Decoding
Phrase-Based Decoding

- **Inputs:**
  - Language model that scores $P(e_i|e_1, \ldots, e_{i-1}) \approx P(e_i|e_{i-n-1}, \ldots, e_{i-1})$
  - Phrase table: set of phrase pairs $(e, f)$ with probabilities $P(f|e)$
  - What we want to find: $e$ produced by a series of phrase-by-phrase translations from an input $f$
Phrase lattices are big!

7 people including by some the Russian the astronauts.
7 people included by France and the Russian international astronomical rapporteur.

7 out including the French and the Russian the fifth.
7 among including from the French and of the Russian of space members.
7 persons including from the French and to Russian of the Aerospace members.
7 include from the French and Russian astronauts.
7 numbers include from France and Russian of astronauts who.
7 populations include those from France and Russian astronauts.
7 deportees included come from France and Russia in Astronautical personnel.
7 phyltrum including those from France and Russia a space Astronaut member.
include came from France and Russia by cosmonauts.
include came from French and Russia's cosmonauts.
includes coming from French and Russia's cosmonaut astronavigation member.
French and Russia's special rapporteur.
French and Russia's rapporteur.
French and Russia's rapporteur.
### Monotonic Translation

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio</th>
<th>una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
</table>

- Mary
- not
give
a
slap
to
the
witch
green
did not
a
slap
by
green
witch
nc
slap
to
the
did not
give
to
the

**If we translate with beam search, what state do we need to keep in the beam?**

- What have we translated so far?
  $$\arg\max_e \left[ \prod P(\tilde{f}|\tilde{e}) \cdot \prod_{i=1}^{\lvert e \rvert} P(e_i|e_{i-1}, e_{i-2}) \right]$$

- What words have we produced so far?

- When using a 3-gram LM, only need to remember the last 2 words!
Monotonic Translation

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio</th>
<th>una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>witch</td>
<td>green</td>
</tr>
<tr>
<td>did not</td>
<td></td>
<td>a slap</td>
<td>by</td>
<td>green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>slap</td>
<td>to the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did not give</td>
<td></td>
<td></td>
<td>to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>slap</td>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>the witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...did not

idx = 2

<table>
<thead>
<tr>
<th>score</th>
<th>4.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td></td>
</tr>
<tr>
<td>TM</td>
<td></td>
</tr>
</tbody>
</table>

...and TM is broken down into several features

In reality: score = α log P(LM) + β log P(TM)
Monotonic Translation

Several paths can get us to this state, max over them (like Viterbi)

Variable-length translation pieces = semi-HMM
Non-Monotonic Translation

- Non-monotonic translation: can visit source sentence “out of order”
- State needs to describe which words have been translated and which haven’t
- Big enough phrases already capture lots of reorderings, so this isn’t as important as you think

Maria    no    dio    una    bofetada    a    la    bruja    verde

Mary     not    give    a    slap    to    the    witch    green

Did not    gave    a    slap    by    the    green    witch

No    did not    give    slap    to    the    witch

Big enough phrases already capture lots of reorderings, so this isn’t as important as you think
Training Decoders

score = α log P(t) + β log P(s|t)

...and P(s|t) is in fact more complex

- Usually 5-20 feature weights to set, want to optimize for BLEU score which is not differentiable

- MERT (Och 2003): decode to get 1000-best translations for each sentence in a small training set (<1000 sentences), do line search on parameters to directly optimize for BLEU
Moses

- Toolkit for machine translation due to Philipp Koehn + Hieu Hoang
  - Pharaoh (Koehn, 2004) is the decoder from Koehn’s thesis
- Moses implements word alignment, language models, and this decoder, plus *a ton* more stuff
  - Highly optimized and heavily engineered, could more or less build SOTA translation systems with this from 2007-2013
<table>
<thead>
<tr>
<th>SOURCE</th>
<th>Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUMAN</td>
<td>That would be an interim solution which would make it possible to work towards a binding charter in the long term.</td>
</tr>
<tr>
<td>1x DATA</td>
<td>[this] [constituerait] [assistance] [transitoire] [who] [permettrait] [licences] [to] [terme] [to] [a] [charter] [to] [value] [contraignante] [.]</td>
</tr>
<tr>
<td>10x DATA</td>
<td>[it] [would] [a solution] [transitional] [which] [would] [of] [lead] [to] [term] [to a] [charter] [to] [value] [binding] [.]</td>
</tr>
<tr>
<td>100x DATA</td>
<td>[this] [would be] [a transitional solution] [which would] [lead to] [a charter] [legally binding] [.]</td>
</tr>
<tr>
<td>1000x DATA</td>
<td>[that would be] [a transitional solution] [which would] [eventually lead to] [a binding charter] [.]</td>
</tr>
</tbody>
</table>
Rather than use phrases, use a synchronous context-free grammar

NP → [DT₁ JJ₂ NN₃; DT₁ NN₃ JJ₂]
DT → [the, la]
DT → [the, le]
NN → [car, voiture]
JJ → [yellow, jaune]

Translation = parse the input with “half” of the grammar, read off the other half
Assumes parallel syntax up to reordering
Use lexicalized rules, look like “syntactic phrases”

Leads to HUGE grammars, parsing is slow

**Grammar**

\[
S \rightarrow \langle \text{ VP } \; ; \; \text{ I VP } \rangle \quad \text{or} \quad S \rightarrow \langle \text{ VP } \; ; \; \text{ you VP } \rangle \\
\text{ VP } \rightarrow \langle \text{ lo haré ADV } \; ; \; \text{ will do it ADV } \rangle \\
S \rightarrow \langle \text{ lo haré ADV } \; ; \; \text{ I will do it ADV } \rangle \\
\text{ ADV } \rightarrow \langle \text{ de muy buen grado } \; ; \; \text{ gladly } \rangle
\]

Slide credit: Dan Klein