An Unsupervised Model for Joint Phrase Alignment and Extraction

Graham Neubig¹,², Taro Watanabe², Eiichiro Sumita², Shinsuke Mori¹, Tatsuya Kawahara¹

¹Graduate School of Informatics, Kyoto University
²National Institute of Information and Communication Technology
Phrase Table Construction
The Phrase Table

- The most important element of phrase-based SMT
- Consists of scored bilingual phrase pairs

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>it</td>
<td>0.05 0.20 0.005 1</td>
</tr>
<tr>
<td>le admettre</td>
<td>admit it</td>
<td>1.0 1.0 1e-05 1</td>
</tr>
<tr>
<td>admettre</td>
<td>admit</td>
<td>0.4 0.5 0.02 1</td>
</tr>
</tbody>
</table>

- Usually learned from a parallel corpus aligned at the sentence level
  
  → Phrases must be aligned
Traditional Phrase Table Construction: 1-to-1 Alignment, Combination, Extraction

+ Generally quite effective, default for Moses
- Complicated, with lots of heuristics
- Does not directly acquire phrases, which are the final goal of alignment
- Phrase table is exhaustively extracted and thus large
Previous Work: Many-to-Many Alignment

- Significant recent research on many-to-many alignment [Zhang+ 08, DeNero+ 08, Blunsom+ 10]
  + Model is simplified, gains in accuracy
- Short phrases are aligned, then combined into longer phrases during the extraction step
- Some issues still remain
  - Large phrase table, heuristics, no direct modeling of extracted phrases
Proposed Model for Joint Phrase Alignment and Extraction

- Phrases of multiple granularities directly modeled
  - No mismatch between alignment goal and final goal
  - Completely probabilistic model, no heuristics
  - Competitive accuracy, smaller phrase table
- Uses a hierarchical model for Inversion Transduction Grammars (ITG)
Phrasal Inversion
Transduction Grammars
(Previous Work)
Inversion Transduction Grammar (ITG)

• Like a CFG over two languages
  • Have non-terminals for regular and inverted productions
  • One pre-terminal
  • Terminals specifying phrase pairs

```
                reg
                   
      term          term
              /     \      /     \
             I/il me  hate/coûte

English  I hate  French il me coûte

                inv
                   
      term          term
              /     \      /     \
             admit/admettre  it/le

English admit it  French le admettre
```
Biparsing-based Alignment with ITGs

- Non/pre-terminal distribution $P_x$, and phrase distribution $P_t$

Sentence Pair $<e,f>$

Viterbi parsing and sampling both possible in $O(n^6)$
Learning Phrasal ITGs with Blocked Gibbs Sampling [Blunsom+ 10]

1) Choose sentence to sample

2) Subtract current $d_i$

3) Perform biparsing using $P_x$ and $P_t$

4) Add new $d_i$

5) Replace $d_i$ in the corpus

... and get a new sample for $d_i$
Calculating Probabilities given Counts

- Adapt Bayesian approach, assume that probabilities were generated from Pitman-Yor process, Dirichlet distribution

\[
P_t \sim PY \left(d, \theta, P_{\text{base}}\right)
\]
\[
P_x \sim \text{Dirichlet} \left(\alpha = 1, 1/3\right)
\]

- Marginal probabilities can be calculated (in example, ignoring \(d\) for the PY process)

\[
P_t(f,e) = \frac{c_t(f,e) + \theta_t P_{\text{base}}(f,e)}{\sum_{f,e} c_t(f,e) + \theta_t}
\]
\[
P_x(x) = \frac{c_x(x) + \alpha_x/3}{\sum_x c_x(x) + \alpha_x}
\]
Base Measure

\[ P_t(f, e) = \frac{c_t(f, e) + \theta_t P_{base}(f, e)}{\sum_{f, e} c_t(f, e) + \theta_t} \]

- \( P_{base} \) has an effect of smoothing probabilities
  - Particularly for low frequency pairs
- To bias towards good phrase pairs, use geometric mean of word-based Model 1 probabilities [DeNero+ 08]

\[ P_{base}(e, f) = \left( P_{m1}(f | e) P_{uni}(e) P_{m1}(e | f) P_{uni}(f) \right)^{\frac{1}{2}} \]

- Good word match in both directions = good phrase match
Calculating Counts given Derivations

- Elements generated from each distribution $P_x$ and $P_t$ added to the counts used to calculate the probabilities

\[
\begin{align*}
&c_x(\text{reg}) += 3 \\
&c_x(\text{inv}) += 1 \\
&c_x(\text{term}) += 5
\end{align*}
\]

- Problem: only minimal phrases are added

→ Must still heuristically combine into multiple granularities
Joint Phrase Alignment and Extraction
(Our Work)
Basic Idea

- Generative story in reverse order

- Traditional ITG Model:
  - Generate branches (reordering structure) from $P_x$
  - Generate leaves (phrase pairs) from $P_t$

- Proposed ITG Model:
  - From the top, try to generate phrase pair from $P_t$
  - Divide and conquer using $P_x$ to handle sparsity
Derivation in the Proposed Model

- Phrases of many granularities generated from $P_t$, added to $c_t$

$$P_t(base)$$

$$c_t(i 	ext{ hate to admit } it/il 	ext{ me coûte de le admettre})++$$

$$c_x(reg) += 3$$
$$c_x(inv) += 1$$
$$c_x(base) += 1$$

- No extraction needed, as multiple granularities are included!
Recursive Base Measure

- Previous work: high prob. words = high prob. phrases
- **Proposed**: Build new phrase pairs by combining existing phrase pairs in $P_{dac}$ ("divide-and-conquer")

$P_t(l/il me) \leftarrow \text{high}$

$P_t(hate/coûte) \leftarrow \text{high}$

$P_{dac}(I hate/il me coûte) \leftarrow \text{high}$

$$P_t(f, e) = \frac{c_t(f, e) + \alpha_t P_{dac}(f, e)}{\sum_{f, e} c_t(f, e) + \alpha_t}$$

- High probability sub-phrases $\rightarrow$ high probability phrases
- $P_t$ is included in $P_{dac}$, $P_{dac}$ is included in $P_t$
Details of $P_{\text{dac}}$

- Choose from $P_x$ one of three patterns for $P_{\text{dac}}$, like ITG

**Regular:** $P_x(\text{reg}) \times P_t(\text{l/il me}) \times P_t(\text{hate/coûte}) \rightarrow I\text{ hate/il me coûte}$

**Inverted:** $P_x(\text{inv}) \times P_t(\text{admit/admettre}) \times P_t(\text{it/le}) \rightarrow \text{admit it/le admettre}$

**Base:** $P_x(\text{base}) \times P_{\text{base}}(\text{hate/coûte}) \rightarrow \text{hate/coûte}$

- $P_{\text{base}}$ is the same as before
Phrase Extraction

- **Traditional Heuristics:**
  Exhaustively combine and count all neighboring phrases
  - $O(n^2)$ phrases per sent.

- **Model Probabilities:**
  Calculate phrase table from model probabilities where $c(e,f) \geq 1$
  - $O(n)$ phrases per sent.

Phrase Table Scores

- $P(e|f) = \frac{c(e,f)}{c(f)}$
- $P(f|e) = \frac{c(e,f)}{c(e)}$

Phrase Table Scores

- $P(e|f) = \frac{P_t(e,f)}{P_t(f)}$
- $P(f|e) = \frac{P_t(e,f)}{P_t(e)}$
Experiments
## Tasks/Data

- 4 Languages, 2 tasks: es-en, de-en, fr-en, ja-en
  - de-en, es-en, fr-en: WMT10 news-commentary
  - ja-en: NTCIR08 patent translation
- Data was lowercased, tokenized, and sentences of length 40 and under were used

<table>
<thead>
<tr>
<th></th>
<th>WMT</th>
<th>NTCIR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>de</td>
<td>es</td>
</tr>
<tr>
<td>TM</td>
<td>1.85M</td>
<td>1.82M</td>
</tr>
<tr>
<td>LM</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tune</td>
<td>47.2k</td>
<td>52.6k</td>
</tr>
<tr>
<td>Test</td>
<td>62.7k</td>
<td>68.1k</td>
</tr>
</tbody>
</table>
Setting

- Used Moses as a decoder
- Evaluated using BLEU score
- **3 Alignment Methods:**
  - GIZA++ and *grow-diag-final-and* heuristic
  - Traditional ITG model (FLAT)
  - Proposed ITG model (HIER)
- **2 Phrase Extraction Methods:**
  - Heuristic phrase extraction
  - Using the model probabilities $P_t$
Results

- GIZA++ uses heuristic extraction, others use model probabilities
- Same accuracy as GIZA++, phrase table smaller
- Higher accuracy than FLAT (when using model probs.)
Phrase Table: Heuristic Extraction vs. Model Probabilities

- HIER + Model Probabilities has competitive accuracy, smaller table size
Conclusion

• Used a **hierarchical model** to include phrases of multiple granularities in the alignment process

• Able to achieve competitive accuracy **directly using model probabilities** in the phrase table

• Future work:
  • Expansion to tree-based translation
  • Further refinement of modeling and search techniques

• Software is released open source:

  pialign – Phrasal ITG Aligner
  [http://www.phontron.com/pialign](http://www.phontron.com/pialign)
Thank You!