4CSLL5
’Advanced Computational Linguistics’
Phrase Based Machine Trans

Martin Emms

October 31, 2018
The Phrase-Based Translation Model
The Phrase Based Translation Model
The Phrase Translation Model

- Simplest assumption is that all segmentations are equally likely, so there are no explicit probability terms for choice of segmentation.
The Phrase Translation Model

- Simplest assumption is that all segmentations are equally likely, so there are no explicit probability terms for choice of segmentation.
- For any particular segmentations of $o$ and $s$ into $K$ segments, the model of $p(\bar{o}, \tau, \bar{s})$ is:

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

$$= \left[ \prod_{k=1}^{K} tr(\bar{o}_{\tau(k)}|\bar{s}_k) d(\bar{o}_{\tau(k-1)}, \bar{o}_{\tau(k)}) \right] LM(s)$$
The Phrase Translation Model

- Simplest assumption is that all segmentations are equally likely, so there are no explicit probability terms for choice of segmentation.
- For any particular segmentations of $o$ and $s$ into $K$ segments, the model of $p(\bar{o}, \tau, \bar{s})$ is:

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

$$= \left[ \prod_{k=1}^{K} tr(\bar{o}_{\tau(k)}|\bar{s}_k) \ d(\bar{o}_{\tau(k-1)}, \bar{o}_{\tau(k)}) \right] \ LM(s)$$

- **Phrase translation**: there’s a term $tr(\bar{o}_{\tau(k)}|\bar{s}_k)$ for the phrase-translation probabilities for an observed phrase $\bar{o}_{\tau(k)}$ to be generated from source phrase $\bar{s}_k$
The Phrase Translation Model

- Simplest assumption is that all segmentations are equally likely, so there are no explicit probability terms for choice of segmentation.
- For any particular segmentations of \( o \) and \( s \) into \( K \) segments, the model of \( p(\bar{o}, \tau, \bar{s}) \) is:

\[
p(\bar{o}, \tau, \bar{s}) = p(\bar{o}, \tau|\bar{s}) \times p(\bar{s})
\]

\[
= \left[ \prod_{k=1}^{K} tr(\bar{o}_{\tau(k)}|\bar{s}_k) \ d(\bar{o}_{\tau(k-1)}, \bar{o}_{\tau(k)}) \right] \ LM(s)
\]

- **Phrase translation**: there’s a term \( tr(\bar{o}_{\tau(k)}|\bar{s}_k) \) for the phrase-translation probabilities for an observed phrase \( \bar{o}_{\tau(k)} \) to be generated from source phrase \( \bar{s}_k \)
- **Reordering**: there is re-ordering term concerning how likely the destination for a \( \bar{s}_k \) phrase is given destination for previous \( \bar{s}_{k-1} \) phrase: \( d(\bar{o}_{\tau(k-1)}, \bar{o}_{\tau(k)}) \)
The Phrase Translation Model

- Simplest assumption is that all segmentations are equally likely, so there are no explicit probability terms for choice of segmentation.
- For any particular segmentations of $o$ and $s$ into $K$ segments, the model of $p(\bar{o}, \tau, \bar{s})$ is:

\[
p(\bar{o}, \tau, \bar{s}) = p(\bar{o}, \tau|\bar{s}) \times p(\bar{s})
\]

\[
= \left[ \prod_{k=1}^{K} tr(\bar{o}_{\tau(k)}|\bar{s}_k) \right] d(\bar{o}_{\tau(k-1)}, \bar{o}_{\tau(k)}) \quad LM(s)
\]

- **Phrase translation**: there’s a term $tr(\bar{o}_{\tau(k)}|\bar{s}_k)$ for the phrase-translation probabilities for an observed phrase $\bar{o}_{\tau(k)}$ to be generated from source phrase $\bar{s}_k$

- **Reordering**: there is re-ordering term concerning how likely the destination for a $\bar{s}_k$ phrase is given destination for previous $\bar{s}_{k-1}$ phrase: $d(\bar{o}_{\tau(k-1)}, \bar{o}_{\tau(k)})$

- **Language model**: the probability of the source phrases $\bar{s}$ is equated to simply the probability of the source sequence $s$ as given by an $n$-gram model $LM(s)$
using (2), preferred translation $s_{\text{best}}$ is defined as the source part of

$$\langle \bar{s}, \tau \rangle_{\text{best}} = \arg \max_{\bar{s}, \tau} p(\bar{o}, \tau, \bar{s})$$
The Phrase Translation Model contd

- using (2), preferred translation $s_{\text{best}}$ is defined as the source part of

$$\langle \bar{s}, \tau \rangle_{\text{best}} = \arg \max_{\bar{s}, \tau} p(\bar{0}, \tau, \bar{s})$$

- note you could seek to define $s_{\text{best}}$ as

$$s_{\text{best}} = \arg \max_{\bar{s}} \sum_{\tau} p(\bar{0}, \tau, \bar{s})$$

but this is not standardly done; the above is regarded as a 'Viterbi' approximation of the sum
The distortion term

- The distortion term $d(\overline{o}_{\tau(k-1)}, \overline{o}_{\tau(k)})$ is not learned via EM
- instead just standardly defined as an exponentially decaying function of the 'distance' $x$ between end of $\overline{o}_{\tau(k-1)}$ and start of $\overline{o}_{\tau(k)}$
- in particular where $x = |fst(\overline{o}_{\tau(k)}) - lst(\overline{o}_{\tau(k-1)}) - 1|$, the $d$ term is $\alpha^x$ for some $\alpha < 1$. 
illustration of Distance-Based Reordering

\[ src \quad \boxed{1} \quad \boxed{2} \quad \boxed{3} \quad \boxed{4} \]

\[ obs \quad \boxed{1 \ 2 \ 3} \quad \boxed{4 \ 5} \quad \boxed{6} \quad \boxed{7} \]

<table>
<thead>
<tr>
<th>phrase ( \bar{s}_i )</th>
<th>translated to ( \bar{o}_{\tau(i)} )</th>
<th>change ( \bar{o}<em>{\tau(i-1)} ) to ( \bar{o}</em>{\tau(i)} )</th>
<th>'displacement' ( \times )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{s}_1 )</td>
<td>( \bar{o}_{\tau(1)} = o_1 : o_3 )</td>
<td>start at beginning</td>
<td>0</td>
</tr>
<tr>
<td>( \bar{s}_2 )</td>
<td>( \bar{o}_{\tau(2)} = o_6 : o_6 )</td>
<td>( o_3 ) to ( o_6 )</td>
<td>+2 (( = ) ( \text{fst}(o_6 : o_6) - \text{lst}(o_1 : o_3) - 1 ))</td>
</tr>
<tr>
<td>( \bar{s}_3 )</td>
<td>( \bar{o}_{\tau(3)} = o_4 : o_5 )</td>
<td>( o_6 ) to ( o_4 )</td>
<td>-3 (( = ) ( \text{fst}(o_4 : o_5) - \text{lst}(o_6 : o_6) - 1 ))</td>
</tr>
<tr>
<td>( \bar{s}_4 )</td>
<td>( \bar{o}_{\tau(4)} = o_7 : o_7 )</td>
<td>( o_5 ) to ( o_7 )</td>
<td>+1 (( = ) ( \text{fst}(\bar{o}_7 : \bar{o}_7) - \text{lst}(o_4, o_5) - 1 ))</td>
</tr>
</tbody>
</table>

Scoring function: \( \alpha |x| \) — exponential with distance