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: ~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...)

NMT systems can hallucinate words, especially when not using attention — phrase-based doesn’t do this

Luong et al. (2015)
## MT Examples

<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</em>, 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.</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)
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
Similar to human-level performance on *English-Spanish*
<table>
<thead>
<tr>
<th>Source</th>
<th>Translation</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>She was spotted three days later by a dog walker trapped in the quarry</td>
<td></td>
</tr>
<tr>
<td>PBMT</td>
<td>Elle a été repéré trois jours plus tard par un promeneur de chien piégé dans la carrière</td>
<td>6.0</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>
<td>2.0</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>
<td>5.0</td>
</tr>
</tbody>
</table>

Gender is correct in GNMT but not in PBMT

“sled” “walker”

Wu et al. (2016)
QA Intro
Combinatory Categorial Grammar
Combinatory Categorial Grammar

- Steedman+Szabolcsi (1980s): formalism bridging syntax and semantics
- Parallel derivations of syntactic parse and lambda calculus expression
- Syntactic categories (for this lecture): S, NP, “slash” categories
- S\NP: “if I combine with an NP on my left side, I form a sentence” — verb
- When you apply this, there has to be a parallel instance of function application on the semantics side

\[
\lambda y. \text{sings}(y)
\]

Eminem \text{sings}
Combinatory Categorial Grammar

- Steedman+Szabolcsi 1980s: formalism bridging syntax and semantics
- Syntactic categories (for this lecture): S, NP, “slash” categories
  - S\NP: “if I combine with an NP on my left side, I form a sentence” — verb
  - (S\NP)/NP: “I need an NP on my right and then on my left” — verb with a direct object


<table>
<thead>
<tr>
<th>What</th>
<th>states</th>
<th>border</th>
<th>Texas</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\frac{(S/((S\backslash NP))/N}{\lambda f.\lambda g.\lambda x.f(x) \land g(x)})</td>
<td>(N)</td>
<td>(\frac{(S\backslash NP)/NP}{\lambda x.\lambda y.borders(y, x)})</td>
<td>(NP)</td>
</tr>
<tr>
<td>(\frac{(S\backslash NP)}{\lambda x.\lambda y.borders(y, x)})</td>
<td>(\lambda x.\lambda y.borders(y, x))</td>
<td>(\lambda y.borders(y, texas))</td>
<td></td>
</tr>
</tbody>
</table>

- “What” is a very complex type: needs a noun and needs a S\NP to form a sentence. S\NP is basically a verb phrase (\textit{border Texas})
“What” is a very complex type: needs a noun and needs a S\NP to form a sentence. S\NP is basically a verb phrase (border Texas).

What in this case knows that there are two predicates (states and border Texas). This is not a general thing. Zettlemoyer and Collins (2005)
CCG Parsing

- These question are *compositional*: we can build bigger ones out of smaller pieces

  \[
  \text{What states border Texas?}
  \]

  \[
  \text{What states border states bordering Texas?}
  \]

  \[
  \text{What states border states bordering states bordering Texas?}
  \]

- In general, answering this does require parsing and not just slot-filling
Training CCG Parsers

- Training data looks like pairs of sentences and logical forms

\[
\begin{align*}
\text{What states border Texas} & \quad \lambda x. \text{state}(x) \land \text{borders}(x, e89) \\
\text{What borders Texas} & \quad \lambda x. \text{borders}(x, e89)
\end{align*}
\]

- Unlike PCFGs, we don’t know which words yielded which fragments of CCG

- Requires an “unsupervised” approach like Model 1 for word alignment

Zettlemoyer and Collins (2005)
Seq2seq Semantic Parsing
Semantic Parsing as Translation

“what states border Texas”

\[
\text{lambda } x ( \text{state}(x) \text{ and border}(x, e89) )
\]

- Write down a linearized form of the semantic parse, train seq2seq models to directly translate into this representation

- What are some benefits of this approach compared to grammar-based?

- What might be some concerns about this approach? How do we mitigate them?

Jia and Liang (2016)
Handling Invariances

“what states border Texas”   “what states border Ohio”

- Parsing-based approaches handle these the same way
  - Possible divergences: features, different weights in the lexicon
- Can we get seq2seq semantic parsers to handle these the same way?
- Key idea: do data augmentation by synthetically creating more data from a single example
Semantic Parsing as Translation

- Prolog
- Lambda calculus
- Other DSLs

- Handle all of these with uniform machinery!

Jia and Liang (2016)
### Semantic Parsing as Translation

<table>
<thead>
<tr>
<th>Previous Work</th>
<th>GEO</th>
<th>ATIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zettlemoyer and Collins (2007)</td>
<td>88.9</td>
<td>84.6</td>
</tr>
<tr>
<td>Kwiatkowski et al. (2010)</td>
<td>91.1</td>
<td></td>
</tr>
<tr>
<td>Liang et al. (2011)</td>
<td>88.6</td>
<td>82.8</td>
</tr>
<tr>
<td>Kwiatkowski et al. (2011)</td>
<td></td>
<td>83.5</td>
</tr>
<tr>
<td>Poon (2013)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhao and Huang (2015)</td>
<td>88.9</td>
<td>84.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Our Model</th>
<th>GEO</th>
<th>ATIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Recombination</td>
<td>85.0</td>
<td>76.3</td>
</tr>
<tr>
<td>ABS ENTITIES</td>
<td>85.4</td>
<td>79.9</td>
</tr>
<tr>
<td>ABS WHOLE PHRASES</td>
<td>87.5</td>
<td></td>
</tr>
<tr>
<td>CONCAT-2</td>
<td>84.6</td>
<td>79.0</td>
</tr>
<tr>
<td>CONCAT-3</td>
<td></td>
<td>77.5</td>
</tr>
<tr>
<td>AWP + AE</td>
<td>88.9</td>
<td></td>
</tr>
<tr>
<td>AE + C2</td>
<td></td>
<td>78.8</td>
</tr>
<tr>
<td>AWP + AE + C2</td>
<td>89.3</td>
<td></td>
</tr>
<tr>
<td>AE + C3</td>
<td></td>
<td>83.3</td>
</tr>
</tbody>
</table>

- Three forms of data augmentation all help
- Results on these tasks are still not as strong as hand-tuned systems from 10 years ago, but the same simple model can do well at all problems

Jia and Liang (2016)
Applications

- GeoQuery (Zelle and Mooney, 1996): answering questions about states (~80% accuracy)
- Jobs: answering questions about job postings (~80% accuracy)
- ATIS: flight search
- Can do well on all of these tasks if you handcraft systems and use plenty of training data: these domains aren’t that rich
Next Time

- QA from raw text: how do we answer a question about a passage?
- Neural networks for QA
- Final project discussion