Spring School
Day 3: Decoding / Phrase-based models

MT Marathon
14 May 2008

Statistical Machine Translation

• Components: Translation model, language model, decoder

foreign/English parallel text

English text

statistical analysis

statistical analysis

Translation Model

Language Model

Decoding Algorithm
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            | $\phi(e|f)$ | 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

- 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

- Many to one translation

Decoding Process

- 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>slap</td>
<td>by</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>to</td>
<td>the</td>
<td>slap</td>
<td>the witch</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>slap</td>
<td>by</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>to</td>
<td>the</td>
<td>slap</td>
<td>the witch</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
14

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></td>
<td>did not</td>
<td>give</td>
<td>a slap</td>
<td>by to</td>
<td>the green witch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>did not give</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>slap</td>
<td>to</td>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>slap</td>
<td>to</td>
<td>the 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

15

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

- Add another hypothesis

Further hypothesis expansion
Hypothesis Expansion

- 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
Explosion of Search Space

- Number of hypotheses is \textit{exponential} with respect to sentence length
  \[ \Rightarrow \text{Decoding is NP-complete} \quad \text{[Knight, 1999]} \]
  \[ \Rightarrow \text{Need to reduce search space} \]
  - risk free: hypothesis \textit{recombination}
  - risky: \textit{histogram/threshold pruning}

Hypothesis Recombination

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

⇒ **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
- 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

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

dio una bofetada
bruja verde
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
  \[ LM \times TM = p(\text{to}) \times p(\text{the}|\text{to}) \times p(\text{to \, the}|\text{a \, la}) \]

Future Cost Estimation: Step 2

- Step 2: find cheapest cost among translation options

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

- 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</th>
<th>TM WordPenalty</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>this is a small house</td>
<td>0 -27.0908</td>
<td>-1.83258 -5</td>
<td>-28.9234</td>
</tr>
<tr>
<td>this is a little house</td>
<td>0 -28.1791</td>
<td>-1.83258 -5</td>
<td>-30.0117</td>
</tr>
<tr>
<td>it is a small house</td>
<td>0 -27.108</td>
<td>-3.21888 -5</td>
<td>-30.5268</td>
</tr>
<tr>
<td>it is a little house</td>
<td>0 -28.1963</td>
<td>-3.21888 -5</td>
<td>-31.4152</td>
</tr>
<tr>
<td>it is an small house</td>
<td>0 -32.3094</td>
<td>-3.21888 -5</td>
<td>-35.5283</td>
</tr>
<tr>
<td>it is an little house</td>
<td>0 -33.7639</td>
<td>-3.21888 -5</td>
<td>-35.5965</td>
</tr>
<tr>
<td>this is a house small</td>
<td>-3 -31.4861</td>
<td>-1.83258 -5</td>
<td>-36.3176</td>
</tr>
<tr>
<td>this is a house little</td>
<td>-3 -31.5689</td>
<td>-1.83258 -5</td>
<td>-36.6015</td>
</tr>
<tr>
<td>this is a small house</td>
<td>0 -34.3439</td>
<td>-3.21888 -5</td>
<td>-37.5628</td>
</tr>
<tr>
<td>this is a little house</td>
<td>0 -34.5022</td>
<td>-3.21888 -5</td>
<td>-37.7211</td>
</tr>
<tr>
<td>this is an small house</td>
<td>0 -32.8999</td>
<td>-1.83258 -5</td>
<td>-37.7325</td>
</tr>
<tr>
<td>this is an little house</td>
<td>0 -33.1868</td>
<td>-3.21888 -5</td>
<td>-37.8049</td>
</tr>
<tr>
<td>this is a small house</td>
<td>-3 -32.9837</td>
<td>-1.83258 -5</td>
<td>-37.8163</td>
</tr>
<tr>
<td>this is a little house</td>
<td>-3 -33.3689</td>
<td>-3.21888 -5</td>
<td>-38.0364</td>
</tr>
<tr>
<td>the house is a small house</td>
<td>0 -35.6999</td>
<td>-2.52573 -5</td>
<td>-38.2156</td>
</tr>
<tr>
<td>the house is a little house</td>
<td>0 -36.3663</td>
<td>-3.91202 -5</td>
<td>-38.5723</td>
</tr>
<tr>
<td>the house is a small house</td>
<td>0 -38.7683</td>
<td>-2.52573 -5</td>
<td>-38.7294</td>
</tr>
<tr>
<td>it ’s a small house</td>
<td>0 -34.8557</td>
<td>-3.91202 -5</td>
<td>-38.7677</td>
</tr>
<tr>
<td>it ’s a little house</td>
<td>0 -34.1446</td>
<td>-3.91202 -5</td>
<td>-39.0566</td>
</tr>
<tr>
<td>this house is a small house</td>
<td>-7 -28.3018</td>
<td>-3.91202 -5</td>
<td>-39.2139</td>
</tr>
</tbody>
</table>
Moses: Open Source Toolkit

- **Open source** statistical machine translation system (developed from scratch 2006)
  - state-of-the-art *phrase-based* approach
  - novel methods: *factored translation models*, *confusion network decoding*
  - support for **very large models** through *memory-efficient* data structures

- Documentation, source code, binaries **available** at [http://www.statmt.org/moses/](http://www.statmt.org/moses/)
- Development also **supported by**
  - EC-funded *TC-STAR* project
  - *US* funding agencies DARPA, NSF
  - universities (Edinburgh, Maryland, MIT, ITC-irst, RWTH Aachen, ...)

Phrase-based models
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-based translation model

- Major components of phrase-based model
  - phrase translation model \( \phi(f|e) \)
  - reordering model \( \omega^d(\text{start}_i - \text{end}_{i-1} - 1) \)
  - language model \( p_{LM}(e) \)
- Bayes rule
  \[
  \arg\max_{e} p(e|f) = \arg\max_{e} p(f|e)p(e) = \arg\max_{e} \phi(f|e) p_{LM}(e) \omega^d(\text{start}_i - \text{end}_{i-1} - 1)
  \]
- Sentence \( f \) is decomposed into \( I \) phrases \( f^I_1 = f_1, ..., f_I \)
- Decomposition of \( \phi(f|e) \)
  \[
  \phi(f^I_1|e^I_1) = \prod_{i=1}^{I} \phi(f_i|e_i) \omega^d(\text{start}_i - \text{end}_{i-1} - 1))
  \]
Advantages of phrase-based translation

- *Many-to-many* translation can handle non-compositional phrases
- Use of *local context* in translation
- The more data, the *longer phrases* can be learned

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       | ...          | ...             |
How to learn the phrase translation table?

- Start with the *word alignment*:

\[
\begin{array}{cccc}
\text{Maria no} & \text{daba} & \text{una} & \text{bofetada} \\
\text{a} & \text{la} & \text{bruja} & \text{verde} \\
\text{Mary} & \text{did} & \text{not} & \text{slap} \\
\text{the} & \text{green} & \text{witch} & \\
\end{array}
\]

- Collect all phrase pairs that are **consistent** with the word alignment

**Consistent with word alignment**

\[
\begin{array}{ccc}
\text{Maria no} & \text{daba} & \text{una} \\
\text{Mary} & \text{did} & \text{not} \\
\text{slap} & \text{consistent} \\
\text{Mary} & \text{did} & \text{not} \\
\text{slap} & \text{inconsistent} \\
\text{Mary} & \text{did} & \text{not} \\
\text{slap} & \text{inconsistent} \\
\end{array}
\]

- **Consistent with the word alignment** := phrase alignment has to contain all alignment points for all covered words

\[
(\tau, \overline{f}) \in BP \iff \forall e_i \in \tau : (e_i, f_j) \in A \rightarrow f_j \in \overline{f} \\
\text{AND} \quad \forall f_j \in \overline{f} : (e_i, f_j) \in A \rightarrow e_i \in \tau
\]
Word alignment induced phrases

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch)
Word alignment induced phrases

(María, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(María no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (María no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)
Word alignment induced phrases (5)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch),  (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch),  (no daba una bofetada a la bruja verde, did not slap the green witch),
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)

Probability distribution of phrase pairs

- We need a probability distribution \( \phi(f|e) \) over the collected phrase pairs
  - Possible choices
    - relative frequency of collected phrases: 
      \[
      \phi(f|e) = \frac{\text{count}(f,e)}{\sum_{f} \text{count}(f,e)}
      \]
    - or, conversely \( \phi(e|f) \)
    - use lexical translation probabilities
Reordering

- **Monotone** translation
  - do not allow any reordering
  - → worse translations
- **Limiting** reordering (to movement over max. number of words) helps
- **Distance-based** reordering cost
  - moving a foreign phrase over $n$ words: cost $\omega^n$
- **Lexicalized** reordering model

Lexicalized reordering models

- Three orientation types: **monotone**, **swap**, **discontinuous**
- Probability $p(swap|e, f)$ depends on foreign (and English) **phrase** involved
Learning lexicalized reordering models

- Orientation type is *learned during phrase extractions*
- *Alignment point* to the *top left* (monotone) or *top right* (swap)?
- For more, see [Tillmann, 2003] or [Koehn et al., 2005]