Introduction

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
Phrase-Based Model

- The source sentence is segmented into phrases.
- Each source phrase is translated into an observed phrase.
- Observed phrases are reordered.

Compared to IBM Model

- Recall IBM models assumed a hidden alignment between the source $s$ and observed $o$, 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 the source $s$ and observed $o$ 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 a formula for $p(\bar{o}, \tau, \bar{s})$ as $p(\bar{o}, \tau | \bar{s}) \times p(\bar{s})$.

Example

- Task: translate a certain German 'observed' sentence into 'source' English.
- The observed sentence is 'er geht ja nicht nach hause'.
- Possible segmentation of the source $s_{1:5} = he does not go home$ and the observed $o_{1:6} = er geht ja nicht nach hause$.
- Possible segmentation of $s_{1:5}$ into $\bar{s}_{1:4}$ is $\bar{s}_1 = s_{1:1} = he$, $\bar{s}_2 = s_{2:3} = does not$, $\bar{s}_3 = s_{4:4} = go$, $\bar{s}_4 = s_{5:5} = home$.
- Possible segmentation of $o_{1:6}$ into $\bar{o}_{1:4}$ is $\bar{o}_1 = o_{1:1} = er$, $\bar{o}_2 = o_{2:2} = geht$, $\bar{o}_3 = o_{3:4} = ja nicht$, $\bar{o}_4 = o_{5:6} = nach hause$.
- 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

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

$\begin{array}{|c|c|c|c|}
\hline
\text{er} & \text{geht} & \text{ja} & \text{nicht} \\
\hline
\text{he} & \text{is} & \text{yes} & \text{not} \\
\text{he} & \text{goes} & \text{is} & \text{does not} \\
\text{he} & \text{does go} & \text{not} & \text{in} \\
\text{he} & \text{goes} & \text{does not} & \text{not at all} \\
\text{he} & \text{is not} & \text{not} & \text{do not} \\
\text{he} & \text{are not} & \text{not} & \text{do not} \\
\text{he} & \text{is not} & \text{is not} & \text{is not} \\
\hline
\end{array}$

The phrase-based translation will be built with these ingredients.

$\begin{array}{|c|c|c|c|}
\hline
\phi & \psi_1 & \psi_2 & \psi_3 \\
\hline
\text{er} & \text{he} & \text{ja nicht} & \text{go} \\
\text{geht} & \text{goes} & \text{he} & \text{he goes} \\
\text{nach} & \text{after} & \text{ja} & \text{does not} \\
\text{hause} & \text{house} & \text{nicht} & \text{not} \\
\hline
\end{array}$

Pick a phrase $\phi = \text{‘er’}$ in observed, choose ‘he’ as $\psi_1$ in source.

Pick a phrase $\phi = \text{‘ja nicht’}$ in observed, choose ‘does not’ as $\psi_2$ in source.

$\text{er geht ja nicht nach hause}$

$\text{Pick a phrase } \phi = \text{‘geht’} \text{ in observed, choose ‘go’ as } \psi_3 \text{ in source}$

$\text{er geht ja nicht nach hause}$

$\text{NB: allowed to choose } \phi \text{ phrases out of sequence; } \psi \text{ phrases chosen in sequence}$

$\text{NB: phrases may have multiple words: many-to-many translation}$
Constructing a Phrase-Based Translation

- Pick a phrase $\delta = \text{‘nach hause’}$ in observed, choose ‘home’ as $s_4$

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

- 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 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

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

(\textit{that \textendash dass})

(\textit{assumes \textendash geht davon aus})

(\textit{in the \textendash im})

Learning ctd: extract consistent phrase pairs

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

eg. \textit{(in the \textendash im) \textendash (house \textendash haus)}

\ \rightarrow \ \textit{(in the house \textendash im haus)}

Learning ctd: extract consistent phrase pairs

N-to-N cases: \textit{(will stay \textendash bleibt) \textendash (in the house \textendash im haus)}

\ \rightarrow \ \textit{(will stay in the house \textendash im haus bleibt)}
Learning ctd: extract consistent phrase pairs

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

Learning ctd: what you can’t extract

no pair (\( \overline{e} \rightarrow \text{er im} \))
no pair (\( \text{he stay} \rightarrow \overline{g} \))
because corresponding parts not adjacent

Learning ctd: extract consistent phrase pairs

N-to-N cases\(^1\): \( \text{assumes} \rightarrow \text{geht davon aus} \) + \( (\epsilon \rightarrow ,) + (\text{that} \rightarrow \text{dass}) \)
\[ \rightarrow (\text{assumes that} \rightarrow \text{geht davon aus , dass}) \]

\(^1\) here the unaligned German , is swept up; tantamount to treating it as paired with empty string \( \epsilon \)

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:
\[
\begin{align*}
tr(e|g) &= \frac{\text{count}(\overline{e}, \overline{g})}{\sum_{g'} \text{count}(\overline{e}, \overline{g'})} \\
tr(g|\overline{e}) &= \frac{\text{count}(\overline{e}, \overline{g})}{\sum_{\overline{g'}} \text{count}(\overline{e}, \overline{g'})}
\end{align*}
\]
- 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 \( tr(\bar{e}|\bar{g}) \) for \( \bar{g} = \text{den Vorschlag} \) and various English 'phrases'

| English          | \( \phi(\bar{e}|\bar{g}) \) | English                  | \( \phi(\bar{e}|\bar{g}) \) |
|------------------|-------------------------------|---------------------------|-------------------------------|
| 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                        | ...                       | ...                           |

- lexical variation (\textit{proposal} vs \textit{suggestions})
- morphological variation (\textit{proposal} vs \textit{proposals})
- included function words (\textit{the}, \textit{a}, ...)
- noise (\textit{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 \textit{if} attempts are made to limit to grammatically motivated 'linguistic' phrases, overall translation quality goes \textit{down}