4CSLL5 'Advanced Computational Linguistics’ Phrase Based Machine Trans

Martin Emms

November 16, 2017
Introduction

Learning the Phrase Translation Table
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 $p(\bar{o}, \bar{s})$ is:

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

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

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

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

- **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: $d(\bar{o}_{\tau(k-1)}, \bar{o}_{\tau(k)})$

- **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 $LM(s)$
The Phrase Translation Model contd

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

\[
\langle \tilde{s}, \tau \rangle_{\text{best}} = \arg \max_{\tilde{s}, \tau} p(\tilde{o}, \tau, \tilde{s})
\]

\( p \)
The Phrase Translation Model contd

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

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

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

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

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

- The distortion term $d(\bar{o}_{\tau(k-1)}, \bar{o}_{\tau(k)})$ is not learned via EM

- instead just standardly defined as an exponentially decaying function of the 'distance' $x$ between end $\bar{o}_{\tau(k-1)}$ and start of $\bar{o}_{\tau(k)}$

- in particular where $x = |fst(\bar{o}_{\tau(k)}) - lst(\bar{o}_{\tau(k-1)}) - 1|$, the $d$ term is $a^x$ for some $\alpha < 1$. 
**illustration of Distance-Based Reordering**

<table>
<thead>
<tr>
<th>phrase</th>
<th>translated to</th>
<th>change $\bar{d}<em>{\tau(i-1)}$ to $\bar{d}</em>{\tau(i)}$</th>
<th>'displacement' $\times$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{s}_1$</td>
<td>$\bar{d}_{\tau(1)} = o_1 : o_3$</td>
<td>start at beginning</td>
<td>0</td>
</tr>
<tr>
<td>$\bar{s}_2$</td>
<td>$\bar{d}_{\tau(2)} = o_6 : o_6$</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>$\bar{d}_{\tau(3)} = o_4 : o_5$</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>$\bar{d}_{\tau(4)} = o_7 : o_7$</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