Statistical Machine Translation
Lecture 2
Theory and Praxis of Decoding

Philipp Koehn
pkoehn@inf.ed.ac.uk
School of Informatics
University of Edinburgh

---

Phrases-Based Systems

- A number of research groups developed phrase-based systems (RWTH Aachen, Univ. of Southern California/ISI, CMU, IBM, Johns Hopkins Univ., Cambridge Univ., Univ. of Catalunya, ITC-irst, Univ. Edinburgh, Univ. of Maryland...)

- Systems differ in
  - training methods
  - model for phrase translation table
  - reordering models
  - additional feature functions

- Currently best method for SMT (MT?)
  - top systems in DARPA/NIST evaluation are phrase-based
  - best commercial system for Arabic-English is phrase-based

---

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 | 0.1668         | 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

- One to many translation

- Many to one translation

- Many to one translation

- Reordering
Decoding Process

- Translation finished

Translation Options

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

Hypothesis Expansion

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

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:
    - $p(\text{Mary did not}) = p(\text{Mary} <s> ) \times p(\text{did}|\text{Mary},<s>) \times p(\text{not}|\text{Mary did})$
Hypothesis Expansion

- Further hypothesis expansion

Addition of more hypotheses

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

⇒ Combine paths
  - Drop weaker hypothesis
  - Keep pointer from worse 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)

Pruning

- Hypothesis recombination is not sufficient
  ⇒ Heuristically discard weak hypotheses
- Organize Hypothesis in stacks, e.g. by
  - same foreign words covered
  - same number of foreign words covered (Pharaoh does this)
  - 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 α times the cost of best hypothesis in stack (e.g., α = 0.001)

Comparing Hypotheses

- Comparing hypotheses with same number of foreign words covered
  Maria no dio una bofetada a la bruja verde
  
  \[
  \begin{align*}
  e: & \text{ Mary did not} \\
  f: & \text{ **-------} \\
  p: & 0.154
  \end{align*}
  \]
  
  \[
  \begin{align*}
  e: & \text{ the} \\
  f: & \text{ -------**} \\
  p: & 0.354
  \end{align*}
  \]
  
  better partial translation
  covers easier part
  \[
  \text{⇒ Need to consider future cost of uncovered parts}
  \]

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
    \[
    \text{LM} \times \text{TM} = p(\text{to}) \times p(\text{the|to}) \times p(\text{the|a la})
    \]
Future Cost Estimation: Step 2

- Step 2: find cheapest cost among translation options

\[
\begin{align*}
\text{to the} & \quad \text{cost} = 0.0372 \\
\text{to} & \quad \text{cost} = 0.0299 \\
\text{the} & \quad \text{cost} = 0.0354
\end{align*}
\]

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 precomputed

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
  - factor them to actually accumulated cost for translation option for pruning

Pharaoh

- A beam search decoder for phrase-based models
  - works with various phrase-based models
  - beam search algorithm
  - time complexity roughly linear with input length
  - good quality takes about 1 second per sentence

- Very good performance in DARPA/NIST Evaluation
- Freely available for researchers
  [http://www.isi.edu/license/pharaoh/](http://www.isi.edu/license/pharaoh/)

Running the decoder

- An example run of the decoder:

  % echo 'das ist ein kleines haus' | pharaoh -f pharaoh.ini > out
  Pharaoh v1.2.9, written by Philipp Koehn
  a beam search decoder for phrase-based statistical machine translation models
  (c) 2002-2003 University of Southern California
  (c) 2004 Massachusetts Institute of Technology
  (c) 2005 University of Edinburgh, Scotland
  loading language model from europarl.srilm
  loading phrase translation table from phrase-table, stored 21, pruned 0, kept 21
  loaded data structures in 2 seconds
  translating 1 sentences.translated 1 sentences in 0 seconds
  % cat out
  this is a small house

Phrase Translation Table

- Core model component is the phrase translation table:
### Trace

- **Running the decoder with switch “-t”**

  ```
  % echo 'das ist ein kleines haus' | pharaoh -f pharaoh.ini -t
  this is a small house
  ```

- **Trace for each applied phrase translation**:
  - **output phrase**: (there is)
  - **cost incurred by this phrase**: 0.014086
  - **coverage of foreign words**: 0-1

### Hypothesis Accounting

- **The switch “-v” allows for detailed run time information**:

  ```
  % echo 'das ist ein kleines haus' | pharaoh -f pharaoh.ini -v 2
  ```

  **Statistics over how many hypothesis were generated**:
  - 114 hypotheses were added to hypothesis stacks
  - 284 hypotheses were discarded because they were too bad
  - 0 hypotheses were pruned, because a stack got too big
  - 58 hypotheses were merged due to recombination

- **Probability of the best translation**: \(\exp(-28.9234)\)

### Future Cost Estimation

- **Pre-computation of the future cost estimates**:

  ```
  future costs from 0 to 0 is -5.78855
  future costs from 0 to 1 is -10.207
  ```

### Reordering Example

- **Sometimes phrases have to be reordered**:

  ```
  % echo 'ein kleines haus ist das' | pharaoh -f pharaoh.ini -t -d 0.5
  ```

  **First output phrase (this) is translation of the 4th word**

- **Translation model cost (pC) and future cost estimates (c)**

### Translation Options

- **Even more run time information is revealed with “-v 3”**:

  ```
  [das;2]
  the<1>, pC=-0.916291, c=-5.78855
  it<2>, pC=-2.30259, c=-8.0761
  this<3>, pC=-2.30259, c=-8.00205
  [ist;4]
  is<4>, pC=0, c=-4.92223
  ’s<5>, pC=0, c=-6.11591
  [ein;7]
  an<5>, pC=0, c=-5.55151
  [kleines;9]
  small<10>, pC=-1.60944, c=-9.72116
  little<11>, pC=-1.60944, c=-10.0953
  [haus;10]
  house<12>, pC=0, c=-9.26607
  ```

### Hypothesis Expansion

- **Start of beam search: First hypothesis (das → the)**

  ```
  creating hypothesis 1 from 0 [...]
  base score 0
  covered 0-1: das
  translated as: the => translation cost -0.916291
  distance 0 ⇒ distortion cost 0
  language model cost for ‘the’ -2.03434
  word penalty -0
  score -2.95564 + futureCost -29.4246 = -32.3752
  new best estimate for this stack
  ```
Hypothesis Expansion

- Another hypothesis (das ist $\rightarrow$ this)

creating hypothesis 12 from 0 ( ... $\langle s \rangle$ $\langle s \rangle$ )
base score 0
covering 0-1: das ist
translated as: this $\Rightarrow$ translation cost -0.223144
distance 0 $\Rightarrow$ distortion cost 0
language model cost for 'this' -3.06276
language model cost for 'is' -0.976669
word penalty -0
score -8.56259 $+$ futureCost -24.5023 = -33.0649
new best estimate for this stack
merged hypothesis on stack 2, now size 2

Another hypothesis (this is)
creating hypothesis 27 from 3 ( ... $\langle s \rangle$ this )
base score -5.36535
covering 1-1: ist
translated as: is $\Rightarrow$ translation cost 0
distance 0 $\Rightarrow$ distortion cost 0
language model cost for 'is' -0.976669
word penalty -0
score -6.34202 $+$ futureCost -24.5023 = -30.8443
worse than existing path to 12, discarding

Bad hypothesis that falls out of the beam
creating hypothesis 52 from 6 ( ... $\langle s \rangle$ a )
base score -6.65992
covering 0-0: das
translated as: this $\Rightarrow$ translation cost -2.30259
distance -3 $\Rightarrow$ distortion cost -3
language model cost for 'this' -8.69176
word penalty -0
score -20.6543 $+$ futureCost -23.9095 = -44.5637
estimate below threshold, discarding

Generating Best Translation

- Generating best translation
  - find best final hypothesis (442)
  - trace back path to initial hypothesis

best hypothesis 442
[ 442 $\Rightarrow$ 343 ]
[ 343 $\Rightarrow$ 106 ]
[ 106 $\Rightarrow$ 12 ]
[ 12 $\Rightarrow$ 0 ]

Beam Size

- Trade-off between speed and quality via beam size

$\%$ echo 'das ist ein kleines haus' | pharaoh -f pharaoh.ini -s 10 -v 2 [...]
collected 12 translation options
BYP: 78 added, 122 discarded below threshold, 33 pruned, 20 merged.
BEST: this is a small house -28.9234

<table>
<thead>
<tr>
<th>Beam size</th>
<th>Threshold</th>
<th>Hyp. added</th>
<th>Hyp. discarded</th>
<th>Hyp. pruned</th>
<th>Hyp. merged</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>unlimited</td>
<td>634</td>
<td>0</td>
<td>0</td>
<td>1306</td>
</tr>
<tr>
<td>100</td>
<td>unlimited</td>
<td>557</td>
<td>32</td>
<td>199</td>
<td>572</td>
</tr>
<tr>
<td>100</td>
<td>0.00001</td>
<td>144</td>
<td>284</td>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>10</td>
<td>0.00001</td>
<td>78</td>
<td>122</td>
<td>33</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>0.00001</td>
<td>9</td>
<td>19</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
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

- Current reordering models are weak, so limits improve translation quality

Sample N-Best List

- N-best list from Pharaoh:

<table>
<thead>
<tr>
<th>Translation</th>
<th>Reordering LM TM 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>
</tr>
<tr>
<td>it is a small house</td>
<td>0 -27.104 -3.21888 -5</td>
<td>-30.1121</td>
</tr>
<tr>
<td>it is a little house</td>
<td>0 -28.1963 -3.21888 -5</td>
<td>-31.4125</td>
</tr>
<tr>
<td>this is an small house</td>
<td>0 -31.7294 -1.83258 -5</td>
<td>-33.542</td>
</tr>
<tr>
<td>it is an house little</td>
<td>0 -32.3094 -3.21888 -5</td>
<td>-35.5234</td>
</tr>
<tr>
<td>this is a house small</td>
<td>0 -33.7633 -1.83258 -5</td>
<td>-35.5965</td>
</tr>
<tr>
<td>this is a house little</td>
<td>-3 -31.5689 -1.83258 -5</td>
<td>-36.4015</td>
</tr>
<tr>
<td>it is a little house</td>
<td>0 -34.3439 -3.21888 -5</td>
<td>-37.5269</td>
</tr>
<tr>
<td>it is a house small</td>
<td>0 -33.5689 -1.83258 -5</td>
<td>-37.7211</td>
</tr>
<tr>
<td>it is a house little</td>
<td>0 -34.3439 -3.21888 -5</td>
<td>-37.7235</td>
</tr>
<tr>
<td>this is an house little</td>
<td>-3 -31.5689 -1.83258 -5</td>
<td>-37.8049</td>
</tr>
<tr>
<td>this is an house little</td>
<td>-3 -31.5689 -1.83258 -5</td>
<td>-37.8049</td>
</tr>
<tr>
<td>this is a small house</td>
<td>0 -35.6899 -2.52573 -5</td>
<td>-38.2156</td>
</tr>
<tr>
<td>this is a little house</td>
<td>-4 -30.3603 -3.91202 -5</td>
<td>-38.2723</td>
</tr>
<tr>
<td>the house is a small</td>
<td>0 -38.8557 -3.91202 -5</td>
<td>-38.8767</td>
</tr>
<tr>
<td>this house is a little</td>
<td>-7 -28.5107 -2.52573 -5</td>
<td>-38.8364</td>
</tr>
<tr>
<td>the is a small house</td>
<td>0 -35.6899 -2.52573 -5</td>
<td>-38.8364</td>
</tr>
<tr>
<td>the is a little house</td>
<td>-4 -30.3603 -3.91202 -5</td>
<td>-38.8767</td>
</tr>
<tr>
<td>is it a little house</td>
<td>0 -38.8557 -3.91202 -5</td>
<td>-38.8767</td>
</tr>
<tr>
<td>it’s a small house</td>
<td>0 -34.8557 -3.91202 -5</td>
<td>-38.8767</td>
</tr>
<tr>
<td>it’s a house small</td>
<td>0 -35.6899 -2.52573 -5</td>
<td>-38.8364</td>
</tr>
<tr>
<td>this house is a little</td>
<td>-7 -28.5107 -2.52573 -5</td>
<td>-38.8364</td>
</tr>
</tbody>
</table>

Thank You!

- Questions?