Machine Translation
Phrase Models

Philipp Koehn, University of Edinburgh

12 February 2009
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

Morgen fliege ich nach Kanada zur Konferenz

Tomorrow I will fly to the conference in Canada
Phrase-based translation model

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

- Bayes rule
  $$\arg\max_ep(e|f) = \arg\max_ep(f|e)p(e) = \arg\max_ep(f|e)\ p_{LM}(e)\ \Omega(f|e)$$

- Sentence $f$ is decomposed into $I$ phrases $\bar{f}^I_1 = \bar{f}_1, \ldots, \bar{f}_I$

- Decomposition of $\phi(f|e)$
  $$\phi(\bar{f}^I_1|\bar{e}^I_1) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) \ \omega^{d(\text{start}_i-\text{end}_{i-1}-1)}$$
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*:

  ![Word Alignment Diagram]

- 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

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

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

⇒ 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 \( \omega^n \)

- **Lexicalized** reordering model
Lexicalized reordering models

- Three orientation types: monotone, swap, discontinuous
- Probability $p(\text{swap}|e, f)$ depends on foreign (and English) phrase involved
Learning lexicalized reordering models

- Orientation type is *learned during phrase extractions*

- *Alignment point* to the *top left* (monotone) or *top right* (swap)?

- For more, see [Tillmann, 2003] or [Koehn et al., 2005]
Names and Numbers

- All word tokens are treated the same

- Names and numbers pose special problems
  - there are many different names and numbers
  - if input and output use different scripts, translation is not easy

- Name translation is hard
  - names may not have a properly defined spelling in non-native scripts
  - training data is not always easy to come by
  - treated as special transliteration problem
XML Markup

Er erzielte <NUMBER english='17.55'>17,55</NUMBER> Punkte.

• *Add additional translation options*
  - number translation
  - name translation

• Additional options
  - provide multiple translations
  - provide probability distribution along with translations
  - allow bypassing of provided translations