Scalable Inference and Training of Context-Rich Syntactic Translation Models

Michel Galley, Jonathan Graehl, Keven Knight, Daniel Marcu, Steve DeNeefe Wei Wang and Ignacio Thayar

Presentation by: Nadir Durrani
GHKM : What’s in a Translation Rule? (Recap)

• Given a triple (f, π,a)
  – A source side sentence \( \rightarrow f \)
  – A target-side parsed tree \( \rightarrow \pi \)
  – An alignment between f and leaves of \( \pi \rightarrow a \)

• A process of transforming \( \pi \) into f

• A minimal set of syntactically motivated transformation rules that explain human translation
Contributions of this paper

• Obtain multi-level rules
  – Acquire rules of arbitrary size that condition on more syntactic context
  – Multiple interpretation how unaligned words are accounted in a derivation

• Probability models for multi-level transfer rules
  – Assigning probabilities to very large rule sets
Rule Extraction Algorithm

• Compute a set of frontier nodes $F$ of the alignment graph

• For each $n \in F$, compute the minimal frontier graph rooted at $n$
Rule Extraction Algorithm
Rule Extraction Algorithm

- Compute a set of frontier nodes $F$ of the alignment graph
- A 3 step process – for each node $n \in G$ (direct Graph):
  - Label with its span
  - Label with its compliment span
  - Decide whether $n \in F$
Step-1 : Label with Span

Span : Indexes of first and last words reachable by a node n
Step-II : Label with Compliment Span

Compliment Span (n) = Compliment Span (Parent (n)) + Span (Siblings (n))

Compliment Span (root) = ∅
Step-II : Label with Compliment Span

\[
\text{comp\_span (VP)} = \\
\varnothing + 1-2 + 9 = \\
1-2,9
\]
Step-II : Label with Compliment Span

\[ \text{comp}_\text{span}(\text{VP}) = 1-3,9 + 7-8 = 1-3, 7-9 \]
Computing Frontier Set

• A node n is in frontier set iff compliment_span (n) ∩ closure (span (n)) = ∅

• Closure (span (n)) = Shortest contiguous span which is superset of span(n)
  - Example closure {2,3,5,7} = {2,3,4,5,6,7}
Computing Frontier Set

The diagram represents a syntactic structure of a sentence, with nodes indicating words and phrases. The numbers indicate the span of the text covered by each node.
Rule Extraction Algorithm

- Compute a set of frontier nodes $F$ of the alignment graph
- Compute the minimal frontier graph for all nodes in frontier set
Rule Extraction Algorithm

- Compute a set of frontier nodes F of the alignment graph
- Compute the minimal frontier graph for all nodes in frontier set

Algorithm:
- For each node \( n \in F \)
  - Expand \( n \) then as long as \( n' \notin F \) expand \( n' \)
  - if \( n' \in F \)
    - Replace \( n' \) by a variable \( x_i \)
Computing Frontier Set
Computing Frontier Set
Computing Frontier Set
Tree to String Transducers

\[
S(x_0:NP, x_1:VP, x_2:.) \rightarrow x_0, x_1, x_2
\]

\[
NP(x_0:DT, CD(7), NNS(people)) \rightarrow x_0, 7人
\]

\[
VP(VBG(coming), PP(IN(from), x_0:NP)) \rightarrow 来自, x_0
\]
Minimal Derivation Corresponding to Example

(a) $S(x_0:\text{NP}, x_1:\text{VP}, x_2:.) \rightarrow x_0, x_1, x_2$
(b) $\text{NP}(x_0:\text{DT}, \text{CD}(7), \text{NNS}(\text{people})) \rightarrow x_0, 7\text{人}$
(c) $\text{DT}(\text{these}) \rightarrow \text{这}$
(d) $\text{VP}(x_0:\text{VBP}, x_1:\text{NP}) \rightarrow x_0, x_1$
(e) $\text{VBP}(\text{include}) \rightarrow \text{中包括}$
(f) $\text{NP}(x_0:\text{NP}, x_1:\text{VP}) \rightarrow x_1, \text{的}, x_0$
(g) $\text{NP}(x_0:\text{NNS}) \rightarrow x_0$
(h) $\text{NNS}(\text{astronauts}) \rightarrow \text{宇航, 员}$
(i) $\text{VP}(\text{VBG}(\text{coming}), \text{PP}(\text{IN}(\text{from}), x_0:\text{NP})) \rightarrow \text{来自}, x_0$
(j) $\text{NP}(x_0:\text{NNP}) \rightarrow x_0$
(k) $\text{NNP}(\text{France}) \rightarrow \text{法国}$
(l) $.(.) \rightarrow . $
Acquiring Multi-level Rules

• GHKM : Extract minimal rules
  – Unique derivation for G
  – Rules defined over G cannot be decomposed further induced by the same graph

• This work: Extract multi-level rules
  – Multiple derivations per triple
  – Composition of 2 or more minimal rules to form larger rules
Example

NP(DT(these), CD(7), NNS(people)) → 这，7人
Example

NP(DT(these), CD(7), NNS(people)) → 这, 7人
Multiple Interpretations of Unaligned Words

• Highly frequent phenomenon in Chinese-English
  – 24.1% of Chinese words in 179 million word are unaligned
  – 84.8% of Chinese sentences contain at least 1 unaligned word

• GHKM : Extract minimal rules
  – Attach unaligned words with certain constituent of \( \pi \)
Example

- Heuristic: Attach unaligned words with highest attachment
Multiple Interpretations of Unaligned Words

• This Work
  – No prior assumption about “correct” way of assigning unaligned words to a constituent
  – Consider all possible derivations that are consistent with G
  – Use corpus evidence find more probable unaligned word attachments
6 Minimal Derivations for the Working Example
Representing Derivations as Forest

• Rather than enumerating all possible derivation
  represent as derivation forest
  – Time and space efficient

• For each derivation each unaligned item appears only
  once in the rules of that derivation
  – To avoid biased estimates by disproportional representation
Representing Derivations as Forest
Derivation Building and Rule Extraction Algo

• Preprocessing Step
  – Assign spans
  – Assign complement spans
  – Compute frontier set $F$
  – Extract minimal rules

• Each $n \in F$ has $q_o$ (open queue) and $q_c$ (closed queue) of rules
  – $q_o$ is initialized with minimal rules for each node $n \in F$

• For each node $n \in F$
  – Pick the smallest rule ‘r’ from $q_o$
  – For each variable of ‘r’ discover new rules by composition
  – If $q_o$ becomes empty or threshold on rule size or number of rules in $q_c$ is reached
    • Connect new OR-node to all rules extracted for n
    • Add to or-dforest – table to store OR-nodes with format $[x, y, c]$
Contributions of this paper

• Obtain multi-level rules
  – Acquire rules of arbitrary size that condition on more syntactic context
  – Multiple interpretation how unaligned words are accounted in a derivation

• Probability models for multi-level transfer rules
  – Assigning probabilities to very large rule sets
Probability Models

- Using noisy-channel approach

\[
\hat{e} = \arg \max_{e \in E} \left\{ Pr(e) \cdot Pr(f|e) \right\}
\]

Monolingual Language Model  Translation Model

- Incorporating dependencies on target-side syntax

\[
\hat{e} = \arg \max_{e \in E} \left\{ Pr(e) \cdot \sum_{\pi \in \tau(e)} Pr(f|\pi) \cdot Pr(\pi|e) \right\}
\]

Syntax Based Translation Model  Syntactic Parsing Model

\( T(e) \) is set of all target-trees that yield \( e \)
Syntax Based Translation Model

\[
Pr(f|\pi) = \frac{1}{|\Lambda|} \sum_{\theta_i \in \Theta} \prod_{r_j \in \theta_i} p(rhs(r_j)|lhs(r_j))
\]

- \( \Theta \) is set of all derivations constructible from \( G = (\pi,f,a) \)

- A derivation \( \theta_i = r_1 \circ \ldots \circ r_n \)
  - Independence assumption

- \( \Lambda \) is set of all sub-tree decompositions of corresponding to derivations in \( \Theta \)
  - Normalization factor to keep the distribution tight i.e. sum to 1 over all strings \( f_i \in F \) derivable from \( \pi \)
Example

\[ p(rh_{s}(r)|lh_{s}(r)) = \frac{f(r)}{\sum_{r':lh_{s}(r')=lh_{s}(r)} f(r')} \]
Example

\[ (\pi, f_1, a_1): \]
\[
\begin{array}{ccc}
  & X & \\
  & \downarrow & \\
  a & b & c \\
  \downarrow & \downarrow & \downarrow \\
  a' & b' & c' \\
\end{array}
\]

\[
(\pi, f_2, a_2): \]
\[
\begin{array}{ccc}
  & X & \\
  & \downarrow & \\
  a & b & c \\
  \downarrow & \downarrow & \downarrow \\
  b' & a' & c' \\
\end{array}
\]

\[
r_1: \ X(a, Y(b, c)) \rightarrow a', b', c' \quad p_1 = \frac{1}{2} = 0.5
\]

\[
r_2: \ X(a, Y(b, c)) \rightarrow b', a', c' \quad p_2 = \frac{1}{2} = 0.5
\]

\[
r_3: \ X(a, x_0: Y) \rightarrow a', x_0 \quad p_3 = 1
\]

\[
r_4: \ Y(b, c) \rightarrow b', c' \quad p_4 = 1
\]
Example

Total probability mass distributed across two source strings \(a, b, c\) and \(a', b', c'\)

\[
= p(a', b', c' \mid \pi) + p(b', a', c') = [p_1 + (p_3 \cdot p_4)] + [p_2] = 2
\]
Problems with Relative Frequency Estimator

\[ p(rhs(r) | lhs(r)) = \frac{f(r)}{\sum_{r': lhs(r') = lhs(r)} f(r')} \]

- Biased estimates when extracting only minimal rules

\[ \Lambda = 2 \]
\[ p(a'b'c' | \pi) = 1/2 \quad [p3 \cdot p4] = 0.5 \]
\[ p3 = 99/99 = 1 \]
\[ p4 = 99/99 = 1 \]
\[ p(b'a'c' | \pi) = 1/2 \quad [p2] = 0.5 \]
\[ p2 = 1/1 \]
Problems with Relative Frequency Estimator

$(\pi, f_1, a_1)$:
99 times

$(\pi, f_2, a_2)$:

\[ r_1: \ X(a, Y(b, c)) \rightarrow a', b', c' \]
\[ r_2: \ X(a, Y(b, c)) \rightarrow b', a', c' \]
\[ r_3: \ X(a, x_0; Y) \rightarrow a', x_0 \]
\[ r_4: \ Y(b, c) \rightarrow b', c' \]

$\Lambda = 2$

\[ p(a'b'c') = 1/2 \ [p1 + p3 \cdot p4] = 0.995 \]
\[ p1 = 99/100 = 0.99 \]
\[ p3 = 99/99 = 1 \quad p4 = 99/99 = 1 \]
\[ p(b'a'c') = 1/2 \ [p2] = 0.005 \]
\[ p2 = 1/100 = 0.01 \]

Correct Estimate
\[ p(a'b'c') = 0.99 \]
\[ p(b'a'c') = 0.1 \]

Minimum Rules
\[ p(a'b'c') = 0.5 \]
\[ p(b'a'c') = 0.5 \]

All Rules
\[ p(a'b'c') = 0.995 \]
\[ p(b'a'c') = 0.05 \]
Joint Model Conditioned on Root

\[ p(r | \text{root}(r)) = \frac{f(r)}{\sum_{r' : \text{root}(r') = \text{root}(r)} f(r')} \]

\[(\pi, f_1, a_1): \]
\[
\begin{array}{c}
X \\
\hspace{1cm} Y \\
\hspace{2cm} a \\
\hspace{3cm} b \\
\hspace{4cm} c \\
\hspace{5cm} a' \\
\hspace{6cm} b' \\
\hspace{7cm} c' \\
\end{array}
\]

99 times

\[(\pi, f_2, a_2): \]
\[
\begin{array}{c}
X \\
\hspace{1cm} Y \\
\hspace{2cm} a \\
\hspace{3cm} b \\
\hspace{4cm} c \\
\hspace{5cm} a' \\
\hspace{6cm} b' \\
\hspace{7cm} c' \\
\end{array}
\]

once

\[ r_1: X(a, Y(b, c)) \rightarrow a', b', c' \quad p(a'b'c' | \pi) = 1/2 \quad p_3 = 99/100 = 1 \quad p_4 = 99/99 = 1 \]

\[ r_2: X(a, Y(b, c)) \rightarrow b', a', c' \]

\[ r_3: X(a, x_0: Y) \rightarrow a', x_0 \]

\[ r_4: Y(b, c) \rightarrow b', c' \]

Correct Estimate
\[ p(a'b'c') = 0.99 \]
\[ p(b'a'c') = 0.1 \]
Joint Model Conditioned on Root

\[ p(r | \text{root}(r)) = \frac{f(r)}{\sum_{r': \text{root}(r') = \text{root}(r)} f(r')} \]

(\(\pi, f_1, a_1\)):
- \(99\) times
  - \(r_1: X(a, Y(b, c)) \rightarrow a', b', c'\)
  - \(r_2: X(a, Y(b, c)) \rightarrow b', a', c'\)
  - \(r_3: X(a, x_0; Y) \rightarrow a', x_0\)
  - \(r_4: Y(b, c) \rightarrow b', c'\)

(\(\pi, f_2, a_2\)):
- \(\text{once}\)
  - \(p(a'b'c' | \pi) = \frac{1}{2} [p1 + p3 \cdot p4] = 0.99\)
  - \(p1 = \frac{99}{100}\)
  - \(p3 = \frac{99}{100}\)
  - \(p4 = \frac{99}{99}\)

Correct Estimate
- \(p(a'b'c') = 0.99\)
- \(p(b'a'c') = 0.1\)
EM Training

• Which derivation in the derivation forest is true?
  – Score each derivation with its rule probabilities and find the most likely derivation

• How do we get good rules?
  – Collect the rule counts from the most likely derivation

• Chicken or the Egg problem – Calls for EM training
EM Training

Algorithm

1. Initialize each derivation with uniform rule probabilities
2. Score each derivation \( \theta_i \in \Theta \) with rule probabilities
3. Normalize to find probability \( p_i \) of each derivation
4. Collect the weighted rule counts from each derivation \( \theta_i \) with weight \( p_i \)
5. Normalize the rule counts to obtain new rule probabilities
6. Repeat 2–5 until converge
Evaluation

• Three models $C_m$, $C_3$ and $C_4$ (minimal derivation, composed rules with 3 and 4 internal nodes in lhs respectively)

• NIST 2002 54 million word English-Chinese corpus

• 1 best derivation per sentence pair based on GIZA alignments (Figure 4)
Evaluation

<table>
<thead>
<tr>
<th>Language</th>
<th>Syntactic</th>
<th>AlTemp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic-to-English</td>
<td>40.2</td>
<td>46.6</td>
</tr>
<tr>
<td>Chinese-to-English</td>
<td>24.3</td>
<td>30.7</td>
</tr>
</tbody>
</table>

Table 5: BLEU-4 scores for the 2005 NIST test set.

<table>
<thead>
<tr>
<th>Language</th>
<th>$C_m$</th>
<th>$C_3$</th>
<th>$C_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese-to-English</td>
<td>24.47</td>
<td>27.42</td>
<td>28.1</td>
</tr>
</tbody>
</table>

Table 6: BLEU-4 scores for the 2002 NIST test set, with rules of increasing sizes.
Conclusion

• Acquire larger rules – condition on more syntactic context
  – 3.63 BLEU point gain over baseline minimal rules system

• Using multiple derivations including multiple interpretations of unaligned words in derivation forest

• Probability models to score multi-level transfer rules