Machine Translation
Decoding

Philipp Koehn, University of Edinburgh

9 February 2009
Statistical Machine Translation

- Components: Translation model, language model, decoder
Phrase-Based Translation

- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered

Morgen fliege ich nach Kanada zur Konferenz
Tomorrow I will fly to the conference in Canada
Phrase Translation Table

- Phrase Translations for “den Vorschlag”:

| English            | $\phi(e|f)$ | English        | $\phi(e|f)$ |
|--------------------|------------|----------------|------------|
| the proposal       | 0.6227     | the suggestions| 0.0114     |
| ’s proposal        | 0.1068     | the proposed   | 0.0114     |
| a proposal         | 0.0341     | the motion     | 0.0091     |
| the idea           | 0.0250     | the idea of    | 0.0091     |
| this proposal      | 0.0227     | the proposal , | 0.0068     |
| proposal           | 0.0205     | its proposal   | 0.0068     |
| of the proposal    | 0.0159     | it             | 0.0068     |
| the proposals      | 0.0159     | ...            | ...        |
Decoding Process

- Build translation left to right
  - *select foreign* words to be translated
Decoding Process

<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>

- Build translation *left to right*
  - select foreign words to be translated
  - *find English* phrase translation
  - *add English* phrase to end of partial translation
## Decoding Process

- Build translation left to right
  - select foreign words to be translated
  - find English phrase translation
  - add English phrase to end of partial translation
  - *mark foreign* words as translated

<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
### Decoding Process

<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>

- *One to many* translation
Decoding Process

- Many to one translation
Decoding Process

- *Many to one* translation
Decoding Process

- Reordering

Maria no dio una bofetada a la bruja verde

Mary did not slap the green
Decoding Process

- Translation *finished*
### Translation Options

<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></td>
<td></td>
<td></td>
<td>green witch</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>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>slap</td>
<td></td>
<td></td>
<td>the witch</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Look up *possible phrase translations*
  - many different ways to *segment* words into phrases
  - many different ways to *translate* each phrase
Hypothesis Expansion

<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>a slap</td>
<td>by</td>
<td>green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did not give</td>
<td>to</td>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slap</td>
<td>the</td>
<td>witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Start with **empty hypothesis**
  - e: no English words
  - f: no foreign words covered
  - p: probability 1
### Hypothesis Expansion

<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 give</td>
<td>a slap</td>
<td>to</td>
<td>the</td>
<td>witch</td>
<td>green</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>did not</td>
<td>a slap</td>
<td>by</td>
<td>green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>did not give</td>
<td>to</td>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>slap</td>
<td>the witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- e: Mary
- f: *--------
- p: .534

- Pick **translation option**
- Create **hypothesis**
  - e: add English phrase Mary
  - f: first foreign word covered
  - p: probability 0.534
A Quick Word on Probabilities

• Not going into detail here, but...

• Translation Model
  – phrase translation probability $p(\text{Mary} | \text{Maria})$
  – reordering costs
  – phrase/word count costs
  – ...

• Language Model
  – uses trigrams:
  – $p(\text{Mary did not}) =
    p(\text{Mary} | \text{START}) \times p(\text{did} | \text{Mary,START}) \times p(\text{not} | \text{Mary did})$
Hypothesis Expansion

<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 did not give a slap to the green witch.

- **e:** witch  
  - **f:** -------  
  - **p:** .182

- **e:** Mary  
  - **f:** *--------  
  - **p:** .534

- **e:**  
  - **f:**  
  - **p:** 1

- **e:**  
  - **f:**  
  - **p:** 1

- **e:**  
  - **f:**  
  - **p:** 1

- **e:**  
  - **f:**  
  - **p:** 1

- **e:**  
  - **f:**  
  - **p:** 1

• Add another hypothesis
Hypothesis Expansion

- Further *hypothesis expansion*
Hypothesis Expansion

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

Mary did not give a slap to the green witch.

e: witch 
f: -------*- p: .182
e: slap 
f: *---*---- p: .043
e: Mary 
f: *----*---- p: .534
e: did not 
f: **----- p: .154
e: slap 
f: *****---- p: .015
e: the 
f: *******- p: .004283
e: green witch 
f: ********* p: .000271

- ... until all foreign words *covered*
  - find *best hypothesis* that covers all foreign words
  - *backtrack* to read off translation
Hypothesis Expansion

<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><em>Mary</em></td>
<td><em>not</em></td>
<td><em>give</em></td>
<td><em>a</em></td>
<td><em>slap</em></td>
<td><em>to</em></td>
<td><em>the</em></td>
<td><em>witch</em></td>
<td><em>green</em></td>
</tr>
<tr>
<td><em>did not</em></td>
<td><em>give</em></td>
<td><em>a slap</em></td>
<td><em>by</em></td>
<td><em>green witch</em></td>
<td><em>to</em></td>
<td><em>the</em></td>
<td><em>to</em></td>
<td><em>the</em></td>
</tr>
<tr>
<td><em>no</em></td>
<td><em>slap</em></td>
<td><em>by</em></td>
<td><em>green witch</em></td>
<td><em>to</em></td>
<td><em>the</em></td>
<td><em>to</em></td>
<td><em>the</em></td>
<td></td>
</tr>
</tbody>
</table>

- Adding more hypothesis

⇒ *Explosion* of search space
Explosion of Search Space

- Number of hypotheses is *exponential* with respect to sentence length

⇒ Decoding is NP-complete [Knight, 1999]

⇒ Need to *reduce search space*
  - risk free: hypothesis *recombination*
  - risky: *histogram/threshold pruning*
Hypothesis Recombination

- Different paths to the \textit{same} partial translation
Hypothesis Recombination

- Different paths to the same partial translation

⇒ Combine paths
  - drop weaker path
  - keep pointer from weaker path (for lattice generation)
Hypothesis Recombination

- Recombined hypotheses do *not* have to match completely.
- No matter what is added, weaker path can be dropped, if:
  - *last two English words* match (matters for language model)
  - *foreign word coverage* vectors match (effects future path)
Hypothesis Recombination

- Recombined hypotheses do not have to match completely
- No matter what is added, weaker path can be dropped, if:
  - last two English words match (matters for language model)
  - foreign word coverage vectors match (effects future path)

⇒ *Combine paths*
Pruning

• Hypothesis recombination is not sufficient

⇒ Heuristically discard weak hypotheses early

• Organize Hypothesis in stacks, e.g. by
  – same foreign words covered
  – same number of foreign words covered
  – same number of English words produced

• Compare hypotheses in stacks, discard bad ones
  – histogram pruning: keep top \( n \) hypotheses in each stack (e.g., \( n=100 \))
  – threshold pruning: keep hypotheses that are at most \( \alpha \) times the cost of best hypothesis in stack (e.g., \( \alpha = 0.001 \))
Hypothesis Stacks

- Organization of hypothesis into stacks
  - here: based on *number of foreign words* translated
  - during translation all hypotheses from one stack are expanded
  - expanded Hypotheses are placed into stacks
Comparing Hypotheses

• Comparing hypotheses with *same number of foreign words* covered

Maria no     dio una bofetada     a la     bruja verde

- e: Mary did not
- f: **-------
- p: 0.154

- e: the
- f: ------***
- p: 0.354

\textbf{better}
\textit{partial}
\textbf{translation}

\textbf{covers}
\textbf{easier part}
\textbf{--> lower cost}

• Hypothesis that covers *easy part* of sentence is preferred

⇒ Need to consider \textbf{future cost} of uncovered parts
Future Cost Estimation

- *Estimate cost* to translate remaining part of input

- Step 1: estimate future cost for each *translation option*
  - look up translation model cost
  - estimate language model cost (no prior context)
  - ignore reordering model cost
  \[ LM \times TM = p(\text{to}) \times p(\text{the}|\text{to}) \times p(\text{to the}|\text{a la}) \]
Future Cost Estimation: Step 2

- Step 2: find *cheapest cost* among translation options
Future Cost Estimation: Step 3

- Step 3: find *cheapest future cost path* for each span
  - can be done *efficiently* by dynamic programming
  - future cost for every span can be *pre-computed*
Future Cost Estimation: Application

- Use future cost estimates when *pruning* hypotheses

- For each *uncovered contiguous span*:
  - look up *future costs* for each maximal contiguous uncovered span
  - *add* to actually accumulated cost for translation option for pruning
A* search

- Pruning might drop hypothesis that lead to the best path (search error)

- **A* search**: safe pruning
  - future cost estimates have to be accurate or underestimates
  - **lower bound** for probability is established early by
    depth first search: compute cost for one complete translation
  - if cost-so-far and future cost are worse than lower bound, hypothesis can be safely discarded

- Not commonly done, since not aggressive enough
Limits on Reordering

- Reordering may be **limited**
  - **Monotone** Translation: No reordering at all
  - Only phrase movements of at most \( n \) words

- Reordering limits *speed* up search (polynomial instead of exponential)

- Current reordering models are weak, so limits *improve* translation quality
**Word Lattice Generation**

- **Search graph** can be easily converted into a **word lattice**
  - can be further mined for **n-best lists**
  - enables **reranking** approaches
  - enables **discriminative training**
Sample N-Best List

- Simple **N-best list**:

<table>
<thead>
<tr>
<th>Translation</th>
<th>Reordering LM TM</th>
<th>WordPenalty</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>this is a small house</td>
<td>0 -27.0908 -1.83258 -5</td>
<td>-28.9234</td>
<td></td>
</tr>
<tr>
<td>this is a little house</td>
<td>0 -28.1791 -1.83258 -5</td>
<td>-30.0117</td>
<td></td>
</tr>
<tr>
<td>it is a small house</td>
<td>0 -27.108 -3.21888 -5</td>
<td>-30.3268</td>
<td></td>
</tr>
<tr>
<td>it is a little house</td>
<td>0 -28.163 -3.21888 -5</td>
<td>-31.4152</td>
<td></td>
</tr>
<tr>
<td>this is an small house</td>
<td>0 -31.7294 -1.83258 -5</td>
<td>-33.562</td>
<td></td>
</tr>
<tr>
<td>it is an small house</td>
<td>0 -32.3094 -3.21888 -5</td>
<td>-35.5283</td>
<td></td>
</tr>
<tr>
<td>this is an little house</td>
<td>0 -33.7639 -1.83258 -5</td>
<td>-35.5965</td>
<td></td>
</tr>
<tr>
<td>this is a house small</td>
<td>-3 -31.4851 -1.83258 -5</td>
<td>-36.3176</td>
<td></td>
</tr>
<tr>
<td>this is a house little</td>
<td>-3 -31.5689 -1.83258 -5</td>
<td>-36.4015</td>
<td></td>
</tr>
<tr>
<td>it is an little house</td>
<td>0 -34.3439 -3.21888 -5</td>
<td>-37.5628</td>
<td></td>
</tr>
<tr>
<td>it is a house small</td>
<td>-3 -31.5022 -3.21888 -5</td>
<td>-37.7211</td>
<td></td>
</tr>
<tr>
<td>this is an house small</td>
<td>-3 -32.8999 -1.83258 -5</td>
<td>-37.7325</td>
<td></td>
</tr>
<tr>
<td>it is a house little</td>
<td>-3 -31.586 -3.21888 -5</td>
<td>-37.8049</td>
<td></td>
</tr>
<tr>
<td>this is an house little</td>
<td>-3 -32.9837 -1.83258 -5</td>
<td>-37.8163</td>
<td></td>
</tr>
<tr>
<td>the house is a little</td>
<td>-7 -28.5107 -2.52573 -5</td>
<td>-38.0364</td>
<td></td>
</tr>
<tr>
<td>the is a small house</td>
<td>0 -35.6899 -2.52573 -5</td>
<td>-38.2156</td>
<td></td>
</tr>
<tr>
<td>is it a little house</td>
<td>-4 -30.3603 -3.91202 -5</td>
<td>-38.2723</td>
<td></td>
</tr>
<tr>
<td>the house is a small</td>
<td>-7 -28.7683 -2.52573 -5</td>
<td>-38.294</td>
<td></td>
</tr>
<tr>
<td>it ’s a small house</td>
<td>0 -34.8557 -3.91202 -5</td>
<td>-38.7677</td>
<td></td>
</tr>
<tr>
<td>this house is a little</td>
<td>-7 -28.0443 -3.91202 -5</td>
<td>-38.9563</td>
<td></td>
</tr>
<tr>
<td>it ’s a little house</td>
<td>0 -35.1446 -3.91202 -5</td>
<td>-39.0566</td>
<td></td>
</tr>
<tr>
<td>this house is a small</td>
<td>-7 -28.3018 -3.91202 -5</td>
<td>-39.2139</td>
<td></td>
</tr>
</tbody>
</table>