Synchronous Binarization for Machine Translation
Hao Zhang (University of Rochester), Liang Huang (University of Pennsylvania),
Daniel Gildea (University of Rochester), Kevin Knight (University of Southern California)

Introduction
- Synchronous grammars are the framework of syntax-based machine translation.
- Both alignment and decoding are parsing problems in this framework.
- Factorization of synchronous grammars can drastically reduce parsing complexity.
- Binarization of Synchronous CFG (SCFG) is the focus of this work.

Synchronous CFG vs. CFG

<table>
<thead>
<tr>
<th>CFG</th>
<th>SCFG</th>
</tr>
</thead>
<tbody>
<tr>
<td>terminals</td>
<td>words</td>
</tr>
<tr>
<td>nonterminals</td>
<td>phrases</td>
</tr>
<tr>
<td>productions</td>
<td>sequences</td>
</tr>
<tr>
<td>X → X1 X2 X3</td>
<td>X → X1</td>
</tr>
<tr>
<td>parsing</td>
<td>over spans</td>
</tr>
</tbody>
</table>

Parsing Complexity of n-ary SCFG
Assume the sentence length (on either side) is O(N).

- Naïve parsing is O(N^{n+1}), by enumerating at most 2(n + 1) split points into the two sentences.
- O(N^{n+1+1}) = O(N^{n+2}) is also doable, if we go from left to right on one side and pick up one pebble at a time, allowing discontinuous spans on the other side.
- But even if the strategy allows discontinuous spans on both sides, it is O(N^{n+T}), (Satta and Peserico, 2005), because there exist hard-to-parse permutations out of n!
- When n = 2, the parsing complexity O(N^2) is the theoretical lower bound for synchronous parsing using tabular methods.

Synchronous Binarization Maintains Continuity

An Actual Example: Before Binarization

ADJP → x0:RB x1:IN x2:PP

The left hand side is an English tree fragment. The right hand side is a Chinese string including coindexed variables referring to the translations of the left-hand side nonterminals.

An Actual Example: After Binarization

ADJP → T_{859} (1), T_{859} (1)
T_{859} → V_1 (1) V_2 (2), V_1 (1) V_2 (2)
V_1 → RB (1), RB (1) fuse
V_2 → resp. for the NN (1) V_3 (2), V_3 (2) NN (1)
V_3 → PP (1), PP (1) de

T_{859} (1) is a symbolic representation of the left-hand side tree in the original rule. We can reconstruct the original rule from these rules.

How Many Rules Are Binarizable?

The solid line curve represents the distribution of all rules against permutation lengths. The dashed line stairs indicate the percentage of non-binarizable rules in our initial rule set while the dotted-line denotes that percentage among all permutations. We have 50,879,242 rules.

Faster and More Accurate Decoding

The Highest BLEU Score

system | bleu
--- | ---
monolingual binarization | 36.25
synchronous binarization | 38.44
alignment-template system | 37.00

Conclusion
- We reduced the SCFG binarization problem to the problem of hierarchical binarization of permutations and devised a linear time algorithm.
- The majority of syntactic reorderings, at least between languages like English and Chinese, can be efficiently decomposed into hierarchical binary reorderings.
- From a modeling perspective, on the other hand, it is beneficial to start with a richer representation.
- As a result, decoding with n-gram models can be fast and accurate, making it possible for our syntax-based system to overtake a comparable phrase-based system in BLEU score.