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 hidden segmentations of $s$ and $o$ into $K$ phrases $\bar{s}_{1:K}$ and $\bar{o}_{1:K}$
- phrase-based models also assume a hidden mapping from the phrases $\bar{s}$ to the phrases $\bar{o}$. This 1-to-1, and generally not order preserving.
- we will have a formula for formula for $p(\bar{o},\tau,\bar{s})$ as

$$p(\bar{o},\tau|\bar{s}) \times p(\bar{s})$$

Example

- 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 $\bar{s}_{1:4}$ is

$$\bar{s}_1 = s_{1:1} = \text{he}, \bar{s}_2 = s_{2:3} = \text{does not}, \bar{s}_3 = s_{4:4} = \text{go}, \bar{s}_4 = s_{5:5} = \text{home}$$
- possible segmentation of $o_{1:6}$ into $\bar{o}_{1:4}$ is

$$\bar{o}_1 = o_{1:1} = \text{er}, \bar{o}_2 = o_{2:2} = \text{geht}, \bar{o}_3 = o_{3:4} = \text{ja nicht}, \bar{o}_4 = o_{4:6} = \text{nach hause}$$
- possible mapping $\tau$ from $\bar{s}$ to $\bar{o}$ is

$$\tau(1) = 1, \tau(2) = 3, \tau(3) = 2, \tau(4) = 4$$

Constructing a Phrase-Based Translation

- Task: translate a certain German 'observed' sentence into 'source' English

$$\text{er geht ja nicht nach hause}$$
Constructing a Phrase-Based Translation

- Assume a ‘phrase-table’ giving for many possible ‘phrases’ $\phi$ in the observed German, possible ‘phrases’ $\phi'$ in potential source English

\[
\begin{array}{c|c|c|c|c|c}
\text{er} & \text{geht} & \text{ja} & \text{nicht} & \text{nach} & \text{hause} \\
\hline
\text{he} & \text{is} & \text{yes} & \text{not} & \text{to} & \text{home} \\
\text{it} & \text{is} & \text{does not} & \text{at} & \text{in} & \text{home} \\
\text{he} & \text{does} & \text{not} & \text{at} & \text{not} & \text{under} \\
\text{he} & \text{goes} & \text{not} & \text{not} & \text{not} & \text{under} \\
\end{array}
\]

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

\[
\begin{align*}
\text{he} & \rightarrow \text{er} \\
\text{does not} & \rightarrow \text{ja nicht} \\
\end{align*}
\]

\[
\text{er geht ja nicht nach hause}
\]

- Pick a phrase $\phi = \text{\‘er\’}$ in observed, choose ‘he’ as $\phi_1$ in source

- NB: allowed to choose $\phi$ phrases out of sequence; $\phi'$ phrases chosen in sequence

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

- 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.
  → 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.
  → ’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

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. \((\text{in the} - \text{im}) + (\text{house} - \text{haus})\)
\(\rightarrow (\text{in the house} — \text{im haus})\)
Learning ctd: extract consistent phrase pairs

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

Learning ctd: what you can’t extract

no pair \((e \leftarrow e\text{ im})\)
no pair \((\text{he will stay} \leftarrow \text{he})\)
because corresponding parts not adjacent

Scoring Phrase Translations

\(\triangleright\) 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 . . .

\(\triangleright\) so from huge table of counts of phrase pairs, phrase-translation probabilities are simply defined by relative frequencies:

\[
\begin{align*}
\text{tr}(\bar{e} | \bar{g}) &= \frac{\text{count}(\bar{e}, \bar{g})}{\sum_{\bar{g}'} \text{count}(\bar{e}, \bar{g}')}, \\
\text{tr}(\bar{g} | \bar{e}) &= \frac{\text{count}(\bar{e}, \bar{g})}{\sum_{\bar{e}'} \text{count}(\bar{e}', \bar{g})}
\end{align*}
\]

\(\triangleright\) 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 $tr(\bar{e}\mid\bar{g})$ for $\bar{g} = \text{den Vorschlag}$ and various English 'phrases'

<table>
<thead>
<tr>
<th>English</th>
<th>$\phi(\bar{e}\mid\bar{g})$</th>
<th>English</th>
<th>$\phi(\bar{e}\mid\bar{g})$</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>

- 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