Rule Markov Models for Fast Tree-to-String Translation

Ashish Vaswani  
USC / ISI

Haitao Mi  
CAS / ICT

Liang Huang  
USC / ISI

David Chiang  
USC / ISI
Bushi held talks with Sharon.
Bush held talks with Sharon
Bushi held talks with Sharon.

Japanese: じゅxing 里面 はいたん

Tree Structure:

- VP@2.2
  - VV
  - AS
  - NP
  - juxing
  - le
  - huitan

- VP@2
  - PP@2.1
  - P@2.1.1
  - NP@2.1.2
  - PP@2.1
  - NP@2.1
  - NP@1
  - IP@ε
  - NP@1
  - VP@2
Bush held talks with Shalong.

Shalong held talks with Sharon.
Bush held talks with Sharon

juxing le huitan

held talks with

Bushi yu Shalong juxing le huitan

yu with Shalong
Bush held talks with Sharon

Shalong juxing le huitan

held talks with Sharon
Bush held talks with Sharon

Shalong

Sharon
Bush held talks with Sharon
Bush held talks with Sharon.
<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>Decoding time</th>
<th>Grammar size</th>
</tr>
</thead>
<tbody>
<tr>
<td>composed</td>
<td>✔️</td>
<td></td>
<td></td>
</tr>
<tr>
<td>minimal</td>
<td></td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>??</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
composed rules alleviate independence b/w rules

but we can explicitly model this dependence by conditioning on the history (grammar annotation)

Learn the ancestral dependencies between minimal rules with rule Markov models (RMM)

Use rule Markov models with top down incremental decoder (Huang and Mi, 2010)
Generative Story

P(r₁|ε)

NP@₁ VP@₂

IP@ε

NP@₁

VP@₂

r₁
Generative Story

\[ P(r_2|r_1) \]

Diagram:
- Node \( r_1 \) connected to \( r_2 \)
- Structure labeled with categories: NP@1, VP@2, IP@ε
Generative Story

Bush held talks

P(r_5|r_1, r_3)
Bush held talks
Bush held talks with NP@2.1.2

P(r_6|r_1,r_3,r_4)
Generative Story

Bush held talks with Sharon
Probability of a Derivation Tree

\[ P(T) = P(r_1|\varepsilon) \cdot P(r_2|r_1) \cdot P(r_3|r_1) \]
\[ P(r_4|r_1,r_3) \cdot P(r_5|r_1,r_3) \]
\[ P(r_6|r_1,r_3,r_4) \cdot P(r_7|r_1,r_3,r_4) \]
Probability of a Derivation Tree

\[ P(T) = P(r_1|\varepsilon) \cdot P(r_2|r_1) \cdot P(r_3|r_1) \]
\[ \cdot P(r_5|r_1,r_3) \cdot P(r_4|r_1,r_3) \]
\[ \cdot P(r_6|r_1,r_3,r_4) \cdot P(r_7|r_1,r_3,r_4) \]
Probability of a Derivation Tree

\[ P(T) = P(r_1 | \varepsilon) \cdot P(r_2 | r_1) \cdot P(r_3 | r_1) \]
\[ \quad \cdot P(r_5 | r_1, r_3) \cdot P(r_4 | r_1, r_3) \]
\[ \quad \cdot P(r_6 | r_1, r_3, r_4) \cdot P(r_7 | r_1, r_3, r_4) \]
Probability of a Derivation Tree

\[
P(T) = P(r_1|\epsilon) \cdot P(r_2|r_1) \cdot P(r_3|r_1) \\
P(r_5|r_1,r_3) \cdot P(r_4|r_1,r_3) \\
P(r_6|r_1,r_3,r_4) \cdot P(r_7|r_1,r_3,r_4)
\]
Probability of a Derivation Tree

\[ P(T) = P(r_1|\varepsilon) \cdot P(r_2|r_1) \cdot P(r_3|r_1) \cdot P(r_4|r_1, r_3) \cdot P(r_5|r_1, r_3) \cdot P(r_6|r_1, r_3, r_4) \cdot P(r_7|r_1, r_3, r_4) \]
Probability of a Derivation Tree

\[
P(T) = P(r_1 | \varepsilon) \cdot P(r_2 | r_1) \cdot P(r_3 | r_1) \\
\quad \cdot P(r_5 | r_1, r_3) \cdot P(r_4 | r_1, r_3) \\
\quad \cdot P(r_6 | r_1, r_3, r_4) \cdot P(r_7 | r_1, r_3, r_4)
\]
Probability of Derivation Tree

\[ P(T) = P(r_1 | \varepsilon) \cdot P(r_2 | r_1) \cdot P(r_3 | r_1) \]
\[ P(r_5 | r_1, r_3) \cdot P(r_4 | r_1, r_3) \]
\[ P(r_6 | r_1, r_3, r_4) \cdot P(r_7 | r_1, r_3, r_4) \]
Learning Rule Markov Models

IP@ε

NP@1

VP@2

PP@2.1

P@2.1.1

NP@2.1.2

VV

AS

NP

Bushi

yu

Shalong

juxing

le

huitan

Bush

held

talks

with

Sharon

Bush

held

talks

with

Sharon
Learning Rule Markov Models

Galley et al.
Learning Rule Markov Models
Learning Rule Markov Models
Learning Rule Markov Models

- $r_1$
- $r_2$
- $r_3$
- $r_4$
- $r_5$
- $r_6$
- $r_7$
Smaller rule Markov models
Smaller rule Markov models

• RM-A: Keep only those contexts in which more than $P$ items were observed. Tune $P$ for BLEU on dev
Smaller rule Markov models

- RM-A: Keep only those contexts in which more than $P$ items were observed. Tune $P$ for BLEU on dev

- RM-B: Keep only those contexts which were observed more than $P$ times. Tune $P$ for BLEU on dev
Smaller rule Markov models

- **RM-A:** Keep only those contexts in which more than $P$ items were observed. Tune $P$ for BLEU on dev
- **RM-B:** Keep only those contexts which were observed more than $P$ times. Tune $P$ for BLEU on dev
- **RM-C:** Build Prediction Suffix Trees using the approach of Bejerano and Yona (1999)
Smaller rule Markov models

- RM-A: Keep only those contexts in which more than $P$ items were observed. Tune $P$ for BLEU on dev.
- RM-B: Keep only those contexts which were observed more than $P$ times. Tune $P$ for BLEU on dev.
Incremental decoding with RMMs
Incremental decoding with RMMs

- Rule Markov models using only vertical context are a natural fit for Incremental decoding (Huang and Mi, 2010)
• rule Markov models using only vertical context are a natural fit for Incremental decoding (Huang and Mi, 2010)

• The top-down decoder maintains the vertical context as part of its state which is used in predicting the next rule
Incremental Decoding

- stack (active derivation history): \([.IP@\varepsilon] [NP@1 . VP@2]\)
- three colors for nodes: white (uncovered), grey (partially covered), and black (covered)

(Huang and Mi, 2010)
Incremental Decoding

- stack (active derivation history): \([\cdot IP@\varepsilon] [NP@1 \cdot VP@2]\)

- three colors for nodes: white (uncovered), grey (partially covered), and black (covered)

```
Bushi
```

```
\begin{itemize}
\item \text{``I have finished NP subtree but not started with VP subtree''}
\end{itemize}
```

(Huang and Mi, 2010)
Example Incremental Decoding

<s> · IP@<sup>e</sup> </s>

<s>

(Huang and Mi, 2010)
Example Incremental Decoding

[<s> . IP@ε </s>] [ . NP@1 VP@2]

<s>

stack

hypothesis

action: predict (push)

(Huang and Mi, 2010)
Example Incremental Decoding

Hypothesis:

<s> IP@ε </s> [ NP@1 VP@2 ] [ Bush ]

<s>

Action: predict

(Huang and Mi, 2010)
Example Incremental Decoding

<s>  IP@[ε </s>]  [ NP@[1 ]  VP@[2 ]  [Bush < ]

<s> Bush

action: scan

(Huang and Mi, 2010)
Example Incremental Decoding

<s> · IP@ε </s> [NP@1 · VP@2]

<s> Bush</s>

action: pop

(Huang and Mi, 2010)
Example Incremental Decoding

<s> Bush

(Huang and Mi, 2010)
Example Incremental Decoding

[<s> • IP@ε </s>] [NP@1 • VP@2] [... VP@2.2 PP@2.1] [... held talks]

<s> Bush

(Huang and Mi, 2010)
Example Incremental Decoding

<s> Bush held talks

(Huang and Mi, 2010)
Example Incremental Decoding

<s> Bush held talks

(Huang and Mi, 2010)
Example Incremental Decoding

<s> Bush held talks

stack

hypothesis

action: predict

(Huang and Mi, 2010)
Example Incremental Decoding

<s> Bush held talks

(Huang and Mi, 2010)
Example Incremental Decoding

<s> Bush held talks and

(Huang and Mi, 2010)
Example Incremental Decoding

<s>. IP<sup>e</sup> </s>] [NP@1 VP@2] [VP@2.2 PP@2.1] [P@2.1.1 NP@2.1.2]

<s> Bush held talks and

action: pop

(Huang and Mi, 2010)
Example Incremental Decoding

<s>. IP@ε</s> [NP@1. VP@2] [VP@2.2. PP@2.1] [P@2.1.1. NP@2.1.2] [. Sharon]

<s> Bush held talks and

(Huang and Mi, 2010)
<s> Bush held talks and Sharon

(Huang and Mi, 2010)
<s> Bush held talks and Sharon

(Huang and Mi, 2010)
Example Incremental Decoding

<s> Bush held talks and Sharon

(Huang and Mi, 2010)
Example Incremental Decoding

<s> Bush held talks and Sharon

(Huang and Mi, 2010)
Example Incremental Decoding

<s> Bush held talks and Sharon

(Huang and Mi, 2010)
Example Incremental Decoding

stack

hypothesis

<s> IP@ε </s>

<s>

NP@1

IP@ε

VP@2

Bushi

PP@2.1

P@2.1.1

yu

NP@2.1.2

Shalong

VP@2.2

juxing

AS

le

NP

huitan

this work: adding rule LM
Example Incremental Decoding

\[
[<s> \text{IP}^\varepsilon <</s>] [\text{NP}^1 \text{VP}^2]
\]

\text{r}_1

\text{IP}^\varepsilon

\text{NP}^1

\text{VP}^2

\text{Bushi}

\text{PP}^{2.1}

\text{VP}^{2.2}

\text{P}^{2.1.1}

\text{NP}^{2.1.2}

\text{VV}

\text{AS}

\text{NP}

\text{yu}

\text{Shalong}

\text{juxing}

\text{le}

\text{huitan}

\text{stack}

\text{hypothesis}

\text{rule probability}

\text{P}(\text{r}_1|\varepsilon)

\text{action: predict (push)}

this work: adding rule LM
Example Incremental Decoding

<s> . IP@ε </s> [ . NP@1 VP@2 ] [ . Bush]  

<s>

stack

hypothesis

rule probability

P(r2|r1)

action: predict

this work: adding rule LM
Example Incremental Decoding

Stack

Hypothesis

Action: scan

This work: adding rule LM
Example Incremental Decoding

<s>. IP@ε </s> [NP@1 . VP@2]

<s> Bush

stack

hypothesis

action: pop

this work: adding rule LM
Example Incremental Decoding

stack

<s> Bush

hypothesis

rule probability

P(r₃|r₁)

action: predict

this work: adding rule LM
Example Incremental Decoding

[<s> • IP}@e <</s>] [NP}@1 • VP}@2] [. VP}@2.2 PP}@2.1] [. held talks]

<s> Bush

this work: adding rule LM
Example Incremental Decoding

<s> • IP@ε </s>] [NP@1 • VP@2] [. VP@2.2 PP@2.1] [held talks •]

<s> Bush held talks

action: scan

this work: adding rule LM
Example Incremental Decoding

<s> Bush held talks

action: pop

this work: adding rule LM
Example Incremental Decoding

<s> Bush held talks

<s> IPÆ </s> [NP@1 . VP@2 ] [VP@2.2 . PP@2.1 ] [ . P@2.1.1 NP@2.1.2 ]

stack

<s> hypothesis

<s> rule probability

P(r₄|r₁,r₃)

action: predict

this work: adding rule LM
Example Incremental Decoding

<s> Bush held talks

<s> IP@ε</s> [NP@1, VP@2] [VP@2.2, PP@2.1] [P@2.1.1, NP@2.1.2] [with/and] r6/r’6

This work: adding rule LM
Example Incremental Decoding

<s> Bush held talks with

this work: adding rule LM
Example Incremental Decoding

[<s> IP@ε </s>] [NP@1 VP@2] [VP@2.2 PP@2.1] [P@2.1.1 NP@2.1.2]

<s> Bush held talks with

action: pop

this work: adding rule LM
Example Incremental Decoding

<s> IP@ε </s> [NP@1 NP@2 VP@2 VP@2.2 PP@2.1 P@2.1.1 NP@2.1.2] [• Sharon]

<s> Bush held talks with

stack

hypothesis

rule probability

P(r7 |r3,r4)

action: predict

this work: adding rule LM
Example Incremental Decoding

<s> Bush held talks with Sharon

This work: adding rule LM
Example Incremental Decoding

<s> Bush held talks with Sharon

This work: Adding rule LM
Example Incremental Decoding

<s> Bush held talks with Sharon

action: pop

this work: adding rule LM
Example Incremental Decoding

<s> Bush held talks with Sharon </s>

This work: adding rule LM
Example Incremental Decoding

<s> IP@ε . </s>

<s> Bush held talks with Sharon </s>

action: pop

this work: adding rule LM
Experiments

- 1.5 million sentence pairs with 38/32 million words of Chinese/English
- Dev set: 616 sentences of the Newswire portion of 2006 NIST MT evaluation test set
- Test set: 619 sentences of the Newswire portion of 2006 NIST MT evaluation test set
Results

- minimal
- minimal + trigram RMM (RM-A)
- composed Rule
Results

- minimal
- minimal + trigram RMM (RM-A)
- composed Rule

\[
\begin{array}{c}
27 \\
26.25 \\
25.5 \\
24.75 \\
24 \\
\end{array}
\]

BLEU
Results

minimal
minimal + trigram RMM (RM-A)
composed Rule
Results

- minimal
- minimal + trigram RMM (RM-A)
- composed Rule
Results

- minimal
- minimal + trigram RMM (RM-A)
- composed Rule

<table>
<thead>
<tr>
<th>BLEU</th>
<th>Minimal</th>
<th>Minimal + Trigram RMM (RM-A)</th>
<th>Composed Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>24.75</td>
<td>26.25</td>
<td>27</td>
</tr>
</tbody>
</table>

Bar chart showing BLEU scores for different models.
Results

- minimal
- minimal + trigram RMM (RM-A)
- composed Rule

![Bar chart showing BLEU scores and decoding time for different models.](chart.png)
Results

- minimal
- minimal + trigram RMM (RM-A)
- composed Rule

Bar chart showing BLEU scores and decoding time per sentence.
Results

- BLEU:
  - minimal
  - minimal + trigram RMM (RM-A)
  - composed Rule

- Decoding time (sec/sent):
  - minimal
  - minimal + trigram RMM (RM-A)
Results

- minimal
- minimal + trigram RMM (RM-A)
- composed Rule

**BLEU**

- 24.75
- 25.5
- 26.25
- 27

**Decoding time sec/sent**

- 0
- 0.75
- 1.5
- 2.25
- 3
Results

color context probabilities
rules
Results

context probabilities
rules

2000000
1500000
1000000
5000000
0

minimal minimal+ RMM (RM-A) composed
Results

- minimal
- minimal + RMM (RM-A)
- composed

**context probabilities**

**rules**

- 0
- 500000
- 1000000
- 1500000
- 2000000
Results

The graph shows the comparison between different methods: minimal, minimal + RMM (RM-A), and composed. The y-axis represents the number of rules and context probabilities. The minimal method has the lowest number of rules and context probabilities, while the composed method has the highest. The minimal + RMM (RM-A) method is in between.
Results

- minimal
- minimal + RMM (RM-A)
- composed

- context probabilities
- rules

- rules count: 0
- composed count: 500000
- minimal + RMM (RM-A) count: 2000000
- minimal count: 1000000
Results

context probabilities
rules

minimal
minimal+ RMM (RM-A)
composed

0
2000000
4000000
6000000
8000000
10000000
12000000
14000000
16000000
18000000
20000000

rules
context probabilities
RMMs with composed Rules

\[ r_c \]

```
    VP[2]
   /    |
VP[2.2] VP[2.1]
P[2.1.1] P[2.1.2]
P[2.1.1] NP[2.1.2]
```

```
    IP[ε]
   /    |
```

```
Bush held talks with Sharon

Bushi yu Shalong juxing le huitan
```

```
with Sharon
```
RMMs with composed Rules

\[ r_c \]

\[ \text{pp@2.1} \]

\[ \text{p@2.1.1} \quad \text{NP@2.1.2} \]

\[ \text{yu} \quad \text{Shalong} \]

\[ \text{with Sharon} \]
RMMs with composed Rules

\[
rc
\]

\[
pp@2.1
\]

\[
p@2.1.1
d@2.1.2
\]

yu Shalong

with Sharon
The image represents RMMs (Regularized Markov Models) with composed Rules. The diagram illustrates a tree structure with nodes labeled as follows:

- $r_c$
- $pp@2.1$
- $p@2.1.1$
- $np@2.1.2$
- $yu$
- $Shalong$
- $with$
- $Sharon$

The tree is structured as follows:

- The root node is $r_c$.
- The node $pp@2.1$ is the parent of $p@2.1.1$ and $np@2.1.2$.
- The node $p@2.1.1$ has two children: $yu$ and another node.
- The node $np@2.1.2$ has two children: $Shalong$ and another node.
- There is an additional node $r_4$, $r_6$, and $r_7$ forming a separate subtree.

The text below the diagram reads, "with composed Rules."
RMMs with composed Rules

\[ r_c \]

\[ \text{pp@2.1} \]

\[ \text{p@2.1.1} \quad \text{NP@2.1.2} \]

\[ \text{yu} \quad \text{Shalong} \]

with Sharon

\[ r_4 \]

\[ r_6 \quad r_7 \]

\[ r_x \]

\[ r_c \]
RMMs with composed Rules

\[
\begin{align*}
    r_{c} & \quad \text{pp@2.1} \\
    p_{@2.1.1} & \quad \text{NP@2.1.2} \\
    \text{yu} & \quad \text{Shalong} \\
    \text{with} & \quad \text{Sharon}
\end{align*}
\]

= 

\[
\begin{align*}
    r_{x} & \quad r_{4} \\
    r_{6} & \quad r_{7}
\end{align*}
\]
RMMs with composed Rules

\[ r_c \]

\[ \text{pp@2.1} \]

\[ \text{p@2.1.1} \quad \text{NP@2.1.2} \]

\[ \text{yu} \quad \text{Shalong} \]

\[ \text{with} \quad \text{Sharon} \]

\[ P(r_c|r_x) = P(r_4|r_x) \cdot P(r_6|r_4,r_x) \cdot P(r_7|r_4,r_x) \]
RMMs with composed Rules

\[ P(r_c|r_x) = P(r_4|r_x) \cdot P(r_6|r_4,r_x) \cdot P(r_7|r_4,r_x) \]
RMMs with composed Rules

\[ P(r_c|r_x) = P(r_4|r_x) \cdot P(r_6|r_4,r_x) \cdot P(r_7|r_4,r_x) \]
Results: RMMs improve over composed rules

- composed
- RMM trigram + composed
Results: RMMs improve over composed rules

<table>
<thead>
<tr>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
</tr>
<tr>
<td>28.25</td>
</tr>
<tr>
<td>27.5</td>
</tr>
<tr>
<td>26.75</td>
</tr>
<tr>
<td>26</td>
</tr>
</tbody>
</table>

- composed
- RMM trigram + composed
Results: RMMs improve over composed rules

- composed
- RMM trigram + composed

![Bar chart showing BLEU scores]

- BLEU 26.75
- BLEU 27.5
- BLEU 28.25
- BLEU 29
Results: RMMs improve over composed rules

- BLEU scores for composed rules: 26
- BLEU scores for RMM trigram + composed: 28.25

Bar chart showing the comparison between composed rules and RMM trigram + composed rules.
Related Work

- Quirk and Menezes (2006)
- Ding and Palmer (2005)
- Liu and Gildea (2008)
Conclusion

- Using rule Markov models, we are able to get significant improvements in BLEU score.
- The grammar size and decoding time is less than the composed rule grammar
- Using rule Markov models with composed rule grammars further improves the BLEU score.
THANKS