**Phrase-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”:

<table>
<thead>
<tr>
<th>English</th>
<th>(\phi(e,f))</th>
<th>English</th>
<th>(\phi(e,f))</th>
</tr>
</thead>
<tbody>
<tr>
<td>the proposal</td>
<td>0.6227</td>
<td>the suggestions</td>
<td>0.0114</td>
</tr>
<tr>
<td>’s proposal</td>
<td>0.1068</td>
<td>the proposed</td>
<td>0.0114</td>
</tr>
<tr>
<td>a proposal</td>
<td>0.0341</td>
<td>the motion</td>
<td>0.0091</td>
</tr>
<tr>
<td>the idea</td>
<td>0.0250</td>
<td>the idea of</td>
<td>0.0091</td>
</tr>
<tr>
<td>this proposal</td>
<td>0.0227</td>
<td>the proposal ,</td>
<td>0.0068</td>
</tr>
<tr>
<td>proposal</td>
<td>0.0205</td>
<td>its proposal</td>
<td>0.0068</td>
</tr>
<tr>
<td>of the proposal</td>
<td>0.0159</td>
<td>it</td>
<td>0.0068</td>
</tr>
<tr>
<td>the proposals</td>
<td>0.0159</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

---

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

- Further hypothesis expansion

**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 $\alpha$ times the cost of best hypothesis in stack (e.g., $\alpha = 0.001$)

Comparing Hypotheses

- Comparing hypotheses with same number of foreign words covered

Comparing partial translation

- 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
  ⇒ $LM \times TM = p(to) \times p(\text{the}) \times p(\text{to the}) / p(\text{a la})$
Future Cost Estimation: Step 2

- Step 2: find cheapest cost among translation options

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

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/d-sw/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:
Running the decoder with switch "-t"

```
% echo 'das ist ein kleines haus' | pharaoh -f pharaoh.ini -t
this is a small house -28.9234
```

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

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
future costs from 0 to 2 is -15.7221
future costs from 0 to 3 is -25.4433
distance 0 => distortion cost 0
language model cost for 'the' -2.03434
word penalty -0
score -2.95564 + futureCost -29.4246 = -32.3752
```

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

```
creating hypothesis 1 from 0 [...] <a>
base score 0
covering 0-1: das
translated as: the => translation cost -0.916291
word penalty -0
score -2.95564 + futureCost -29.4246 = -32.3752
new best estimate for this stack
merged hypothesis on stack 1, now size 1
```
Hypothesis Expansion

- Another hypothesis (das ist → this is)

creating hypothesis 12 from 0 (...) <s> <s> <s>
base score 0
covering 0-1: das ist
translated as: this is → translation cost -0.223144
distance 0 → distortion cost 0
language model cost for ‘this’ -3.06276
language model cost for ‘is’ -0.976669
word penalty -0
new best estimate for this stack
merged hypothesis on stack 2, now size 2

Hypothesis Expansion

- Hypothesis recombination

creating hypothesis 27 from 3 (...) <s> this

base score -5.36535
covering 1-1: ist
translated as: is → translation cost 0
distance 0 → 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

Hypothesis Expansion

- Bad hypothesis that falls out of the beam

creating hypothesis 52 from 6 (...) <s> a

base score -6.65992
covering 0-0: das
translated as: this → translation cost -2.30259
distance -3 → 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 => 343 ]
[ 343 => 106 ]
[ 106 => 12 ]
[ 12 => 0 ]

Translation Table Pruning

- Limiting translation table size speeds up search
- Histogram pruning: keeping only top $n$ entries
- Threshold pruning: keep only entries that score $n$ times worse than best

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>

Philipp Koehn, University of Edinburgh
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

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

- N-best list from Pharaoh:

<table>
<thead>
<tr>
<th>Translation</th>
<th>LM</th>
<th>TM</th>
<th>WordPenalty</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>this is a small house</td>
<td>0.27098</td>
<td>-1.83258</td>
<td>-5</td>
<td>-30.0117</td>
</tr>
<tr>
<td>it is a small house</td>
<td>0.28179</td>
<td>-1.83258</td>
<td>-5</td>
<td>-31.4152</td>
</tr>
<tr>
<td>this is a small house</td>
<td>0.31729</td>
<td>-1.83258</td>
<td>-5</td>
<td>-33.562</td>
</tr>
<tr>
<td>it is an small house</td>
<td>0.32304</td>
<td>-3.21888</td>
<td>-5</td>
<td>-35.5283</td>
</tr>
<tr>
<td>this is a house small</td>
<td>0.33763</td>
<td>-1.83258</td>
<td>-5</td>
<td>-36.3176</td>
</tr>
<tr>
<td>this is a small house</td>
<td>-3.315689</td>
<td>-1.83258</td>
<td>-5</td>
<td>-36.4015</td>
</tr>
<tr>
<td>it is a small house</td>
<td>-3.43349</td>
<td>-3.21888</td>
<td>-5</td>
<td>-37.562</td>
</tr>
<tr>
<td>this is a small house</td>
<td>-3.31502</td>
<td>-3.21888</td>
<td>-5</td>
<td>-37.7211</td>
</tr>
<tr>
<td>it is an house small</td>
<td>-3.328999</td>
<td>-1.83258</td>
<td>-5</td>
<td>-37.7325</td>
</tr>
<tr>
<td>it is a small house</td>
<td>-3.315689</td>
<td>-1.83258</td>
<td>-5</td>
<td>-37.8049</td>
</tr>
<tr>
<td>it is a small house</td>
<td>-3.329837</td>
<td>-1.83258</td>
<td>-5</td>
<td>-37.8163</td>
</tr>
<tr>
<td>the house is a small</td>
<td>-7.285107</td>
<td>-2.52557</td>
<td>-5</td>
<td>-38.0364</td>
</tr>
<tr>
<td>the is a small house</td>
<td>-35.6899</td>
<td>-2.52557</td>
<td>-5</td>
<td>-38.2156</td>
</tr>
<tr>
<td>it is a small house</td>
<td>-30.3603</td>
<td>-3.91202</td>
<td>-5</td>
<td>-38.2723</td>
</tr>
<tr>
<td>the house is a small</td>
<td>-28.7683</td>
<td>-3.91202</td>
<td>-5</td>
<td>-38.294</td>
</tr>
<tr>
<td>it is a small house</td>
<td>-34.8557</td>
<td>-3.91202</td>
<td>-5</td>
<td>-38.7677</td>
</tr>
<tr>
<td>this house is a little</td>
<td>-29.0443</td>
<td>-3.91202</td>
<td>-5</td>
<td>-38.9563</td>
</tr>
<tr>
<td>this house is a small</td>
<td>-35.1446</td>
<td>-3.91202</td>
<td>-5</td>
<td>-39.0566</td>
</tr>
<tr>
<td>this house is a small</td>
<td>-28.3018</td>
<td>-3.91202</td>
<td>-5</td>
<td>-39.2139</td>
</tr>
</tbody>
</table>

XML Interface

- Add additional translation options
  - number translation
  - noun phrase translation [Koehn, 2003]
  - name translation

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

Thank You!

- Questions?