4CSLL5
‘Advanced Computational Linguistics’
Phrase Based Machine Trans

Martin Emms

November 16, 2017
Introduction

Learning the Phrase Translation Table
Intro and Learning
Motivation
Motivation

- Word-Based Models translate \textit{words} as atomic units
Motivation

- Word-Based Models translate *words* as atomic units
- Phrase-Based Models translate *phrases* as atomic units
Motivation

- 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
Motivation

- **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

source

observed

he

does not

goes

home

er

geht

ja nicht

nach hause
Compared to IBM Model
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

$$p(o, a|s) \times p(s)$$
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 $\bar{s}_{1:K}$ and $\bar{o}_{1:K}$
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 $\bar{s}_{1:K}$ 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.
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 $\bar{s}_{1:K}$ 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 $p(\bar{o}, \tau, \bar{s})$ as

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

source

observed

he does not go home

er geht ja nicht nach hause
Example

\[ source \]

\[ \text{he, does not, go, home} \]

\[ observed \]

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

\[
\text{assume } s_{1:5} = \text{he does not go home} \text{ and } o_{1:6} = \text{er geht ja nicht nach hause}
\]
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
  \( \bar{s}_1 = s_{1:1} = \text{he}, \bar{s}_2 = s_{2:3} = \text{does not}, \bar{s}_3 = s_{4:4} = \text{go}, \bar{s}_4 = s_{5:5} = \text{home} \)
Example

\[ \text{assumption: } s_{1:5} = \text{he does not go home} \text{ and } o_{1:6} = \text{er geht ja nicht nach hause} \]

- possible segmentation of \( s_{1:5} \) into \( \bar{s}_{1:4} \) is
  \[ \bar{s}_1 = s_{1:1} = \text{he}, \quad \bar{s}_2 = s_{2:3} = \text{does not}, \quad \bar{s}_3 = s_{4:4} = \text{go}, \quad \bar{s}_4 = s_{5:5} = \text{home} \]

- possible segmentation of \( o_{1:6} \) into \( \bar{o}_{1:4} \) is
  \[ \bar{o}_1 = o_{1:1} = \text{er}, \quad \bar{o}_2 = o_{2:2} = \text{geht}, \quad \bar{o}_3 = o_{3:4} = \text{ja nicht}, \quad \bar{o}_4 = o_{5:6} = \text{nach hause} \]
Example

assume \( s_{1:5} = \) *he does not go home* and \( o_{1:6} = \) *er geht ja nicht nach hause*

possible segmentation of \( s_{1:5} \) into \( \bar{s}_{1:4} \) is
\[
\bar{s}_1 = s_{1:1} = \text{he}, \quad \bar{s}_2 = s_{2:3} = \text{does not}, \quad \bar{s}_3 = s_{4:4} = \text{go}, \quad \bar{s}_4 = s_{5:5} = \text{home}
\]

possible segmentation of \( o_{1:6} \) into \( \bar{o}_{1:4} \) is
\[
\bar{o}_1 = o_{1:1} = \text{er}, \quad \bar{o}_2 = o_{2:2} = \text{geht}, \quad \bar{o}_3 = o_{3:4} = \text{ja nicht}, \quad \bar{o}_4 = o_{5:6} = \text{nach hause}
\]

possible mapping \( \tau \) from \( \bar{s} \) to \( \bar{o} \) is
\[
\tau(1) = 1, \quad \tau(2) = 3, \quad \tau(3) = 2, \quad \tau(4) = 4
\]
Constructing a Phrase-Based Translation

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

  er geht ja nicht nach hause
Constructing a Phrase-Based Translation

- Assume a 'phrase-table' giving for many possible 'phrases' \( \overline{\text{e}} \) in the observed German, possible 'phrases' \( \overline{s} \) in potential source English.

- the phrase-based translation will be built with these ingredients.
Constructing a Phrase-Based Translation

- Pick a phrase $\bar{o} = 'er'$ in observed, choose 'he' as $s_1$ in source
Constructing a Phrase-Based Translation

er geht ja nicht nach hause

he does not

er ja nicht
Constructing a Phrase-Based Translation

- Pick a phrase $\bar{o} = 'ja nicht'$ in observed, choose 'does not' as $\bar{s}_2$ in source

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

- Pick a phrase $\tilde{o} = \text{`ja nicht'}$ in observed, choose 'does not' as $\tilde{s}_2$ in source.
- NB: allowed to choose $\tilde{o}$ phrases out of sequence; $\tilde{s}$ phrases chosen in sequence.
Constructing a Phrase-Based Translation

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

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

$\text{he does not}$

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

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

\[
\begin{array}{ccc}
\text{he} & \text{does not} & \text{go} \\
\text{er} & \text{geht} & \text{ja nicht} \\
\end{array}
\]

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

- Pick a phrase $\bar{o} = 'nach hause'$ in observed, choose 'home' as $\bar{s}_4$
just constructed one particular translation, could have constructed many, many others using the available phrases 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
Introduction

- 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
Introduction

◮ 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

\[
a : \text{Ger} \rightarrow \text{Eng}
\]

\[
a : \text{Eng} \rightarrow \text{Ger}
\]

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

for each training pair, *merge* these alignments
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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>assumes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>that</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>he</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>will</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stay</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>in</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>house</td>
<td></td>
<td></td>
<td></td>
<td></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

- Michael
- Assumes
- That
- He
- Will
- Stay
- In
- The
- House

- Geh
- Davon
- Aus
- Dass
- Er
- Im
- Haus
- Bleibt

- Michael geht davon aus, dass er im Haus bleibt.
Learning ctd: extract consistent phrase pairs

<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 assumes that he will stay in the house.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

obvious 1-to-N, N-to-1 cases eg:
Learning ctd: extract consistent phrase pairs

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

(\textit{that} – \textit{dass})
(\textit{assumes} – \textit{geht davon aus})
(\textit{in the} – \textit{im})
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)*
Learning ctd: extract consistent phrase pairs

N-to-N cases: (will stay – bleibt) + (in the house — im Haus) → (will stay in the house — im Haus bleibt)
Learning ctd: extract consistent phrase pairs

N-to-N cases: \((michael \rightarrow michael) + (assumes \rightarrow geht davon aus)\)  
\(\rightarrow (michael\text{ assumes} \rightarrow michael\text{ geht davon aus})\)
Learning ctd: extract consistent phrase pairs

N-to-N cases¹: \((\text{assumes} \rightarrow \text{geht davon aus}) + (\epsilon \rightarrow ,) + (\text{that} \rightarrow \text{dass})\) → \((\text{assumes that} \rightarrow \text{geht davon aus, dass})\)

¹here the unaligned German , is swept up; tantamount to treating it as paired with empty string \(\epsilon\)
Learning ctd: what you can’t extract

no pair ($\bar{e} -$ er im)
no pair ($he will stay -$ $\bar{g}$)
because corresponding parts not adjacent
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:

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

- 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 $tr(\bar{e}|\bar{g})$ 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 (*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, ...).
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
  
  \[
  \text{spass am} \rightarrow \text{fun with the}
  \]
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
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
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*