Data Intensive Linguistics — Lecture 17
Machine translation (IV): Phrase-Based Models
Philipp Koehn
9 March 2006

Phrase-based translation
Morgen fliege ich nach Kanada zur Konferenz
Tomorrow I will fly to the conference in Canada

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 can not return multiple values for one input
    (no many-to-one mapping)
• But we need many-to-many mappings
  – Intersection of GIZA++ bidirectional alignments

Growing heuristic
GROW-DIAG-FINAL(e2f,f2e):
neighboring = ((-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1))
for foreign word f = 0 ... fn
if ( e aligned with f )
  add alignment point ( e-new, f-new )
  ( e-new, f-new ) in union( e2f, f2e )
for english word e = 0 ... en
if ( e-new not aligned or f-new not aligned ) and
( e-new, f-new ) in alignment a )
  add alignment point ( e-new, f-new )
  ( e-new, f-new ) in union( e2f, f2e )
neighboring = ((-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1))
iterate until no new points added

Symmetrizing word alignments
• Intersection of GIZA++ bidirectional alignments

Phrase-based translation model
• Major components of phrase-based model
  – phrase translation model \( \phi(f|e) \)
  – reordering model \( _{\omega} \text{length}(x) \)
  – language model \( _{\omega} p_{\omega}(e) \)
• Bayes rule
  \[ \arg \max_{f} \phi(f|e) = \arg \max_{f} \phi(f|e)p(e) = \arg \max_{f} \phi(f|e)p_{\omega}(e)_{\omega} \text{length}(x) \]
• Sentence f is decomposed into l phrases \( f_{1}, \ldots, f_{l} \)
• Decomposition of \( \phi(f|e) \)
  \[ \phi(f_{i}|e) = \prod_{i=1}^{l} \phi(f_{i}|e_{i})d(a_{i} - b_{i-1}) \]
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

| English    | $\phi(e|f)$ | English    | $\phi(e|f)$ |
|------------|------------|------------|------------|
| the proposal | 0.0227     | the suggestions | 0.0114     |
| s' proposal | 0.1008     | the proposed  | 0.0114     |
| a proposal  | 0.0341     | the motion   | 0.0091     |
| the idea    | 0.0059     | the idea of  | 0.0008     |
| this proposal | 0.0027     | its proposal | 0.0068     |
| proposal    | 0.0205     | it          | 0.0068     |
| of the proposal | 0.0159     | ...         | ...        |

Consistent with word alignment

$$\mathcal{F} \ni \forall e_i \in e : (e_i, f_j) \in A \rightarrow f_j \in f$$
$$\forall f_j \in f : (e_i, f_j) \in A \rightarrow e_i \in e$$

Phrase translation table

| English       | $\phi(e|f)$ | English       | $\phi(e|f)$ |
|---------------|------------|---------------|------------|
| the proposal  | 0.6227     | its proposal  | 0.0205     |
| proposal      | 0.0159     | the proposals | 0.0068     |

Consistent with word alignment

Phrase alignment has to contain all alignment points for all covered words

$$(\pi, \mathcal{F}) \ni \mathcal{F} \ni \forall e_i \in \mathcal{T} : (e_i, f_j) \in A \rightarrow f_j \in \mathcal{F}$$
$$\forall f_j \in \mathcal{F} : (e_i, f_j) \in A \rightarrow e_i \in \mathcal{T}$$

How to learn the phrase translation table?

- Start with the word alignment:

```
Maria no daba una bofetada a la bruja verde.
```

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

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

Word alignment induced phrases

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

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

(Maria, Mary), (no, did not), (daba, 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).

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

(Maria no daba una bofetada a la bruja verde, Mary did not slap 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 \( \phi(T|F) \) over the collected phrase pairs
  \[ \Rightarrow \text{Possible choices} \]
  - relative frequency of collected phrases: \( \phi(T|F) = \frac{\text{count}(T|F)}{\sum_{T'} \text{count}(T'|F)} \)
  - or, conversely \( \phi(F|T) \)
  - use lexical translation probabilities

Knowledge sources

- Many different knowledge sources useful
  - language model
  - reordering (distortion) model
  - phrase translation model
  - word translation model
  - word count
  - phrase count
  - drop word feature
  - phrase pair frequency
  - additional language models
  - additional features

Philipp Koehn  DIL Lecture 17  9 March 2006

Reordering

- **Monotone** translation
  - do not allow any reordering
  \[ \Rightarrow \text{worse translations} \]
- **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

Log-linear models

- IBM Models provided mathematical justification for factoring components together
  \[ p_{\text{LM}} \times p_{\text{TM}} \times p_{\text{PD}} \]
- These may be weighted
  \[ p_{\text{LM}}^\lambda \times p_{\text{TM}}^\lambda \times p_{\text{PD}}^\lambda \]
- Many components \( p_i \) with weights \( \lambda_i \)
  \[ \Rightarrow \prod p_i^\lambda = \exp(\sum \lambda_i \log(p_i)) \]
  \[ \Rightarrow \log \prod p_i^\lambda = \sum \lambda_i \log(p_i) \]

Set feature weights

- Contribution of components \( p_i \) determined by weight \( \lambda_i \)
- Methods
  - manual setting of weights: try a few, take best
  - automate this process
- Learn weights
  - set aside a development corpus
  - set the weights, so that optimal translation performance on this development corpus is achieved
  - requires automatic scoring method (e.g., BLEU)
Learn feature weights

- Model
- generate n-best list
- change feature weights
- score translations
- find feature weights that move up good translations

Discriminative vs. generative models

- Generative models
  - translation process is broken down to steps
  - each step is modeled by a probability distribution
  - each probability distribution is estimated from the data by maximum likelihood

- Discriminative models
  - model consist of a number of features (e.g. the language model score)
  - each feature has a weight, measuring its value for judging a translation as correct
  - feature weights are optimized on development data, so that the system output matches correct translations as close as possible

Learning task

- Task: find weights, so that feature vector of the correct translations ranked first

Methods to adjust feature weights

- Maximum entropy [Och and Ney, ACL2002]
  - match expectation of feature values of model and data

- Minimum error rate training [Och, ACL2003]
  - try to rank best translations first in n-best list
  - can be adapted for various error metrics, even BLEU

- Ordinal regression [Shen et al., NAACL2004]
  - separate k worst from the k best translations