Recall: Phrase-Based MT

Phrases:
- cat ||| chat ||| 0.9
- the cat ||| le chat ||| 0.8
- dog ||| chien ||| 0.8
- house ||| maison ||| 0.6
- my house ||| ma maison ||| 0.9
- language ||| langue ||| 0.9

Language model $P(e)$
Phrase table $P(f|e)$

Noisy channel model:
- combine scores from translation model +
- language model to translate foreign to
  - English

"Translate faithfully but make fluent English"

Recall: HMM for Alignment

Sequential dependence between a's to capture monotonicity

$$P(f, a|e) = \prod_{i=1}^{n} P(f_i|e_{ai})P(a_i|a_{i-1})$$

- e: Thank you, I shall do so gladly.
- a: 0 2 6 5 7 7 7 7 8
- f: Gracias, lo hare de muy buen grado.

Alignment dist parameterized by jump size: $P(a_j - a_{j-1})$

$P(f_i|e_{ai})$: word translation table

Brown et al. (1993)
Recall: Decoding

Recall: Decoding

<table>
<thead>
<tr>
<th>Mary</th>
<th>no</th>
<th>did</th>
<th>us</th>
<th>boktads</th>
<th>o</th>
<th>la</th>
<th>bruja</th>
<th>verds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>no</td>
<td>did</td>
<td>us</td>
<td>boktads</td>
<td>o</td>
<td>la</td>
<td>bruja</td>
<td>verds</td>
</tr>
<tr>
<td>Mary</td>
<td>no</td>
<td>did</td>
<td>us</td>
<td>boktads</td>
<td>o</td>
<td>la</td>
<td>bruja</td>
<td>verds</td>
</tr>
</tbody>
</table>

...did not give a slap to the green witch

...and TM is broken down into several features

4.2  -1.2  -2.9

\[
\text{score} = \log \left[ P(\text{Mary}) \cdot P(\text{not | Mary}) \cdot P(\text{Mary | Maria}) \cdot P(\text{not | no}) \right]
\]

LM  TM

In reality: score = \(\alpha\) log P(LM) + \(\beta\) log P(TM)

This Lecture

- Syntactic MT
- Neural MT details
- Dilated CNNs for MT
- Transformers for MT

Syntactic MT

Levels of Transfer: Vauquois Triangle

- Is syntax a “better” abstraction than phrases?

Slide credit: Dan Klein
**Syntactic MT**

- Rather than use phrases, use a **synchronous context-free grammar**: constructs “parallel” trees in two languages simultaneously

  \[
  \begin{align*}
  \text{NP} & \rightarrow [\text{DT}_1 \text{JJ}_2 \text{NN}_3; \text{DT}_1 \text{NN}_3 \text{JJ}_2] \\
  \text{DT} & \rightarrow [\text{the, } \text{la}] \\
  \text{NN} & \rightarrow [\text{car, } \text{voiture}] \\
  \text{JJ} & \rightarrow [\text{yellow, } \text{jaune}] \\
  \end{align*}
  \]

- Assumes parallel syntax up to reordering
- Translation = parse the input with “half” the grammar, read off other half

**Encoder-Decoder MT**

- Sutskever seq2seq paper: first major application of LSTMs to NLP
- Basic encoder-decoder with beam search

<table>
<thead>
<tr>
<th>Method</th>
<th>test BLEU score (ntst14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bahdanau et al. [2]</td>
<td>28.45</td>
</tr>
<tr>
<td>Baseline System [29]</td>
<td>33.30</td>
</tr>
<tr>
<td>Single forward LSTM, beam size 12</td>
<td>26.17</td>
</tr>
<tr>
<td>Single reversed LSTM, beam size 12</td>
<td>30.59</td>
</tr>
<tr>
<td>Ensemble of 5 reversed LSTMs, beam size 12</td>
<td><strong>34.81</strong></td>
</tr>
</tbody>
</table>

**SOTA = 37.0 — not all that competitive...**

**Slide credit:** Dan Klein
Encoder-Decoder MT

- Better model from seq2seq lectures: encoder-decoder with attention and copying for rare words
- Distribution over vocab + copying

Results: WMT English-French

- 12M sentence pairs
- Classic phrase-based system: ~33 BLEU, uses additional target-language data
- Rerank with LSTMs: 36.5 BLEU (long line of work here; Devlin+ 2014)
- Sutskever+ (2014) seq2seq single: 30.6 BLEU
- 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 not really comparable across languages
- French, Spanish = easiest
- German, Czech = harder
- Japanese, Russian = hard (grammatically different, lots of morphology…)

MT Examples

| src | In einem Interview sagte Bloom jedoch, dass er und Kerr sich noch immer lieben. |
| ref | However, in an interview, Bloom has said that he and Kerr still love each other. |
| best | In an interview, however, Bloom said that he and Kerr still love. |
| base | However, in an interview, Bloom said that he and Tina were still <unk>. |

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)
### MT Examples

**src**

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.

**ref**

The austerity imposed by Berlin and the European Central Bank, coupled with the straitjacket imposed on national economies through adherence to the common currency, has led many people to think Project Europe has gone too far.

**best**

Because of the strict austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket 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.

- best = with attention, base = no attention

Luong et al. (2015)

**MT Examples**

<table>
<thead>
<tr>
<th>Source</th>
<th>Such changes in reaction conditions include, but are not limited to, an increase in temperature or change in pH.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>The (such) of (reaction) condition of (of) change (include) but (but) not (not) limited (limit)....</td>
</tr>
<tr>
<td>PBM</td>
<td>随着(such) 反应(reaction) 条件(condition) 的(of) 改变(change) 包括(include) 但(but) 不(not) 限于(limit)....</td>
</tr>
<tr>
<td>NMT</td>
<td>这种(such) 反应(reaction) 条件(condition) 的(of) 改变(change) 包括(include) 但(but) 不(not) 限于(limit) pH或or pH的(of)变化(change)....</td>
</tr>
</tbody>
</table>

- NMT can repeat itself if it gets confused (pH or pH)
- Phrase-based MT often gets chunks right, may have more subtle ungrammaticalities

Zhang et al. (2017)

### Rare Words: Word Piece Models

- Use Huffman encoding on a corpus, keep most common $k$ (~10,000) character sequences for source and target

  **Input:** _the _eco tax _port i co in _Po nt - de - Bu i s...
  
  **Output:** le _port ique _éco taxe _de _Pon t - de - Bui s

- Captures common words and parts of rare words
- Subword structure may make it easier to translate
- Model balances translating and transliterating without explicit switching

Wu et al. (2016)

### Rare Words: Byte Pair Encoding

- Simpler procedure, based only on the dictionary
- Input: a dictionary of words represented as characters

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

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

- Final size = initial vocab + num merges. Often do 10k - 30k merges
- Most SOTA NMT systems use this on both source + target

Sennrich et al. (2016)
Google’s NMT System

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

Wu et al. (2016)

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)

Human Evaluation (En-Es)

- Similar to human-level performance on English-Spanish

Wu et al. (2016)

Gender is correct in GNMT but not PBMT

“sled” “walker”

Wu et al. (2016)
Backtranslaction

- Classical MT methods used a bilingual corpus of sentences $B = (S, T)$ and a large monolingual corpus $T'$ to train a language model. Can neural MT do the same?
- Approach 1: force the system to generate $T'$ as targets from null inputs
  - $s_1, t_1$
  - $s_2, t_2$
  - ...$
  - [null], $t_1$
  - [null], $t_2$
  - ...

- Approach 2: generate synthetic sources with a T->S machine translation system (backtranslation)
  - $s_1, t_1$
  - $s_2, t_2$
  - ...
  - MT($t_1'), t_1'$
  - MT($t_2'), t_2'$
  - ...

Sennrich et al. (2015)

<table>
<thead>
<tr>
<th>name</th>
<th>training data</th>
<th>instances</th>
<th>BLEU 2011</th>
<th>BLEU 2012</th>
<th>BLEU 2013</th>
<th>BLEU 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline (Gulcehre et al., 2015)</td>
<td>parallel</td>
<td>7.2m</td>
<td>18.4</td>
<td>18.8</td>
<td>19.9</td>
<td>18.7</td>
</tr>
<tr>
<td>deep fusion (Gulcehre et al., 2015)</td>
<td>parallel/parallel</td>
<td>6m/6m</td>
<td>19.9</td>
<td>20.4</td>
<td>20.1</td>
<td>20.0</td>
</tr>
<tr>
<td>baseline</td>
<td>parallel</td>
<td>7.6m/7.6m</td>
<td>19.9</td>
<td>19.6</td>
<td>19.4</td>
<td>18.2</td>
</tr>
<tr>
<td>Gigawordmono</td>
<td>parallel/Gigawordmono</td>
<td>8.4m/8.4m</td>
<td>21.2</td>
<td>21.1</td>
<td>21.8</td>
<td>20.4</td>
</tr>
</tbody>
</table>
CNNs for Machine Translation

- “ByteNet”: operates over characters (bytes)
- Encode source sequence w/dilated convolutions
- Predict nth target character by looking at the nth position in the source and a dilated convolution over the n-1 target tokens so far
- To deal with divergent lengths, tn actually looks at sna where α is a heuristically-chosen parameter
- Assumes mostly monotonic translation

Compare: CNNs vs. LSTMs

- LSTM: looks at previous word + hidden state, attention over input
- CNN: source encoding at this position gives us “attention”, target encoding gives us decoder context

Attention from CNN

- Model is character-level, this visualization shows which words’s characters impact the convolutional encoding the most
- Largely monotonic but does consult other information

Advantages of CNNs

- LSTM with attention is quadratic: compute attention over the whole input for each decoded token
- CNN is linear!
- CNN is shallower too in principle but the conv layers are very sophisticated (3 layers each)
### English-German MT Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Inputs</th>
<th>Outputs</th>
<th>WMT Test ’14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase Based MT (Freitag et al., 2014; Williams et al., 2015)</td>
<td>phrases</td>
<td>phrases</td>
<td>20.7</td>
</tr>
<tr>
<td>RNN Enc-Dec (Luong et al., 2015)</td>
<td>words</td>
<td>words</td>
<td>11.3</td>
</tr>
<tr>
<td>Reverse RNN Enc-Dec (Luong et al., 2015)</td>
<td>words</td>
<td>words</td>
<td>14.0</td>
</tr>
<tr>
<td>RNN Enc-Dec Att (Zhou et al., 2016)</td>
<td>words</td>
<td>words</td>
<td>20.6</td>
</tr>
<tr>
<td>RNN Enc-Dec Att (Luong et al., 2015)</td>
<td>words</td>
<td>words</td>
<td>20.0</td>
</tr>
<tr>
<td>GNMT (RNN Enc-Dec Att) (Wu et al., 2016a)</td>
<td>word-pieces</td>
<td>word-pieces</td>
<td>24.61</td>
</tr>
<tr>
<td>RNN Enc-Dec Att (Chung et al., 2016b)</td>
<td>BPE</td>
<td>BPE</td>
<td>19.98</td>
</tr>
<tr>
<td>RNN Enc-Dec Att (Chung et al., 2016b)</td>
<td>BPE</td>
<td>char</td>
<td>21.33</td>
</tr>
<tr>
<td>GNMT (RNN Enc-Dec Att) (Wu et al., 2016a)</td>
<td>char</td>
<td>char</td>
<td>23.62</td>
</tr>
<tr>
<td>ByteNet</td>
<td>char</td>
<td>char</td>
<td>25.75</td>
</tr>
</tbody>
</table>

Kalchbrenner et al. (2016)

### Transformers for MT

#### Self-Attention

- Each word forms a “query” which then computes attention over each word
  \[
  \alpha_{i,j} = \text{softmax}(x_i^T x_j) \quad \text{scalar}
  \]
  \[
  x'_i = \sum_{j=1}^n \alpha_{i,j} x_j \quad \text{vector} = \text{sum of scalar} \times \text{vector}
  \]
- Multiple “heads” analogous to different convolutional filters. Use parameters \( W_k \) and \( V_k \) to get different attention values + transform vectors
  \[
  \alpha_{k,i,j} = \text{softmax}(x_i^T W_k x_j) \quad x'_{k,i} = \sum_{j=1}^n \alpha_{k,i,j} V_k x_j
  \]

Vaswani et al. (2017)

#### Transformers

- Positional encoding: augment word embedding with position embeddings, each dim is a sine wave of a different frequency. Closer points = higher dot products

Vaswani et al. (2017)
Transformers

- Encoder and decoder are both transformers
- Decoder consumes the previous generated token (and attends to input), but has no recurrent state

Transformers

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>ByteNet [18]</td>
<td>23.75</td>
</tr>
<tr>
<td>GNMT + RL [38]</td>
<td>24.6</td>
</tr>
<tr>
<td>ConvS2S [9]</td>
<td>39.92</td>
</tr>
<tr>
<td>MoE [32]</td>
<td>25.16</td>
</tr>
<tr>
<td>GNMT + RL Ensemble [38]</td>
<td>40.4</td>
</tr>
<tr>
<td>ConvS2S Ensemble [9]</td>
<td>40.56</td>
</tr>
<tr>
<td>Transformer (base model)</td>
<td>26.30</td>
</tr>
<tr>
<td>Transformer (big)</td>
<td>41.16</td>
</tr>
<tr>
<td><strong>Big</strong></td>
<td>41.29</td>
</tr>
<tr>
<td><strong>base</strong></td>
<td>28.4</td>
</tr>
</tbody>
</table>

Big = 6 layers, 1000 dim for each token, 16 heads, base = 6 layers + other params halved

Visualization
Takeaways

- Can build MT systems with LSTM encoder-decoders, CNNs, or transformers
- Word piece / byte pair models are really effective and easy to use
- State of the art systems are getting pretty good, but lots of challenges remain, especially for low-resource settings