wu et al. (2016)