Rule Markov Models for Fast Tree-to-String Translation

Ashish Vaswani
USC / ISI

Haitao Mi
CAS / ICT

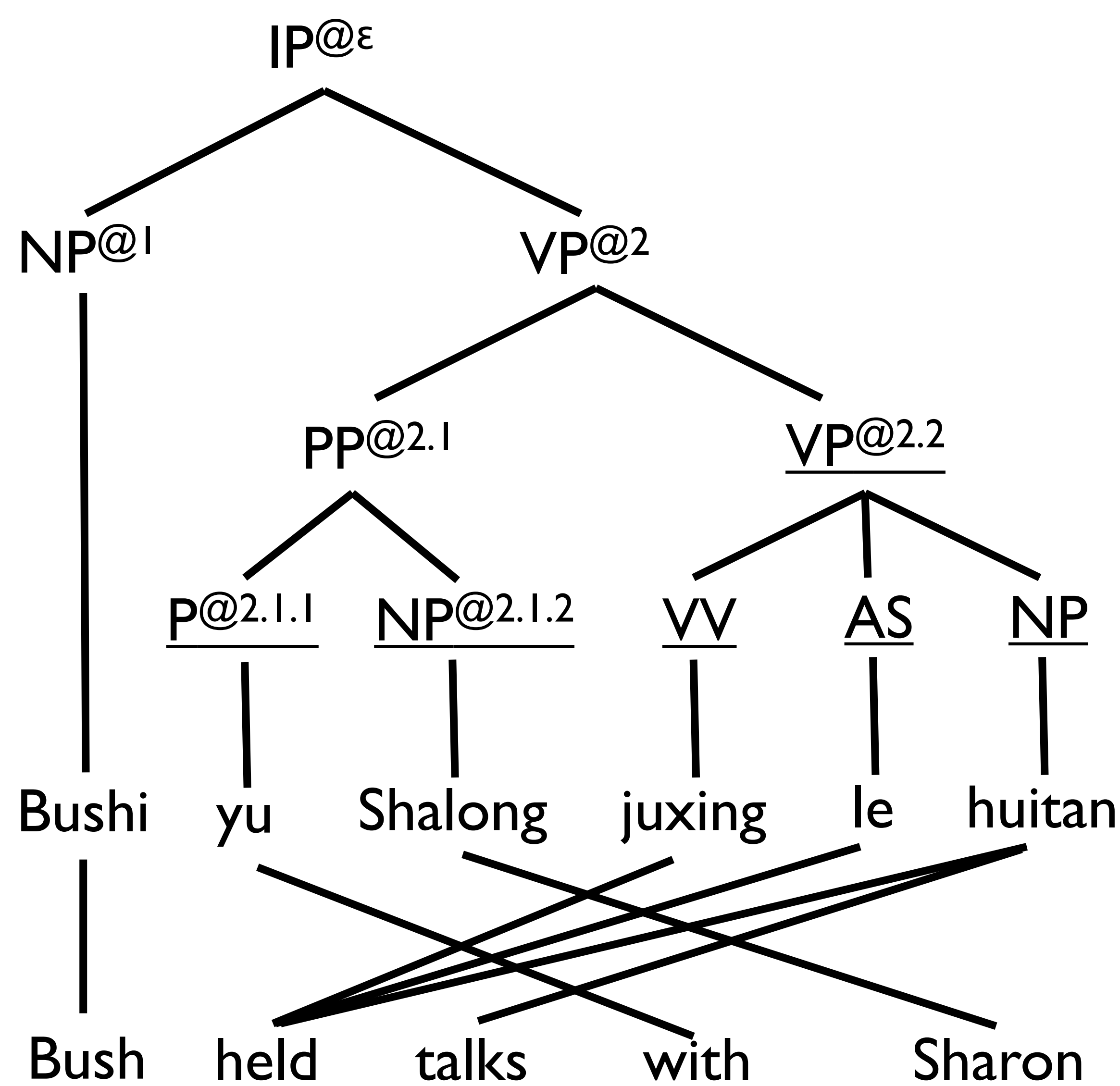
Liang Huang
USC / ISI

David Chiang
USC / ISI

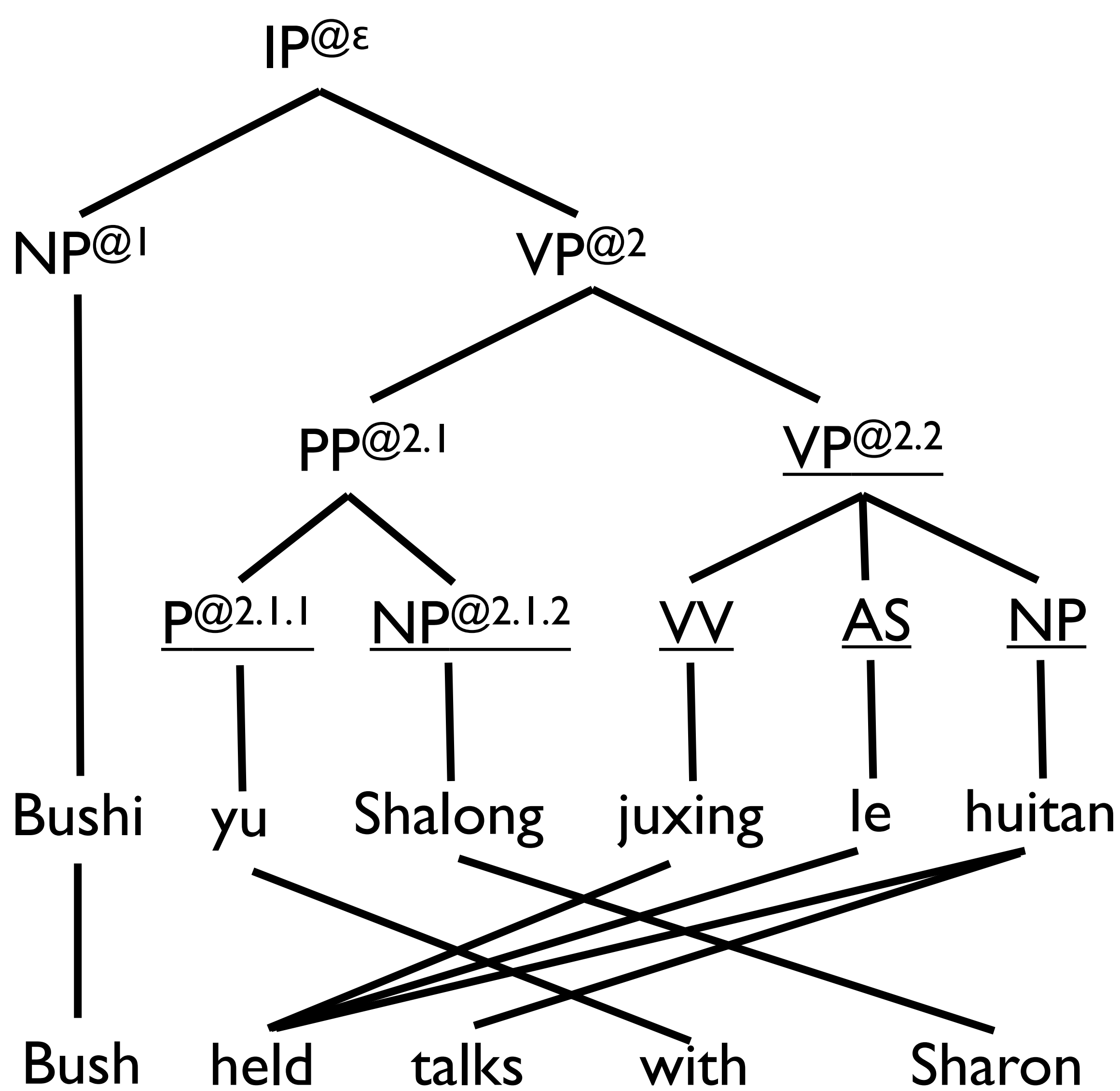
ISI
Information Sciences Institute

USC Viterbi
School of Engineering

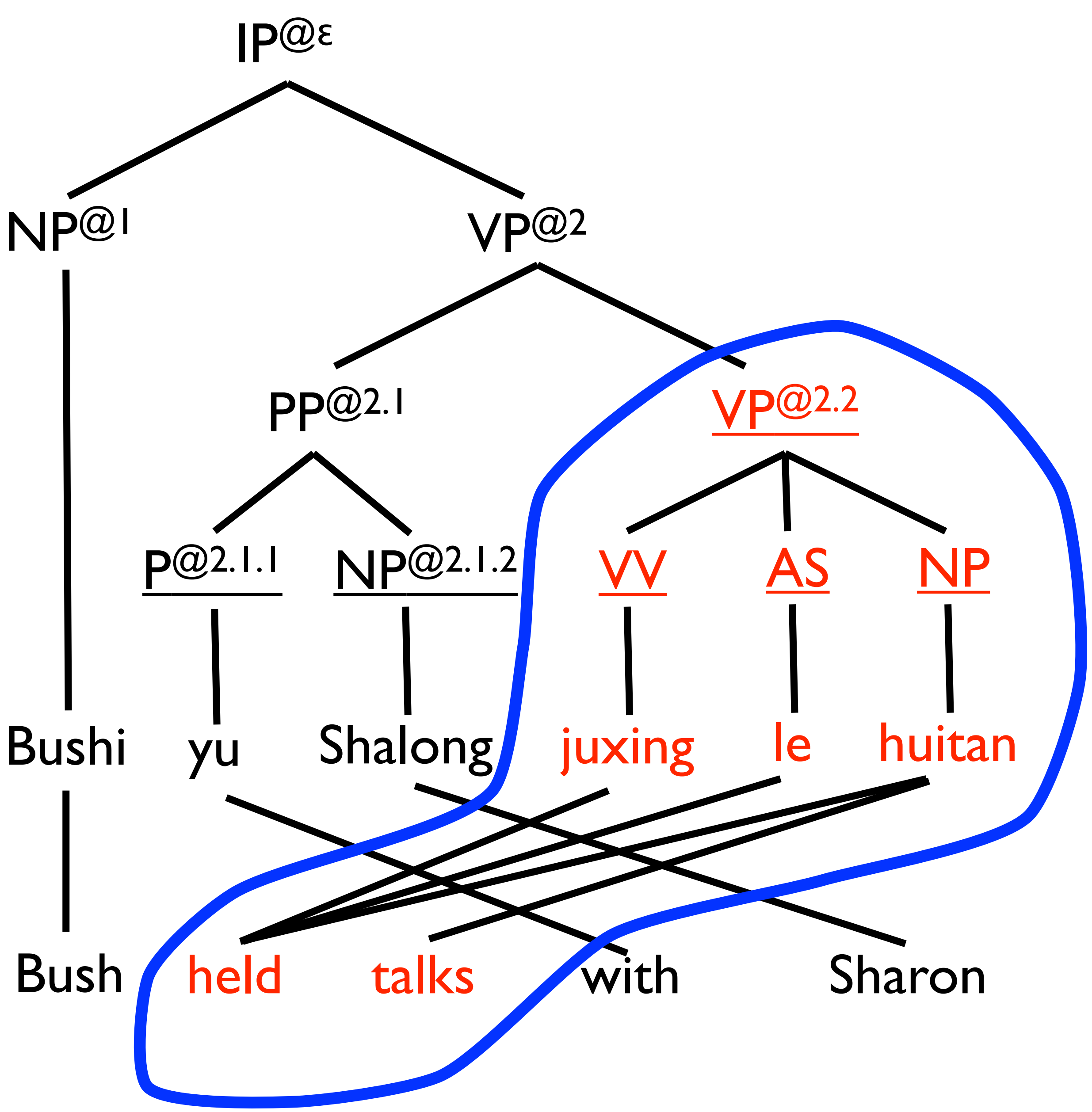
Syntax Based Machine Translation (SBMT)

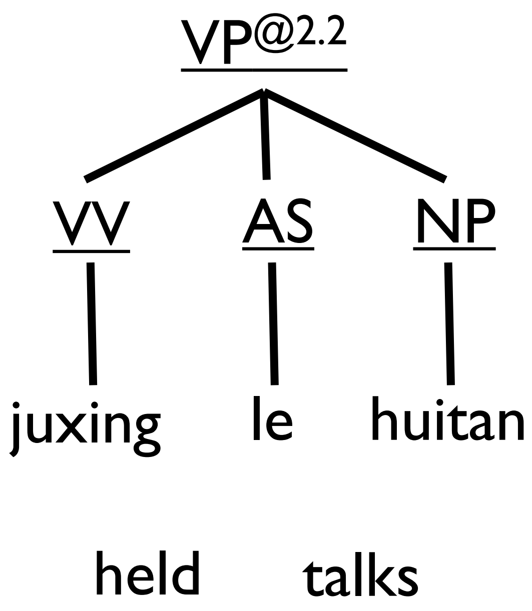


Syntax Based Machine Translation (SBMT)

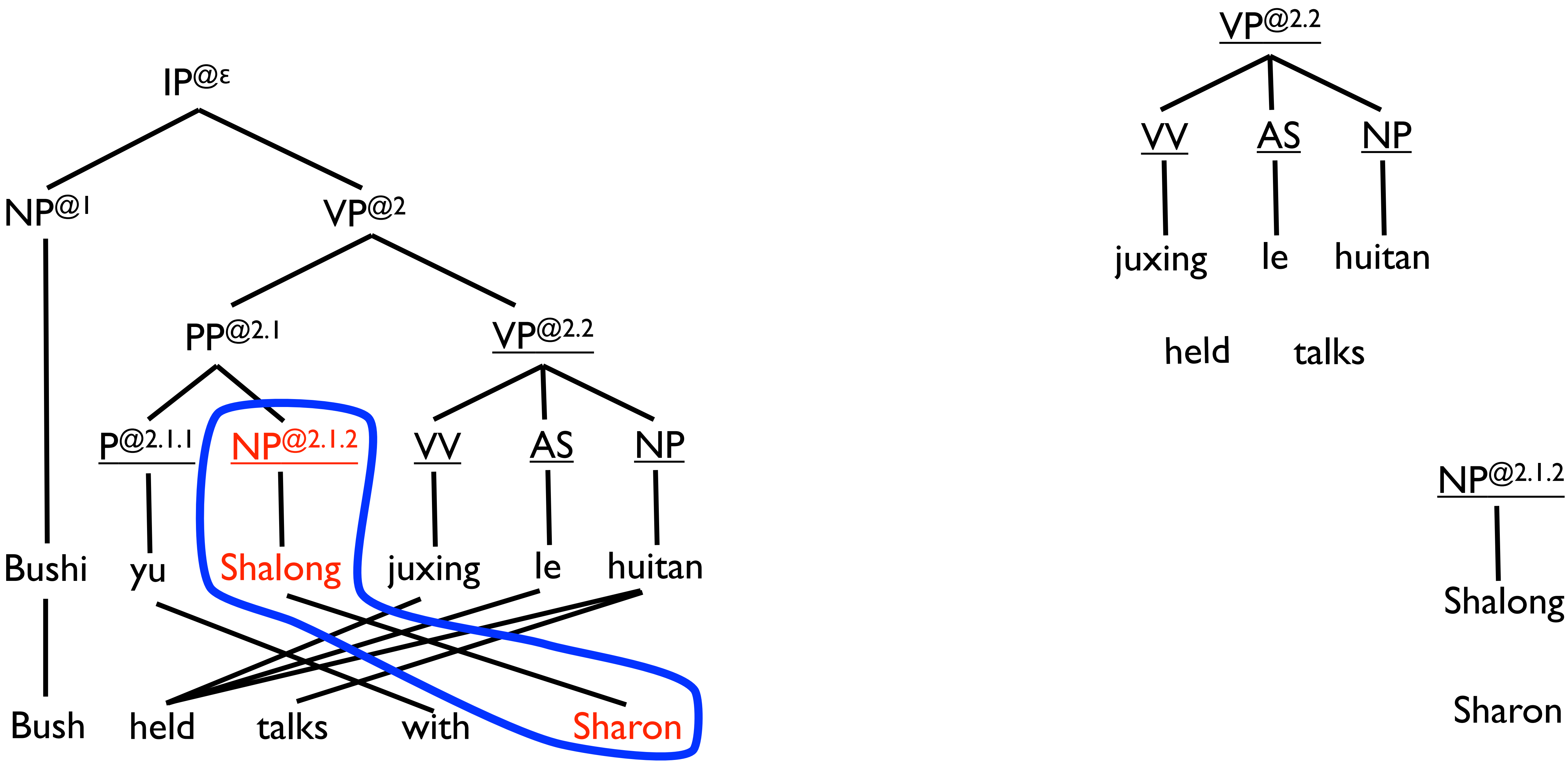


Syntax Based Machine Translation

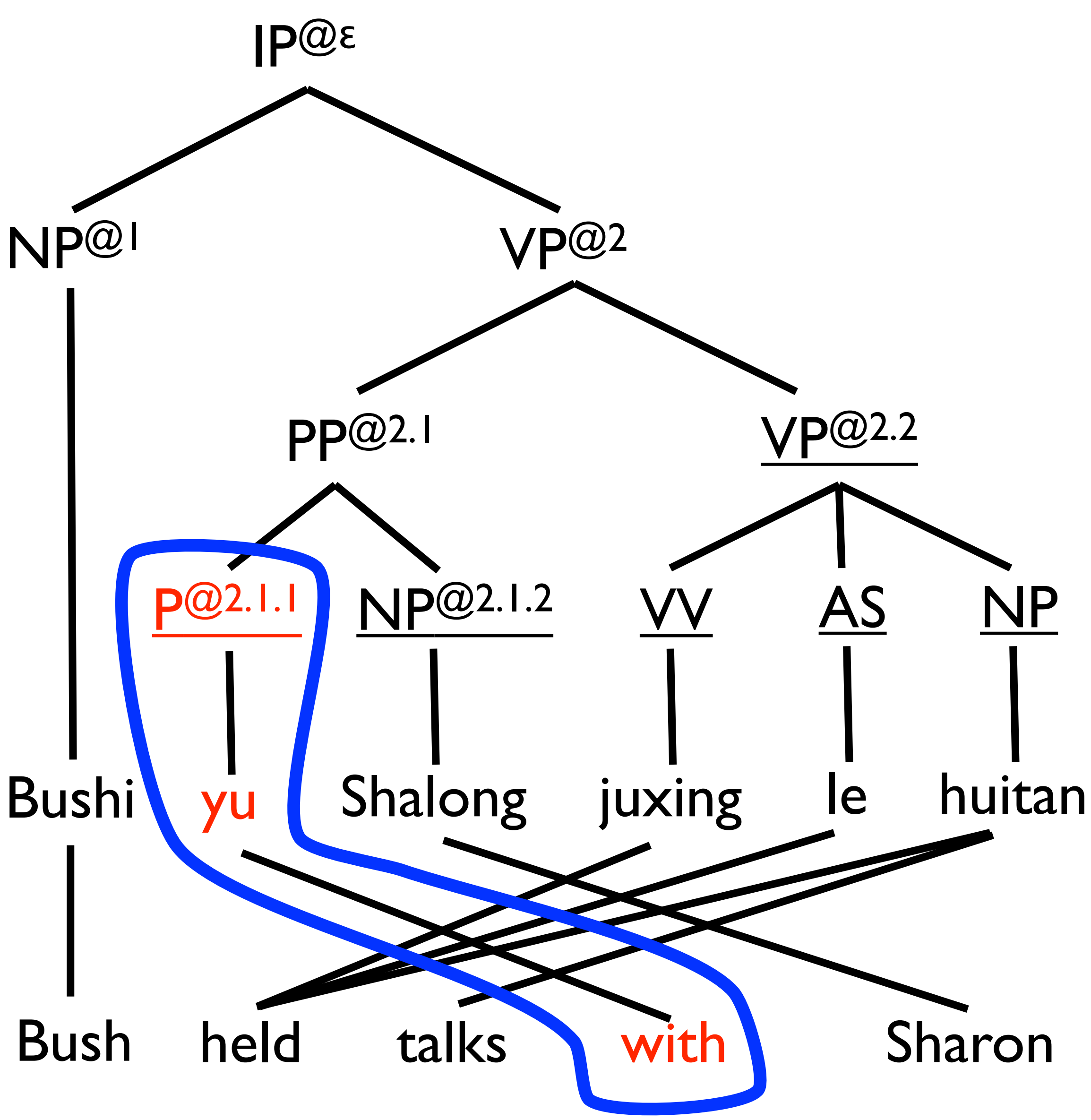


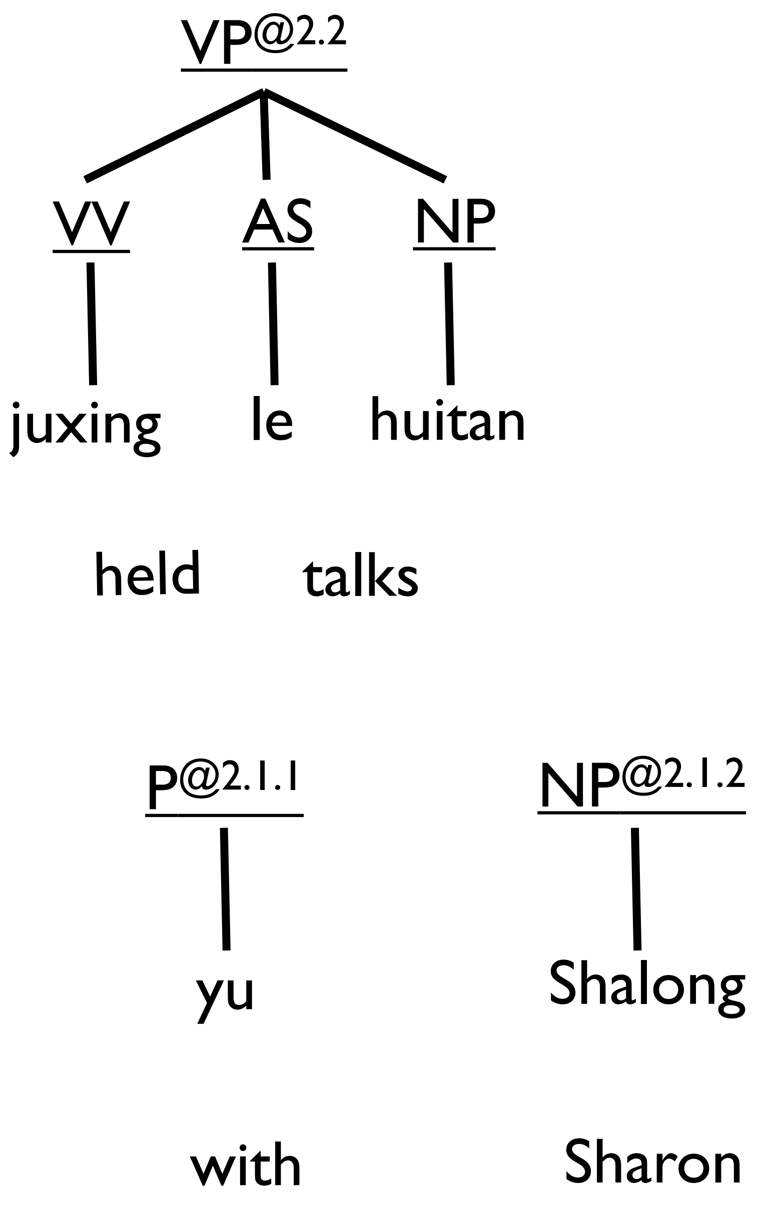


Syntax Based Machine Translation

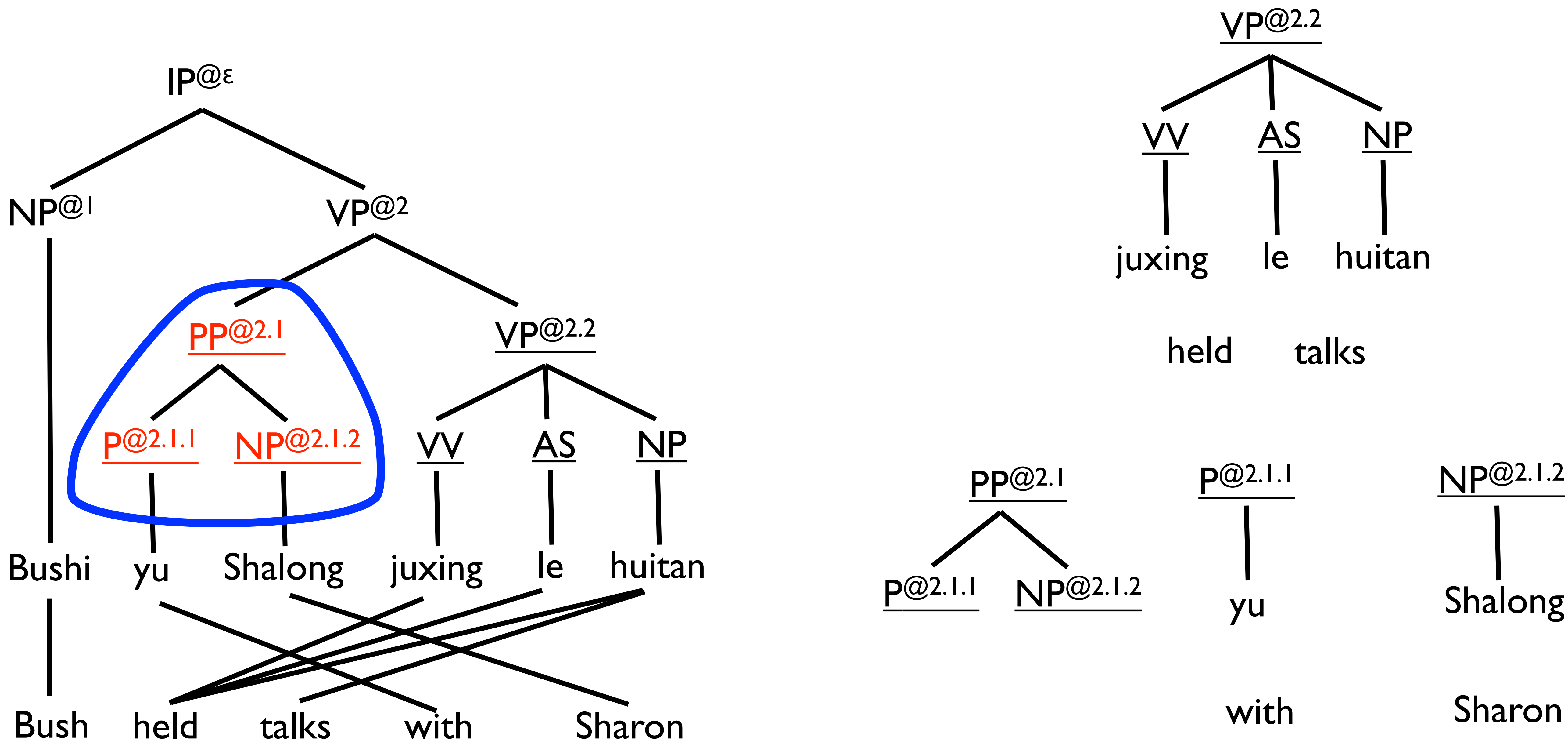


Syntax Based Machine Translation

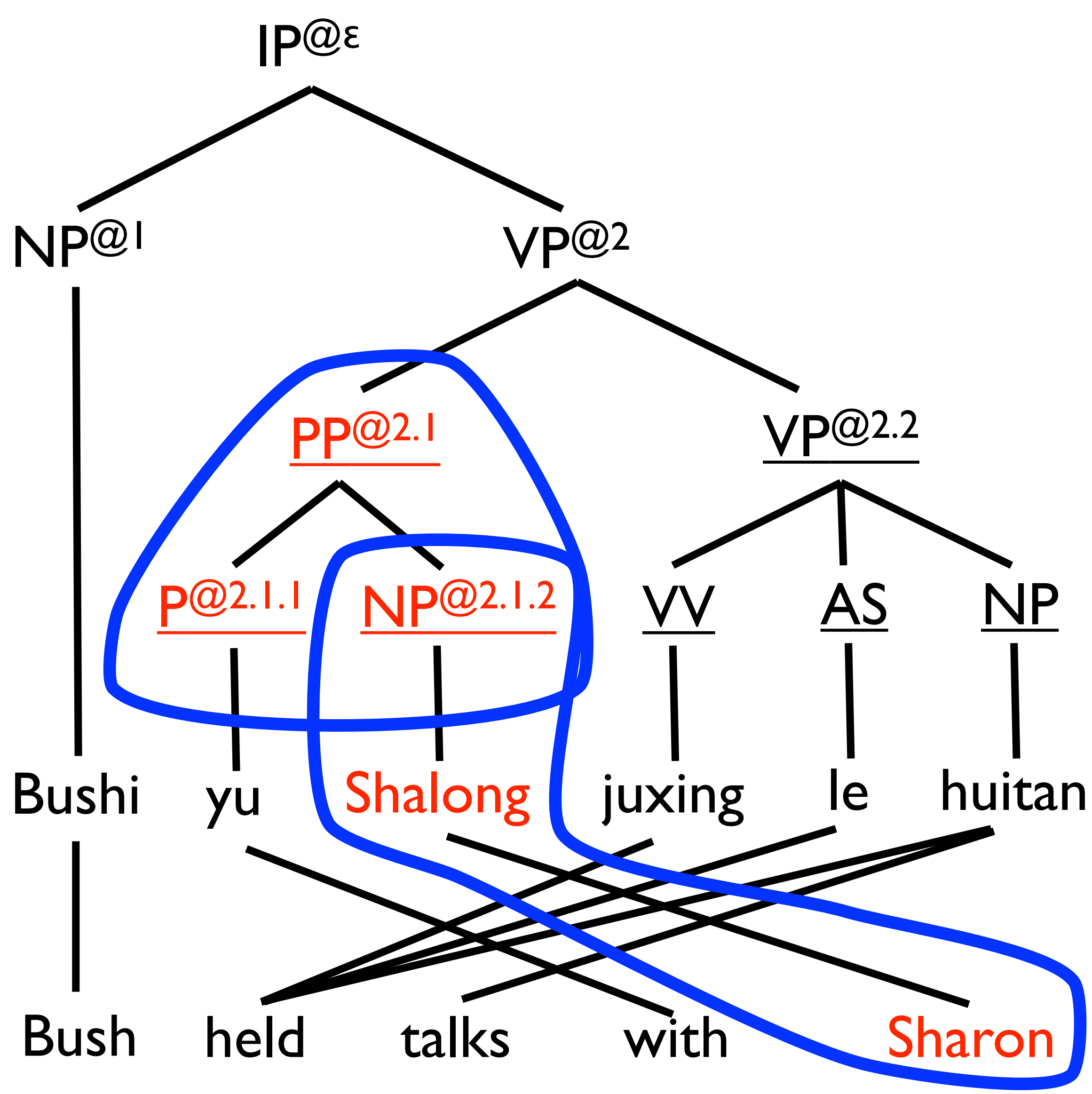


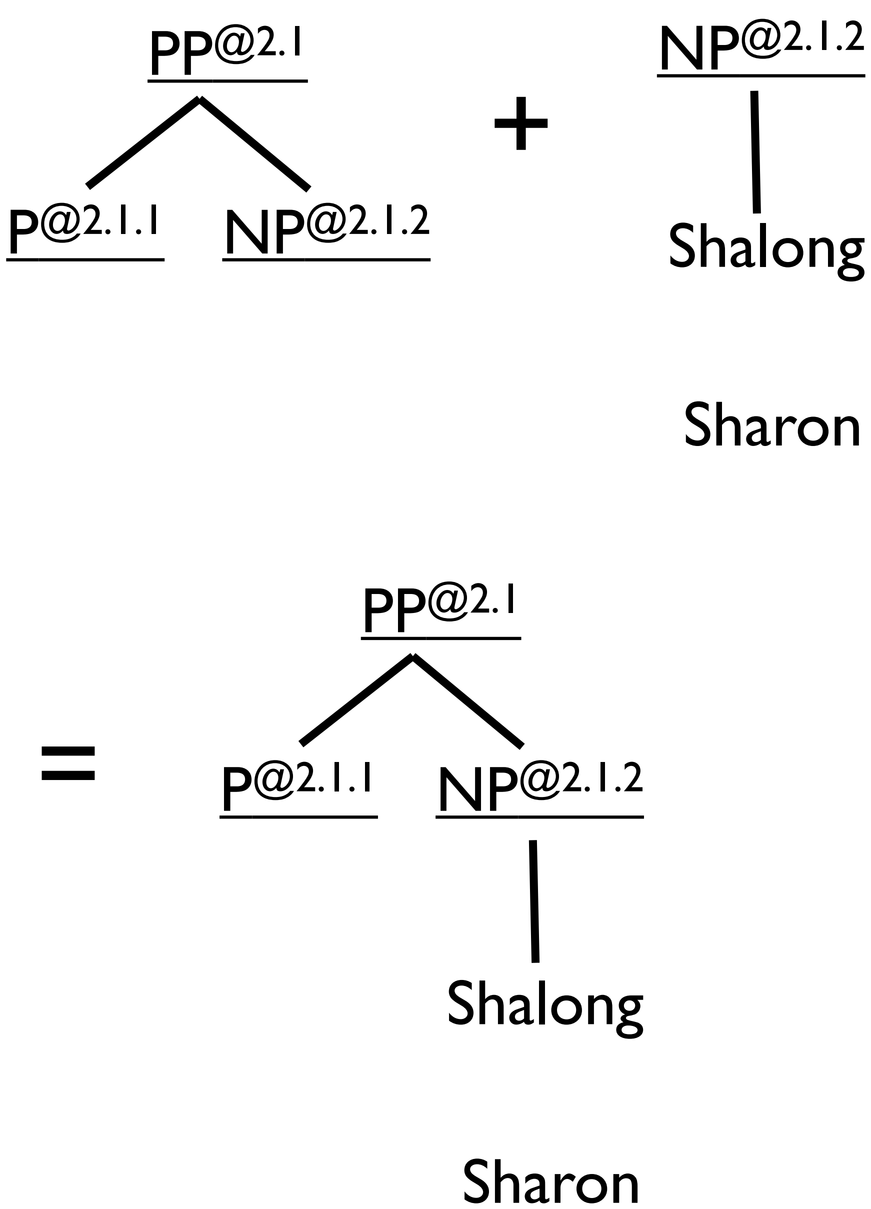


Syntax Based Machine Translation

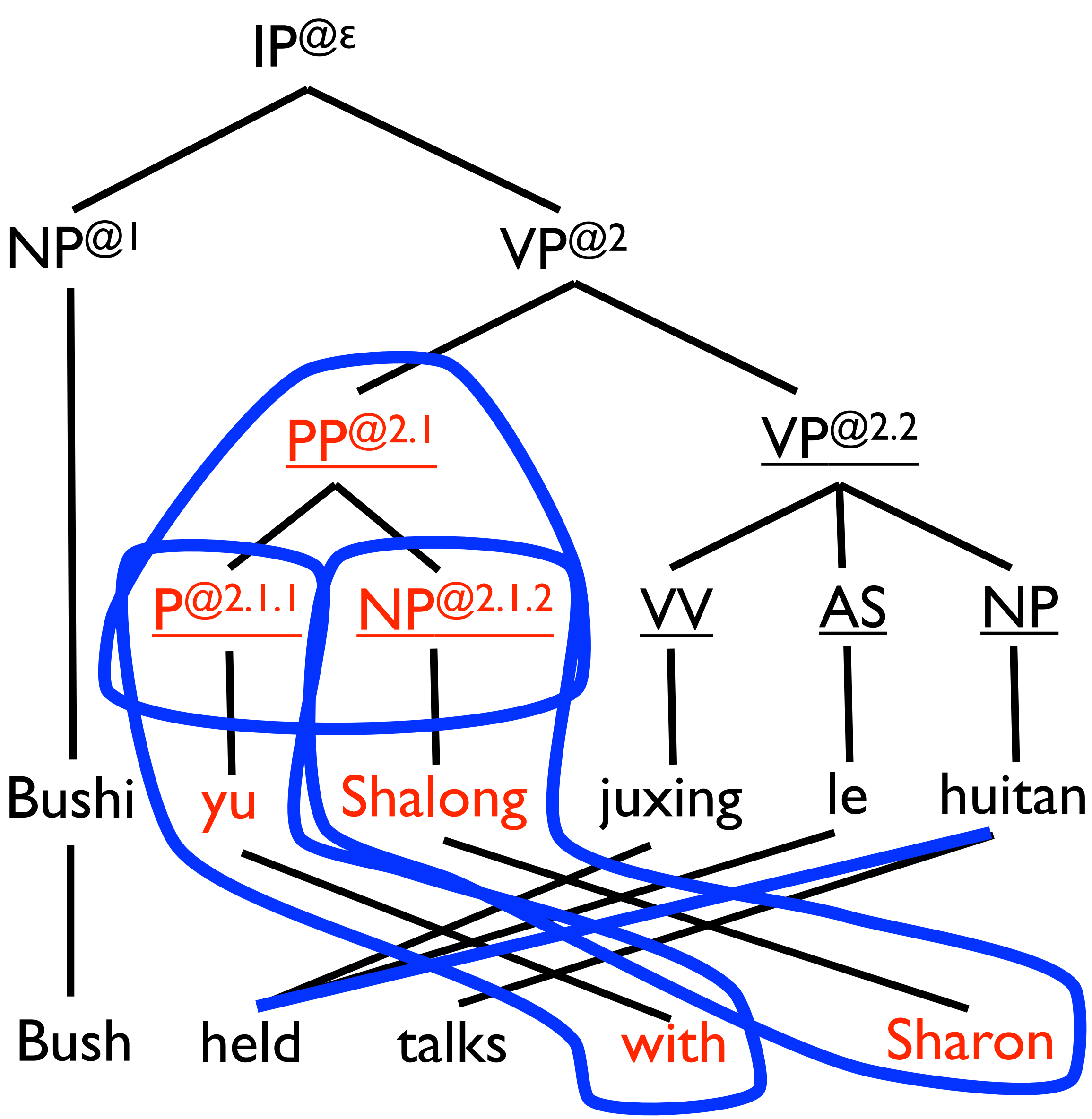


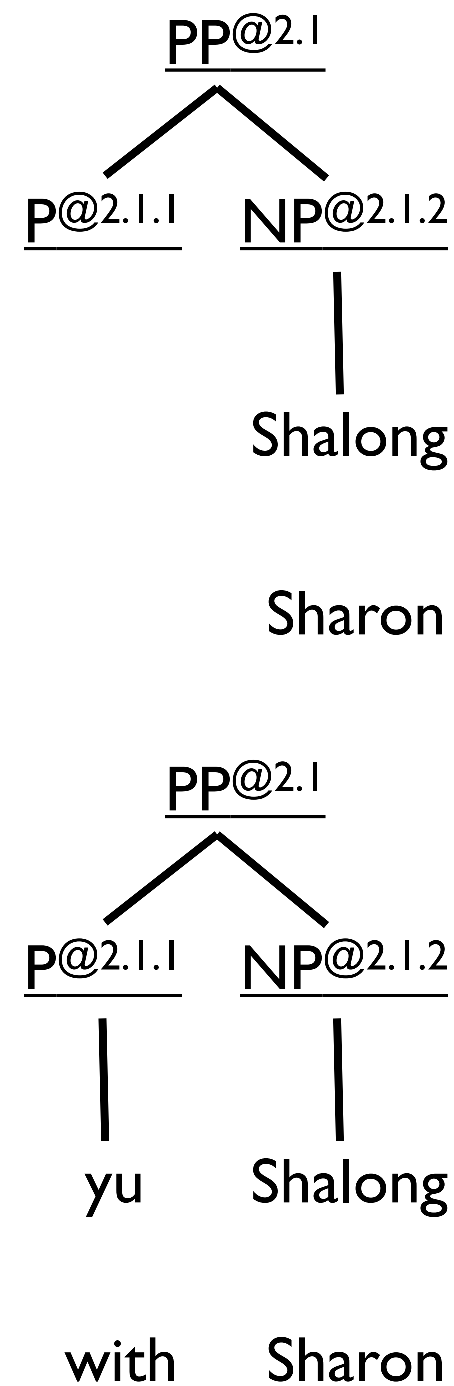
Syntax Based Machine Translation



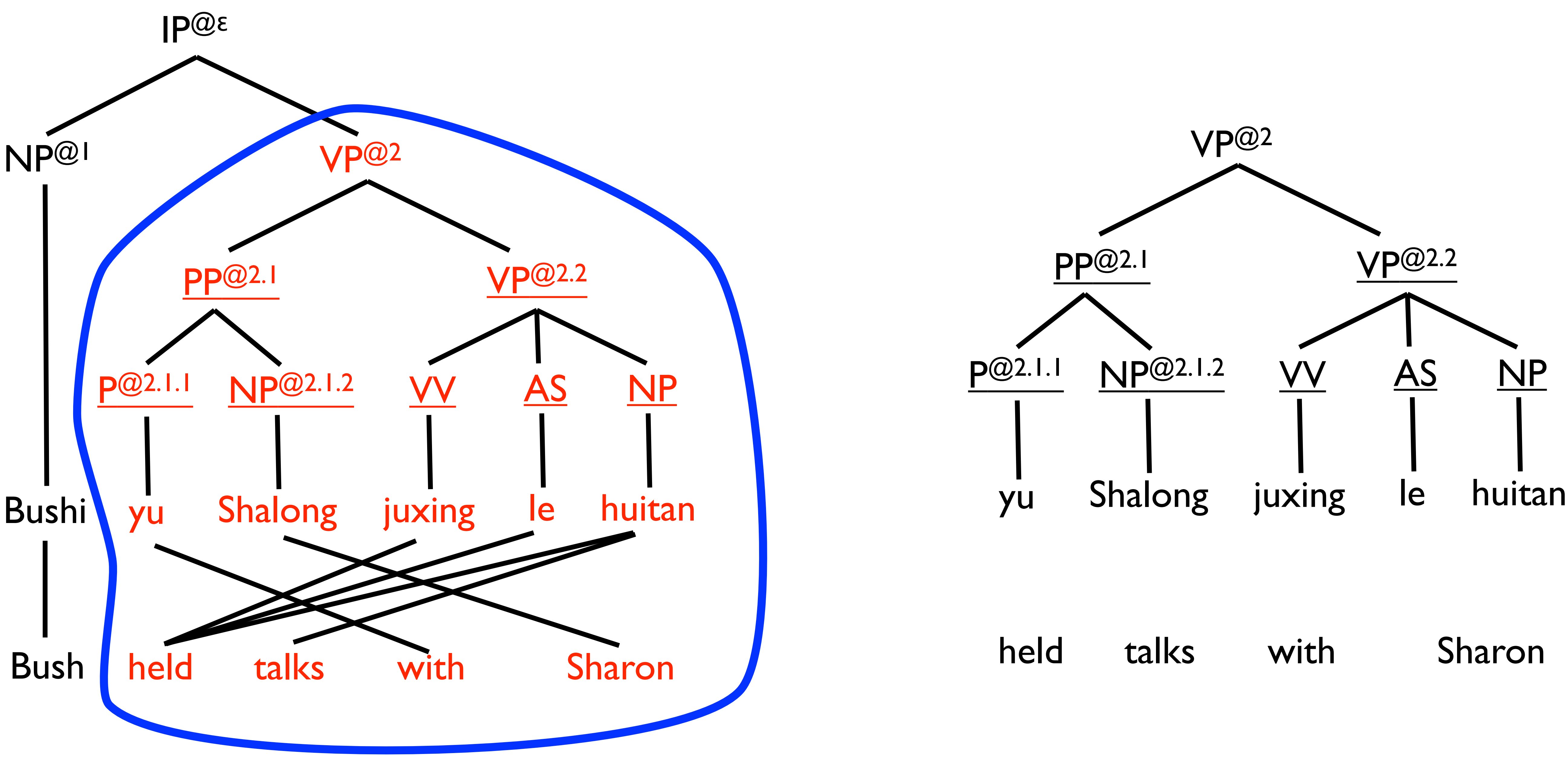


Syntax Based Machine Translation





Syntax Based Machine Translation



Goal

	BLEU	Decoding time	Grammar size
composed	✓		
minimal		✓	✓
??	✓	✓	✓

Our Approach: Rule Markov Models

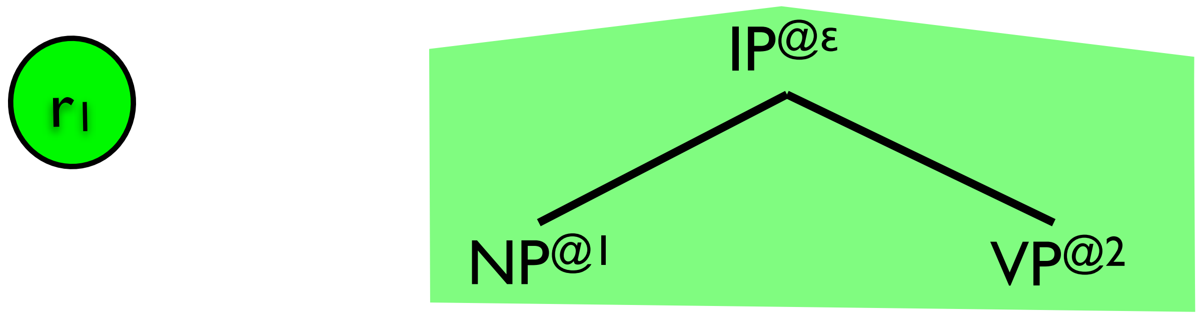
- composed rules alleviate independence b/w rules
- but we can explicitly model this dependence by conditioning on the history (grammar annotation)

- 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

IP@ε

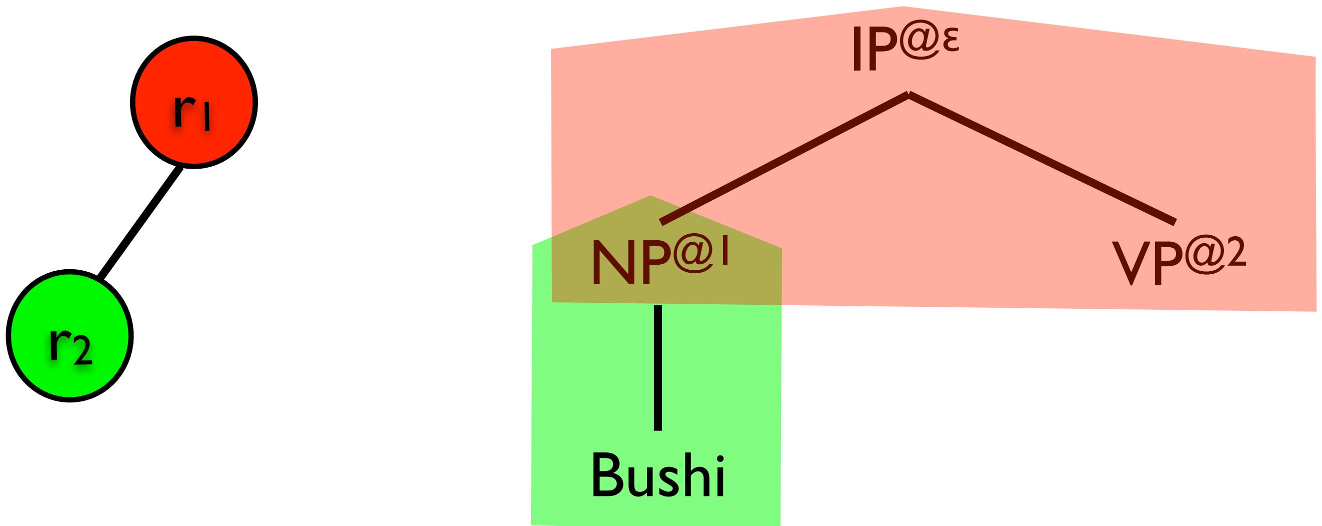
Generative Story



$P(r_1|\varepsilon)$

NP@1 VP@2

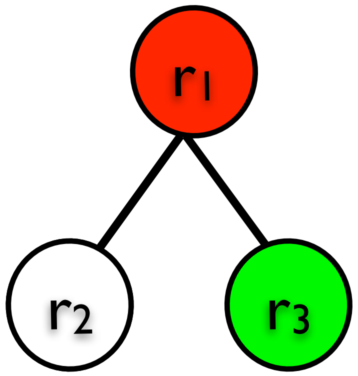
Generative Story

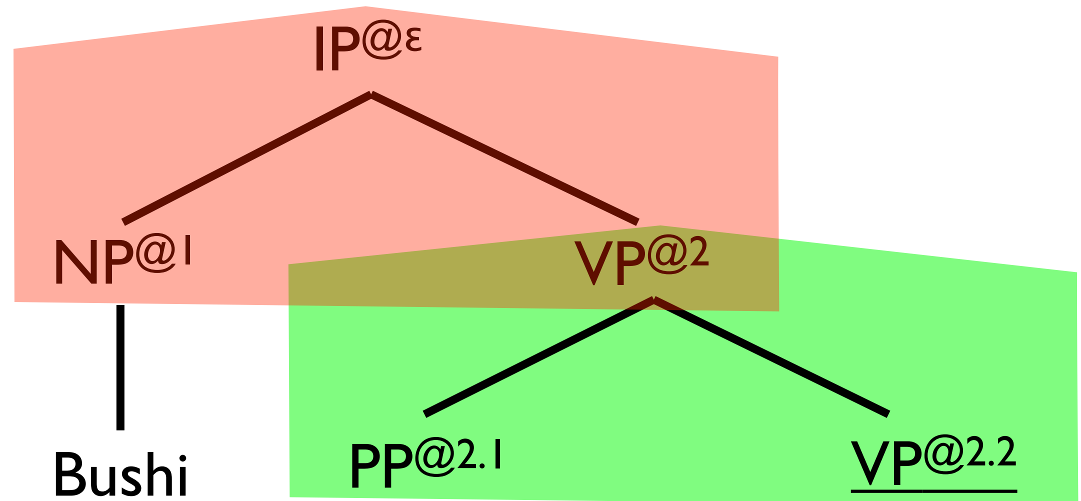


$P(r_2|r_1)$

Bush VP@2

Generative Story

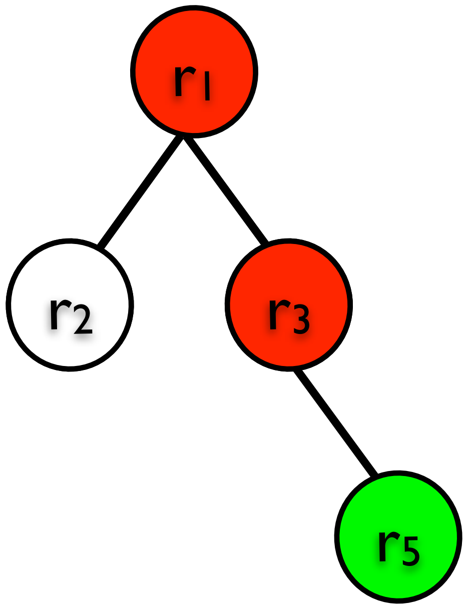


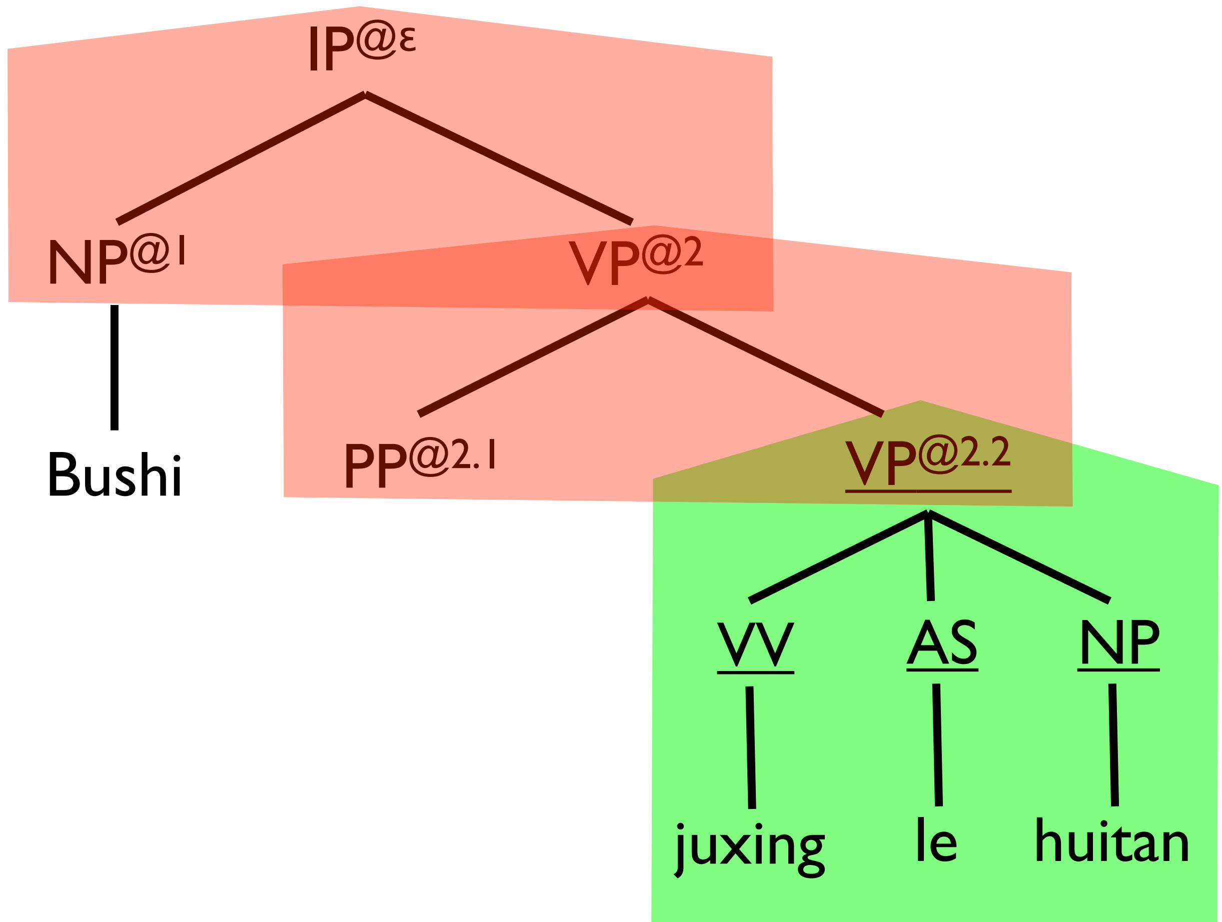


$P(r_3|r_1)$

Bush VP@2.2 PP@2.1

Generative Story

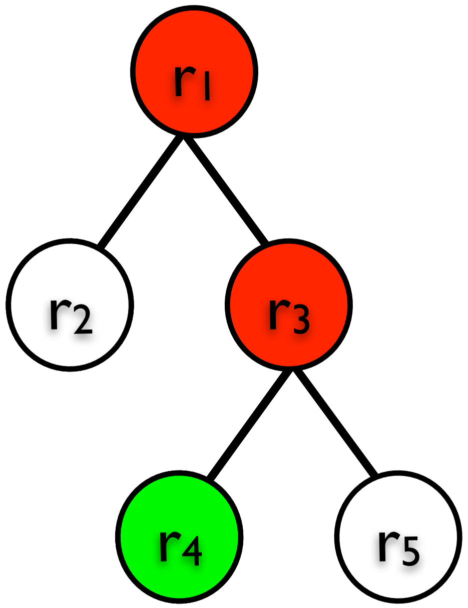


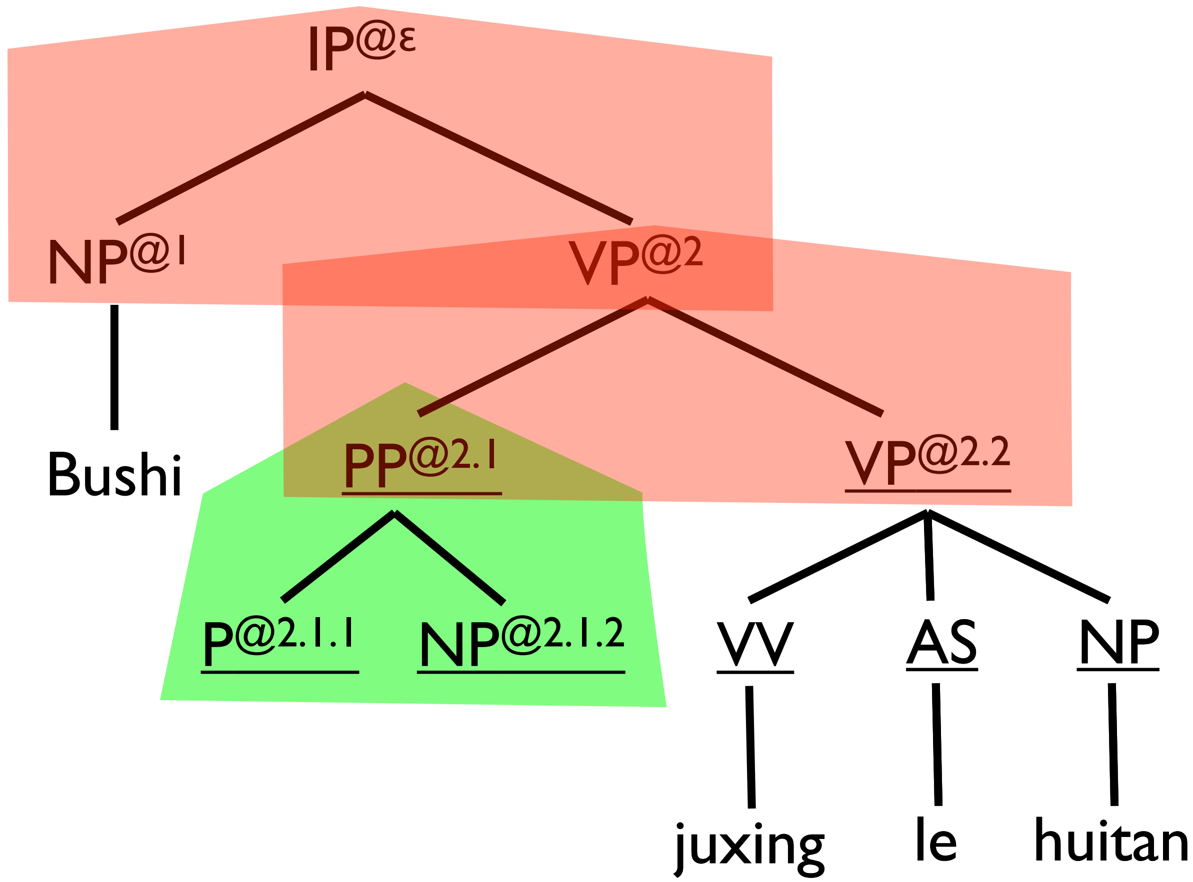


$P(r_5|r_1,r_3)$

Bush held talks PP@2.1

Generative Story

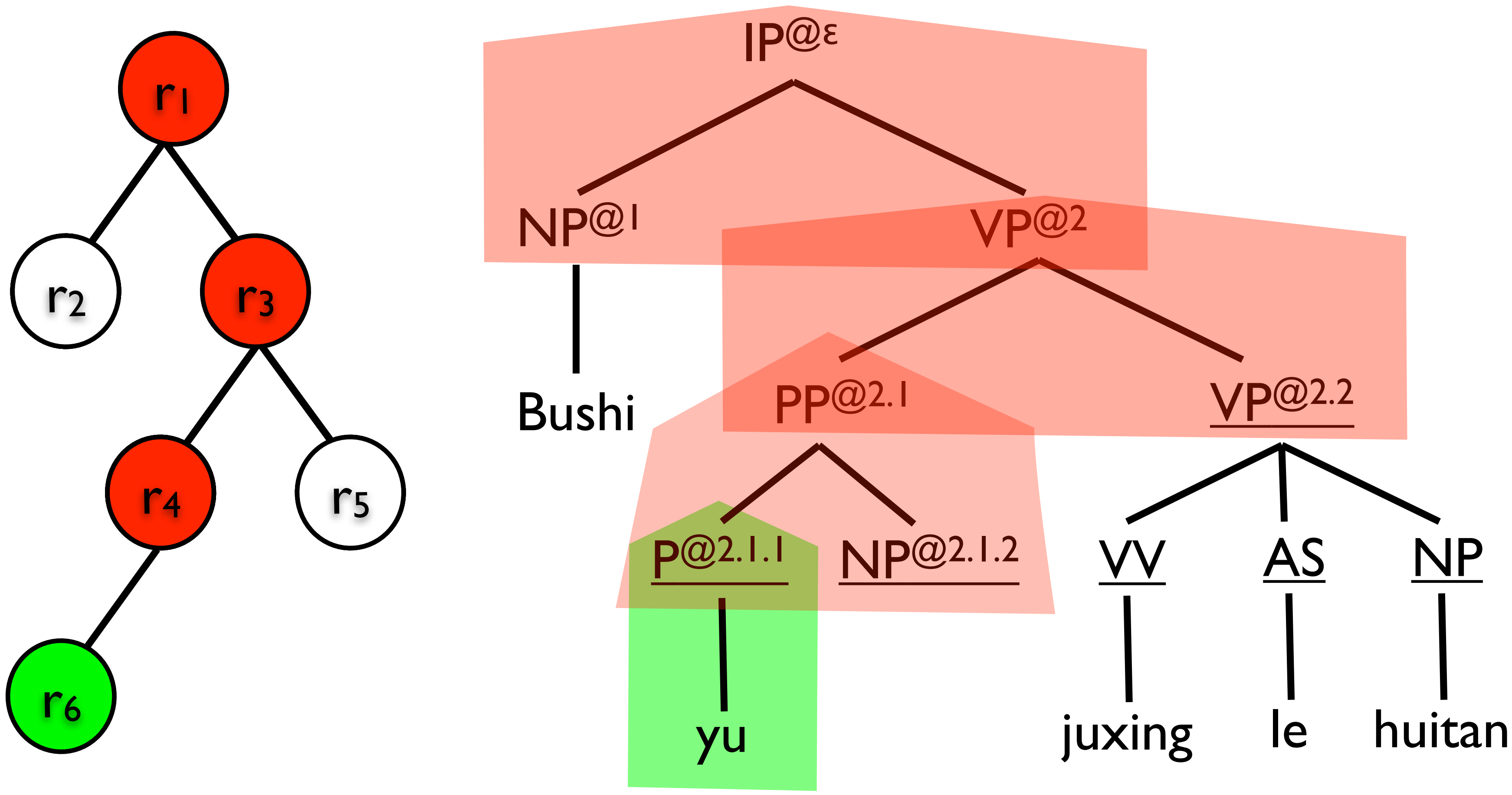




$P(r_4|r_1,r_3)$

Bush held talks P@2.1.1 NP@2.1.2

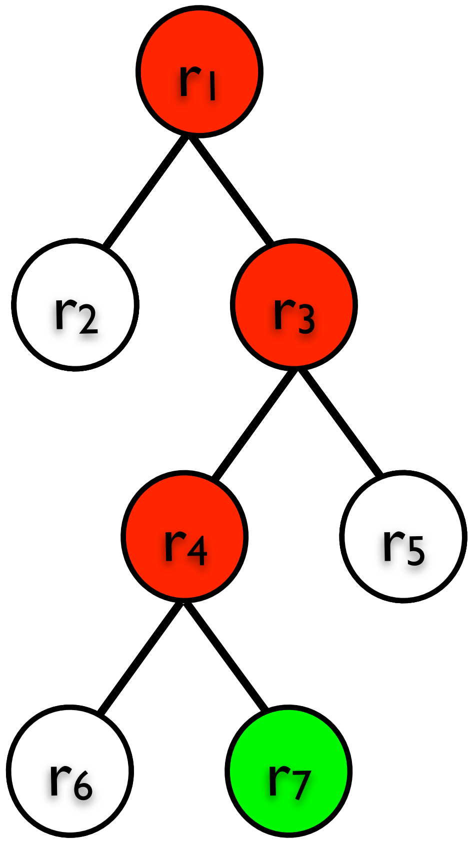
Generative Story

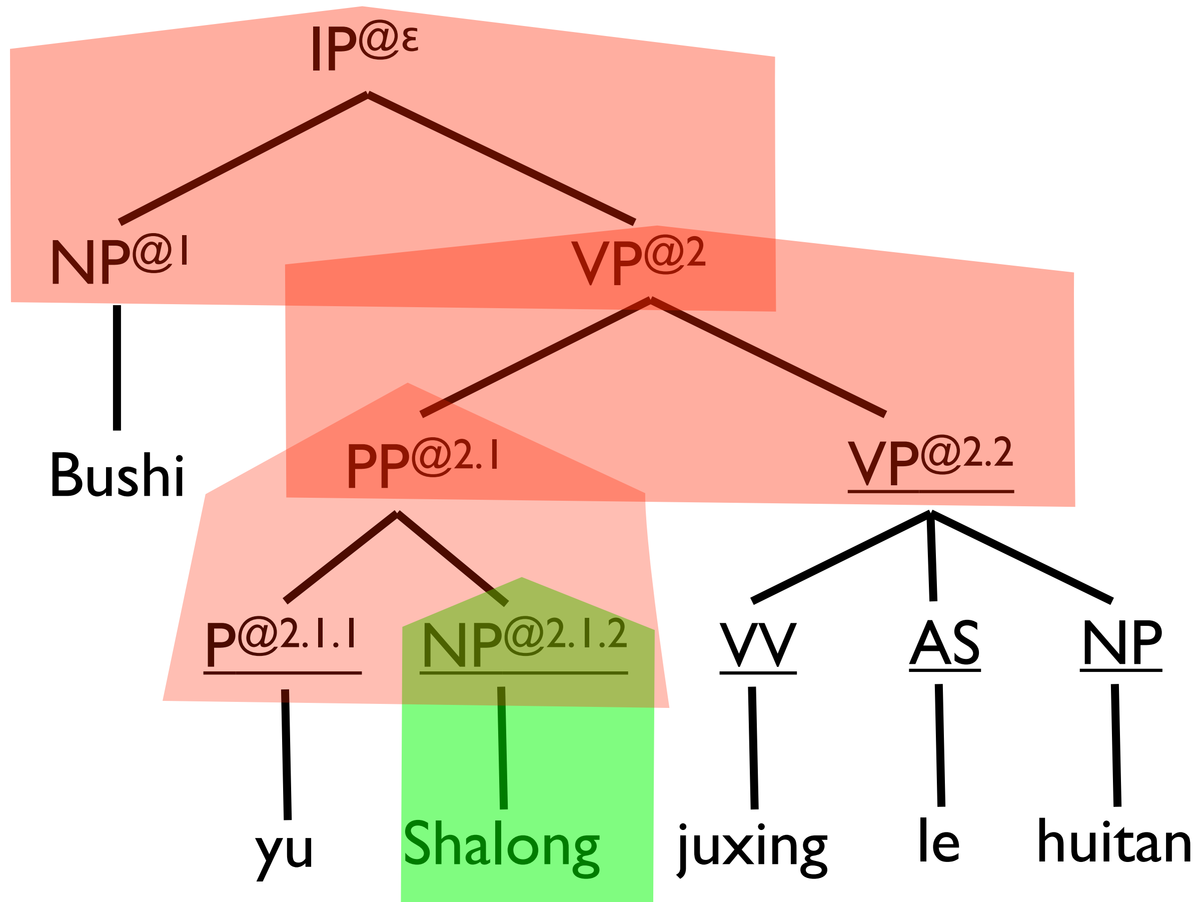


$P(r_6|r_1,r_3,r_4)$

Bush held talks with NP@2.1.2

Generative Story

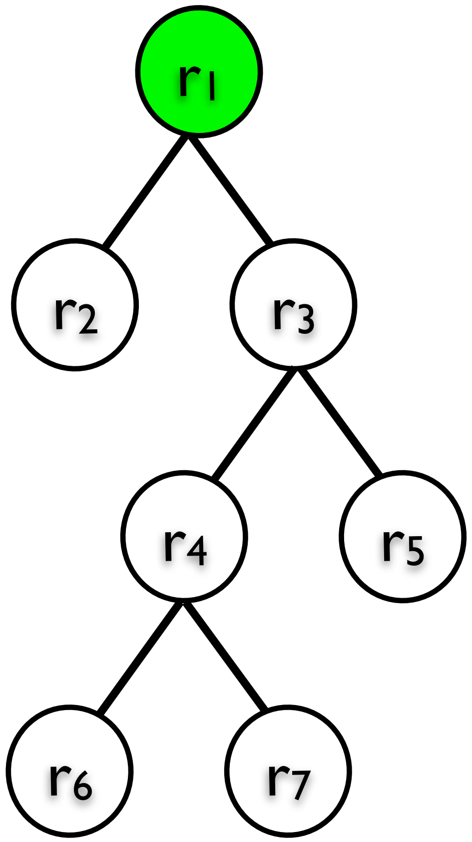




$P(r_7|r_1,r_3,r_4)$

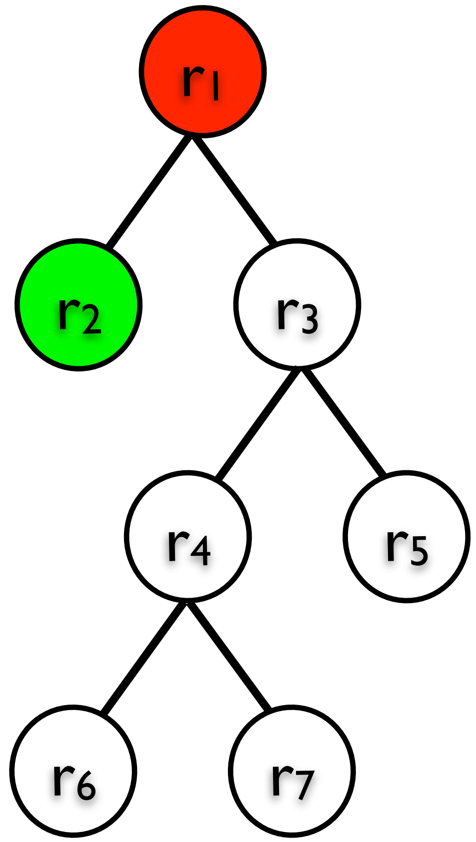
Bush held talks with Sharon

Probability of a Derivation Tree



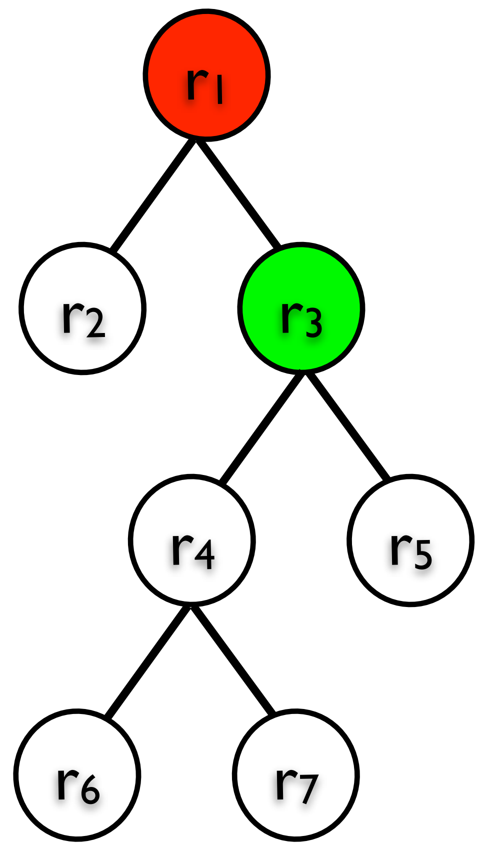
$$P(T) = P(r_1|\varepsilon) \cdot P(r_2|r_1) \cdot P(r_3|r_1)$$
$$P(r_5|r_1,r_3) \cdot P(r_4|r_1,r_3)$$
$$P(r_6|r_1,r_3,r_4) \cdot P(r_7|r_1,r_3,r_4)$$

Probability of a Derivation Tree



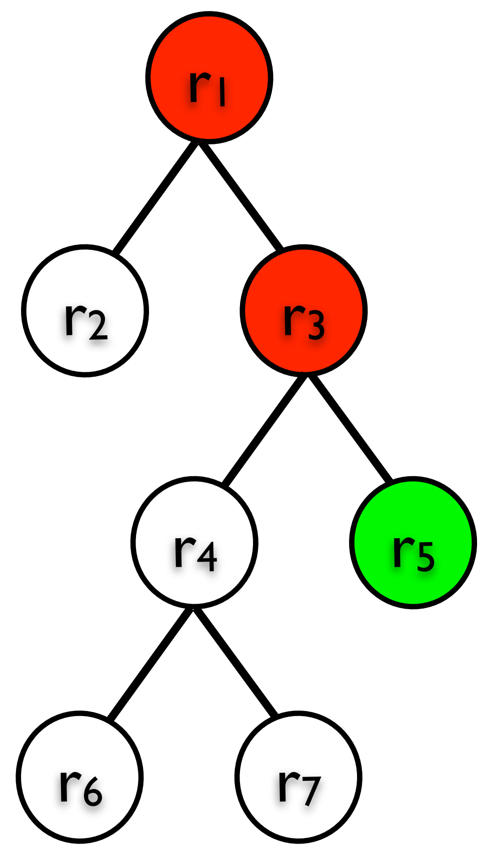
$$P(T) = P(r_1|\varepsilon) \cdot P(r_2|r_1) \cdot P(r_3|r_1)$$
$$P(r_5|r_1,r_3) \cdot P(r_4|r_1,r_3)$$
$$P(r_6|r_1,r_3,r_4) \cdot P(r_7|r_1,r_3,r_4)$$

Probability of a Derivation Tree



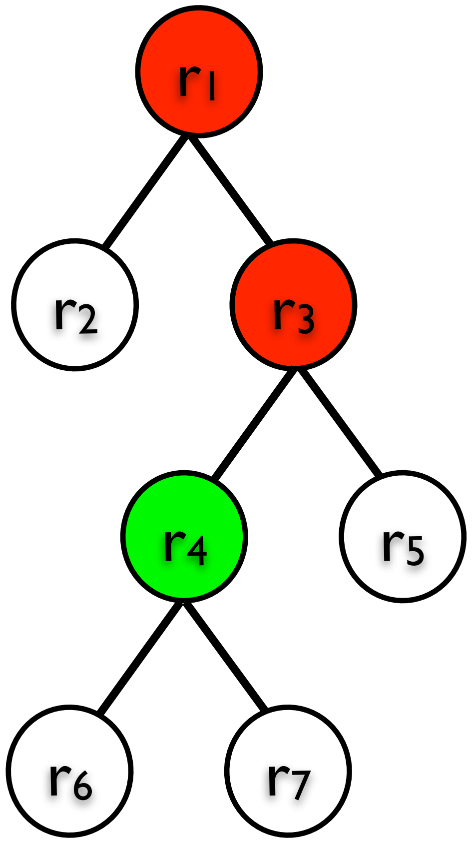
$$P(T) = P(r_1|\varepsilon) \cdot P(r_2|r_1) \cdot P(r_3|r_1)$$
$$P(r_5|r_1,r_3) \cdot P(r_4|r_1,r_3)$$
$$P(r_6|r_1,r_3,r_4) \cdot P(r_7|r_1,r_3,r_4)$$

Probability of a Derivation Tree



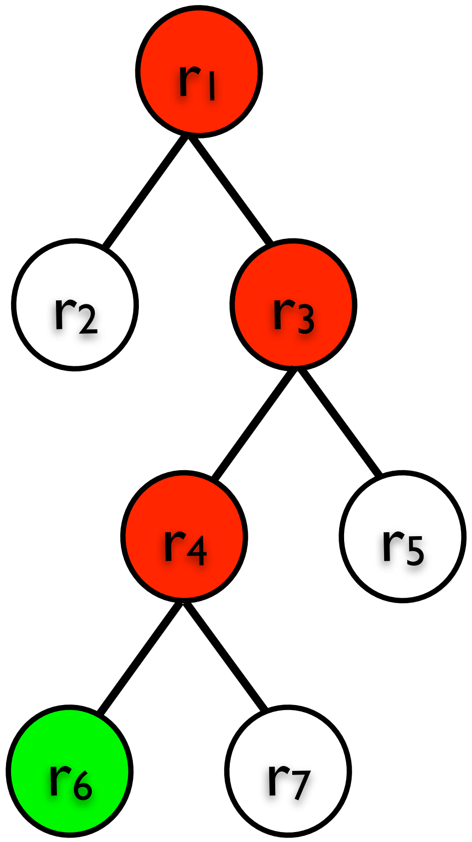
$$P(T) = P(r_1|\varepsilon) \cdot P(r_2|r_1) \cdot P(r_3|r_1)$$
$$P(r_5|r_1,r_3) \cdot P(r_4|r_1,r_3)$$
$$P(r_6|r_1,r_3,r_4) \cdot P(r_7|r_1,r_3,r_4)$$

Probability of a Derivation Tree



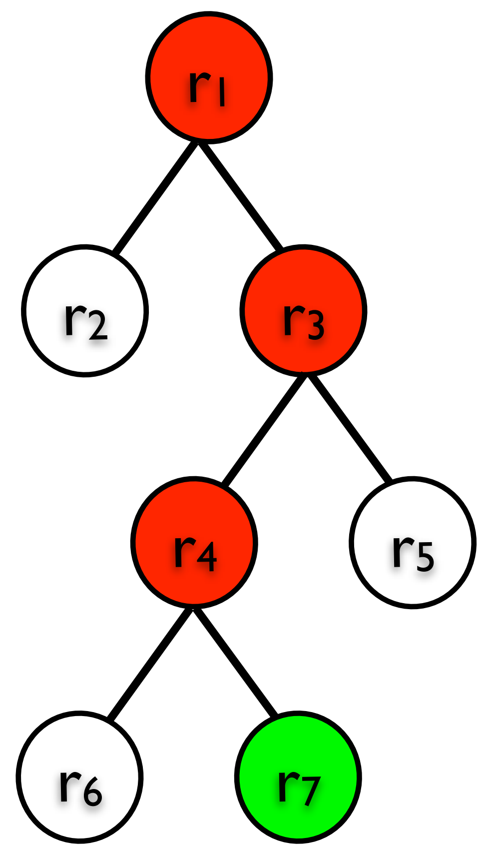
$$P(T) = P(r_1|\varepsilon) \cdot P(r_2|r_1) \cdot P(r_3|r_1)$$
$$P(r_5|r_1,r_3) \cdot P(r_4|r_1,r_3)$$
$$P(r_6|r_1,r_3,r_4) \cdot P(r_7|r_1,r_3,r_4)$$

Probability of a Derivation Tree



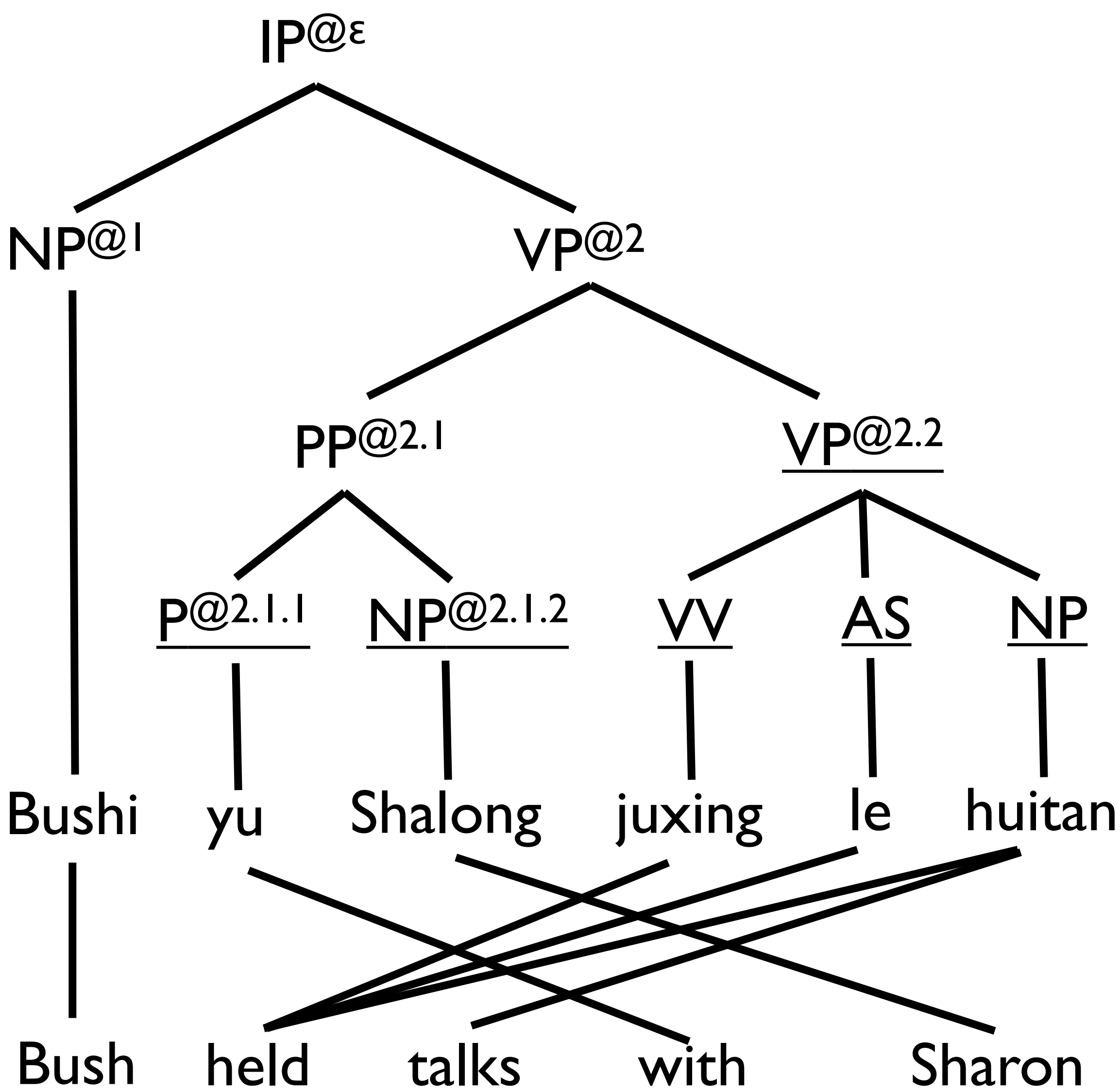
$$P(T) = P(r_1|\varepsilon) \cdot P(r_2|r_1) \cdot P(r_3|r_1)$$
$$P(r_5|r_1,r_3) \cdot P(r_4|r_1,r_3)$$
$$P(r_6|r_1,r_3,r_4) \cdot P(r_7|r_1,r_3,r_4)$$

Probability of a Derivation Tree

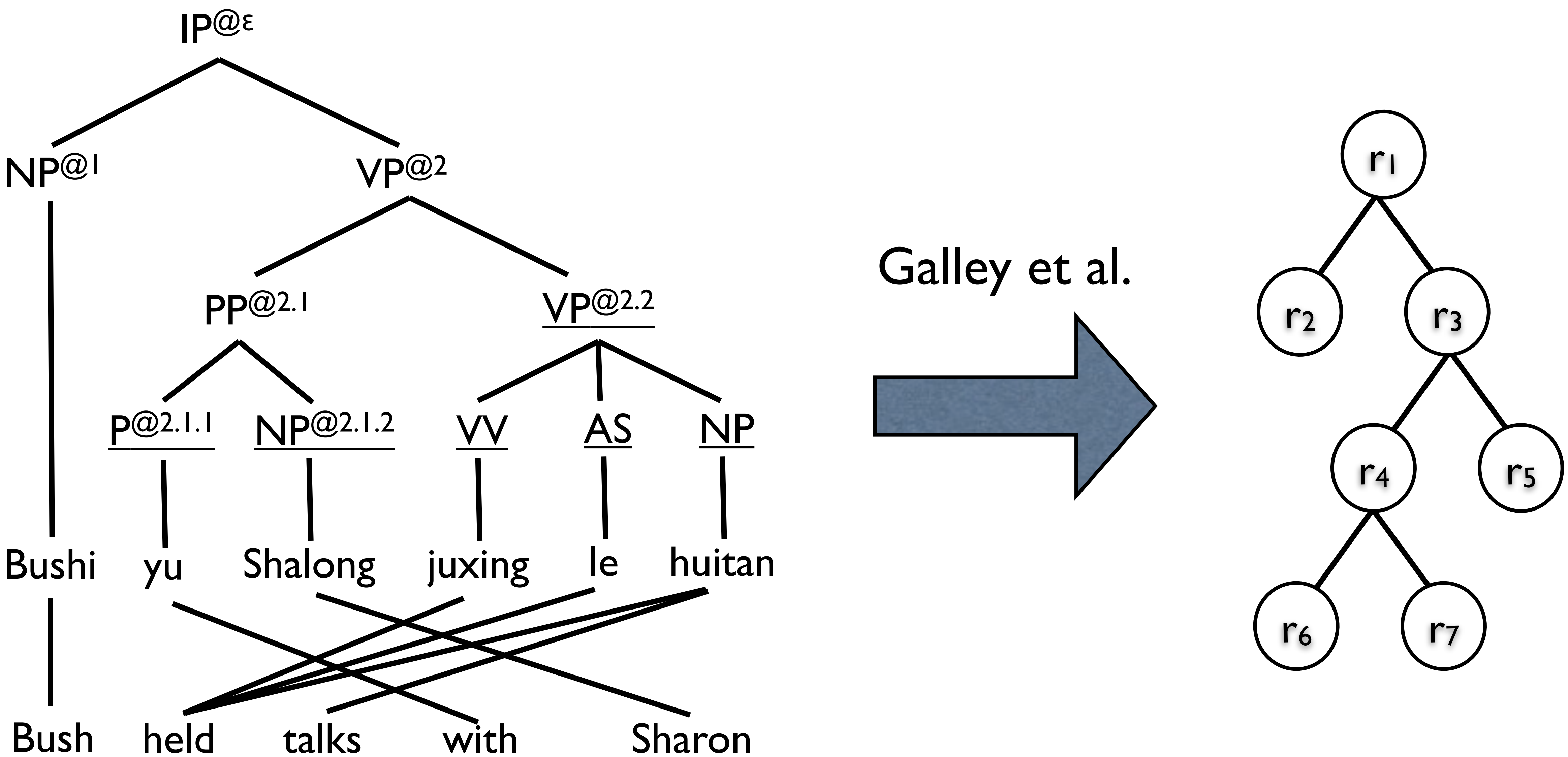


$$P(T) = P(r_1|\varepsilon) \cdot P(r_2|r_1) \cdot P(r_3|r_1)$$
$$P(r_5|r_1,r_3) \cdot P(r_4|r_1,r_3)$$
$$P(r_6|r_1,r_3,r_4) \cdot P(r_7|r_1,r_3,r_4)$$

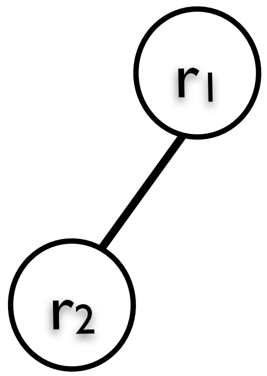
Learning Rule Markov Models

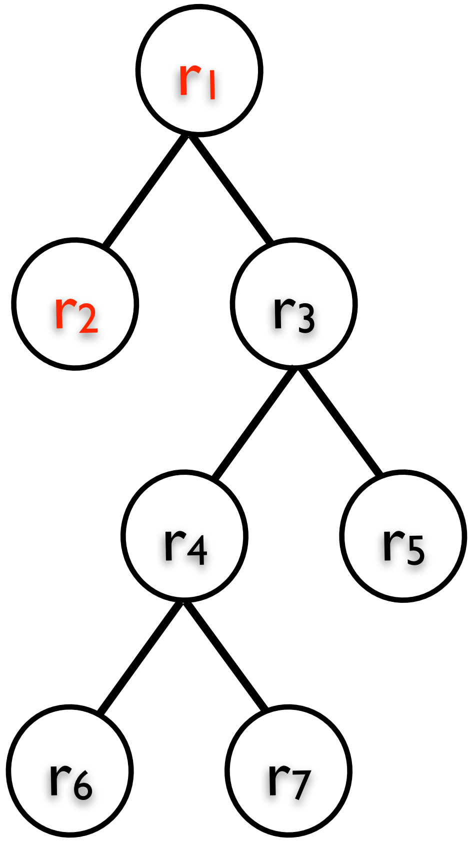


Learning Rule Markov Models

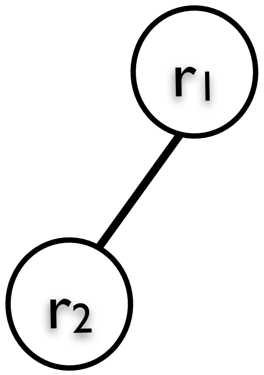


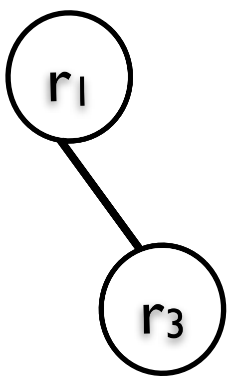
Learning Rule Markov Models

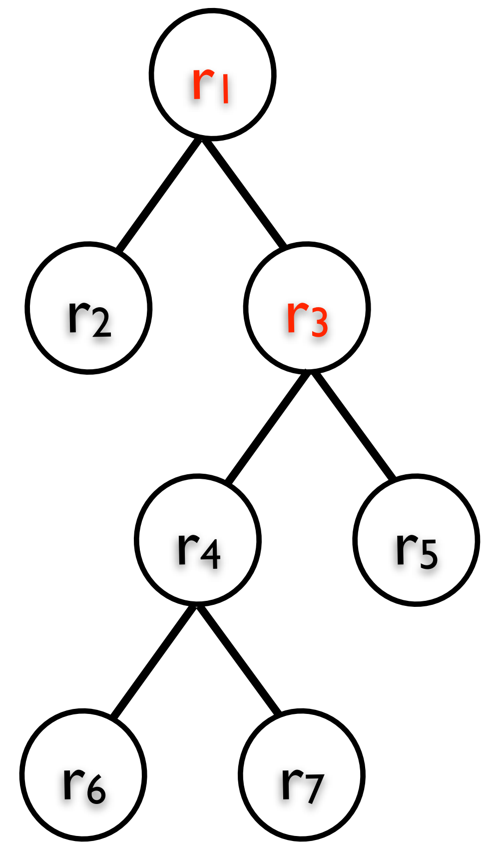




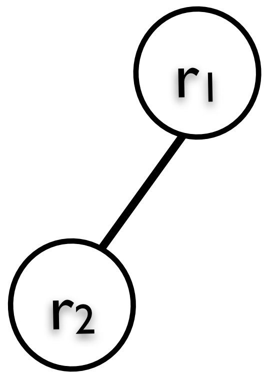
Learning Rule Markov Models

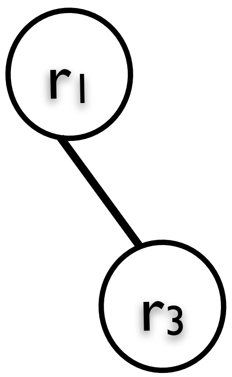


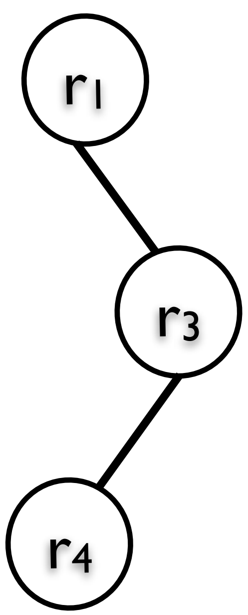


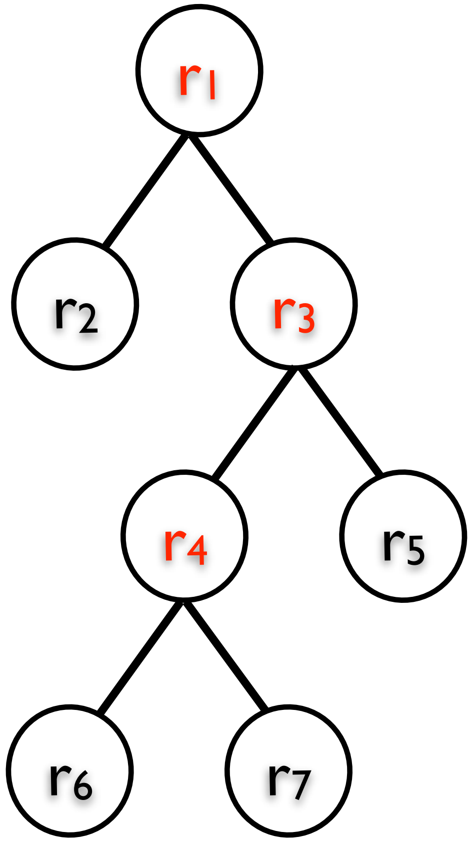


Learning Rule Markov Models









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
- RM-C: Build Prediction Suffix Trees using the approach of Bejerano and Yona (1999)

Incremental decoding with RMMs

Incremental decoding with RMMs

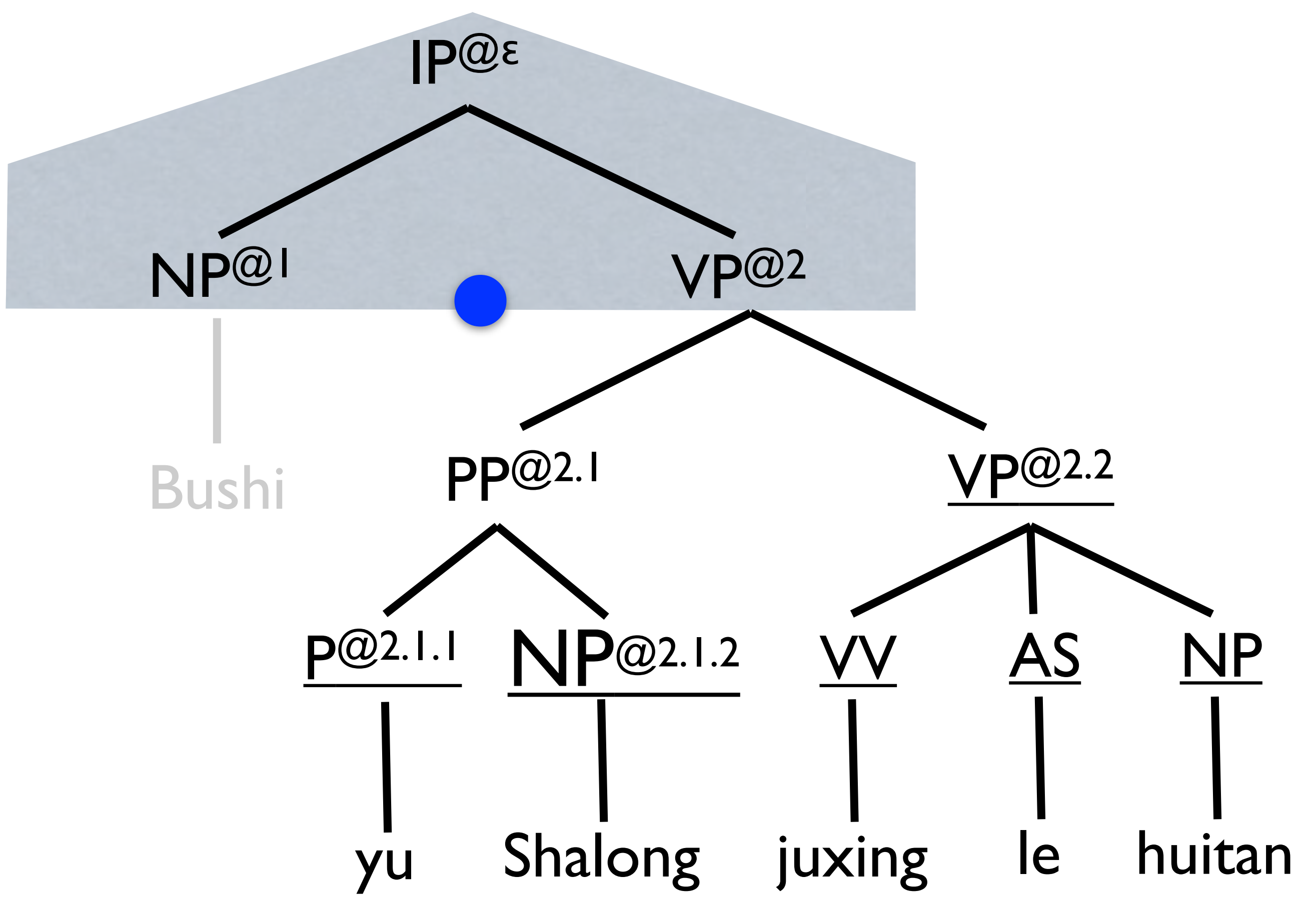
- rule Markov models using only vertical context are a natural fit for Incremental decoding (Huang and Mi, 2010)

Incremental decoding with RMMs

- 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

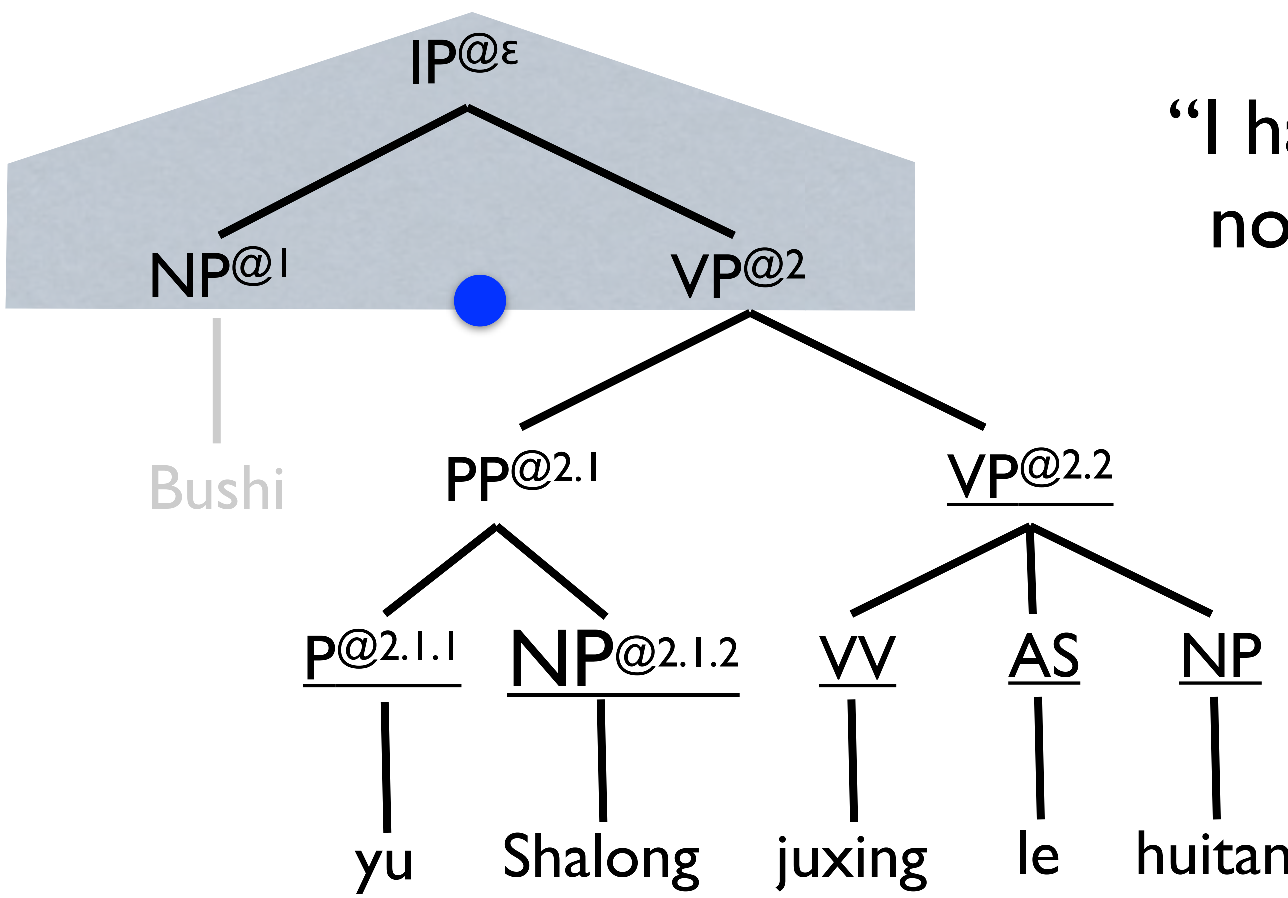
- stack (*active* derivation history): [.IP@ε] [NP@1 .VP@2]
- three colors for nodes: white (uncovered), grey (partially covered), and black (covered)



(Huang and Mi, 2010)

Incremental Decoding

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



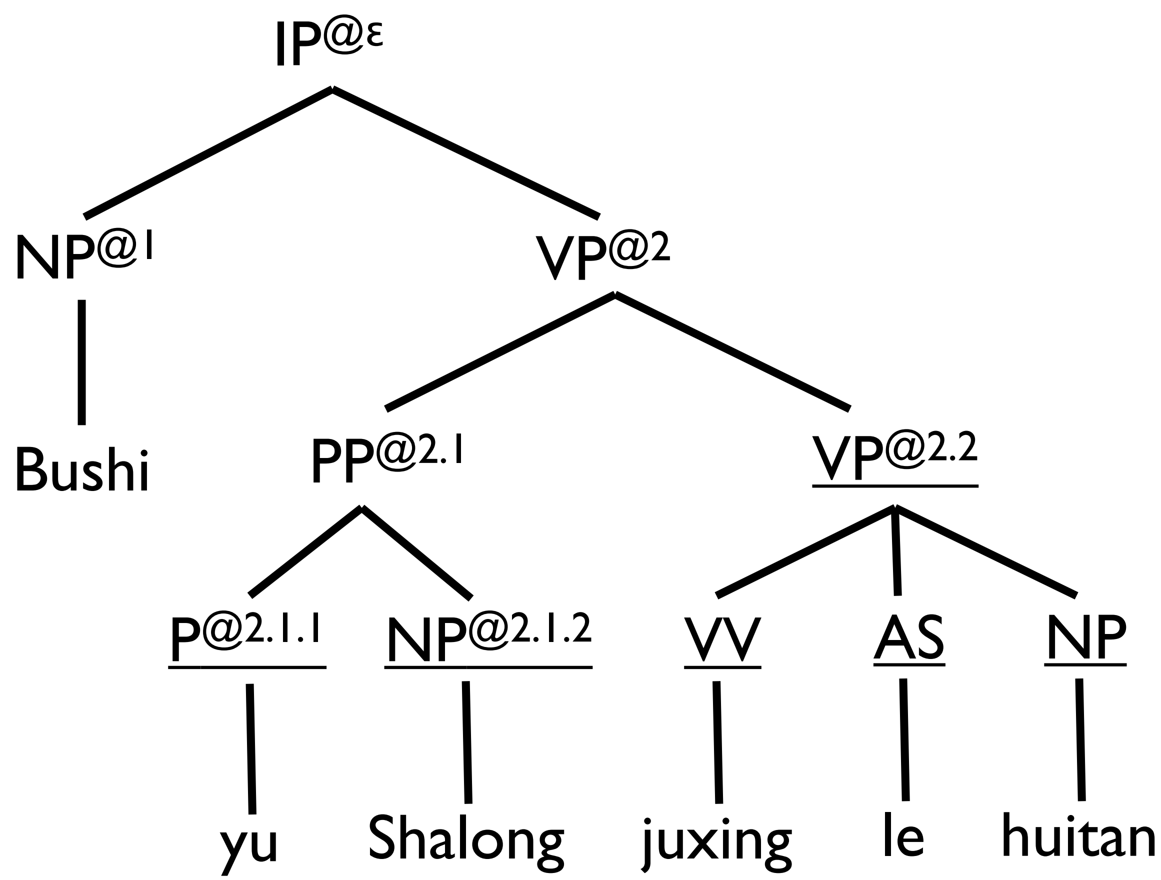
"I have finished NP subtree but not started with VP subtree"

(Huang and Mi, 2010)

Example Incremental Decoding

[<s> . IP@ε </s>] *stack*

<s> *hypothesis*



(Huang and Mi, 2010)

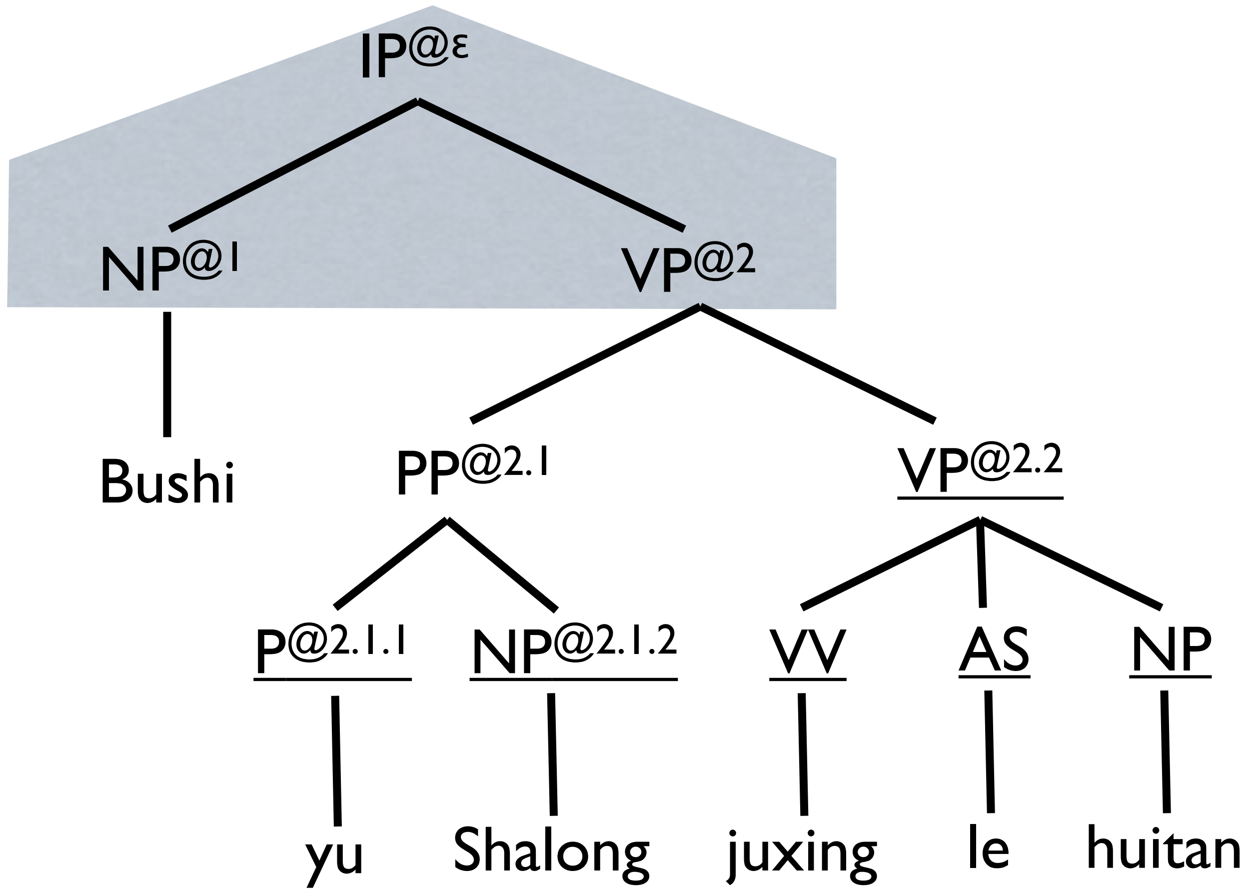
Example Incremental Decoding

[<s> . $IP^{@\varepsilon}$ </s>] [. $NP^{@1}$ $VP^{@2}$]
r_1

stack

<s>

hypothesis



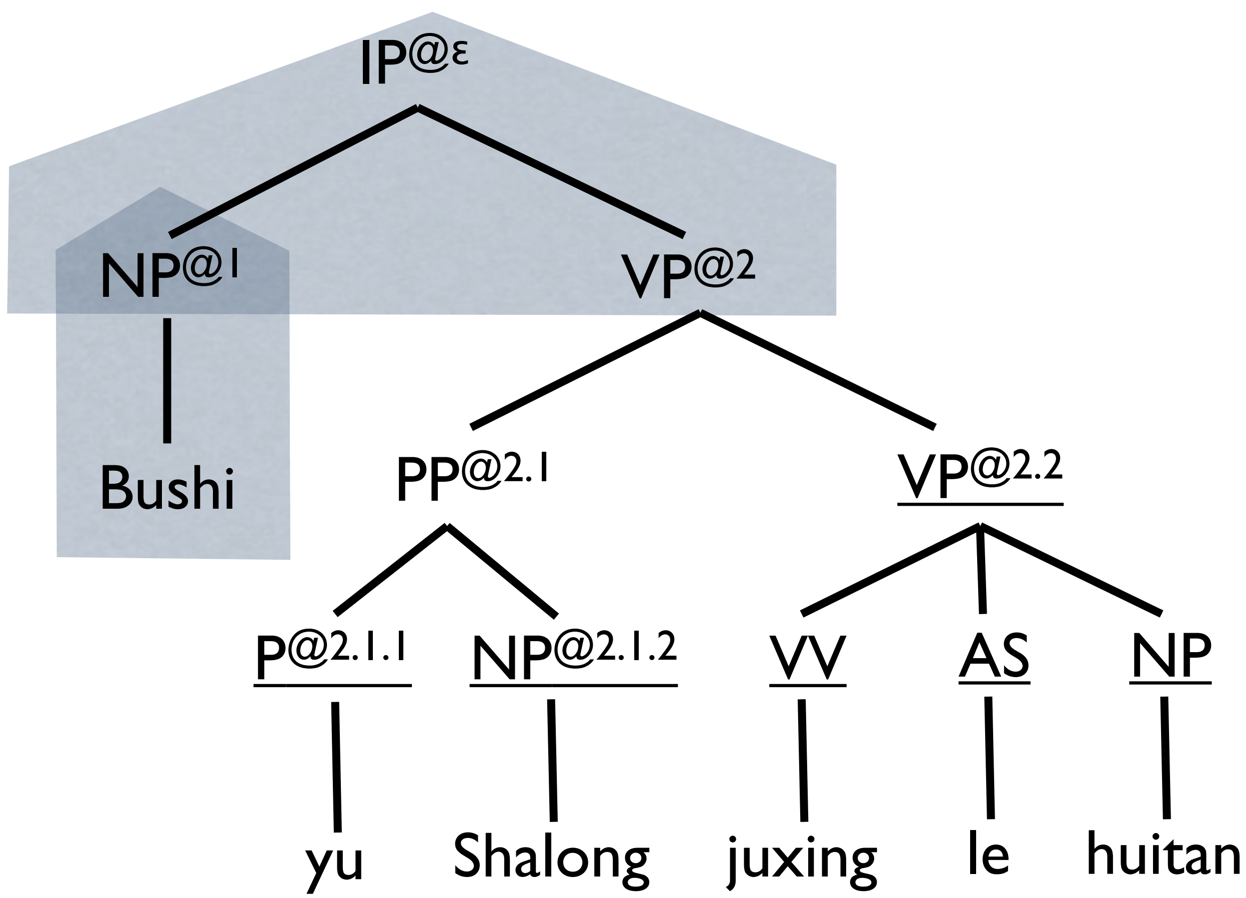
action: predict (push)

(Huang and Mi, 2010)

Example Incremental Decoding

[<s> . IP@ε </s>] [. NP@1 VP@2] [. Bush] *stack*
r1 r2

<s> *hypothesis*



action: predict

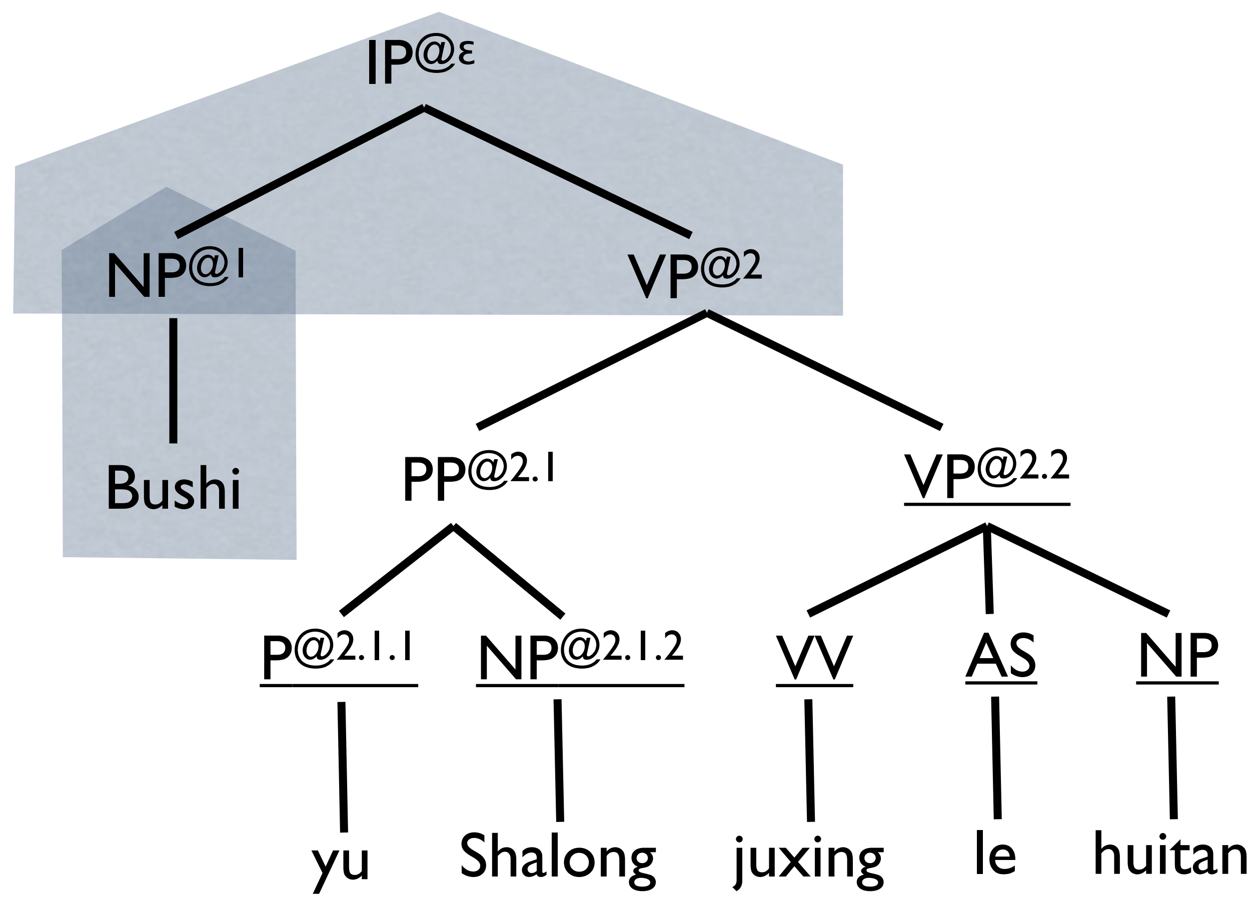
(Huang and Mi, 2010)

Example Incremental Decoding

[<s> . IP$^{@\varepsilon}$ </s>] [. NP$^{@1}$ VP$^{@2}$] [Bush .] *stack*

r_1 r_2

<s> Bush *hypothesis*



action: scan

(Huang and Mi, 2010)

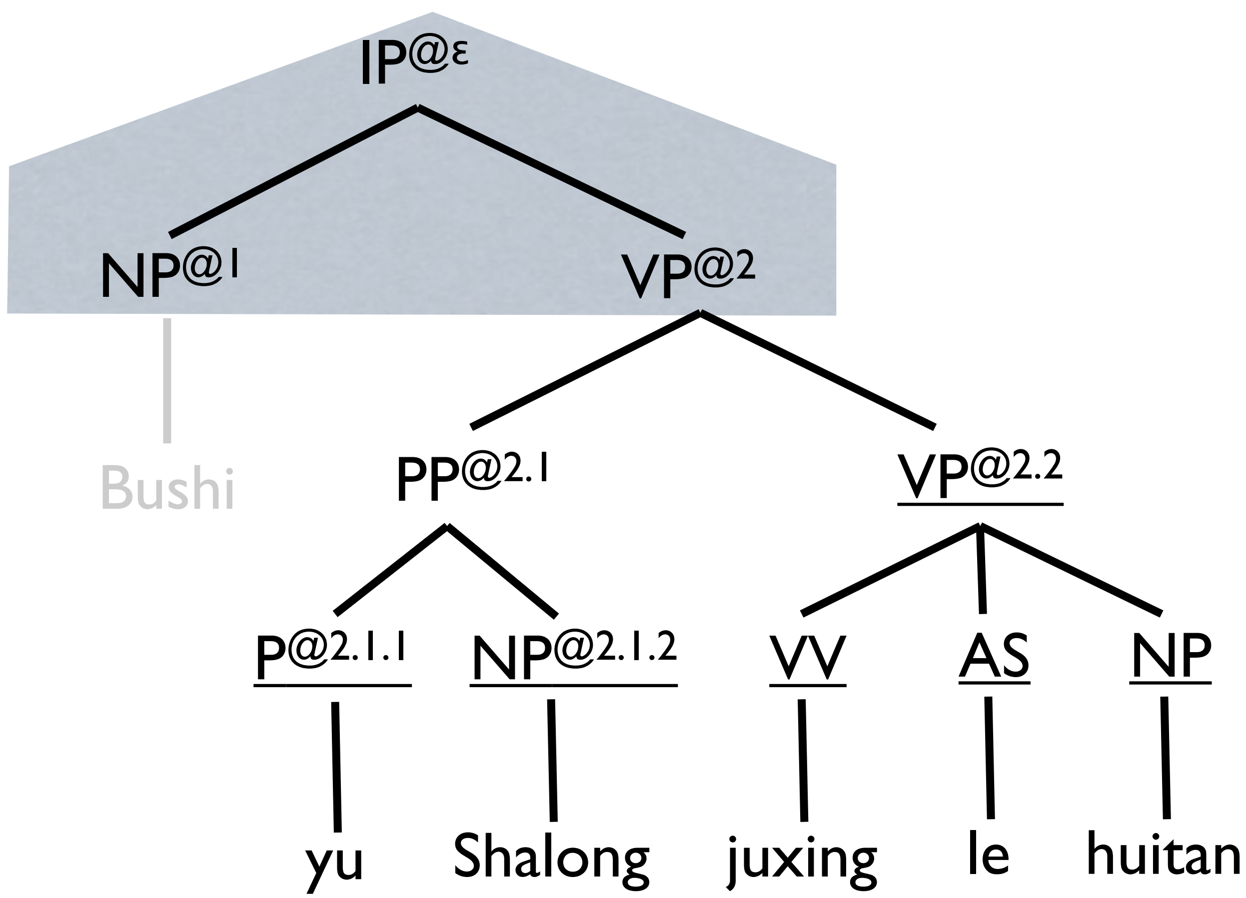
Example Incremental Decoding

[<s> . IP@ε </s>] [NP@1 . VP@2]
r_1

stack

<s> Bush

hypothesis



action: pop

(Huang and Mi, 2010)

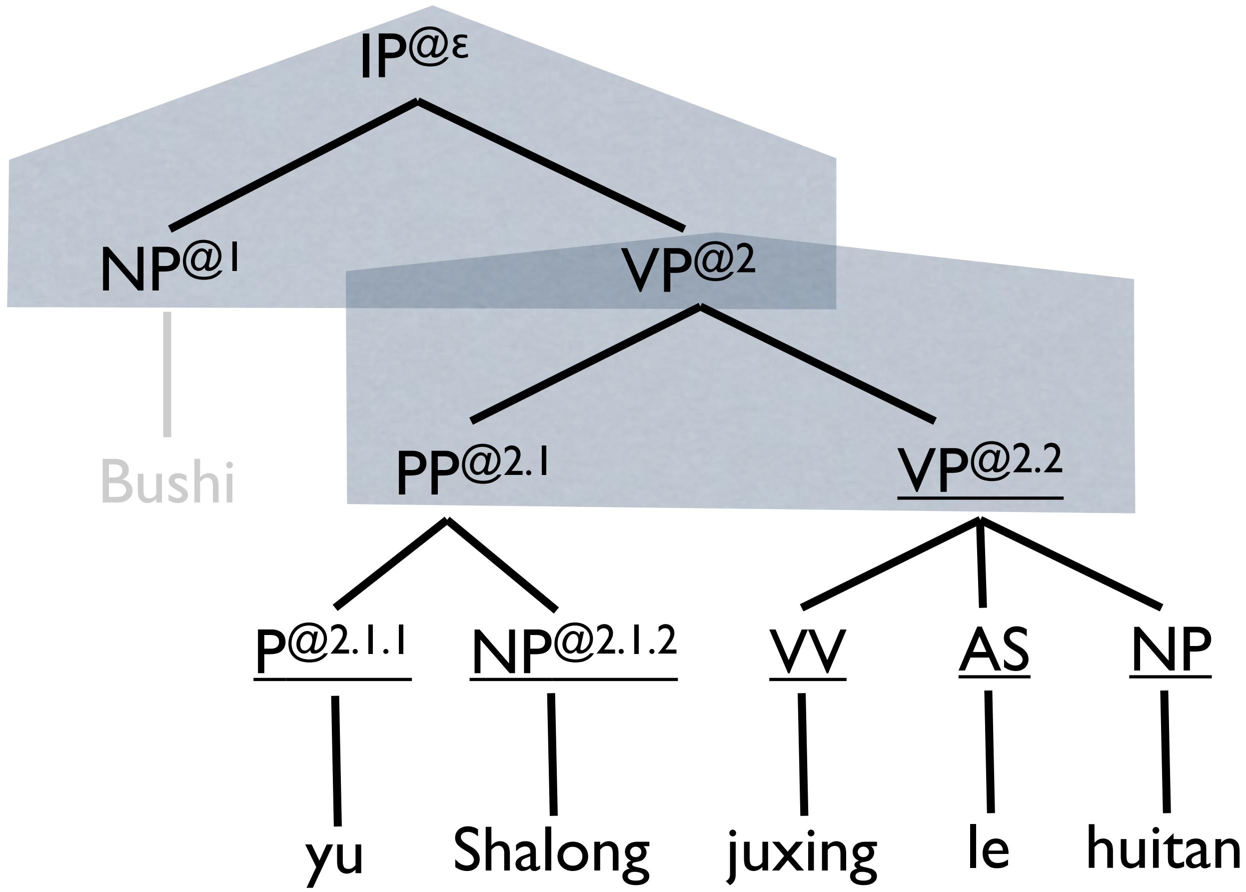
Example Incremental Decoding

[<s> . IP@ε </s>] [NP@1 . VP@2] [. VP@2.2 PP@2.1]
r_1 r_3

stack

<s> Bush

hypothesis



action: predict

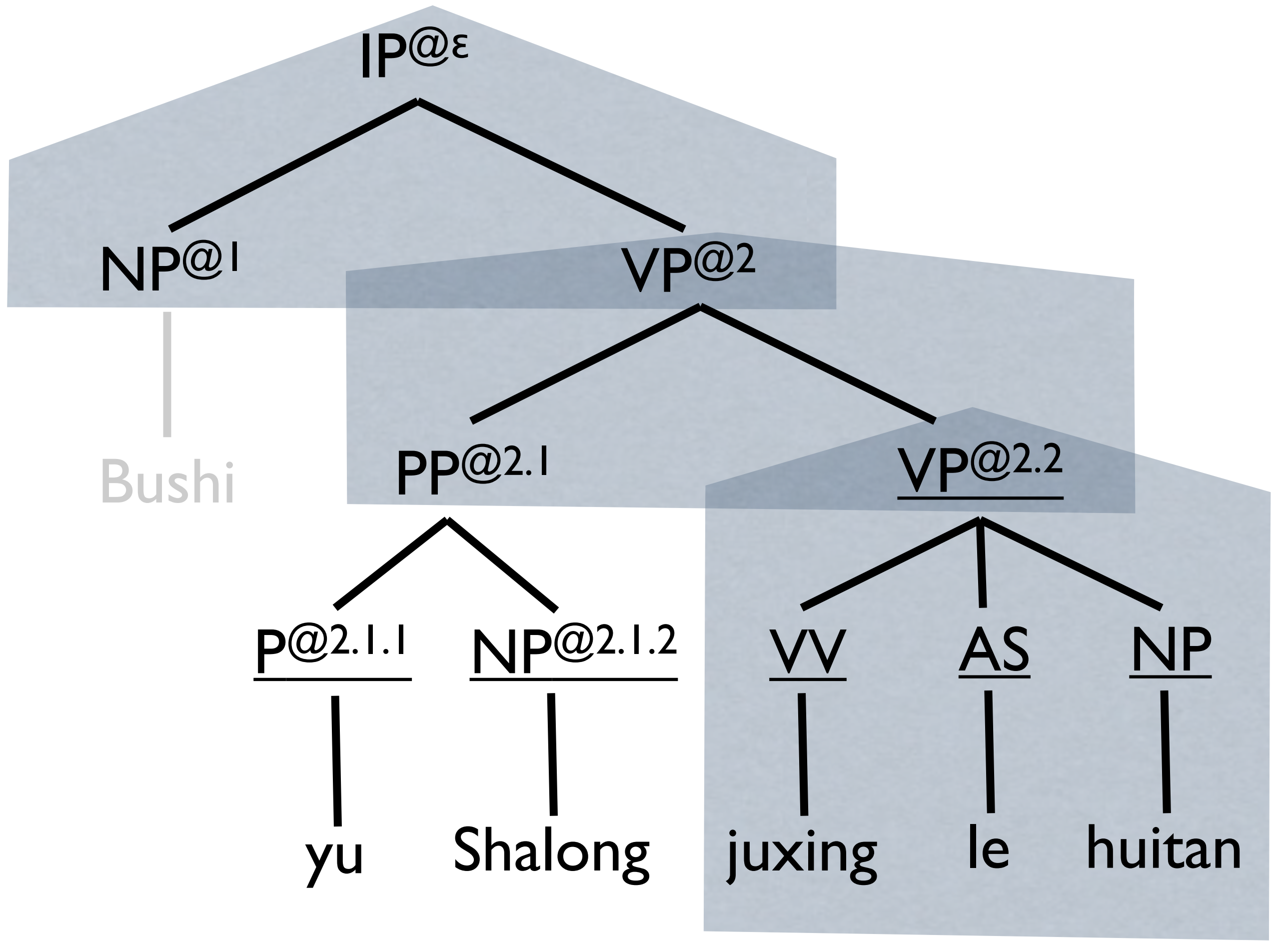
(Huang and Mi, 2010)

Example Incremental Decoding

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

r_1 r_3 r_5

<s> Bush *hypothesis*



action: predict

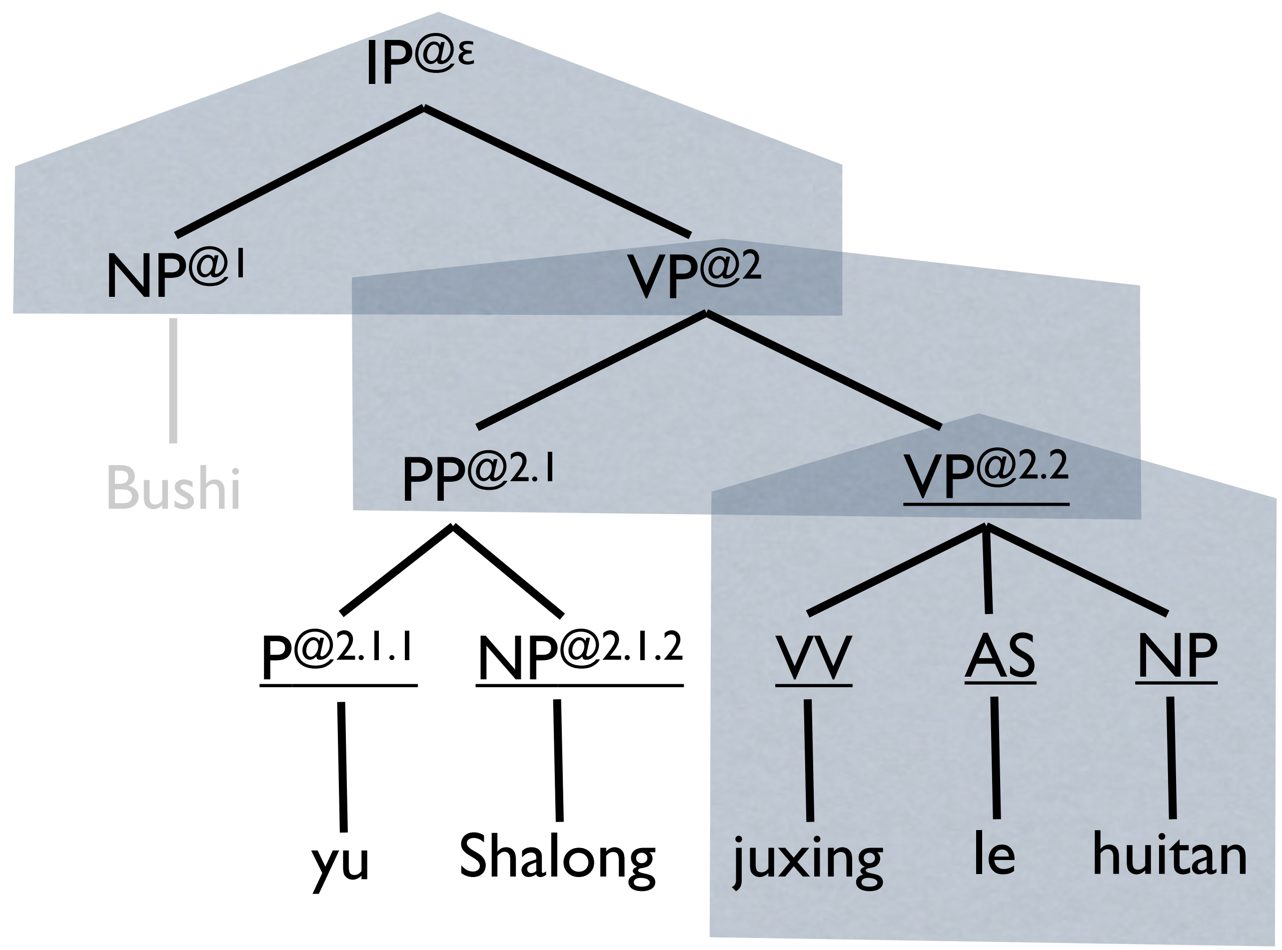
(Huang and Mi, 2010)

Example Incremental Decoding

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

r1 r3 r5

<s> Bush held talks *hypothesis*



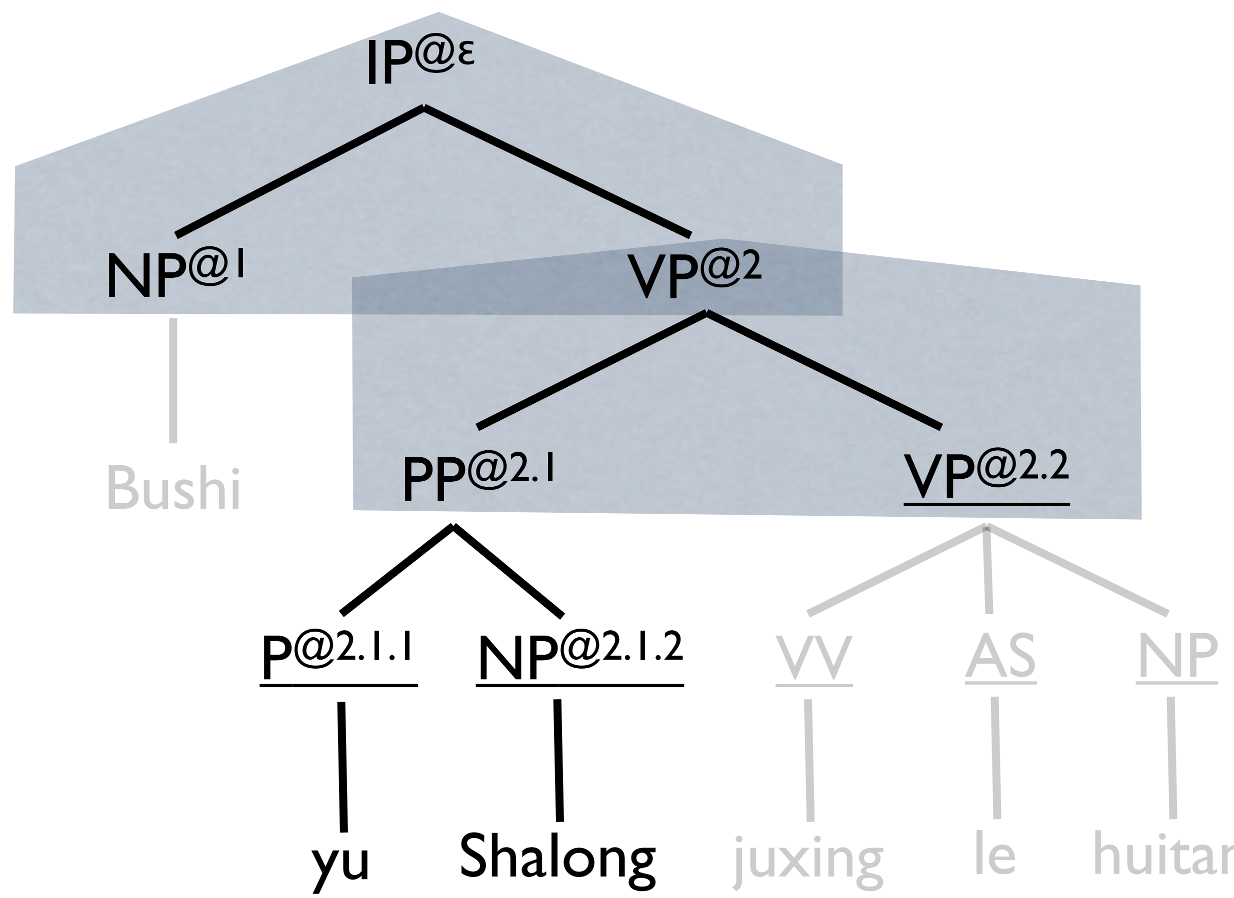
action: scan

(Huang and Mi, 2010)

Example Incremental Decoding

[<s> . IP@ε </s>] [NP@1 $\underset{r_1}{.}$ VP@2] [. $\underline{\text{VP@2.2}}_{\ r_3}$ $\underline{\text{PP@2.1}}$] *stack*

<s> Bush held talks *hypothesis*



action: pop

(Huang and Mi, 2010)

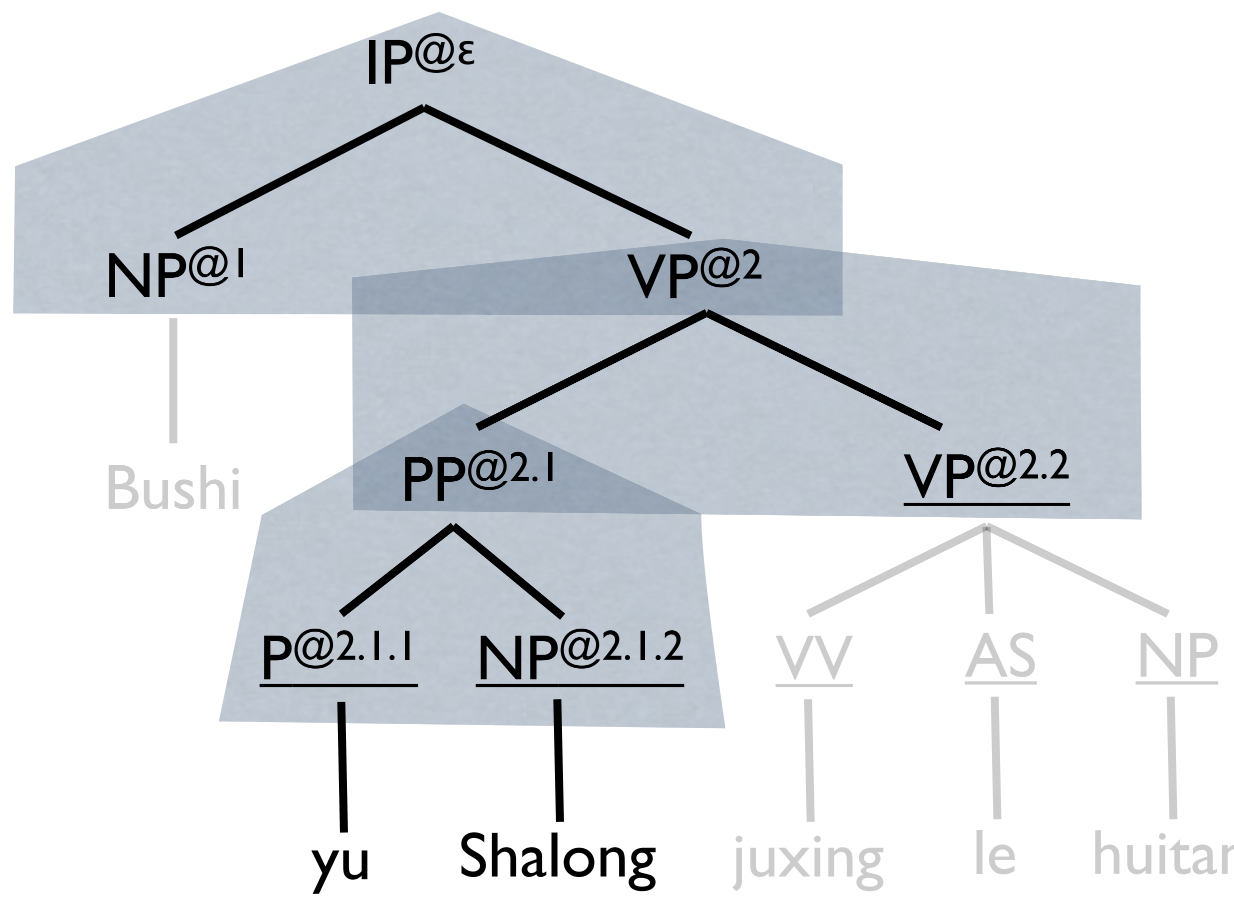
Example Incremental Decoding

$$[\langle s\rangle \bullet \mathrm{IP}^{@\epsilon} \langle /s\rangle]\ \underset{r_1}{[\mathrm{NP}^{@1} \bullet \mathrm{VP}^{@2}]}\ \underset{r_3}{[\underline{\mathrm{VP}^{@2.2}} \bullet \underline{\mathrm{PP}^{@2.1}}]}\ \underset{r_4}{[\bullet\ \underline{\mathrm{P}^{@2.1.1}}\ \underline{\mathrm{NP}^{@2.1.2}}]}$$

stack

<s> Bush held talks

hypothesis



action: predict

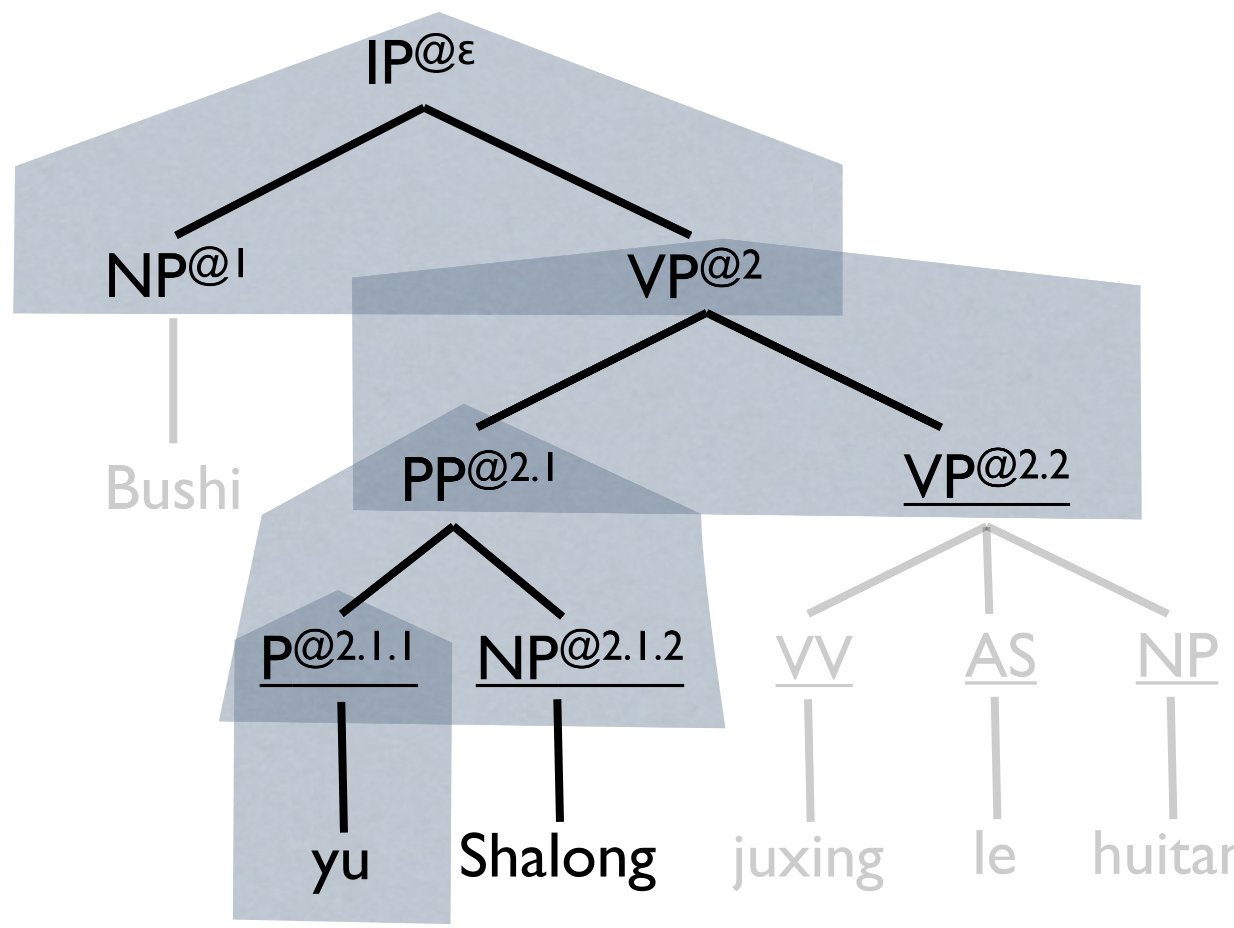
(Huang and Mi, 2010)

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] [. with/and] *stack*

r1 r3 r4 r_6/r'_6

<s> Bush held talks *hypothesis*



action: predict

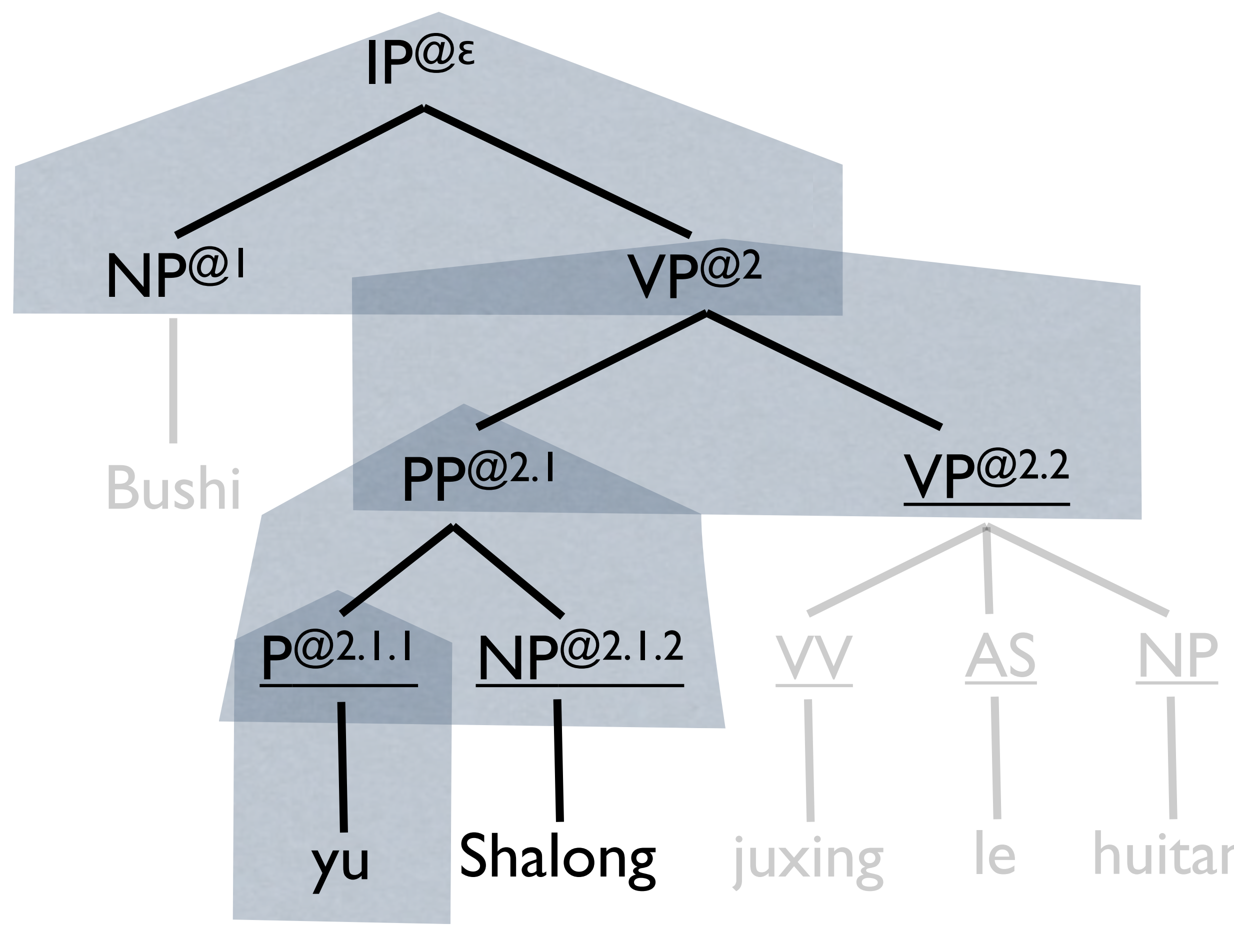
(Huang and Mi, 2010)

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] [and •] *stack*

r1 r3 r4 r'6

<s> Bush held talks and *hypothesis*



action: scan

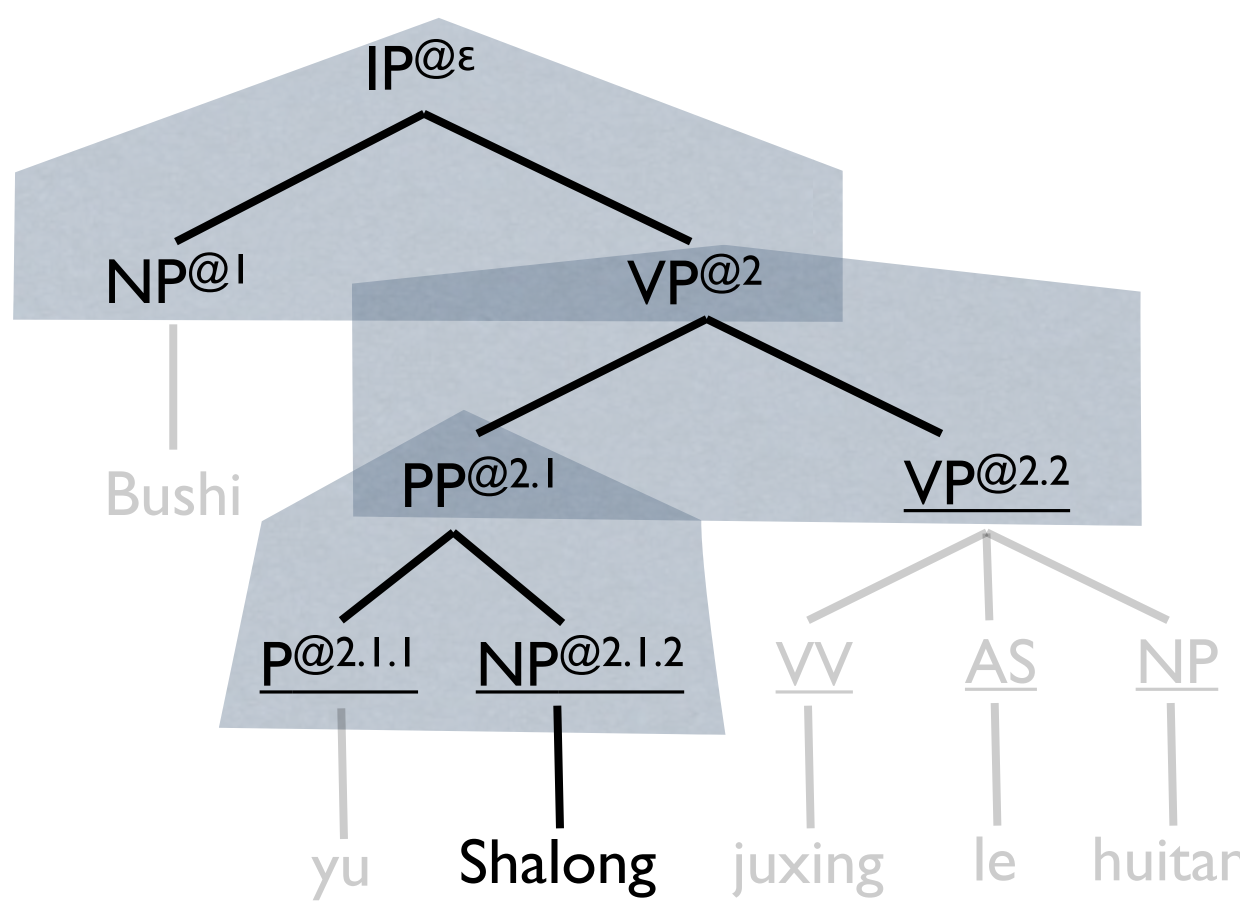
(Huang and Mi, 2010)

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] *stack*

r_1 r_3 r_4

<s> Bush held talks and *hypothesis*



action: pop

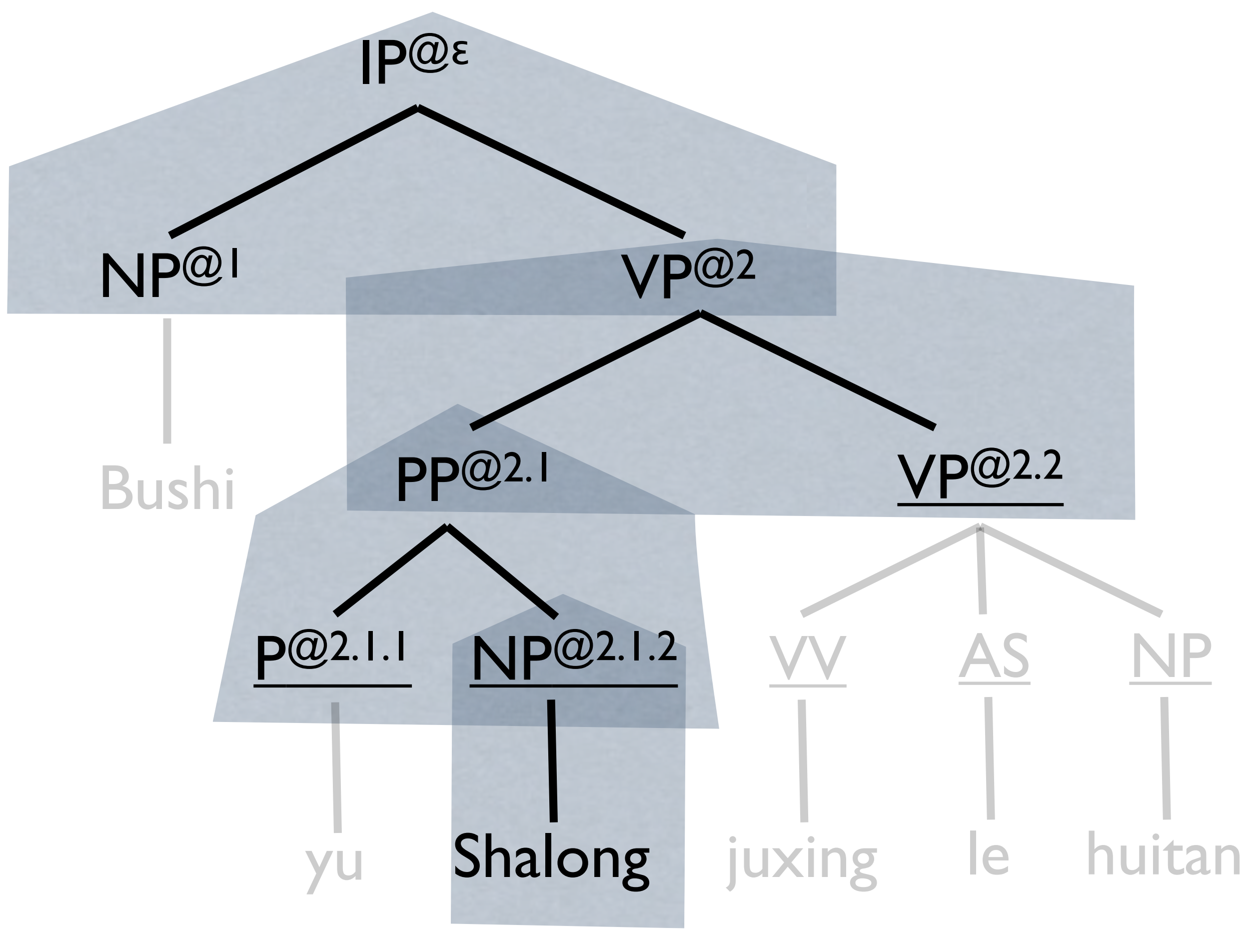
(Huang and Mi, 2010)

Example Incremental Decoding

[<s> . $IP^{@\varepsilon}$ </s>] [$NP^{@1}$. $VP^{@2}$] [$\underline{VP^{@2.2}}$. $\underline{PP^{@2.1}}$] [$\underline{P^{@2.1.1}}$. $\underline{NP^{@2.1.2}}$] [. Sharon] *stack*

r_1 r_3 r_4 r_7

<s> Bush held talks and *hypothesis*



action: predict

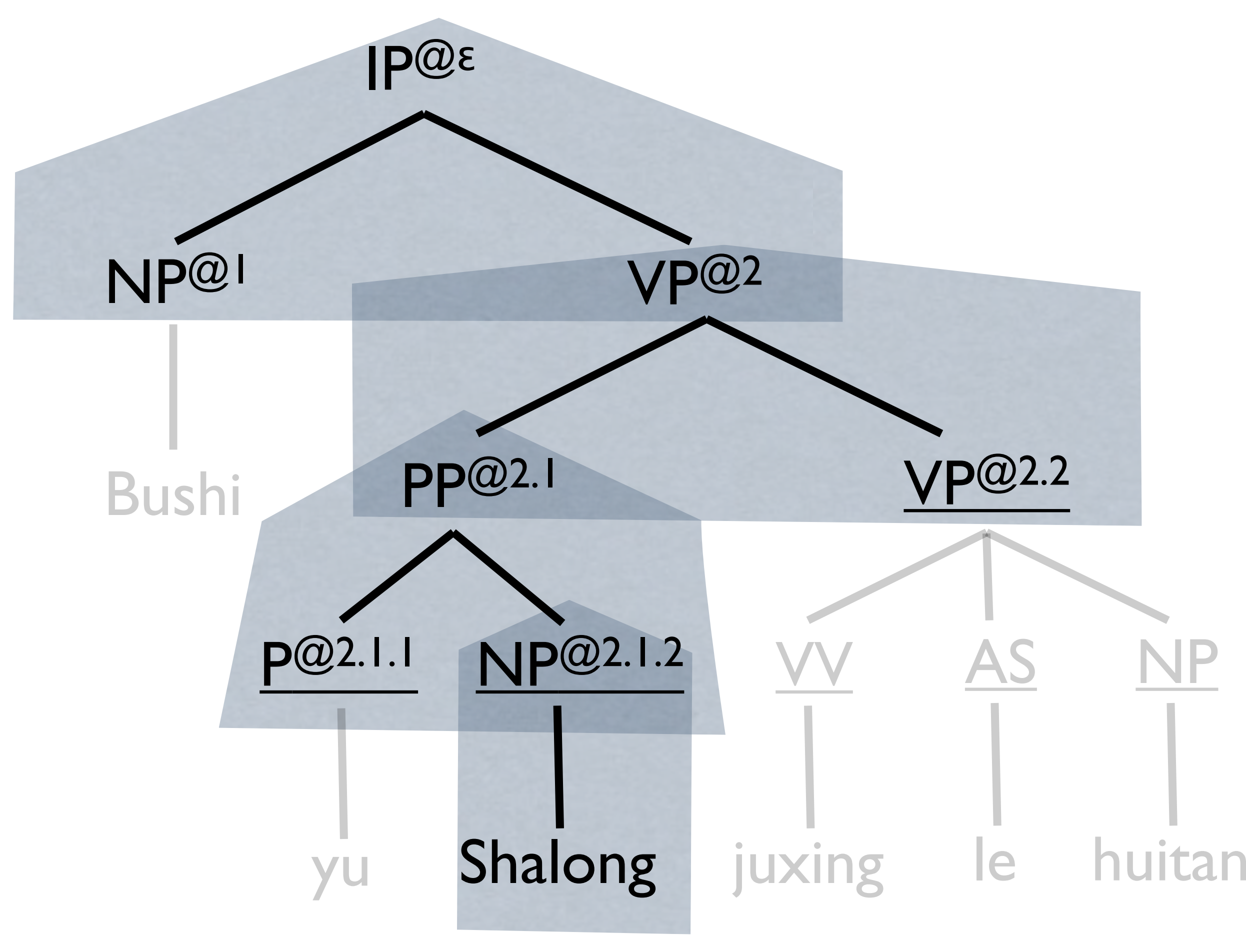
(Huang and Mi, 2010)

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 .] *stack*

r1 r3 r4 r7

<s> Bush held talks and Sharon *hypothesis*



action: scan

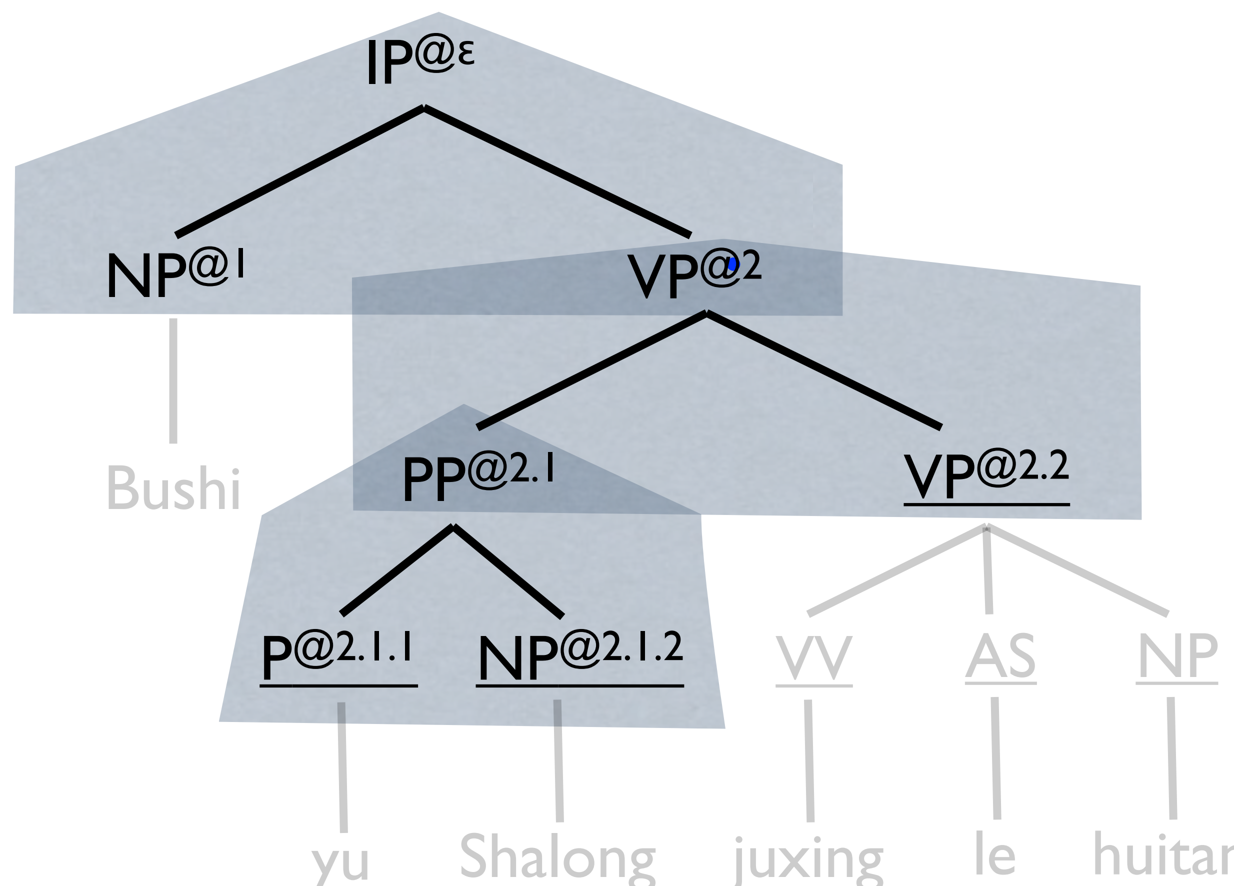
(Huang and Mi, 2010)

Example Incremental Decoding

[<s> . IP@$^{\varepsilon}$ </s>] [NP@^{1} . VP@^{2}] [<u>VP@$^{2.2}$</u> . <u>PP@$^{2.1}$</u>] [<u>P@$^{2.1.1}$</u> <u>NP@$^{2.1.2}$</u> .] *stack*

r$_1$ r$_3$ r$_4$

<s> Bush held talks and Sharon *hypothesis*



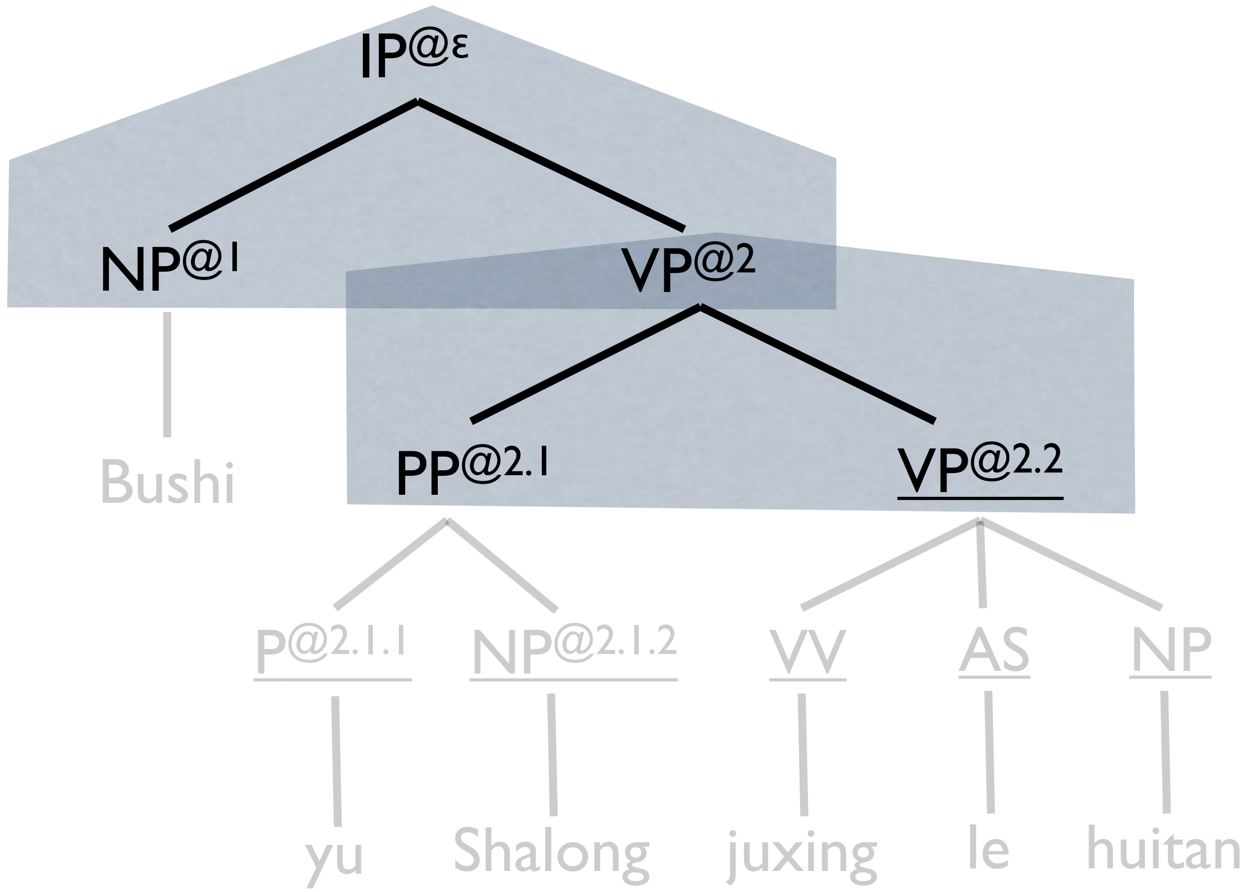
action: pop

(Huang and Mi, 2010)

Example Incremental Decoding

[<s> . $IP^{@\epsilon}$ </s>] [$NP^{@1}$. $VP^{@2}$]$_{r_1}$ [$\underline{VP^{@2.2}}$ $\underline{PP^{@2.1}}$.]$_{r_3}$ *stack*

<s> Bush held talks and Sharon *hypothesis*



action: pop

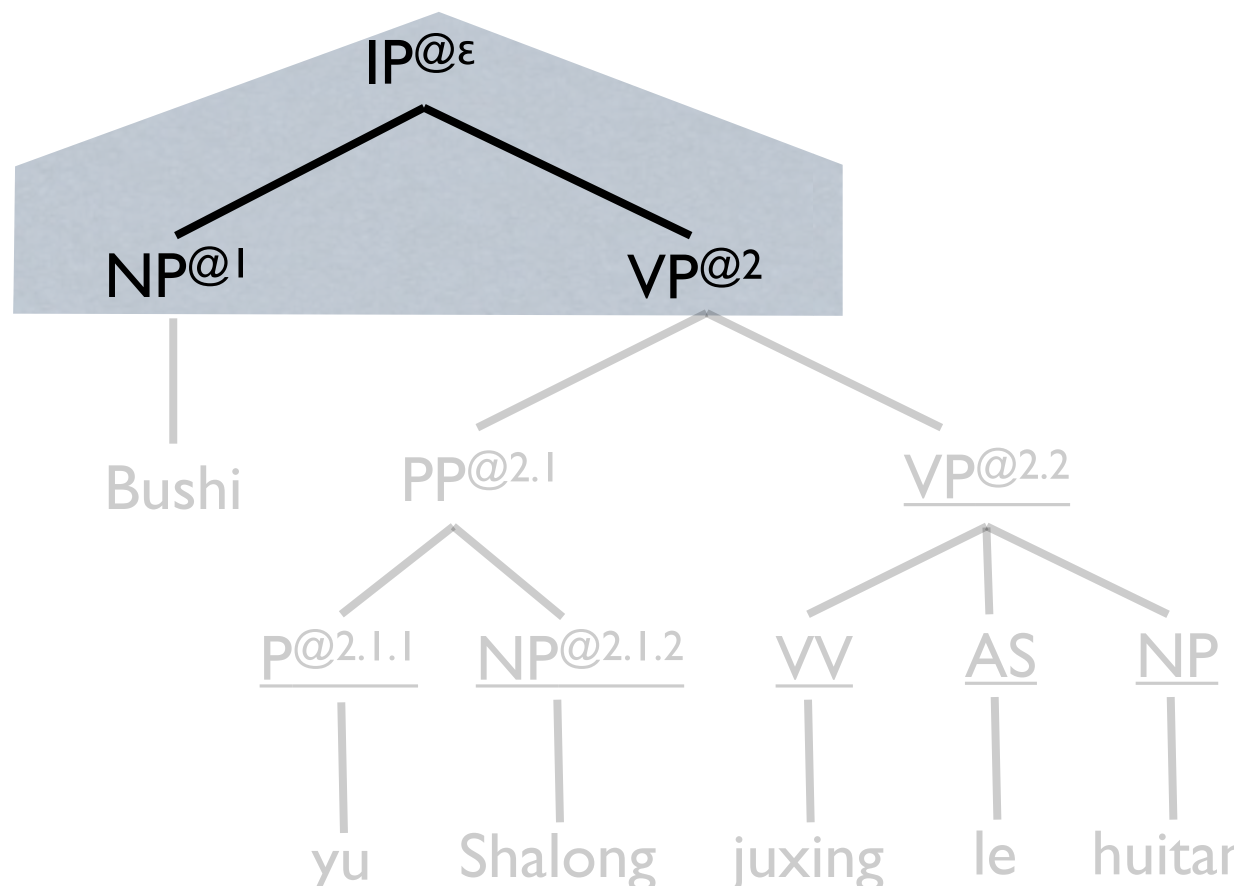
(Huang and Mi, 2010)

Example Incremental Decoding

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

r_1

<s> Bush held talks and Sharon *hypothesis*



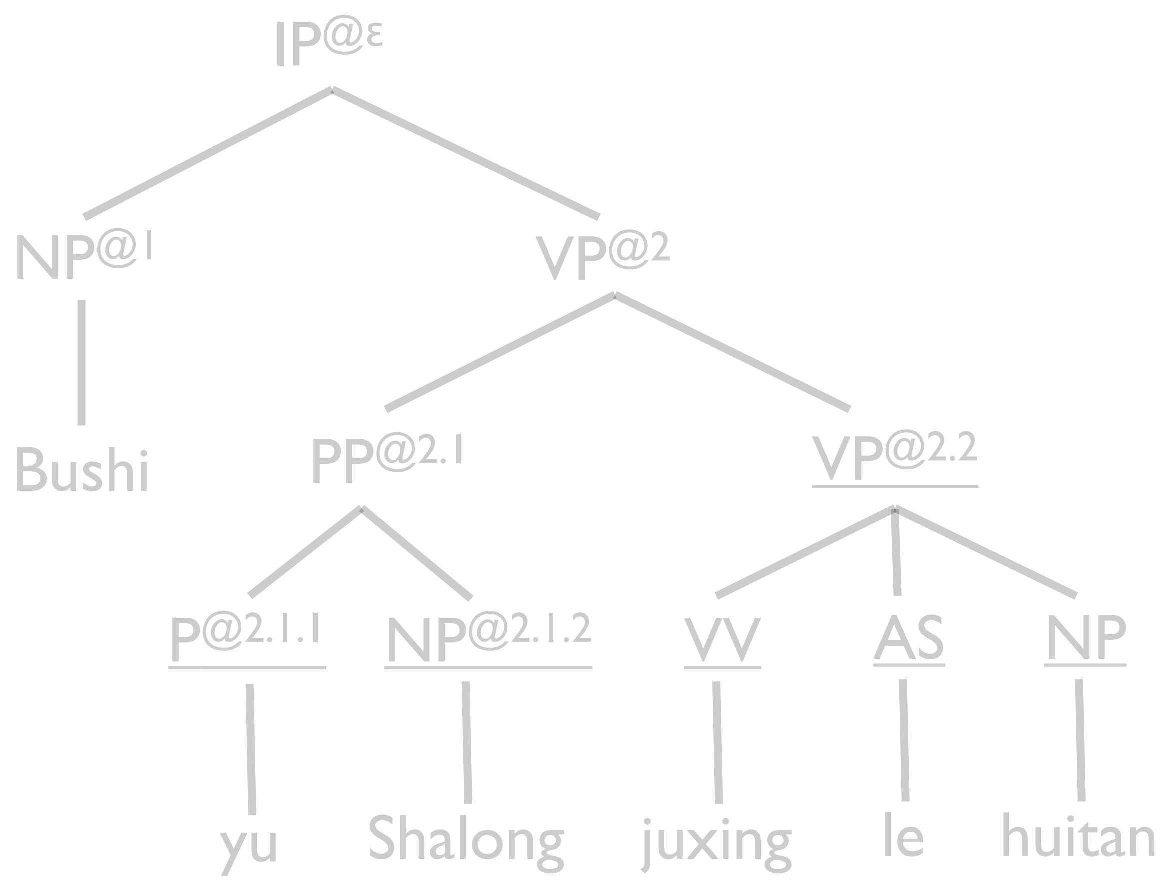
action: pop

(Huang and Mi, 2010)

Example Incremental Decoding

[<s> IP@ε •</s>] *stack*

<s> Bush held talks and Sharon *hypothesis*



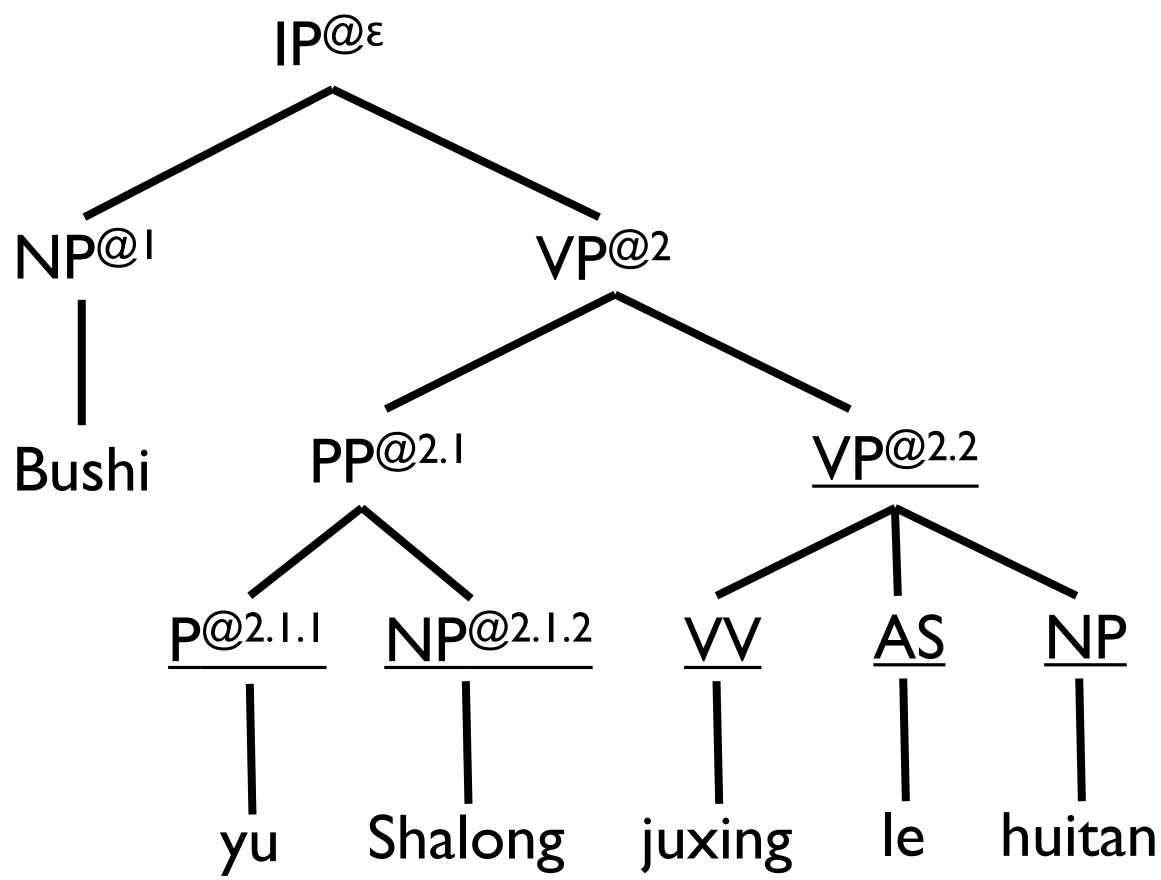
action: pop

(Huang and Mi, 2010)

Example Incremental Decoding

[<s> . $IP^{@\varepsilon}$ </s>] *stack*

<s> *hypothesis*



this work: adding rule LM

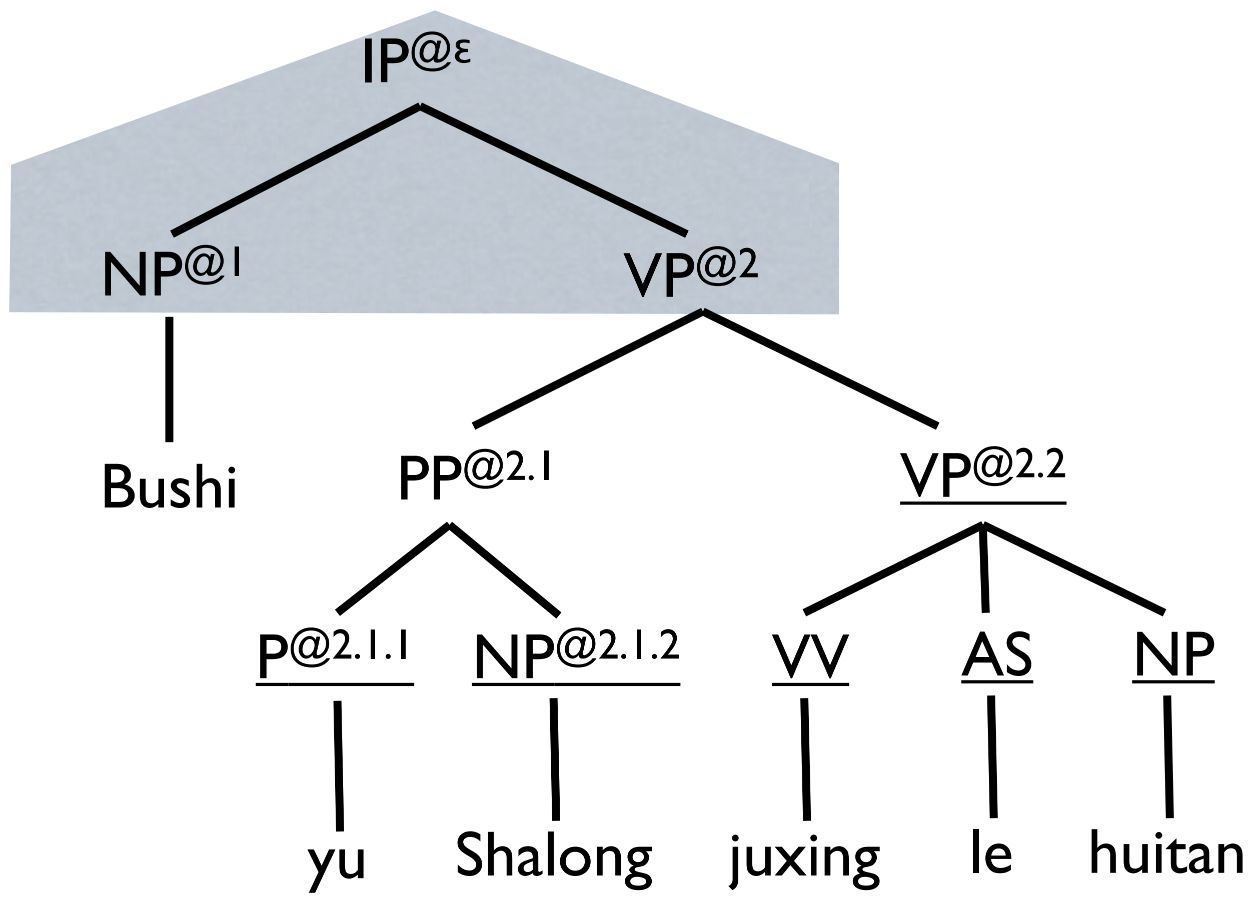
Example Incremental Decoding

[<s> . IP@$^{\epsilon}$ </s>] [. NP@^{1} VP@^{2}]
r_1

stack

<s>

hypothesis



rule probability

$P(r_1|\epsilon)$

action: predict (push)

this work: adding rule LM

Example Incremental Decoding

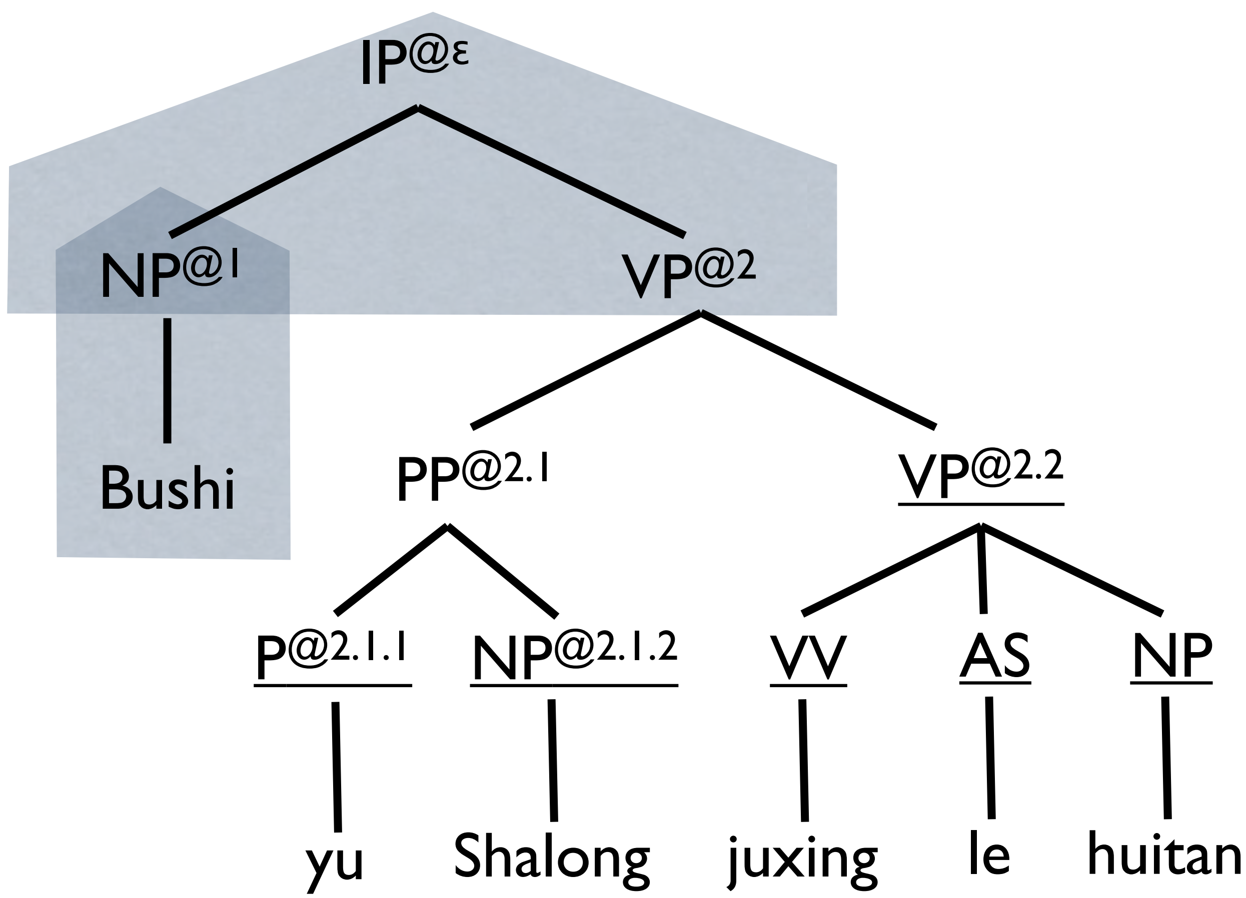
[<s> . $IP^{@\varepsilon}$ </s>] [. $NP^{@1}$ $VP^{@2}$] [. Bush]

r_1 r_2

stack

<s>

hypothesis



rule probability

$P(r_2|r_1)$

action: predict

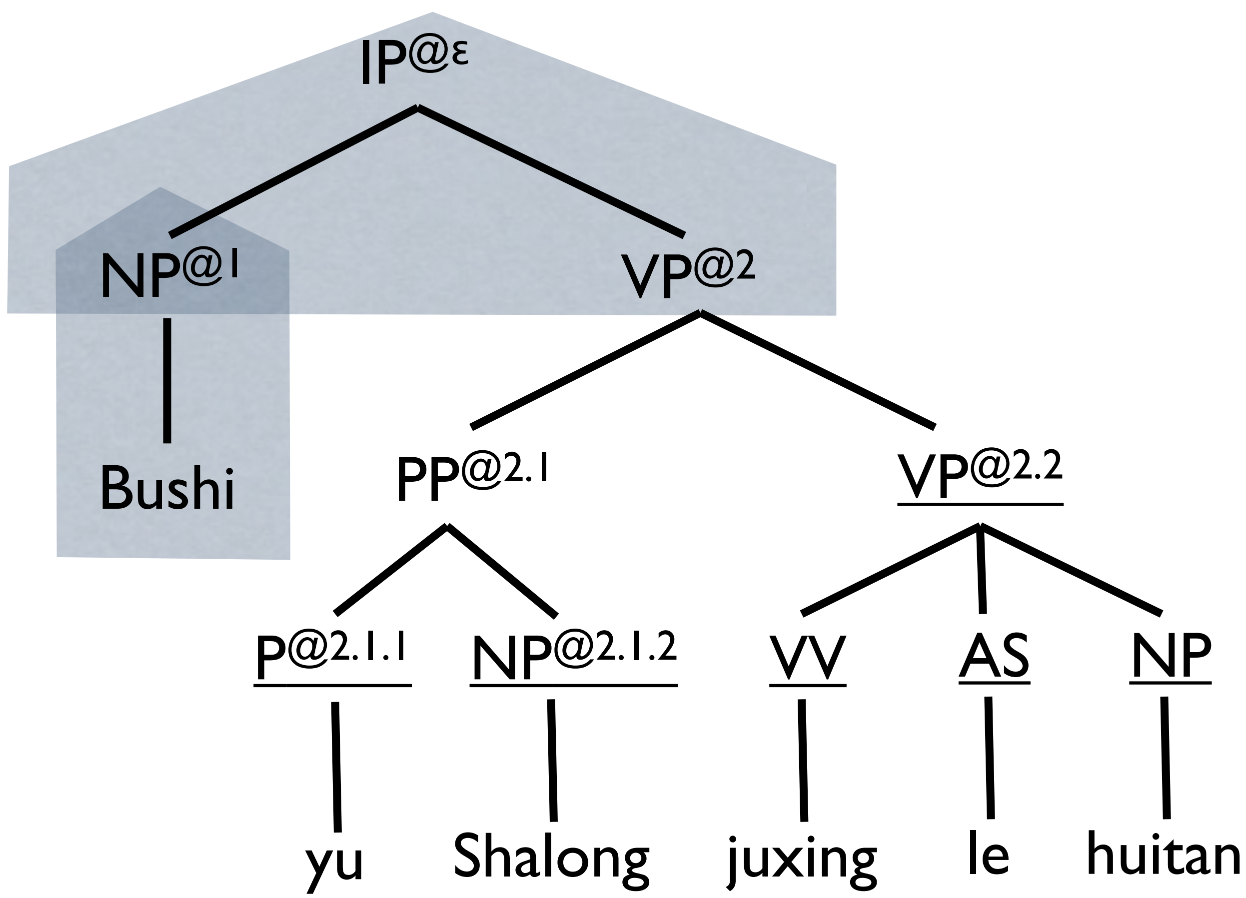
this work: adding rule LM

Example Incremental Decoding

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

r_1 r_2

<s> Bush *hypothesis*



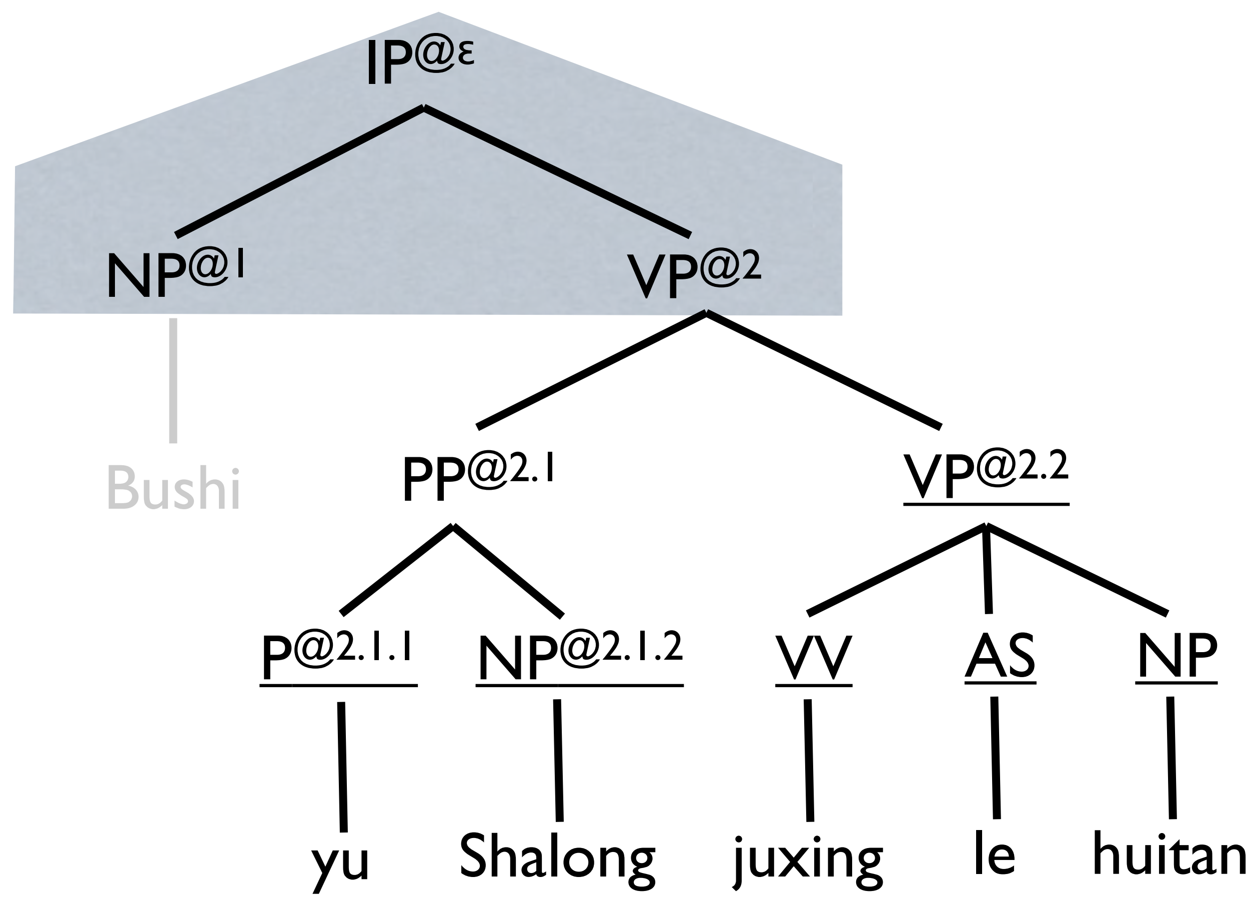
action: scan

this work: adding rule LM

Example Incremental Decoding

[<s> . IP@ε </s>] [NP@1 $_{r_1}$. VP@2] *stack*

<s> Bush *hypothesis*



action: pop

this work: adding rule LM

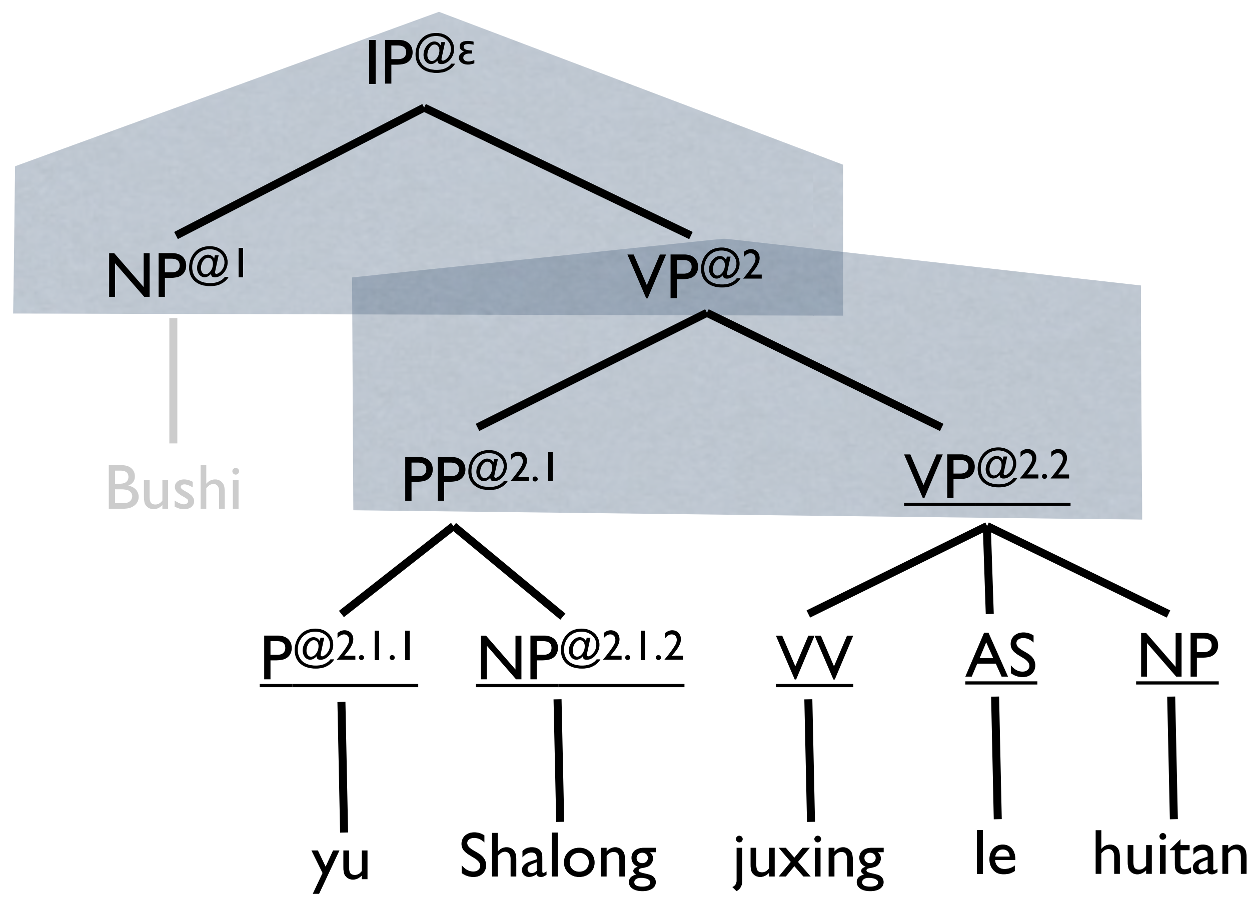
Example Incremental Decoding

[<s> . IP@ε </s>] [NP@1 . VP@2] [. VP@2.2 PP@2.1]
r_1 r_3

stack

<s> Bush

hypothesis



rule probability

$P(r_3|r_1)$

action: predict

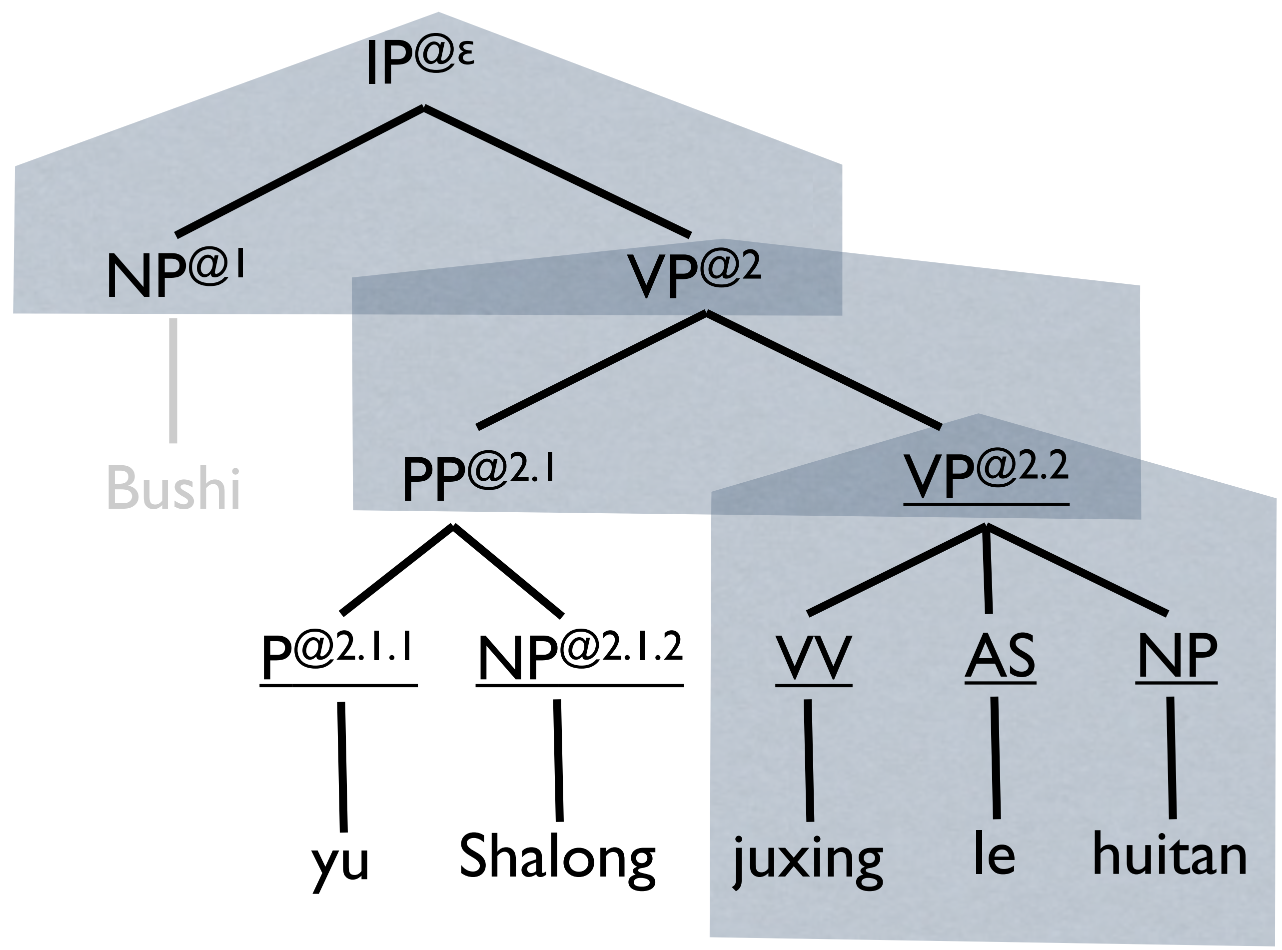
this work: adding rule LM

Example Incremental Decoding

[<s> . $IP^{@\varepsilon}$ </s>] [$NP^{@1}$. $VP^{@2}$] [. $\underline{VP^{@2.2}}$ $\underline{PP^{@2.1}}$] [. held talks] *stack*

r_1 r_3 r_5

<s> Bush *hypothesis*



rule probability

$P(r_5|r_1,r_3)$

action: predict

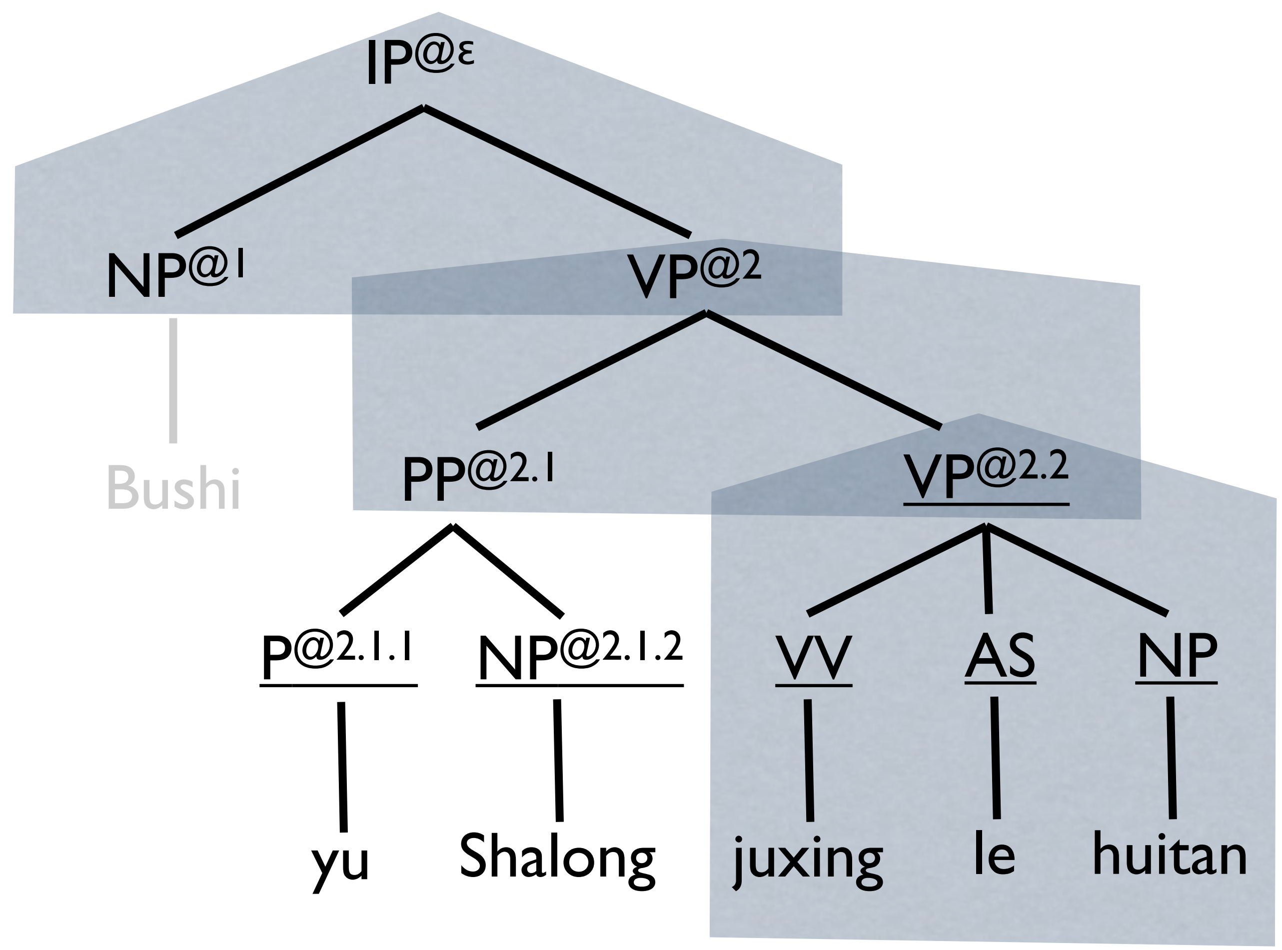
this work: adding rule LM

Example Incremental Decoding

[<s> . IP$^{@\varepsilon}$ </s>] [NP$^{@1}$. VP$^{@2}$] [. $\underline{\text{VP}^{@2.2}}$ $\underline{\text{PP}^{@2.1}}$] [held talks .] *stack*

r_1 r_3 r_5

<s> Bush held talks *hypothesis*



action: scan

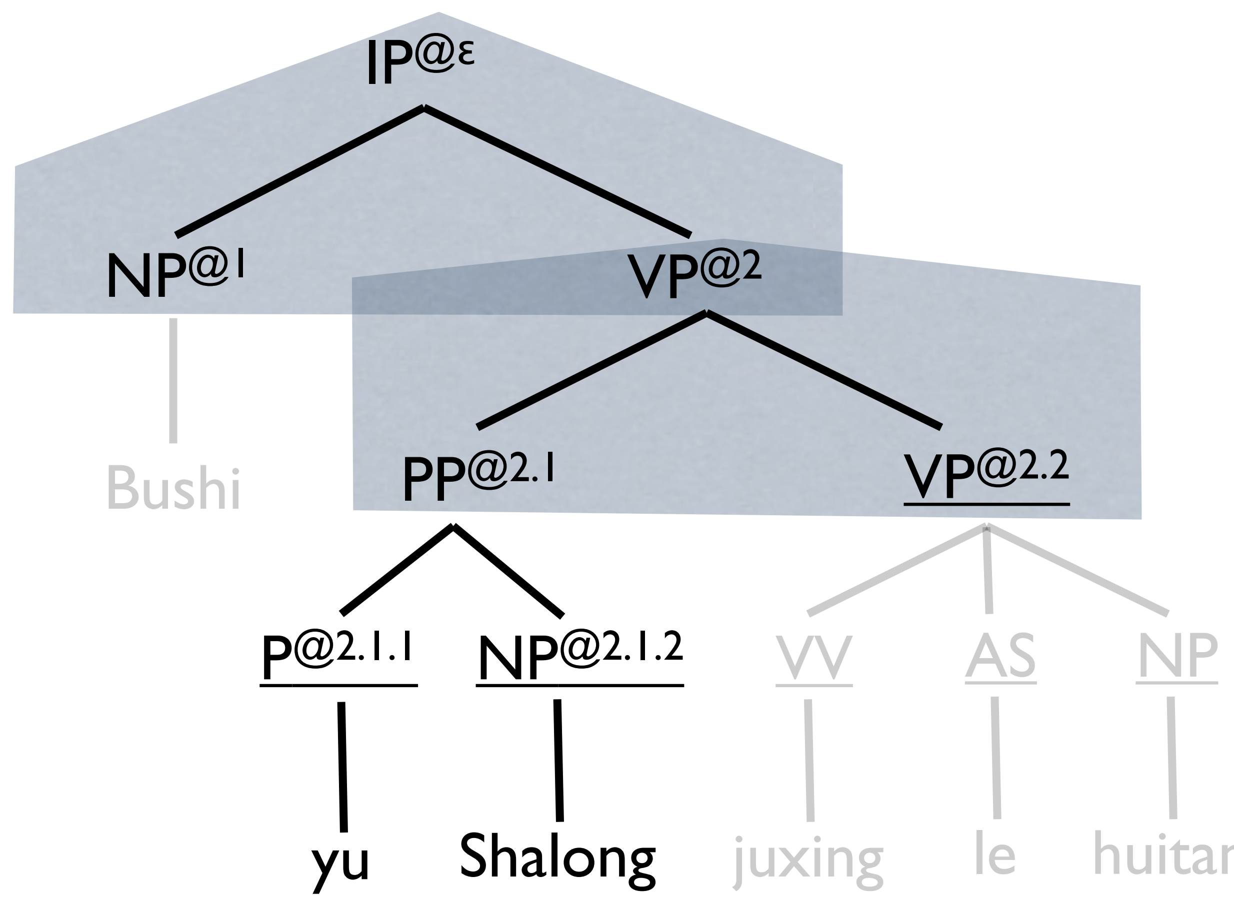
this work: adding rule LM

Example Incremental Decoding

[<s> . IP$^{@\varepsilon}$ </s>] [NP$^{@1}$. VP$^{@2}$] [. VP$^{@2.2}$ PP$^{@2.1}$] *stack*

r_1 r_3

<s> Bush held talks *hypothesis*



action: pop

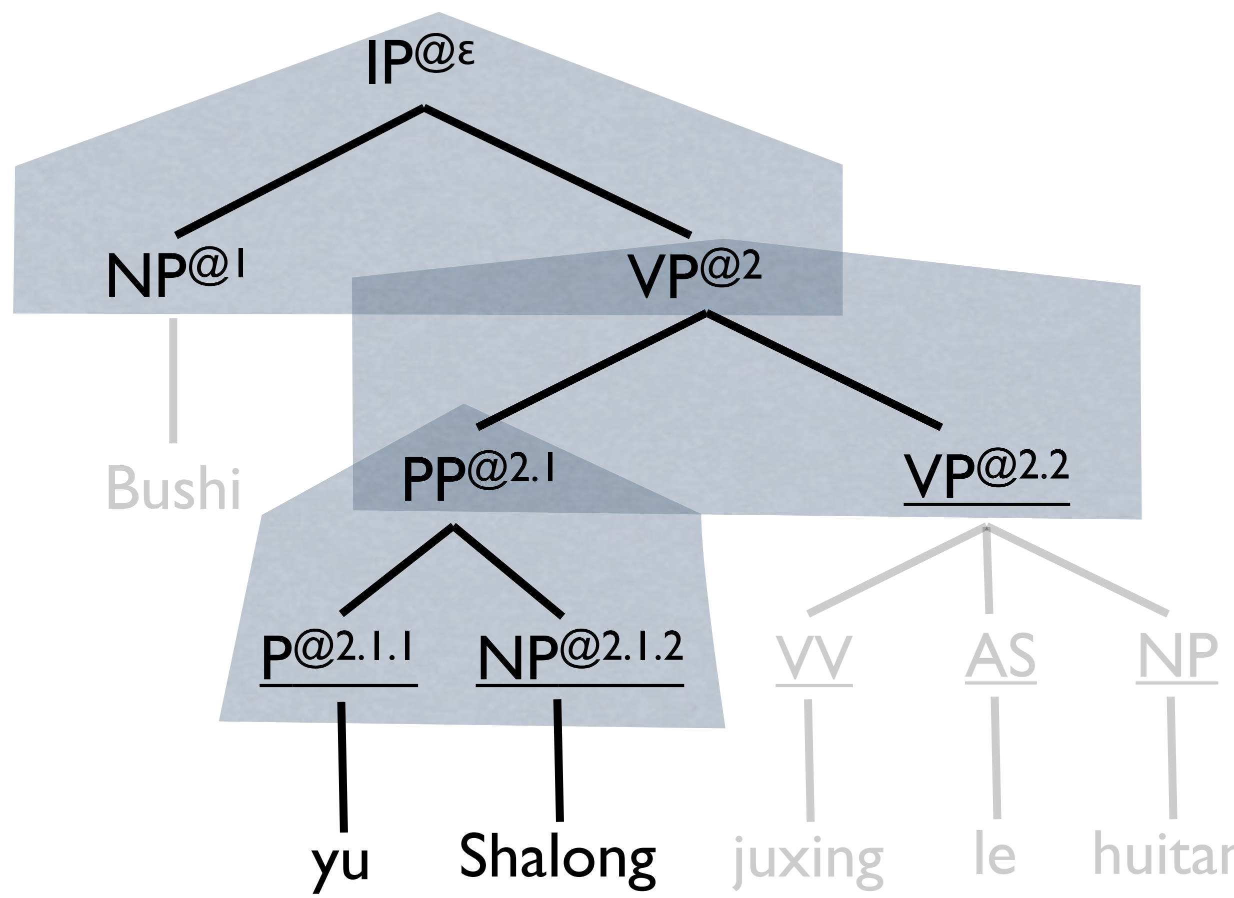
this work: adding rule LM

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] *stack*

r_1 r_3 r_4

<s> Bush held talks *hypothesis*



rule probability

$P(r_4|r_1,r_3)$

action: predict

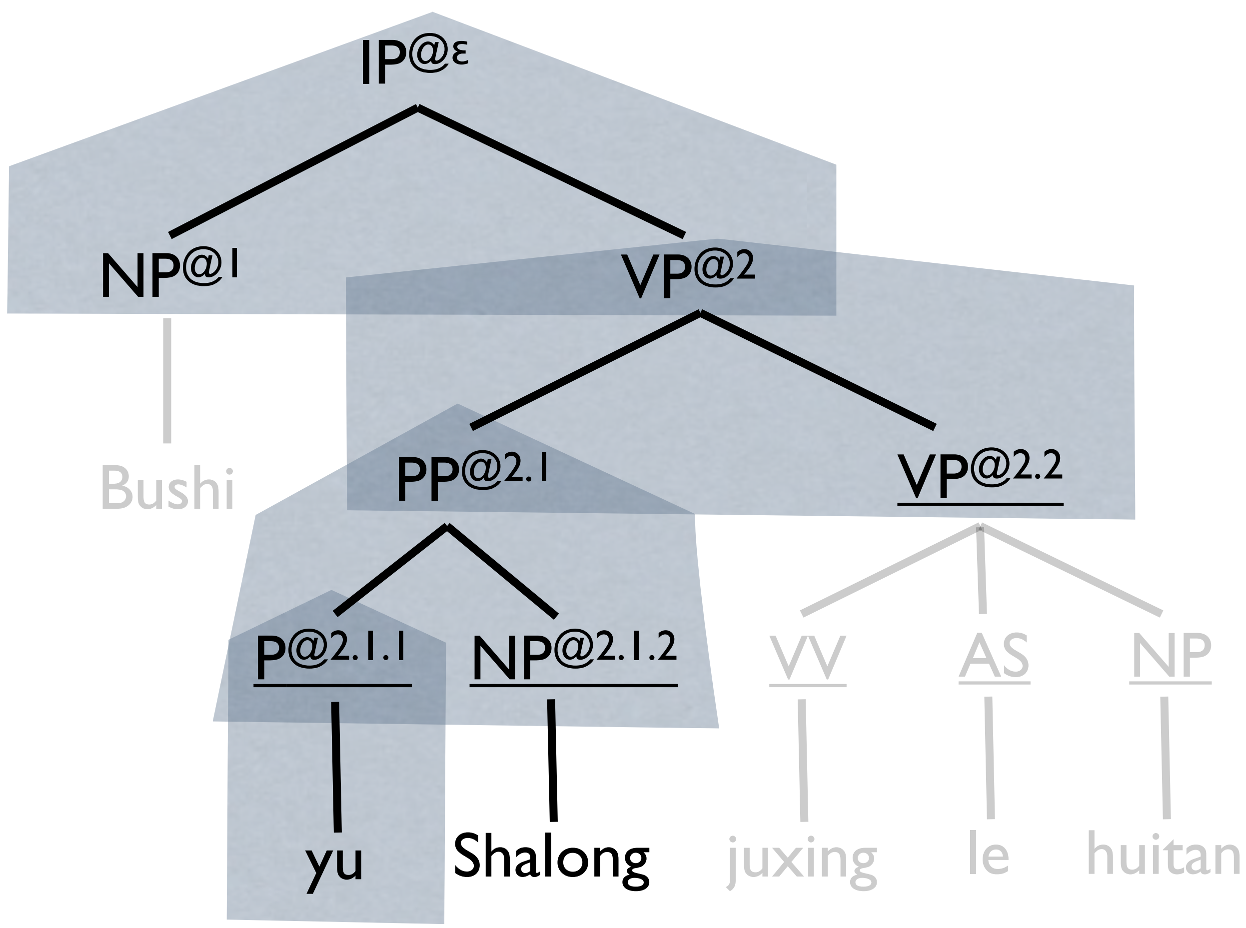
this work: adding rule LM

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] [. with/and] *stack*

r_1 r_3 r_4 r_6/r'_6

<s> Bush held talks *hypothesis*



rule probability

$P(r_6/r'_6 \mid r_3, r_4)$

action: predict

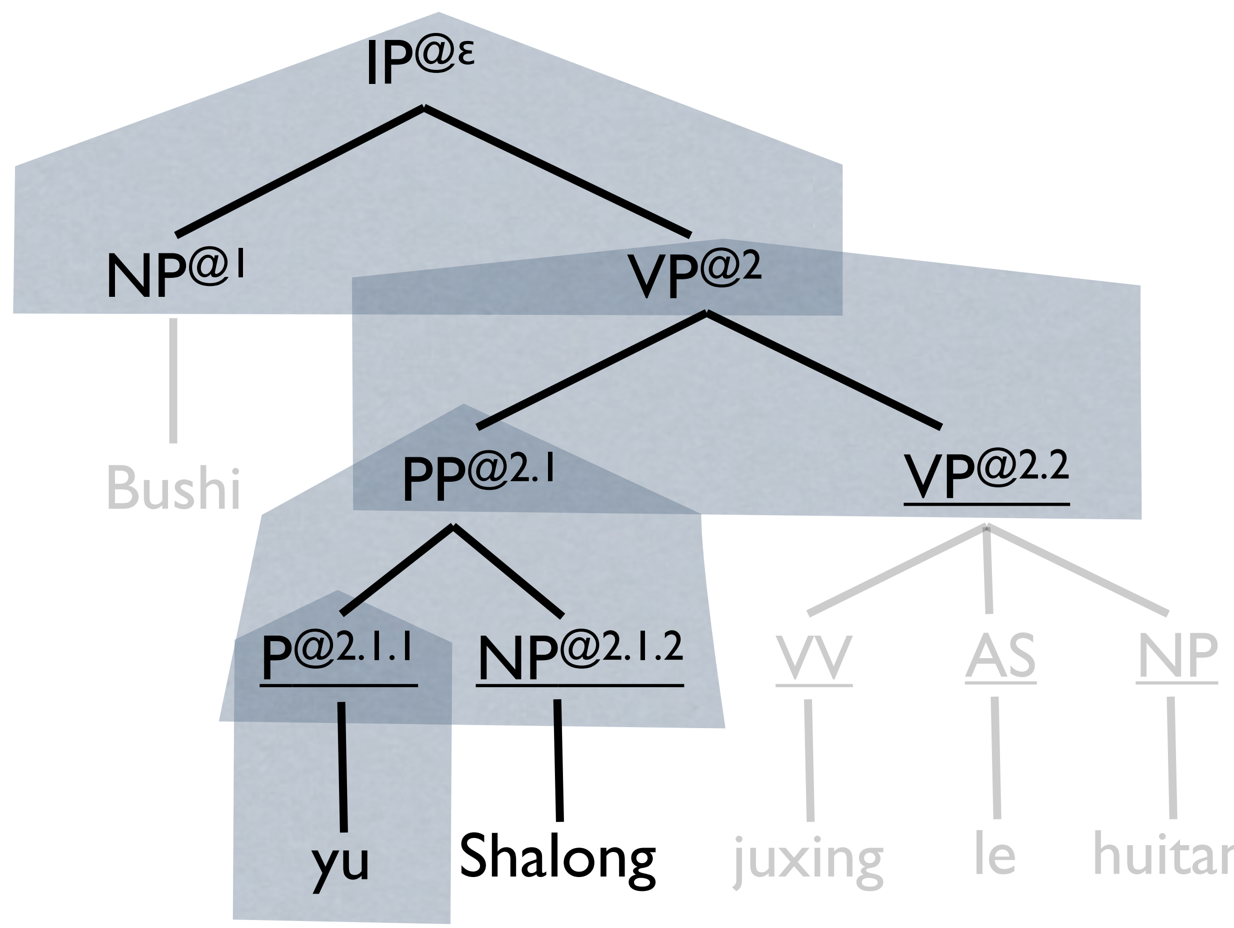
this work: adding rule LM

Example Incremental Decoding

[<s> . $IP^{@\varepsilon}$ </s>] [$NP^{@1}$. $VP^{@2}$] [$\underline{VP^{@2.2}}$. $\underline{PP^{@2.1}}$] [. $\underline{P^{@2.1.1}}$ $\underline{NP^{@2.1.2}}$] [with .] *stack*

r_1 r_3 r_4 r_6

<s> Bush held talks with *hypothesis*



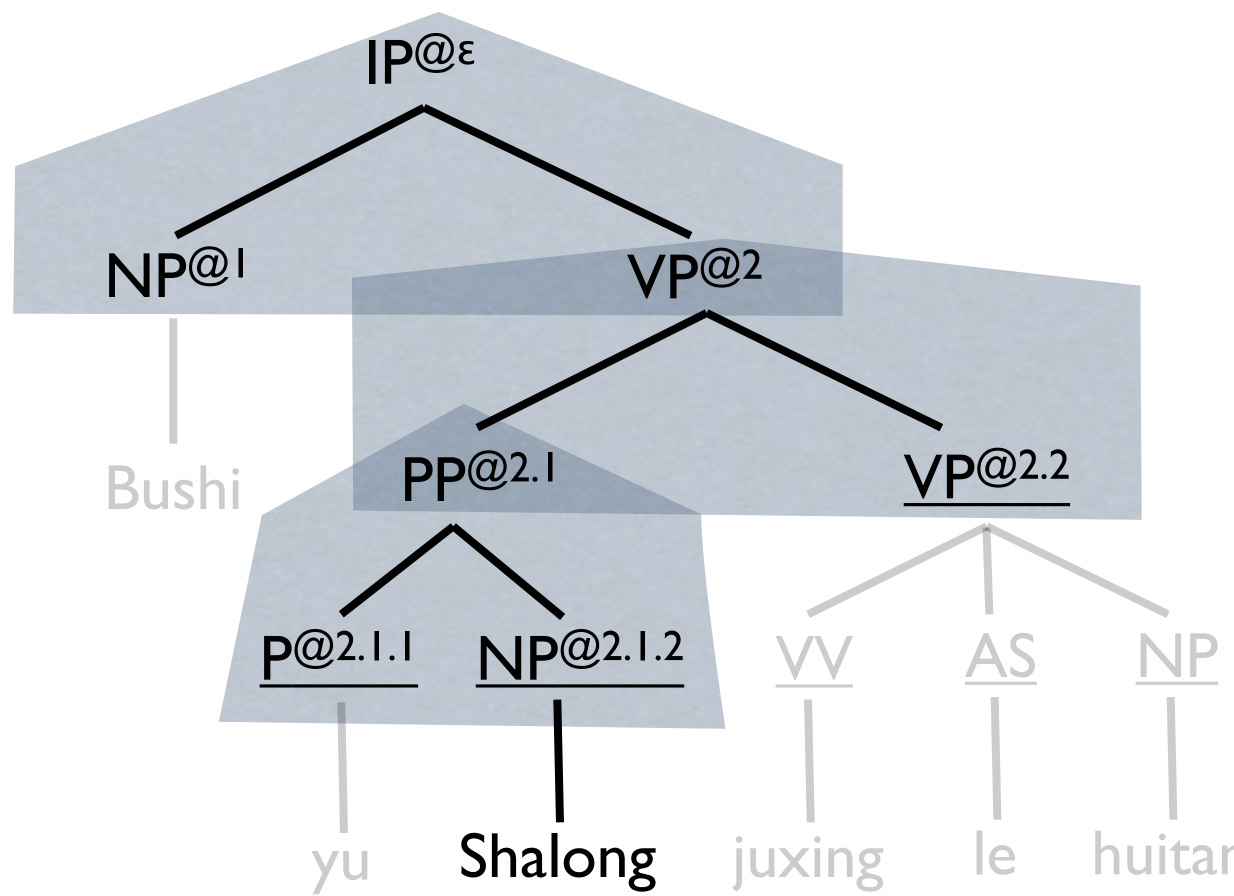
action: scan

this work: adding rule LM

Example Incremental Decoding

$[\langle s\rangle \bullet \text{IP}^{@\varepsilon}\ \langle/s\rangle]\ [\text{NP}^{@1} \underset{r_1}{\bullet} \text{VP}^{@2}]\ [\underline{\text{VP}^{@2.2}} \underset{r_3}{\bullet} \underline{\text{PP}^{@2.1}}]\ [\underline{\text{P}^{@2.1.1}} \underset{r_4}{\bullet} \underline{\text{NP}^{@2.1.2}}]$ *stack*

<s> Bush held talks with *hypothesis*



action: pop

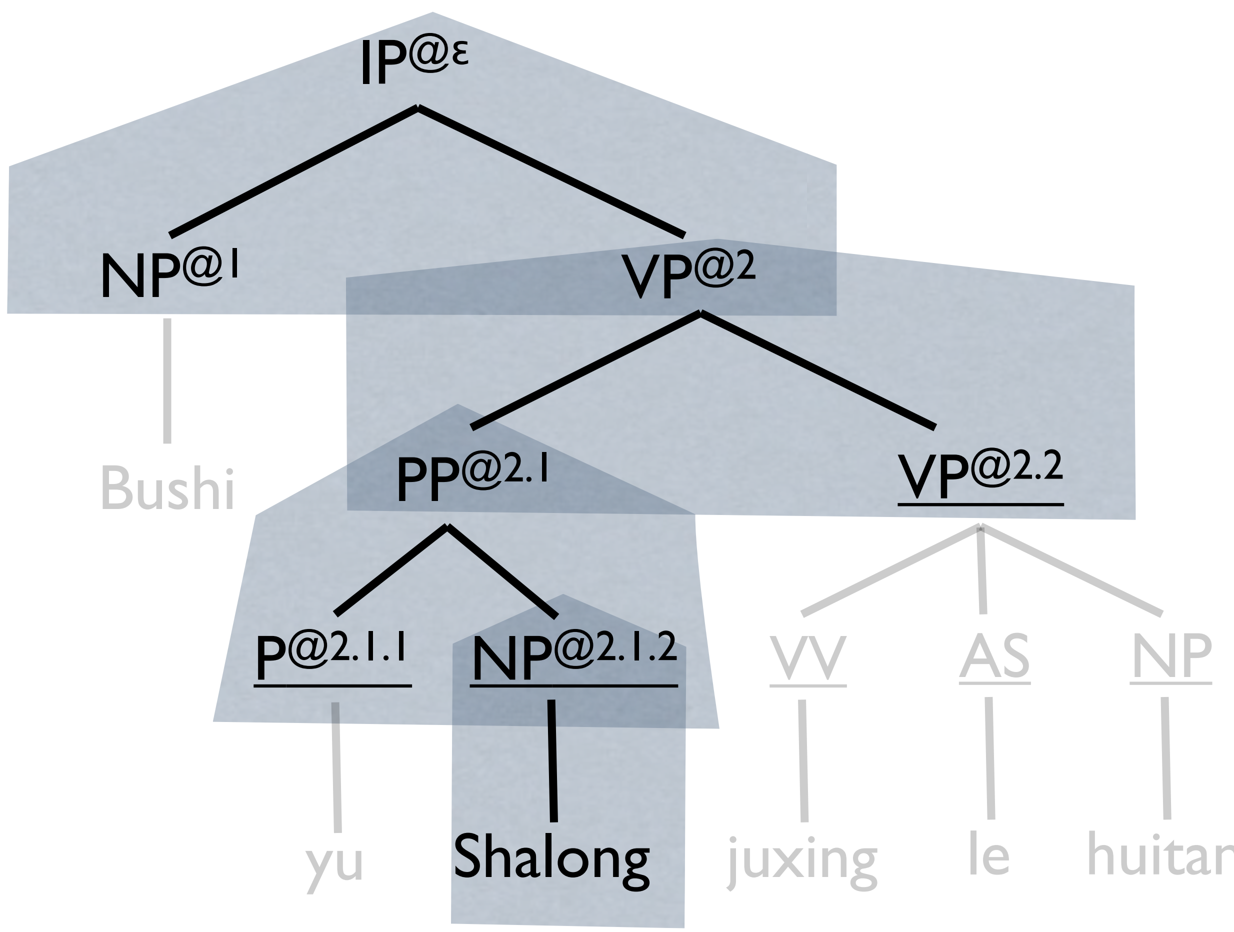
this work: adding rule LM

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] [. Sharon] *stack*

r_1 r_3 r_4 r_7

<s> Bush held talks with *hypothesis*



rule probability

$P(r_7 \mid r_3, r_4)$

action: predict

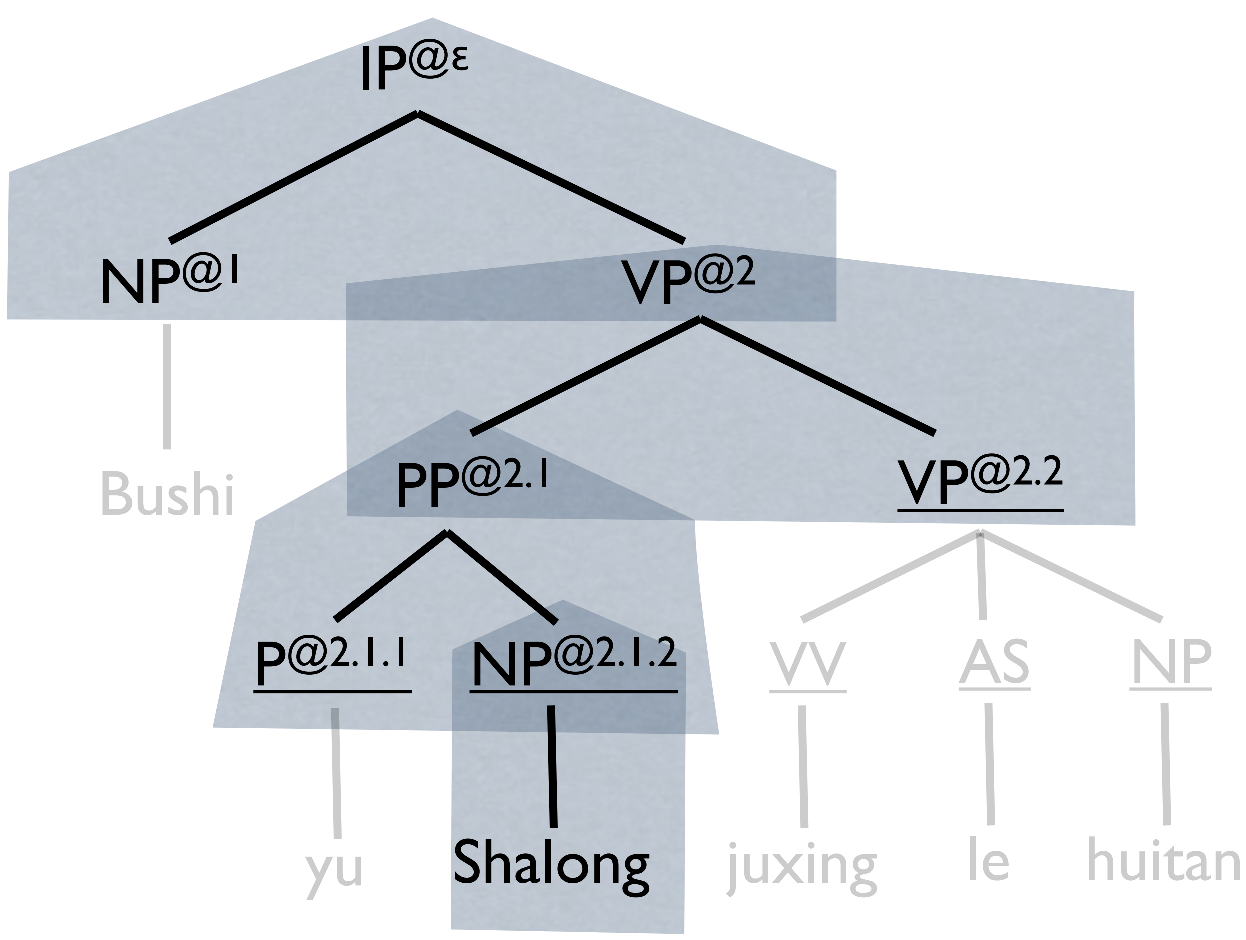
this work: adding rule LM

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] [Sharon .] *stack*

r1 r3 r4 r7

<s> Bush held talks with Sharon *hypothesis*



action: scan

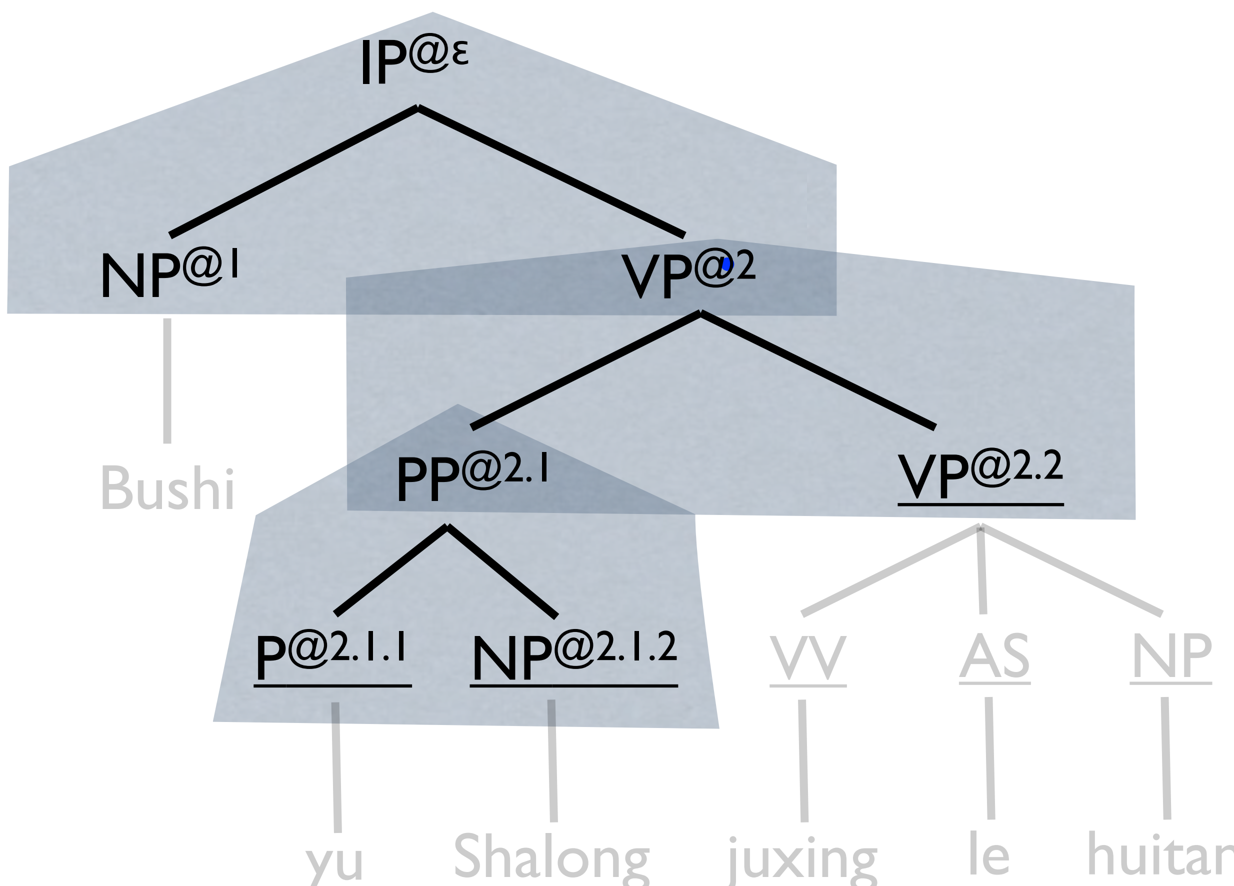
this work: adding rule LM

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 .] *stack*

r1 r3 r4

<s> Bush held talks with Sharon *hypothesis*



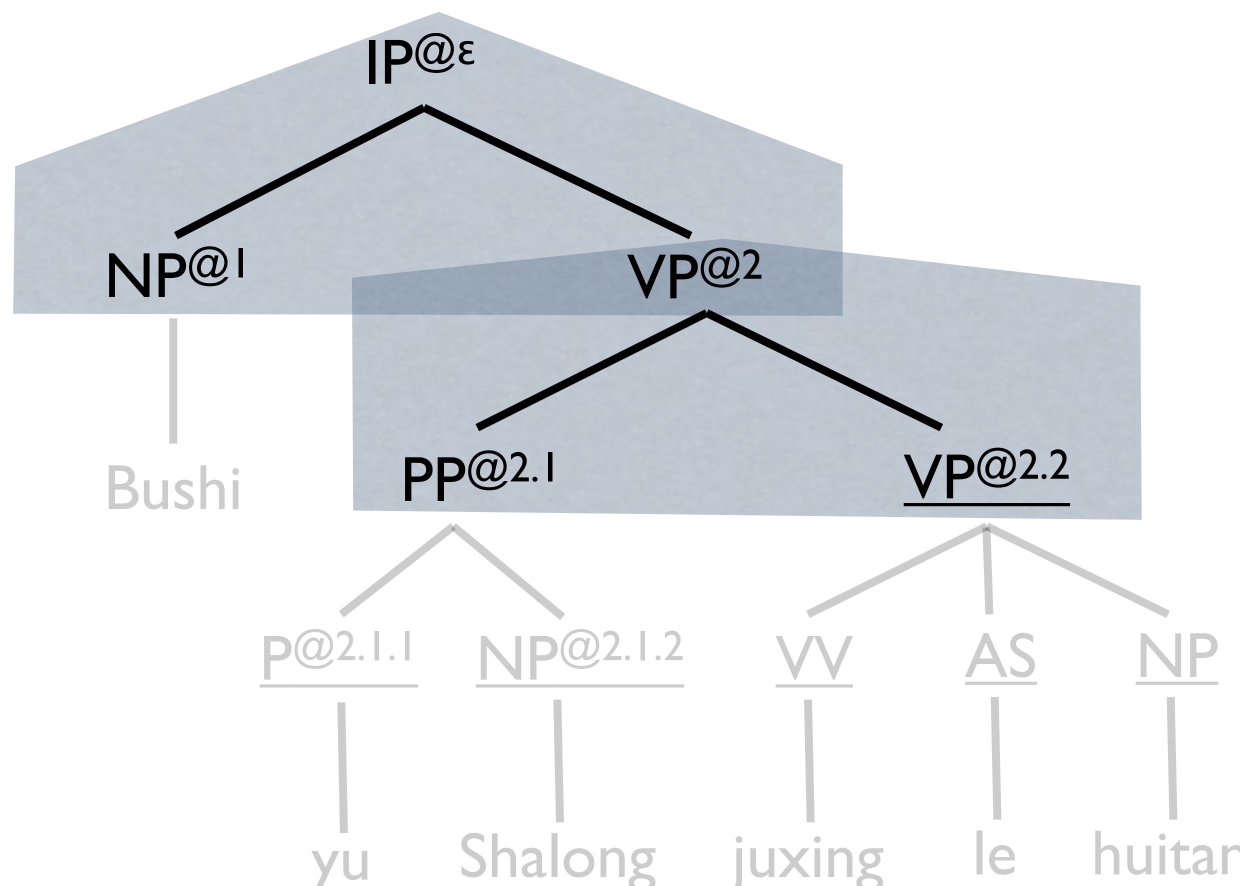
action: pop

this work: adding rule LM

Example Incremental Decoding

$[\langle s\rangle \cdot \mathrm{IP}^{@\varepsilon}\ \langle /s\rangle]\ [\mathrm{NP}^{@1} \underset{r_1}{\cdot} \mathrm{VP}^{@2}]\ [\underline{\mathrm{VP}^{@2.2}}\ \underset{r_3}{\underline{\mathrm{PP}^{@2.1}}} \cdot]$ *stack*

<s> Bush held talks with Sharon *hypothesis*



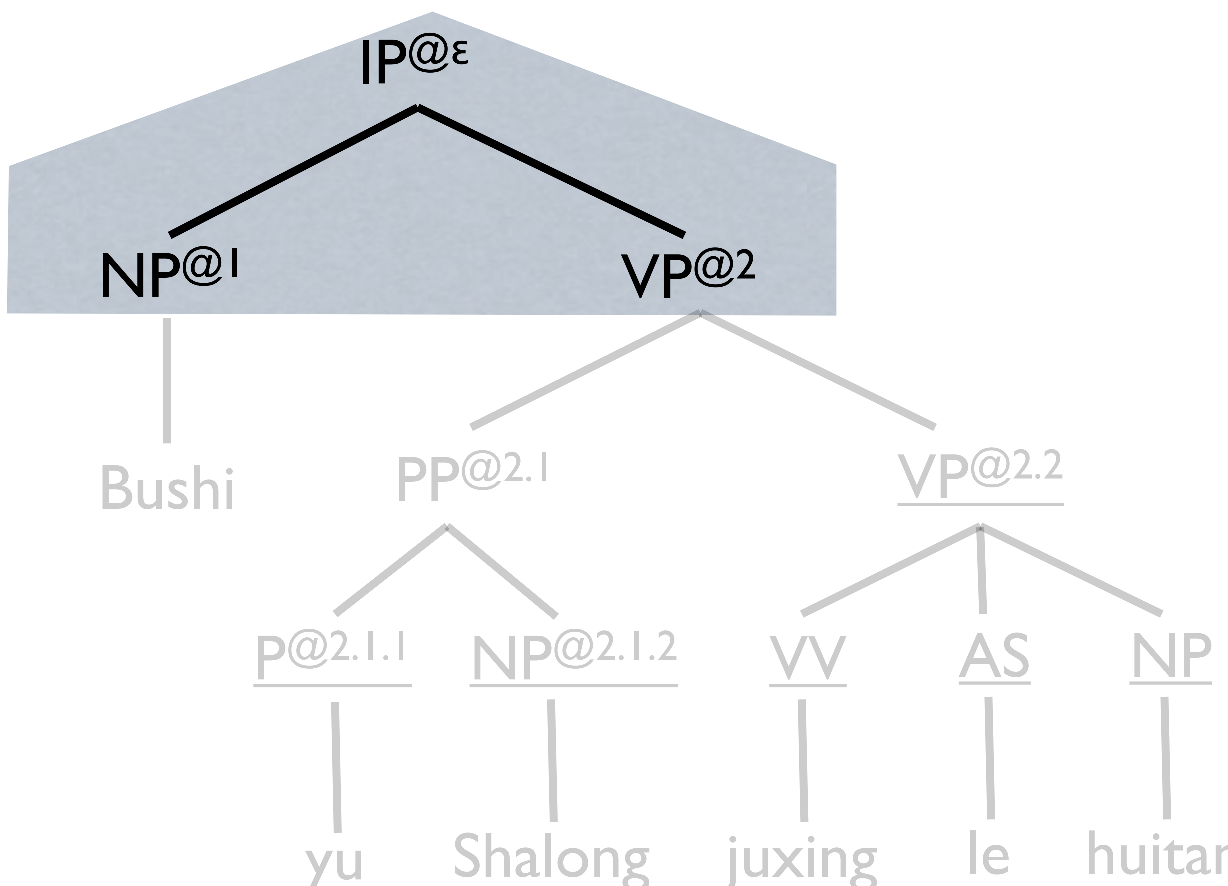
action: pop

this work: adding rule LM

Example Incremental Decoding

[<s> . IP@ε </s>] [NP@1 VP@2 .] *stack*
r_1

<s> Bush held talks with Sharon *hypothesis*



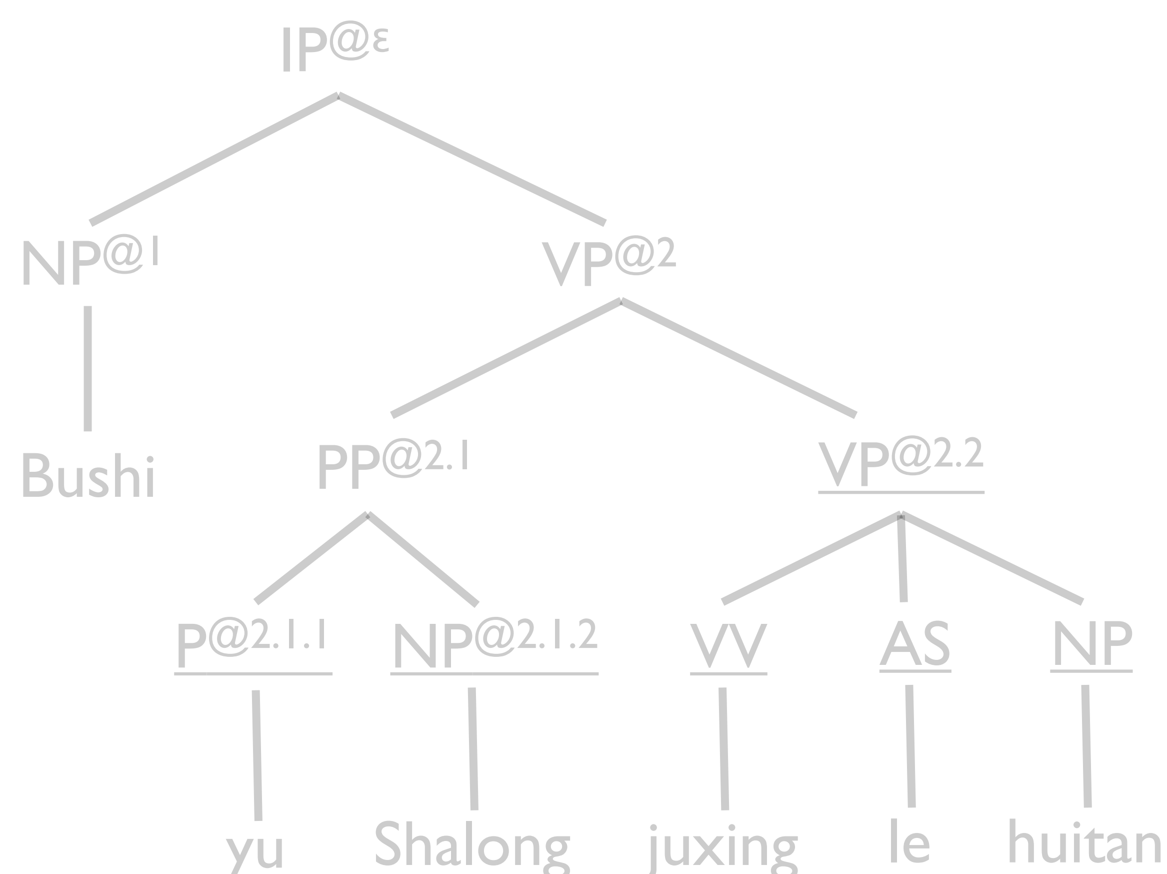
action: pop

this work: adding rule LM

Example Incremental Decoding

[<s> IP@ε .</s>] *stack*

<s> Bush held talks with Sharon *hypothesis*



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

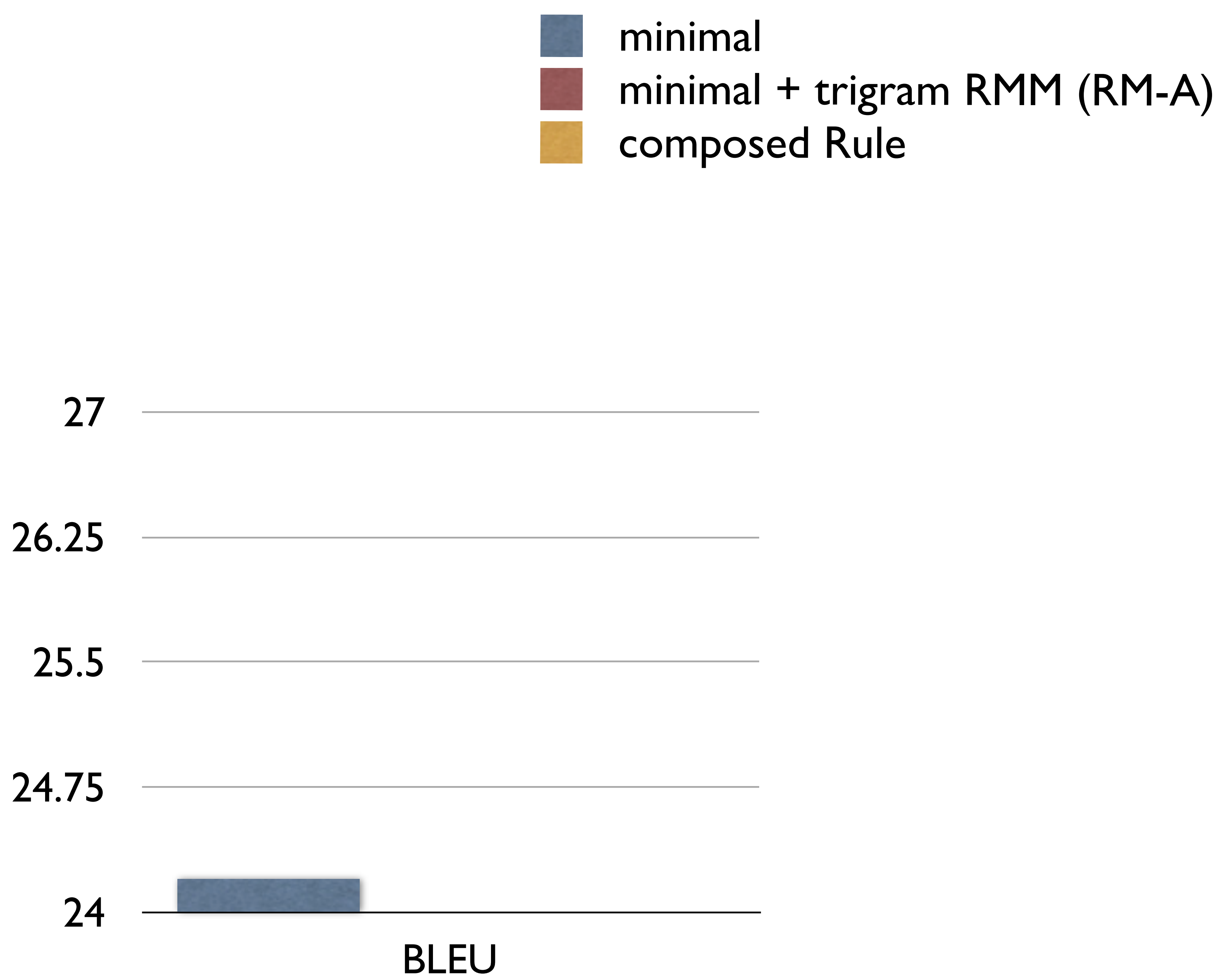
Results

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

Results

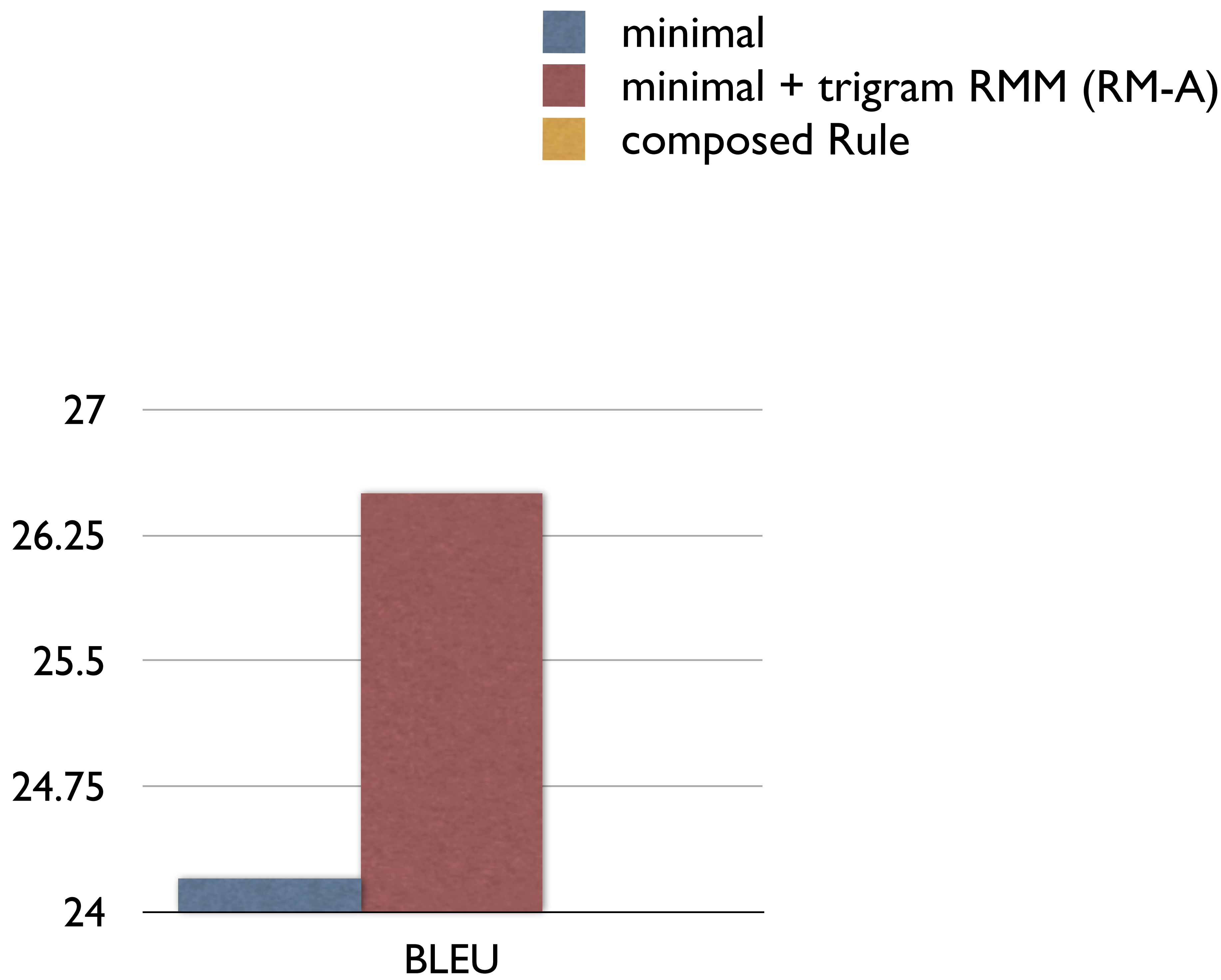

minimal
minimal + trigram RMM (RM-A)
composed Rule
27
26.25
25.5
24.75
24
BLEU

Results

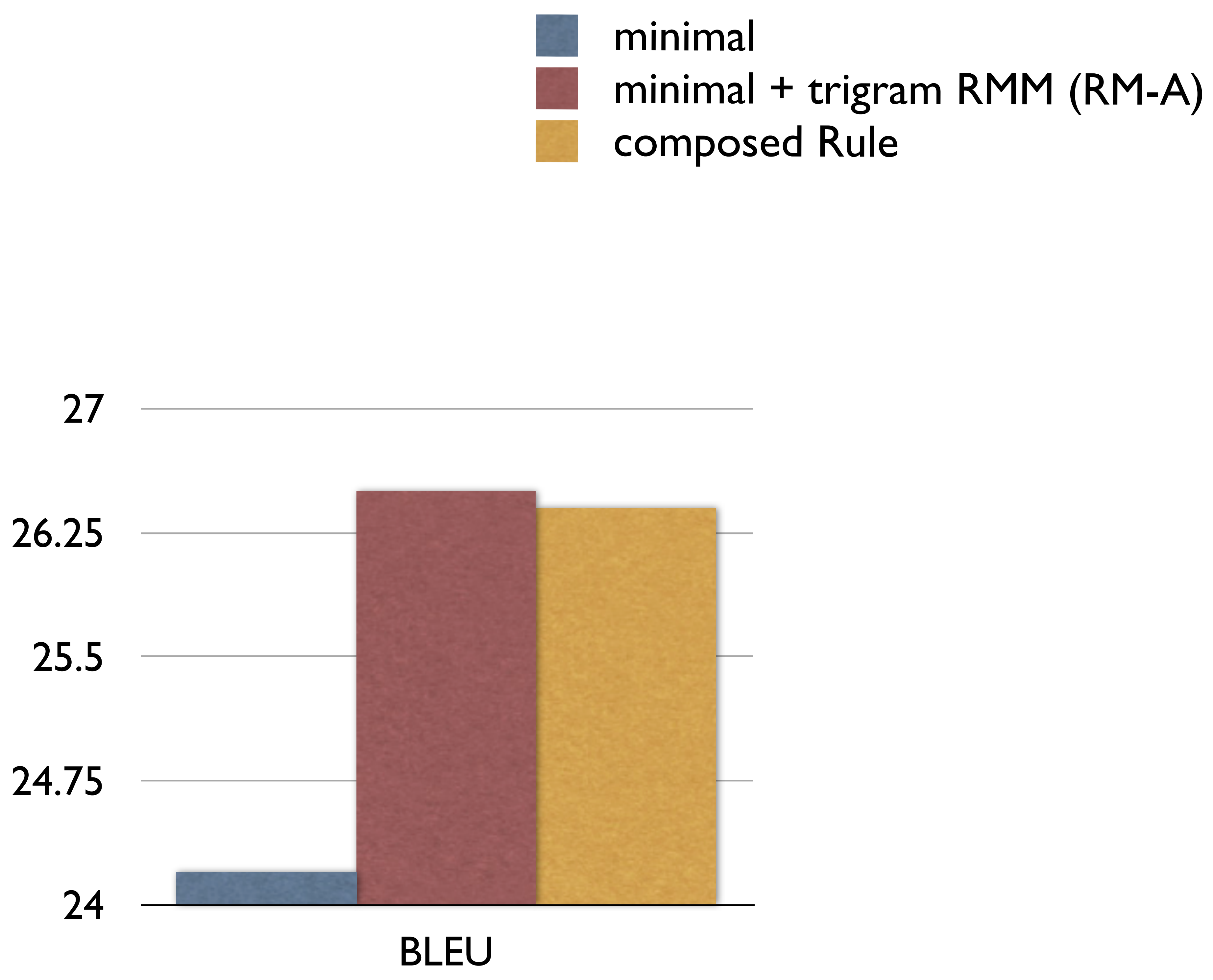


minimal
minimal + trigram RMM (RM-A)
composed Rule
27
26.25
25.5
24.75
24
BLEU

Results

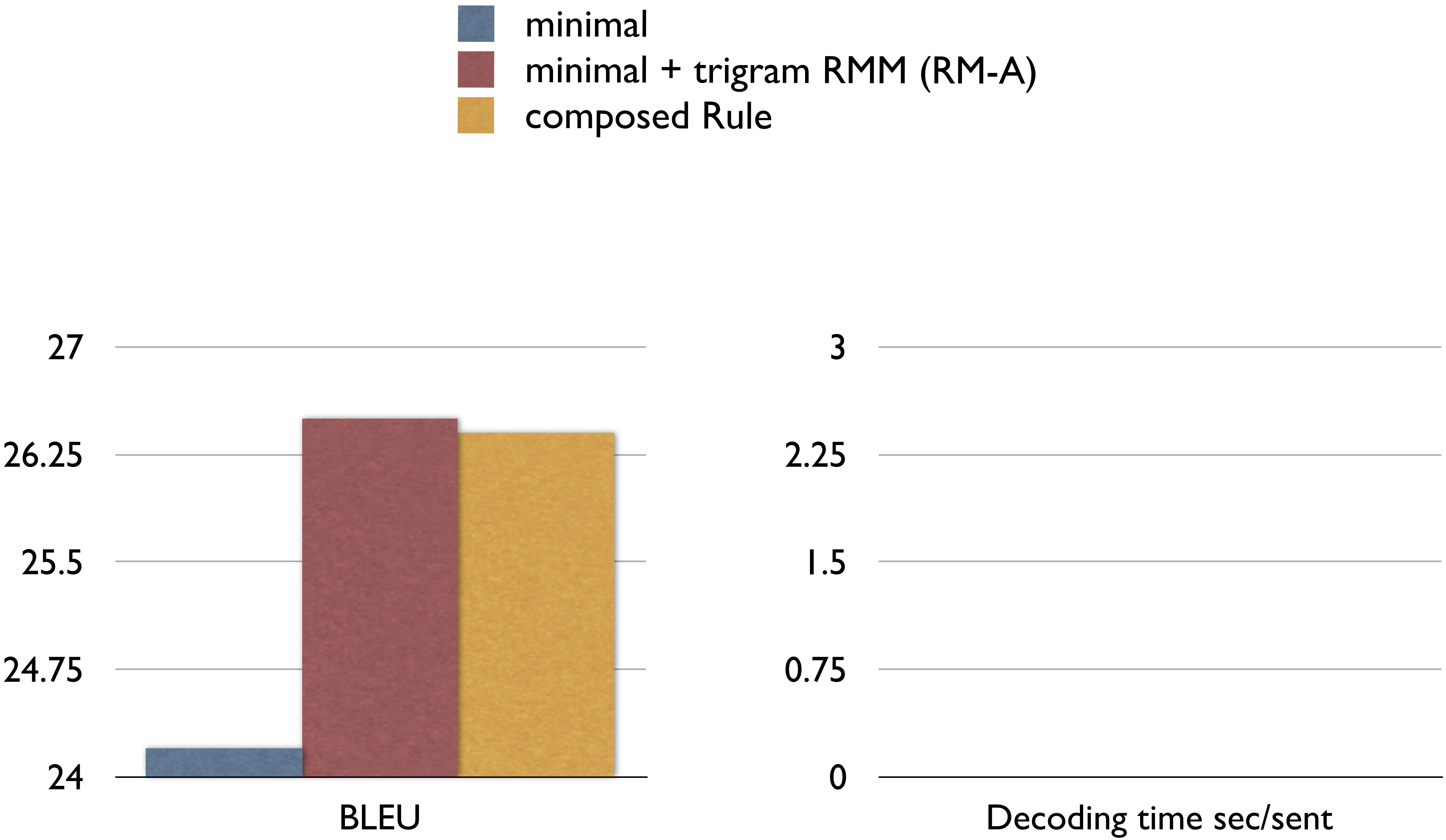

minimal
minimal + trigram RMM (RM-A)
composed Rule
27
26.25
25.5
24.75
24
BLEU

Results



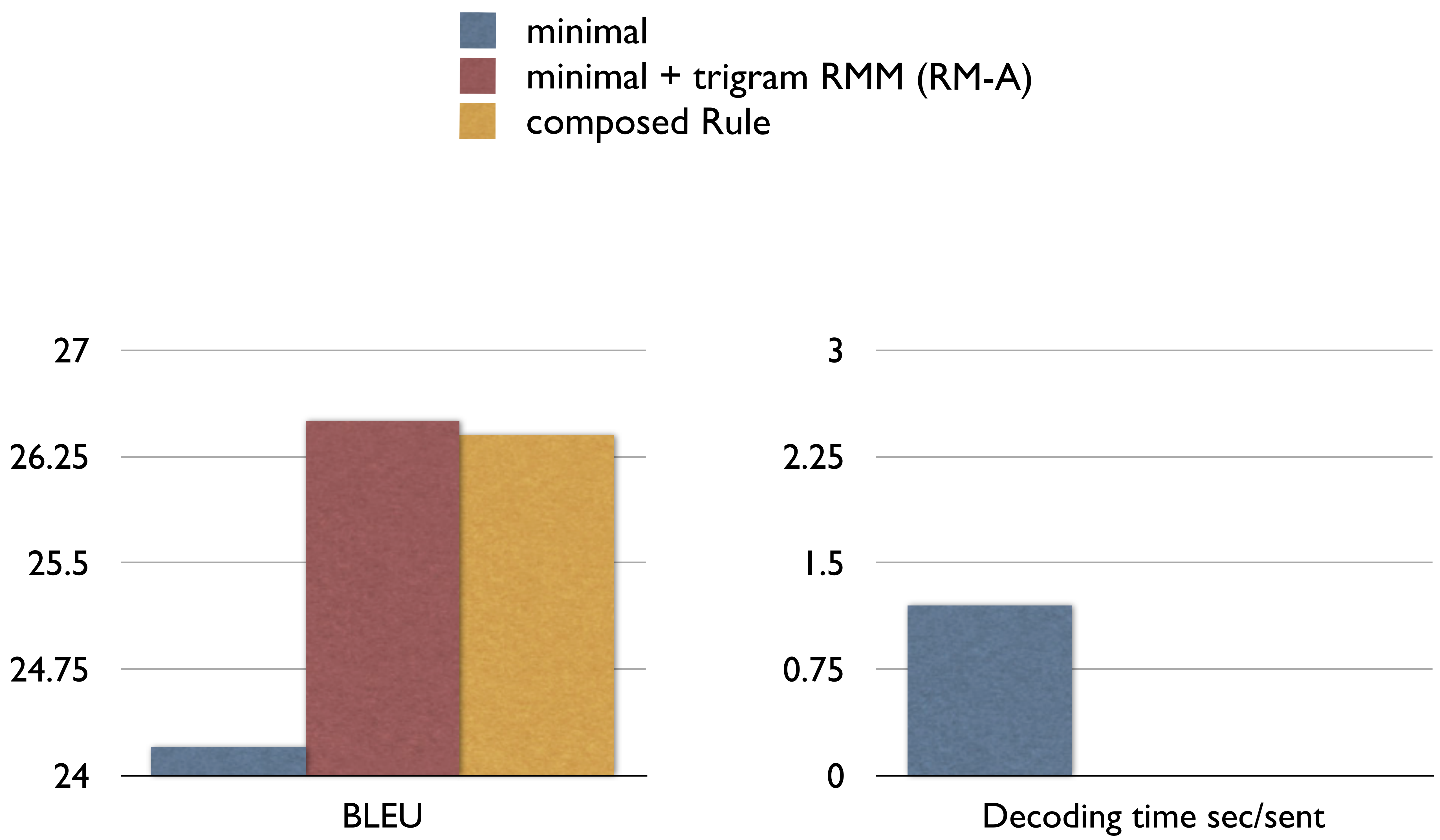
minimal
minimal + trigram RMM (RM-A)
composed Rule
27
26.25
25.5
24.75
24
BLEU

Results



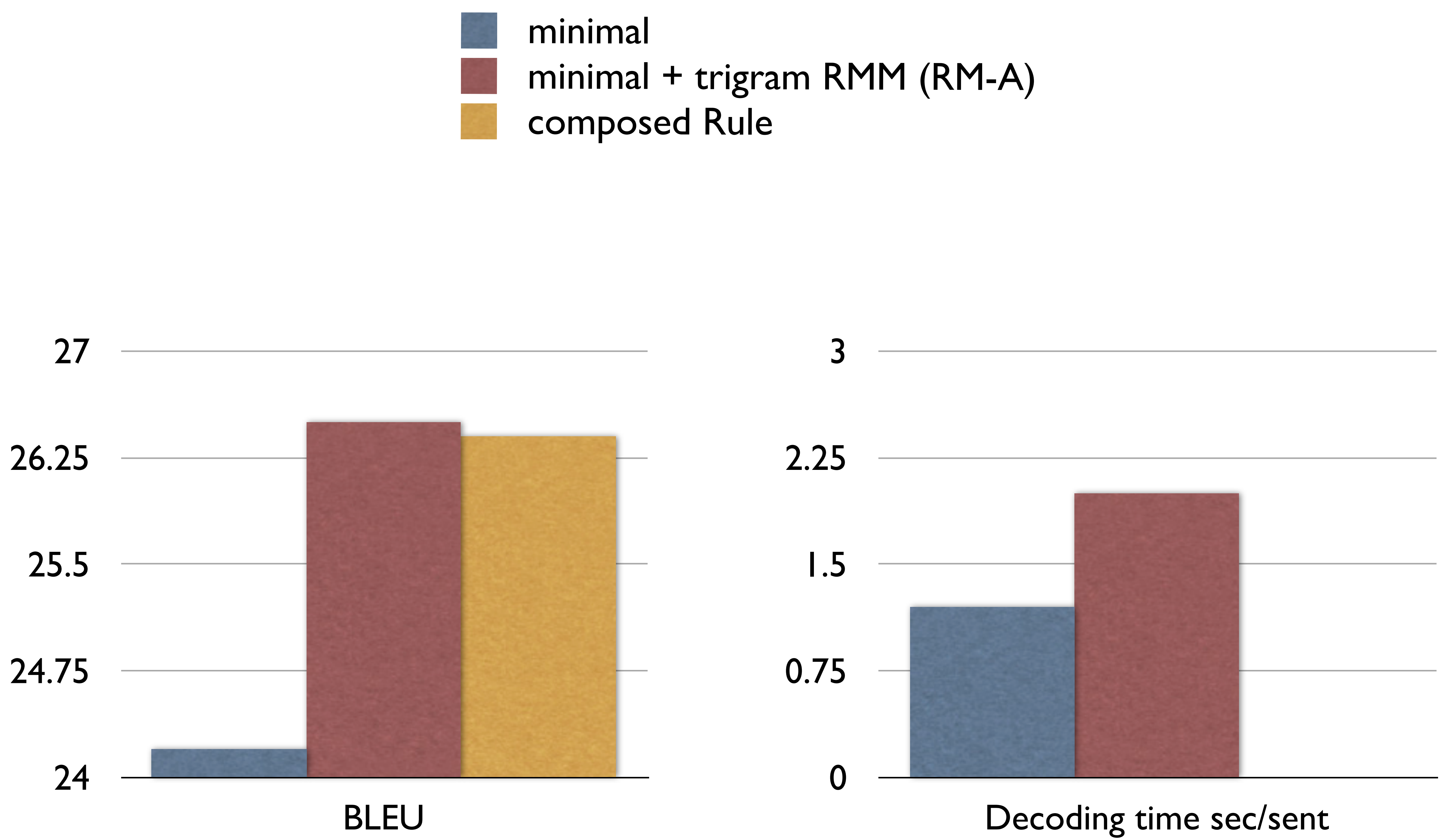
minimal
minimal + trigram RMM (RM-A)
composed Rule
27
26.25
25.5
24.75
24
BLEU
3
2.25
1.5
0.75
0
Decoding time sec/sent

Results

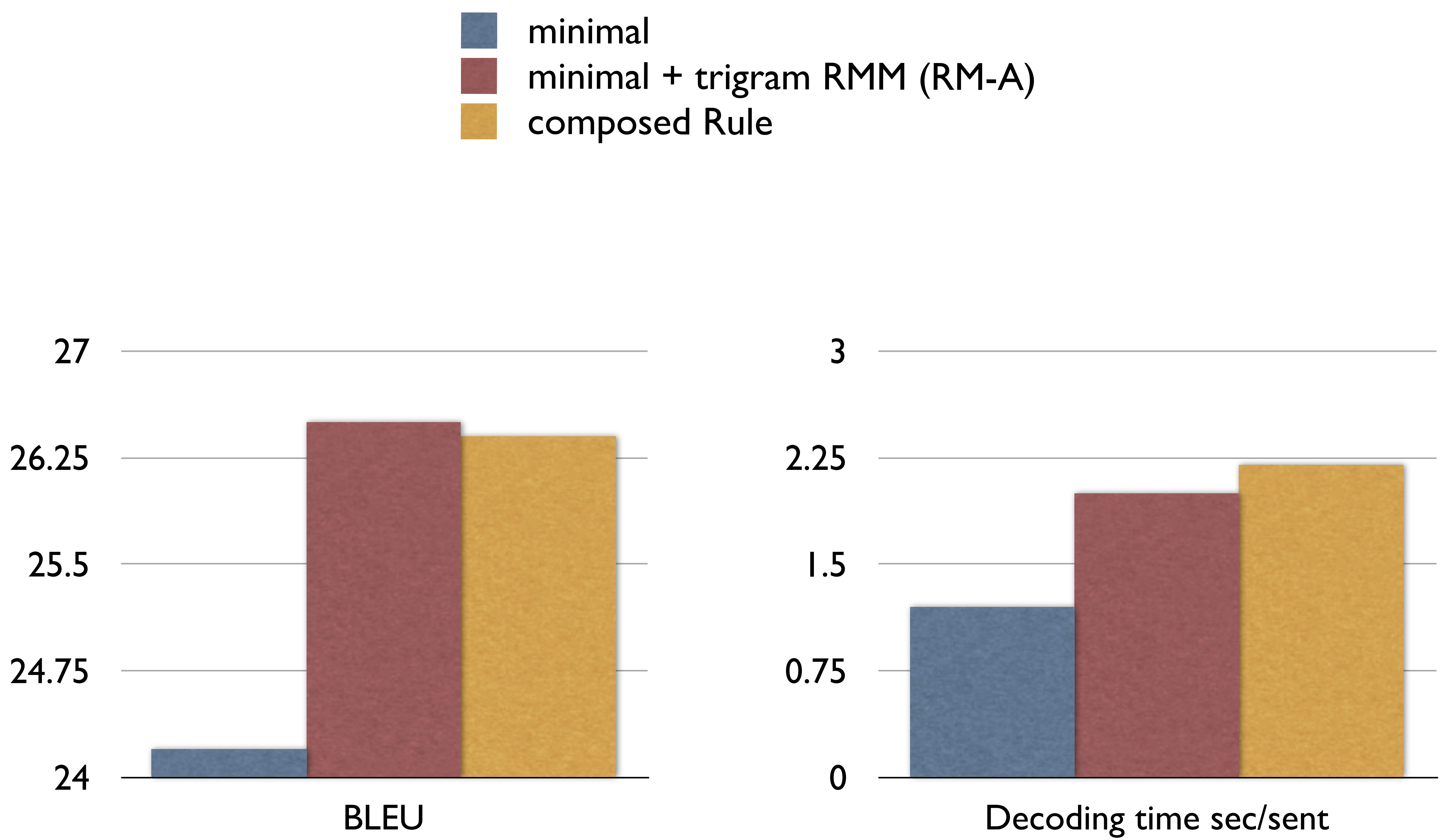


minimal
minimal + trigram RMM (RM-A)
composed Rule
27
26.25
25.5
24.75
24
BLEU
3
2.25
1.5
0.75
0
Decoding time sec/sent

Results


minimal
minimal + trigram RMM (RM-A)
composed Rule
27
26.25
25.5
24.75
24
BLEU
3
2.25
1.5
0.75
0
Decoding time sec/sent

Results



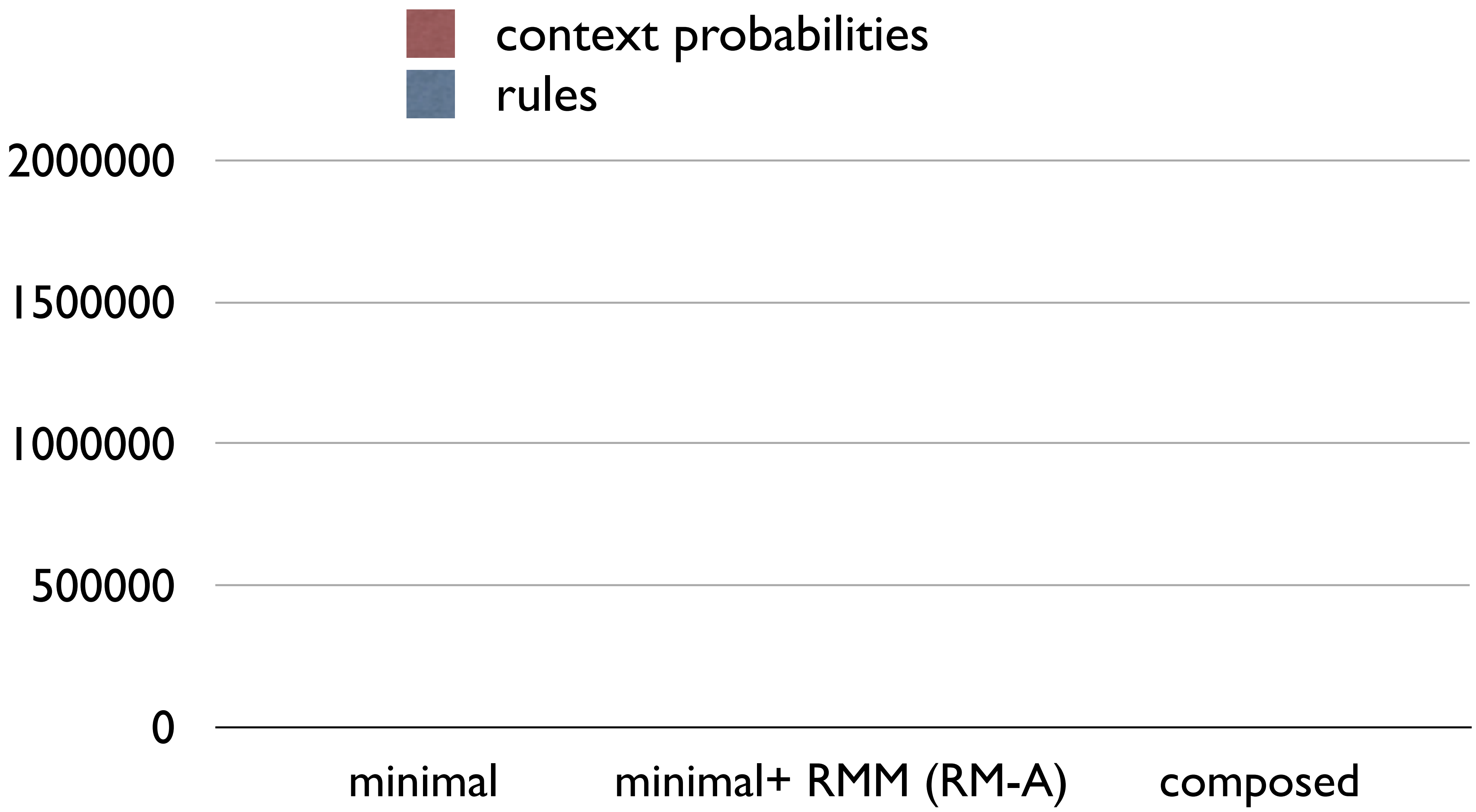
minimal
minimal + trigram RMM (RM-A)
composed Rule
27
26.25
25.5
24.75
24
BLEU
3
2.25
1.5
0.75
0
Decoding time sec/sent

Results

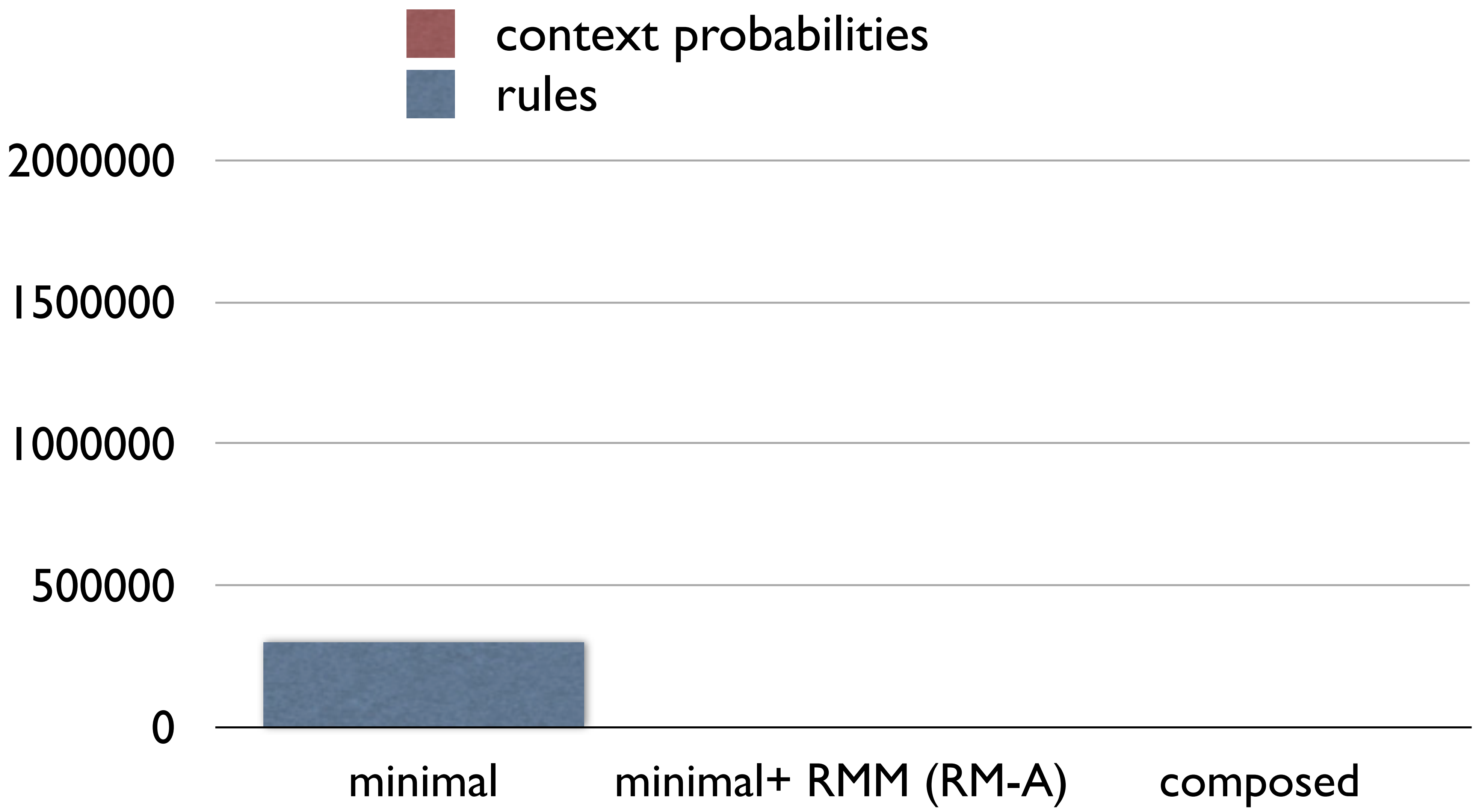
context probabilities

rules

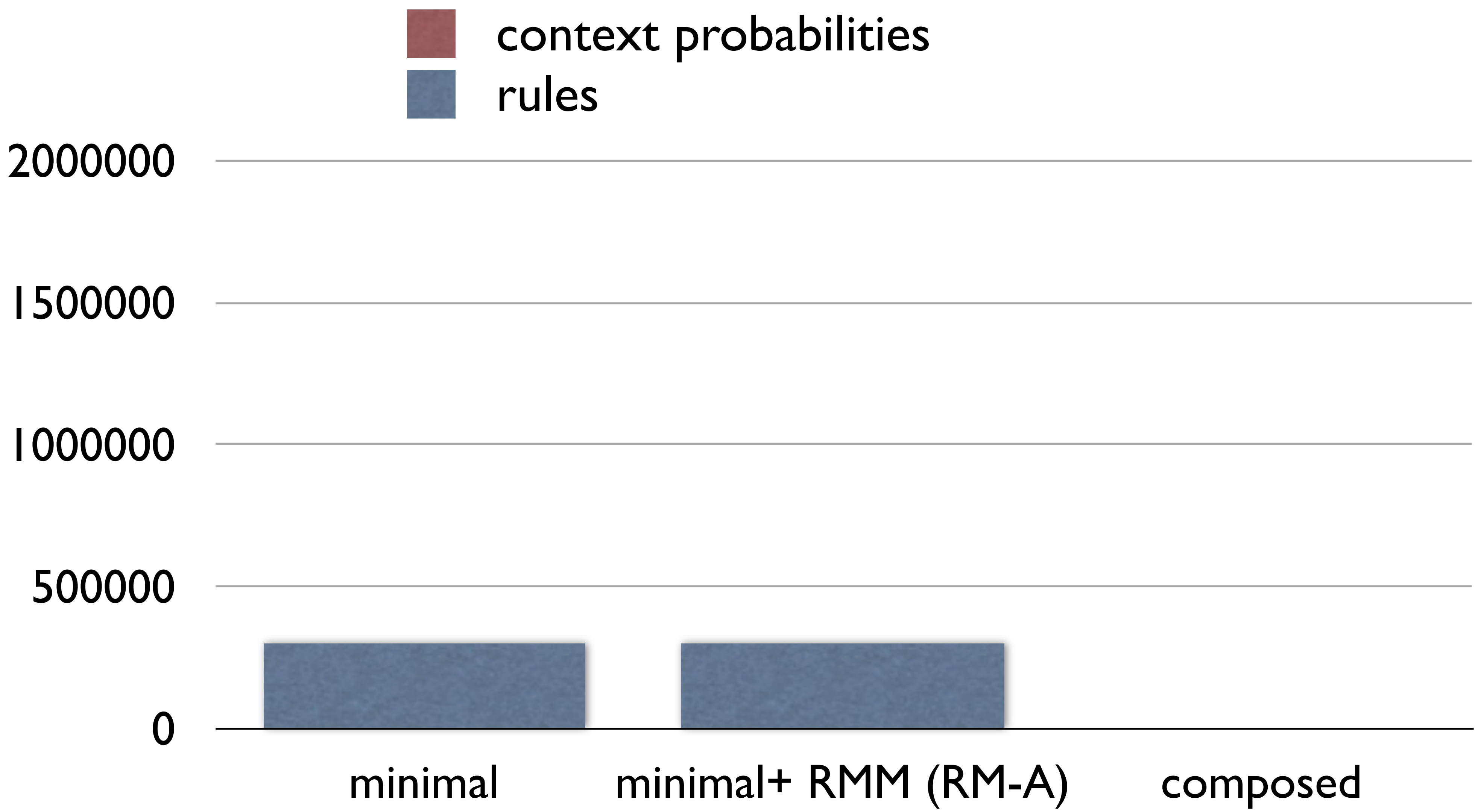
Results


context probabilities
rules
2000000
1500000
1000000
500000
0
minimal
minimal+ RMM (RM-A)
composed

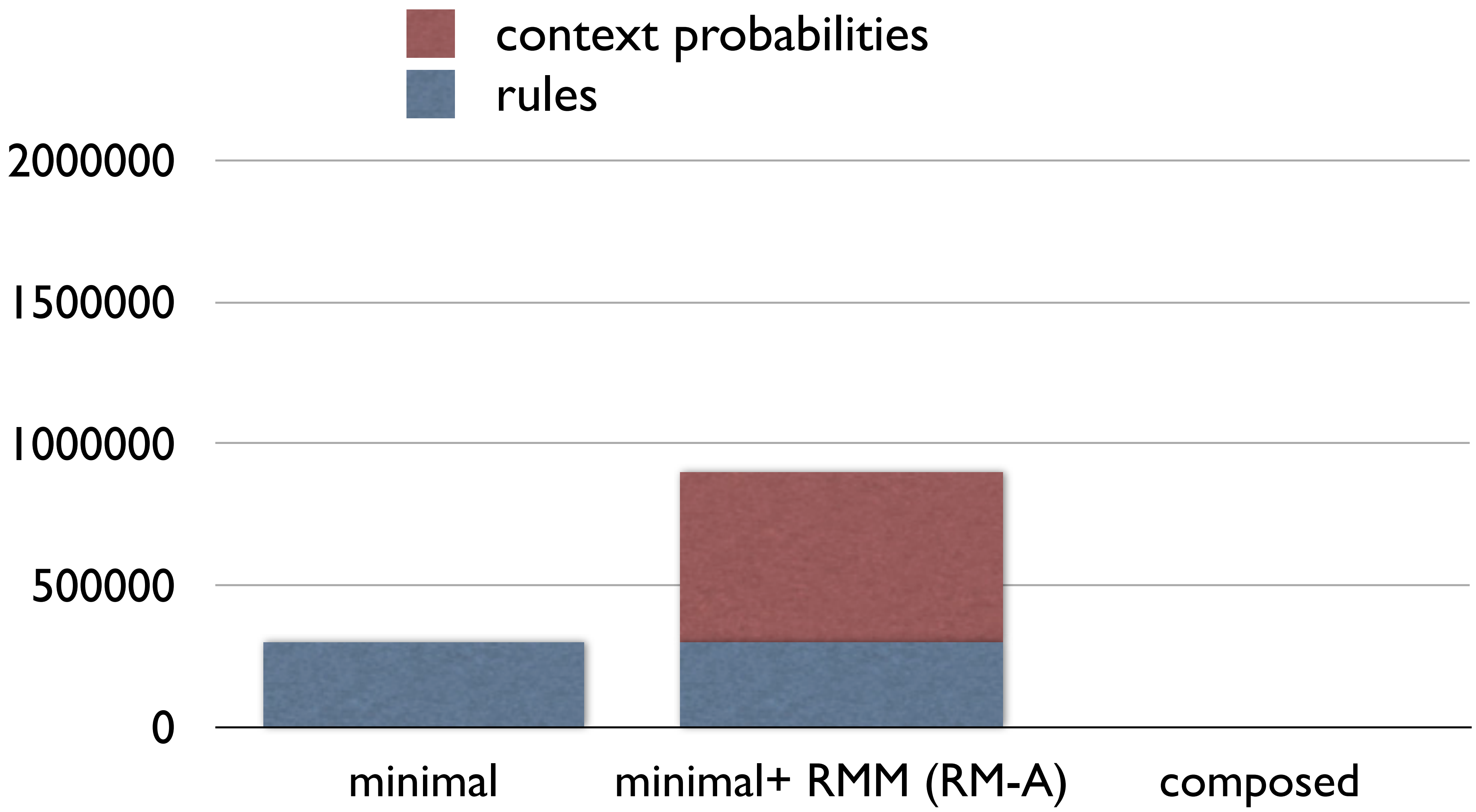
Results


context probabilities
rules
2000000
1500000
1000000
500000
0
minimal
minimal+ RMM (RM-A)
composed

Results

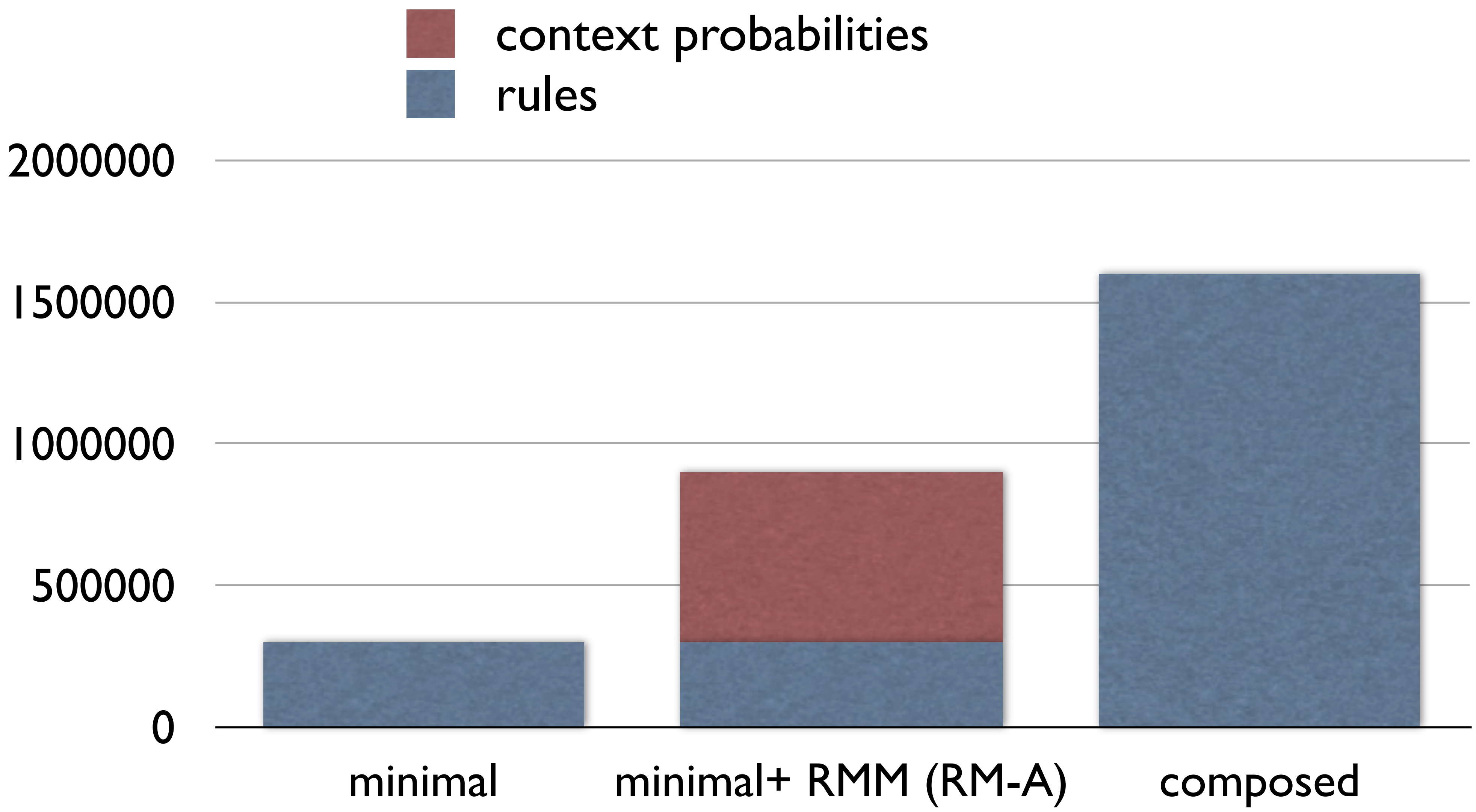

context probabilities
rules
2000000
1500000
1000000
500000
0
minimal
minimal+ RMM (RM-A)
composed

Results



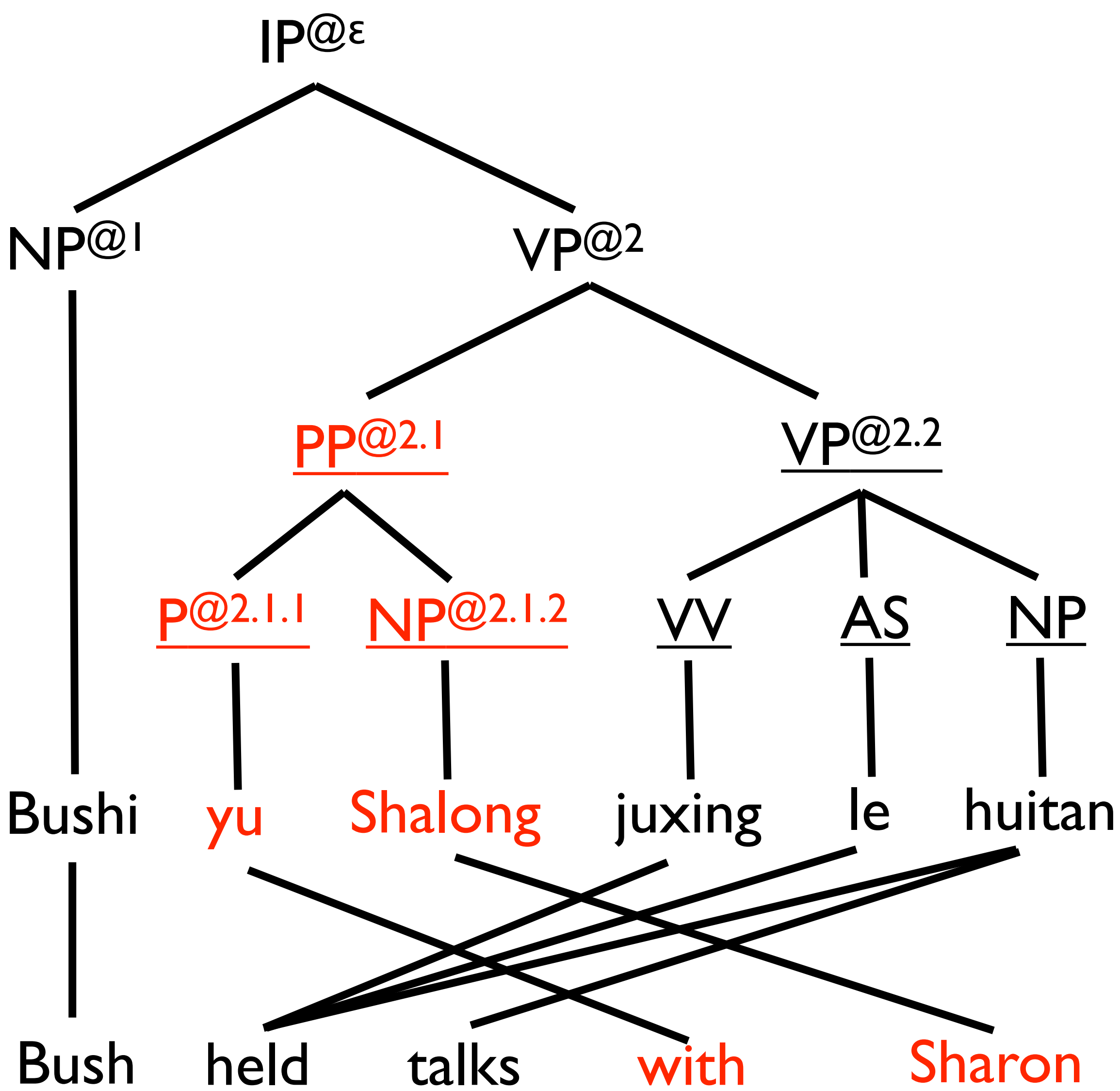
context probabilities
rules
2000000
1500000
1000000
500000
0
minimal
minimal+ RMM (RM-A)
composed

Results

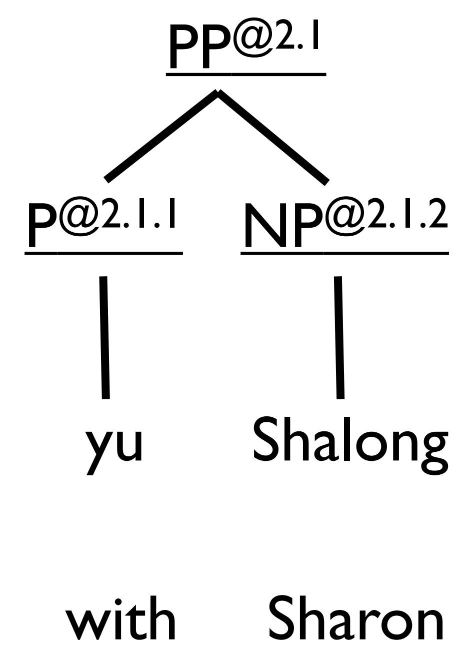


context probabilities
rules
2000000
1500000
1000000
500000
0
minimal
minimal+ RMM (RM-A)
composed

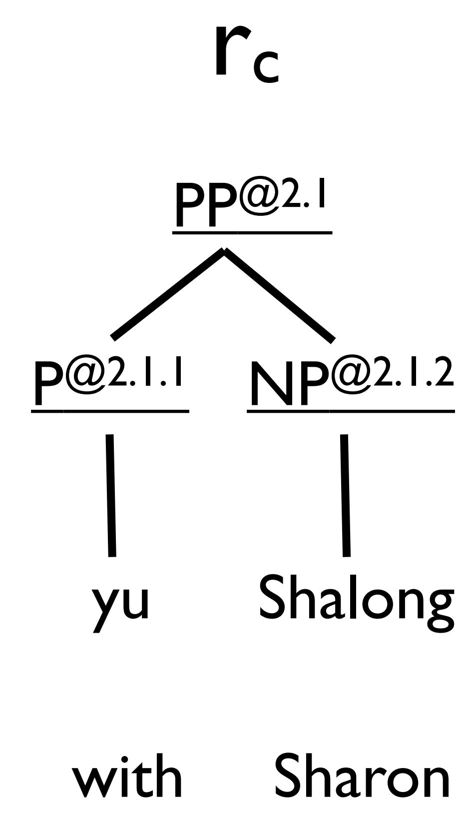
RMMs with composed Rules



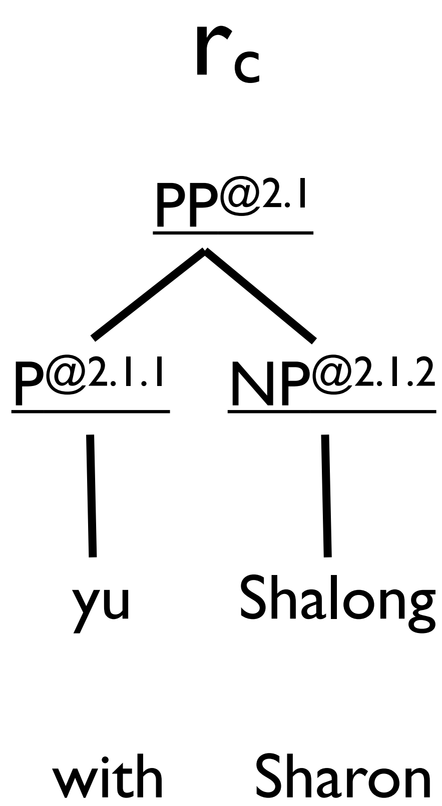
r_c



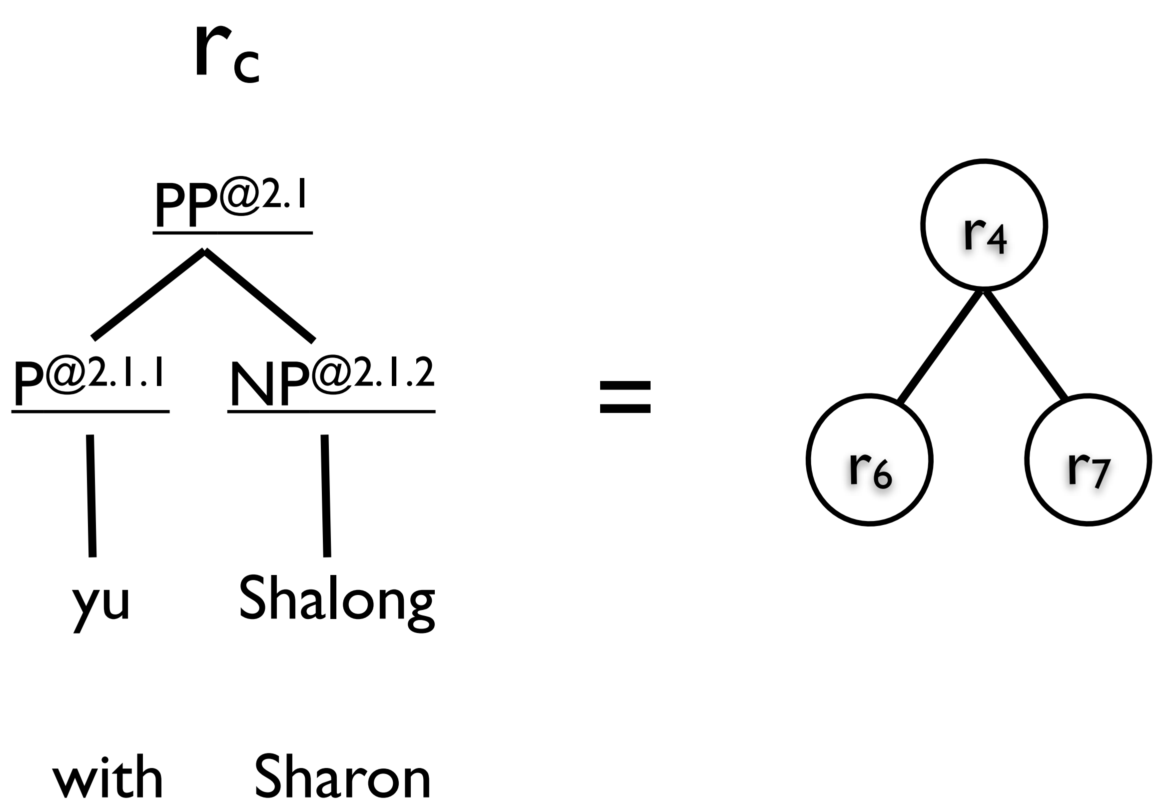
RMMs with composed Rules



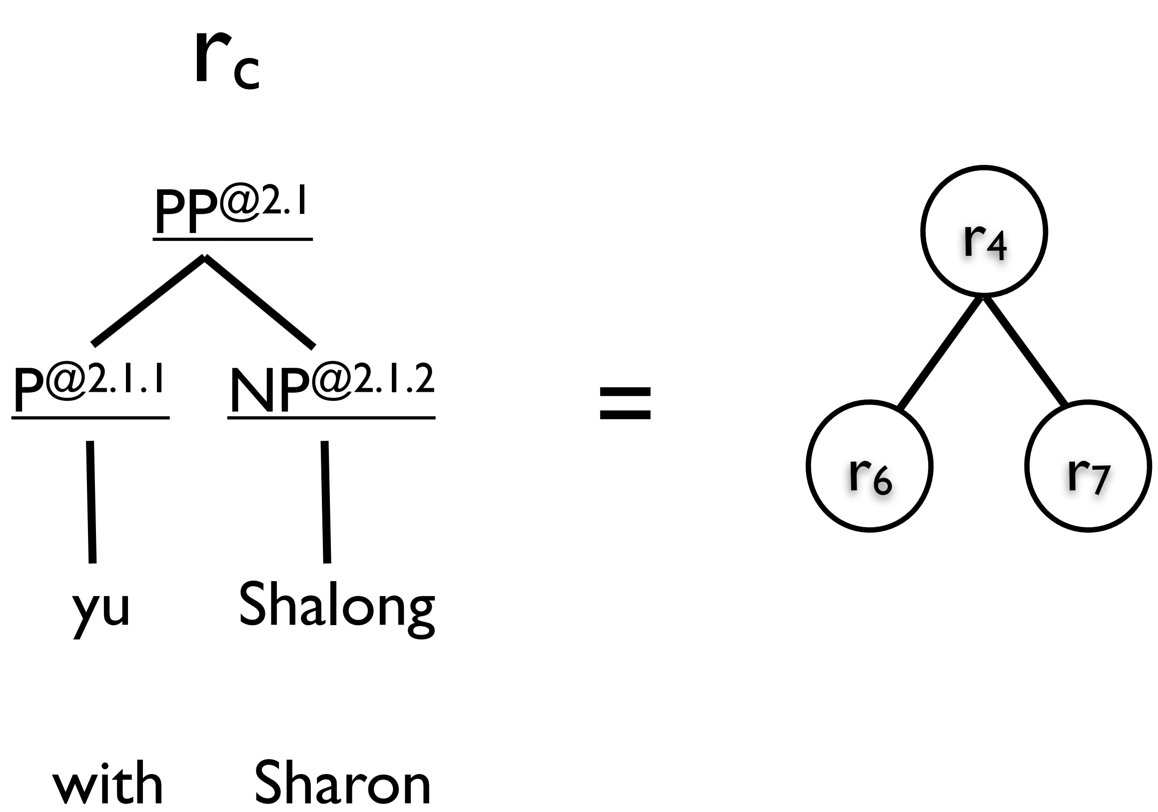
RMMs with composed Rules

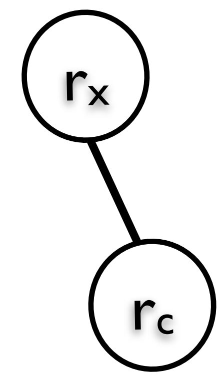


RMMs with composed Rules

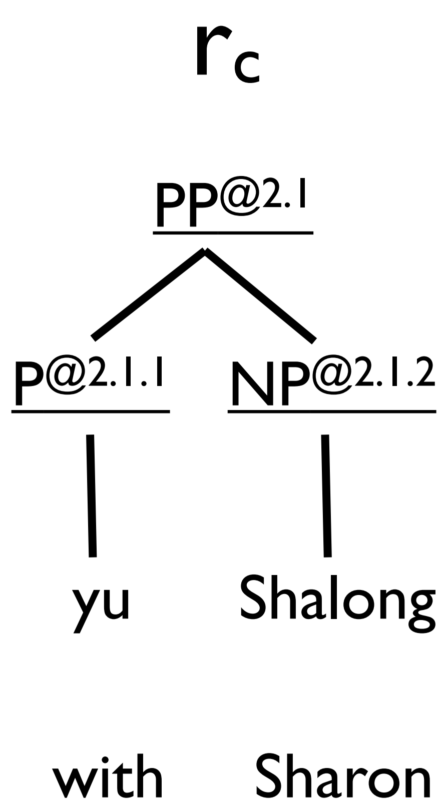


RMMs with composed Rules

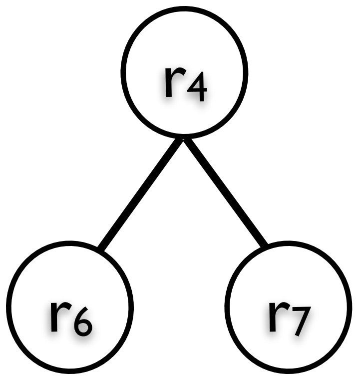


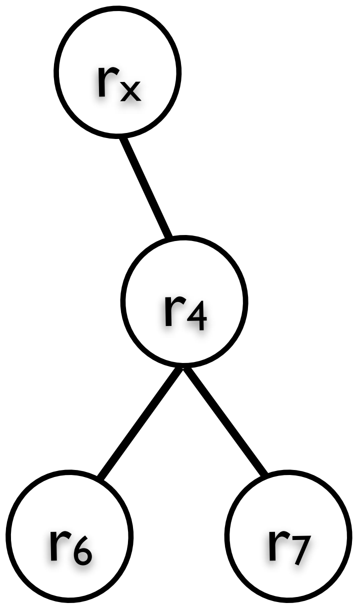


RMMs with composed Rules

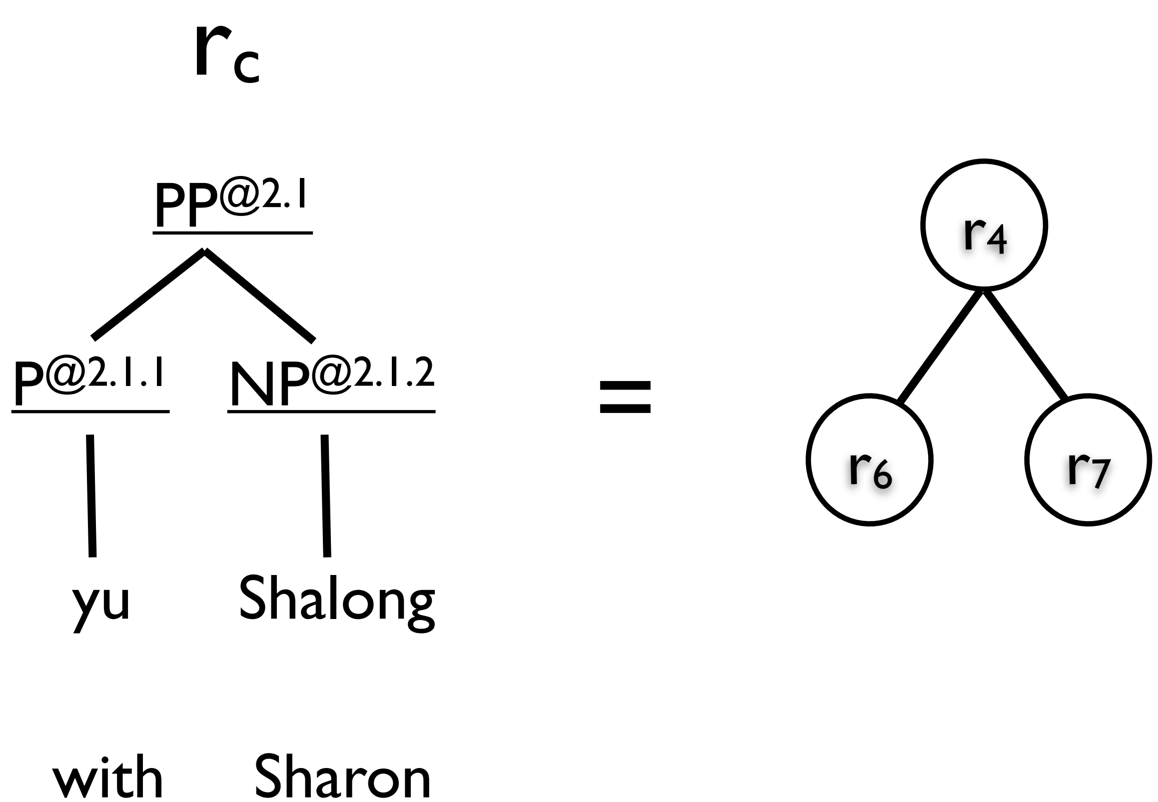


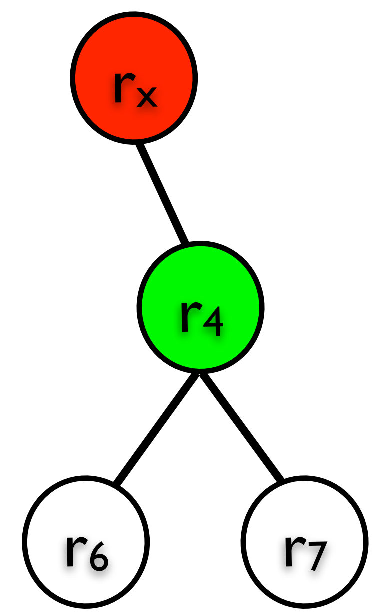
=





RMMs with composed Rules

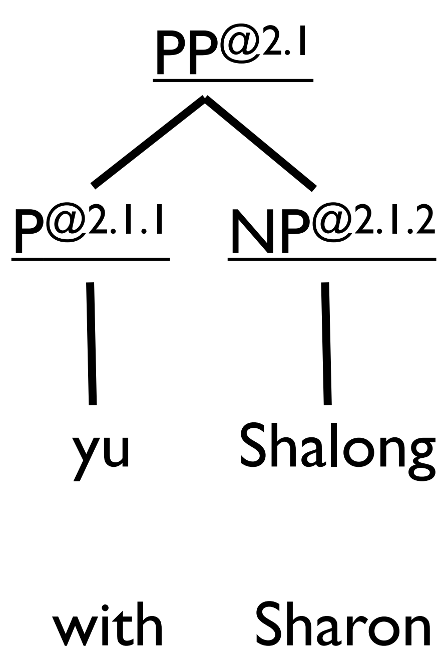




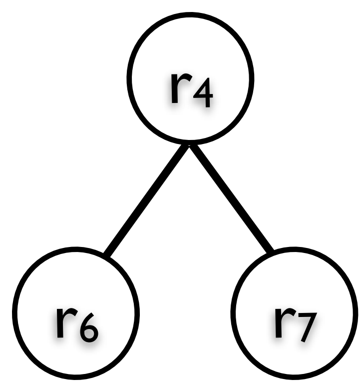
$$P(r_c|r_x) = P(r_4|r_x) \cdot P(r_6|r_4,r_x) \cdot P(r_7|r_4,r_x)$$

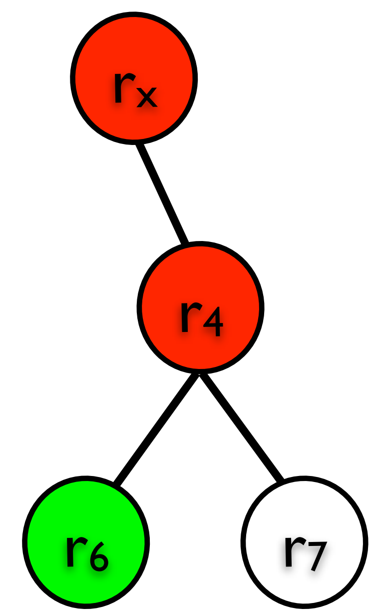
RMMs with composed Rules

r_c



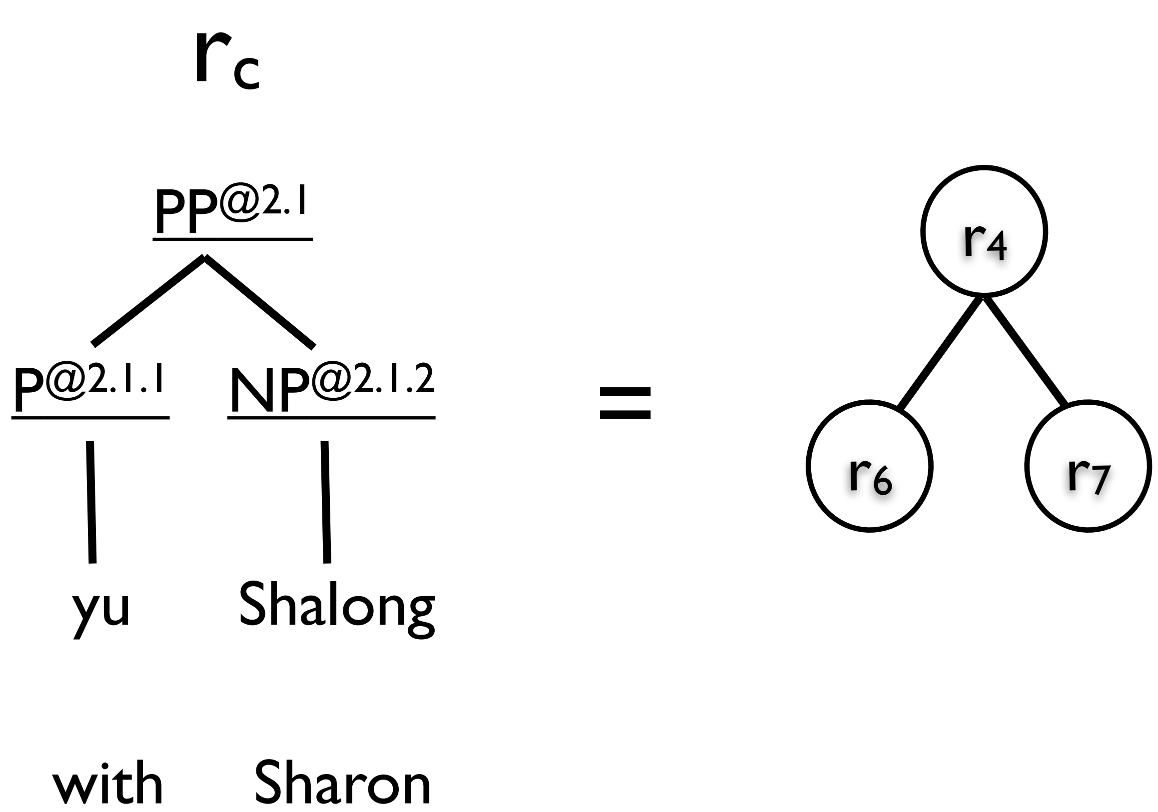
=

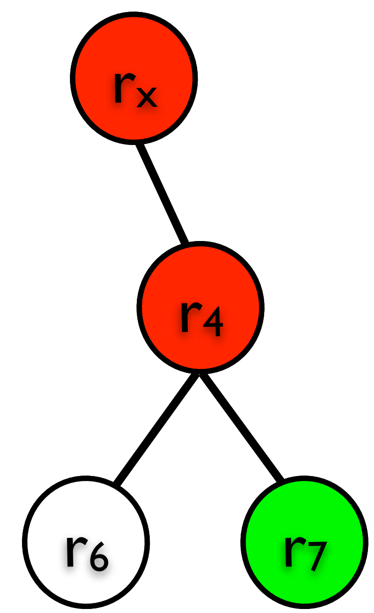




$$P(r_c|r_x) = P(r_4|r_x) \cdot P(r_6|r_4,r_x) \cdot P(r_7|r_4,r_x)$$

RMMs with composed Rules



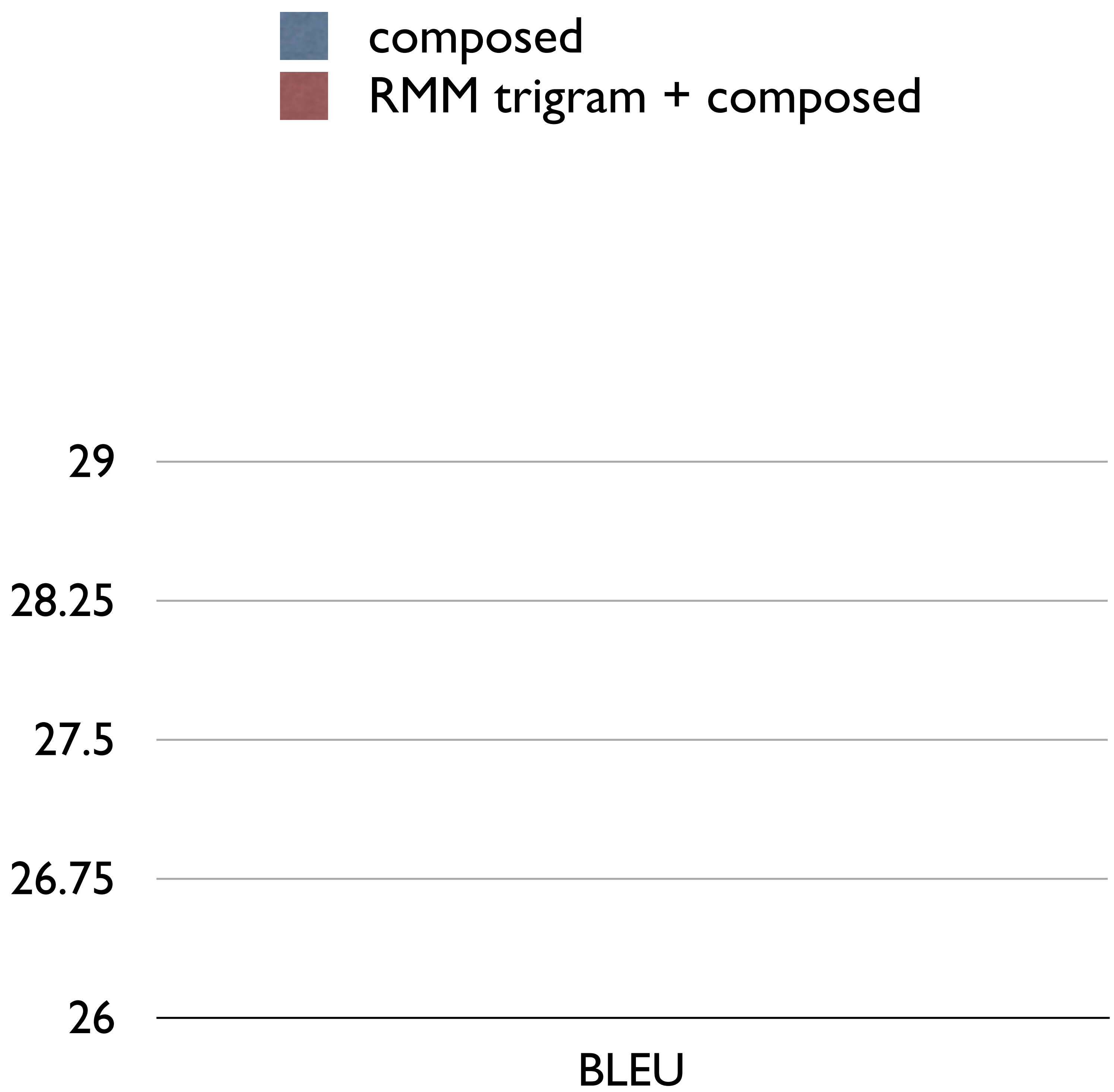


$$P(r_c|r_x) = P(r_4|r_x) \cdot P(r_6|r_4,r_x) \cdot P(r_7|r_4,r_x)$$

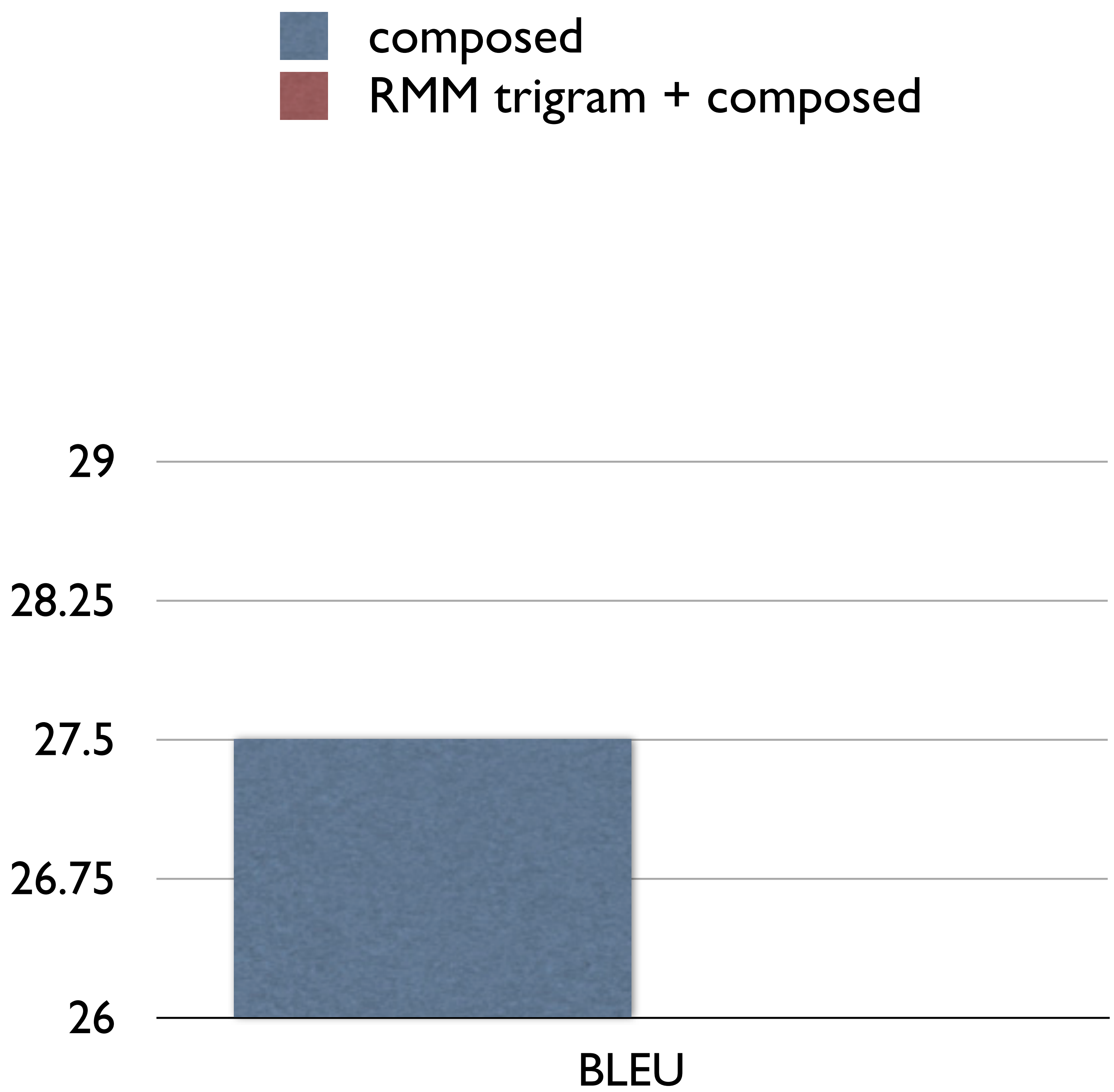
Results: RMMs improve over composed rules

composed
RMM trigram + composed

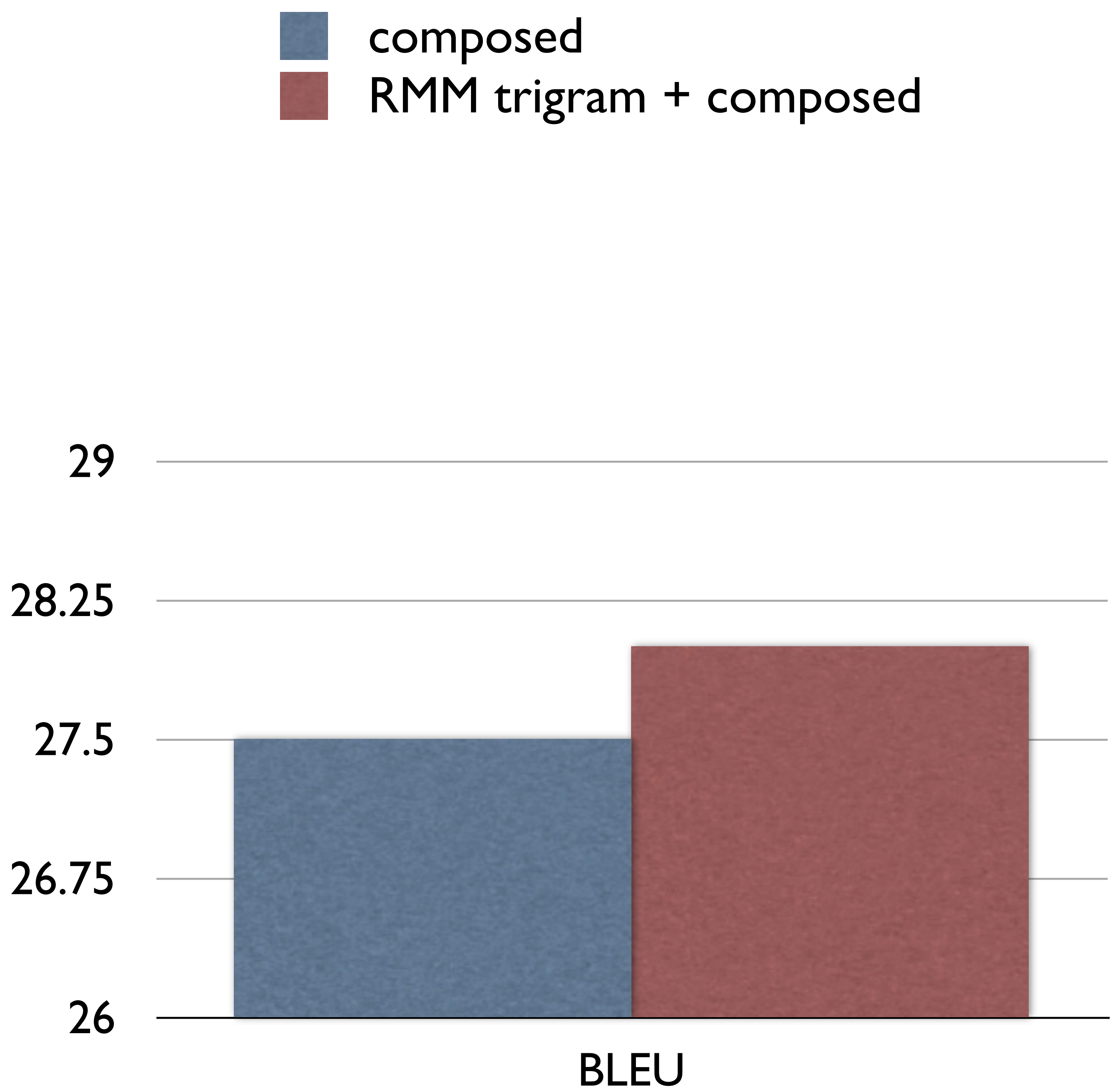
Results: RMMs improve over composed rules



Results: RMMs improve over composed rules



Results: RMMs improve over 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