Introduction

Learning the Phrase Translation Table

Motivation

Intro and Learning

- Word-Based Models translate words as atomic units
- Phrase-Based Models translate phrases as atomic units
- Advantages:
  - many-to-many translation can handle non-compositional phrases
  - use of local context in translation
  - the more data, the longer phrases can be learned
- "Standard Model", used by Google Translate and others
Phrase-Based Model

- source is segmented in phrases
- each source phrase is translated into observed phrase
- observed phrases are reordered

Compared to IBM Model

- recall IBM models assumed a hidden alignment between s and o, giving a formula \( p(o, a|s) \) and so a formula for \( p(o, a, s) \) as
  \[
  p(o, a|s) \times p(s)
  \]
- phrase-based models assume a hidden segmentations of s and o into K phrases \( \tilde{s}_{1:K} \) and \( \tilde{o}_{1:K} \)
- phrase-based models also assume a hidden mapping from the phrases \( \tilde{s} \) to the phrases \( \tilde{o} \). This 1-to-1, and generally not order preserving.
- we will have a formula for formula for \( p(\tilde{o}, \tau, \tilde{s}) \) as
  \[
  p(\tilde{o}, \tau|\tilde{s}) \times p(\tilde{s})
  \]

Example

- Task: translate a certain German ‘observed’ sentence into ‘source’ English
  \[
  \text{er geht ja nicht nach hause}
  \]

Constructing a Phrase-Based Translation

- assume \( s_{1:5} = \text{he does not go home} \) and \( o_{1:6} = \text{er geht ja nicht nach hause} \)
- possible segmentation of \( s_{1:5} \) into \( \tilde{s}_{1:4} \) is
  \[
  \tilde{s}_1 = s_{1:1} = \text{he}, \tilde{s}_2 = s_{2:3} = \text{does not}, \tilde{s}_3 = s_{4:4} = \text{go}, \tilde{s}_4 = s_{5:5} = \text{home}
  \]
- possible segmentation of \( o_{1:6} \) into \( \tilde{o}_{1:4} \) is
  \[
  \tilde{o}_1 = o_{1:1} = \text{er}, \tilde{o}_2 = o_{2:2} = \text{geht}, \tilde{o}_3 = o_{3:4} = \text{ja nicht},
  \tilde{o}_4 = o_{5:6} = \text{nach hause}
  \]
- possible mapping \( \tau \) from \( \tilde{s} \) to \( \tilde{o} \) is
  \[
  \tau(1) = 1, \tau(2) = 3, \tau(3) = 2, \tau(4) = 4
  \]
Constructing a Phrase-Based Translation

- Assume a ‘phrase-table’ giving for many possible ‘phrases’ $\tilde{o}$ in the observed German, possible ‘phrases’ $\tilde{s}$ in potential source English

<table>
<thead>
<tr>
<th>er</th>
<th>geht</th>
<th>ja</th>
<th>nicht</th>
<th>nach</th>
<th>hause</th>
</tr>
</thead>
<tbody>
<tr>
<td>he</td>
<td>is</td>
<td>yes</td>
<td>not</td>
<td>after</td>
<td>home</td>
</tr>
<tr>
<td>it</td>
<td>are</td>
<td>is</td>
<td>does not</td>
<td>in</td>
<td>chamber</td>
</tr>
<tr>
<td>he</td>
<td>go</td>
<td>not</td>
<td>is not</td>
<td>under</td>
<td>home</td>
</tr>
<tr>
<td>he</td>
<td>is not</td>
<td>not</td>
<td>does not</td>
<td>return</td>
<td>home</td>
</tr>
<tr>
<td>he</td>
<td>does</td>
<td>not</td>
<td>to</td>
<td>not to</td>
<td>go</td>
</tr>
<tr>
<td>he</td>
<td>is</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>to</td>
</tr>
<tr>
<td>he</td>
<td>are</td>
<td>not</td>
<td>after</td>
<td>not</td>
<td>to</td>
</tr>
<tr>
<td>he</td>
<td>no</td>
<td>not</td>
<td>after</td>
<td>not</td>
<td>to</td>
</tr>
<tr>
<td>he</td>
<td>not</td>
<td>not</td>
<td>after</td>
<td>not</td>
<td>to</td>
</tr>
<tr>
<td>he</td>
<td>not</td>
<td>not</td>
<td>after</td>
<td>not</td>
<td>to</td>
</tr>
</tbody>
</table>

- the phrase-based translation will be built with these ingredients

Constructing a Phrase-Based Translation

- Pick a phrase $\tilde{o} = \text{‘er’}$ in observed, choose ‘he’ as $\tilde{s}_1$ in source

Constructing a Phrase-Based Translation

- Pick a phrase $\tilde{o} = \text{‘ja nicht’}$ in observed, choose ‘does not’ as $\tilde{s}_2$ in source

Constructing a Phrase-Based Translation

- Pick a phrase $\tilde{o} = \text{‘geht’}$ in observed, choose ‘go’ as $\tilde{s}_3$ in source

- NB: allowed to choose $\tilde{o}$ phrases out of sequence; $\tilde{s}$ phrases chosen in sequence

- NB: phrases may have multiple words: many-to-many translation
### Constructing a Phrase-Based Translation

Pick a phrase $\delta = \text{‘nach hause’} \text{ in observed, choose ‘home’ as } s^4$

- just constructed one particular translation, could have constructed many, many others using the available phrases pairs
- need probabilistic model which favours one over the other
- need to set parameters of that model
  $\rightarrow$ these won’t be learned by EM but instead some are (heuristically) derived from IBM models, and some just set by common sense
- to find high scoring translations need to manage somehow an exponential search space
  $\rightarrow$ ‘beam search’ heuristic

### Learning a Phrase Translation Table

Task: learn the model from a parallel corpus

Three stages:
- word alignment: using IBM models or other method
- extraction of phrase pairs
- scoring phrase pairs

Learning ctd: alignment both ways

do IBM model learning in both directions, and find best alignments both ways
### Learning ctd: unite alignment

<table>
<thead>
<tr>
<th>Michael</th>
<th>geht</th>
<th>davon</th>
<th>aus</th>
<th>dass</th>
<th>er</th>
<th>im</th>
<th>haus</th>
<th>bleibt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael</td>
<td>assumes</td>
<td>that</td>
<td>he</td>
<td>will stay in the house</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For each training pair, merge these alignments.

Then extract phrase pair consistent with this merge:

Next slides show a few cases.

### Learning ctd: extract consistent phrase pairs

#### N-to-N cases: basically taping together adjacent smaller cases.

*eg. (in the — im) + (house — haus)*

→ (in the house — im haus)

#### obvious 1-to-N, N-to-1 cases eg:

- (that — dass)
- (assumes — geht davon aus)
- (in the — im)
Learning ctd: extract consistent phrase pairs

N-to-N cases: \((\text{michael} \rightarrow \text{michael}) + (\text{assumes} \rightarrow \text{geht davon aus})\)
\(\rightarrow (\text{michael assumes} \rightarrow \text{michael geht davon aus})\)

Scoring Phrase Translations

- Preceding slides show some of the phrase pairs extracted from one sentence pair; this is done over all sentence pairs. Some pairs will be frequently extracted, others less so . . .
- so from huge table of counts of phrase pairs, phrase-translation probabilities are simply defined by relative frequencies:

\[
tr(\tilde{e}|\tilde{g}) = \frac{\text{count}(\tilde{e}, \tilde{g})}{\sum_{\tilde{g}'} \text{count}(\tilde{e}, \tilde{g}')} \\
tr(\tilde{g}|\tilde{e}) = \frac{\text{count}(\tilde{e}, \tilde{g})}{\sum_{\tilde{e}'} \text{count}(\tilde{e}', \tilde{g})}
\]

- so phrase probs acquired by exploiting the EM-learned IBM probs
Phrase Translation Probabilities: an example

- below is an extract from table learnt from the Europarl corpus, giving some values of $\text{tr}(\tilde{e}|\tilde{g})$ for $\tilde{g} = \text{den Vorschlag}$ and various English ‘phrases’

| English         | $\phi(\tilde{e}|\tilde{g})$ |
|-----------------|-----------------------------|
| the proposal    | 0.6227                      |
| ‘s proposal     | 0.1068                      |
| a proposal      | 0.0341                      |
| the idea        | 0.0250                      |
| this proposal   | 0.0227                      |
| proposal        | 0.0205                      |
| of the proposal | 0.0159                      |
| the proposals   | 0.0159                      |
| the suggestions | 0.0114                      |
| the motion      | 0.0091                      |
| the idea of     | 0.0091                      |
| the proposal,   | 0.0068                      |
| its proposal    | 0.0068                      |
| it              | 0.0068                      |

- lexical variation (proposal vs suggestions)
- morphological variation (proposal vs proposals)
- included function words (the, a, ...)
- noise (it)

Linguistic Phrases?

- Phrase-table emphatically is not limited to 'linguistic' phrases – that is sequences which are defined by detailed language grammars (noun phrases, verb phrases, prepositional phrases, ...)
- Example non-linguistic phrase pair
  
  $\text{spass am} \rightarrow \text{fun with the}$

- Prior noun often helps with translation of preposition
- ‘phrases’ can include tacked on punctuation
- consensus is that if attempts are made to limit to grammatically motivated ‘linguistic’ phrases, overall translation quality goes down