Statistical Machine Translation

- Components: Translation model, language model, decoder

- foreign/English parallel text
- statistical analysis
- Translation Model
- Decoding Algorithm

- English text
- statistical analysis
- Language Model
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

Phrase Translation Table

- Phrase Translations for “den Vorschlag”:

| English          | φ(e|f) | English          | φ(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

- 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

- One to many translation
Decoding Process

Maria no dio una bofetada

Mary did not slap

• Many to one translation

• Many to one translation
Decoding Process

- **Reordering**

- **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>give</td>
<td>a slap</td>
<td>to the</td>
<td>green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did not give</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did not give slap</td>
<td>to the</td>
<td>the witch</td>
<td></td>
<td></td>
<td></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>give</td>
<td>a slap</td>
<td>to the</td>
<td>green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did not give</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did not give slap</td>
<td>to the</td>
<td>the 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>María</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 a slap to the green witch</td>
<td>Mary not did not give a slap to the green witch</td>
<td>Mary did not give a slap to the green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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 \( p(Mary|Maria) \)
  - reordering costs
  - phrase/word count costs
  - ...

- **Language Model**
  - uses trigrams:
  - \( p(Mary \text{ did not}) = p(Mary|\text{START}) \times p(\text{did}|Mary,\text{START}) \times p(\text{not}|Mary \text{ did}) \)
### 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>
<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</td>
<td>slap</td>
<td>by</td>
<td>the</td>
<td>green</td>
<td>witch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>did not give</td>
<td>to</td>
<td>the</td>
<td>the</td>
<td>green</td>
<td>witch</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Add another hypothesis**

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

- **Further hypothesis expansion**

---

Philipp Koehn  
EMNLP Lecture 16  
29 February 2008
### 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>
<tbody>
<tr>
<td>Mary</td>
<td>not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
</tr>
<tr>
<td>did not</td>
<td>give</td>
<td>a slap</td>
<td>by</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>slap</td>
<td>to</td>
<td>the</td>
</tr>
<tr>
<td>did not</td>
<td>give</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Adding more hypothesis

⇒ *Explosion* of search space

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

Philipp Koehn
EMNLP Lecture 16
29 February 2008
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 *same* partial translation
Hypothesis Recombination

- Different paths to the same partial translation
  \[ \Rightarrow \text{Combine paths} \]
  - drop weaker path
  - keep pointer from weaker path (for lattice generation)

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

  dio una bofetada
bruja verde
e: the
f: -----**--
p: 0.354

  better partial translation

  covers easier part
  --> lower cost

- Hypothesis that covers easy part of sentence is preferred
  ⇒ Need to consider 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
  \[ \text{LM} \times \text{TM} = p(\text{to}) \times p(\text{the}|\text{to}) \times p(\text{to the}|\text{a la}) \]

**Future Cost Estimation: Step 2**

- Step 2: find *cheapest cost* among translation options
  \[
  \begin{align*}
  \text{cost} &= 0.0372 \\
  \text{cost} &= 0.0299 \\
  \text{cost} &= 0.0354
  \end{align*}
  \]
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</th>
<th>TM</th>
<th>WordPenalty</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>this is a small house</td>
<td>0</td>
<td>-27.0908</td>
<td>-1.83258</td>
<td>-5</td>
</tr>
<tr>
<td>this is a little house</td>
<td>0</td>
<td>-28.1791</td>
<td>-1.83258</td>
<td>-5</td>
</tr>
<tr>
<td>it is a small house</td>
<td>0</td>
<td>-27.108</td>
<td>-3.21888</td>
<td>-5</td>
</tr>
<tr>
<td>it is a little house</td>
<td>0</td>
<td>-28.1963</td>
<td>-3.21888</td>
<td>-5</td>
</tr>
<tr>
<td>it is an small house</td>
<td>0</td>
<td>-32.3094</td>
<td>-3.21888</td>
<td>-5</td>
</tr>
<tr>
<td>it is an little house</td>
<td>0</td>
<td>-33.7639</td>
<td>-1.83258</td>
<td>-5</td>
</tr>
<tr>
<td>this is a small house</td>
<td>0</td>
<td>-31.5022</td>
<td>-3.21888</td>
<td>-5</td>
</tr>
<tr>
<td>this is a little house</td>
<td>0</td>
<td>-31.5689</td>
<td>-1.83258</td>
<td>-5</td>
</tr>
<tr>
<td>it is an small house</td>
<td>0</td>
<td>-34.3439</td>
<td>-3.21888</td>
<td>-5</td>
</tr>
<tr>
<td>it is an little house</td>
<td>0</td>
<td>-32.9861</td>
<td>-1.83258</td>
<td>-5</td>
</tr>
<tr>
<td>the is a small house</td>
<td>0</td>
<td>-35.6899</td>
<td>-1.83258</td>
<td>-5</td>
</tr>
<tr>
<td>is it a little house</td>
<td>0</td>
<td>-30.3603</td>
<td>-3.91202</td>
<td>-5</td>
</tr>
<tr>
<td>the house is a small</td>
<td>-7</td>
<td>-28.7683</td>
<td>-2.52573</td>
<td>-5</td>
</tr>
<tr>
<td>the is a small house</td>
<td>0</td>
<td>-34.8557</td>
<td>-3.91202</td>
<td>-5</td>
</tr>
<tr>
<td>it 's a little house</td>
<td>0</td>
<td>-35.1446</td>
<td>-3.91202</td>
<td>-5</td>
</tr>
<tr>
<td>this house is a small</td>
<td>-7</td>
<td>-29.3019</td>
<td>-3.91202</td>
<td>-5</td>
</tr>
</tbody>
</table>
XML Markup

Er erzielte <NUMBER english='17.55'>17,55</NUMBER> Punkte.

- Add additional translation options
  - number translation
  - name translation

- Additional options
  - provide multiple translations
  - provide probability distribution along with translations
  - allow bypassing of provided translations