Statistical Machine Translation
Part III – Many-to-Many Alignments and Phrase-Based SMT

Alexander Fraser
CIS, LMU München

Machine Translation SS 2023
• Last time, we discussed Model 1 and Expectation Maximization
• Today we will discuss getting useful alignments for translation and a translation model
IBM Model 1

- **Generative model**: break up translation process into smaller steps
  - **IBM Model 1** only uses *lexical translation*

- Translation probability
  - for a foreign sentence \( f = (f_1, ..., f_{l_f}) \) of length \( l_f \)
  - to an English sentence \( e = (e_1, ..., e_{l_e}) \) of length \( l_e \)
  - with an alignment of each English word \( e_j \) to a foreign word \( f_i \) according to the alignment function \( a : j \rightarrow i \)

  \[
p(e, a|f) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j | f_{a(j)})
  \]

  - parameter \( \epsilon \) is a *normalization constant*
## Convergence

<table>
<thead>
<tr>
<th>$e$</th>
<th>$f$</th>
<th>initial</th>
<th>1st it.</th>
<th>2nd it.</th>
<th>3rd it.</th>
<th>...</th>
<th>final</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>das</td>
<td>0.25</td>
<td>0.5</td>
<td>0.6364</td>
<td>0.7479</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>book</td>
<td>das</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1818</td>
<td>0.1208</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>house</td>
<td>das</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1818</td>
<td>0.1313</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>the</td>
<td>buch</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1818</td>
<td>0.1208</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>book</td>
<td>buch</td>
<td>0.25</td>
<td>0.5</td>
<td>0.6364</td>
<td>0.7479</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>a</td>
<td>buch</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1818</td>
<td>0.1313</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>book</td>
<td>ein</td>
<td>0.25</td>
<td>0.5</td>
<td>0.4286</td>
<td>0.3466</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>a</td>
<td>ein</td>
<td>0.25</td>
<td>0.5</td>
<td>0.5714</td>
<td>0.6534</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>haus</td>
<td>0.25</td>
<td>0.5</td>
<td>0.4286</td>
<td>0.3466</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>house</td>
<td>haus</td>
<td>0.25</td>
<td>0.5</td>
<td>0.5714</td>
<td>0.6534</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>
Higher IBM Models

<table>
<thead>
<tr>
<th>IBM Model 1</th>
<th>lexical translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Model 2</td>
<td>adds absolute reordering model</td>
</tr>
<tr>
<td>IBM Model 3</td>
<td>adds fertility model</td>
</tr>
<tr>
<td>IBM Model 4</td>
<td>relative reordering model</td>
</tr>
<tr>
<td>IBM Model 5</td>
<td>fixes deficiency</td>
</tr>
</tbody>
</table>

- Only IBM Model 1 has global maximum
  - training of a higher IBM model builds on previous model

- Computationally biggest change in Model 3
  - trick to simplify estimation does not work anymore
  - exhaustive count collection becomes computationally too expensive
  - sampling over high probability alignments is used instead
HMM Model

- Model 4 requires local search (making small changes to an initial alignment and rescoring)
- Another popular model is the HMM model, which is similar to Model 2 except that it uses relative alignment positions (like Model 4)
- Popular because it supports inference via the forward-backward algorithm
Overcoming 1-to-N

• We'll now discuss overcoming the poor assumption behind alignment functions
Word Alignment

Given a sentence pair, which words correspond to each other?

michael  geht  davon  aus  ,  dass  er  in  haus  bleibt

michael  assumes  that  he  will  stay  in  the  house
## Word Alignment?

<table>
<thead>
<tr>
<th></th>
<th>john</th>
<th>wohnt</th>
<th>hier</th>
<th>nicht</th>
</tr>
</thead>
<tbody>
<tr>
<td>john</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>does</td>
<td></td>
<td></td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>not</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>live</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>here</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Is the English word *does* aligned to the German *wohnt* (verb) or *nicht* (negation) or neither?
Word Alignment?

How do the idioms *kicked the bucket* and *biss ins grass* match up? Outside this exceptional context, *bucket* is never a good translation for *grass*. 
Word Alignment with IBM Models

- IBM Models create a many-to-one mapping
  - words are aligned using an alignment function
  - a function may return the same value for different input
    (one-to-many mapping)
  - a function cannot return multiple values for one input
    (no many-to-one mapping)

- Real word alignments have many-to-many mappings
IBM Models: 1-to-N Assumption

• 1-to-N assumption
  • Multi-word “cepts” (words in one language translated as a unit) only allowed on target side. Source side limited to single word “cepts”.
  • Forced to create M-to-N alignments using heuristics
Symmetrizing word alignments

- *Grow* additional alignment points [Och and Ney, CompLing2003]
Symmetrizing Word Alignments

- Intersection of GIZA++ bidirectional alignments
- Grow additional alignment points [Och and Ney, CompLing2003]

Slide from Koehn 2009
Growing heuristic

grow-diag-final(e2f,f2e)
1: neighboring = {(-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1)}
2: alignment $A = \text{intersect}(e2f,f2e)$; grow-diag(); final(e2f); final(f2e);

grow-diag()
1: while new points added do
2: \hspace{1em} for all English word $e \in [1...e_n]$, foreign word $f \in [1...f_n]$, $(e, f) \in A$ do
3: \hspace{2em} for all neighboring alignment points $(e_{new}, f_{new})$ do
4: \hspace{3em} if $(e_{new} \text{ unaligned OR } f_{new} \text{ unaligned}) \text{ AND } (e_{new}, f_{new}) \in \text{union}(e2f,f2e)$ then
5: \hspace{4em} add $(e_{new}, f_{new})$ to $A$
6: \hspace{3em} end if
7: \hspace{2em} end for
8: \hspace{1em} end for
9: end while

final()
1: for all English word $e_{new} \in [1...e_n]$, foreign word $f_{new} \in [1...f_n]$ do
2: \hspace{1em} if $(e_{new} \text{ unaligned OR } f_{new} \text{ unaligned}) \text{ AND } (e_{new}, f_{new}) \in \text{union}(e2f,f2e)$ then
3: \hspace{2em} add $(e_{new}, f_{new})$ to $A$
4: \hspace{1em} end if
5: end for
Discussion

• Most state of the art SMT systems are built as I presented
• Use IBM Models to generate both:
  – one-to-many alignment
  – many-to-one alignment
• Combine these two alignments using symmetrization heuristic
  – output is a many-to-many alignment
  – used for building decoder
• Moses toolkit for implementation: www.statmt.org
  – Uses Och and Ney GIZA++ tool for Model 1, HMM, Model 4
• However, there is newer work on alignment that is interesting!
Where we have been

• We defined the overall problem and talked about evaluation

• We have now covered word alignment
  – IBM Model 1, true Expectation Maximization
  – Briefly mentioned: IBM Model 4, approximate Expectation Maximization
  – Symmetrization Heuristics (such as Grow)
    • Applied to two Viterbi alignments (typically from Model 4)
    • Results in final word alignment
Where we are going

• We will discuss the "traditional" phrase-based model (which noone actually uses, but gives a good intuition)

• Then we will define a high performance translation model (next slide set)

• Finally, we will show how to solve the search problem for this model (= decoding)
Outline

• Phrase-based translation
  – Model
  – Estimating parameters

• Decoding
• We could use IBM Model 4 in the direction $p(f | e)$, together with a language model, $p(e)$, to translate

$$\arg\max_{e} P(e | f) = \arg\max_{e} P(f | e) P(e)$$
• However, decoding using Model 4 doesn’t work well in practice
  – One strong reason is the bad 1-to-N assumption
  – Another problem would be defining the search algorithm
    • If we add additional operations to allow the English words to vary, this will be very expensive
  – Despite these problems, Model 4 decoding was briefly state of the art

• We will now define a better model...
Phrase-based translation

- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated

- Each phrase is translated into English

- Phrases are reordered
Statistical Machine Translation

- Components: Translation model, language model, decoder

foreign/English parallel text

statistical analysis

Translation Model

English text

statistical analysis

Language Model

Decoding Algorithm

Slide from Koehn 2008
Language Model

• Often a trigram language model is used for $p(e)$
  – $P(\text{the man went home}) = p(\text{the} \mid \text{START}) \ p(\text{man} \mid \text{START the}) \ p(\text{went} \mid \text{the man}) \ p(\text{home} \mid \text{man went})$

• Language models work well for comparing the grammaticality of strings of the same length
  – However, when comparing short strings with long strings they favor short strings
  – For this reason, an important component of the language model is the length bonus
    • This is a constant $> 1$ multiplied for each English word in the hypothesis
    • It makes longer strings competitive with shorter strings
Phrase-based translation model

- Major components of phrase-based model
  - phrase translation model \( \phi(f|e) \)
  - reordering model \( d \)
  - language model \( p_{LM}(e) \)

- Bayes rule
  \[
  \arg\max_e p(e|f) = \arg\max_e \phi(f|e)p(e) \\
  = \arg\max_e \phi(f|e)p_{LM}(e)\omega^{\text{length}(e)}
  \]

- Sentence \( f \) is decomposed into \( I \) phrases \( f_1^I = f_1, ..., f_I \)

- Decomposition of \( \phi(f|e) \)
  \[
  \phi(f_1^I|e_1^I) = \prod_{i=1}^{I} \phi(f_i|e_i)d(a_i - b_{i-1})
  \]
Advantages of phrase-based translation

- *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 translation table

- Phrase translations for *den Vorschlag*

| English            | $\phi(e|f)$ | English            | $\phi(e|f)$ |
|--------------------|------------|--------------------|------------|
| 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     | ...                | ...        |
How to learn the phrase translation table?

- Start with the *word alignment*:

- Collect all phrase pairs that are *consistent* with the word alignment
Consistent with word alignment

- **Consistent with the word alignment** :=
  
  phrase alignment has to *contain all alignment points* for all covered words

\[ (\overline{c}, \overline{f}) \in BP \iff \forall e_i \in \overline{c} : (e_i, f_j) \in A \rightarrow f_j \in \overline{f} \]

AND

\[ \forall f_j \in \overline{f} : (e_i, f_j) \in A \rightarrow e_i \in \overline{c} \]

Slide from Koehn 2008
Word alignment induced phrases

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)
Word alignment induced phrases

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
 bruja verde, green witch)
Word alignment induced phrases

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)

Slide from Koehn 2008
Word alignment induced phrases

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the),
(daba una bofetada a la bruja verde, slap the green witch)
Word alignment induced phrases (5)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch),
(no daba una bofetada a la bruja verde, did not slap the green witch),
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)
Probability distribution of phrase pairs

- We need a **probability distribution** $\phi(f|e)$ over the collected phrase pairs

  $\Rightarrow$ Possible **choices**
  
  - **relative frequency** of collected phrases: $\phi(f|e) = \frac{\text{count}(f,e)}{\sum_f \text{count}(f,e)}$
  
  - or, conversely $\phi(e|f)$
  
  - use **lexical translation probabilities**
Reordering

- **Monotone** translation
  - do not allow any reordering
  - worse translations

- **Limiting** reordering (to movement over max. number of words) helps

- **Distance-based** reordering cost
  - moving a foreign phrase over $n$ words: cost $z^n$

- **Lexicalized** reordering model