Hierarchical and Syntax-Based Translation

David Chiang
Australia is one of the few countries that have diplomatic relations with North Korea.

Human translation:
Australia is one of the few countries that have diplomatic relations with North Korea.
Australia is with North Korea is diplomatic relations is one of the few countries.

澳洲是与北韩有邦交的少数国家之一

one of the few countries
Phrases of phrases

与 北 韩 有 邦 交

diplomatic relations with North Korea have
Phrases of phrases

与北韩有邦交

have diplomatic relations with North Korea
Phrases of phrases

have 与 有

with
Australia is with North Korea have diplomatic relations with few countries of the few countries of the minority of countries with China.

Australia is with North Korea have diplomatic relations with few countries of the few countries of the minority of countries with China.
few countries have diplomatic relations with North Korea is one of Australia is the few countries that have diplomatic relations with North Korea.
Australia is one of the few countries that have diplomatic relations with North Korea. 之ー

Australia is one of the few countries that have diplomatic relations with North Korea.
Definition
Context-free grammars

\[ S \rightarrow NP \ VP \]
\[ NP \rightarrow I \]
\[ NP \rightarrow \text{the box} \]
\[ VP \rightarrow V \ NP \]
\[ V \rightarrow \text{open} \]
Context-free grammars

\[
S \rightarrow NP \text{ VP} \\
NP \rightarrow \text{watashi wa} \\
NP \rightarrow \text{hako wo} \\
VP \rightarrow NP \text{ V} \\
V \rightarrow \text{akemasu}
\]
Synchronous CFGs

\[
\begin{align*}
S & \rightarrow NP \ VP \\
S & \rightarrow NP \ VP \\
NP & \rightarrow I \\
NP & \rightarrow \text{the box} \\
NP & \rightarrow \text{watashi wa} \\
NP & \rightarrow \text{hako wo} \\
VP & \rightarrow V \ NP \\
VP & \rightarrow NP \ V \\
V & \rightarrow \text{open} \\
V & \rightarrow \text{akemasu}
\end{align*}
\]
Synchronous CFGs

\[ S \rightarrow \text{NP}_1 \text{VP}_2, \text{NP}_1 \text{VP}_2 \]

\[ \text{NP} \rightarrow \text{I, watashi wa} \]

\[ \text{NP} \rightarrow \text{the box, hako wo} \]

\[ \text{VP} \rightarrow V_1 \text{NP}_2, \text{NP}_2 V_1 \]

\[ V \rightarrow \text{open, akemasu} \]
Synchronous CFGs

NP
VP
NP
VP
NP
VP
NP
VP

I
open the box
watashi wa
hako wo
akemasu

15
Limitations of synchronous CFGs

(English and French tree structures are shown.)
Synchronous tree substitution grammars

S
   NP₁  VP
   │    │
   V    NP₂
misses

S
   NP₂  VP
   │    │
   V    PP
manque P
   à    NP₁

NP  |  NP
John | Jean
NP  |  NP
Mary | Marie
Synchronous tree substitution grammars

S
  NP₁  VP
    John  V  NP₂
        misses  Mary

S
  NP₂  VP
    Marie  V  PP
        manque  à  NP₁
            Jean
Properties
Chomsky normal form

\[ X \rightarrow Y Z \]

\[ X \rightarrow a \]
Chomsky normal form

A → B C D E F

rank 5
Chomsky normal form

A \rightarrow [[[B C] D] E] F \quad \text{rank 5}

A \rightarrow V1 F

V1 \rightarrow V2 E

V2 \rightarrow V3 D

V3 \rightarrow B C \quad \text{rank 2}
A hierarchy of synchronous CFGs

\[ 1\text{-CFG} \subsetneq 2\text{-CFG} = 3\text{-CFG} = 4\text{-CFG} = \ldots \]

\[ 1\text{-SCFG} \subsetneq 2\text{-SCFG} = 3\text{-SCFG} \subsetneq 4\text{-SCFG} \subsetneq \ldots \]
\[ \therefore \therefore \]

ITG
(Wu, 1997)
Synchronous CNF?

\[ A \rightarrow (B_1 \ C_2 \ D_3 , \ C_2 \ D_3 \ B_1) \]

rank 3
Synchronous CNF?

\[ A \rightarrow (B_1 [C_2 \ D_3], [C_2 \ D_3] \ B_1) \quad \text{rank 3} \]

\[ A \rightarrow (B_1 \ V_{12}, \ V_{12} \ B_1) \quad \text{rank 2} \]

\[ V_{1} \rightarrow (C_1 \ D_2, \ C_1 \ D_2) \]
Synchronous CNF?

A → (B₁ C₂ D₃ E₄, C₂ E₄ B₁ D₃)

A → ([B₁ C₂] D₃ E₄, [C₂ E₄ B₁] D₃)

A → (B₁ [C₂ D₃] E₄, [C₂ E₄ B₁ D₃])

A → (B₁ C₂ [D₃ E₄], C₂ [E₄ B₁ D₃])

rank 4
Synchronous CNF?

\[ A \rightarrow (B_1 \ C_2 \ D_3, \ C_2 \ D_3 \ B_1) \]

\[ A \rightarrow (B_1 \ C_2 \ D_3 \ E_4, \ C_2 \ E_4 \ B_1 \ D_3) \]
A hierarchy of synchronous CFGs

\[ 1\text{-CFG} \subsetneq 2\text{-CFG} = 3\text{-CFG} = 4\text{-CFG} = \ldots \]

\[ 1\text{-SCFG} \subsetneq 2\text{-SCFG} = 3\text{-SCFG} \subsetneq 4\text{-SCFG} \subsetneq \ldots \]

\[ \backslash \backslash \]

ITG

(Wu, 1997)
Extraction
Rule extraction

Extract *all possible* rules from each sentence pair that respect:

- Word alignments
- Syntactic structure
- Ad-hoc constraints
Australia is one of the few countries that have diplomatic relations with North Korea.
Grammar extraction

Australia is one of the few countries that have diplomatic relations with North Korea.

(X → with X₁ have X₂, X → have X₂ with X₁)
Ad-hoc constraints

• Hiero:
  ‣ Initial phrases $\leq 10$ words, rules $\leq 6$ symbols
  ‣ Require an aligned terminal
  ‣ Limit to two nonterminals, nonadjacent

• Galley et al:
  ‣ Rule must be minimal or have $\leq 5$ English nodes (non-leaf, non-POS)
Example rules

\[ X \rightarrow \text{的} \quad X \rightarrow \text{的} \]

\[ X \rightarrow X_1 \text{的} X_2 \quad X \rightarrow \text{the} \ X_2 \text{ of} \ X_1 \]

\[ X \rightarrow X_1 \text{的} X_2 \quad X \rightarrow \text{the} \ X_2 \text{ that} \ X_1 \]

\[ X \rightarrow \text{在} \quad X \rightarrow \text{在} \]

\[ X \rightarrow \text{在} \ X_1 \ \text{下} \quad X \rightarrow \text{under} \ X_1 \]

\[ X \rightarrow \text{在} \ X_1 \ \text{前} \quad X \rightarrow \text{before} \ X_1 \]

\[ X \rightarrow \text{今年} \ X_1 \quad X \rightarrow \text{this year} \]

\[ X \rightarrow X_1 \text{ 之一} \quad X \rightarrow \text{one of} \ X_1 \]

\[ X \rightarrow X_1 \text{ 总统} \quad X \rightarrow \text{president} \ X_1 \]
Glue rules

• Plus “glue” rules:

\[
(S \rightarrow S_1X_2, S \rightarrow S_1X_2) \\
(S \rightarrow X_1, S \rightarrow X_1)
\]

• Acts as fallback like phrase-based systems

![Tree diagrams for example rules]
1. Phrases
   * respect word alignments
   * are syntactic constituents

2. Subtract phrases to form rules
Rule extraction

1. Phrases
   * respect word alignments
   * are syntactic constituents

2. Subtract phrases to form rules

- Maria no
- dió una bofetada
- a la bruja verde

- the
- green
- witch

- slap
- NP
- VP

- S
1. Phrases
   * respect word alignments
   * are syntactic constituents

2. Subtract phrases to form rules

Rule extraction

dió una bofetada

slap

X

VB

NP

VP
<table>
<thead>
<tr>
<th>Variety</th>
<th>Method</th>
<th>Source</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>string-to-string</td>
<td>ITG (Wu, 1997)</td>
<td>Hiero (Chiang, 2005)</td>
<td></td>
</tr>
<tr>
<td>tree-to-string</td>
<td>Huang et al, 2006</td>
<td>Liu et al, 2006</td>
<td></td>
</tr>
<tr>
<td>tree-to-tree</td>
<td>DOT (Poutsma, 2000)</td>
<td>Zhang et al., 2008</td>
<td>Liu et al., 2009</td>
</tr>
</tbody>
</table>
dió una bofetada a la bruja verde.
dió una bofetada
台湾 在 顺差 中 贸易 两岸

Taiwan's surplus in the two shores
In trade between the two shores, Taiwan’s surplus in trade between the two shores.

在两岸贸易中，台湾顺差。
台湾在两岸贸易中顺差带来的台湾的贸易 surplus in trade between the two shores.
在两岸贸易中 台湾 顺差

Taiwan’s surplus in trade between the two shores
more than 20 various kinds

JJ

IN

CD

check points of

JJ

NNS

QP

NN

NNS

IN

NP

NP

PP
more than 20 check points of various kinds
more than 20 check points of various kinds
Model
Probability model

- Synchronous CFG gives us possible translations (derivations)
- What makes one translation better than another?
- Log-linear model that combines small number of feature functions
Probability model

Log-linear model (Och and Ney, 2002)

\[ P(D) \propto \prod_{r \in D} \prod_{i} \nu_i(r)^{\lambda_i(r)} \]

- **\( D \)**: derivation
- **\( r \)**: rules in \( D \)
- **\( \nu_i \)**: feature functions
- **\( \lambda_i \)**: feature weights

Weights \( \lambda_i \) learned by maximum-BLEU training (Och 2003)
Model features

- Phrase translation:

  \[ p(X \to X_1 \text{之一} | X \to \text{one of } X_1) \]
  \[ p(X \to \text{one of } X_1 | X \to X_1 \text{之一}) \]

- Problem: rules not actually observed?
Model features

- Lexical weighting (Koehn):
  \[
  \frac{1}{2}[p(\text{之一} | \text{one}) + p(\text{之一} | \text{of})] \\
p(\text{one} | \text{之一}) \times p(\text{of} | \text{之一})
  \]
Model features

- Binary features: extracted rules; glue rule $S \rightarrow SX$; green rules
- Number of English words
- (Word) $n$-gram language model:
  
  $p(\text{Australia} | \text{START}) \times p(\text{is} | \text{START Australia}) \ldots$
Decoding
Review: CKY

\[
\begin{align*}
S & \rightarrow \text{NP} \text{ VP} \\
\text{NP} & \rightarrow \text{I} \\
\text{NP} & \rightarrow \text{the box} \\
\text{VP} & \rightarrow \text{V} \ \text{NP} \\
\text{V} & \rightarrow \text{open}
\end{align*}
\]
Review: CKY

S → NP VP
NP → I
NP → the box
VP → V NP
V → open
Review: CKY

S → NP VP
NP → I
NP → the box
VP → V NP
V → open
Review: CKY

S → NP VP
NP → I
NP → the box
VP → V NP
V → open

I open the box
Review: CKY

S → NP VP
NP → I
NP → the box
VP → V NP
V → open

S
I open the box
Review: CKY

O(n^3) ways of matching
Translation

I open the box

\(O(n^3)\)
Translation

I open the box

watashi wa hako wo akemasu
Finding the best translation

- Just like PCFG parsing
- $n$-best translations is efficient with the right tricks
Bigram language model

I open the box

watashi wa hako wo akemasu
Translation with a LM

NP

V

NP

I

open

the box
Translation with a LM

NP
watashi wa
I

V
akemasu
open

NP
hako wo
the box
Translation with a LM

watashi wa
open
the box

I

S

VP

NP

hako ... akemasu
Translation with a LM

watashi ... akemasu

I  open  the box
Translation with a LM

O($n^3 V^4 (g^{-1})$) combinations
Translation with a LM

$O(n^{3+4(g-1)})$ combinations
Summary

- Synchronous CFG/TSG allows translation based on hierarchical/syntactic structure
- Binarization not free
- Can be extracted from parallel data
- Tree-to-tree translation rare
- LM adds complexity to decoding