

Joint Tokenization, Parsing and Translation

Yang Liu

*Institute of Computing Technology
Chinese Academy of Sciences*



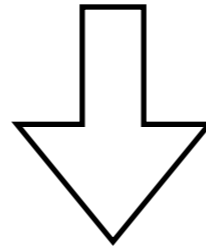
joint work with Xinyan Xiao, Qun Liu, Young-Sook Hwang, and Shouxun Lin

Tokenization

Mr. Smith, please pay 3,000 dollars for the computer.

Tokenization

Mr. Smith, please pay 3,000 dollars for the computer.



Mr. Smith , please pay 3,000 dollars for the computer .

Tokenization Ambiguity

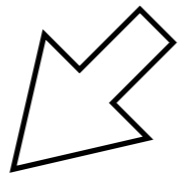
- Tokenization is easy for English, but hard for many languages such as Chinese

下雨天地面积水

Tokenization Ambiguity

- Tokenization is easy for English, but hard for many languages such as Chinese

下雨天地面积水



下雨天

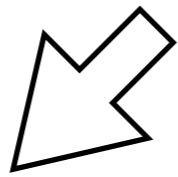
地面

积水

Tokenization Ambiguity

- Tokenization is easy for English, but hard for many languages such as Chinese

下雨天地面积水



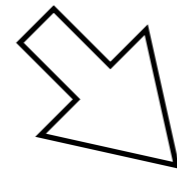
下雨天 地面 积水

in rainy days the ground is wet

Tokenization Ambiguity

- Tokenization is easy for English, but hard for many languages such as Chinese

下雨天地面积水



下雨天 地面 积水

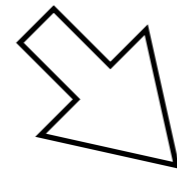
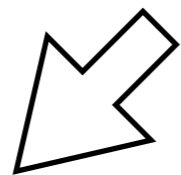
下雨 天地 面积 水

in rainy days the ground is wet

Tokenization Ambiguity

- Tokenization is easy for English, but hard for many languages such as Chinese

下雨天地面积水



下雨天 地面 积水

in rainy days the ground is wet

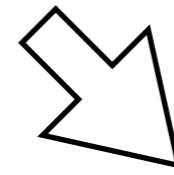
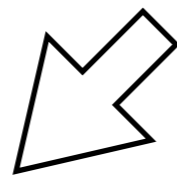
下雨 天地 面积 水

rain world area water

Tokenization Ambiguity

- Tokenization is easy for English, but hard for many languages such as Chinese

下雨天地面积水



下雨天 地面 积水

下雨 天地 面积 水

in rainy days the ground is wet

rain world area water

different tokenizations may lead to different translations

Phrase-based Translation on I-best Tokenization

下雨天

地面

积水

(Koehn et al., 2003)

Phrase-based Translation on I-best Tokenization

(地面, the ground)

下雨天

地面

积水

(Koehn et al., 2003)

Phrase-based Translation on I-best Tokenization

(地面, the ground)

下雨天

地面

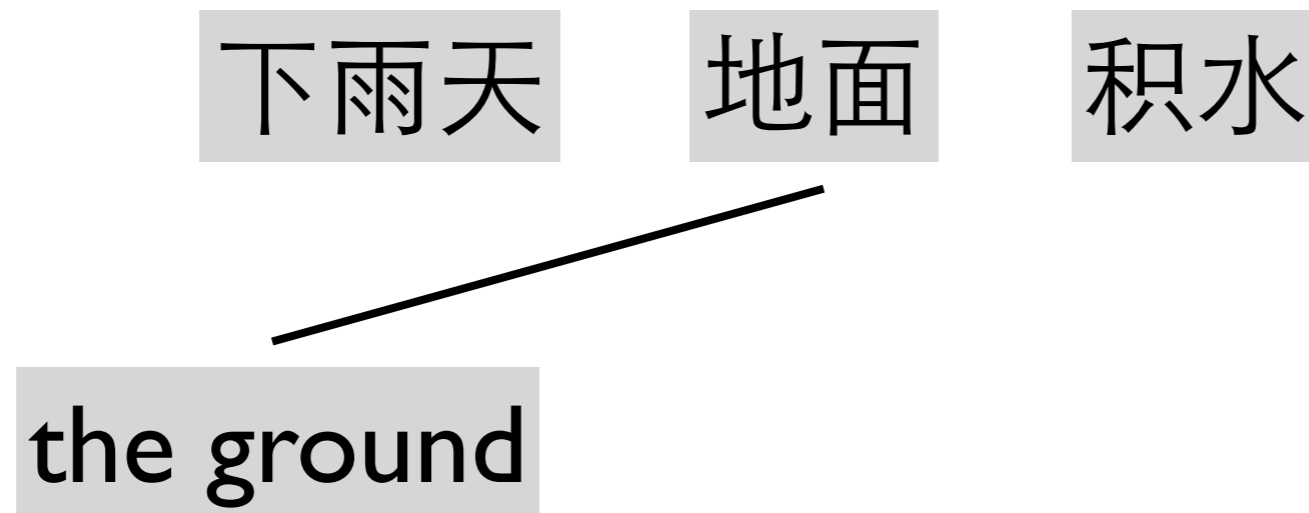
积水

the ground



(Koehn et al., 2003)

Phrase-based Translation on I-best Tokenization



(Koehn et al., 2003)

Phrase-based Translation on I-best Tokenization

(积水, is wet)

下雨天

地面

积水

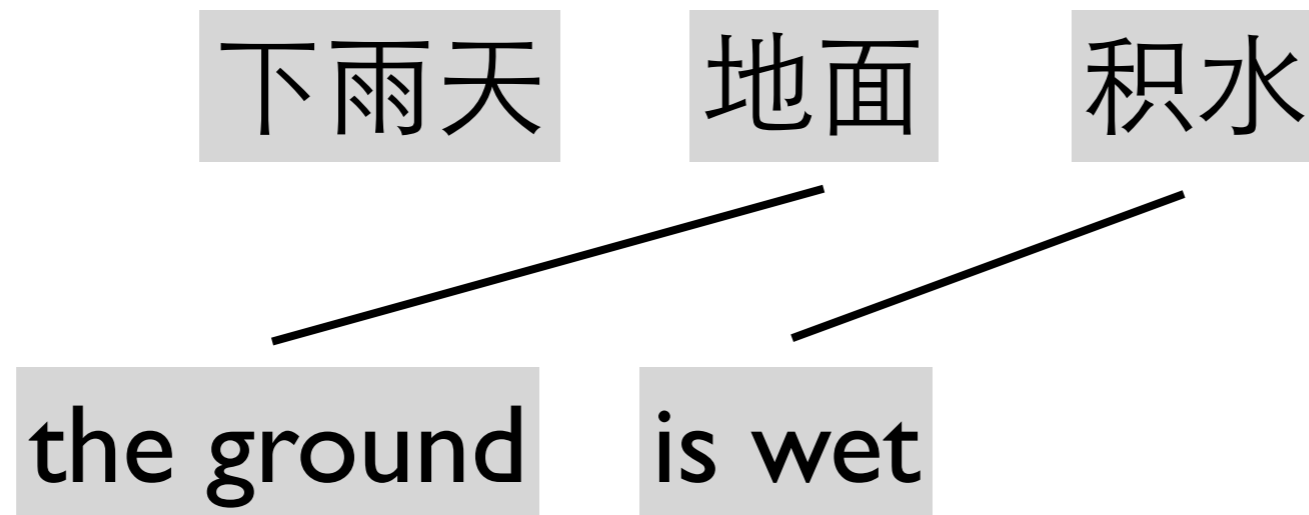
the ground



(Koehn et al., 2003)

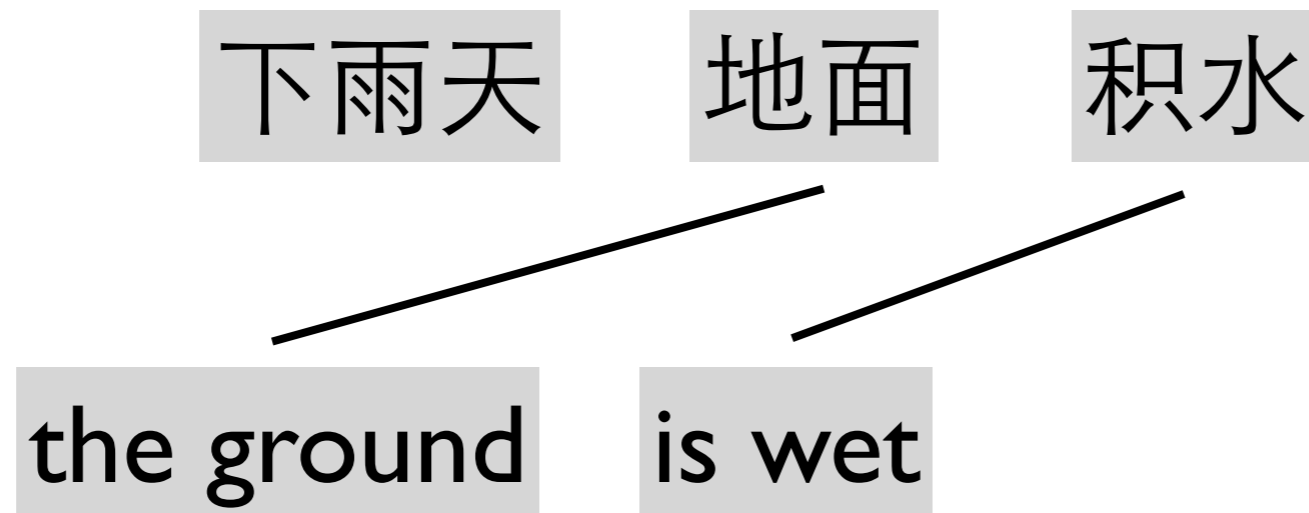
Phrase-based Translation on I-best Tokenization

(积水, is wet)



(Koehn et al., 2003)

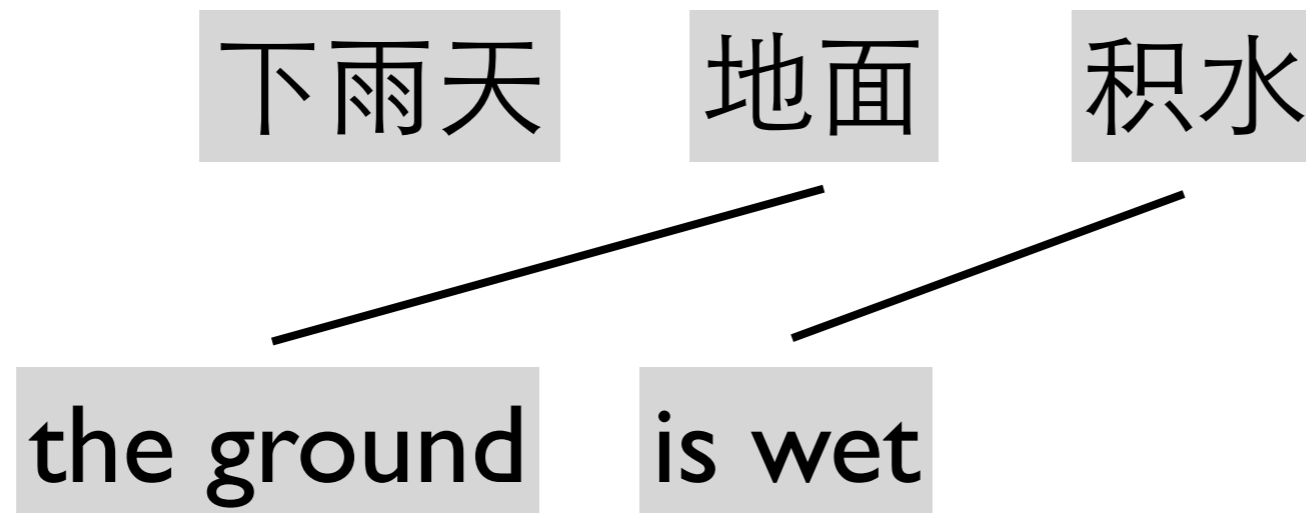
Phrase-based Translation on I-best Tokenization



(Koehn et al., 2003)

Phrase-based Translation on I-best Tokenization

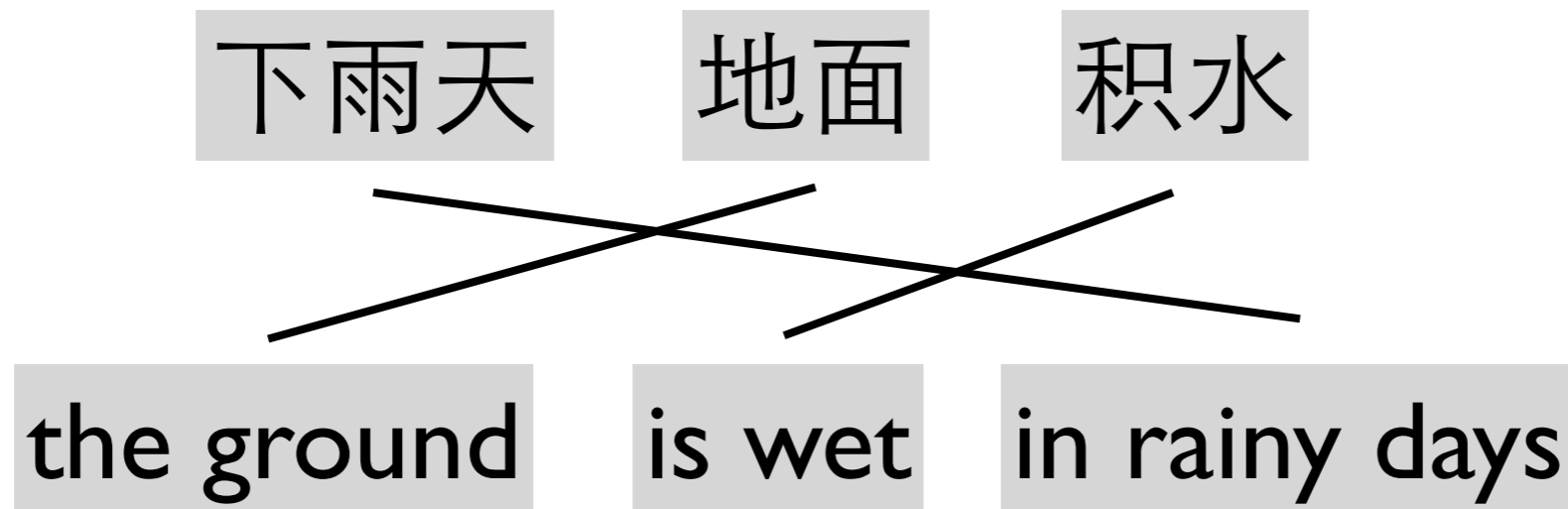
(下雨天, in rainy days)



(Koehn et al., 2003)

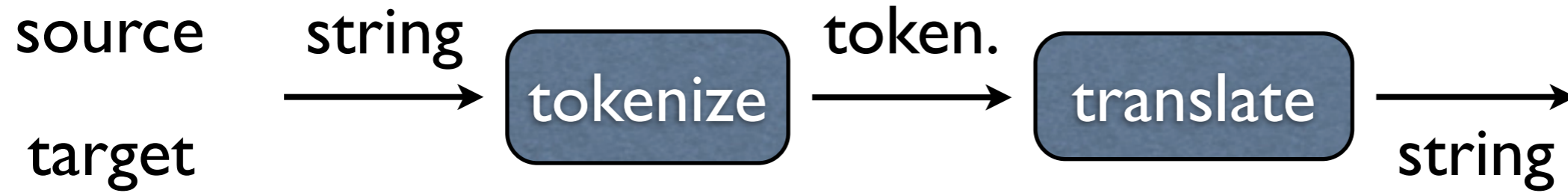
Phrase-based Translation on I-best Tokenization

(下雨天, in rainy days)

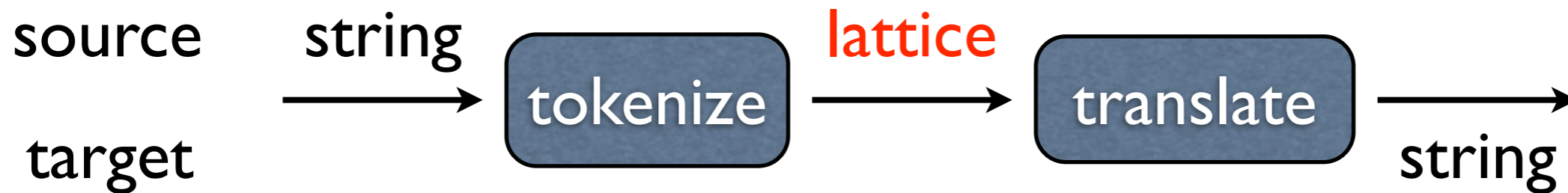
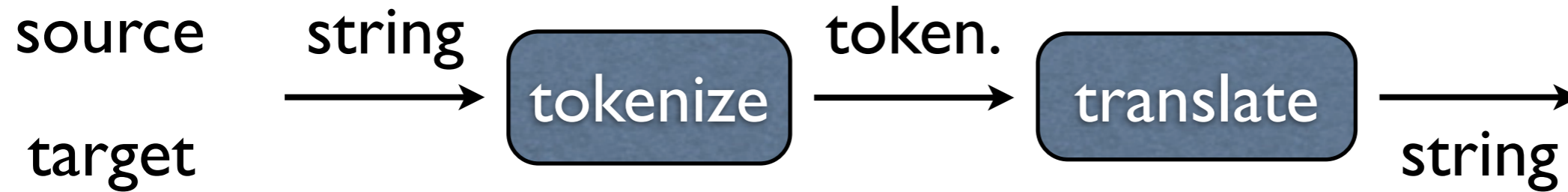


(Koehn et al., 2003)

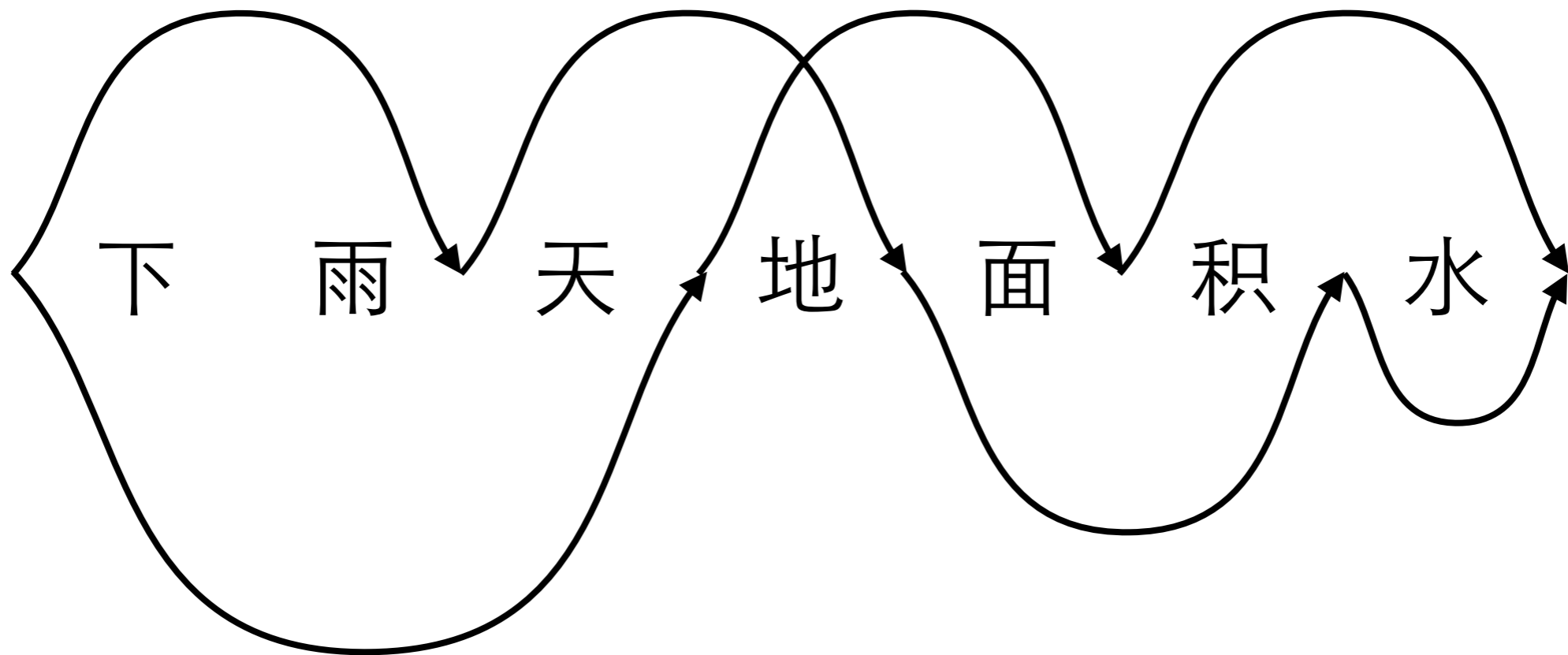
Tokenization Mistake Propagation



Tokenization Mistake Propagation



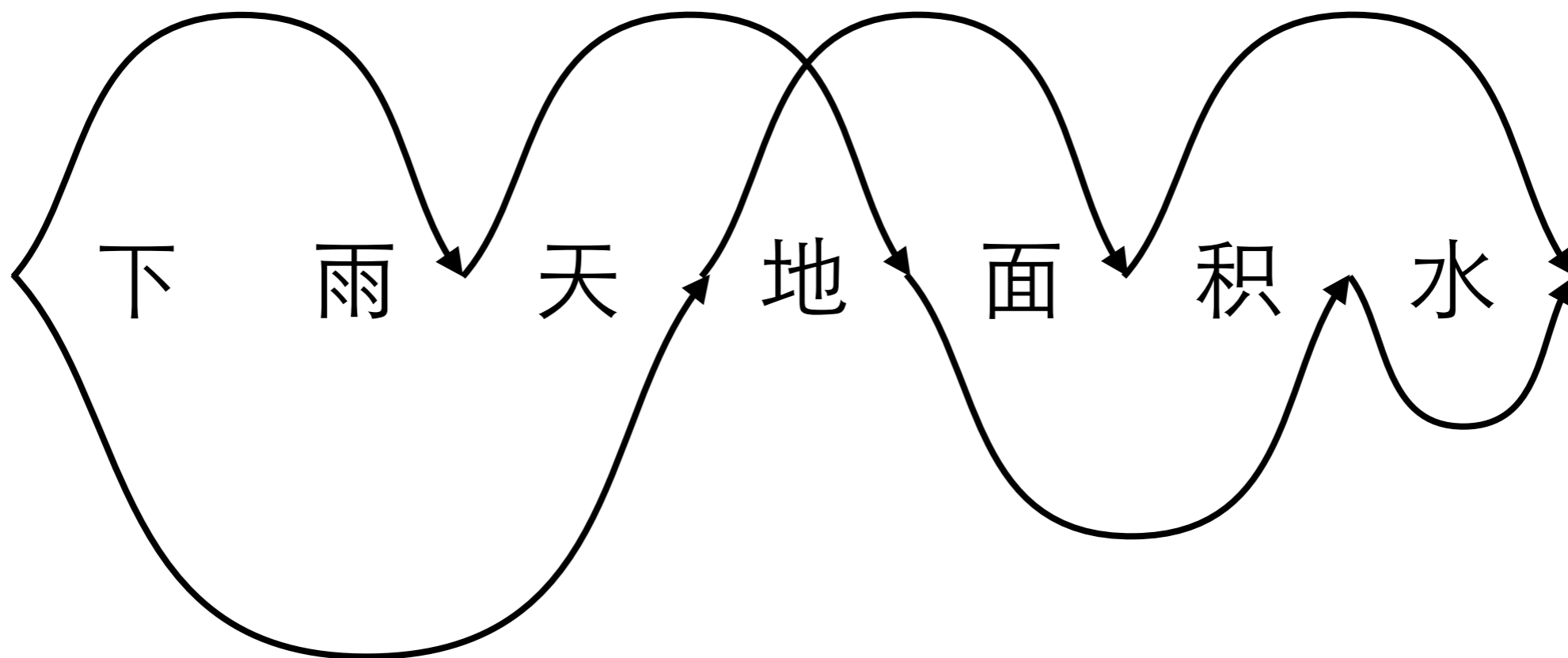
Phrase-based Translation on Lattice



(Dyer et al., 2008)

Phrase-based Translation on Lattice

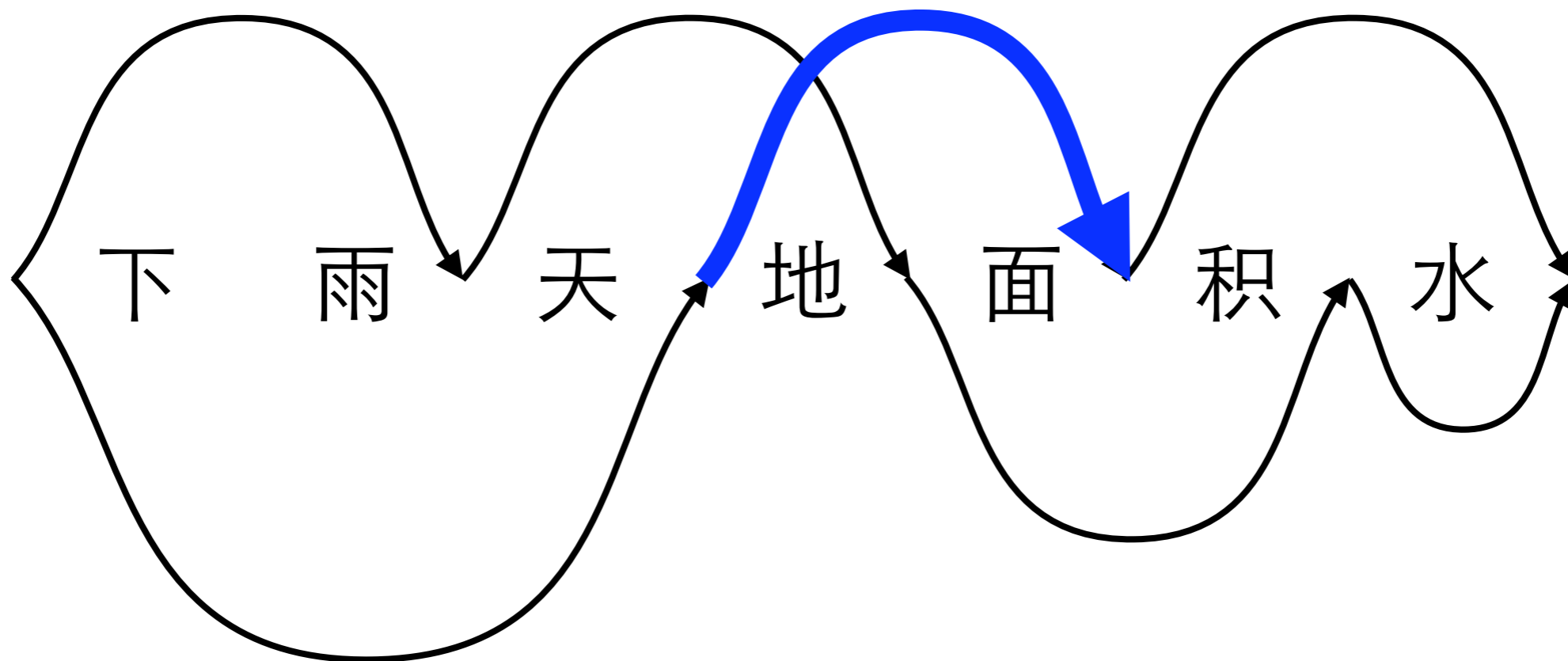
(地面, the ground)



(Dyer et al., 2008)

Phrase-based Translation on Lattice

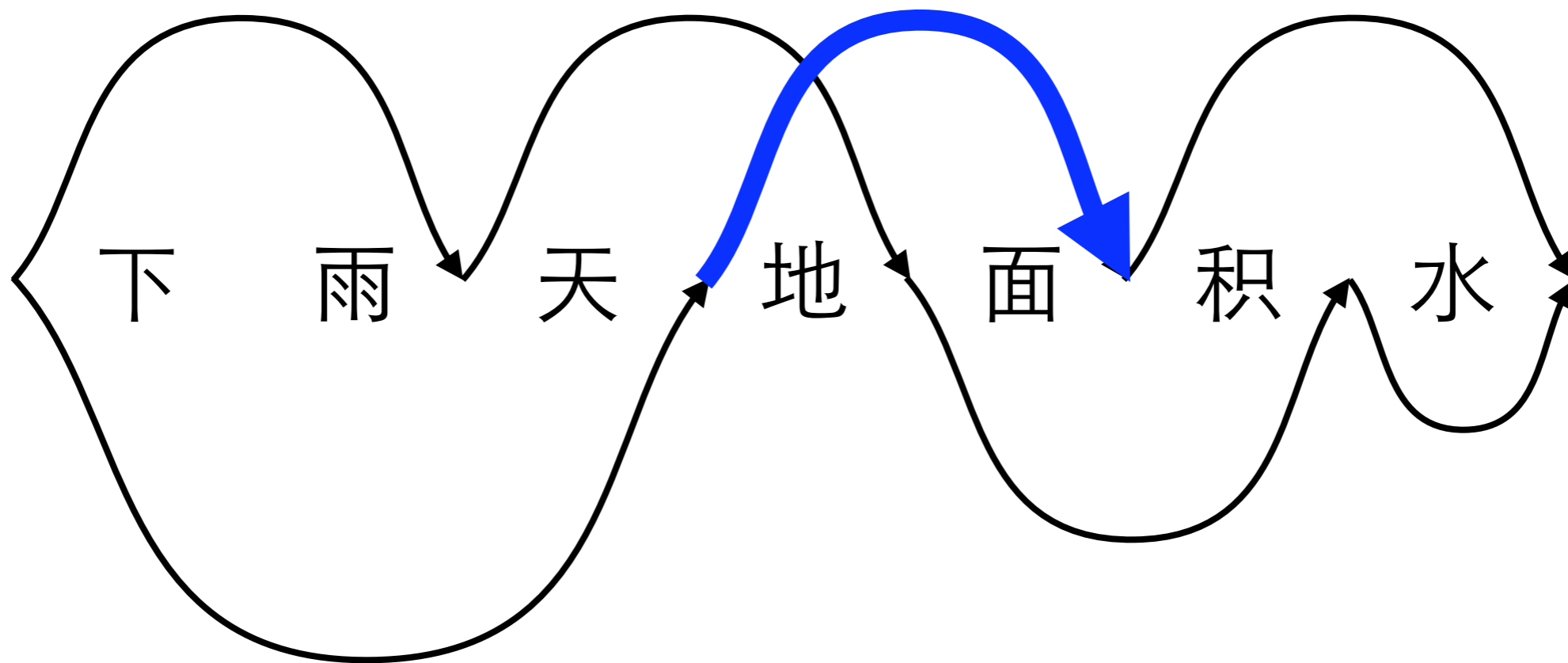
(地面, the ground)



(Dyer et al., 2008)

Phrase-based Translation on Lattice

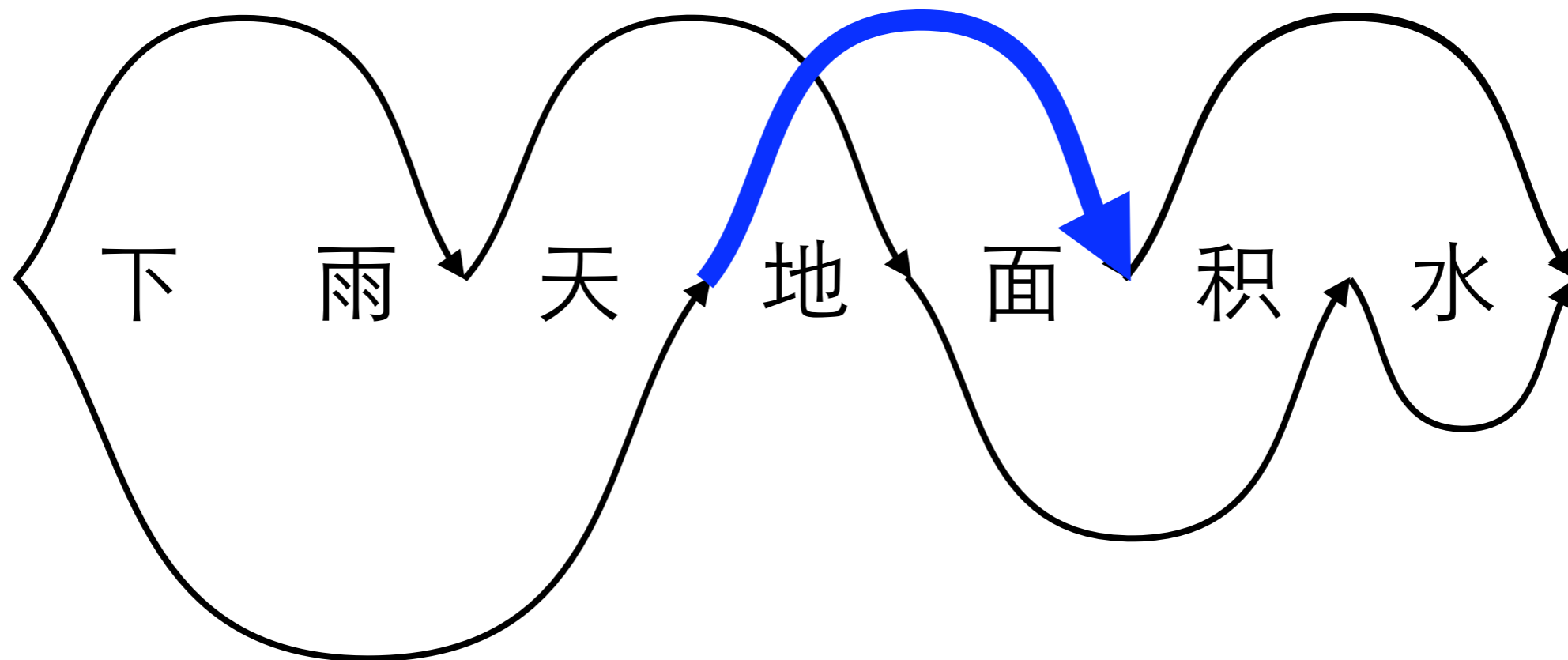
(地面, the ground)



the ground

(Dyer et al., 2008)

Phrase-based Translation on Lattice

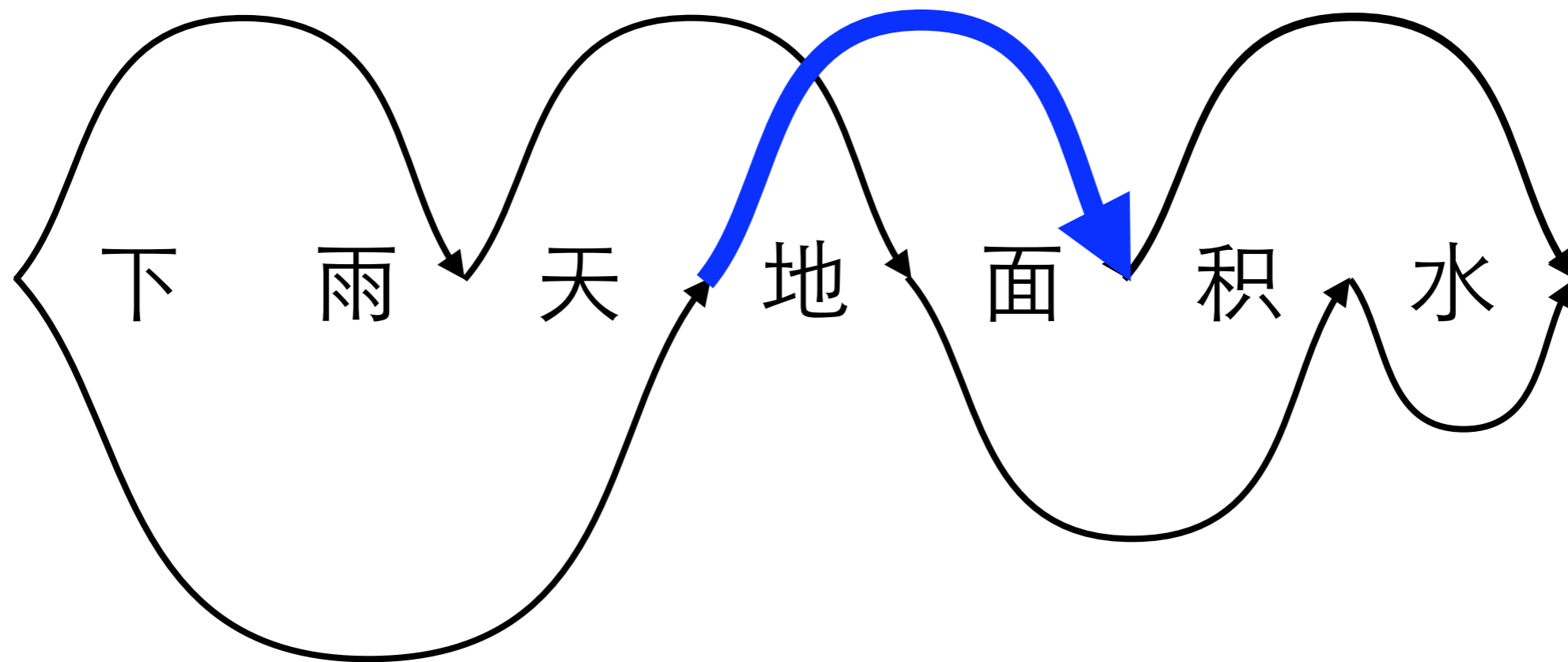


the ground

(Dyer et al., 2008)

Phrase-based Translation on Lattice

(积水, is wet)

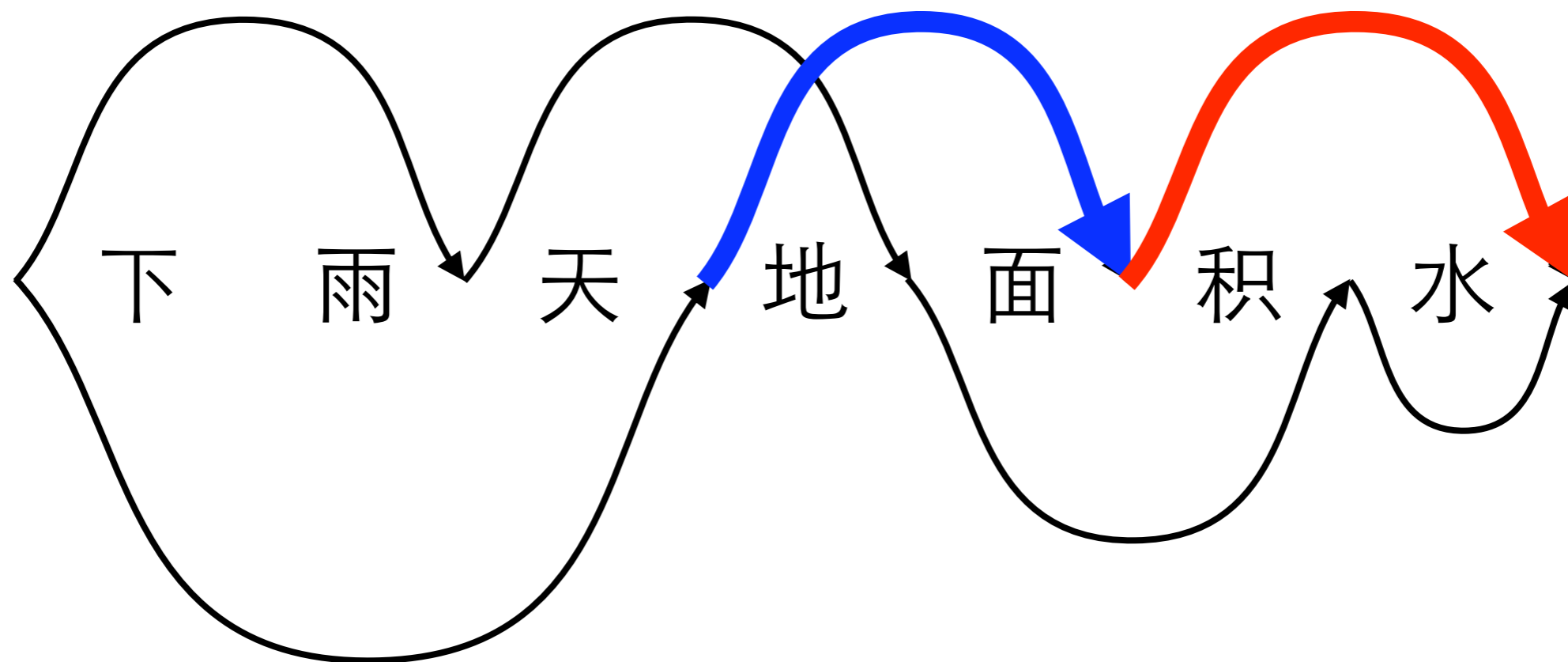


the ground

(Dyer et al., 2008)

Phrase-based Translation on Lattice

(积水, is wet)

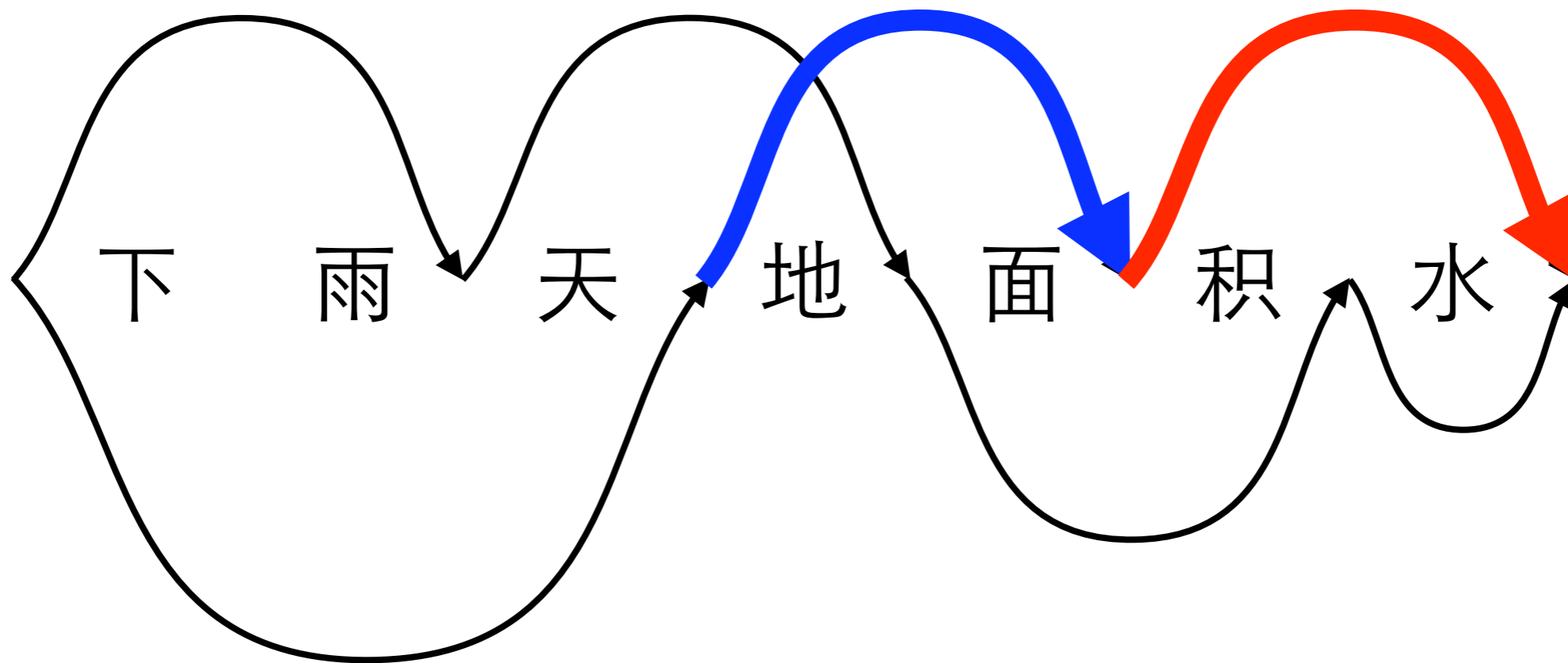


the ground

(Dyer et al., 2008)

Phrase-based Translation on Lattice

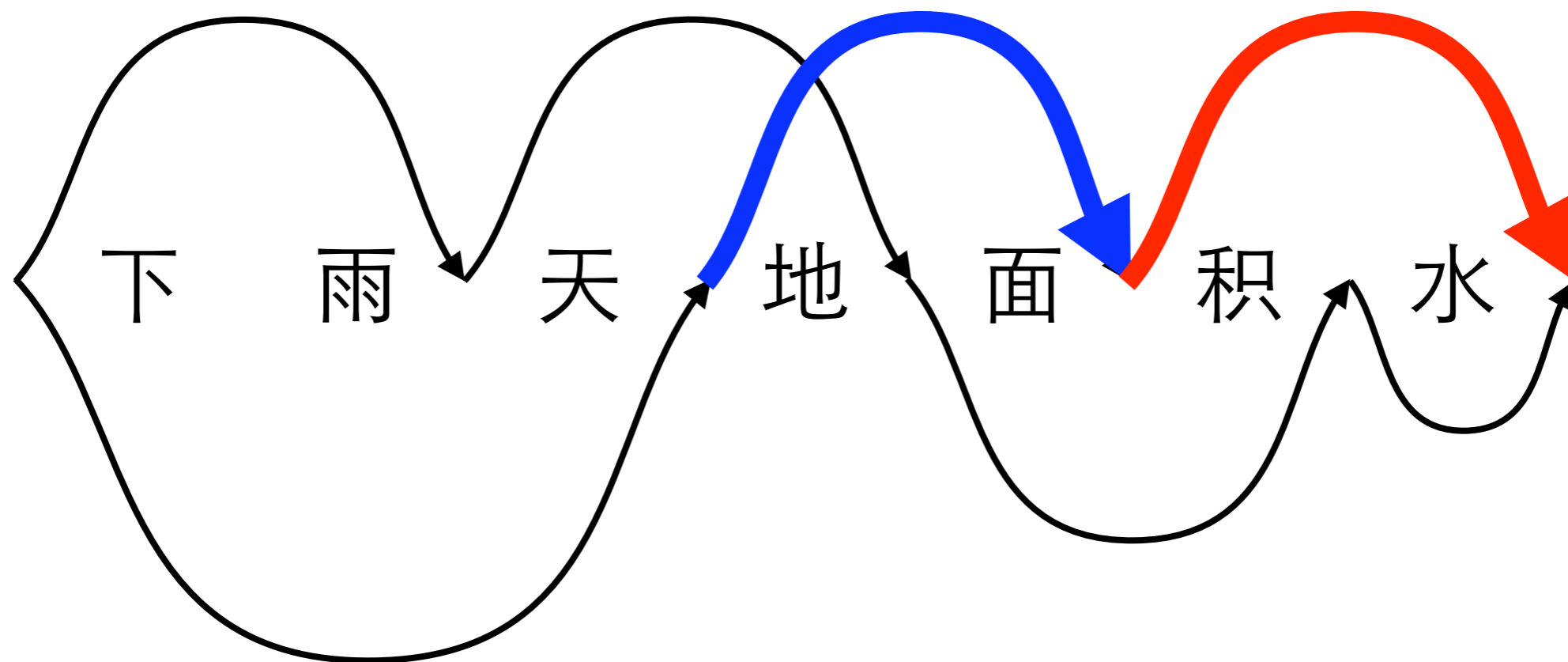
(积水, is wet)



the ground is wet

(Dyer et al., 2008)

Phrase-based Translation on Lattice

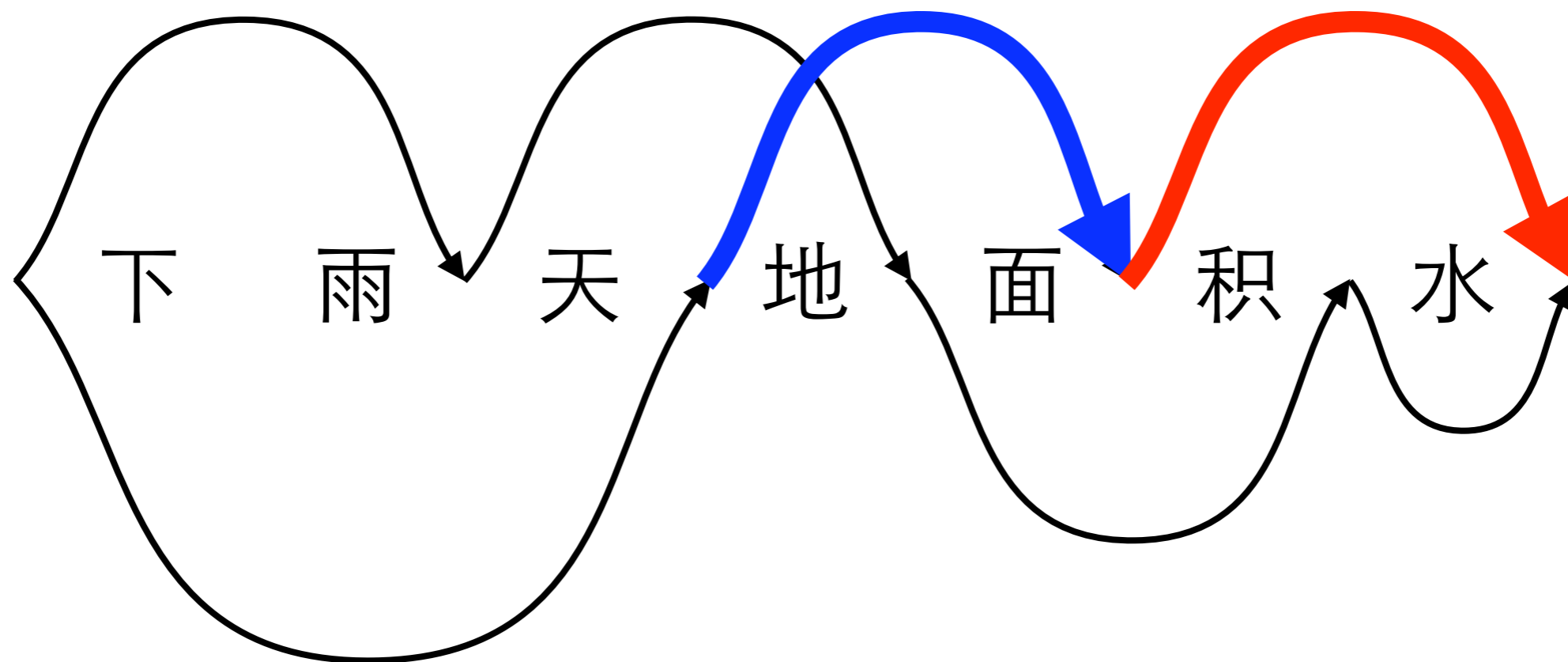


the ground is wet

(Dyer et al., 2008)

Phrase-based Translation on Lattice

(下雨天, in rainy days)

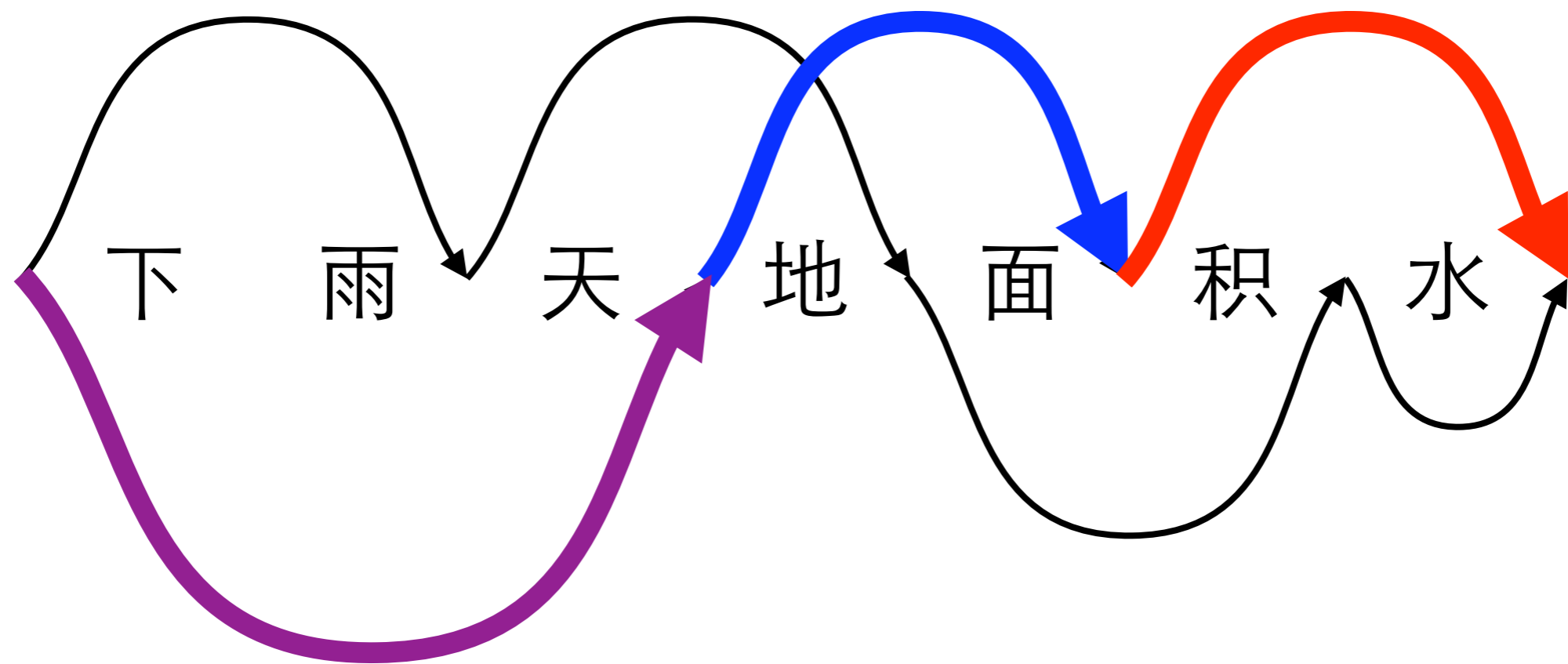


the ground is wet

(Dyer et al., 2008)

Phrase-based Translation on Lattice

(下雨天, in rainy days)

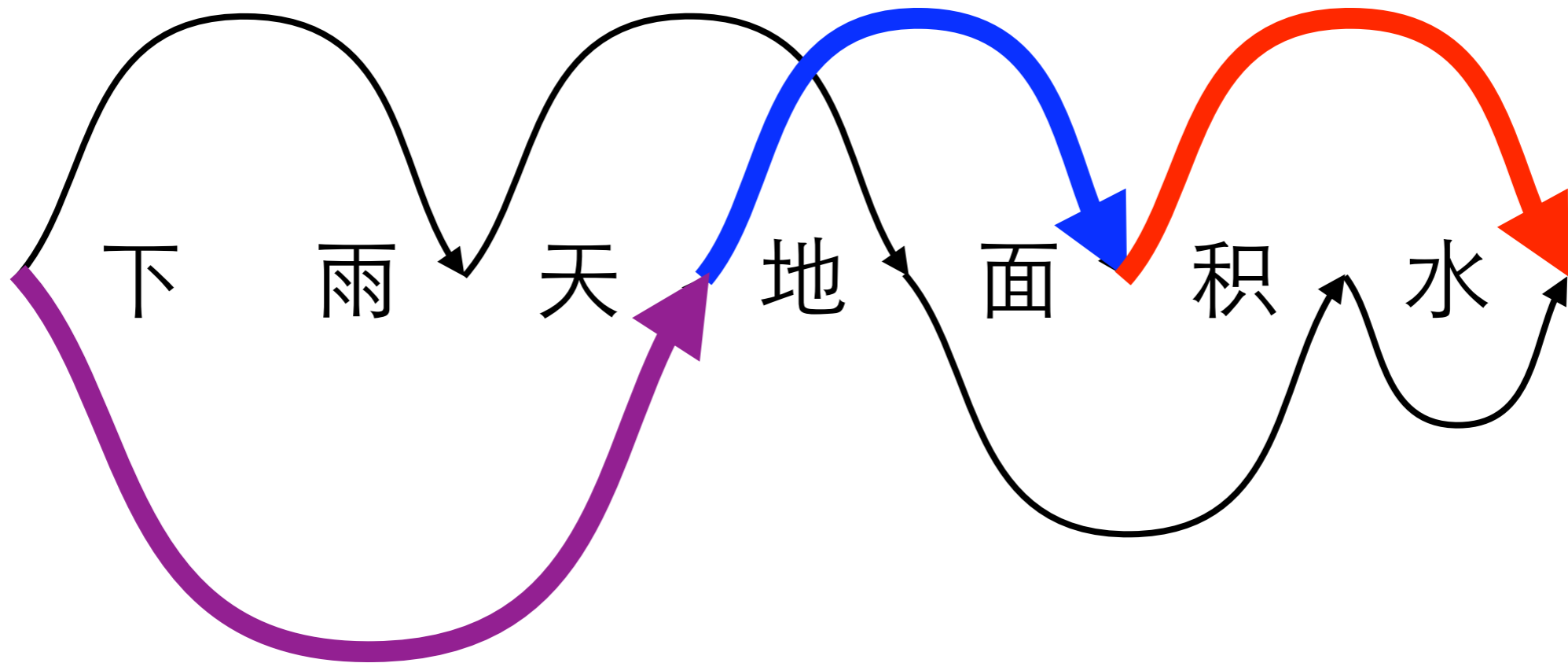


the ground is wet

(Dyer et al., 2008)

Phrase-based Translation on Lattice

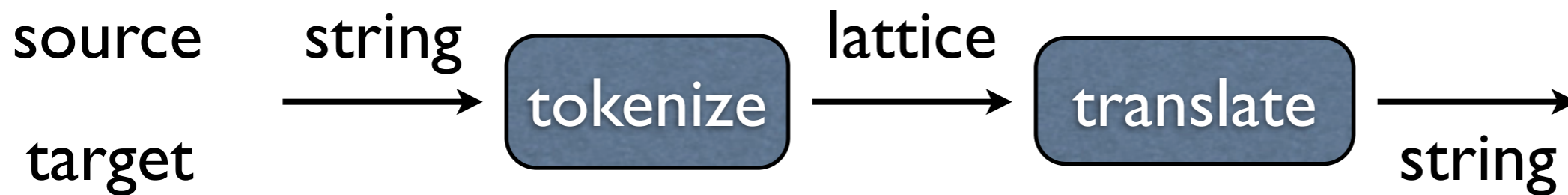
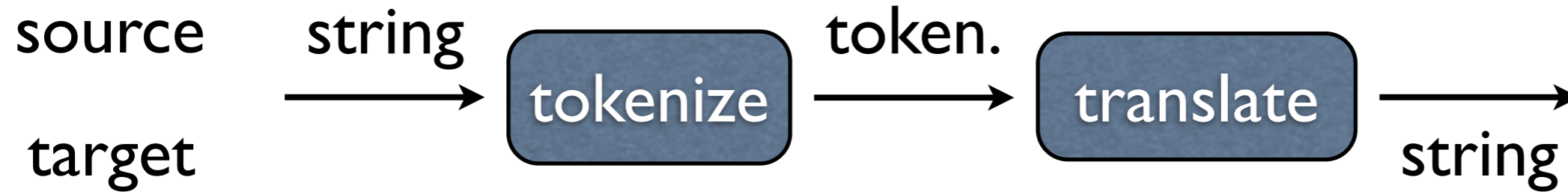
(下雨天, in rainy days)



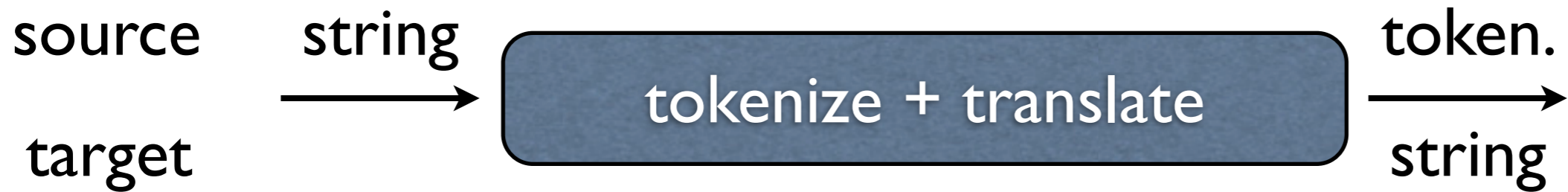
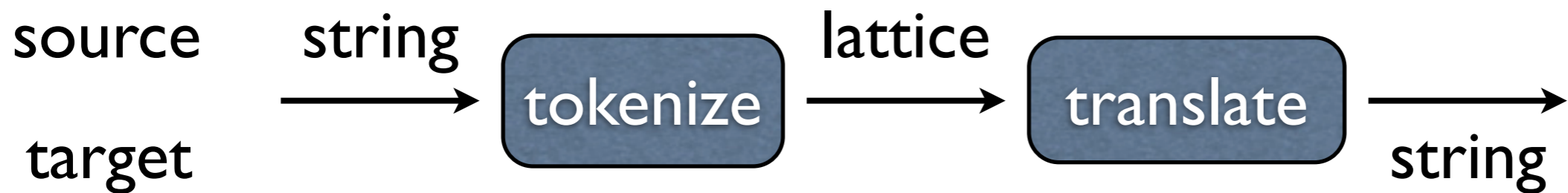
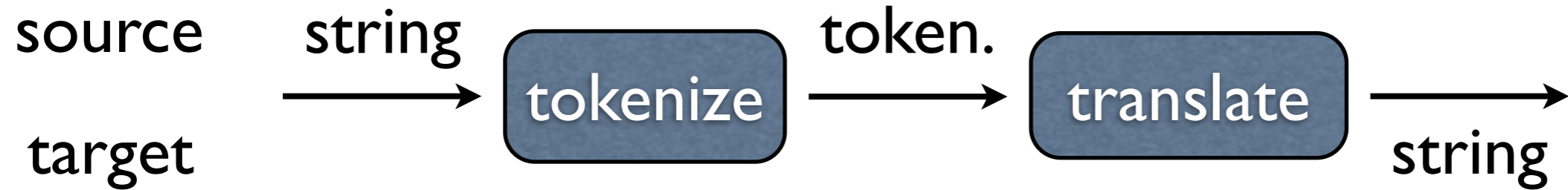
the ground is wet in rainy days

(Dyer et al., 2008)

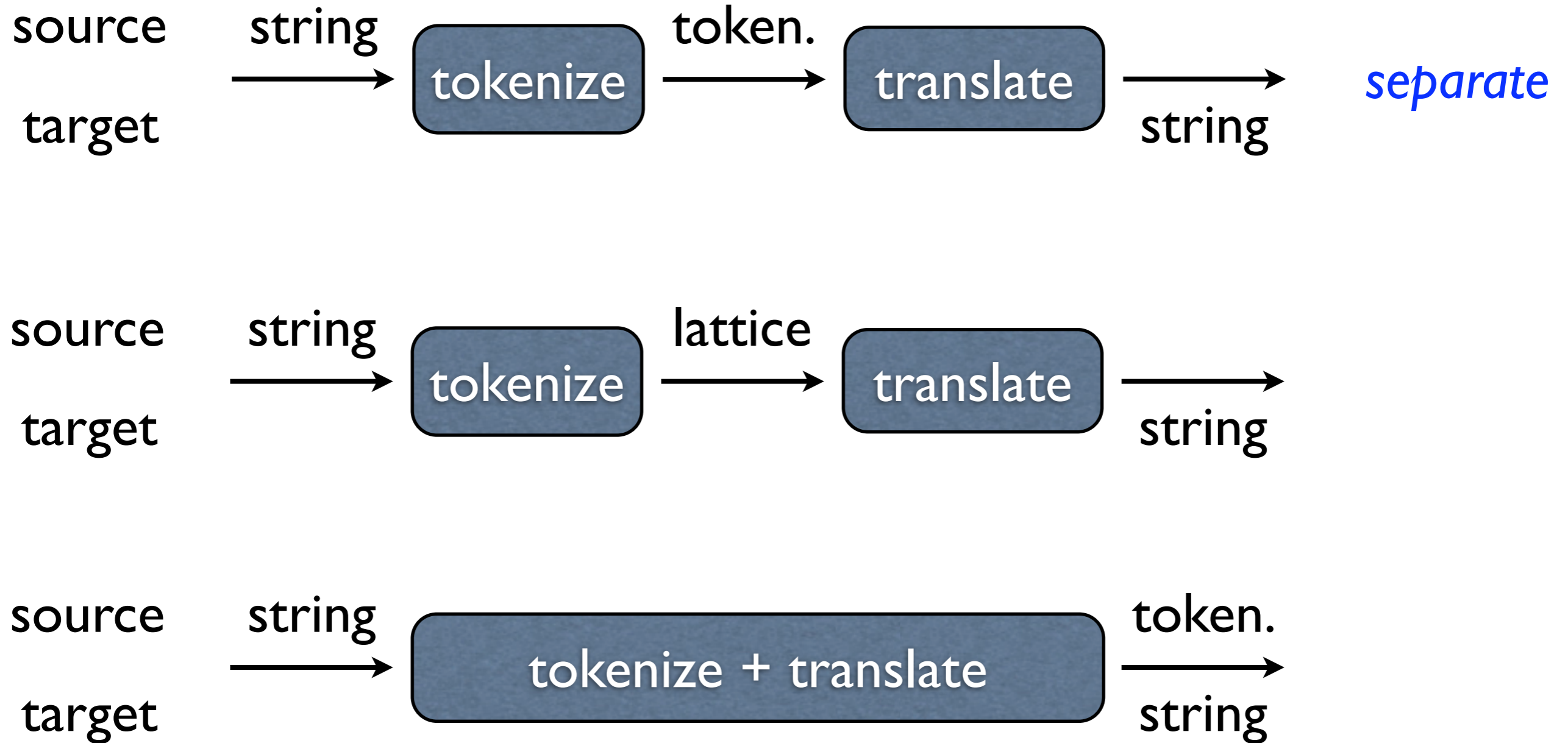
Tokenization and Translation: **Separate** Vs. **Joint**



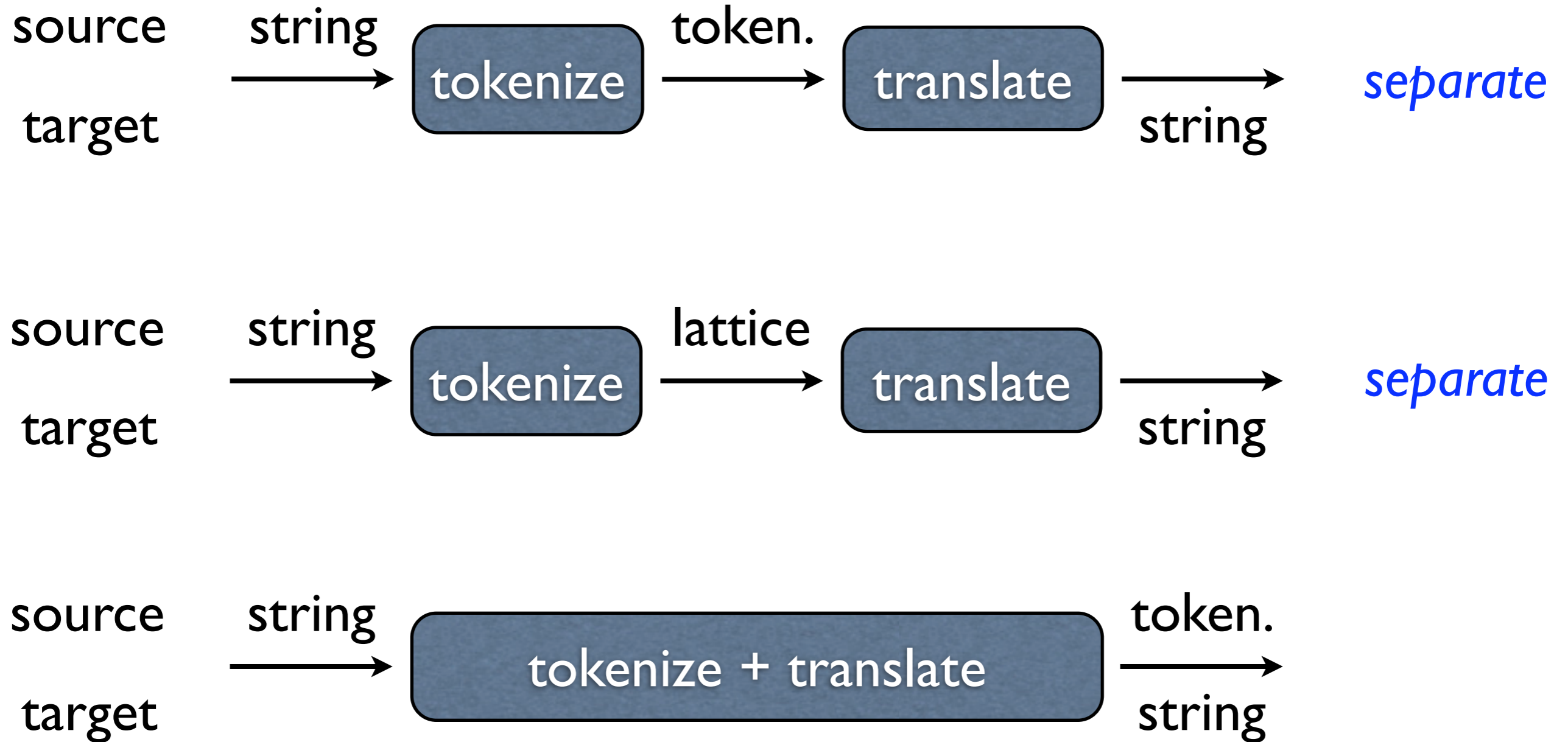
Tokenization and Translation: **Separate** Vs. **Joint**



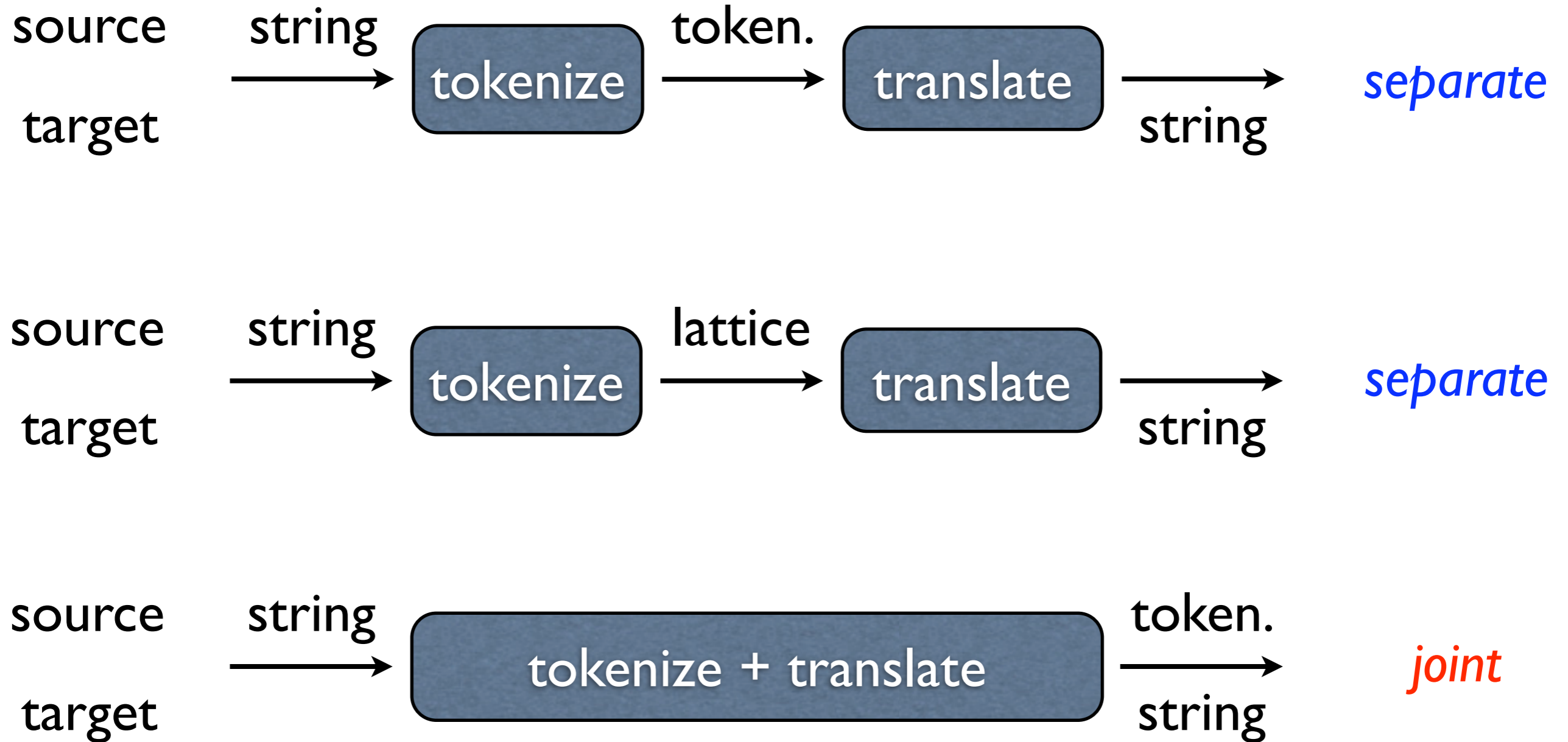
Tokenization and Translation: **Separate** Vs. **Joint**



Tokenization and Translation: **Separate** Vs. **Joint**



Tokenization and Translation: **Separate** Vs. **Joint**



Joint Tokenization and Translation

下 雨 天 地 面 积 水

(Xiao et al., 2010)

Joint Tokenization and Translation

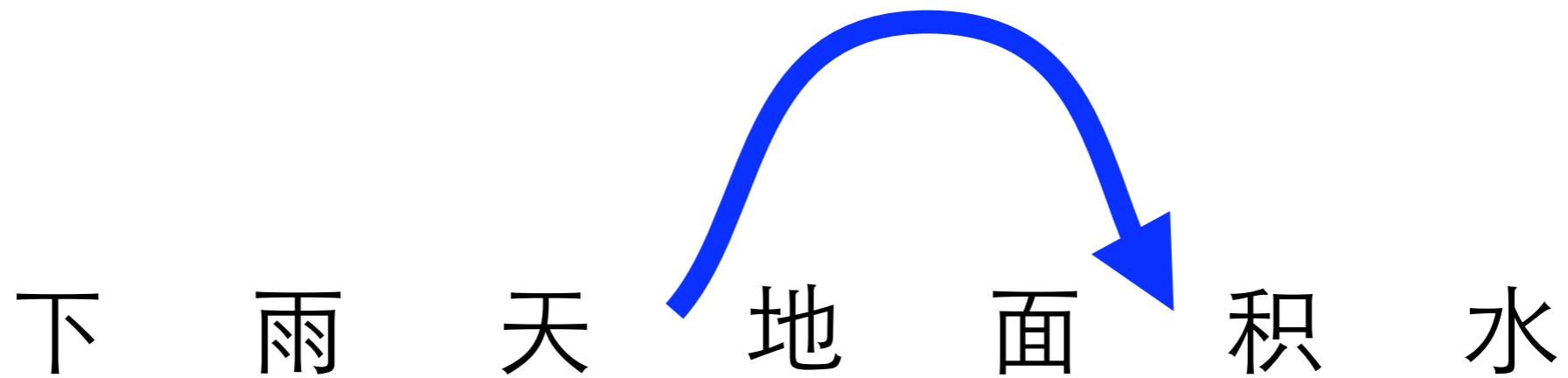
(地面, the ground)

下 雨 天 地 面 积 水

(Xiao et al., 2010)

Joint Tokenization and Translation

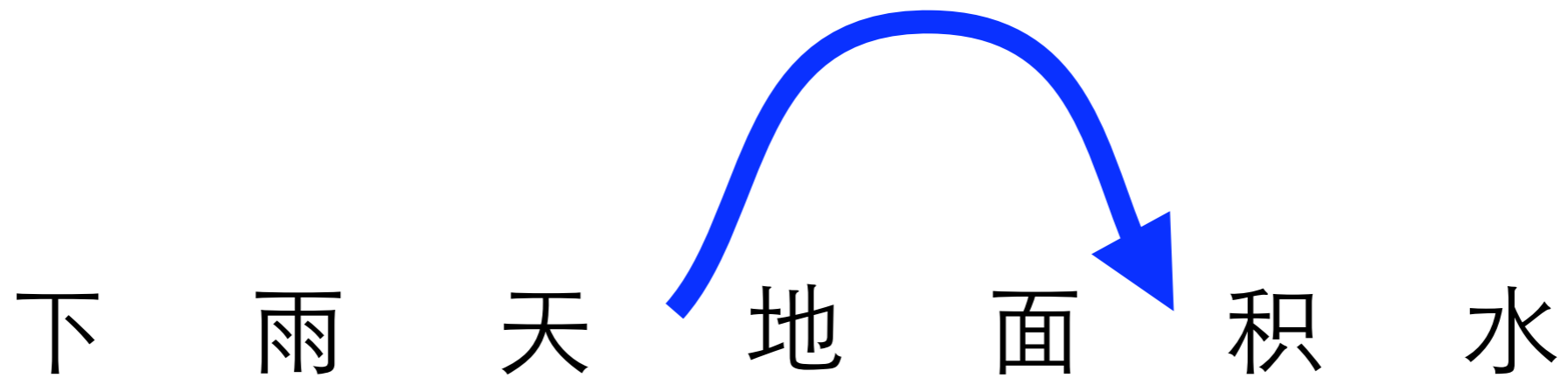
(地面, the ground)



(Xiao et al., 2010)

Joint Tokenization and Translation

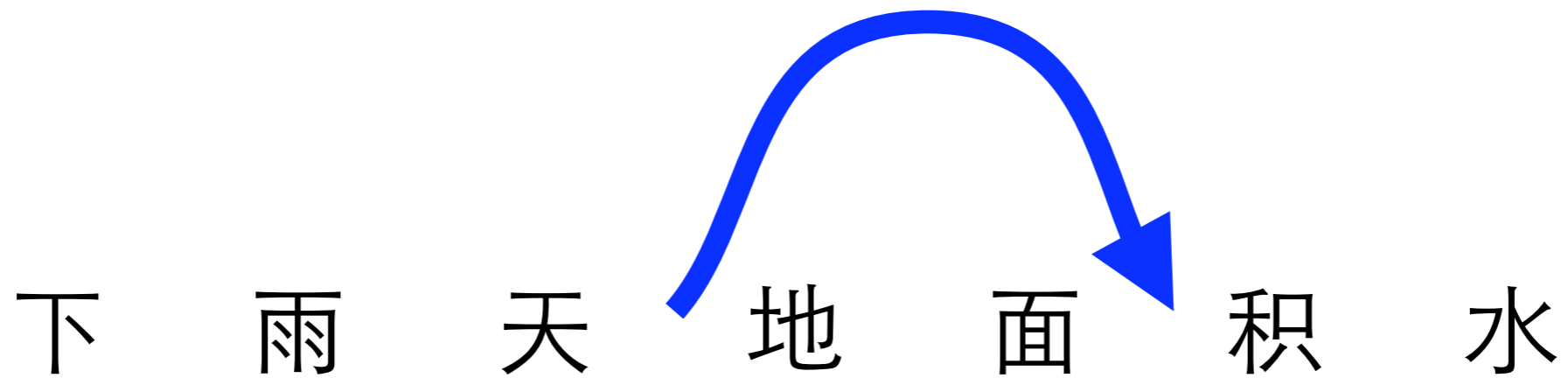
(地面, the ground)



the ground

(Xiao et al., 2010)

Joint Tokenization and Translation



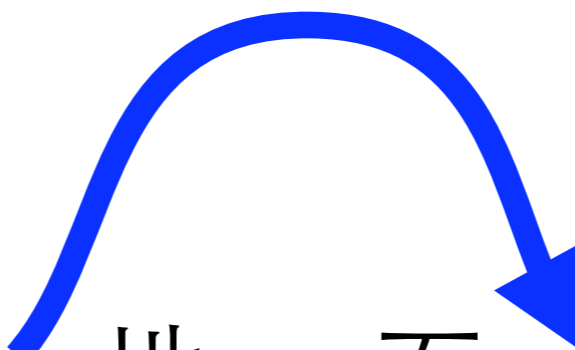
the ground

(Xiao et al., 2010)

Joint Tokenization and Translation

(积水, is wet)

下 雨 天 地 面 积 水

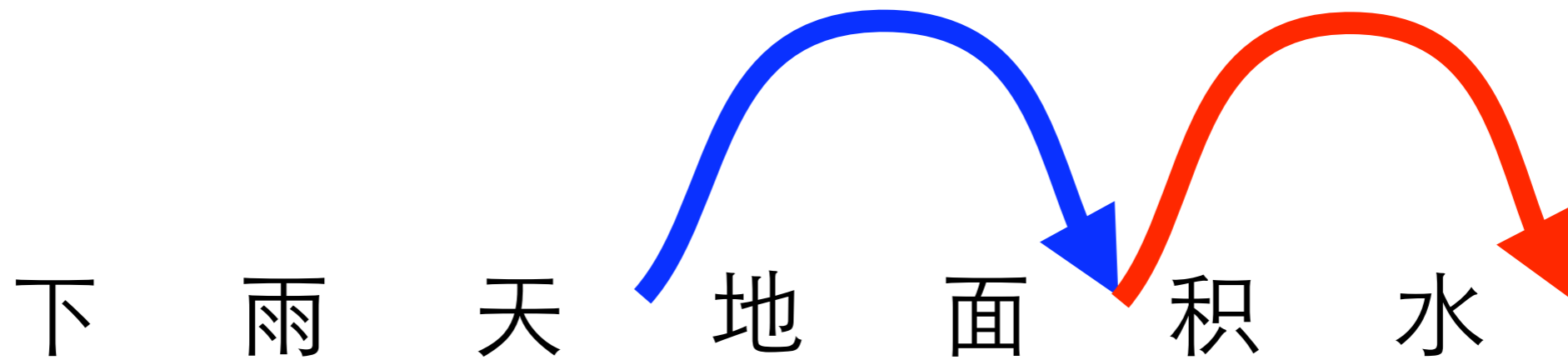


the ground

(Xiao et al., 2010)

Joint Tokenization and Translation

(积水, is wet)

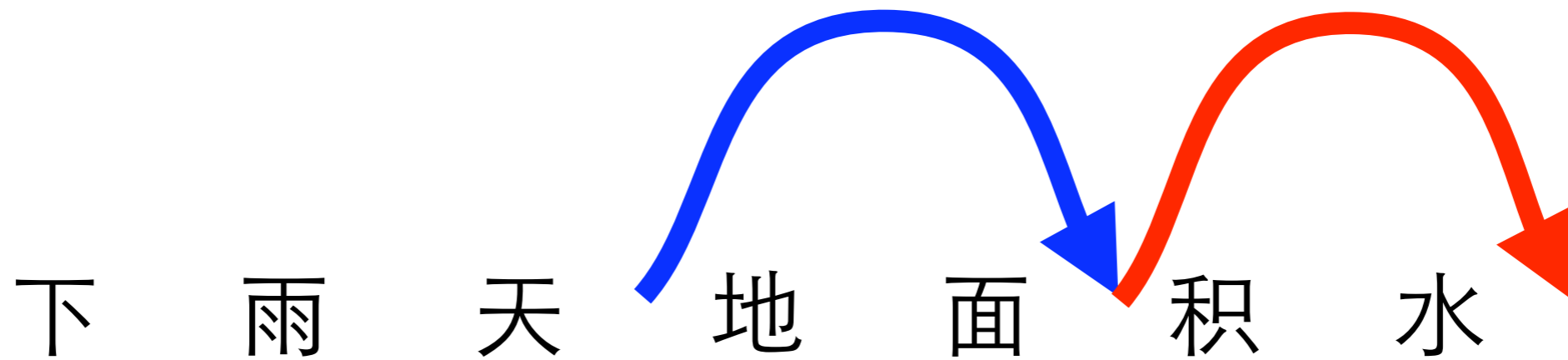


the ground

(Xiao et al., 2010)

Joint Tokenization and Translation

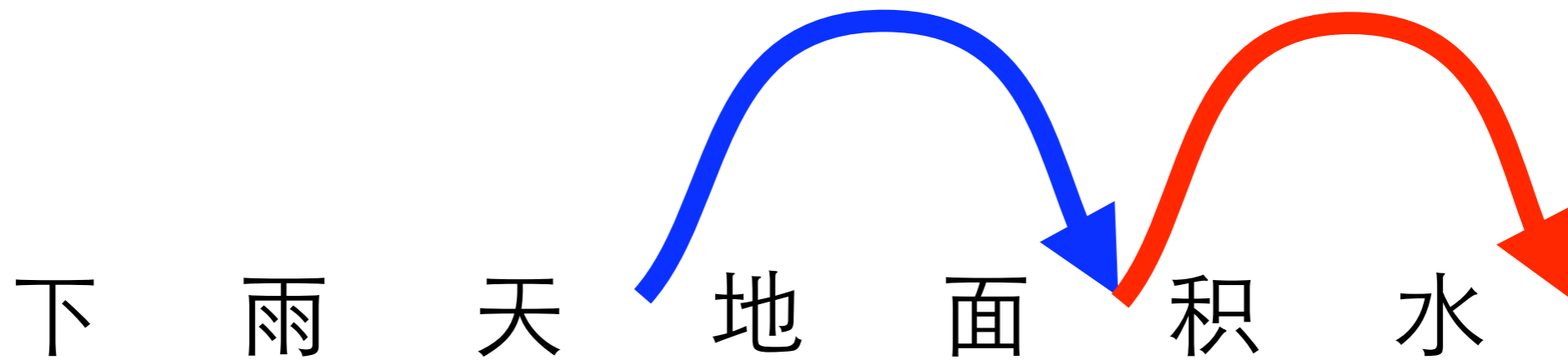
(积水, is wet)



the ground is wet

(Xiao et al., 2010)

Joint Tokenization and Translation

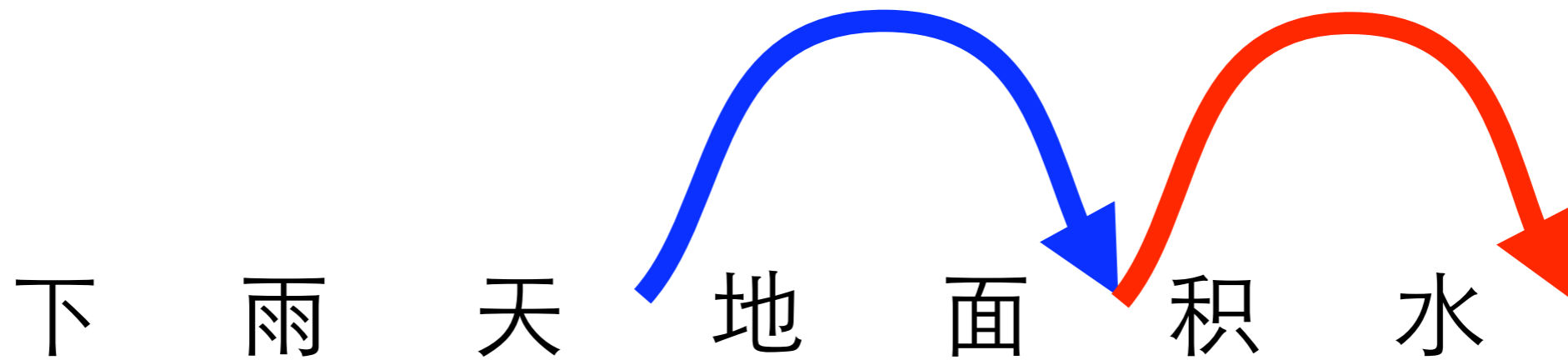


the ground is wet

(Xiao et al., 2010)

Joint Tokenization and Translation

(下雨天, in rainy days)

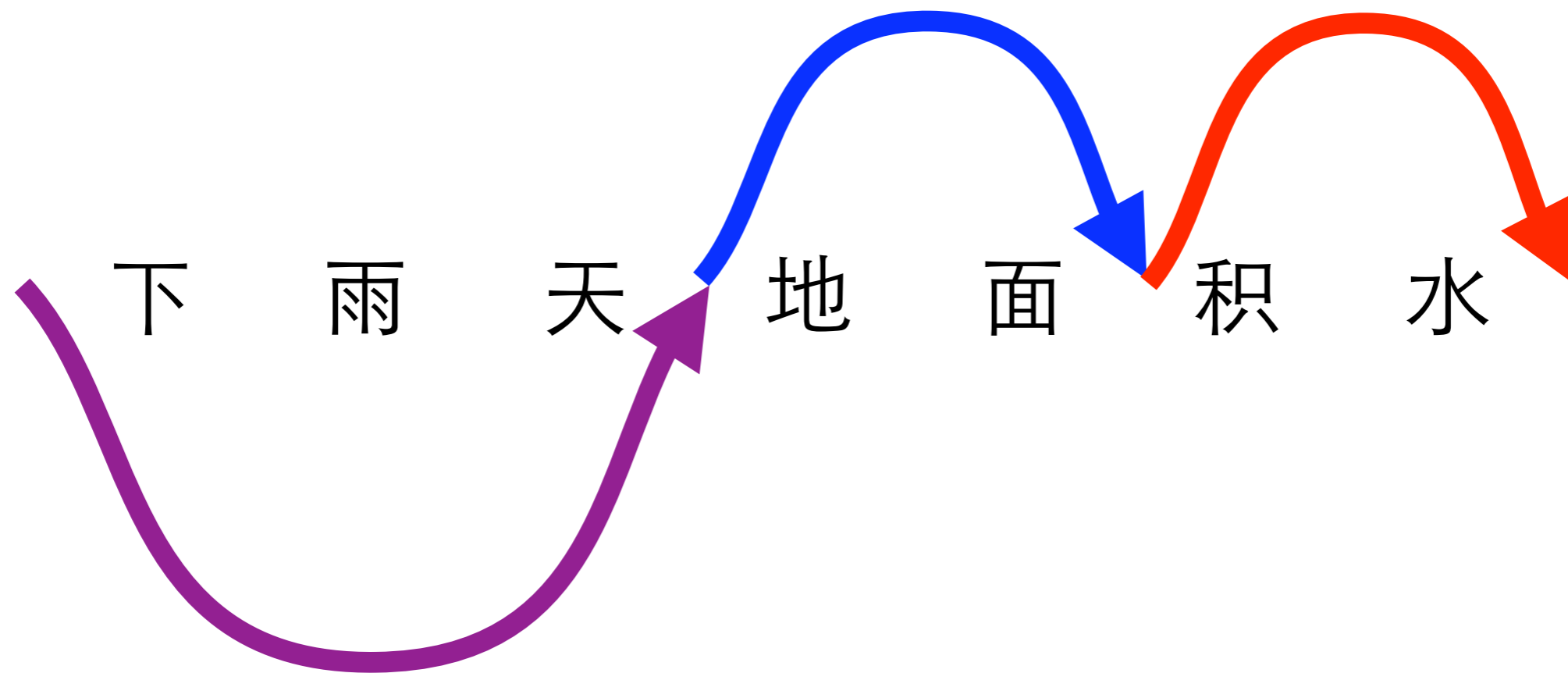


the ground is wet

(Xiao et al., 2010)

Joint Tokenization and Translation

(下雨天, in rainy days)

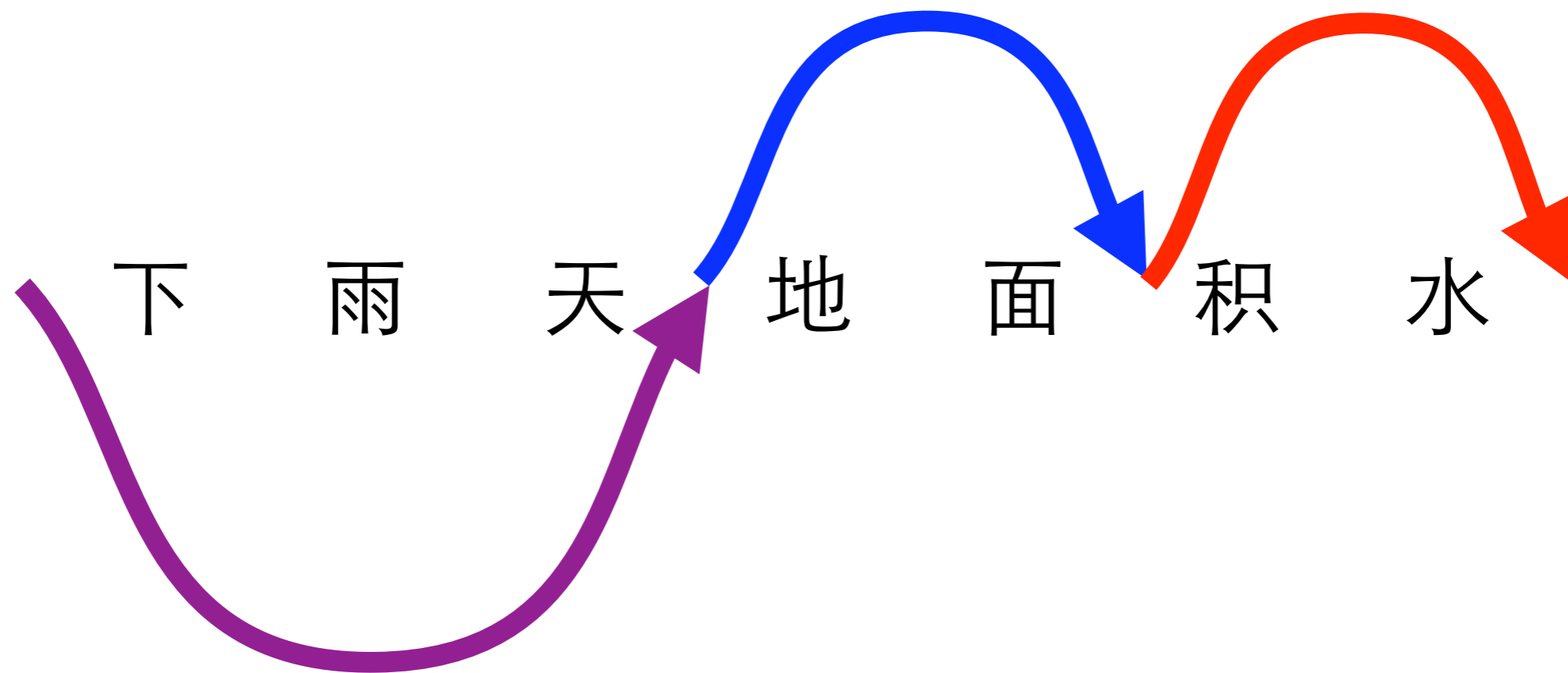


the ground is wet

(Xiao et al., 2010)

Joint Tokenization and Translation

(下雨天, in rainy days)

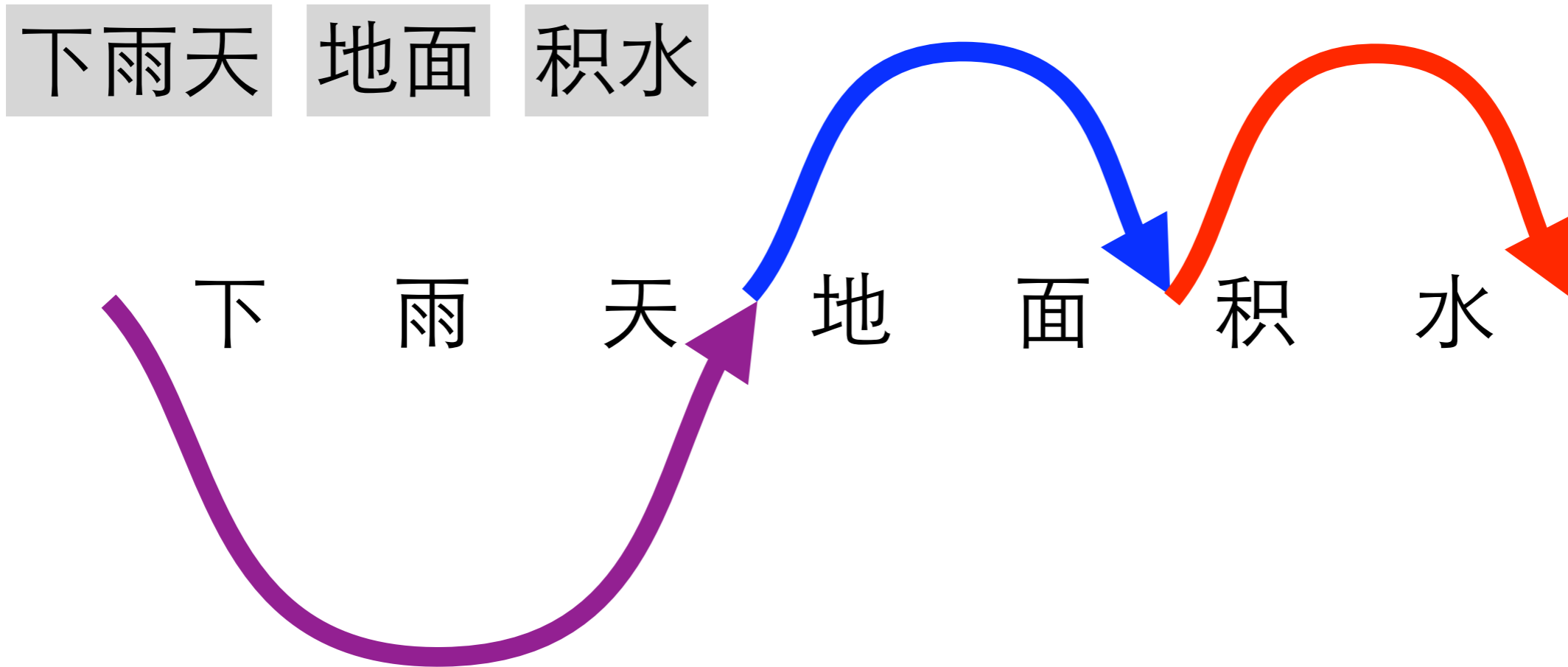


the ground is wet in rainy days

(Xiao et al., 2010)

Joint Tokenization and Translation

(下雨天, in rainy days)



the ground is wet in rainy days

(Xiao et al., 2010)

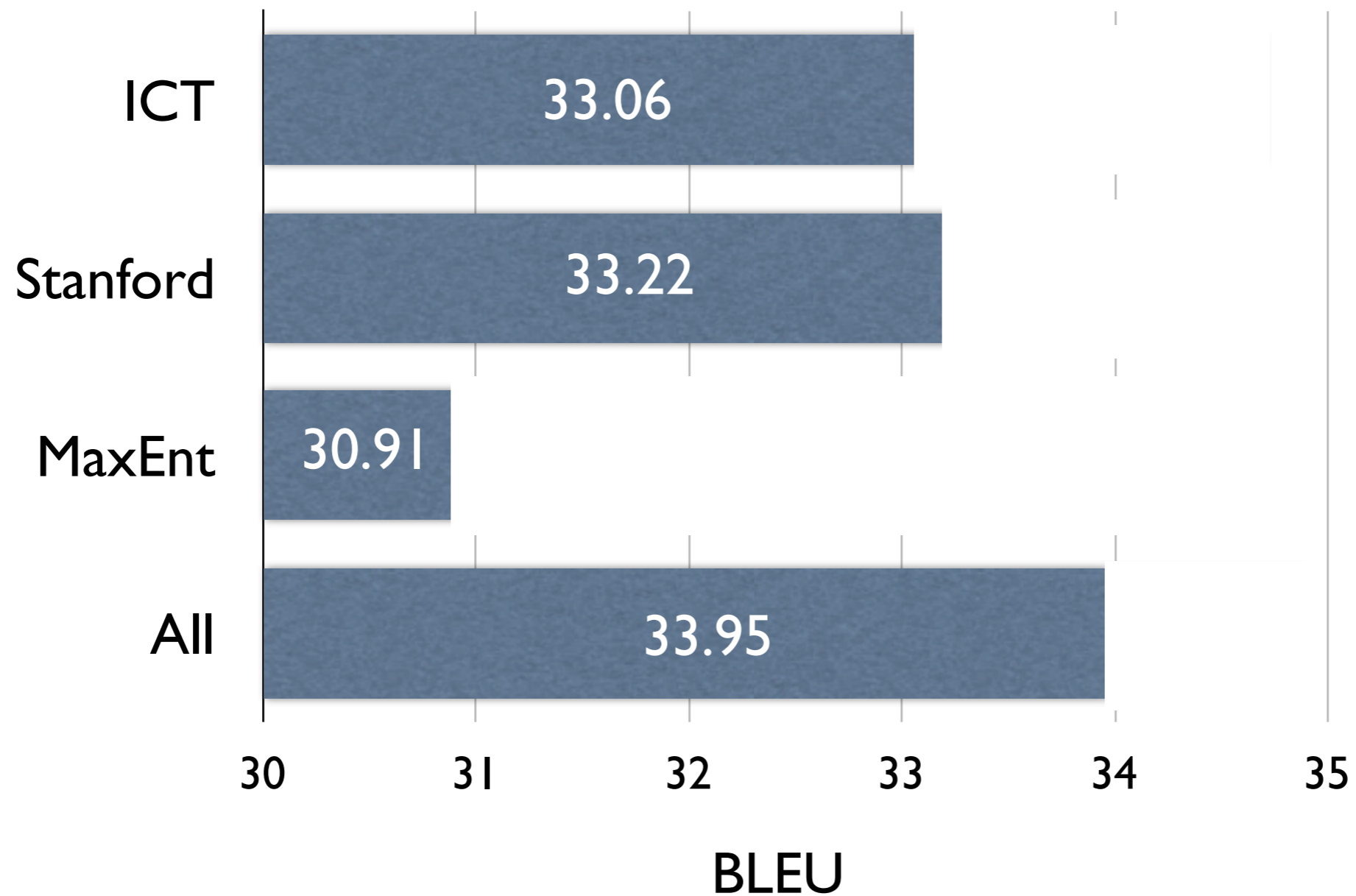
Model and Training

$$score(\mathbf{f}, \mathbf{t}, \mathbf{e}) = \sum_k \lambda_k h_k(\mathbf{f}, \mathbf{t}, \mathbf{e})$$

- We use a linear model to combine **tokenization** and **translation** models as features
- The minimum-error-rate training algorithm is used for training feature weights

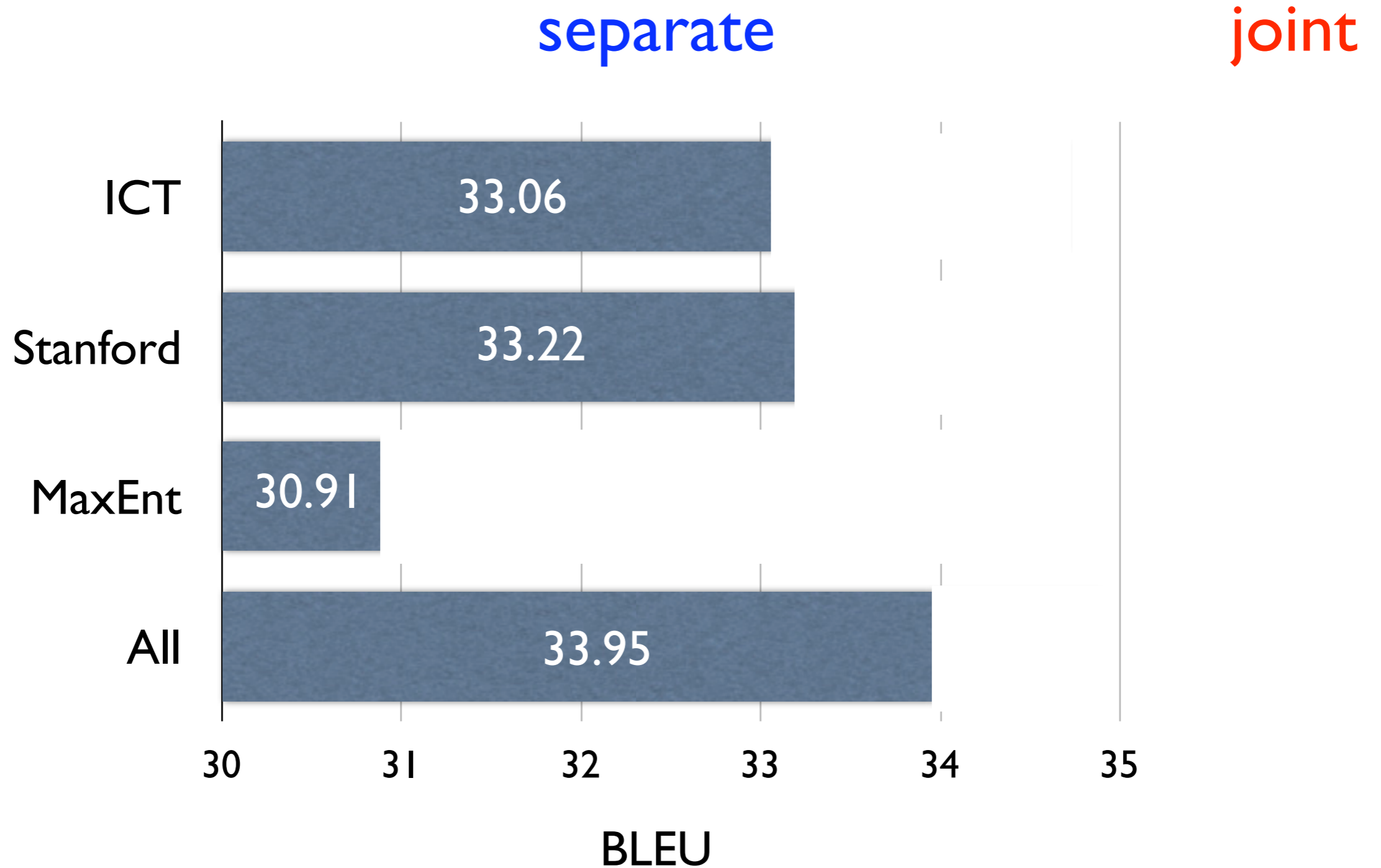
Translation Evaluation

separate



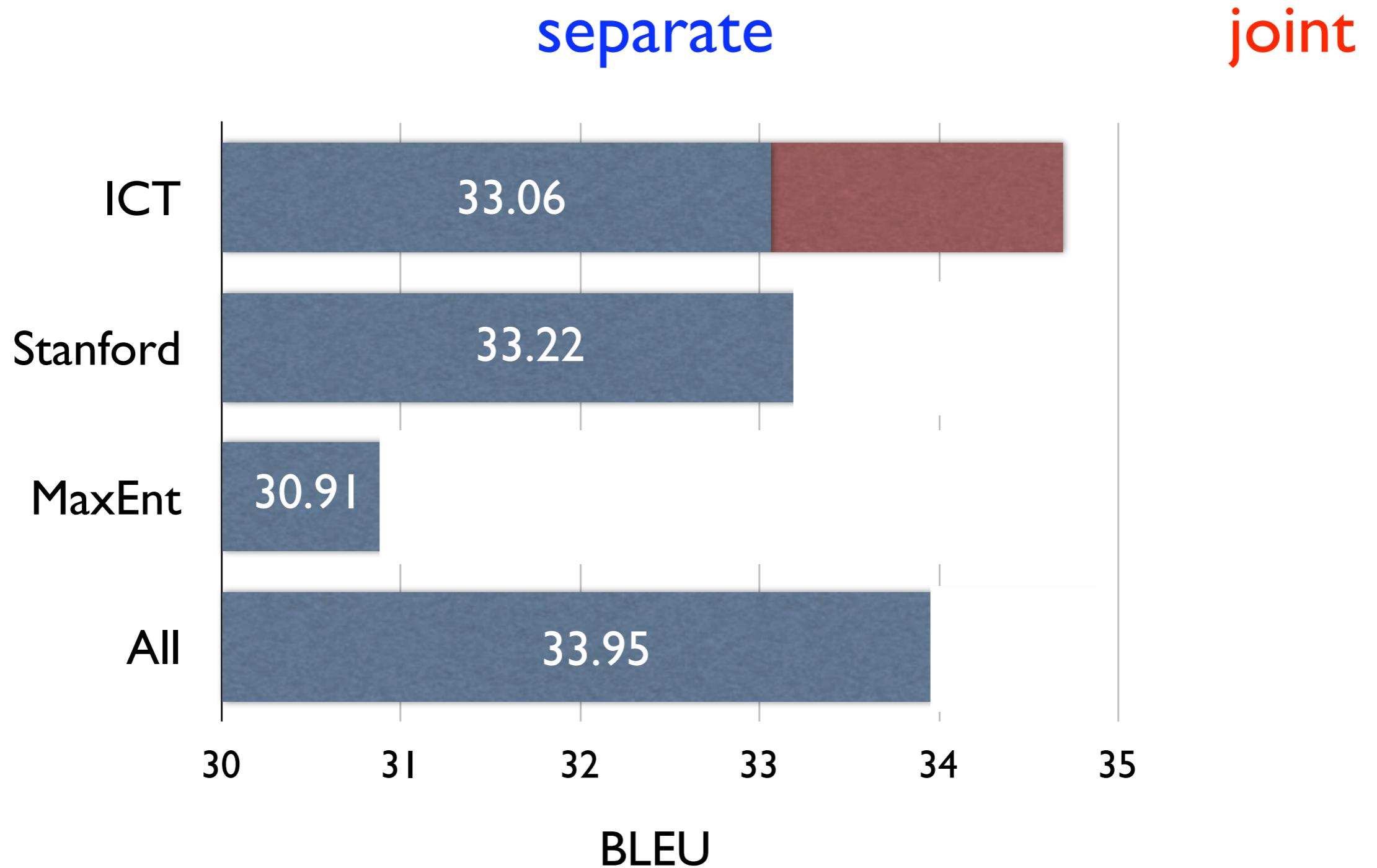
(Xiao et al., 2010)

Translation Evaluation



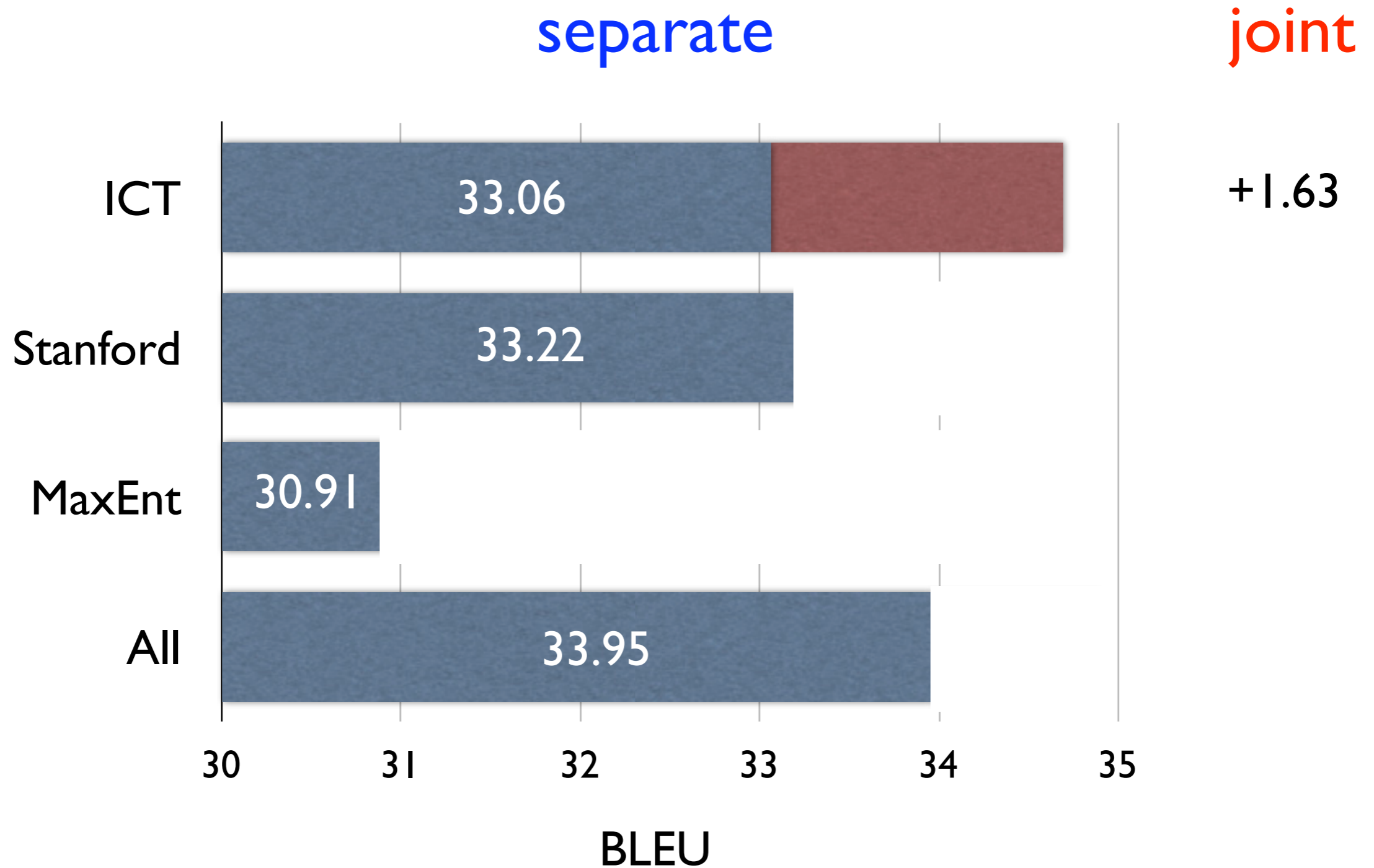
(Xiao et al., 2010)

Translation Evaluation



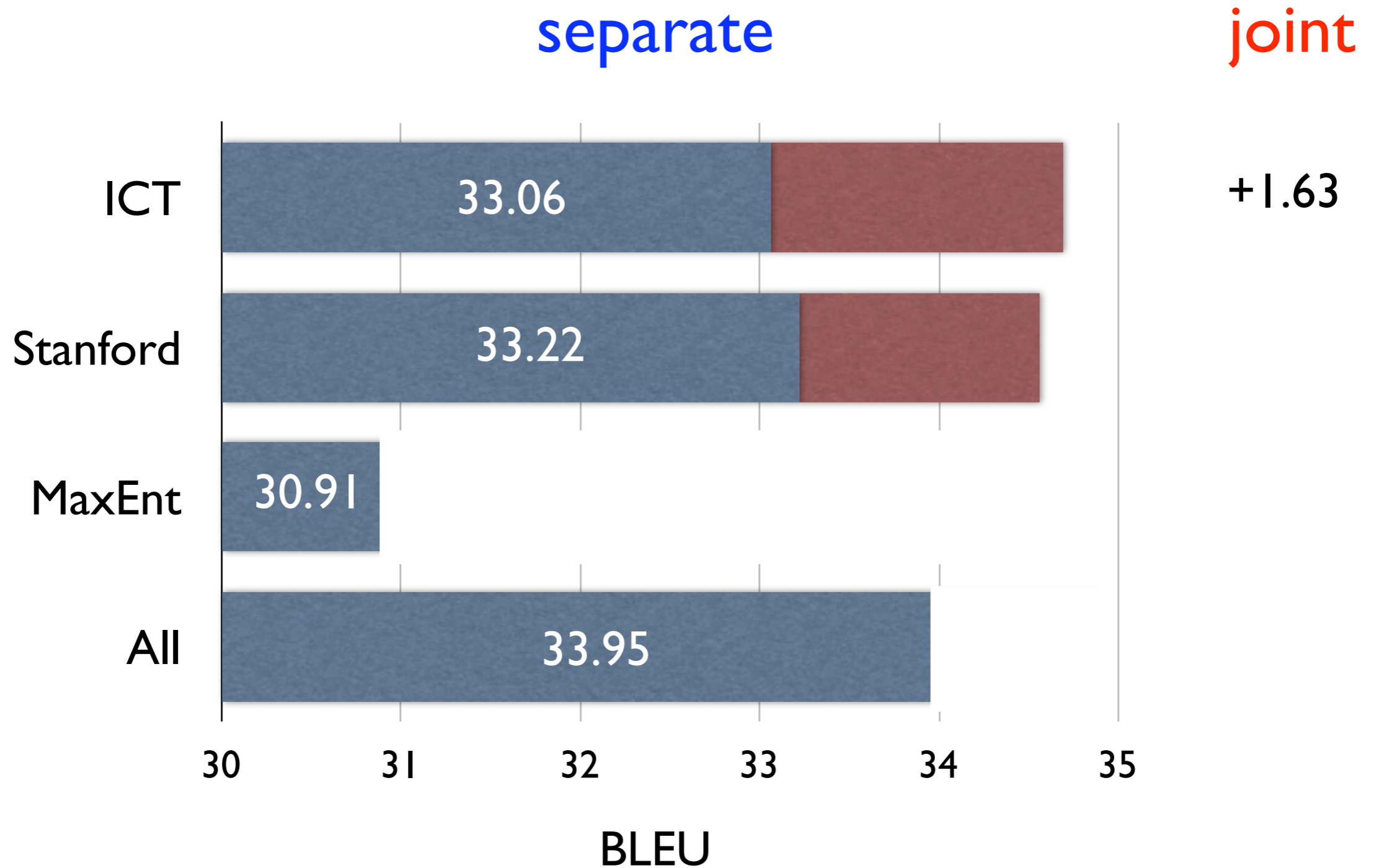
(Xiao et al., 2010)

Translation Evaluation



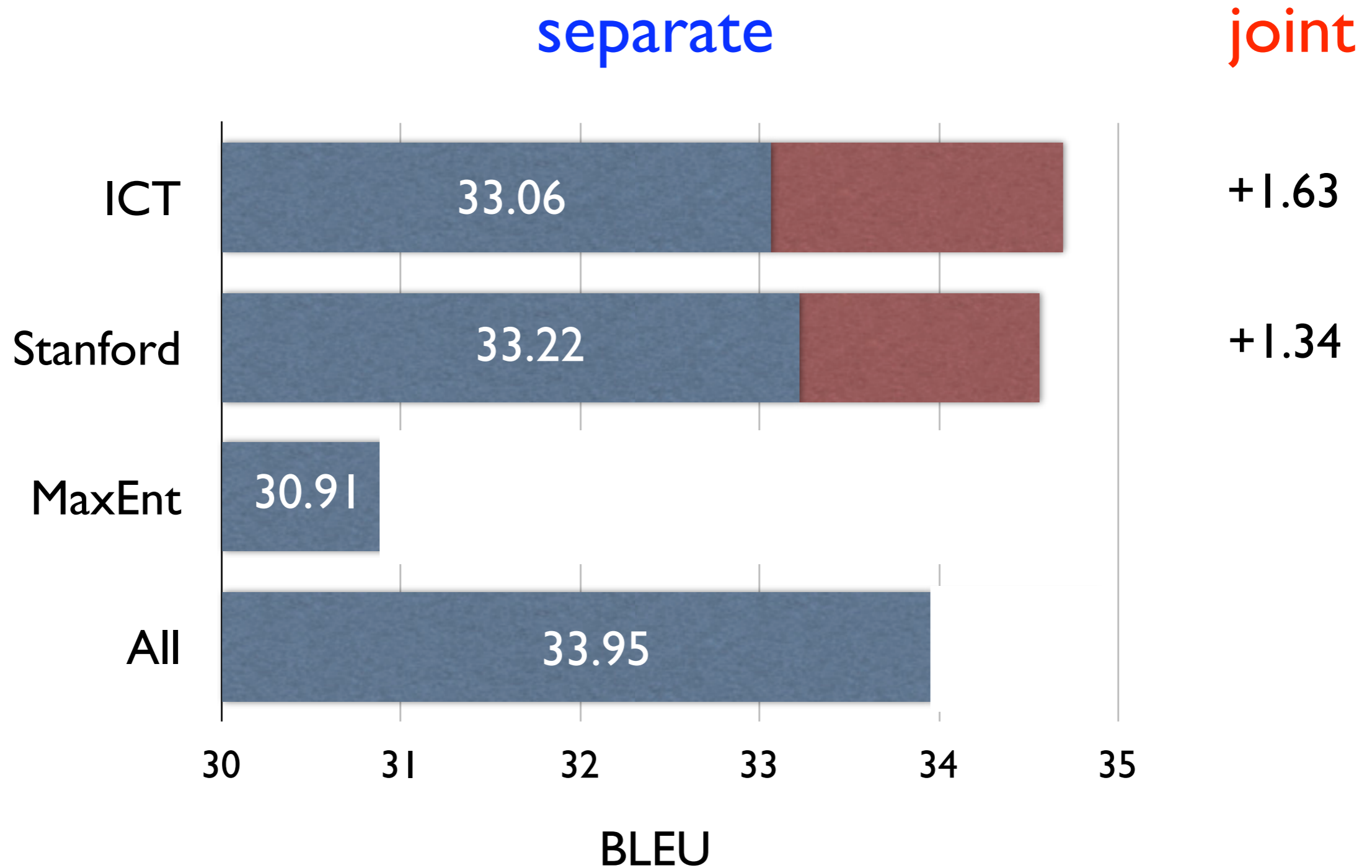
(Xiao et al., 2010)

Translation Evaluation



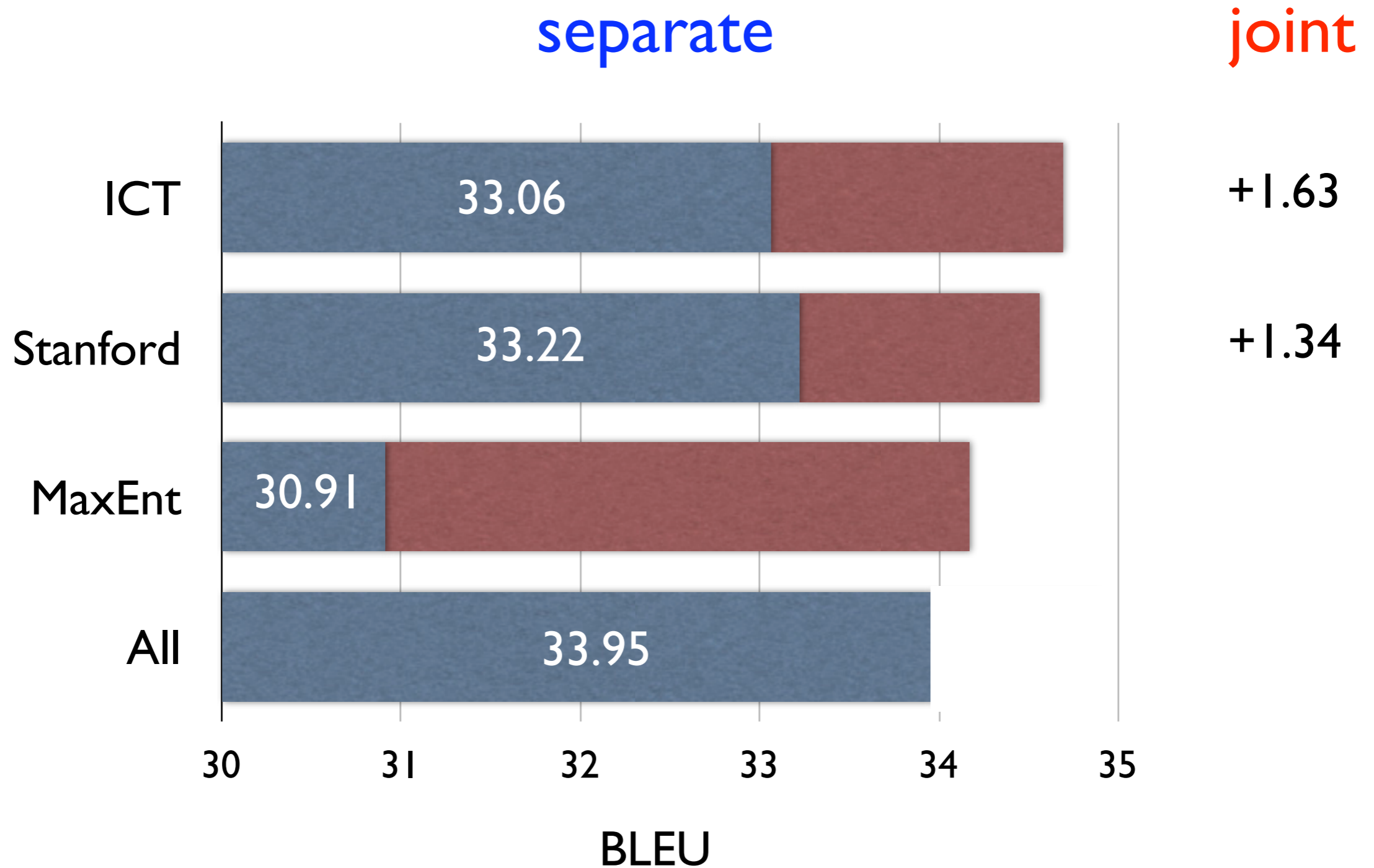
(Xiao et al., 2010)

Translation Evaluation



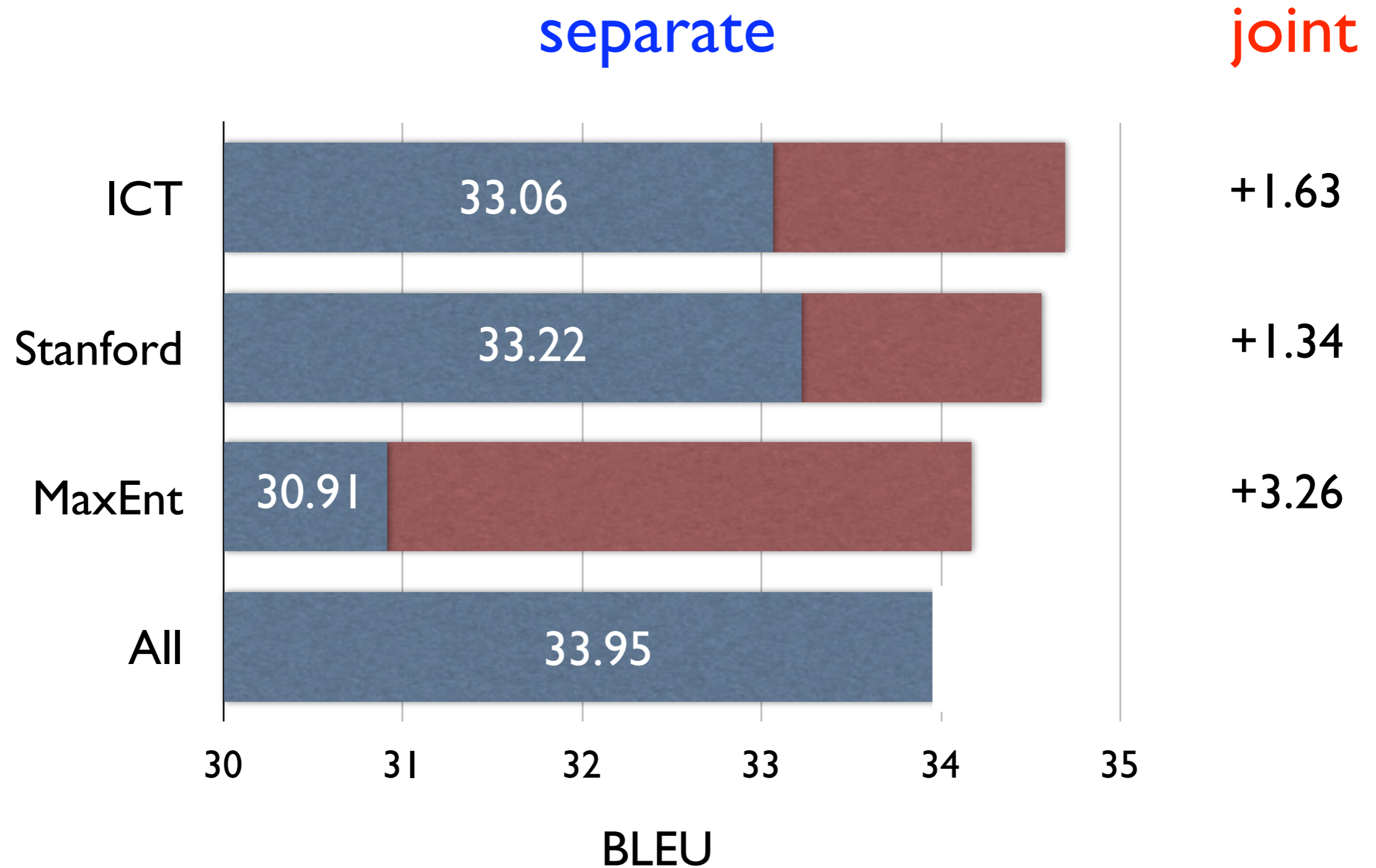
(Xiao et al., 2010)

Translation Evaluation



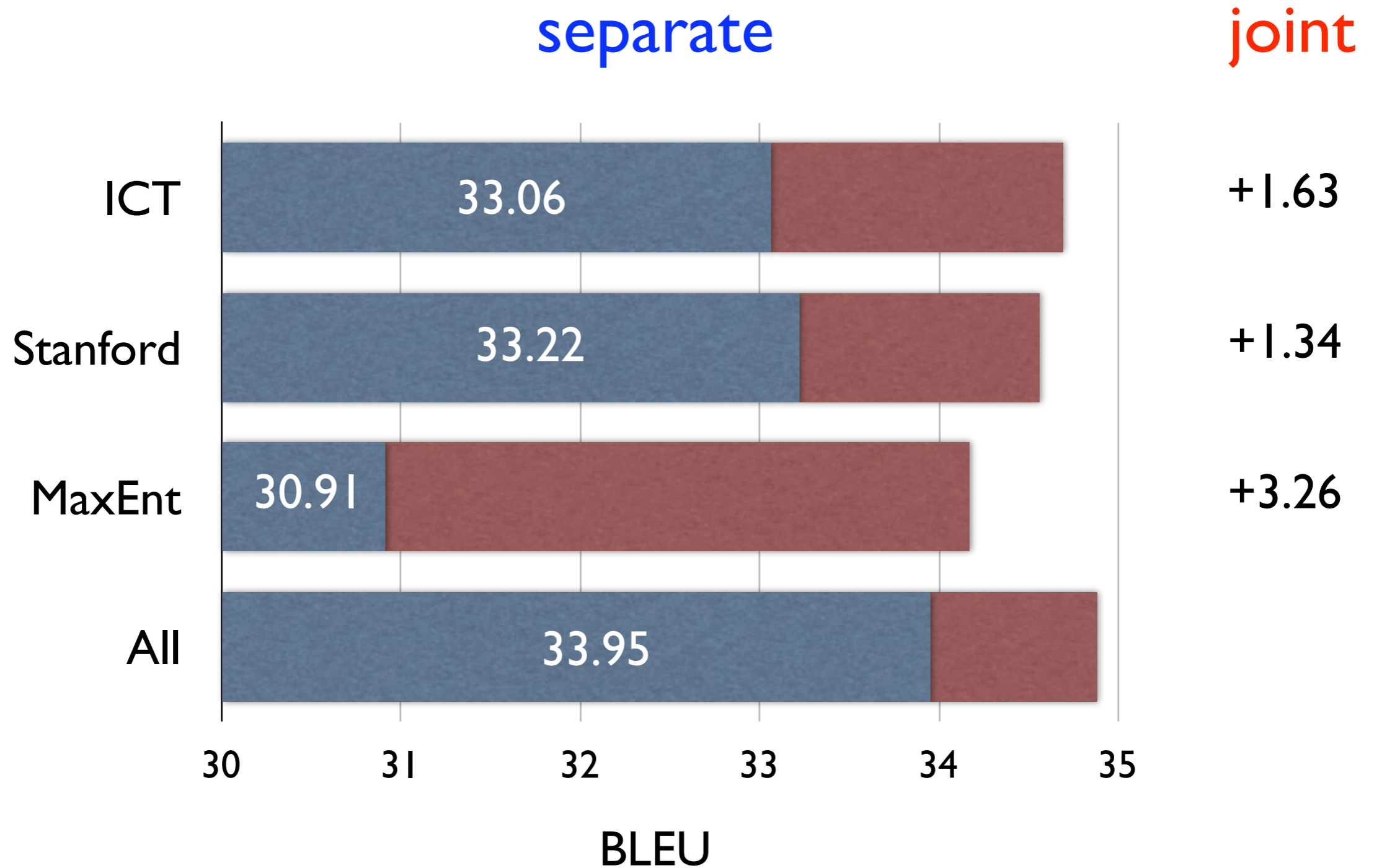
(Xiao et al., 2010)

Translation Evaluation



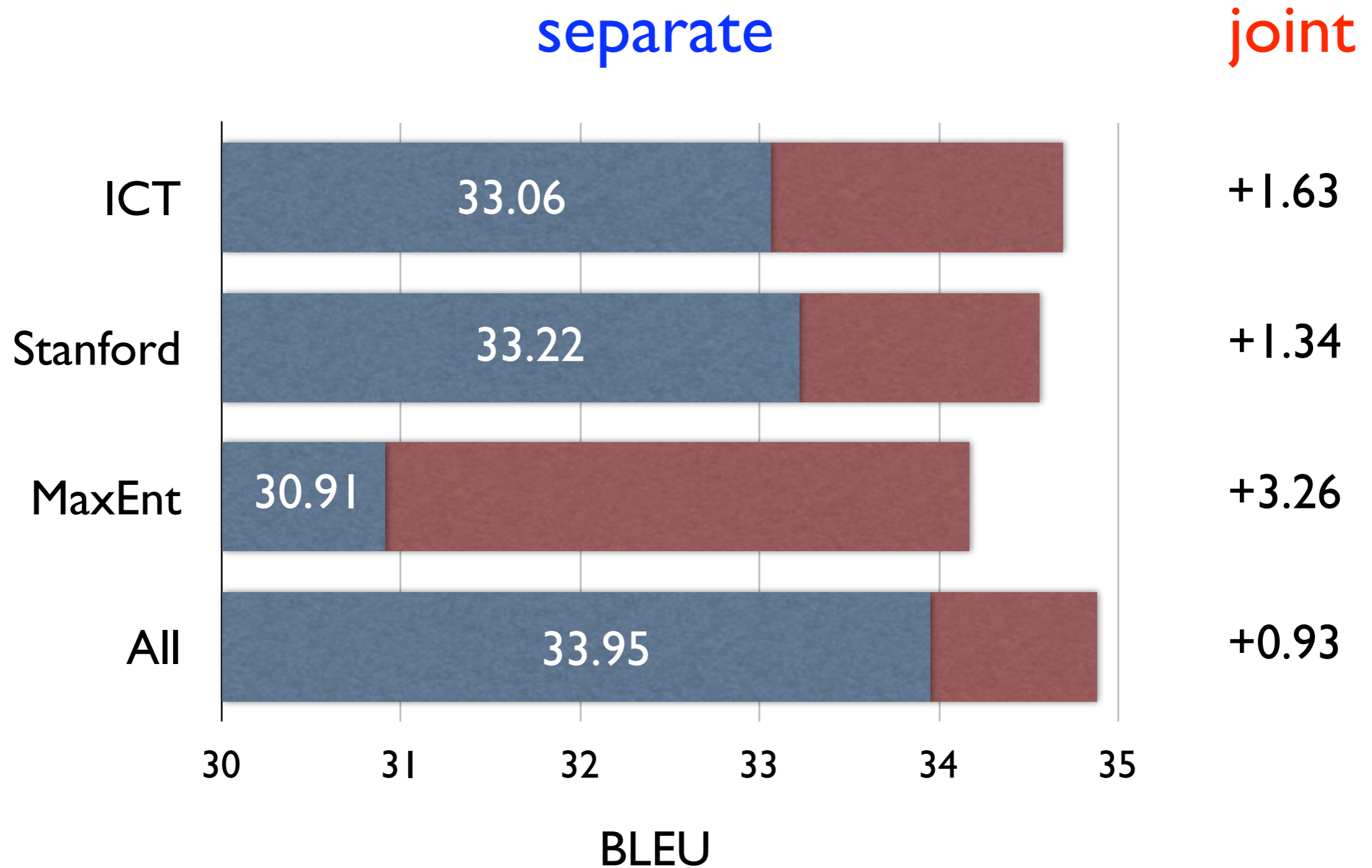
(Xiao et al., 2010)

Translation Evaluation



(Xiao et al., 2010)

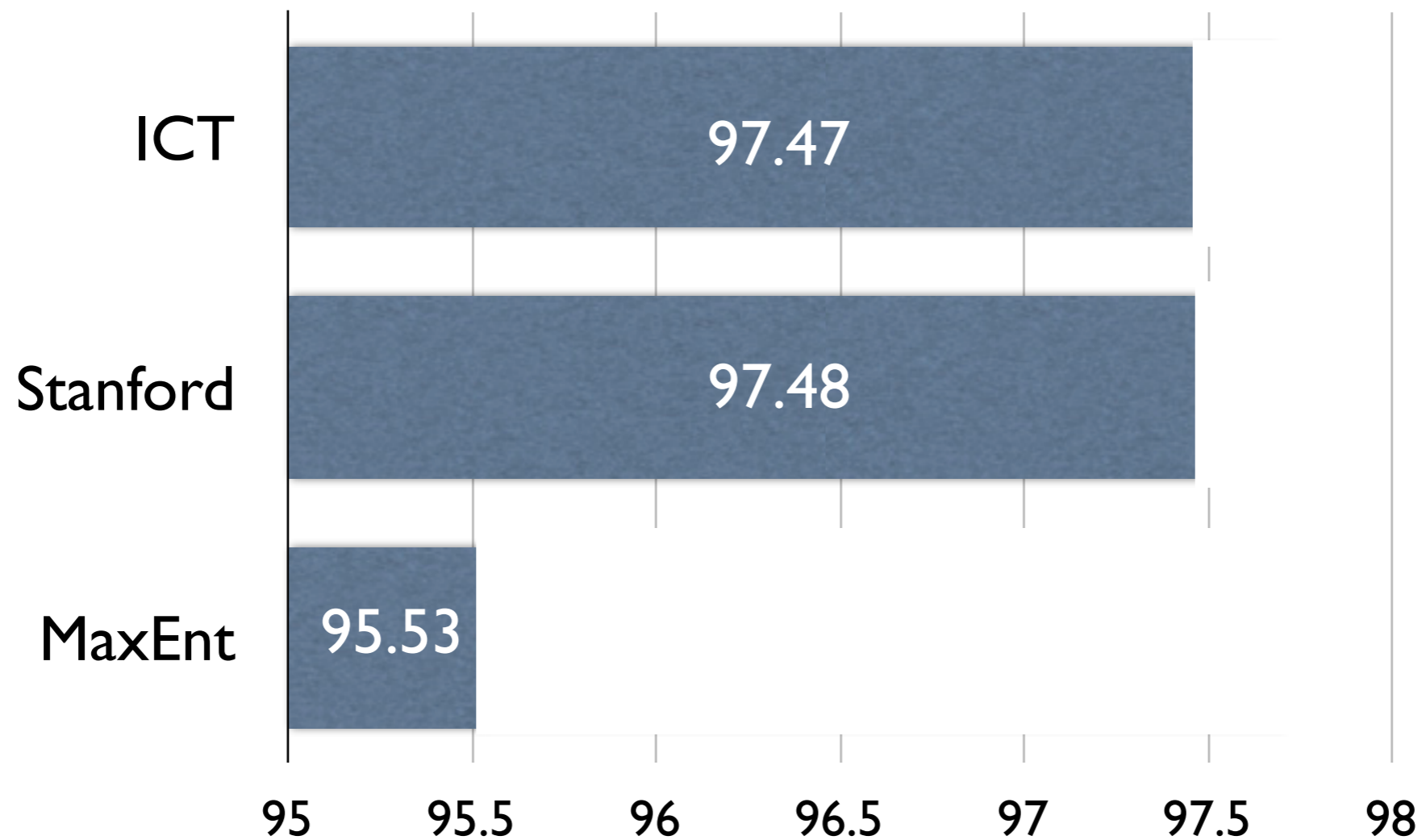
Translation Evaluation



(Xiao et al., 2010)

Tokenization Evaluation

separate



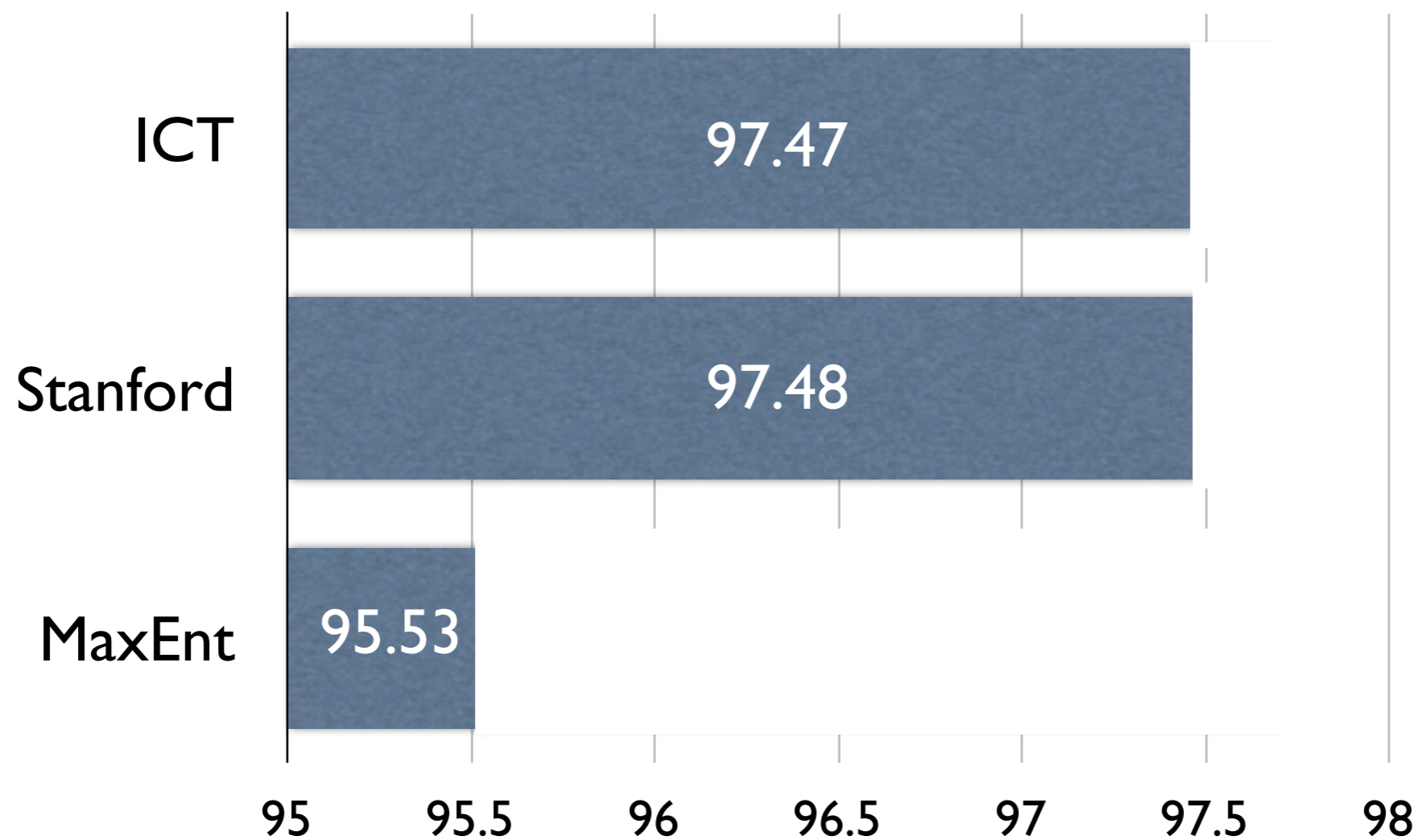
F1

(Xiao et al., 2010)

Tokenization Evaluation

separate

joint



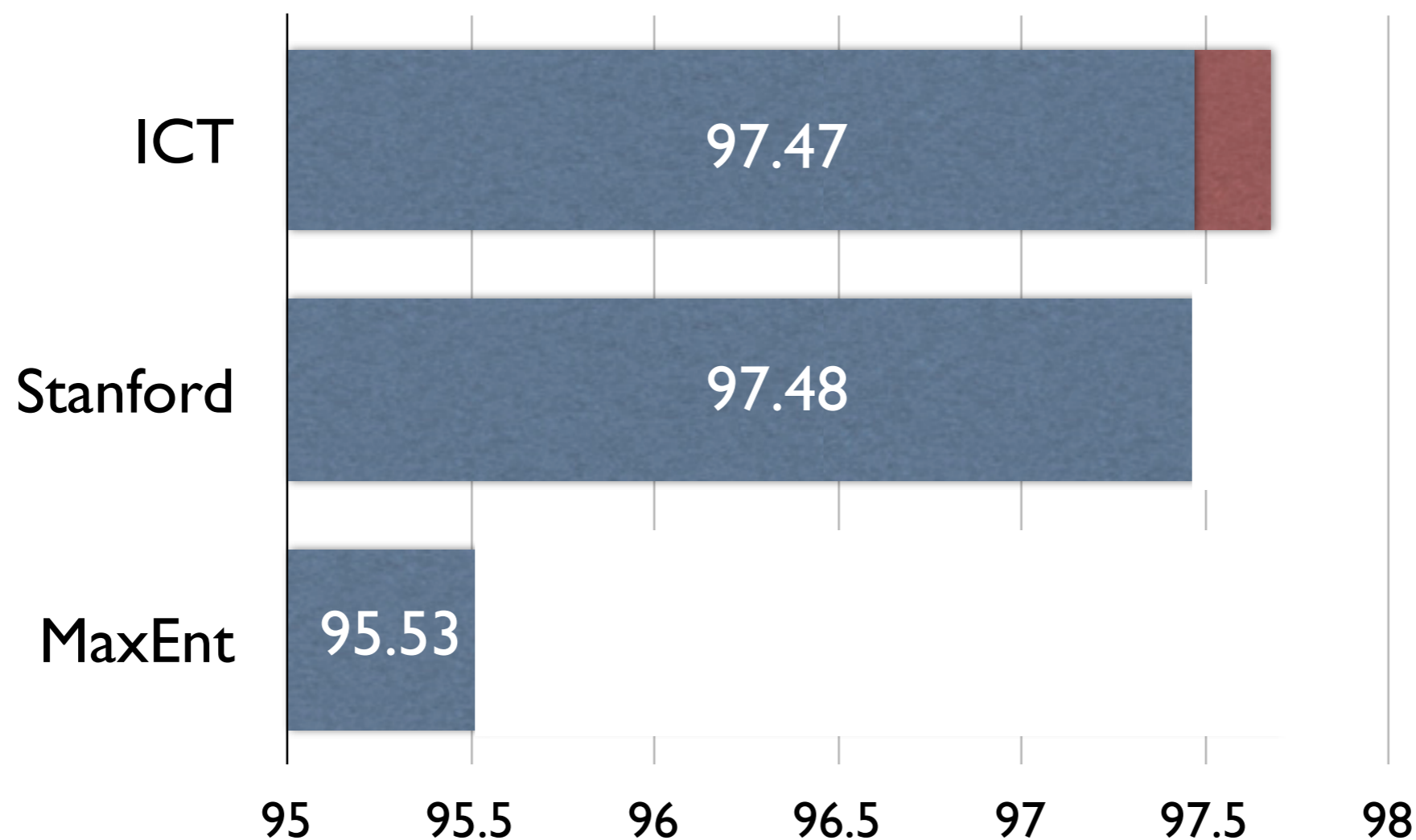
F1

(Xiao et al., 2010)

Tokenization Evaluation

separate

joint



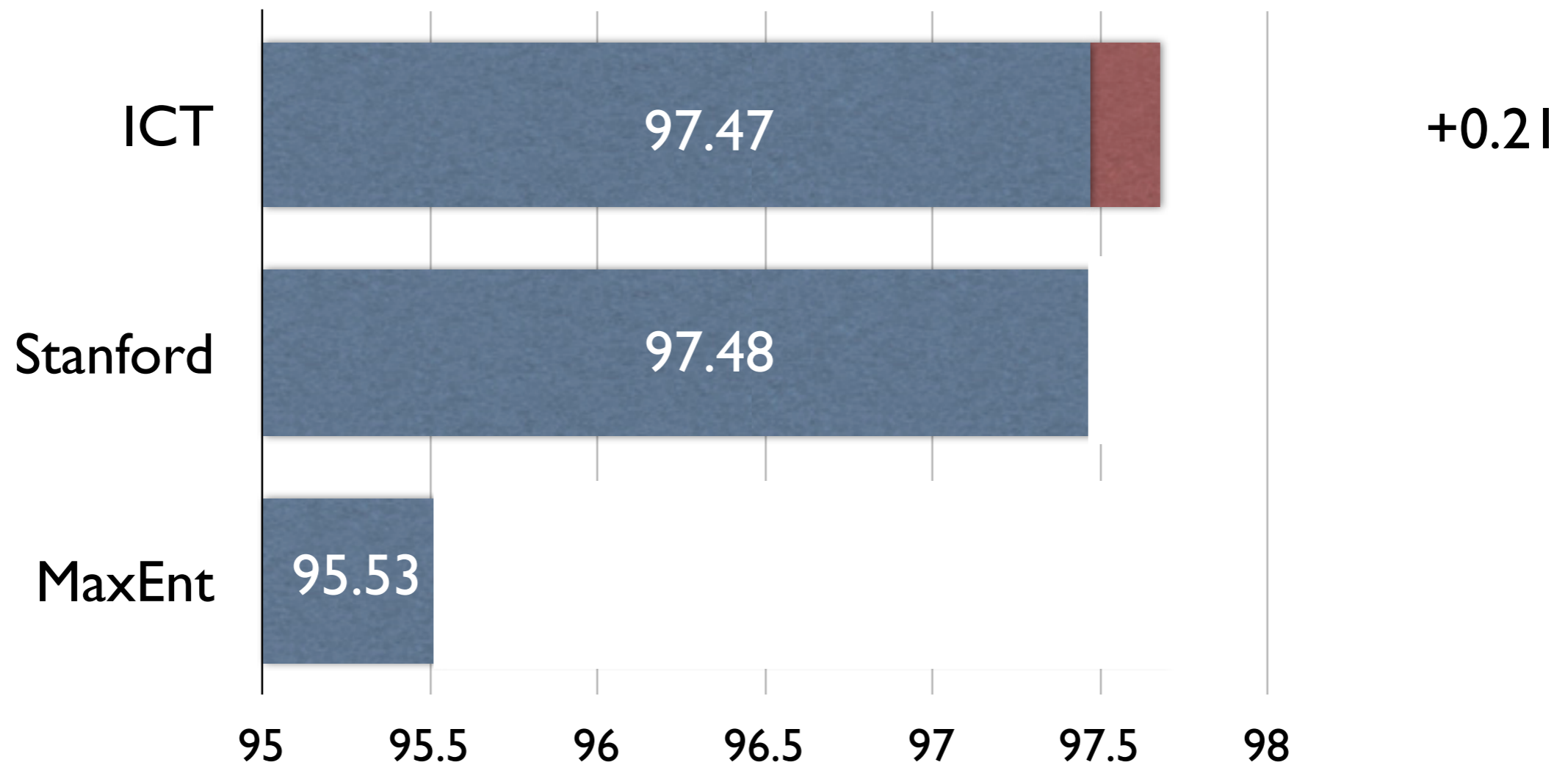
F1

(Xiao et al., 2010)

Tokenization Evaluation

separate

joint



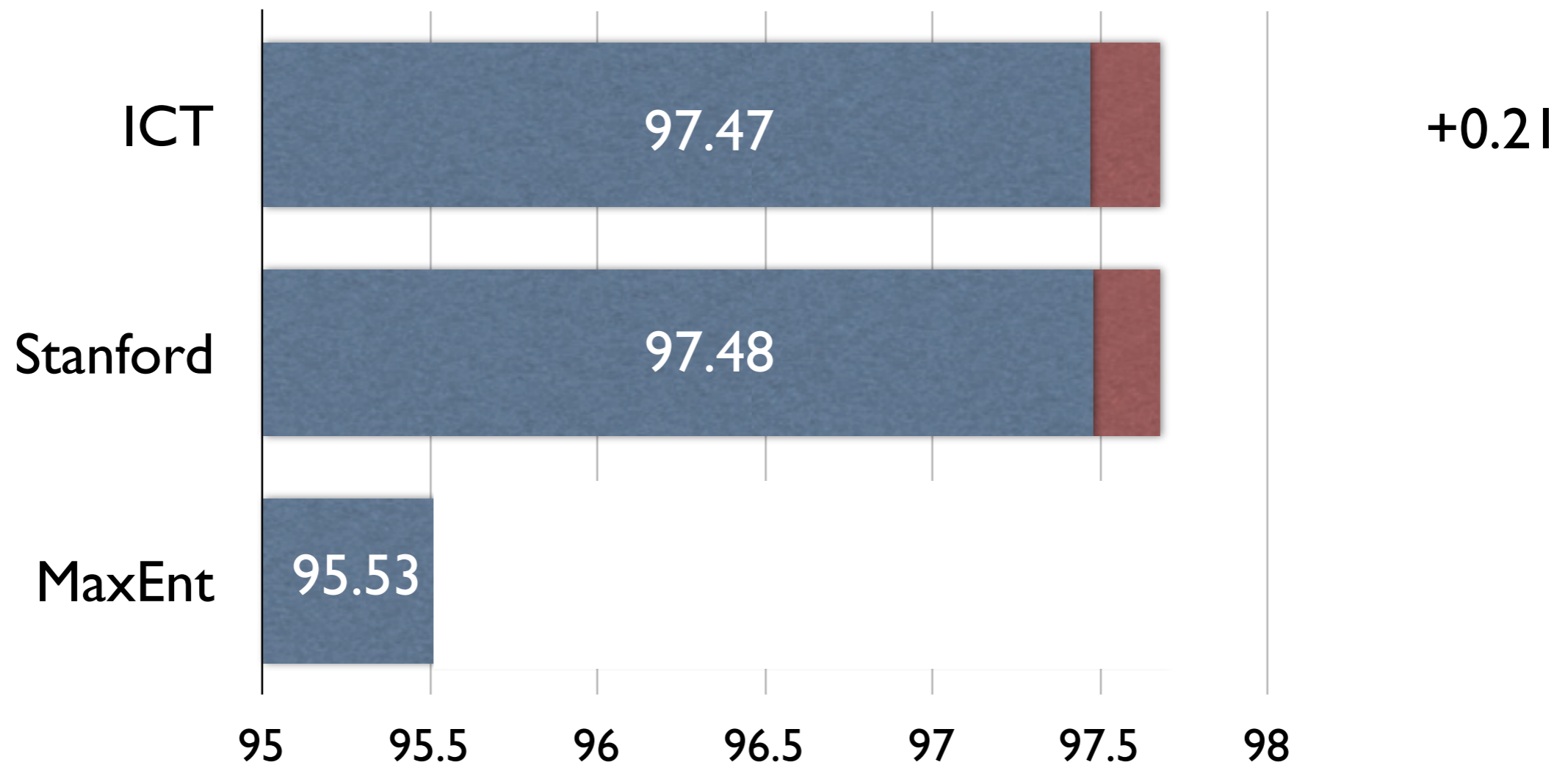
F1

(Xiao et al., 2010)

Tokenization Evaluation

separate

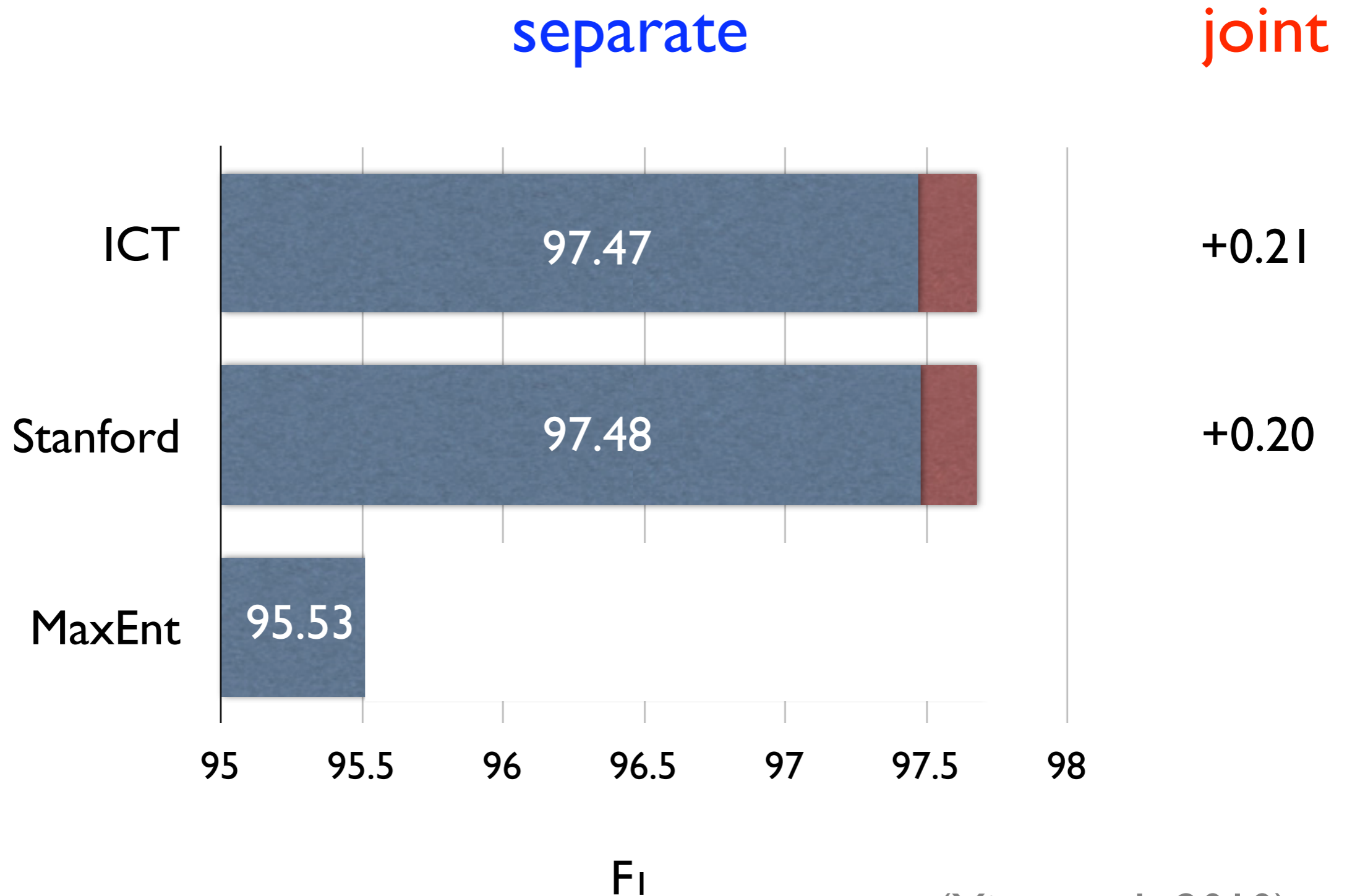
joint



F1

(Xiao et al., 2010)

Tokenization Evaluation

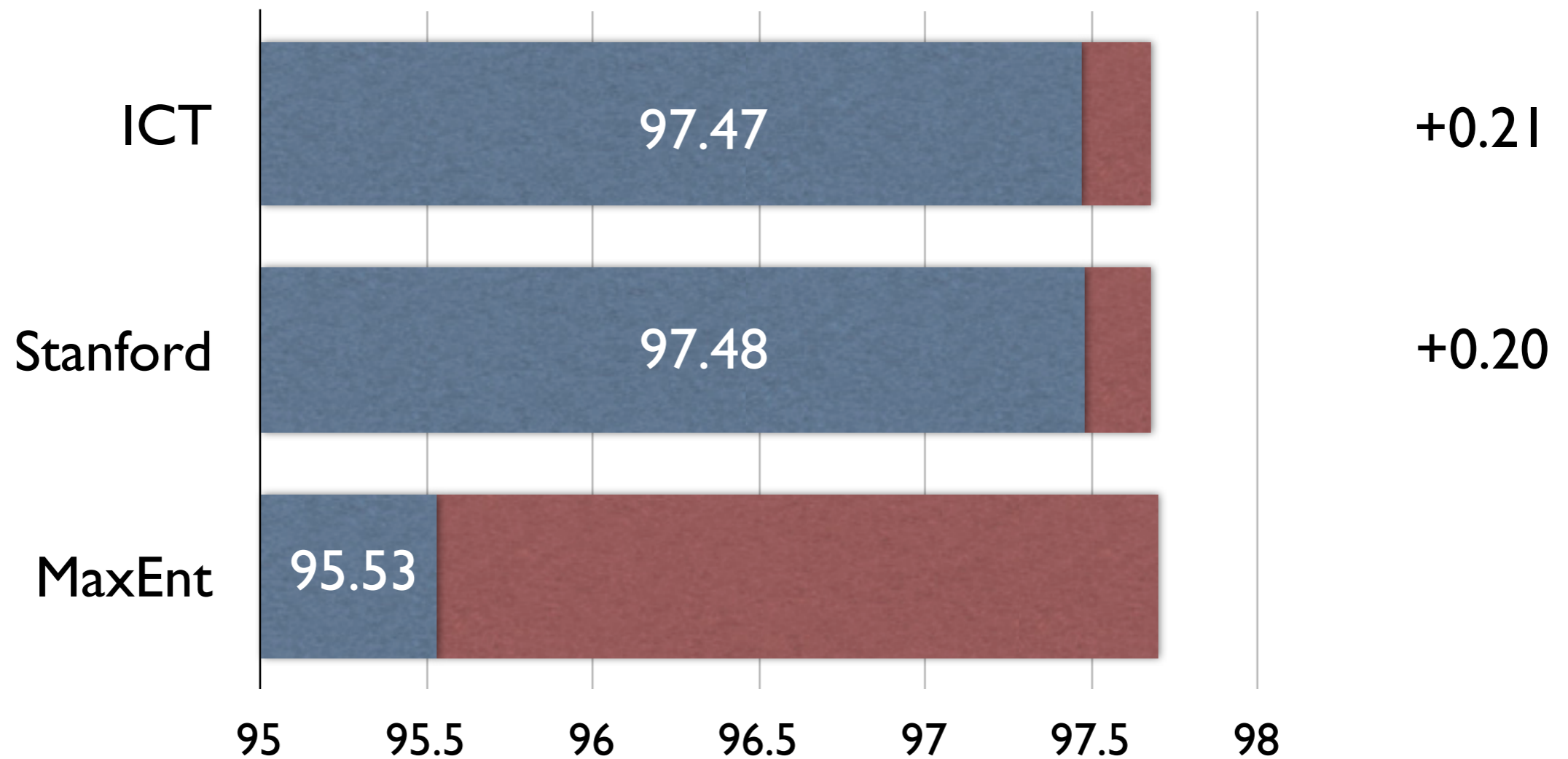


(Xiao et al., 2010)

Tokenization Evaluation

separate

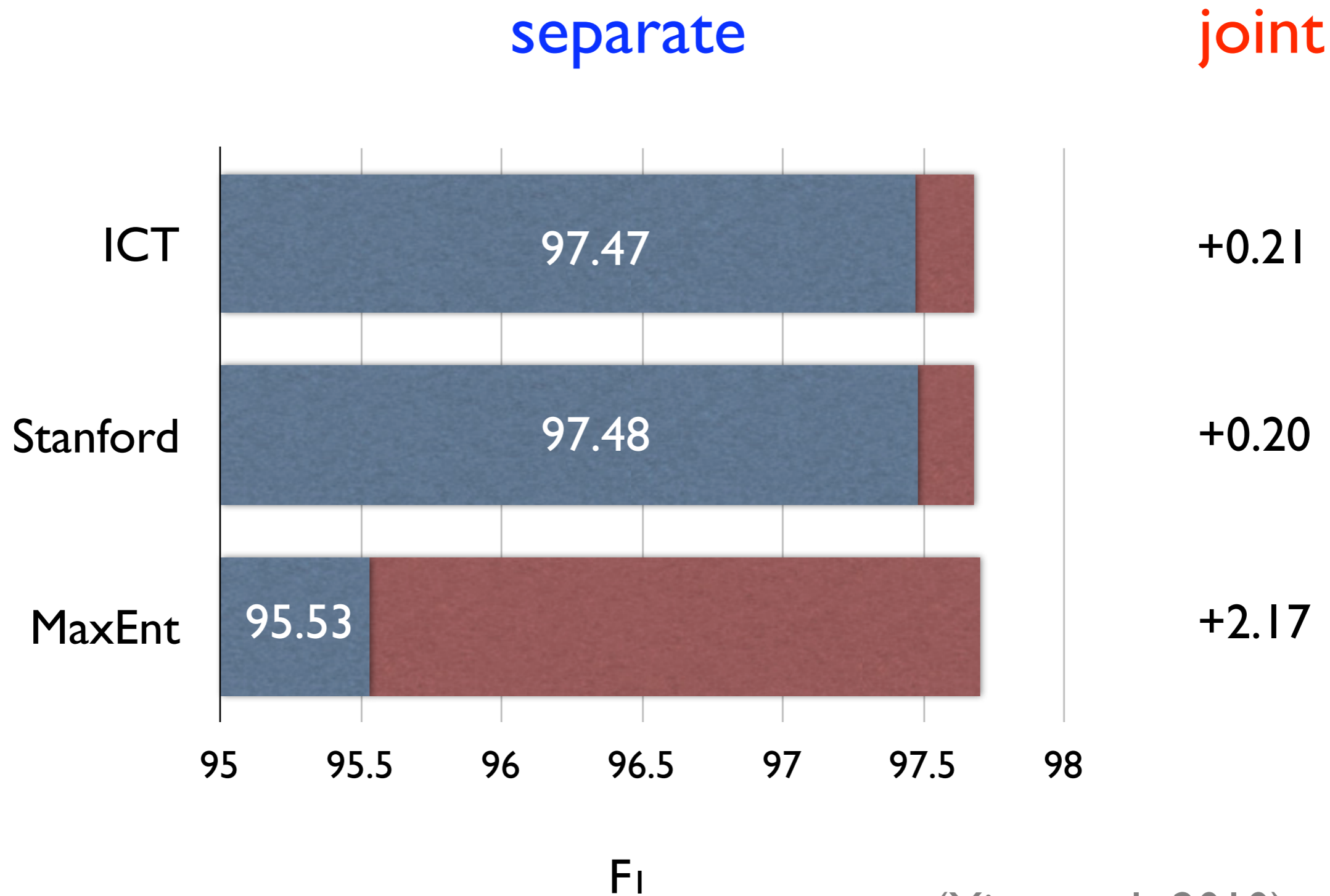
joint



F1

(Xiao et al., 2010)

Tokenization Evaluation



(Xiao et al., 2010)

Search Space Comparison

(Xiao et al., 2010)

Search Space Comparison

tokenization-based

(Xiao et al., 2010)

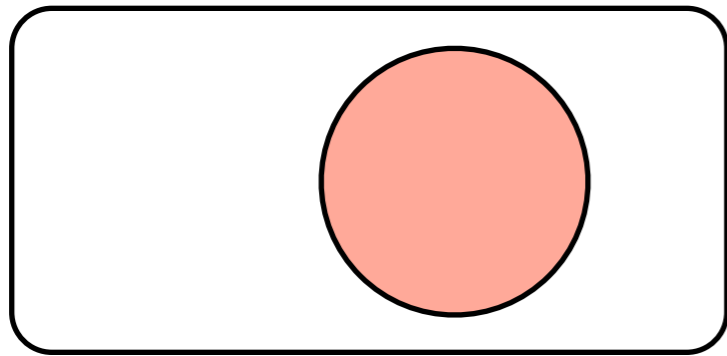
Search Space Comparison



tokenization-based

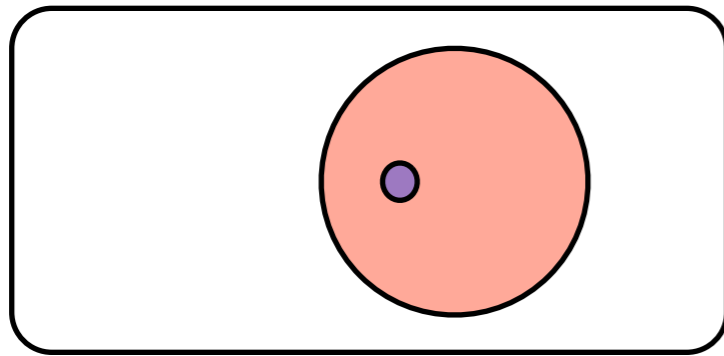
(Xiao et al., 2010)

Search Space Comparison



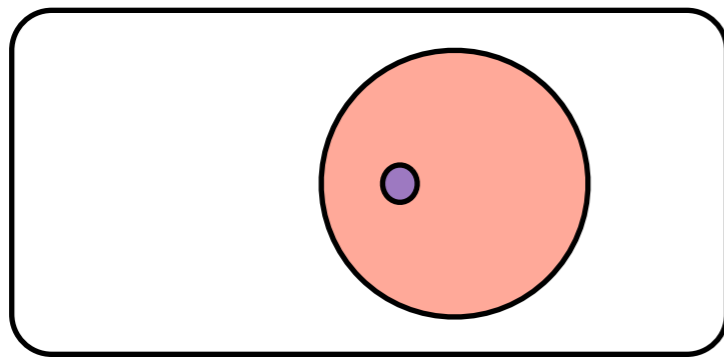
tokenization-based

Search Space Comparison

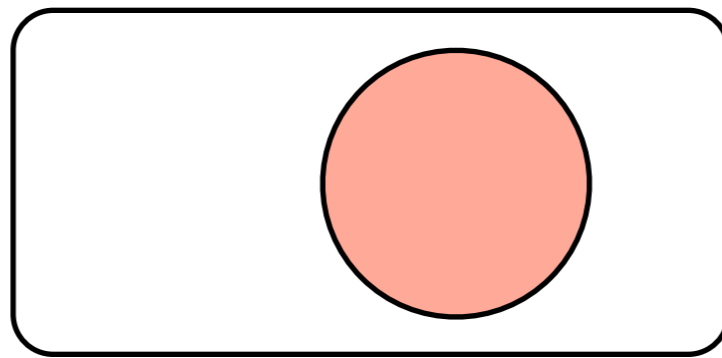


tokenization-based

Search Space Comparison

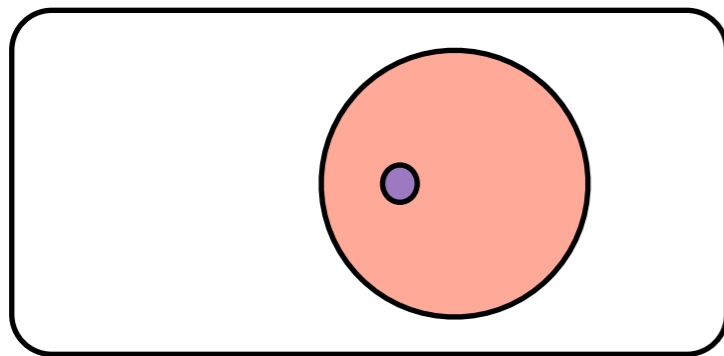


tokenization-based

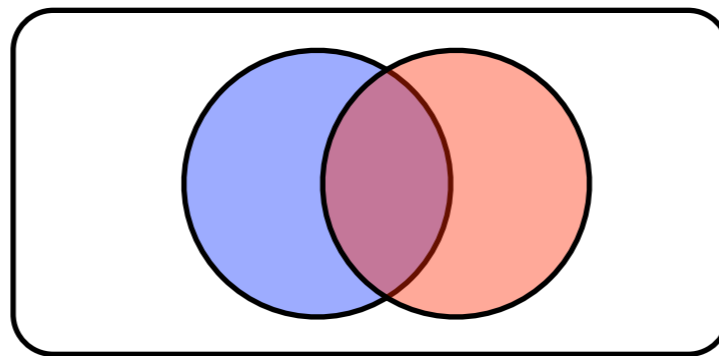


lattice-based

Search Space Comparison

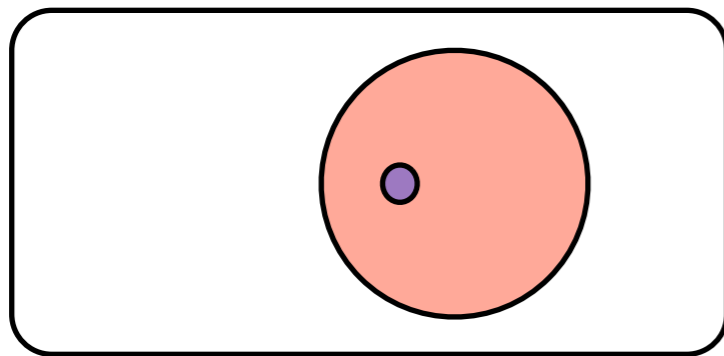


tokenization-based

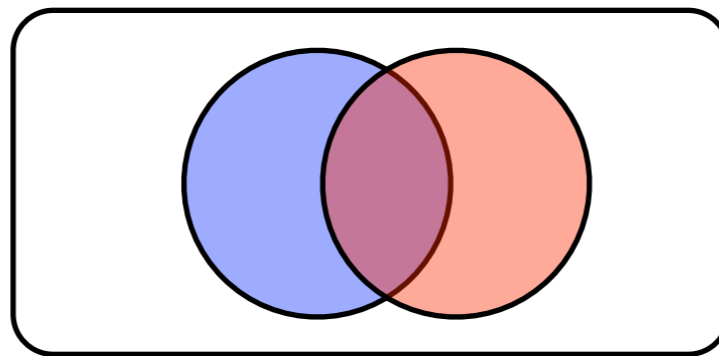


lattice-based

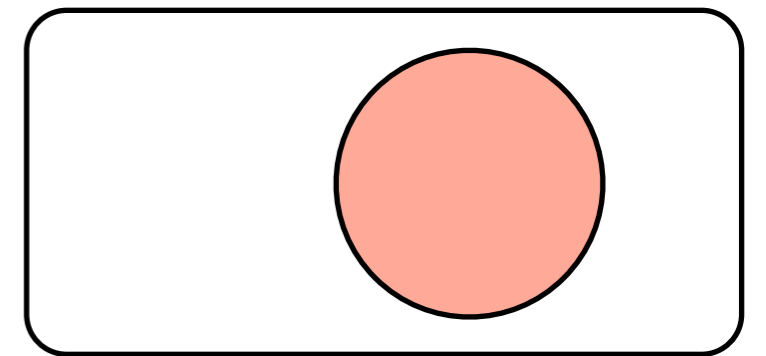
Search Space Comparison



tokenization-based



lattice-based



string-based

Parsing

布什 与 沙龙 举行 了 会谈

Parsing

NPB

|

布什

与

沙龙

举行

了

会谈

Parsing

NPB
|
布什

P
|
与

沙龙

举行

了

会谈

Parsing

NPB
|
布什

P
|
与

NPB
|
沙龙

举行

了

会谈

Parsing

NPB
|
布什

P
|
与

NPB
|
沙龙

VS
|
举行

了

会谈

Parsing

NPB	P	NPB	VS	AS	
布什	与	沙龙	举行	了	会谈

Parsing

NPB
|
布什

P
|
与

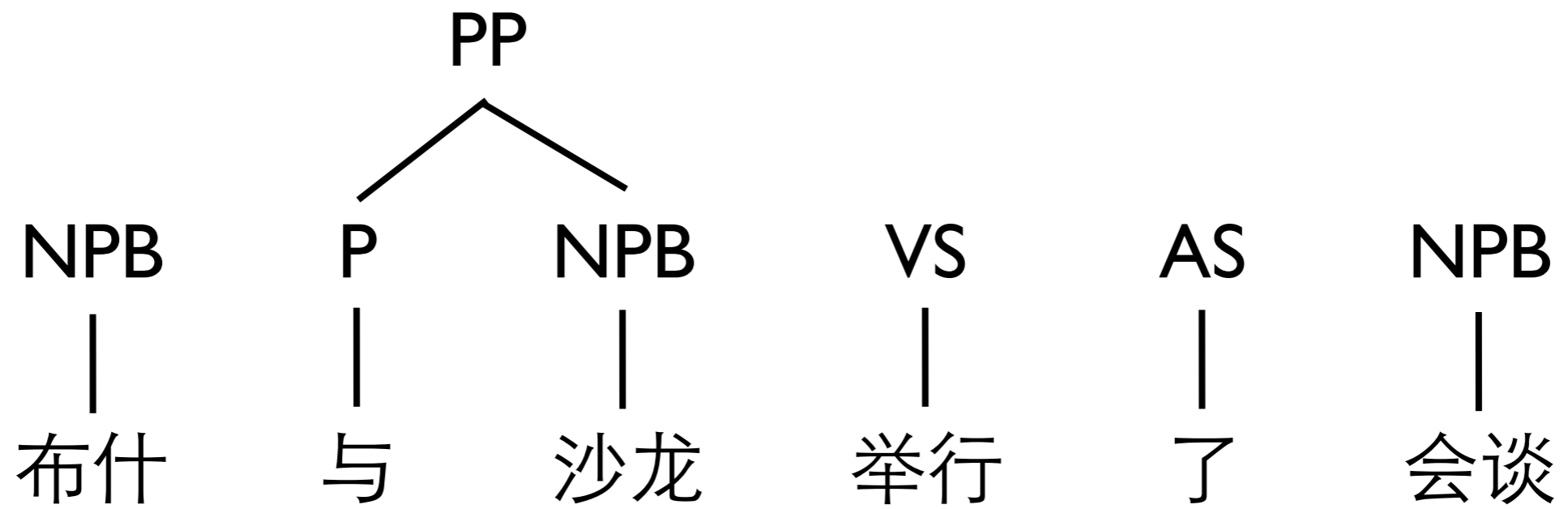
NPB
|
沙龙

VS
|
举行

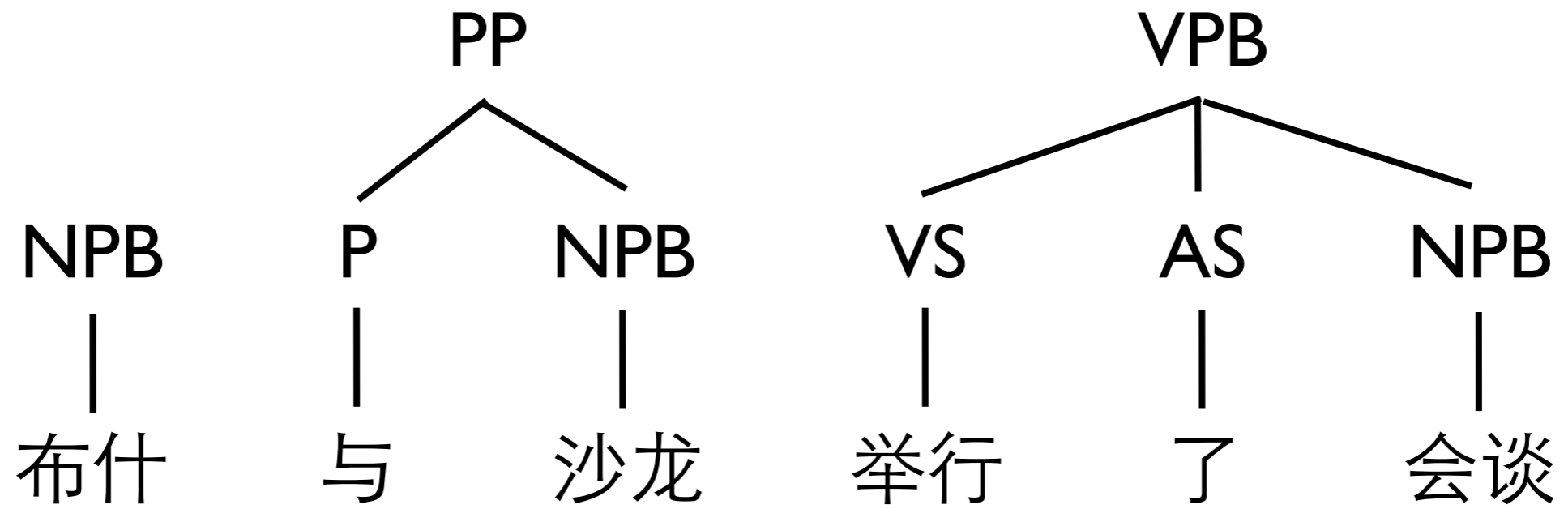
AS
|
了

NPB
|
会谈

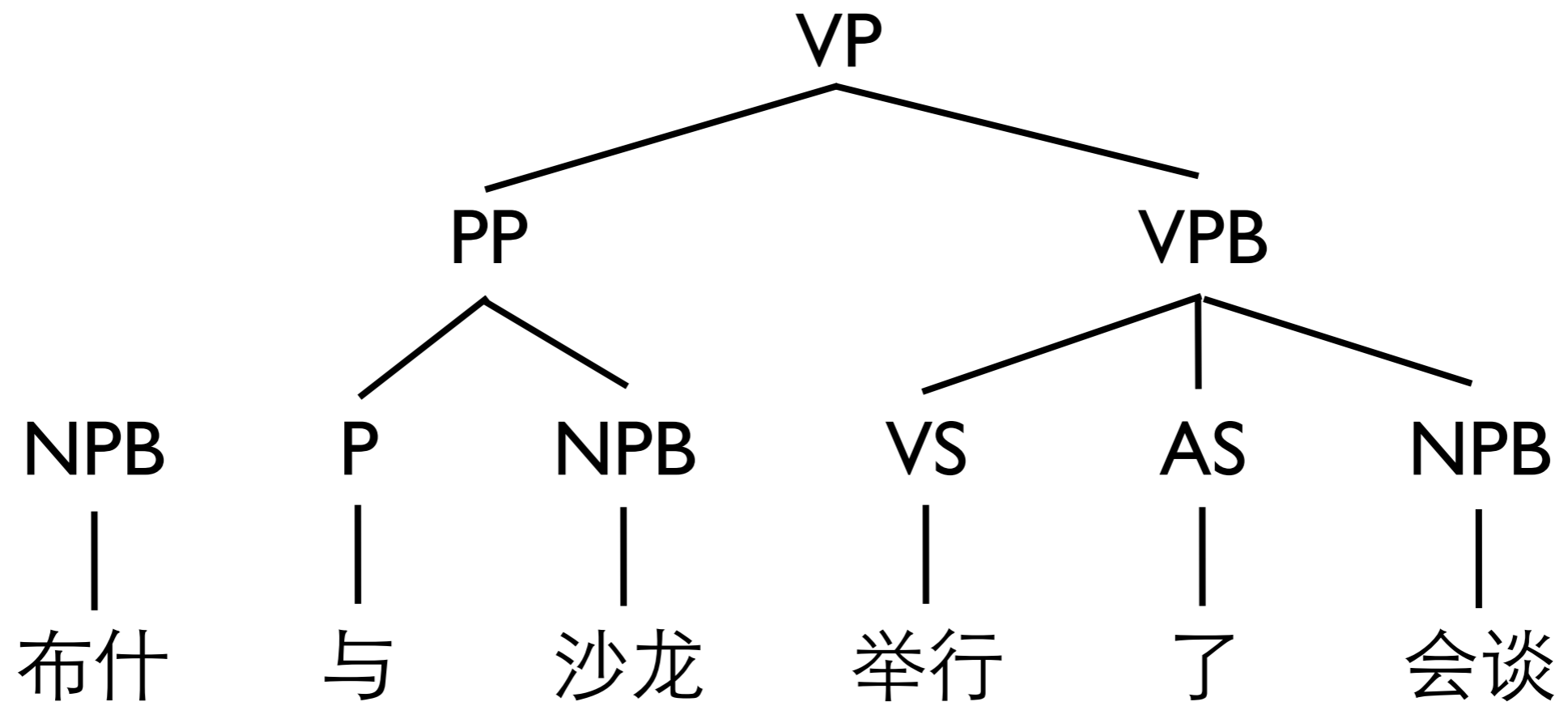
Parsing



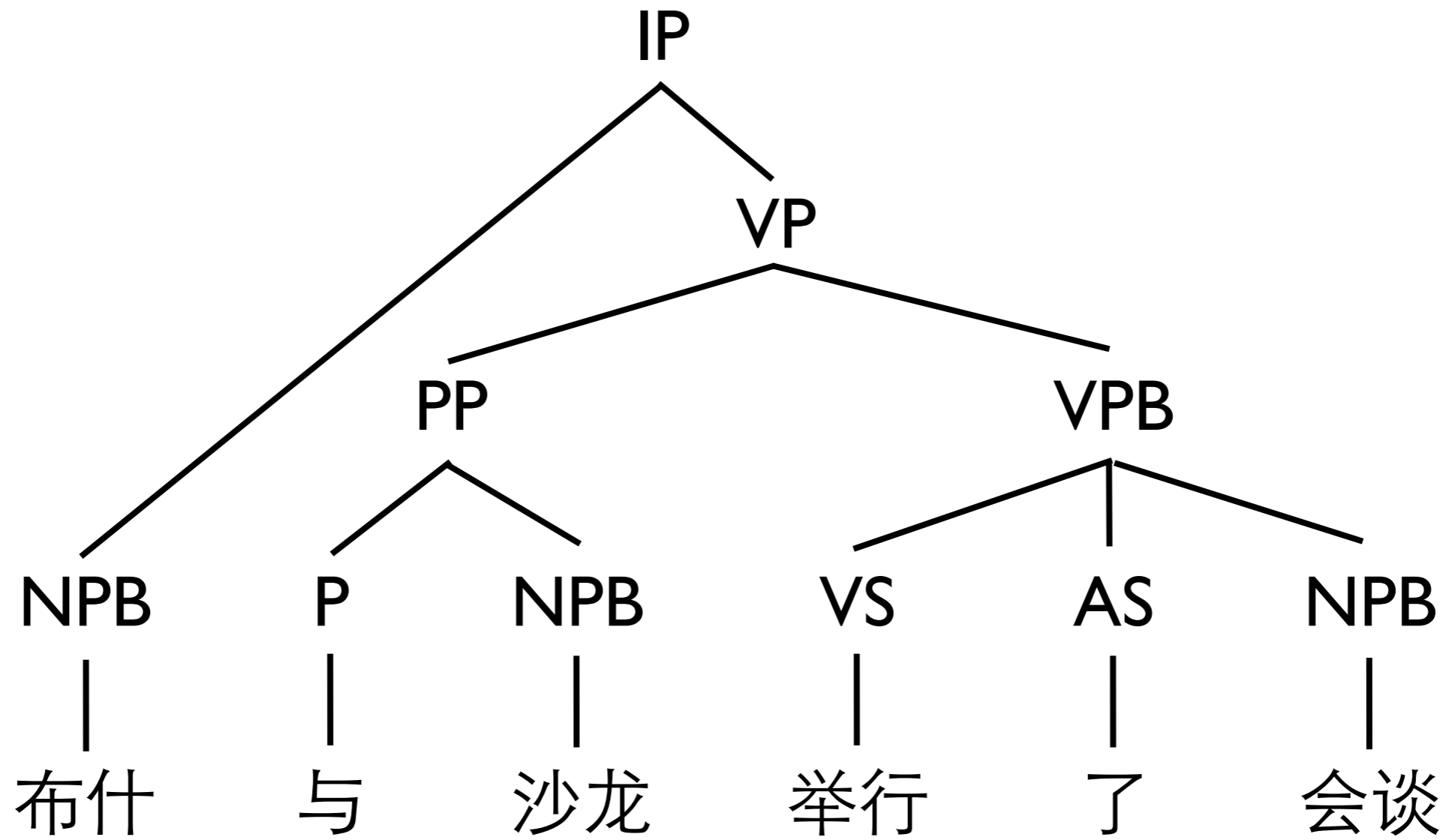
Parsing



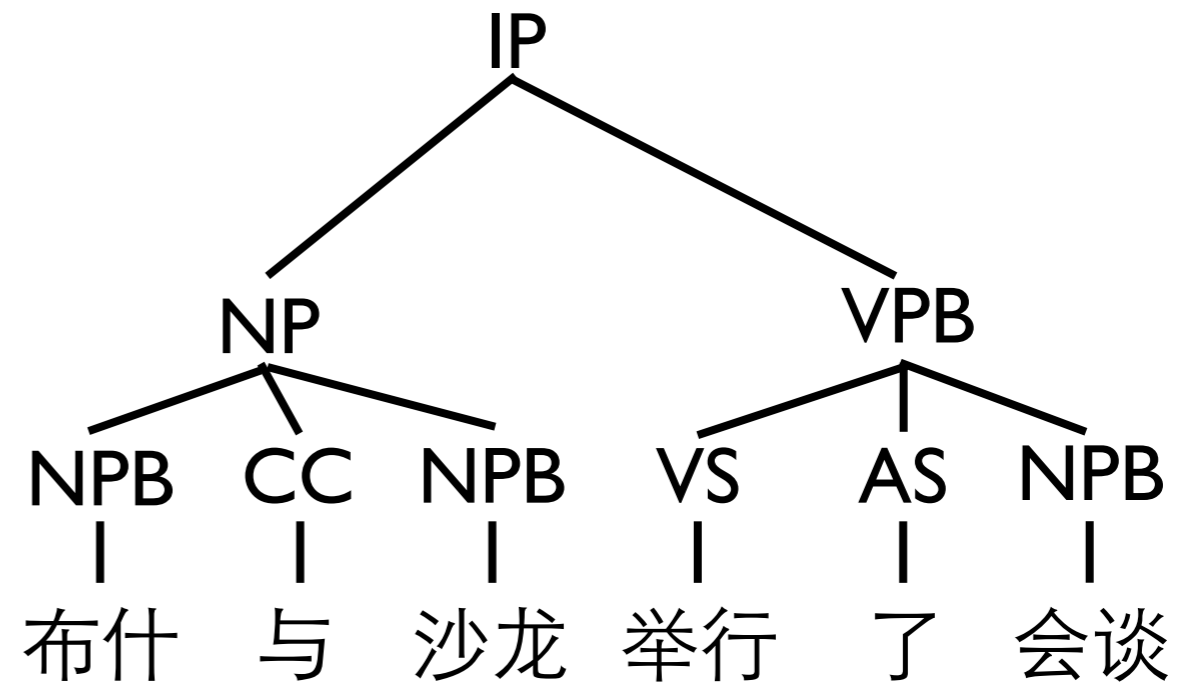
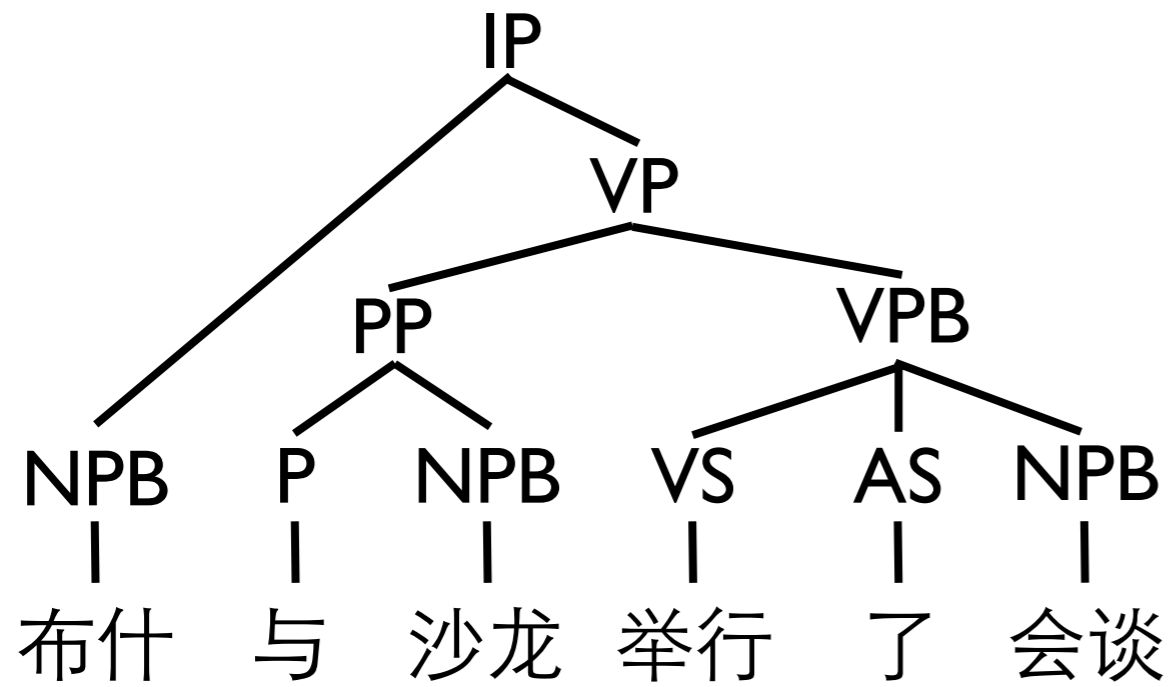
Parsing



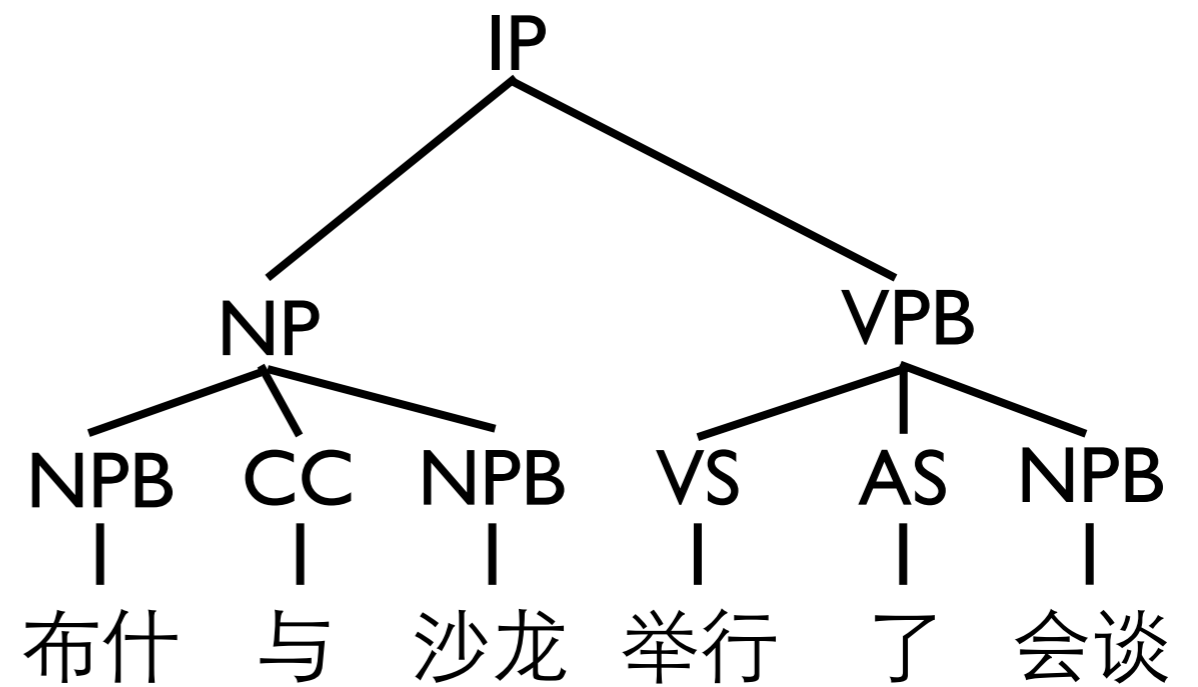
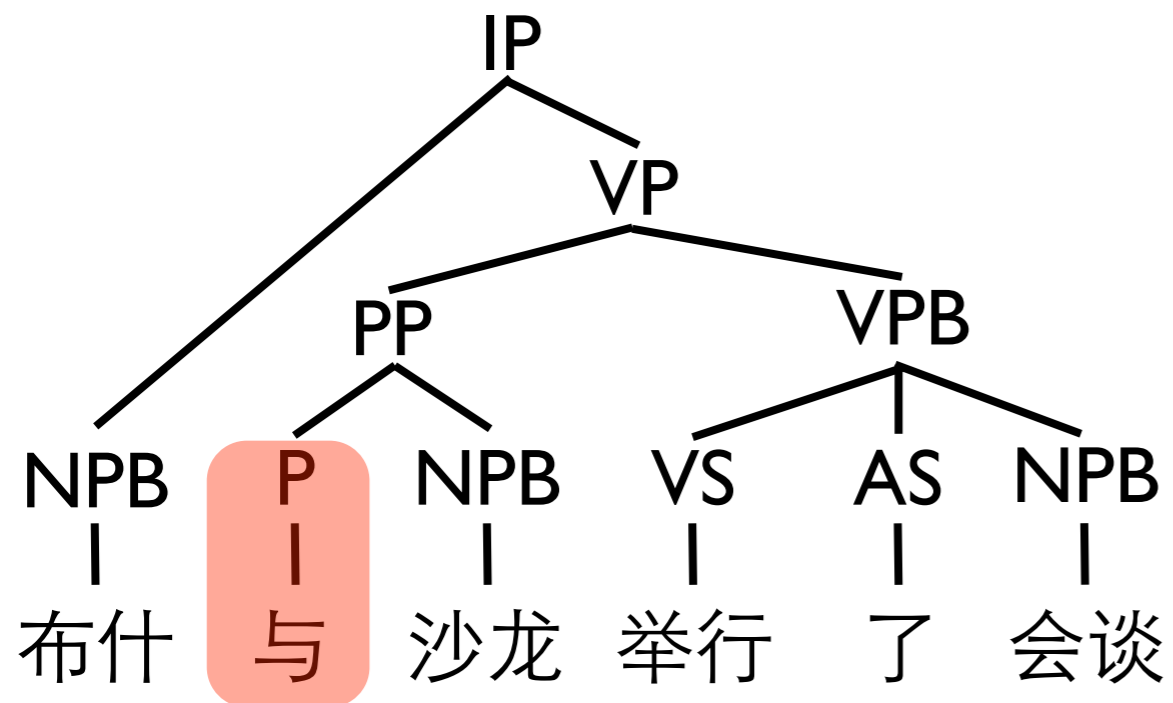
Parsing



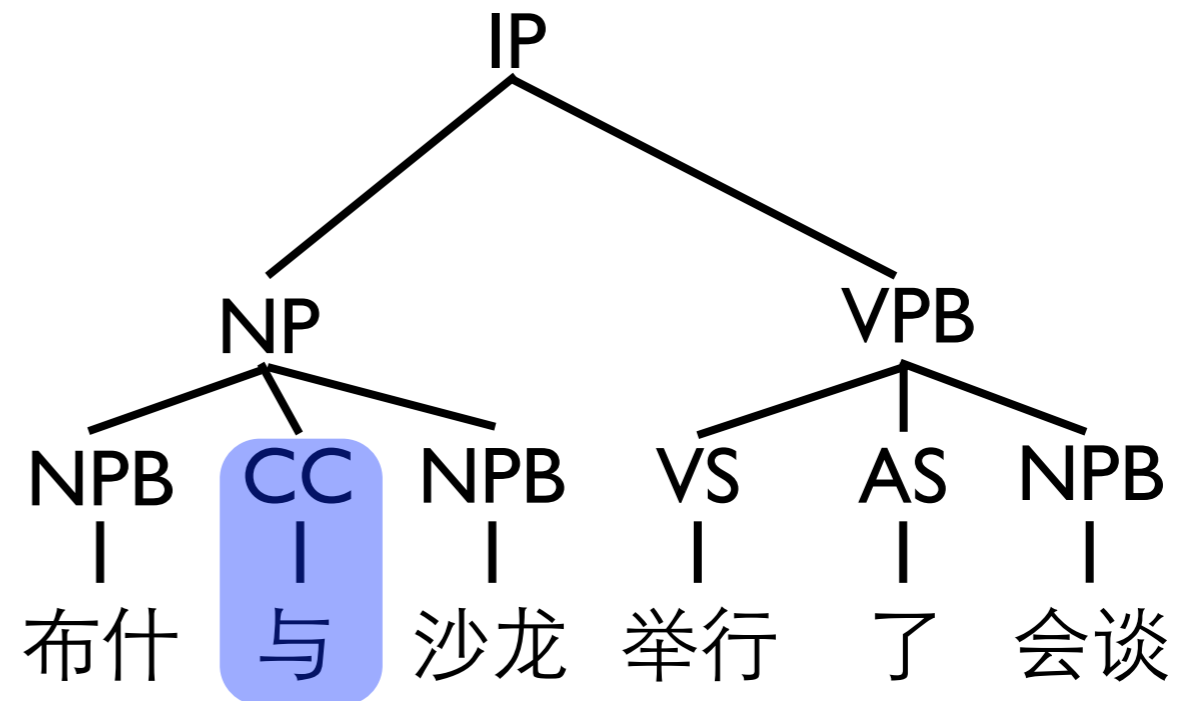
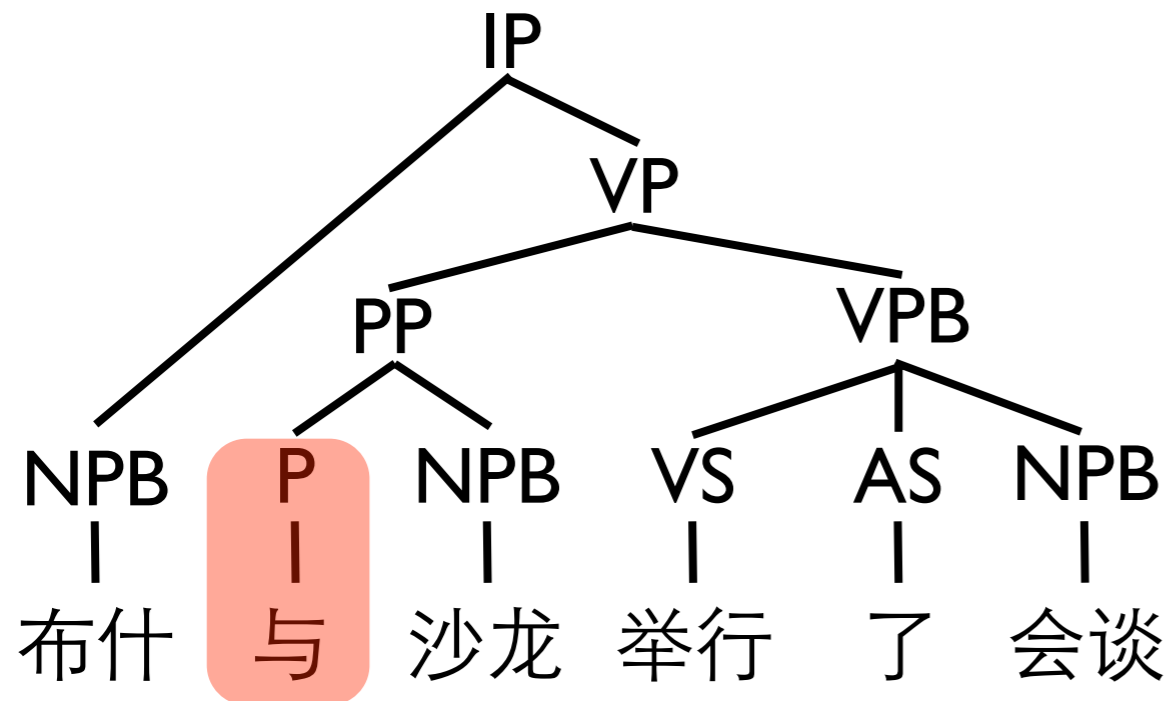
Structural Ambiguity



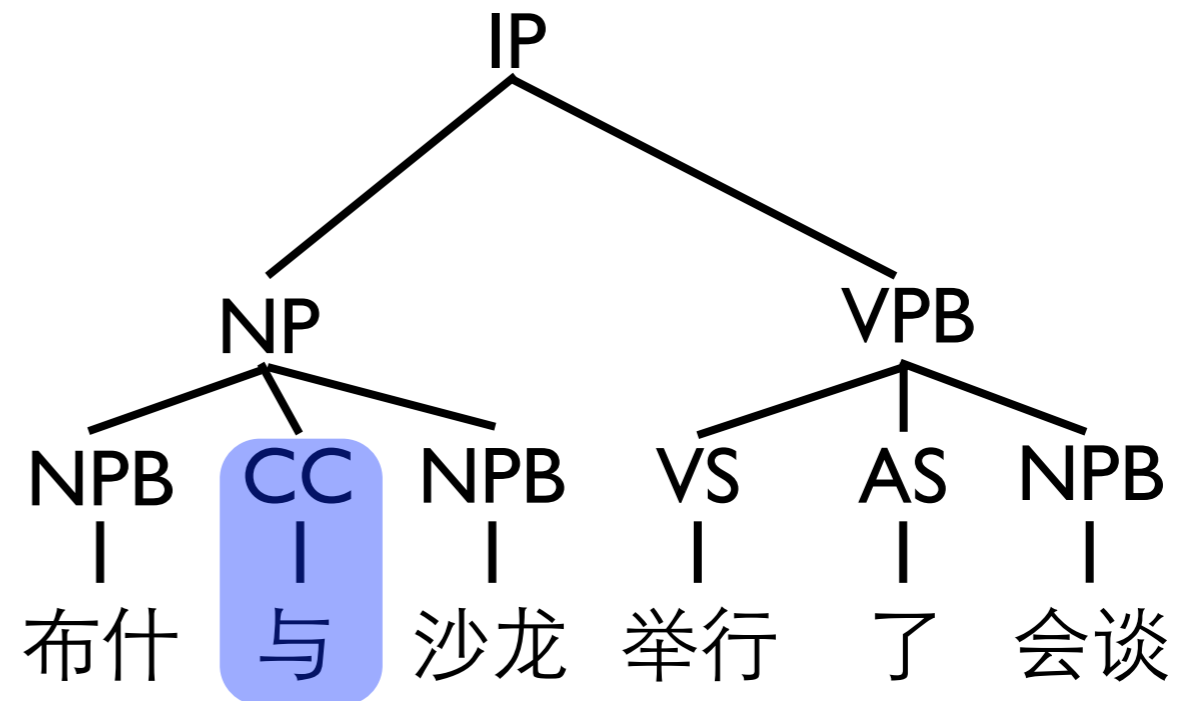
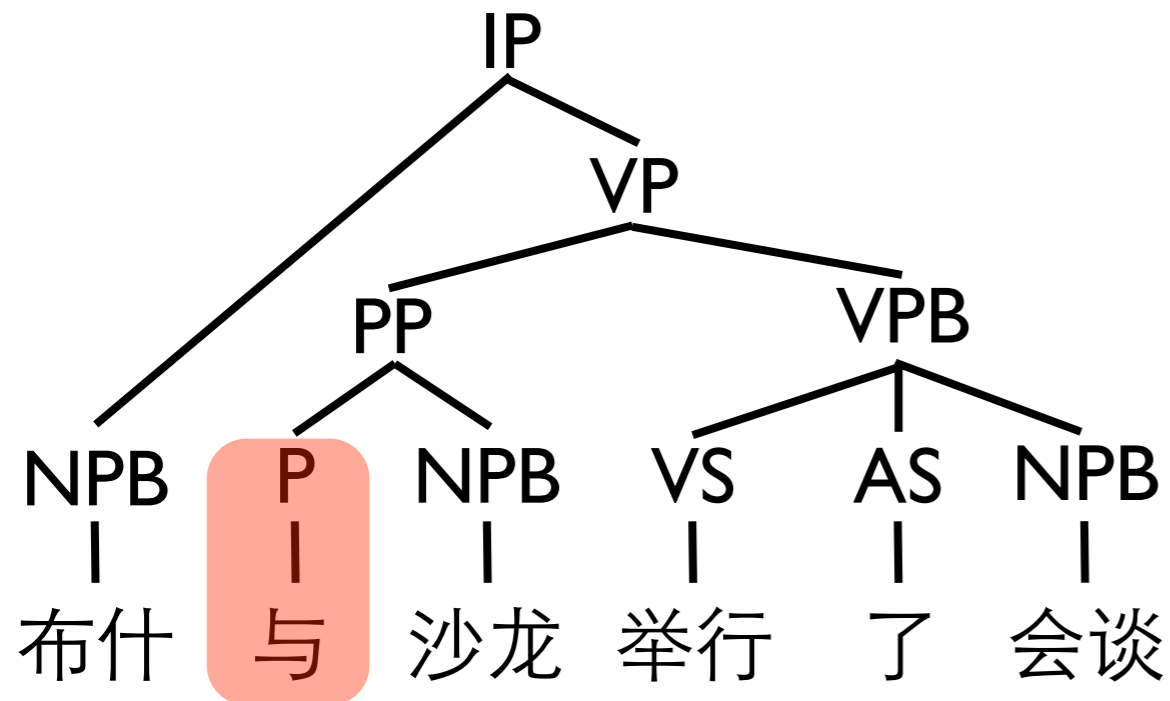
Structural Ambiguity



Structural Ambiguity

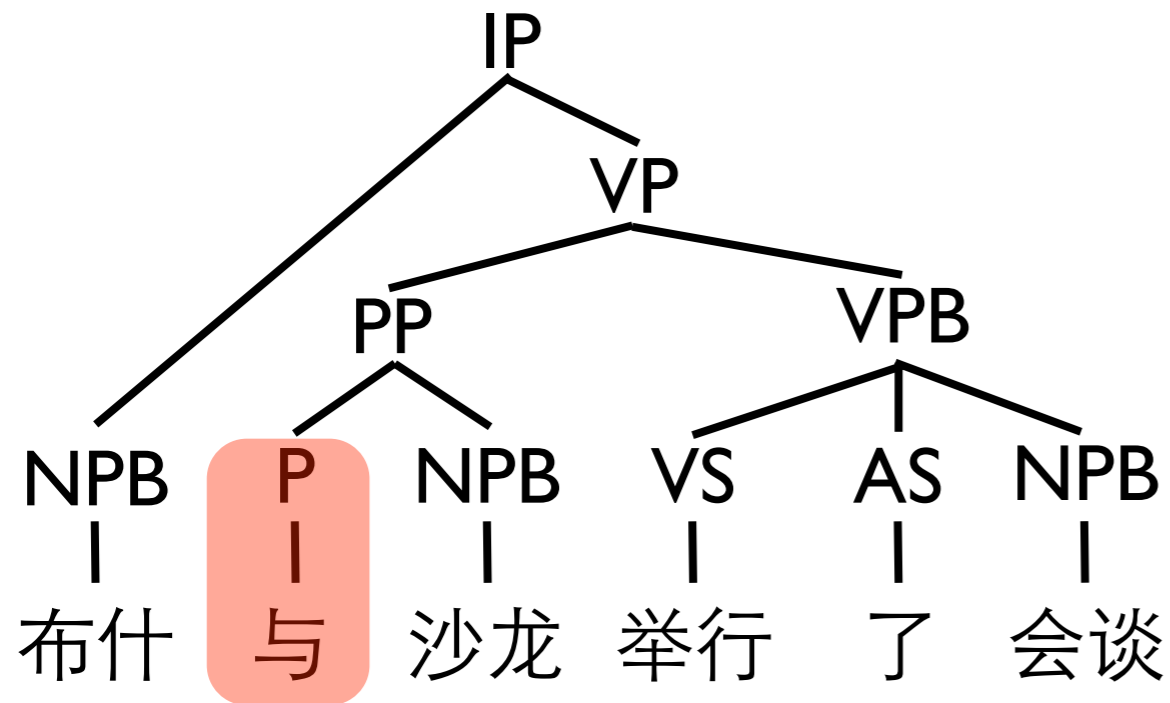


Structural Ambiguity

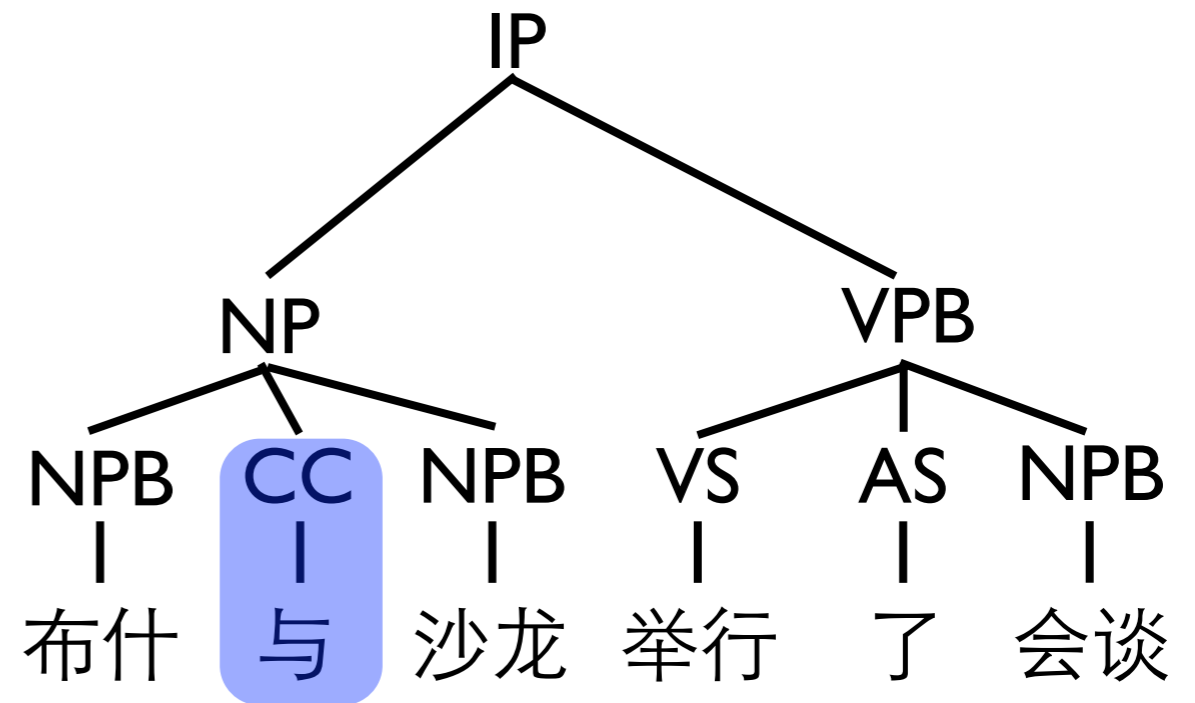


Bush held a talk **with** Sharon

Structural Ambiguity

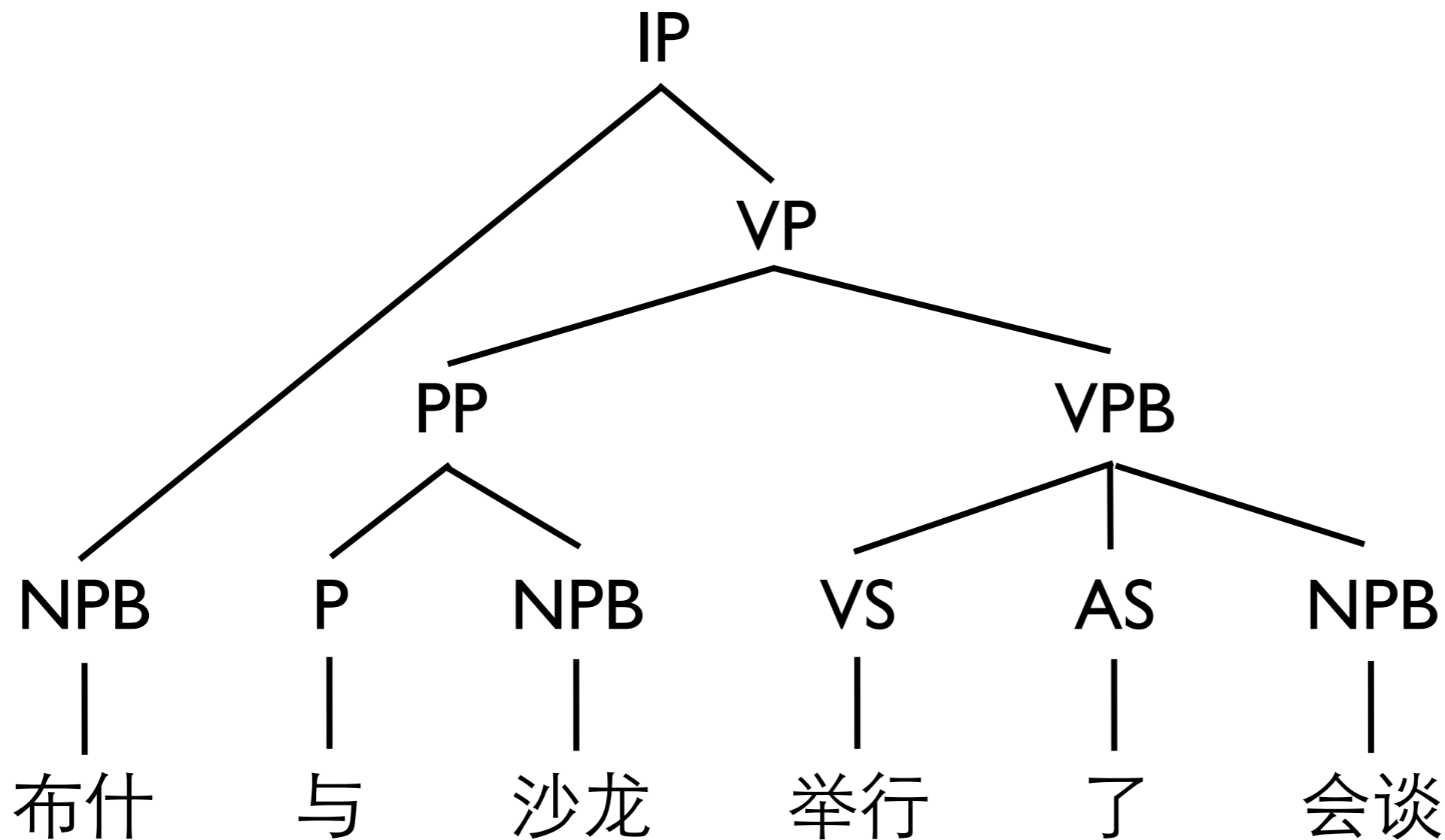


Bush held a talk **with** Sharon



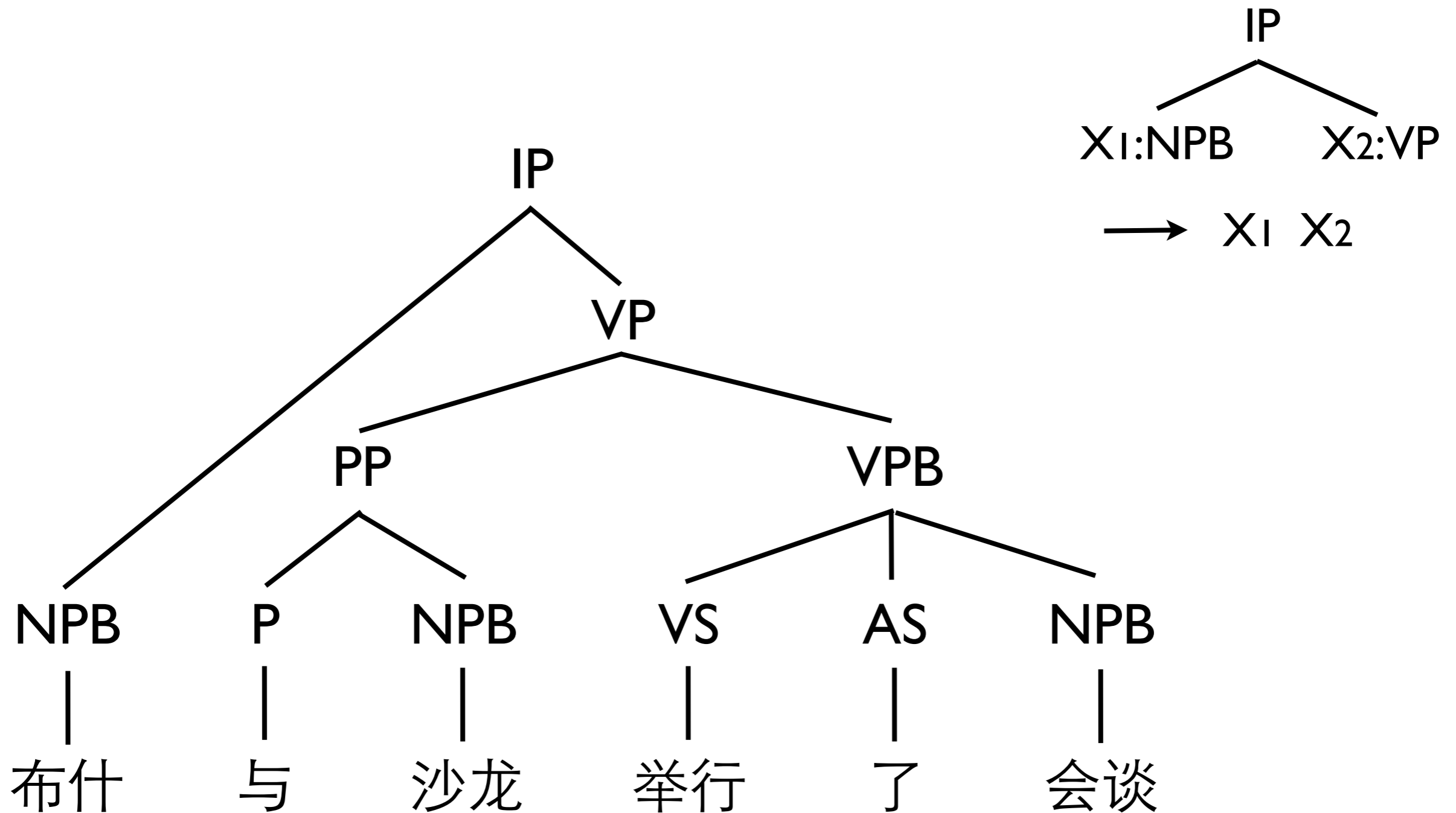
Bush **and** Sharon held a talk

Tree-to-String Translation



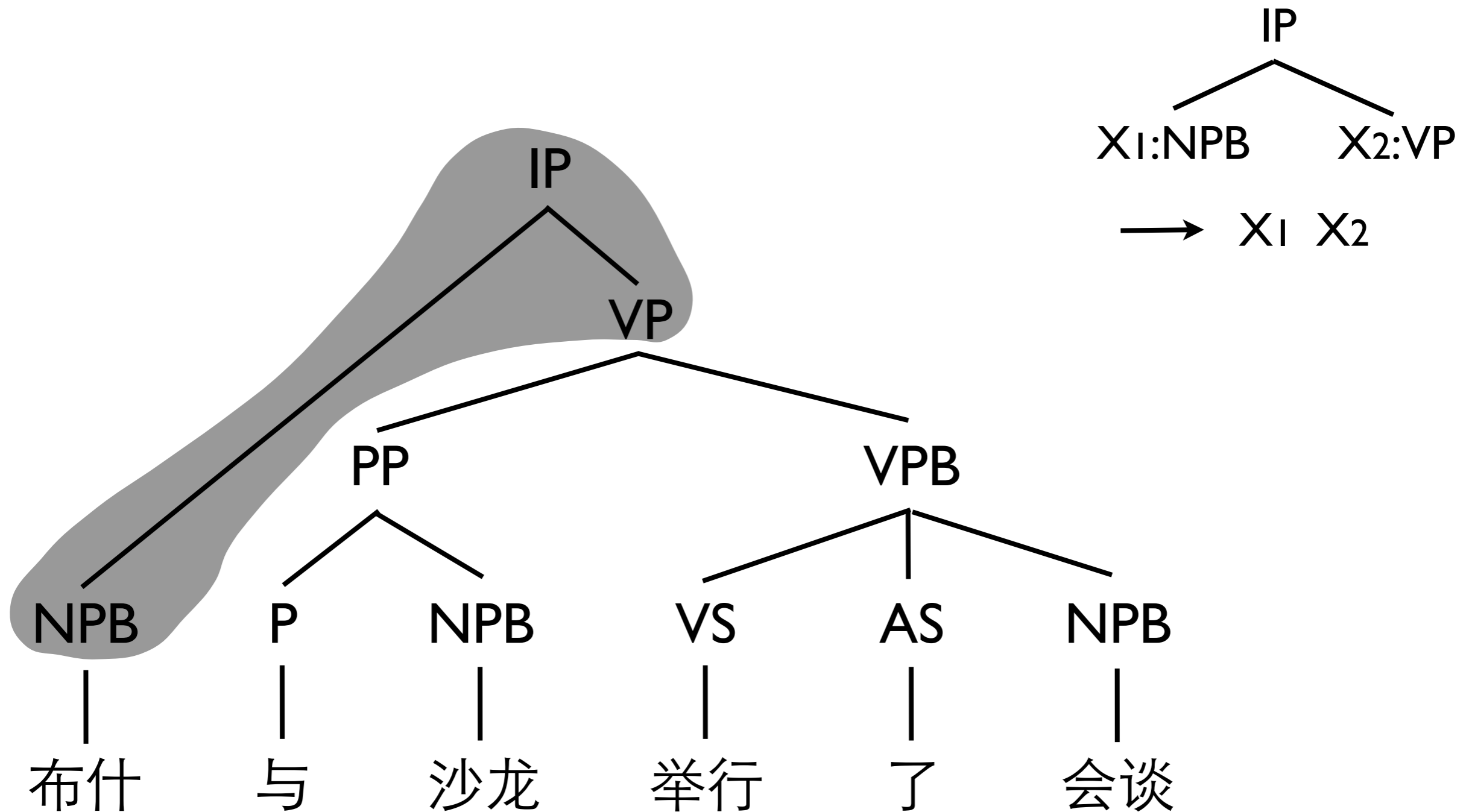
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



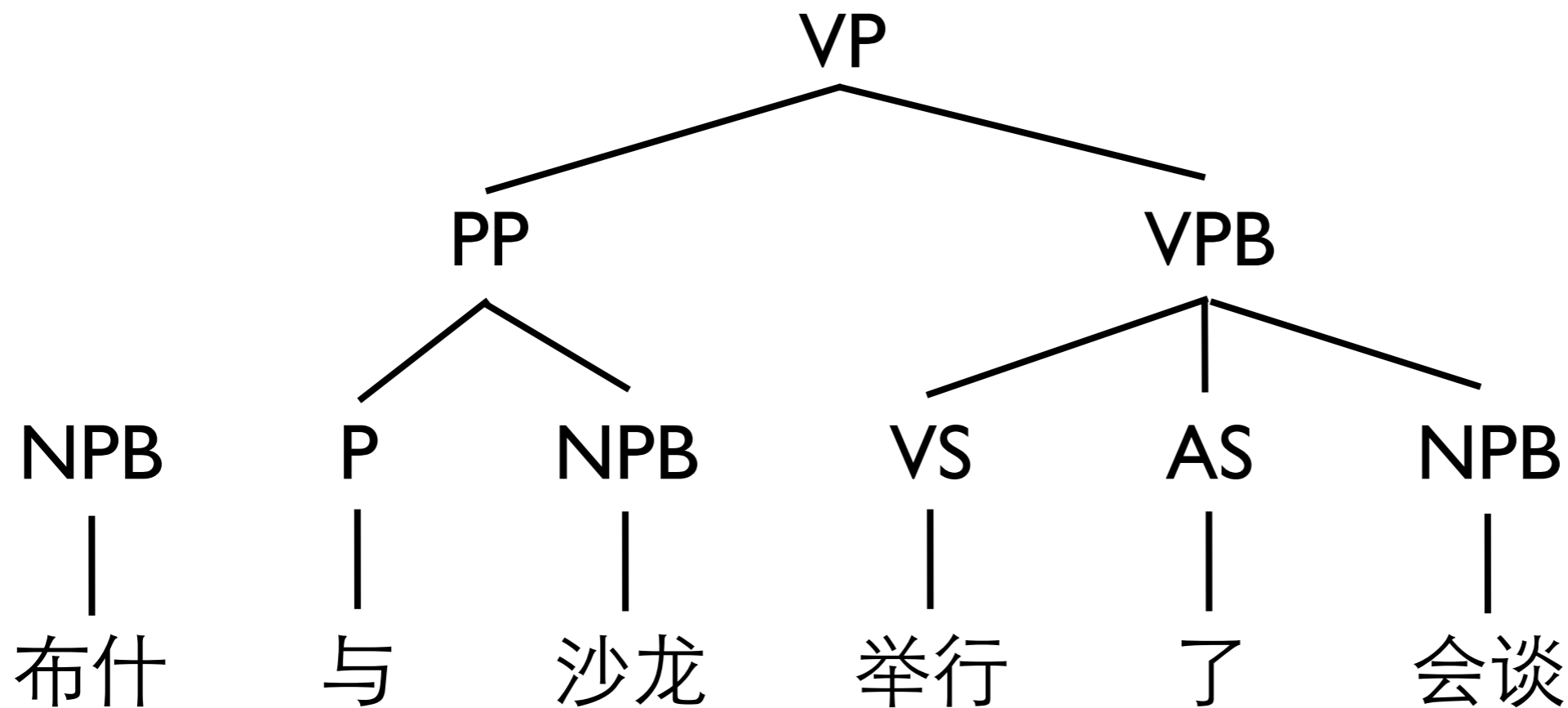
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



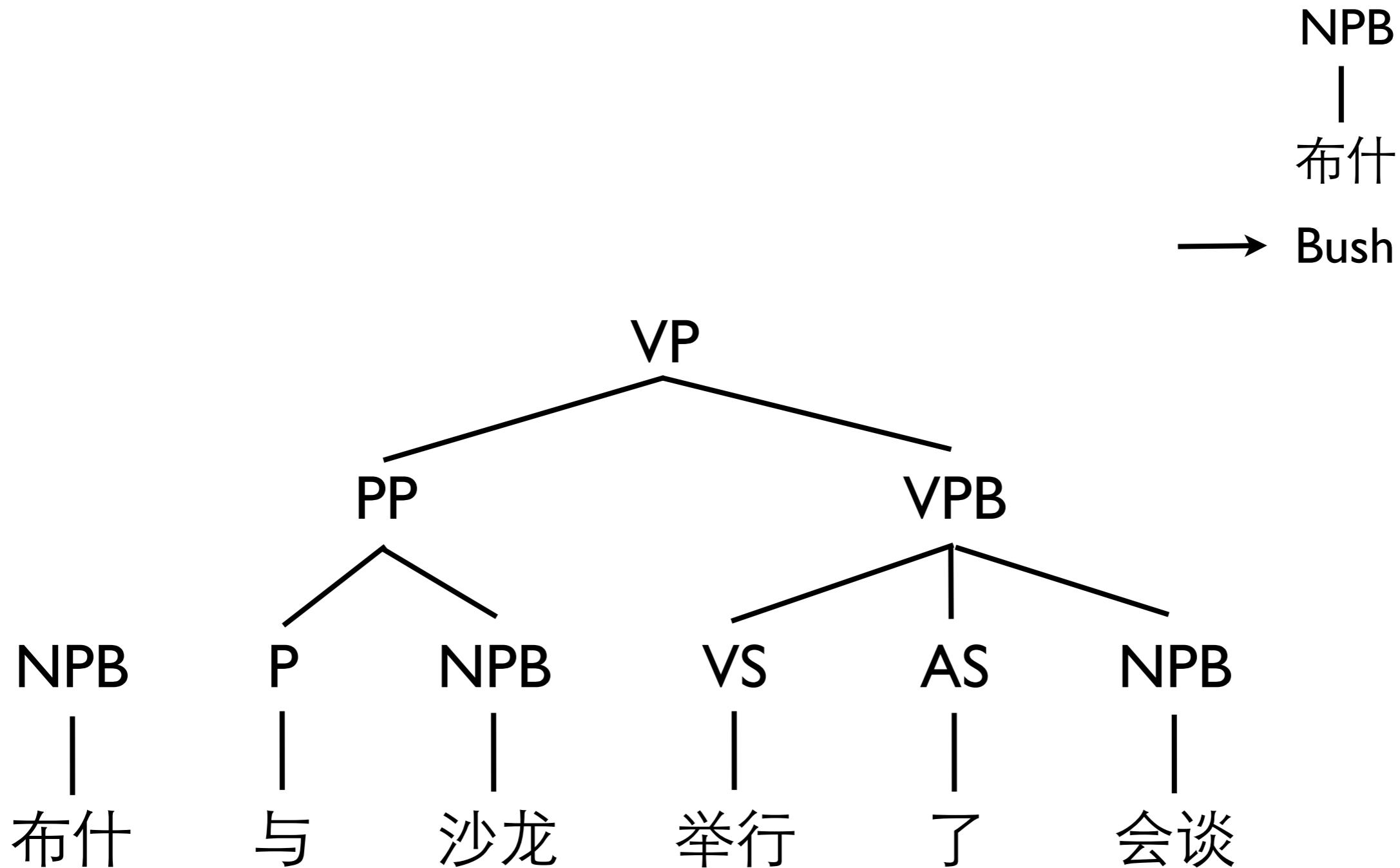
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



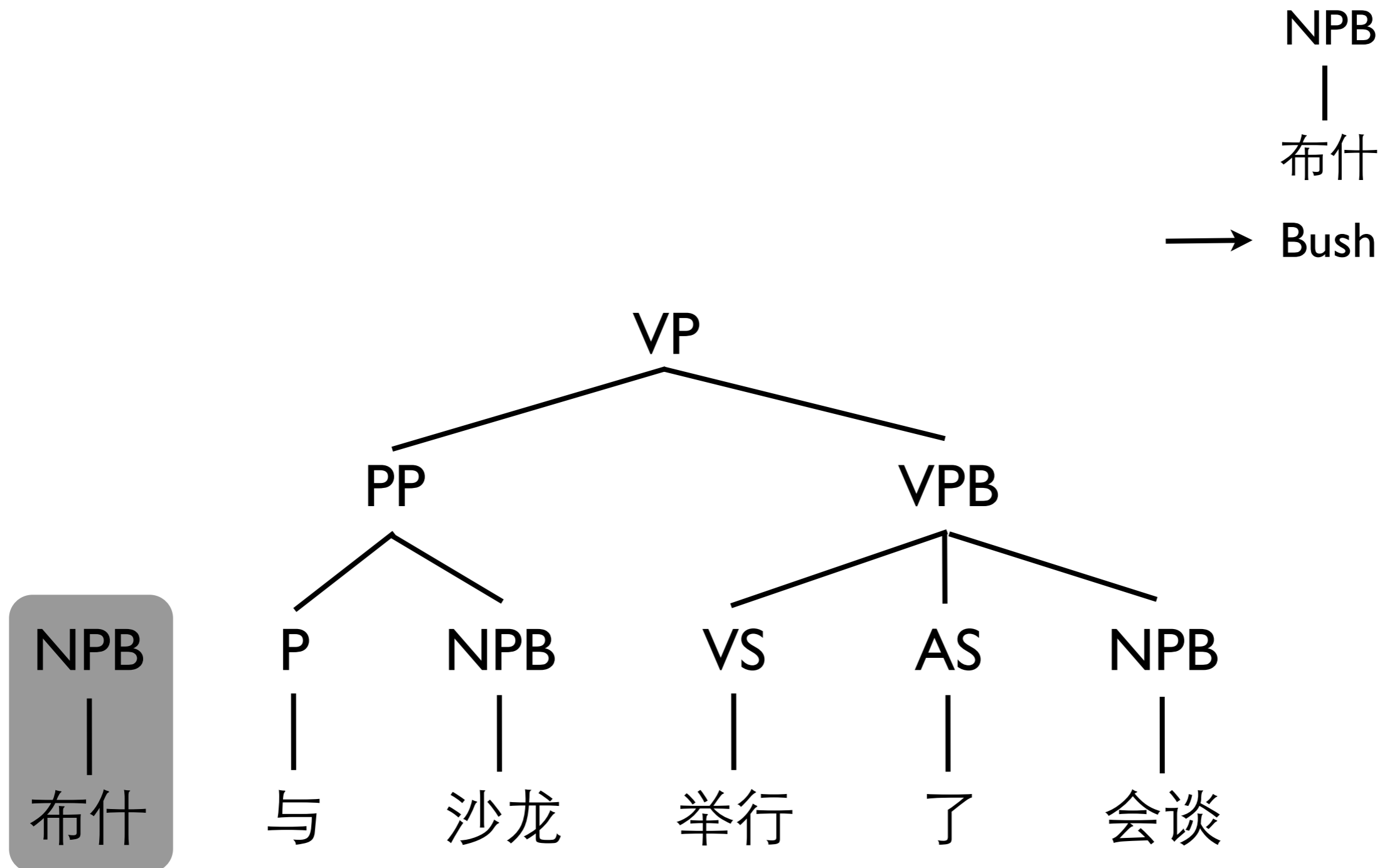
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



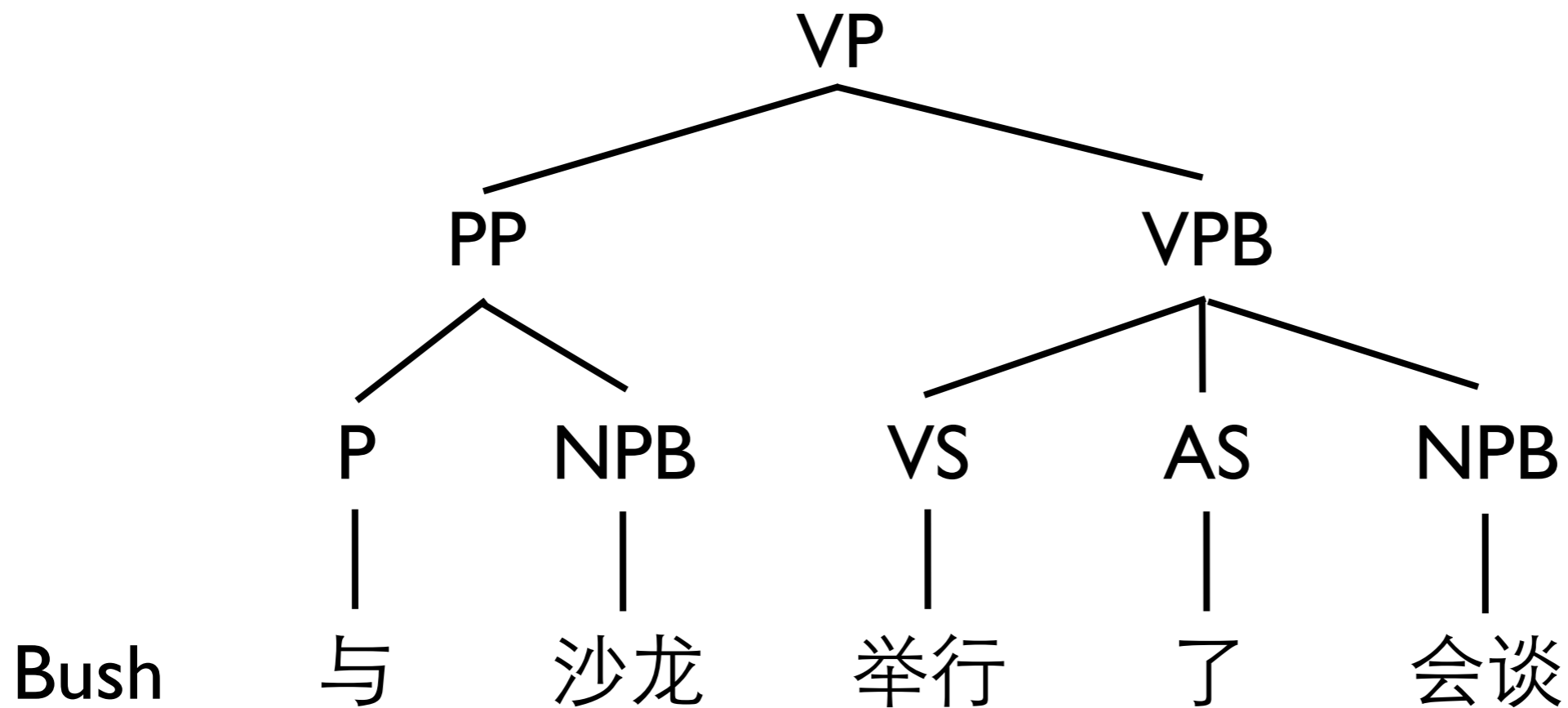
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



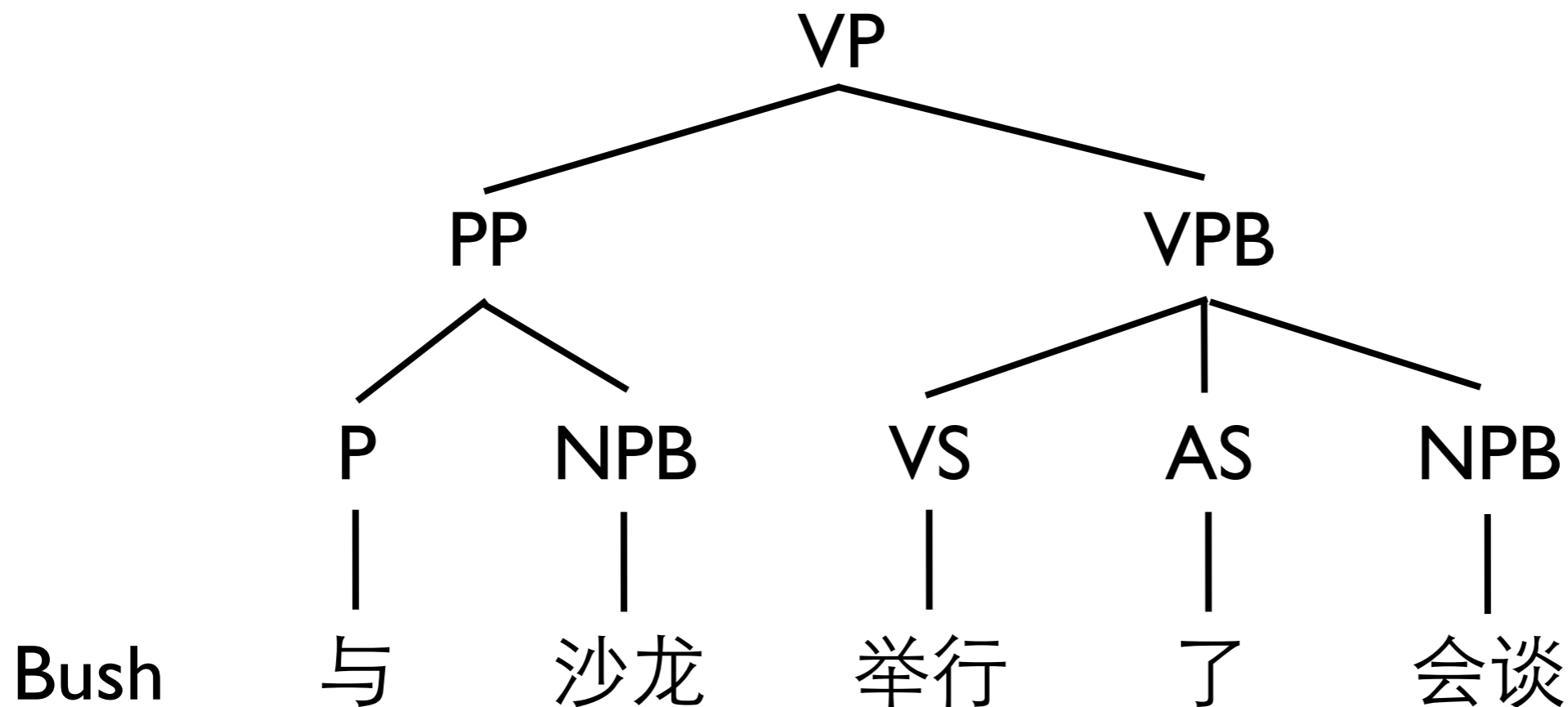
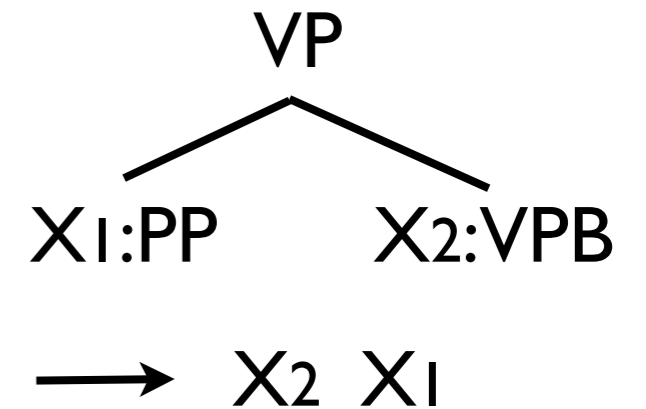
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



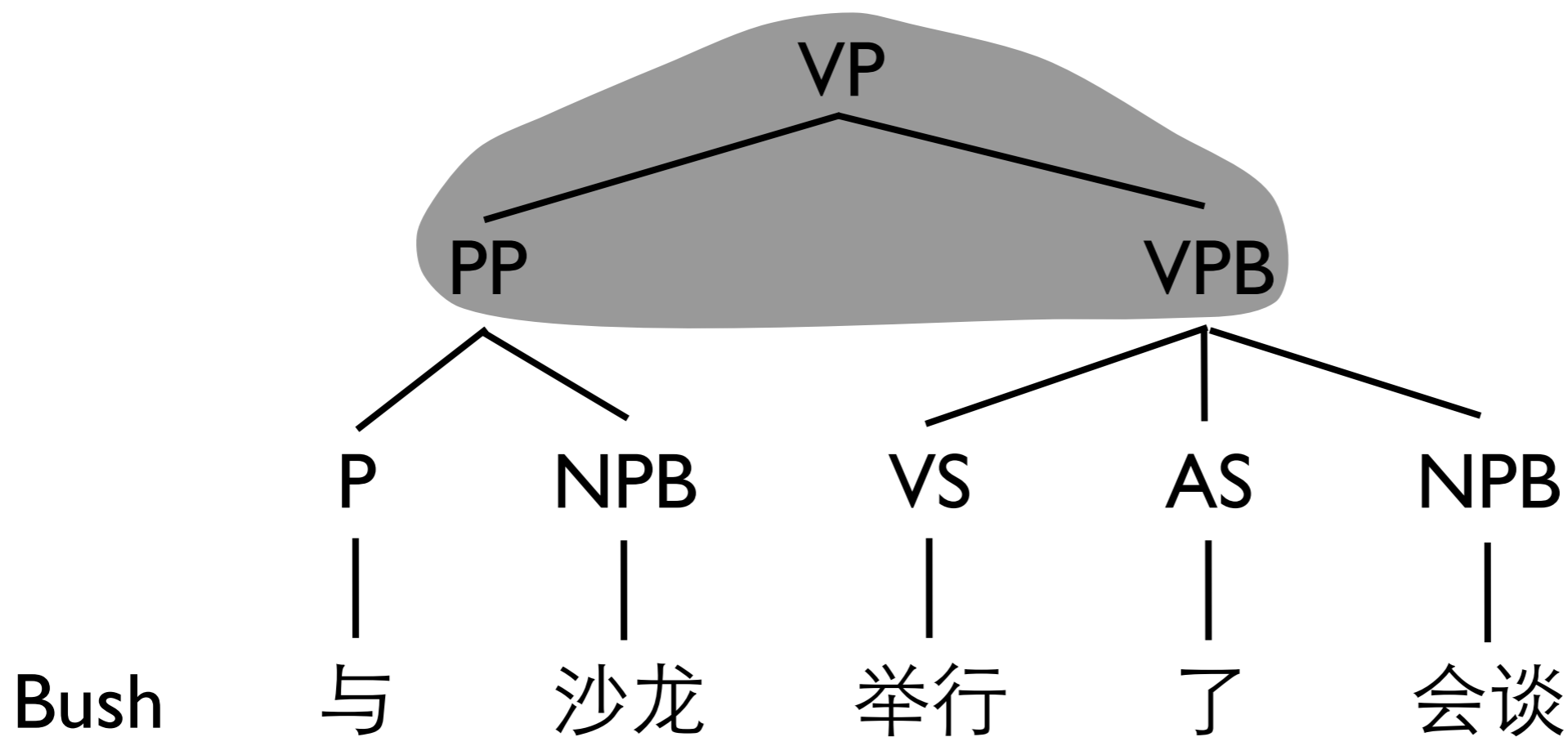
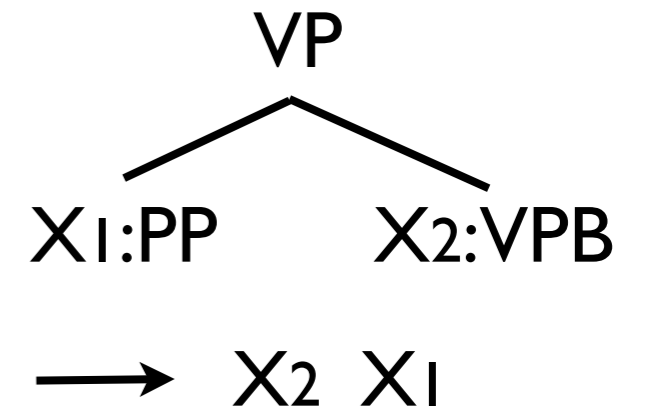
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



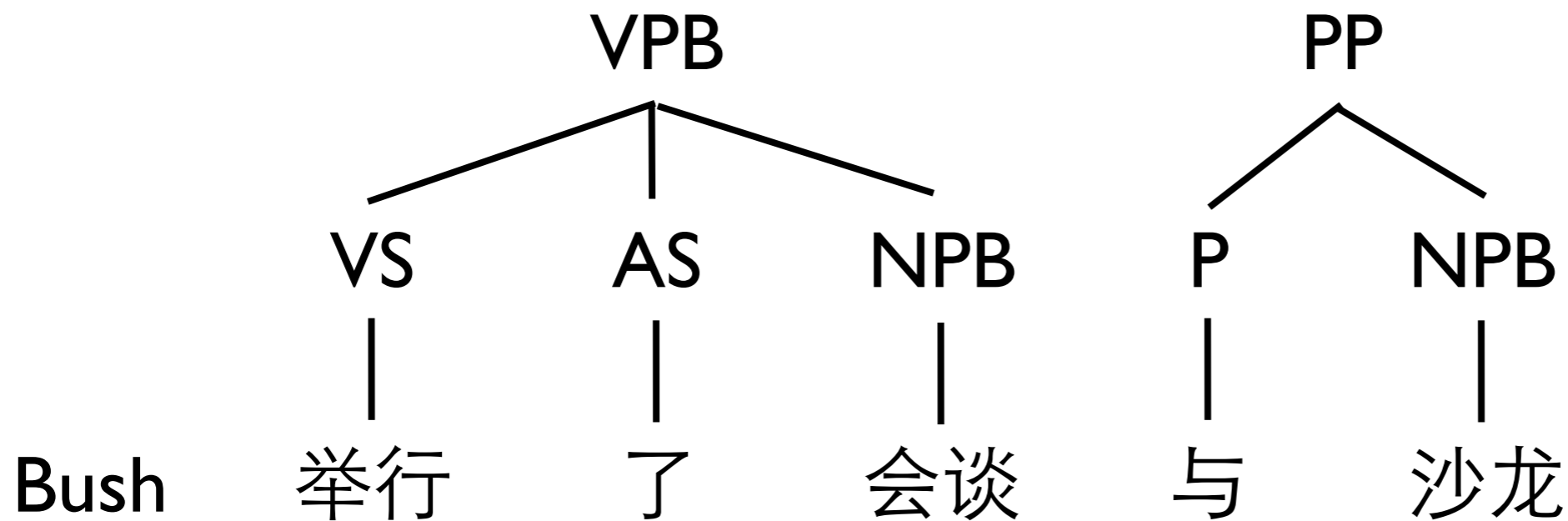
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



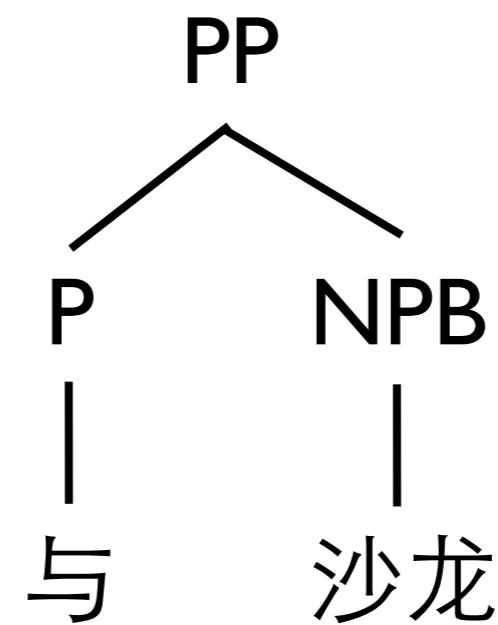
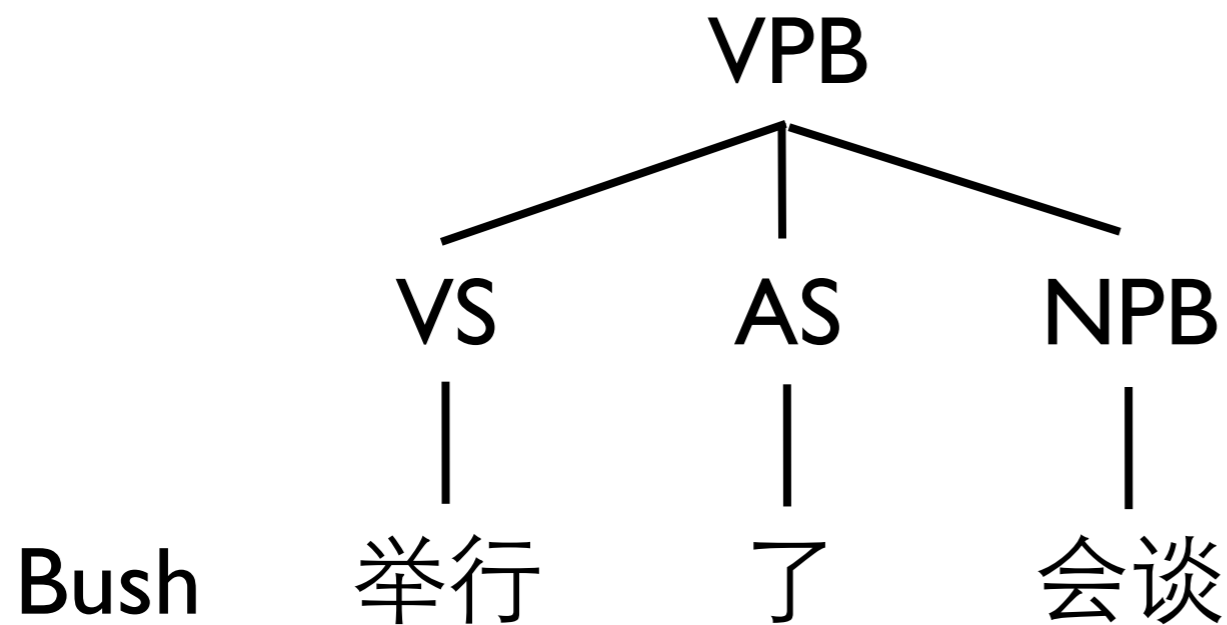
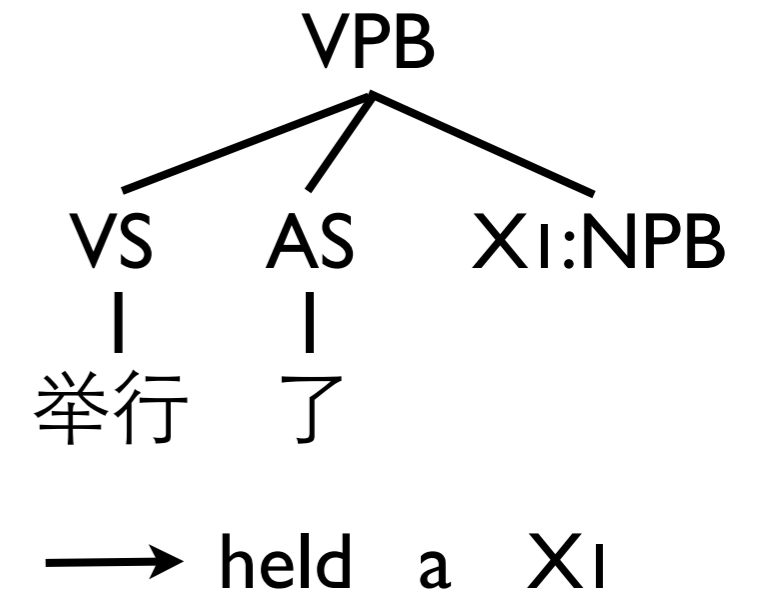
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



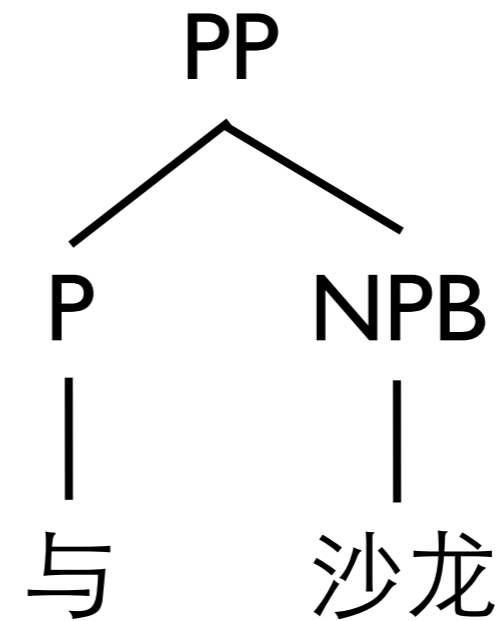
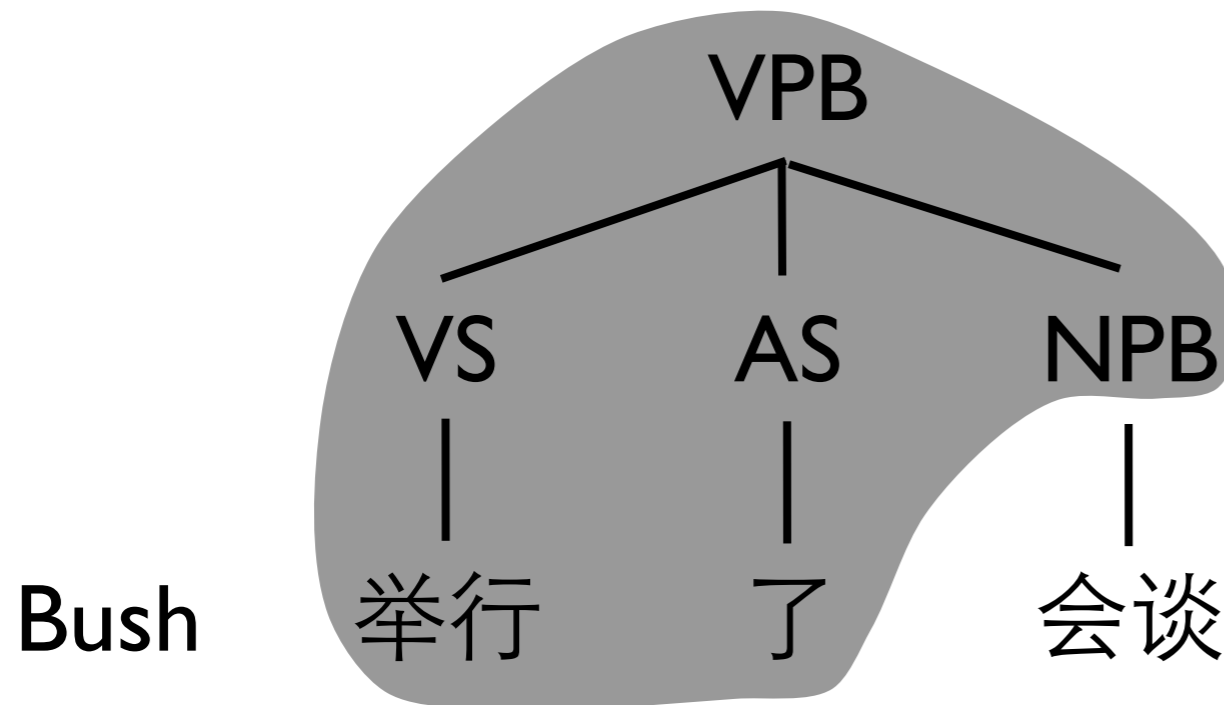
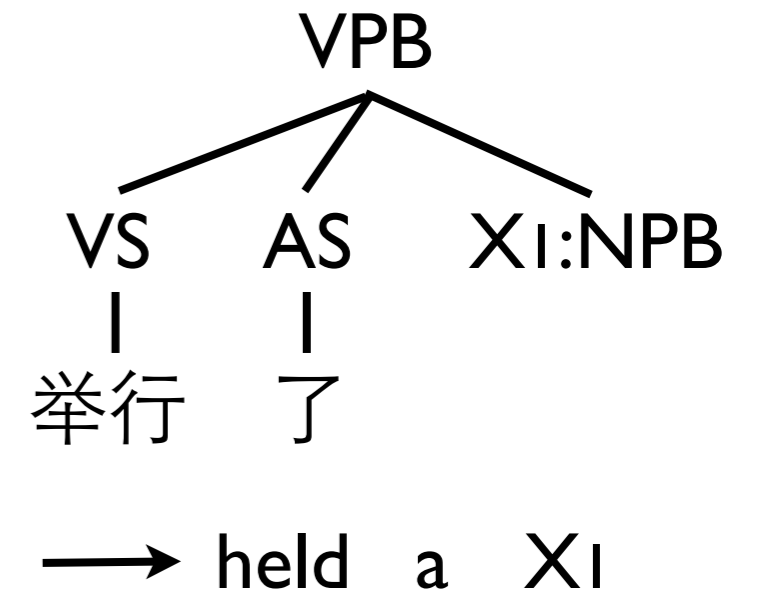
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



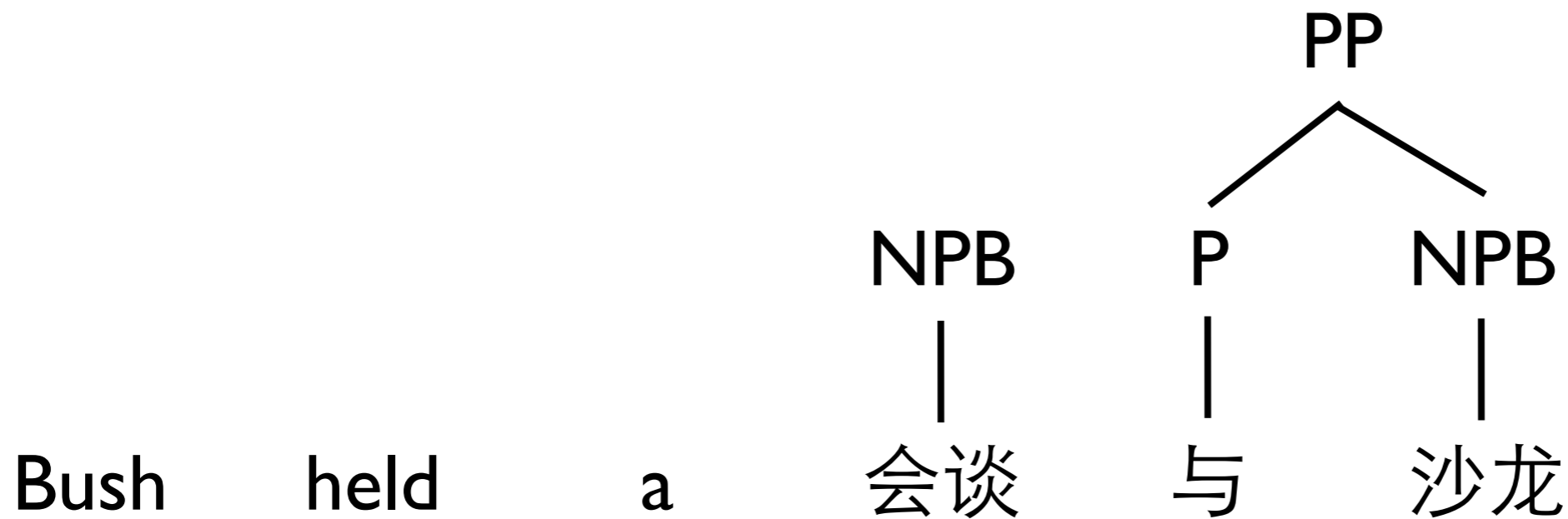
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



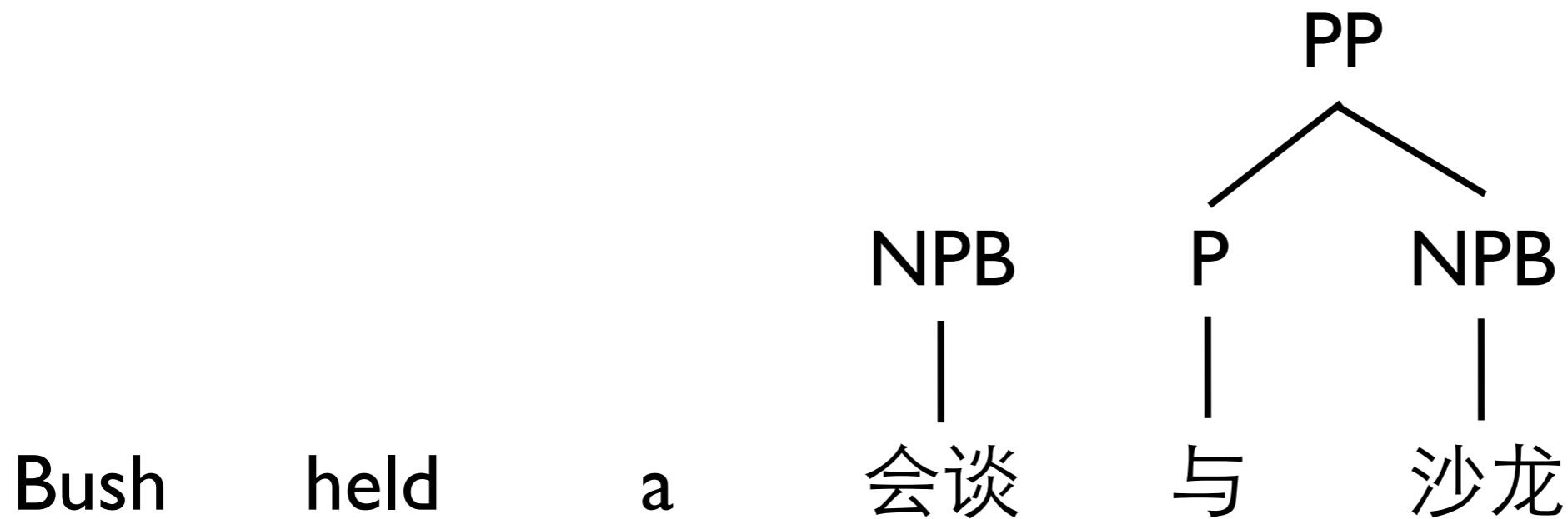
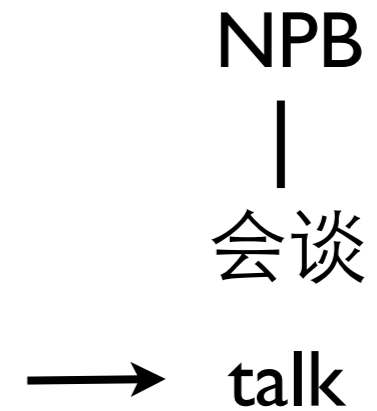
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



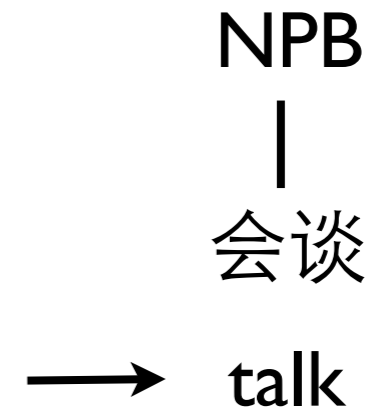
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



(Liu et al., 2006; Huang et al., 2006)

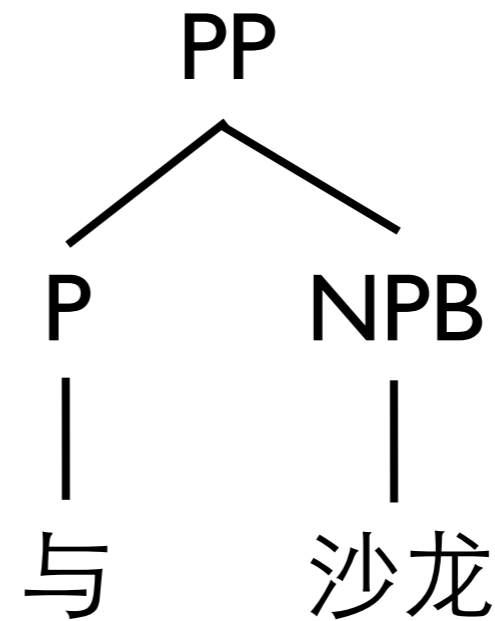
Tree-to-String Translation



Bush

held

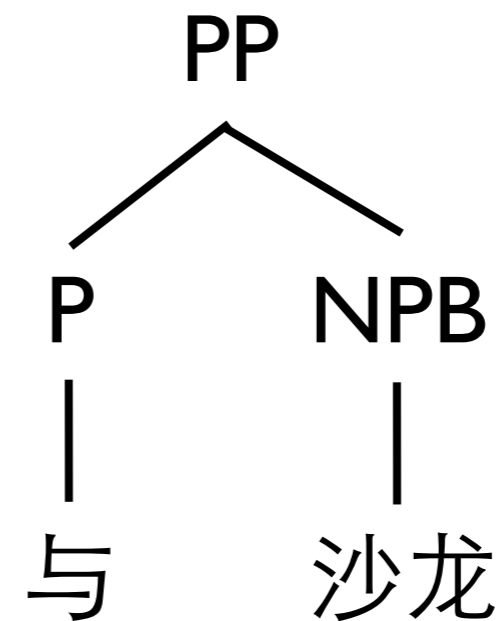
a



(Liu et al., 2006; Huang et al., 2006)

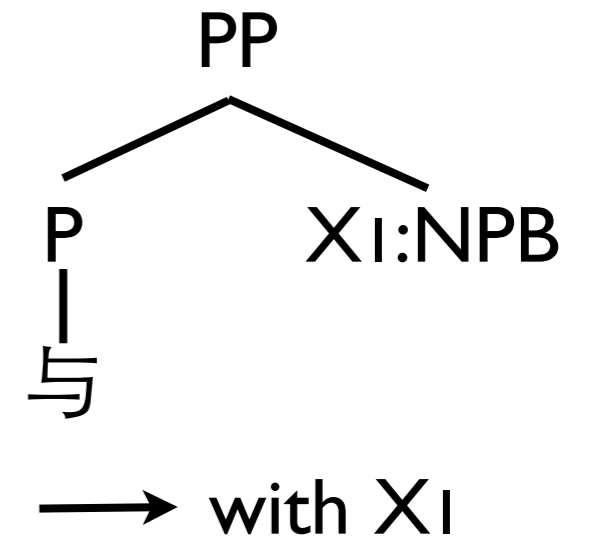
Tree-to-String Translation

Bush held a talk

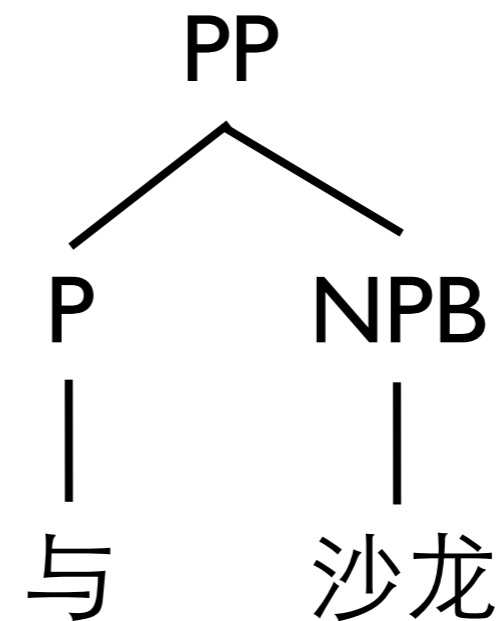


(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation

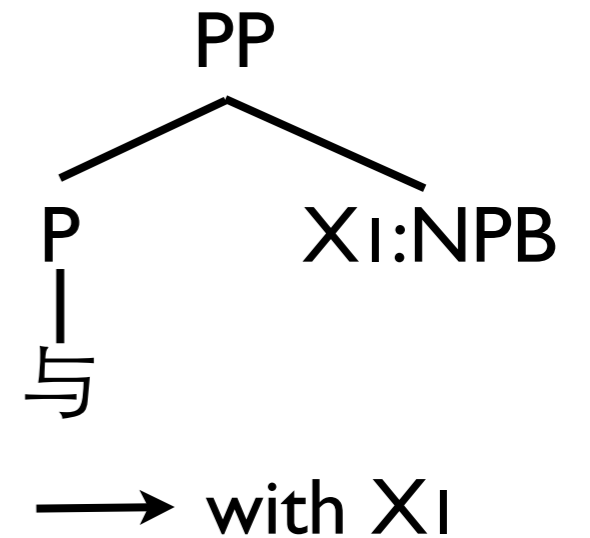


Bush held a talk

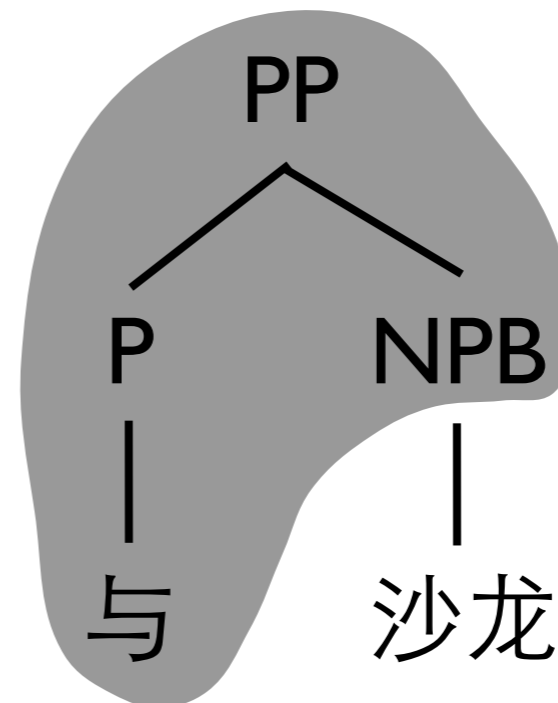


(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



Bush held a talk



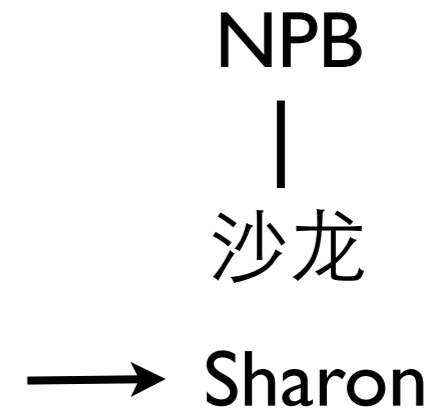
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation

Bush held a talk with NPB
沙龙

(Liu et al., 2006; Huang et al., 2006)

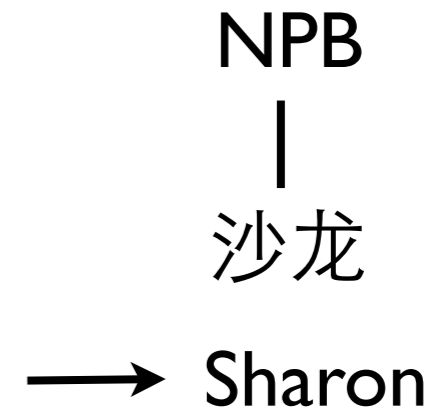
Tree-to-String Translation



Bush held a talk with NPB
沙龙

(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



Bush held a talk with



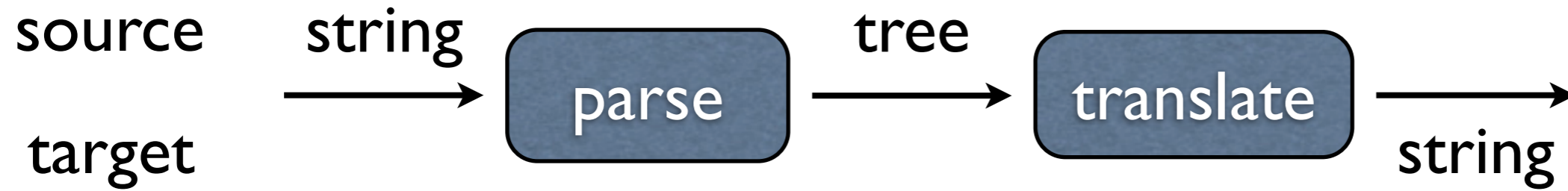
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation

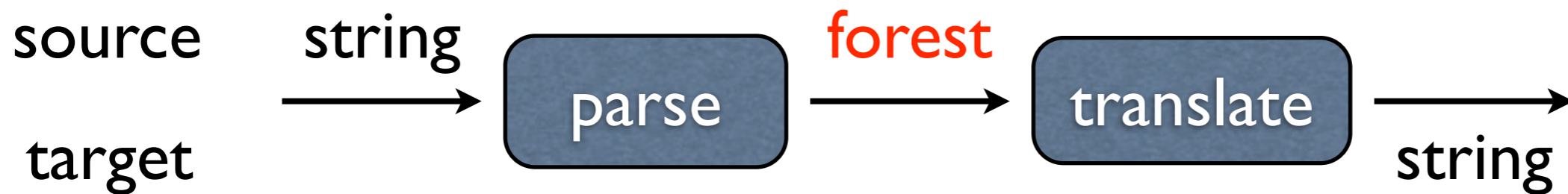
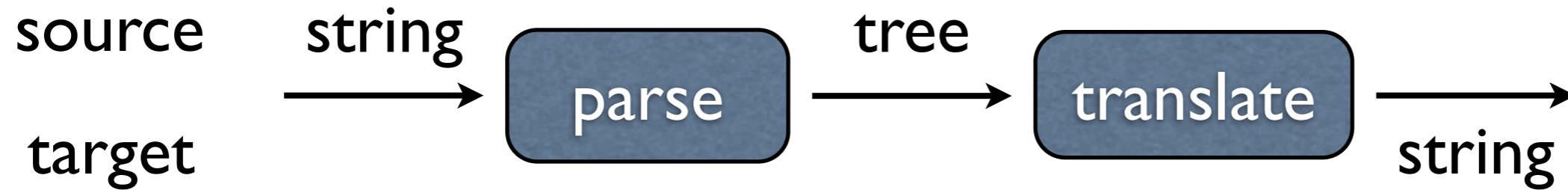
Bush held a talk with Sharon

(Liu et al., 2006; Huang et al., 2006)

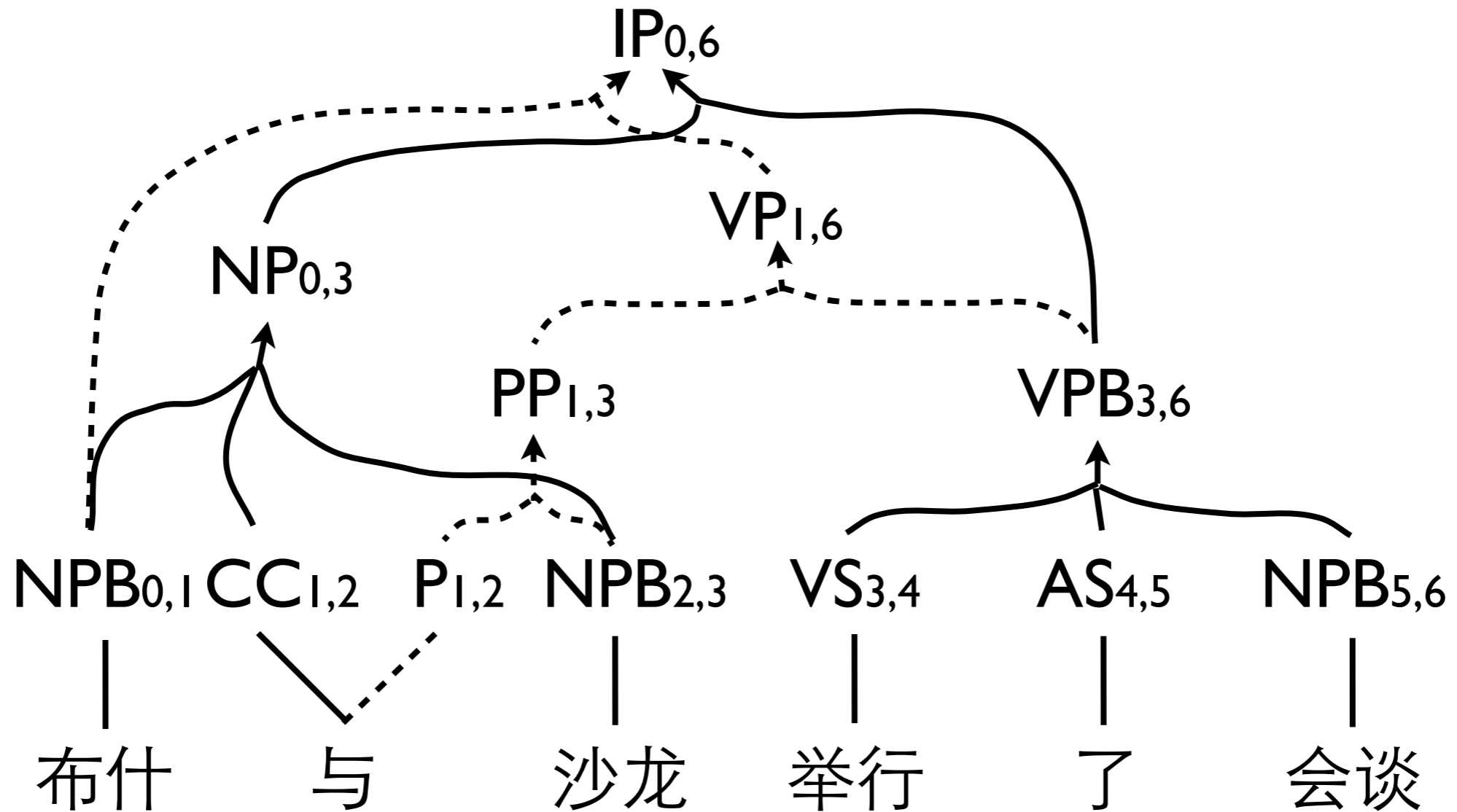
Parsing Mistake Propagation



Parsing Mistake Propagation

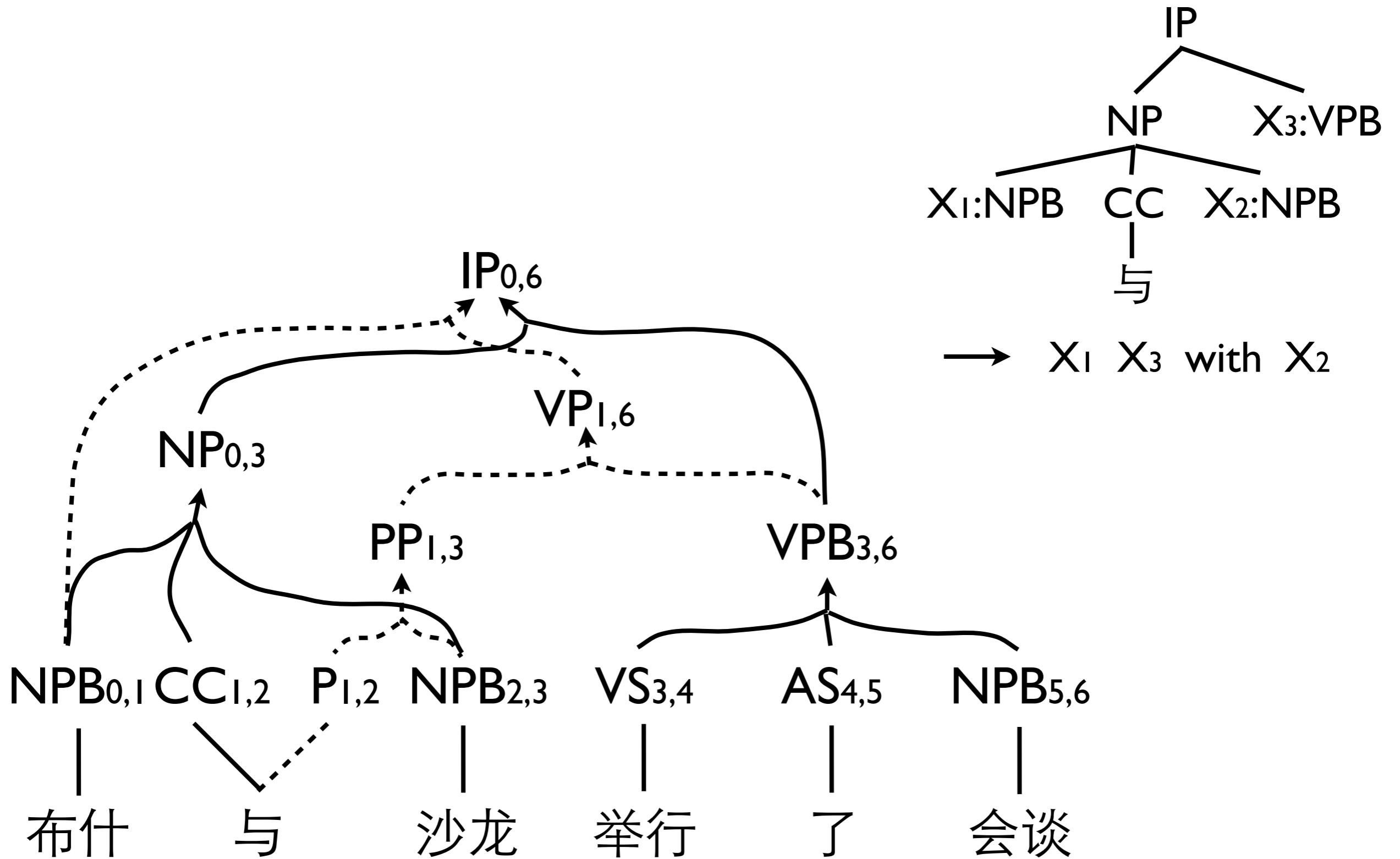


Packed Forest

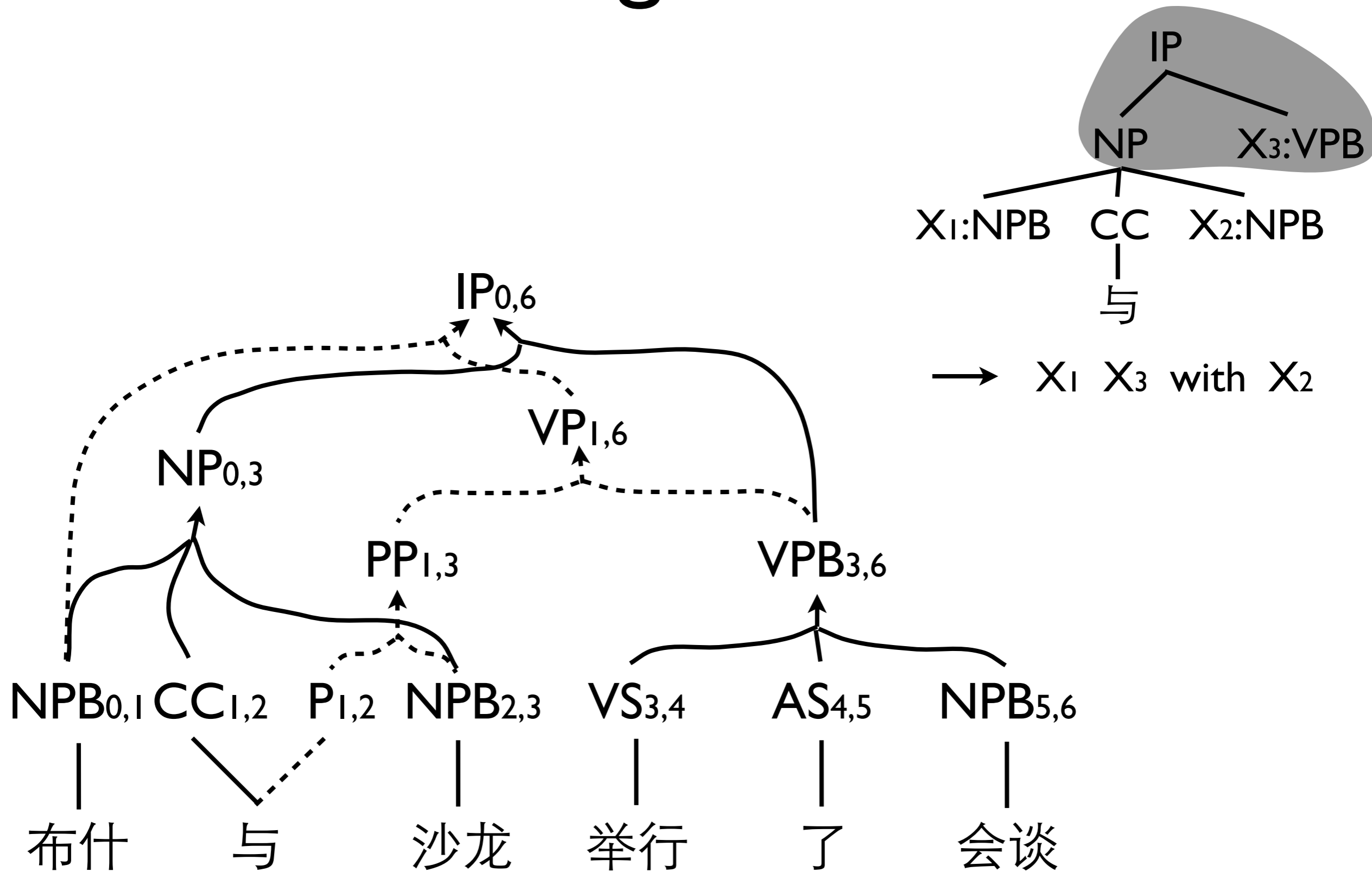


(Billot and Lang, 1989)

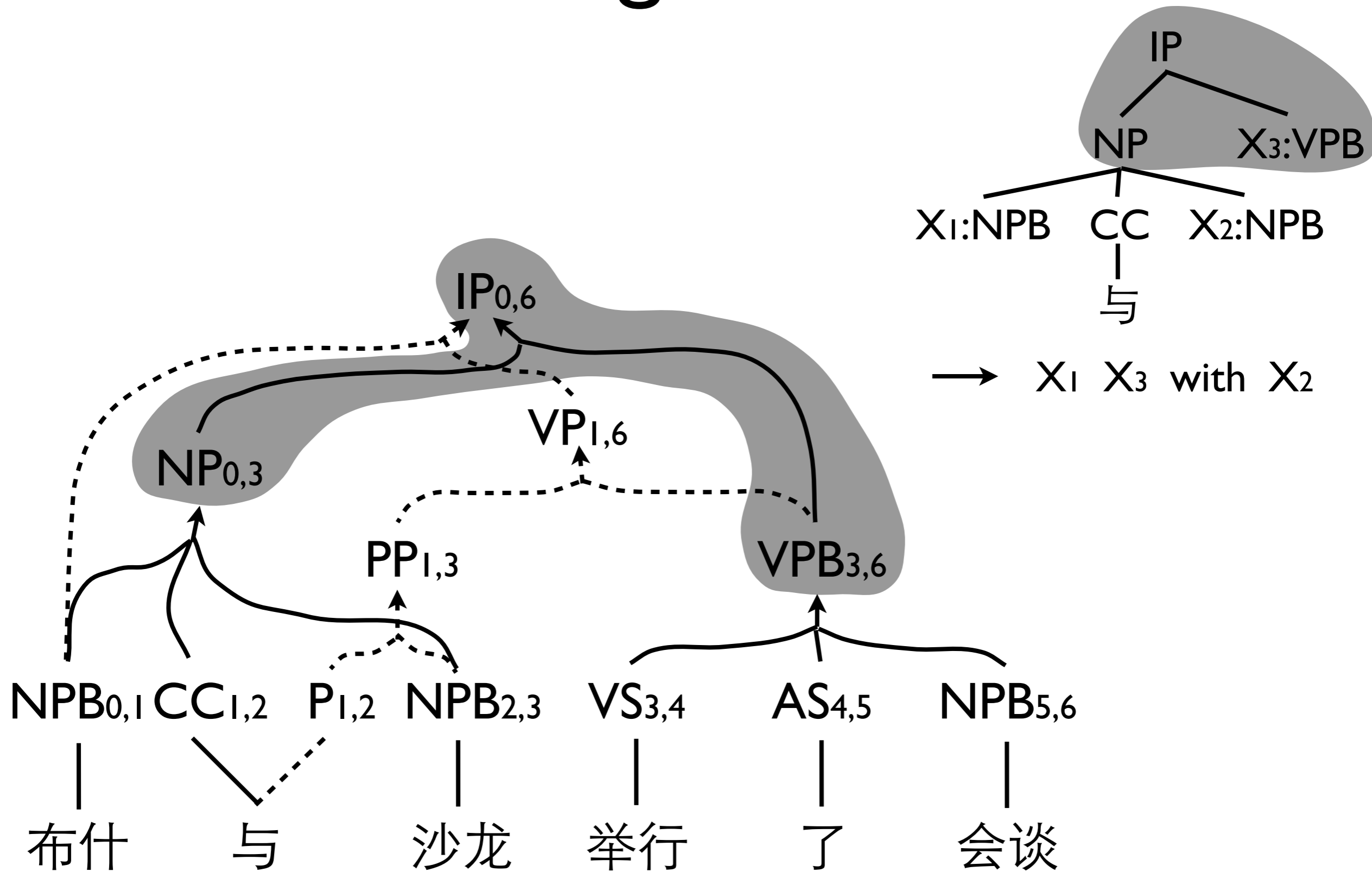
Matching on Forest



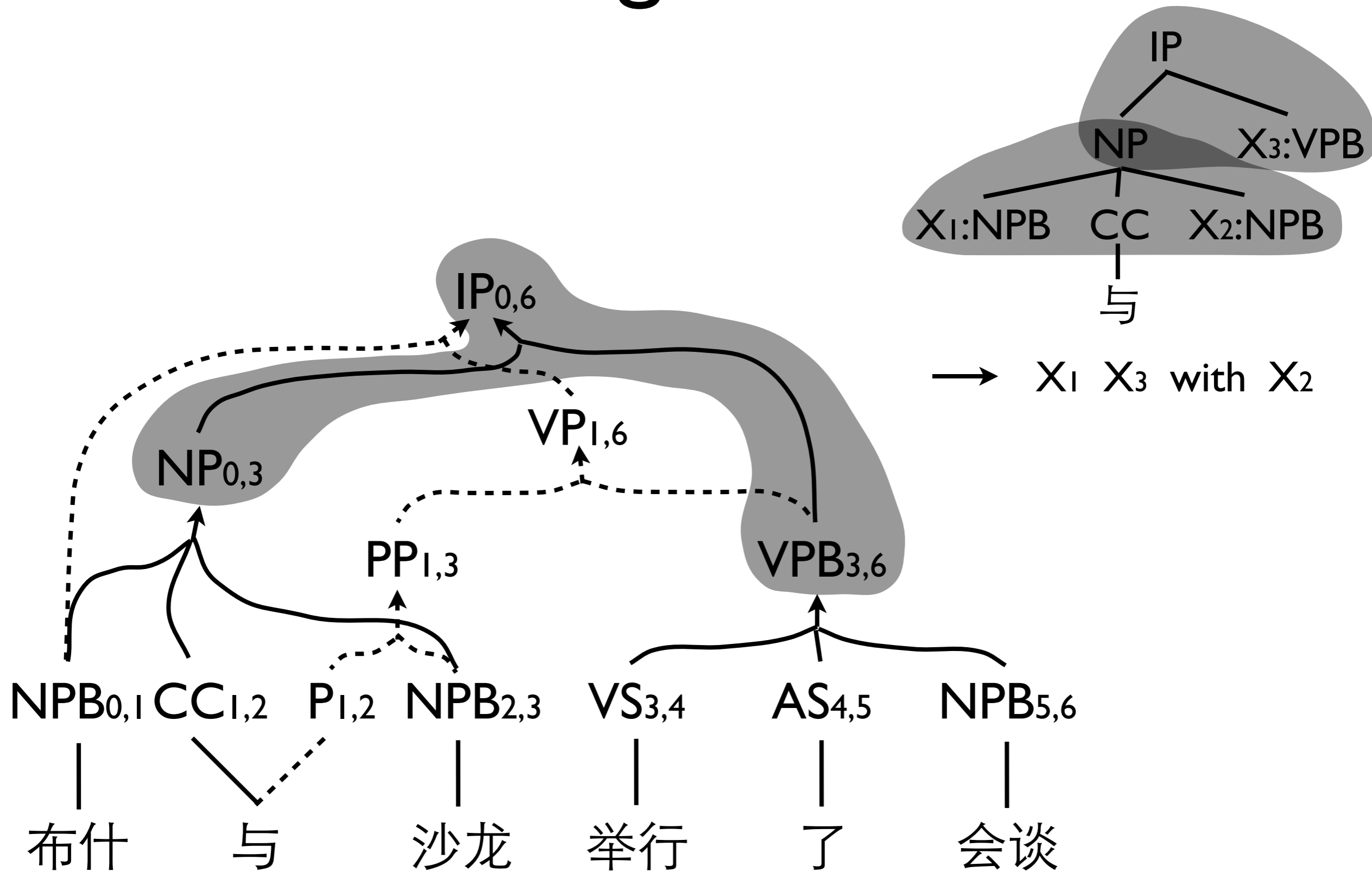
Matching on Forest



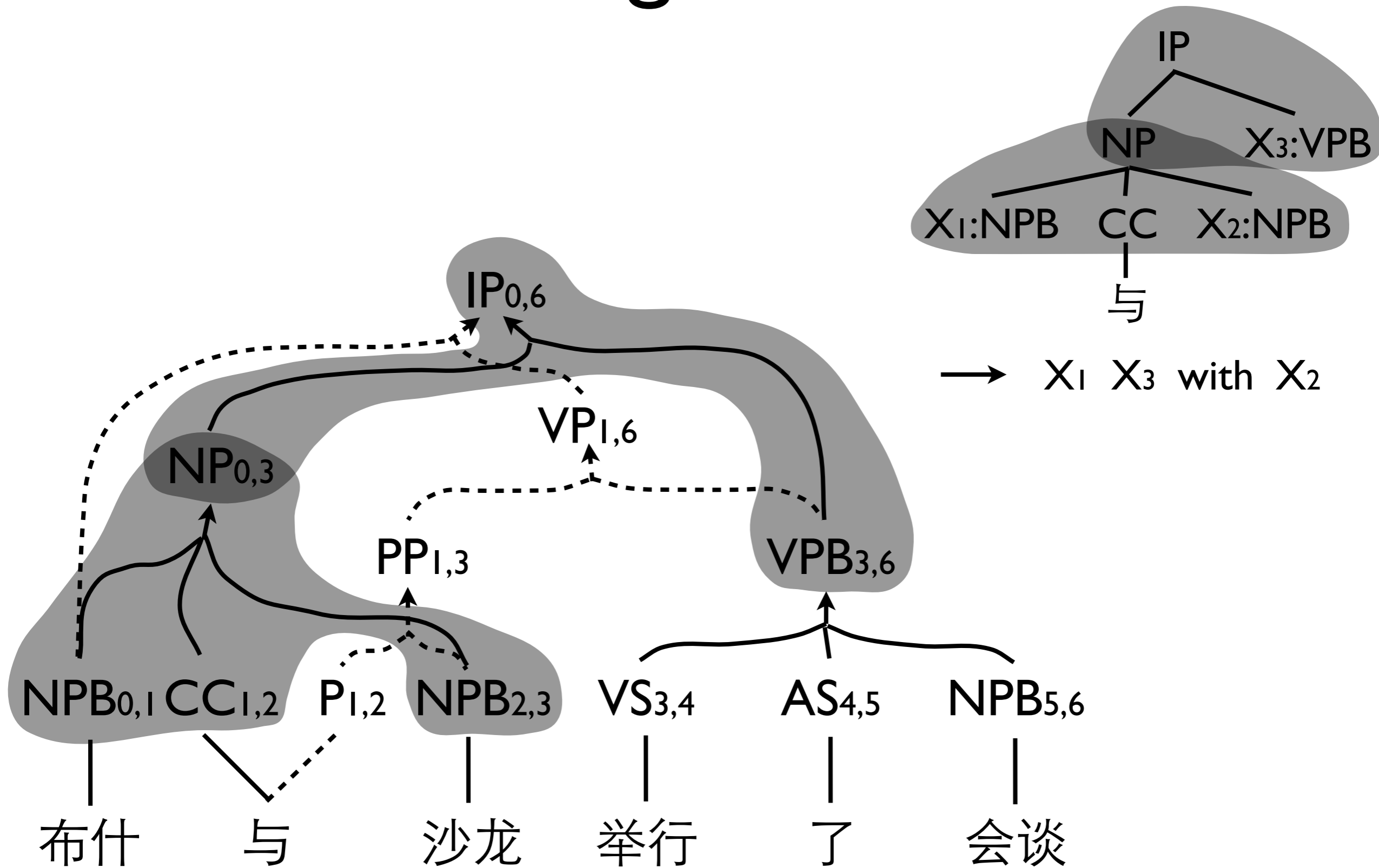
Matching on Forest



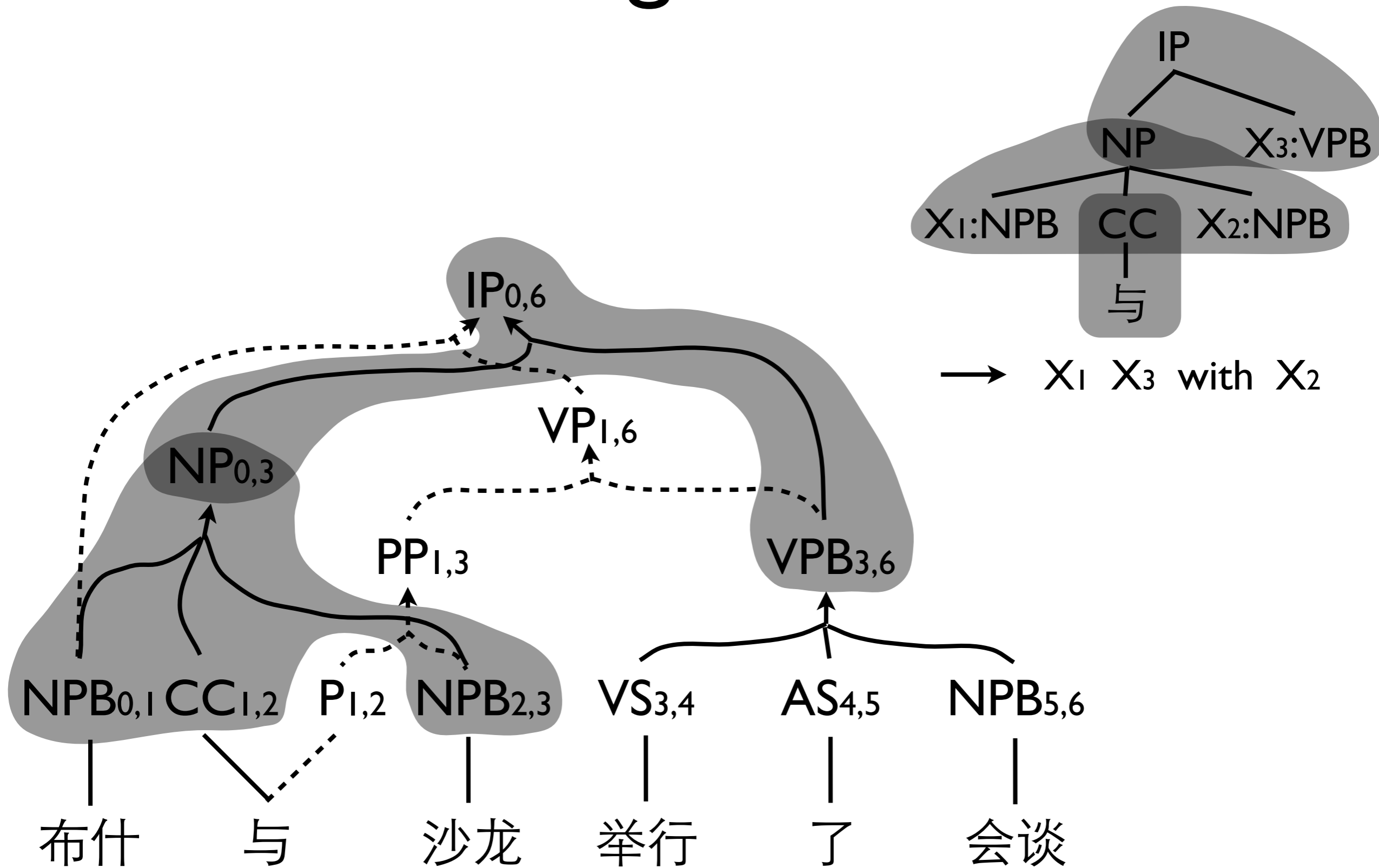
Matching on Forest



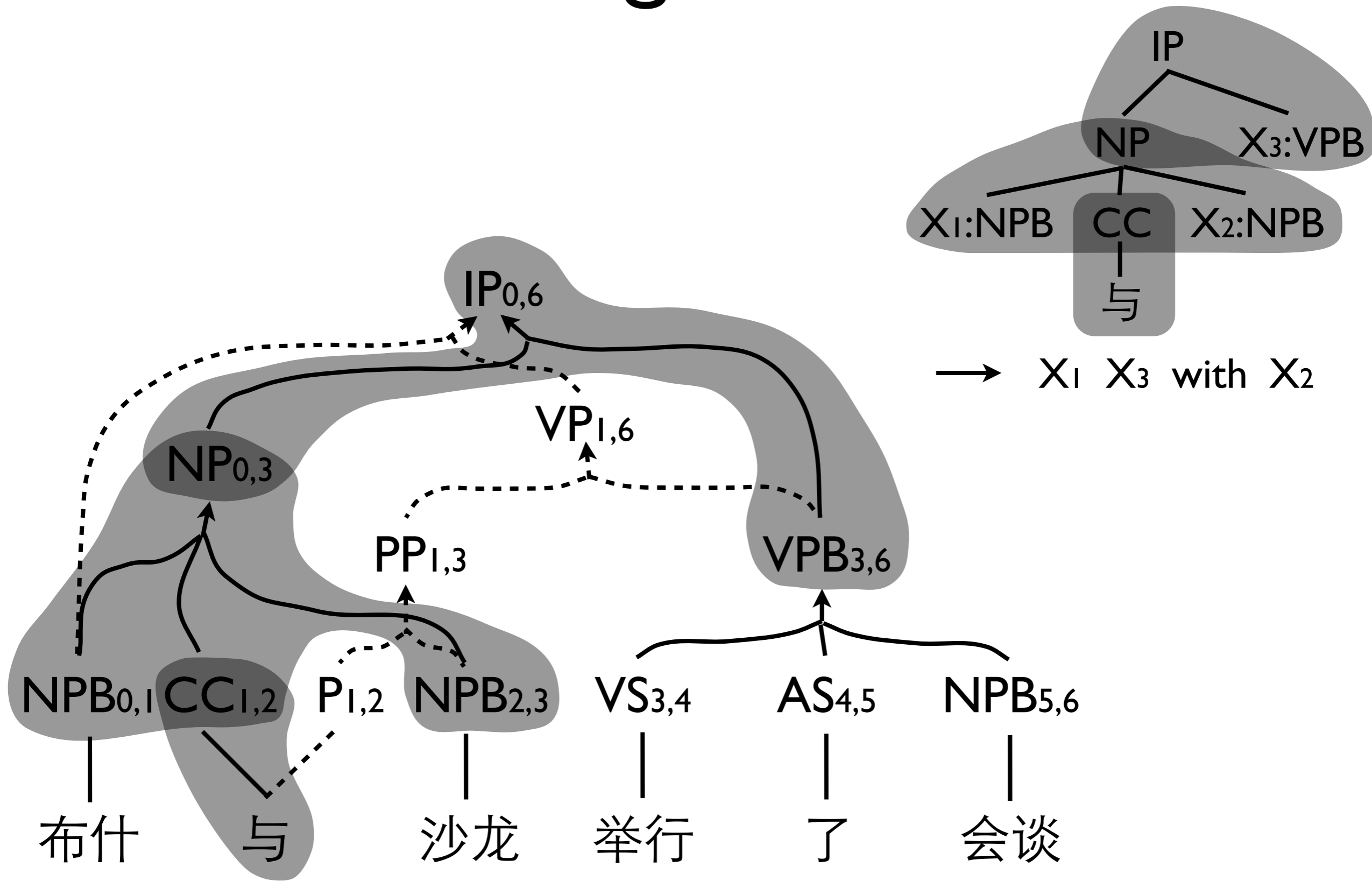
Matching on Forest



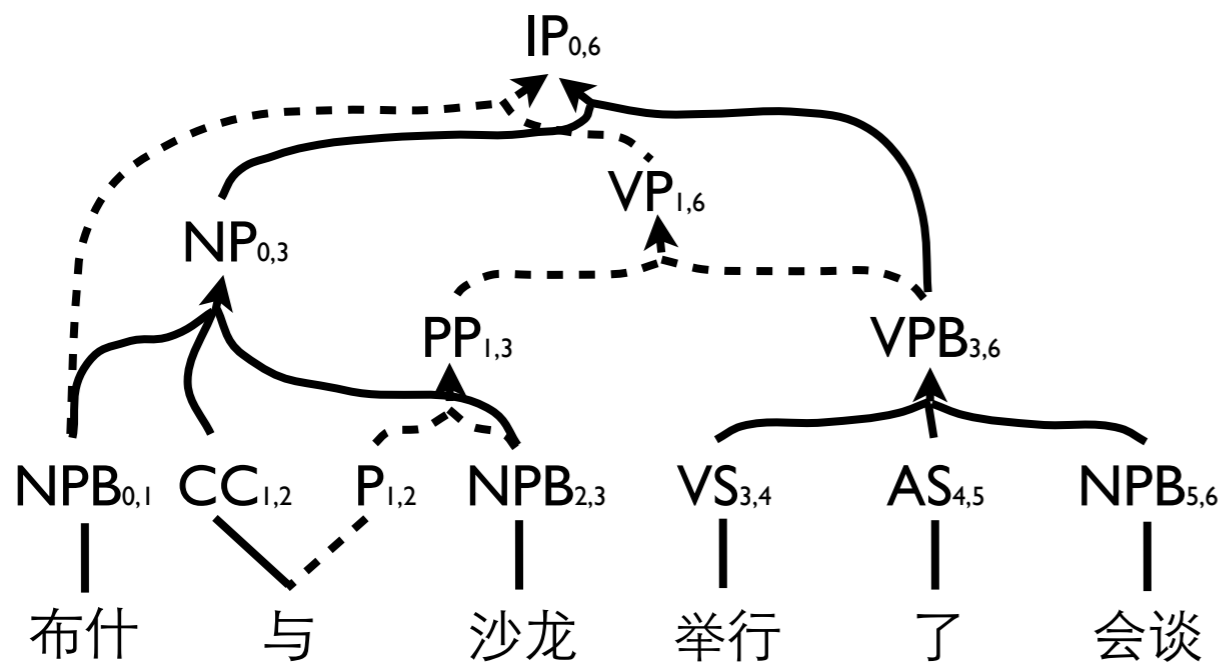
Matching on Forest



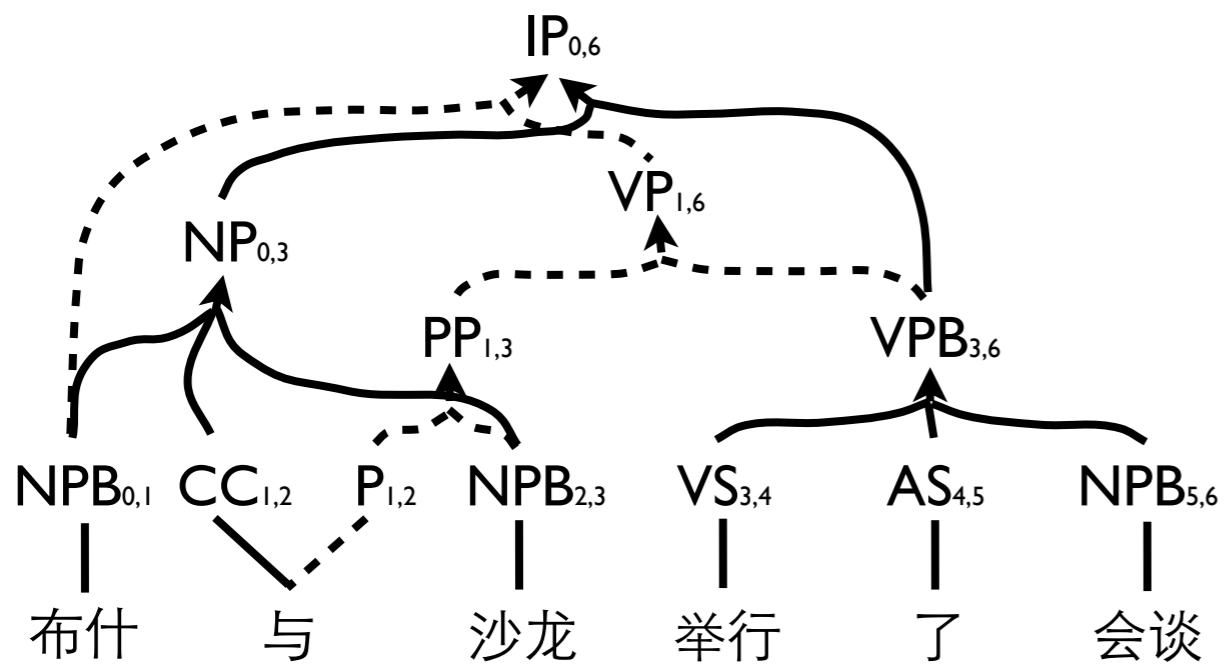
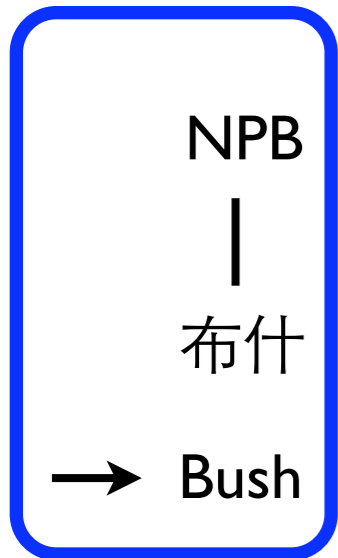
Matching on Forest



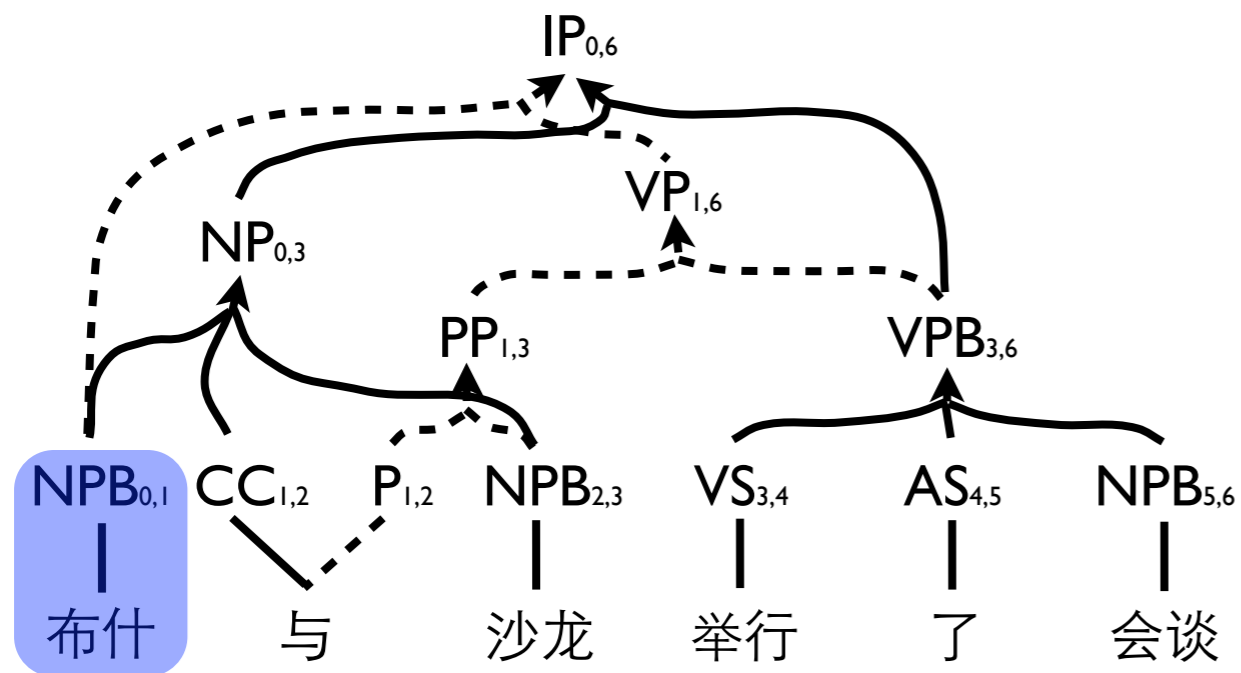
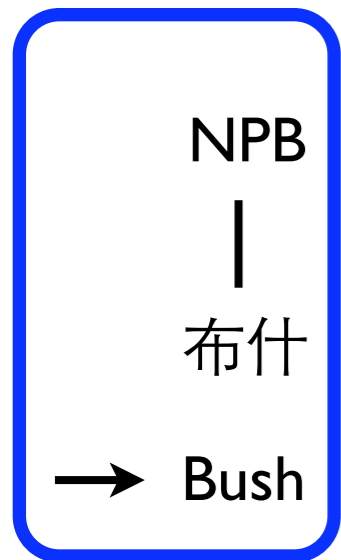
Translation Forest



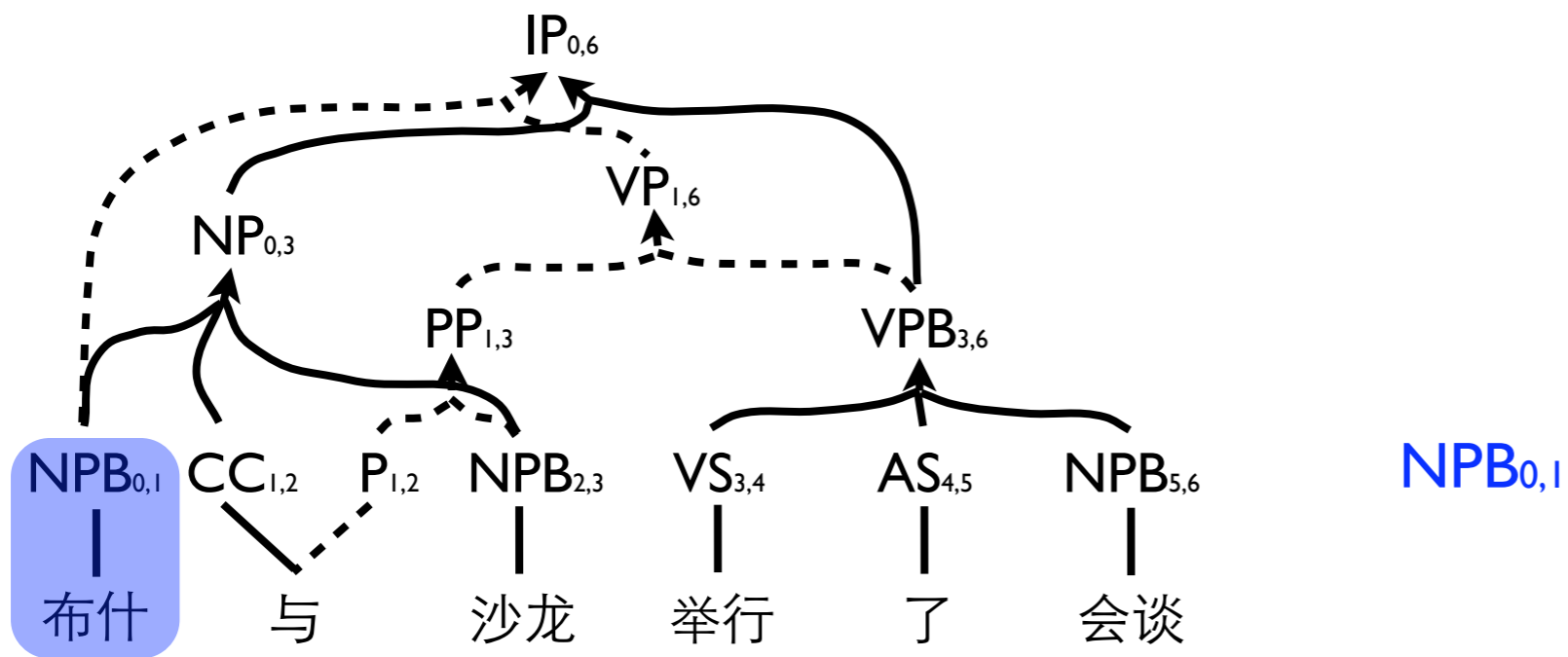
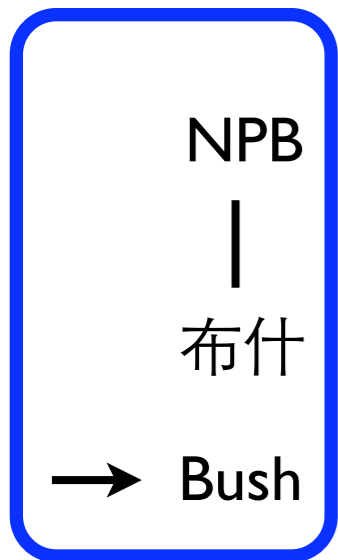
Translation Forest



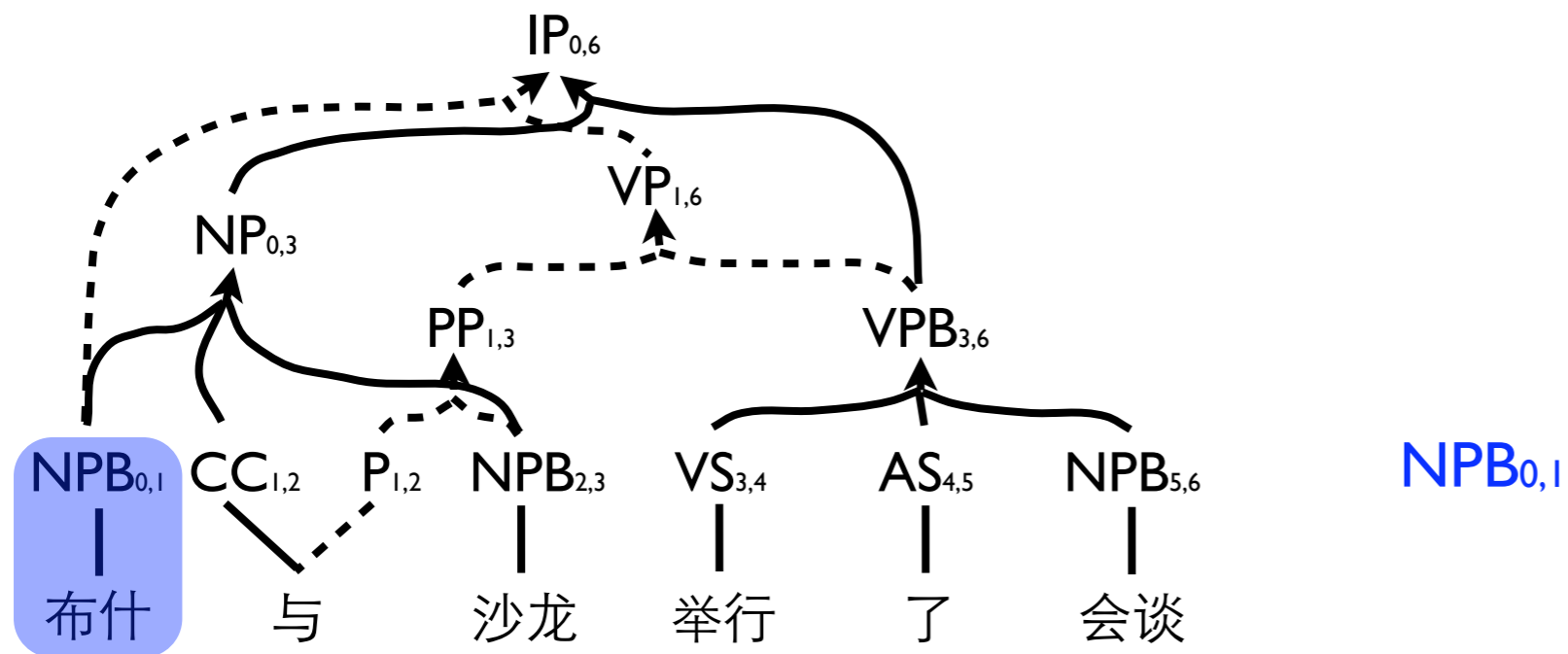
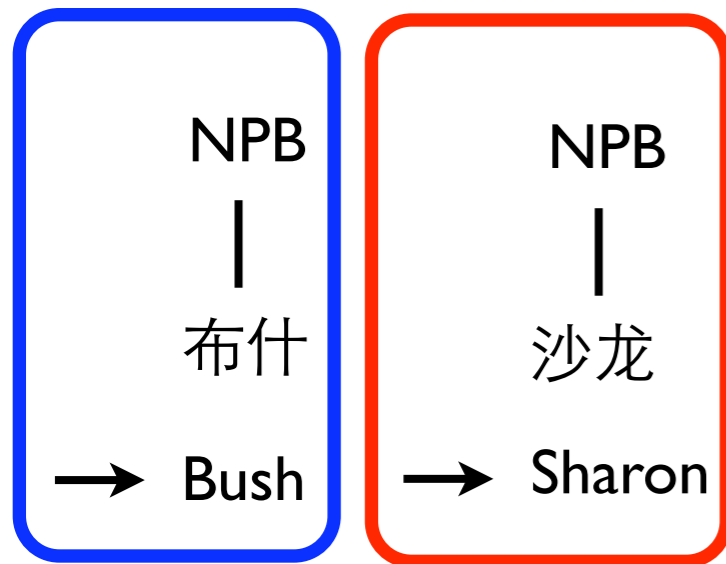
Translation Forest



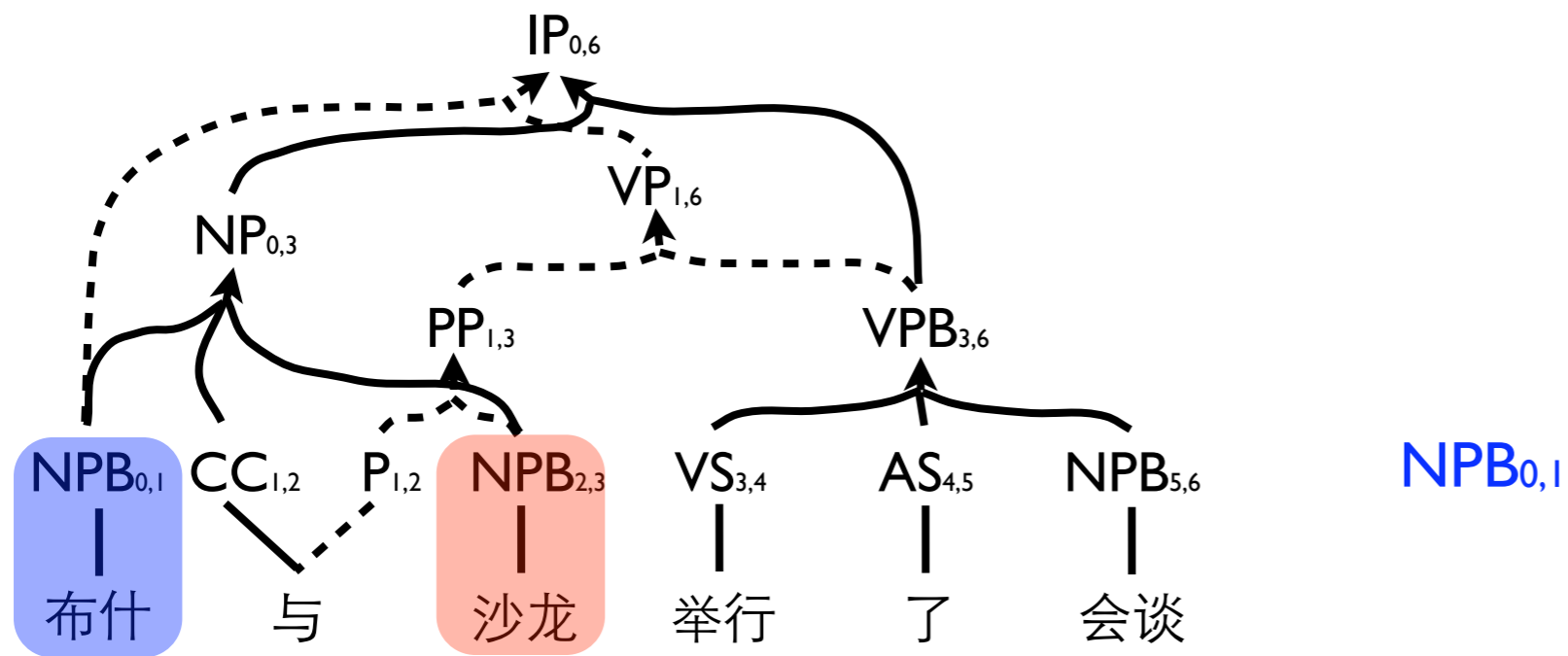
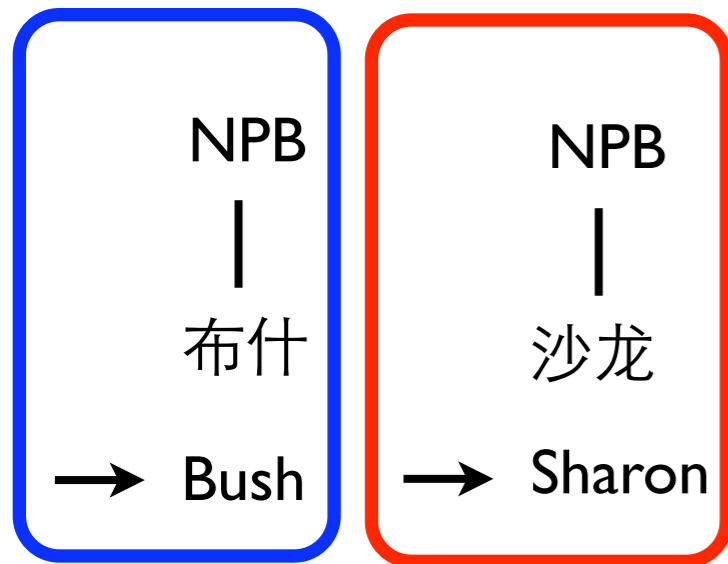
Translation Forest



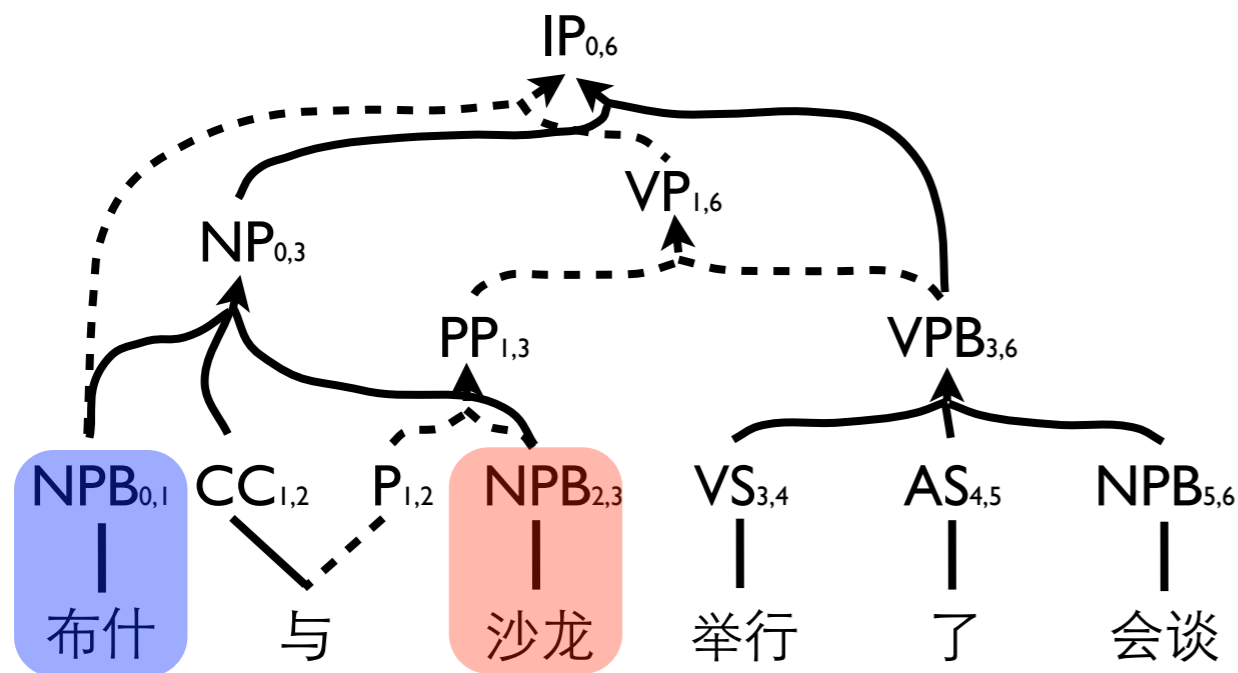
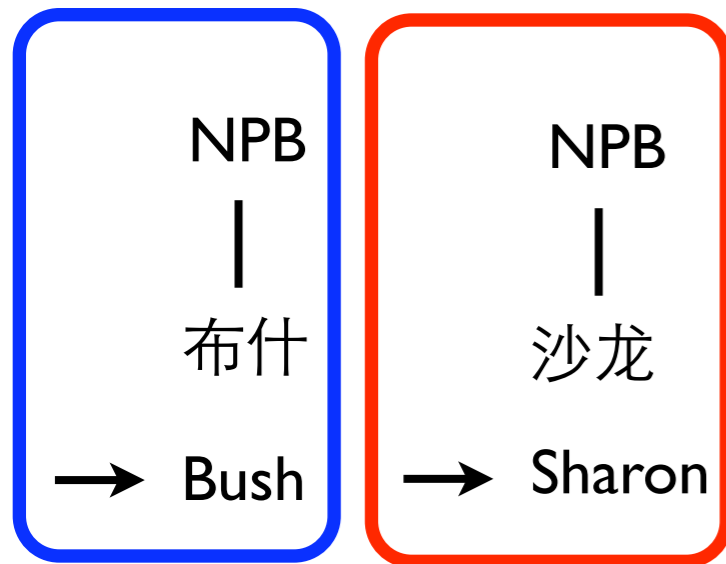
Translation Forest



Translation Forest



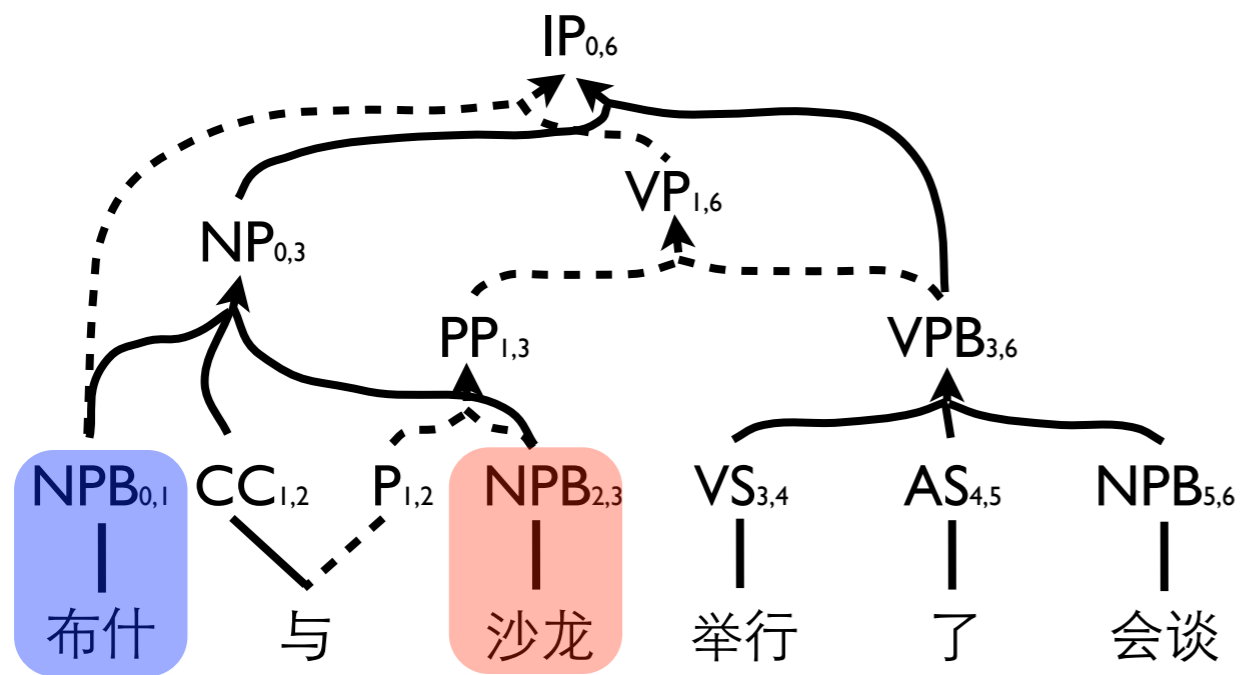
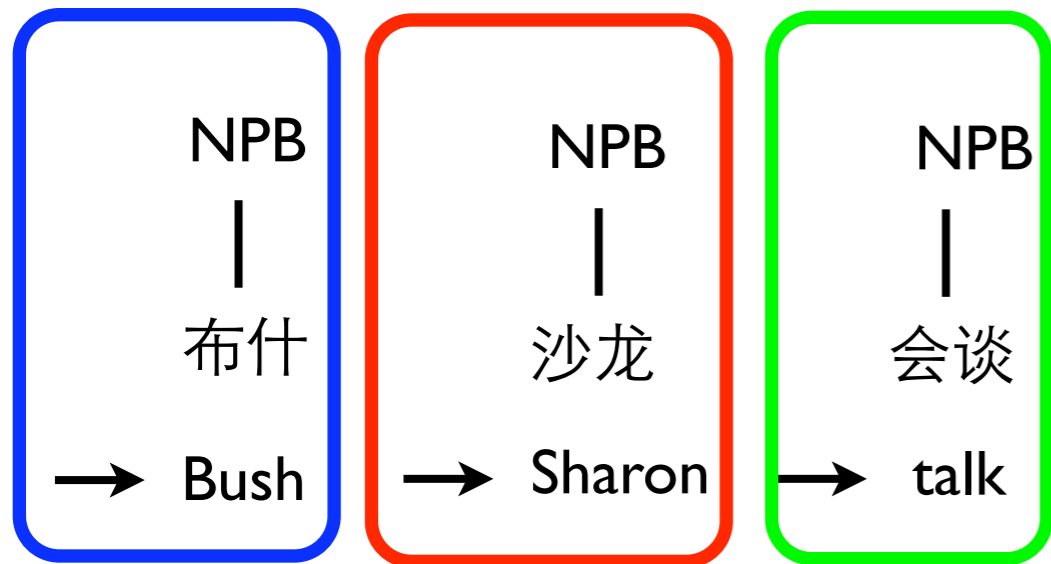
Translation Forest



NPB_{0,1}

NPB_{2,3}

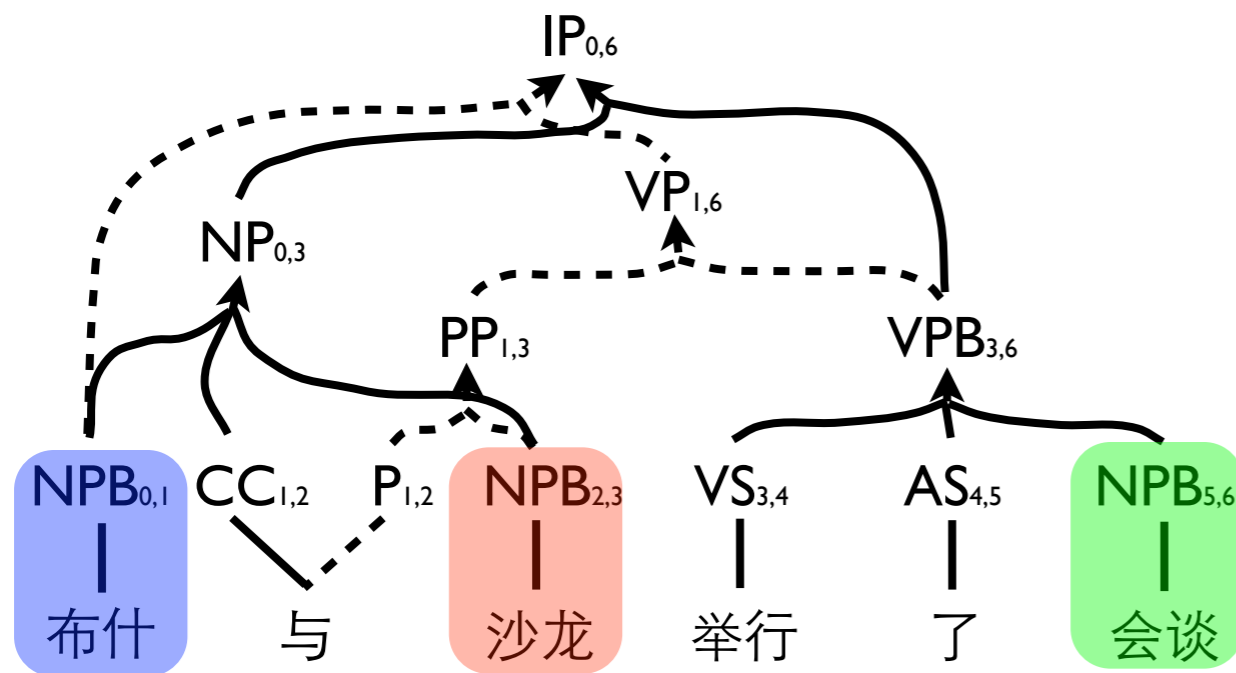
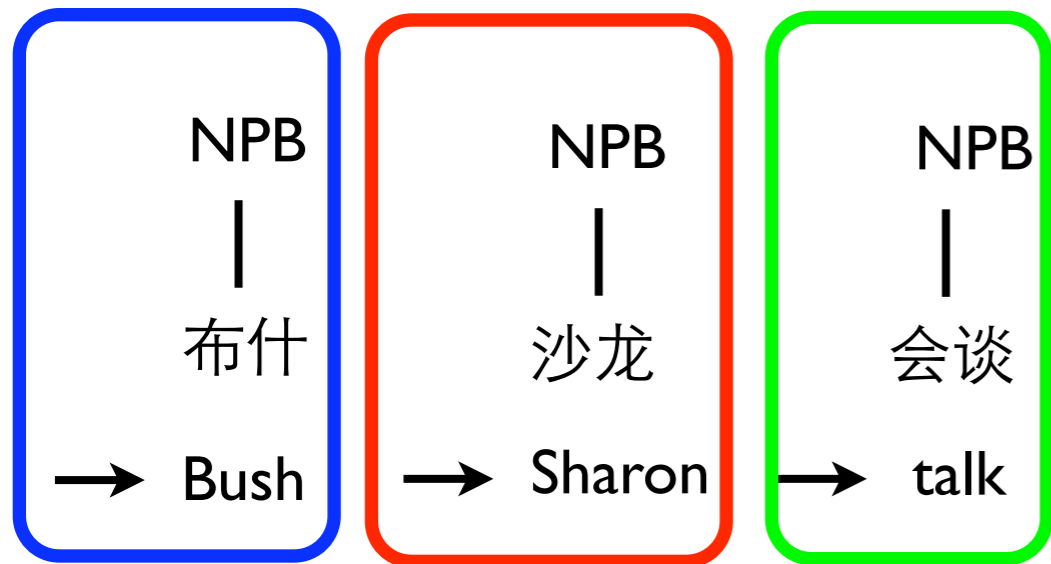
Translation Forest



NPB_{0,1}

NPB_{2,3}

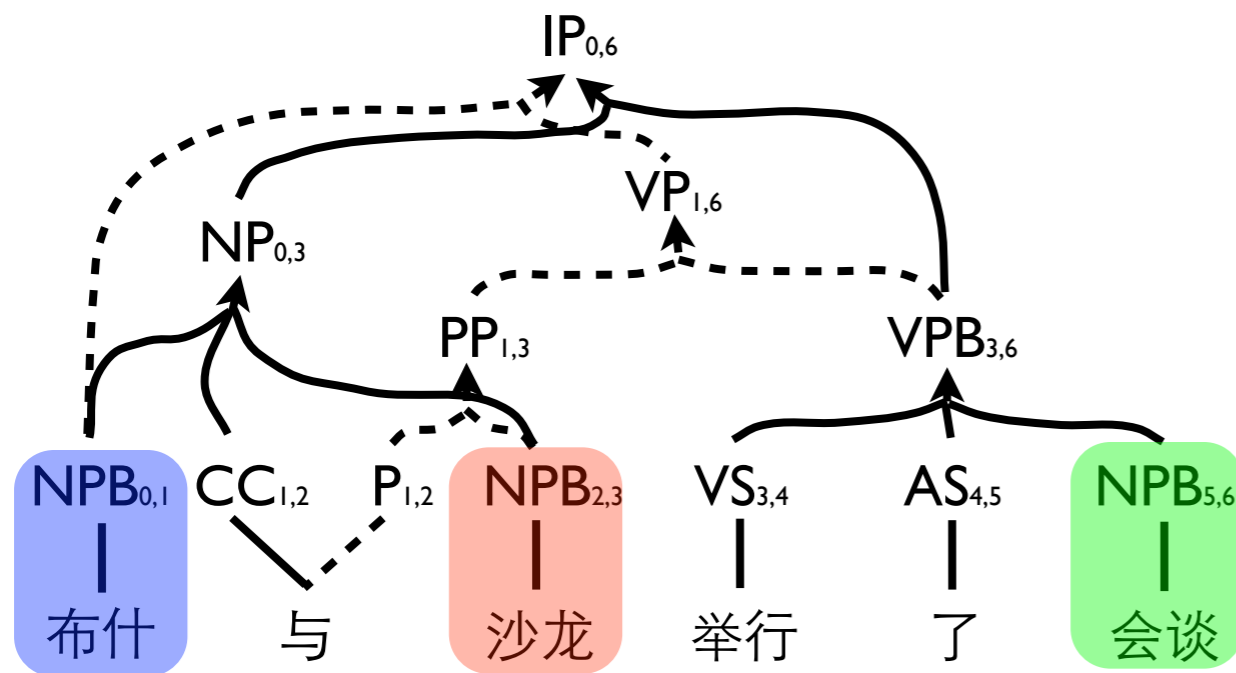
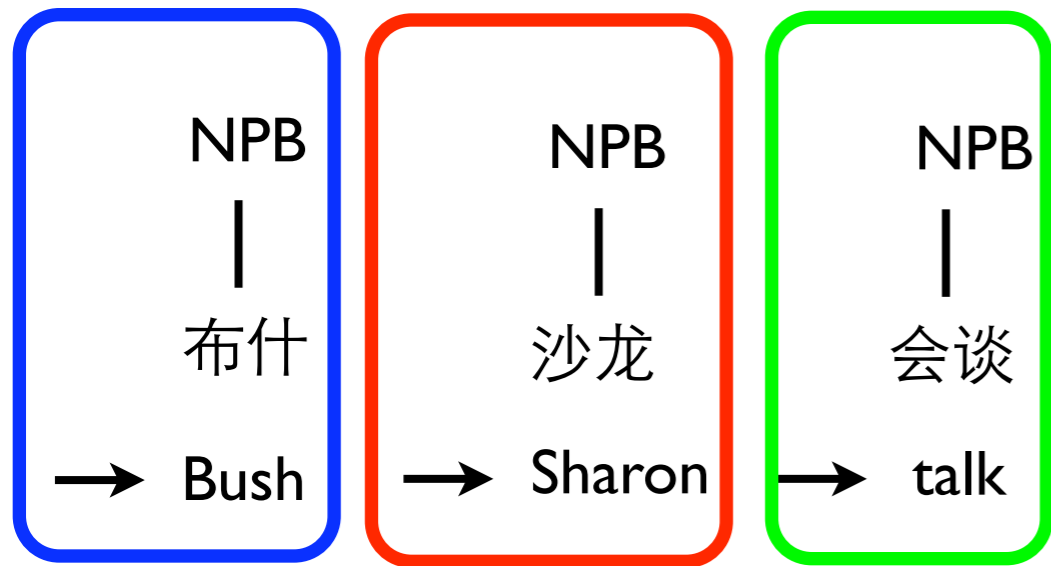
Translation Forest



NPB_{0,1}

NPB_{2,3}

Translation Forest

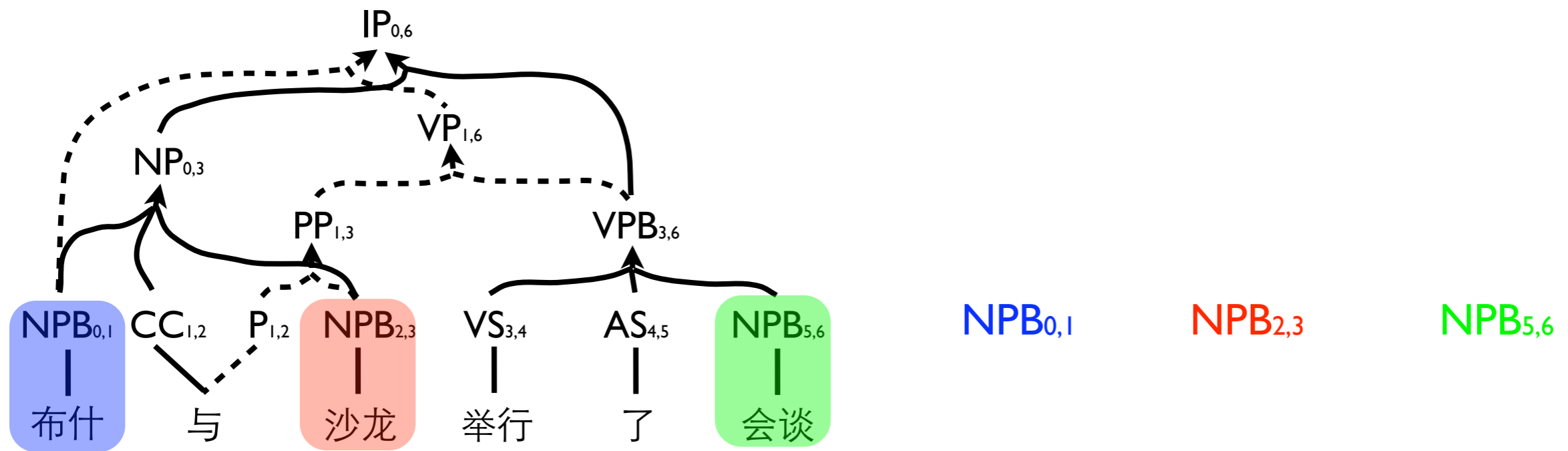
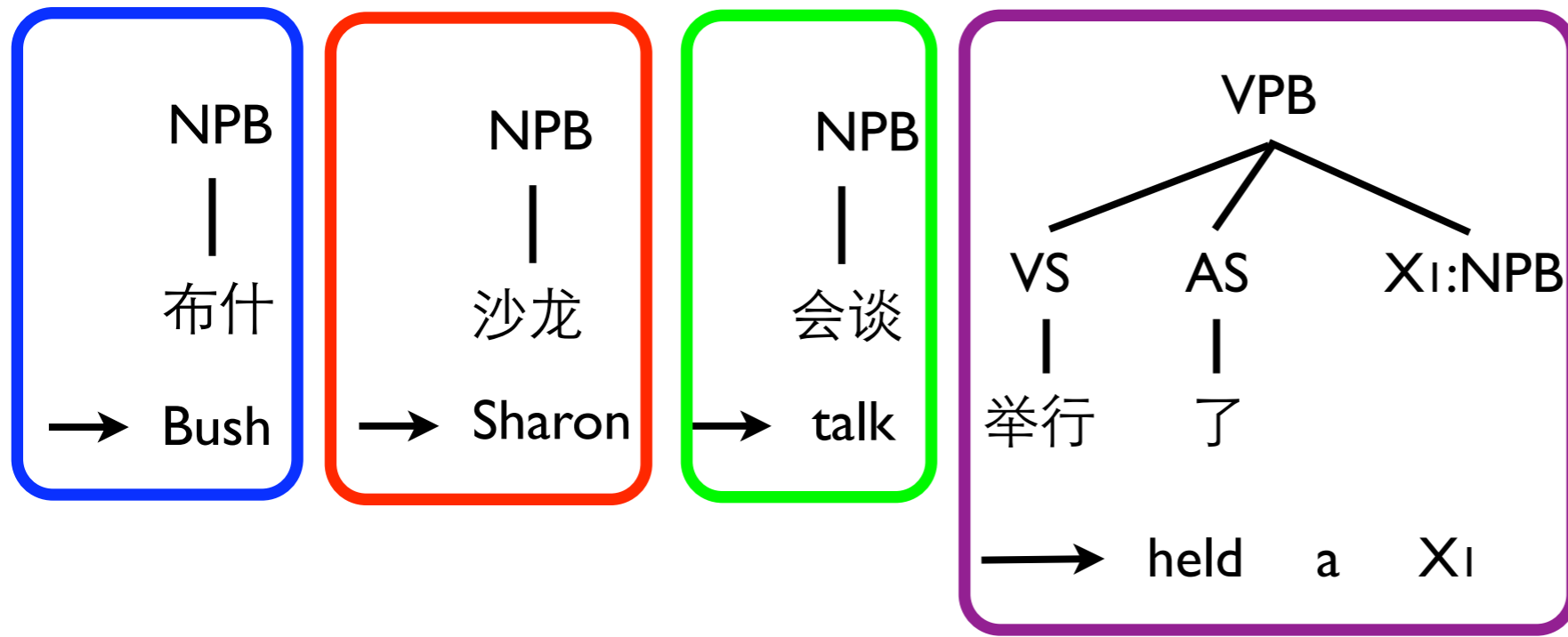


$NPB_{0,1}$

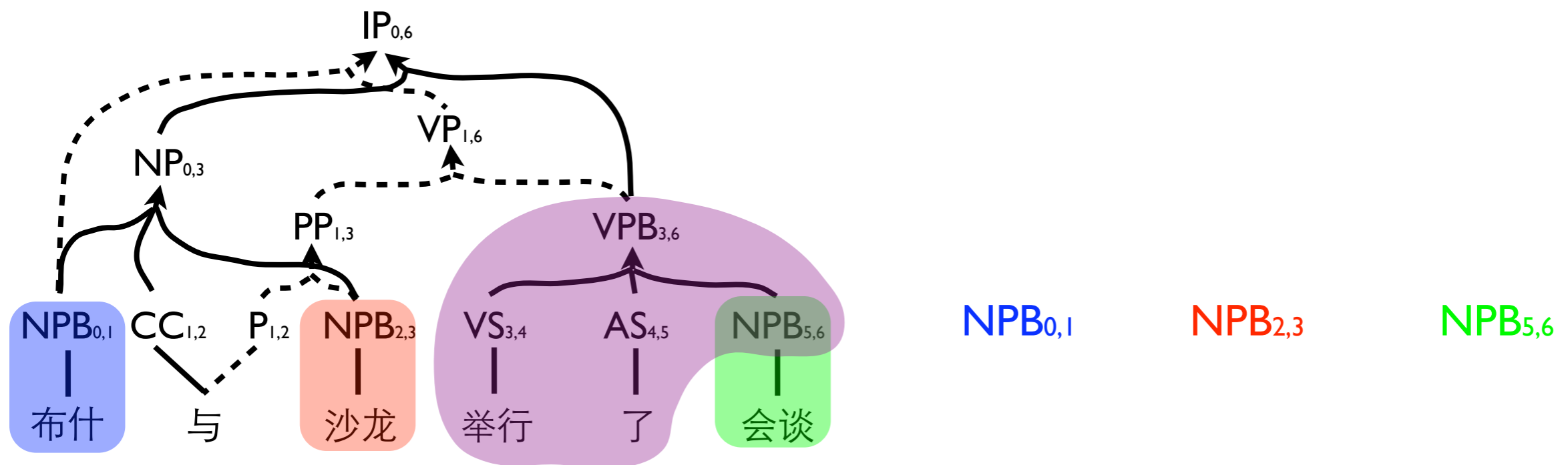
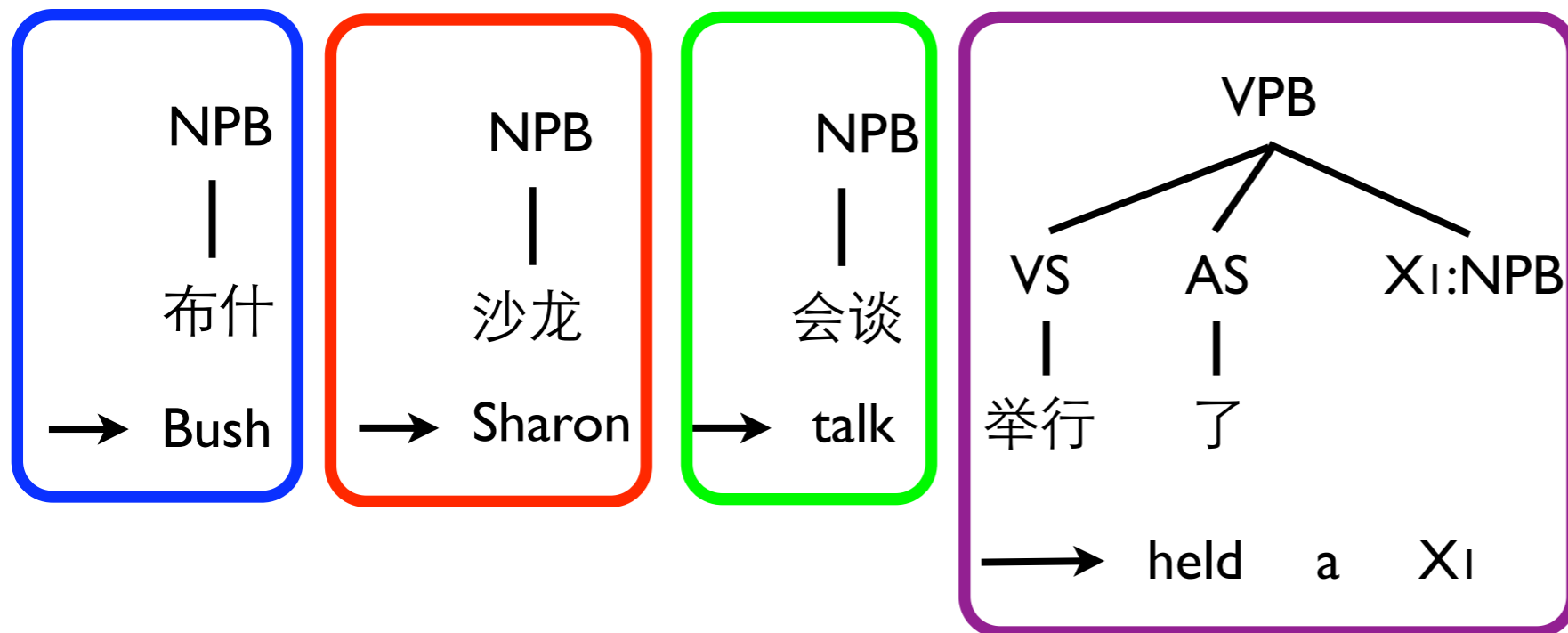
$NPB_{2,3}$

$NPB_{5,6}$

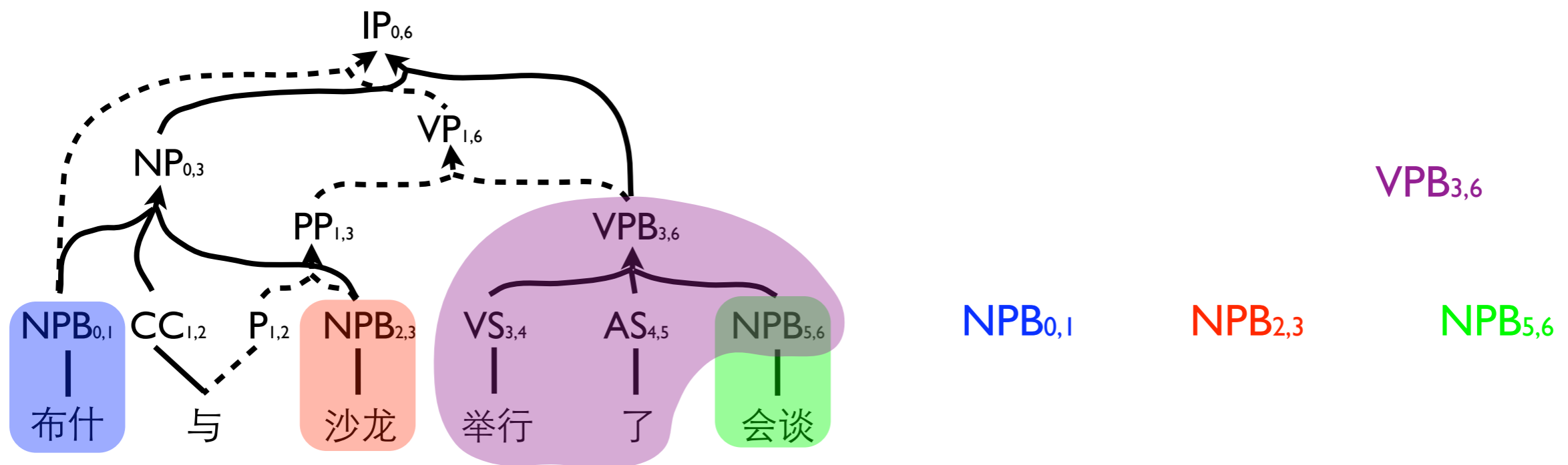
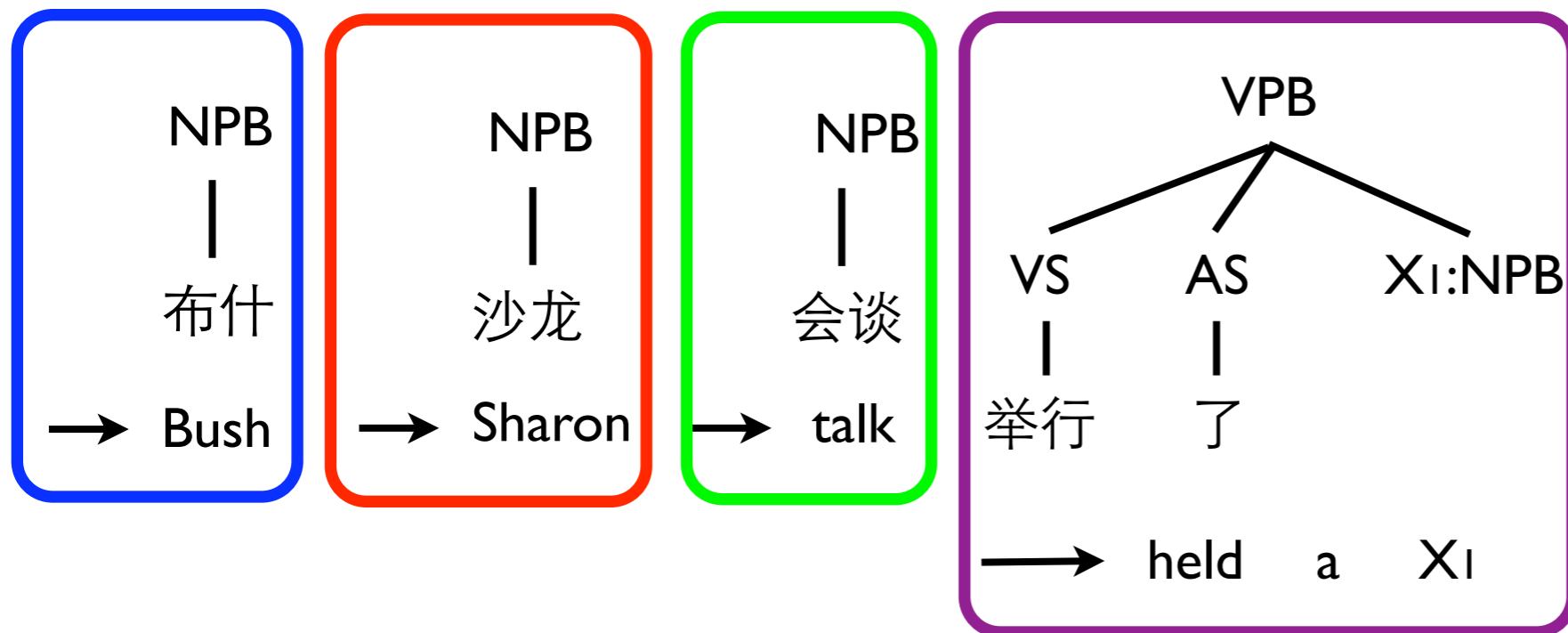
Translation Forest



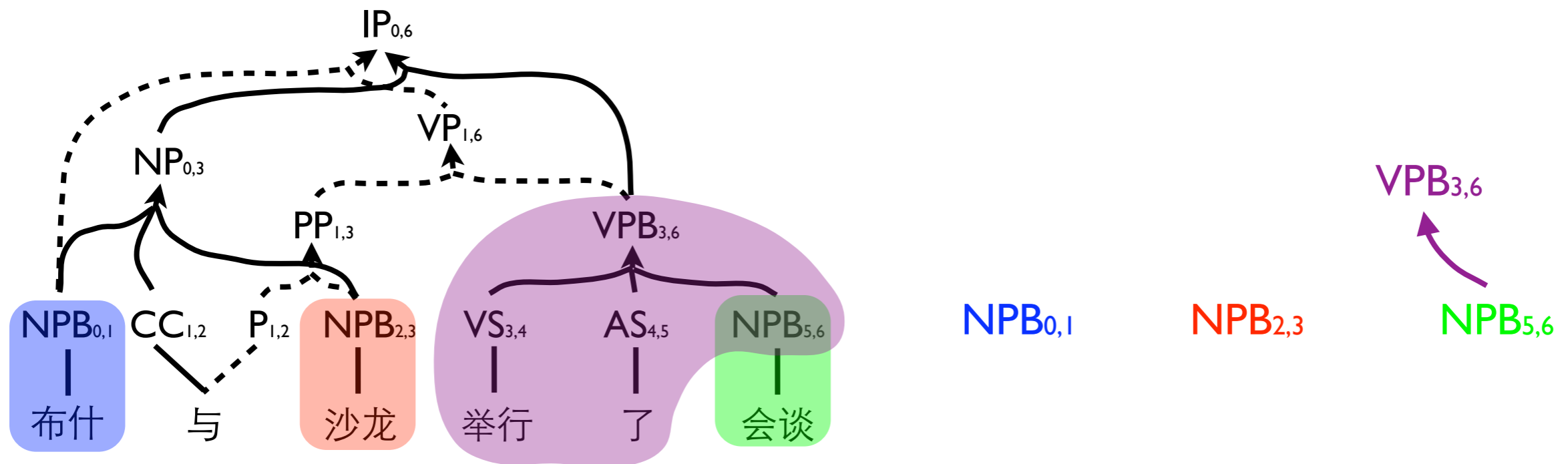
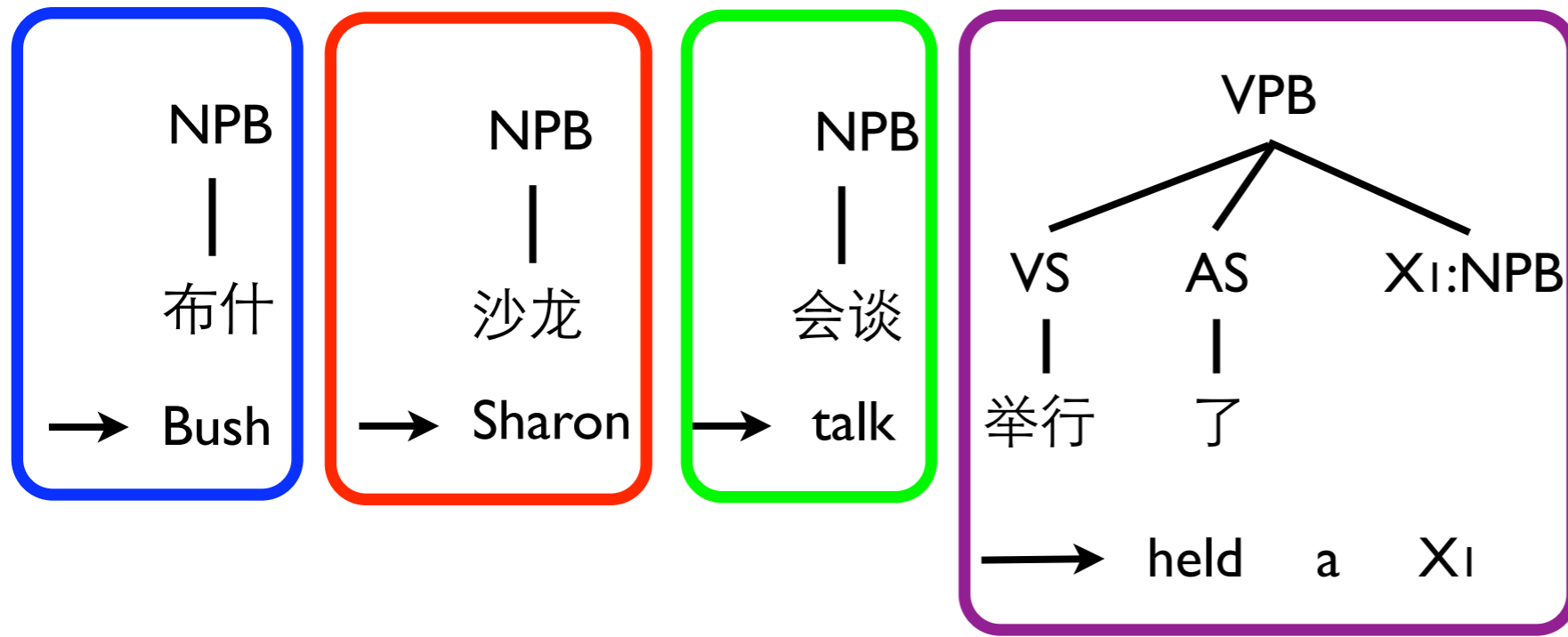
Translation Forest



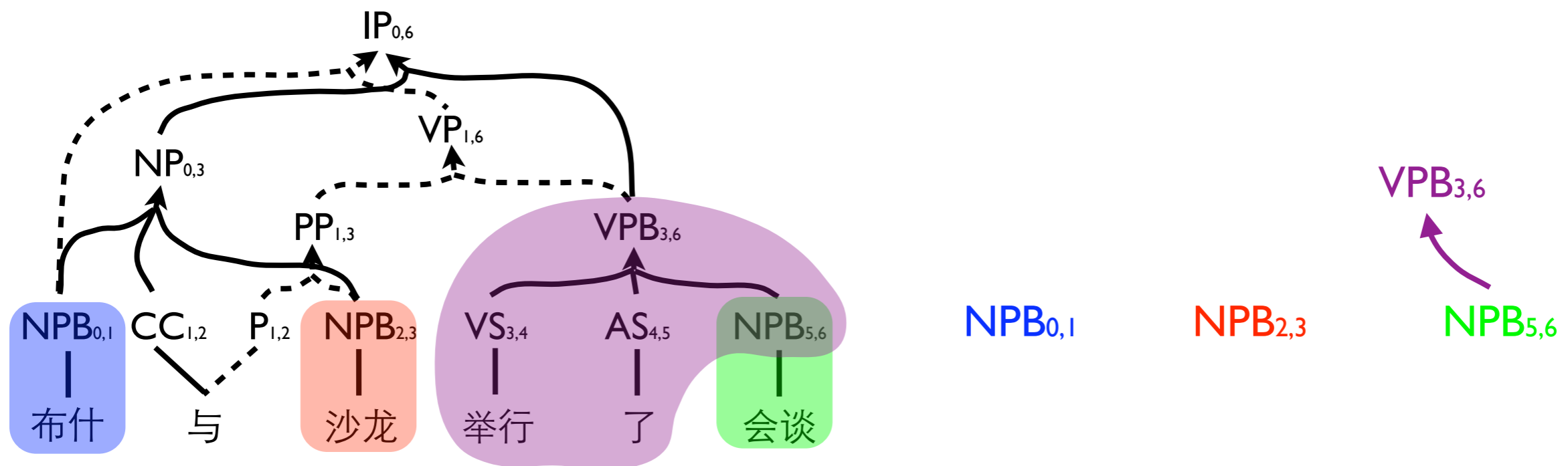
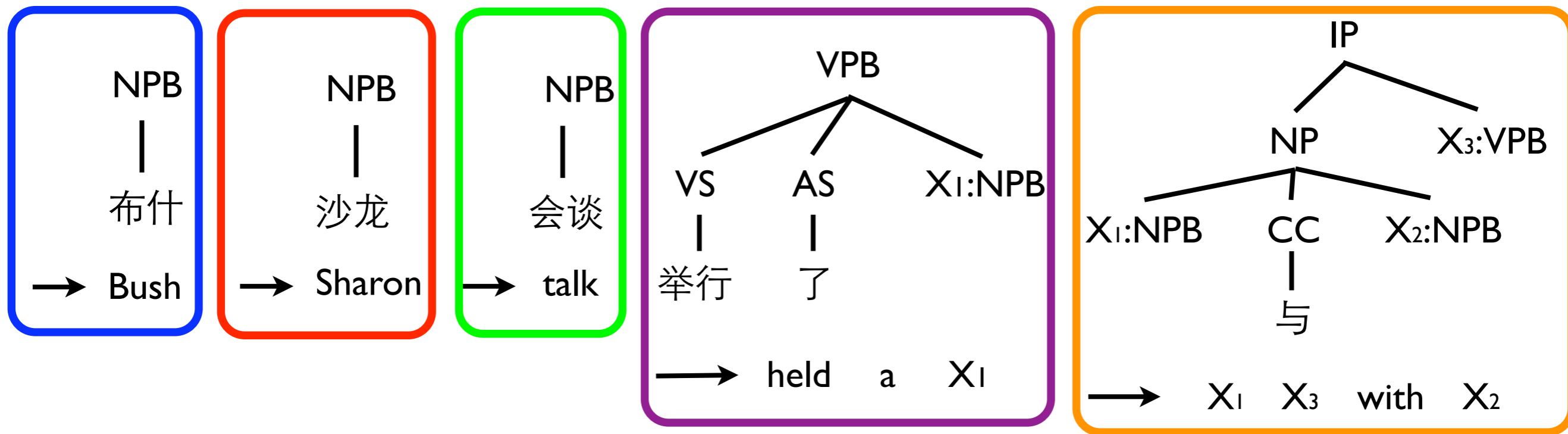
Translation Forest



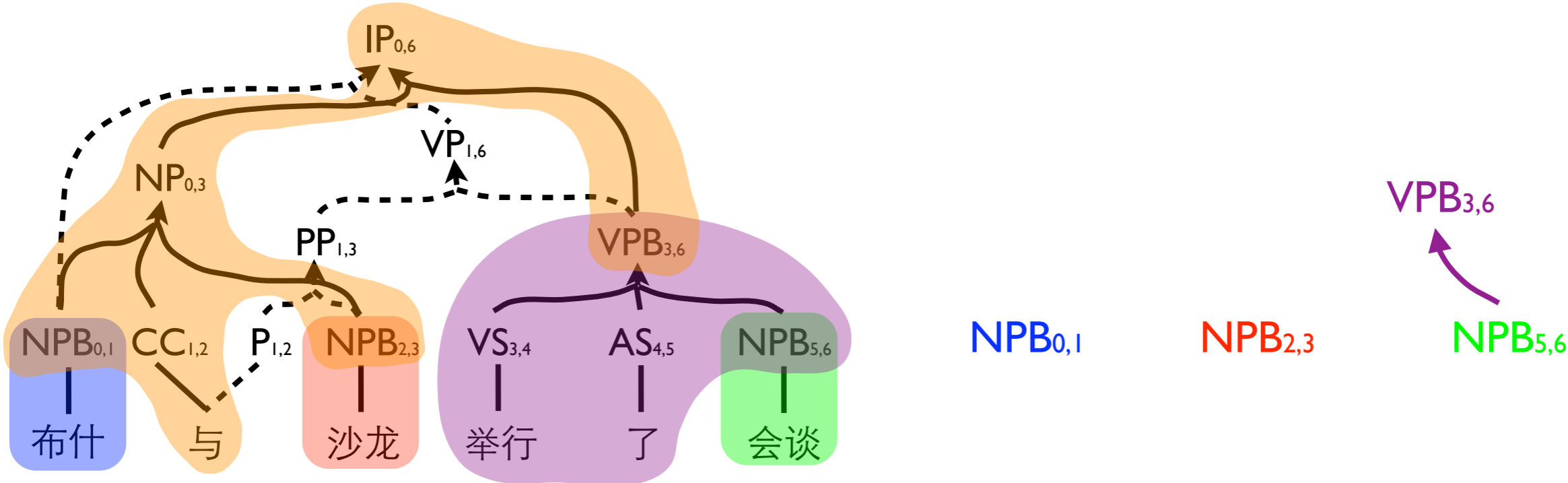
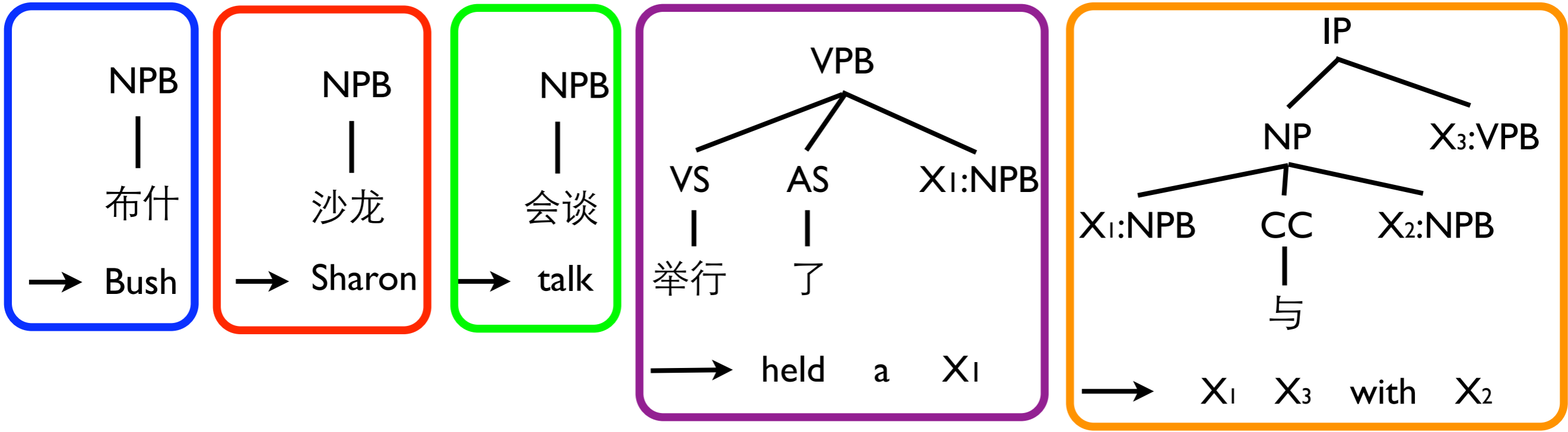
Translation Forest



Translation Forest

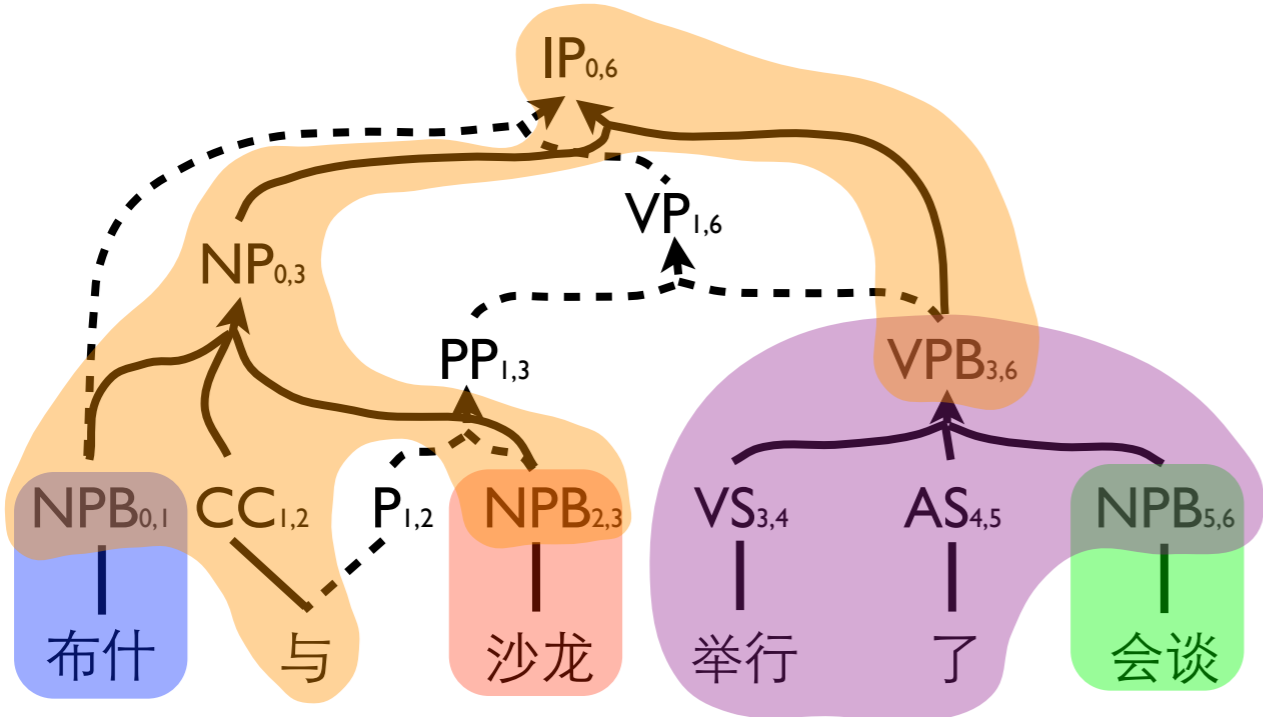
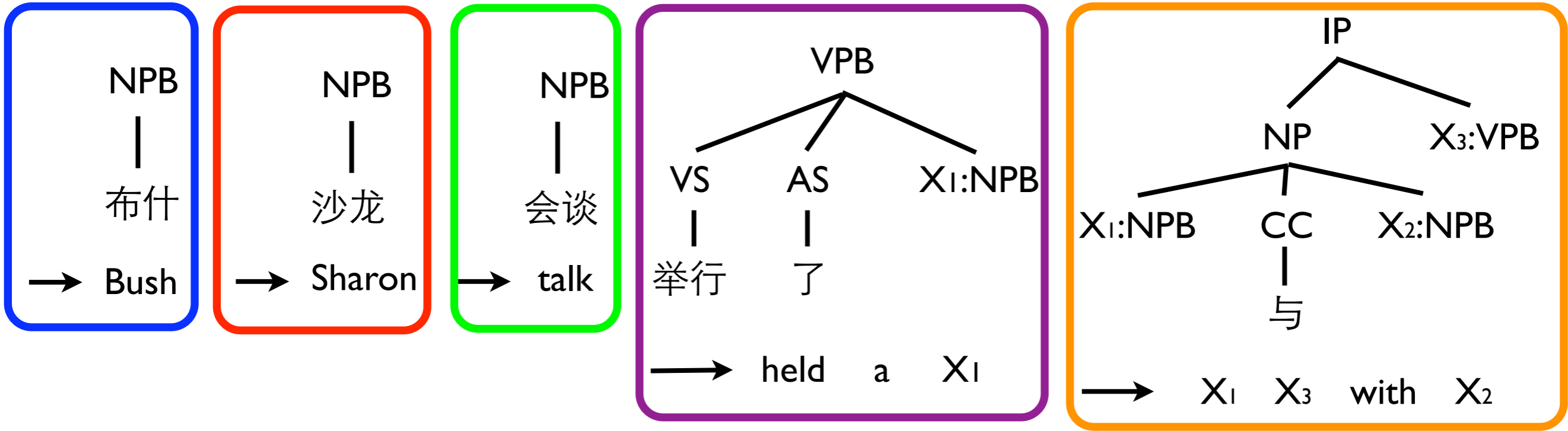


Translation Forest



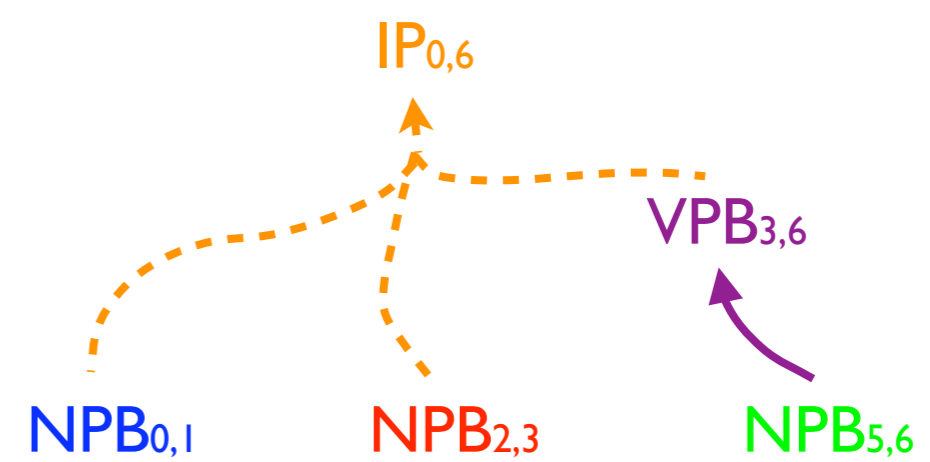
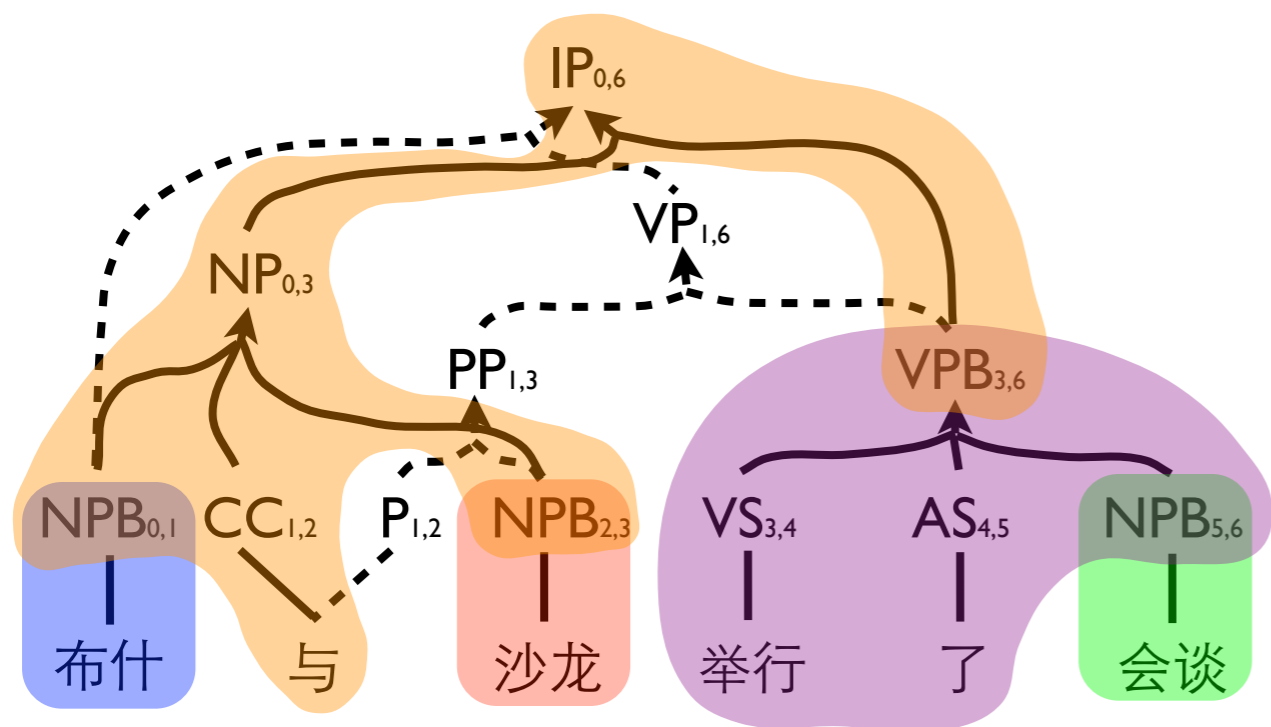
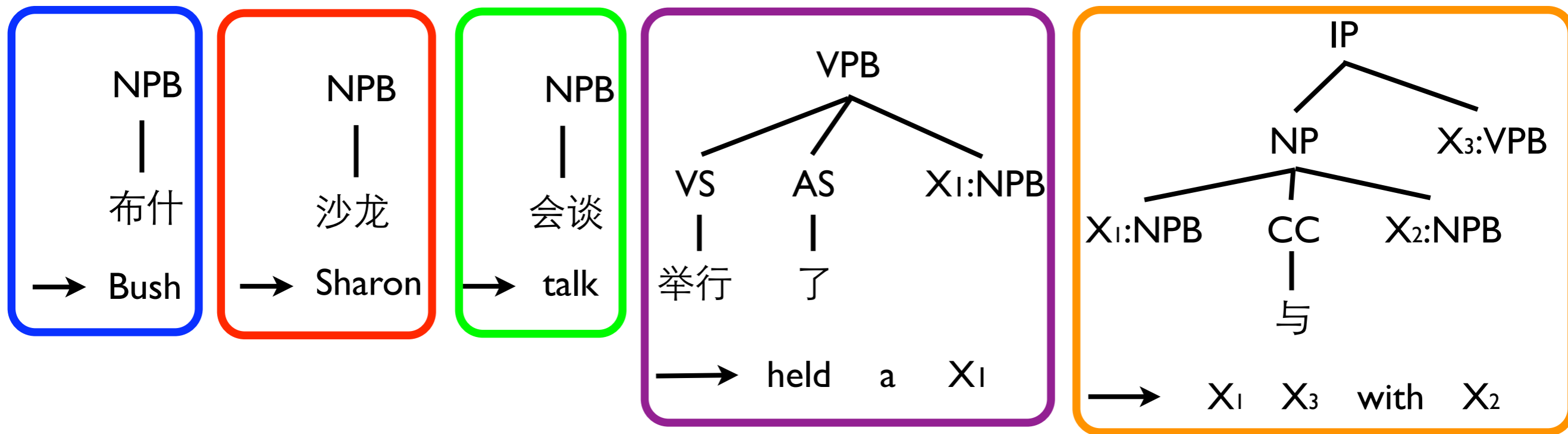
(Mi et al., 2008)

Translation Forest

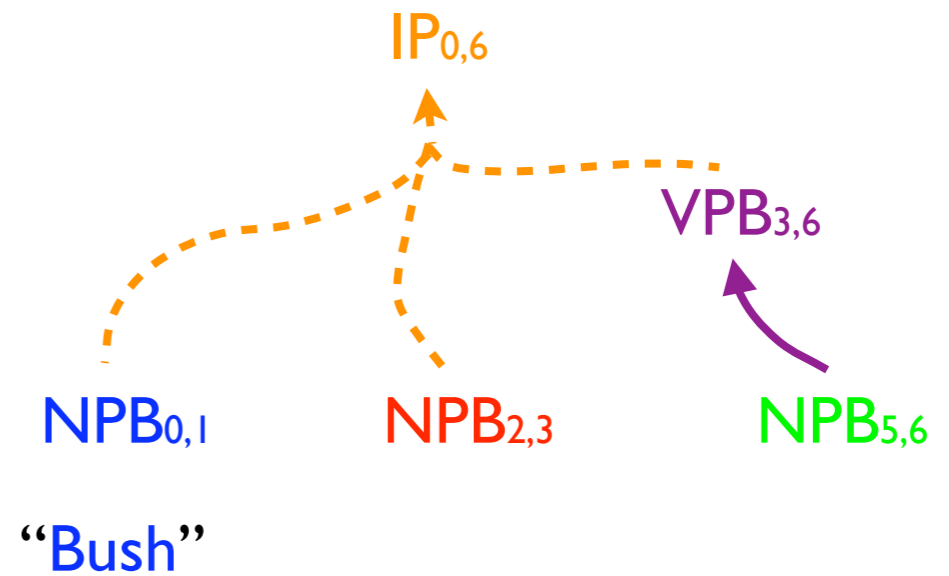
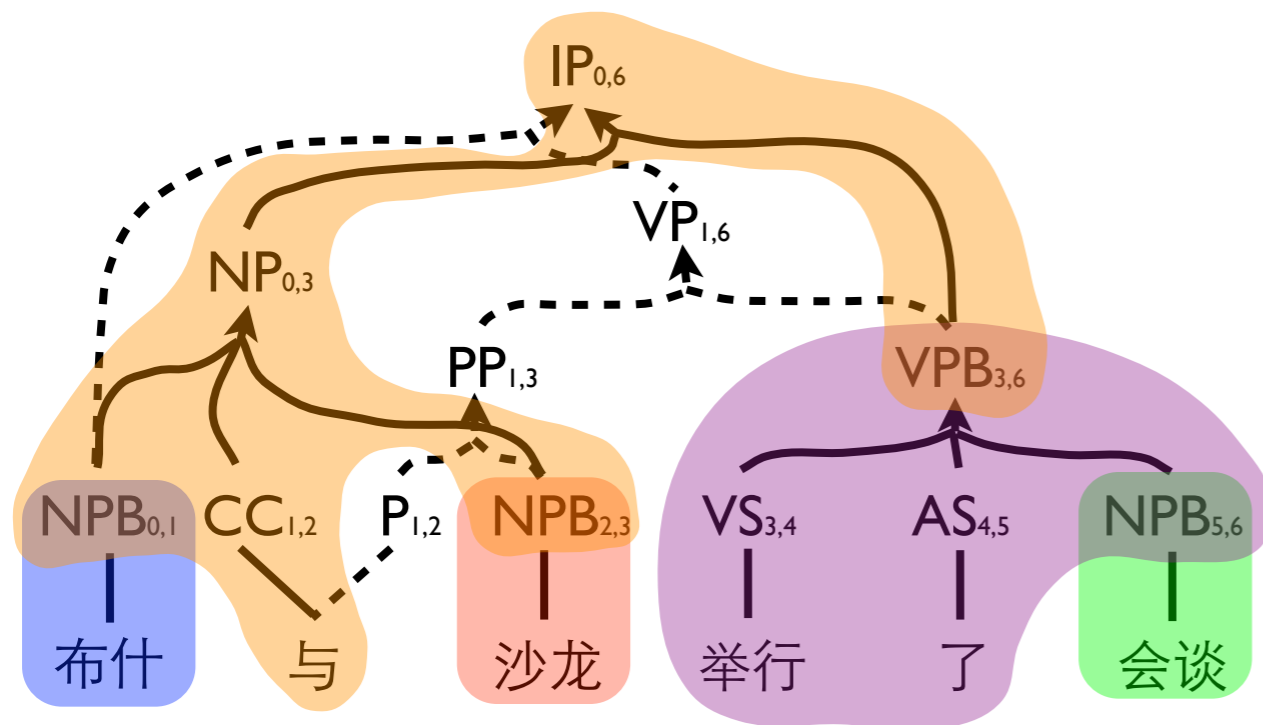
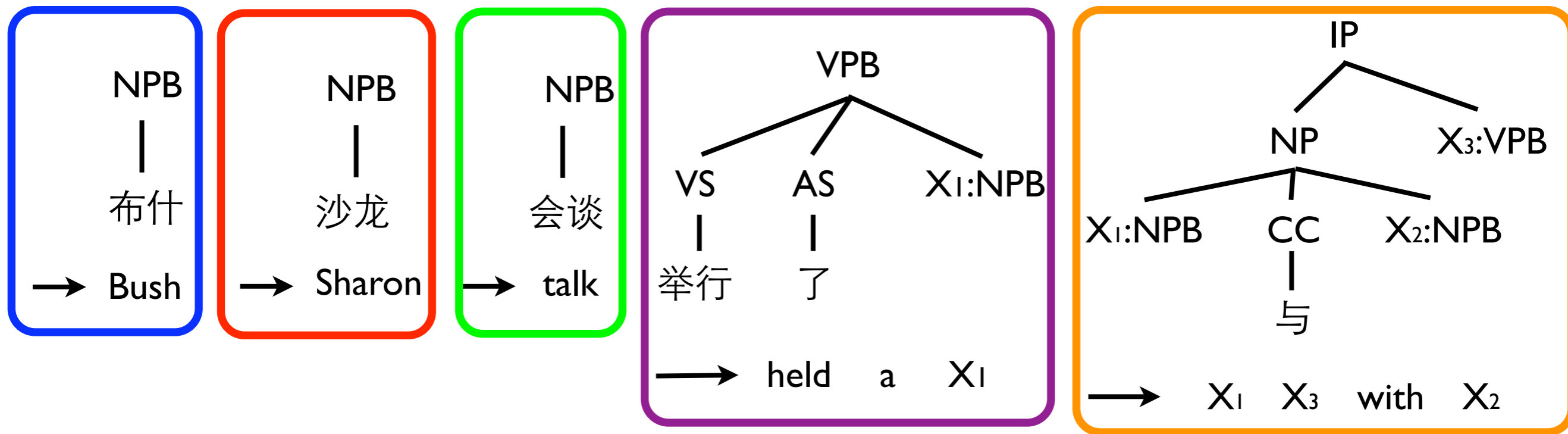


(Mi et al., 2008)

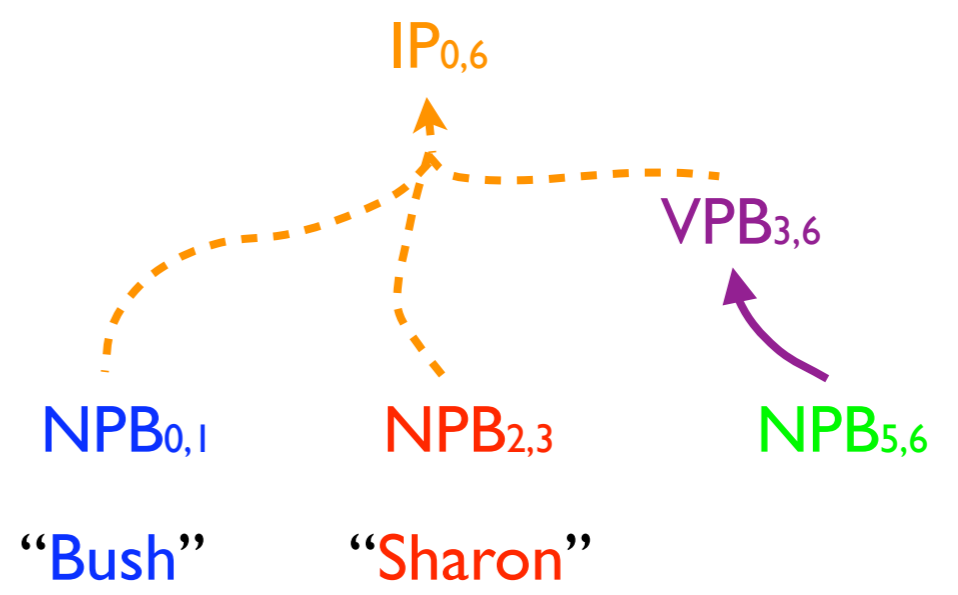
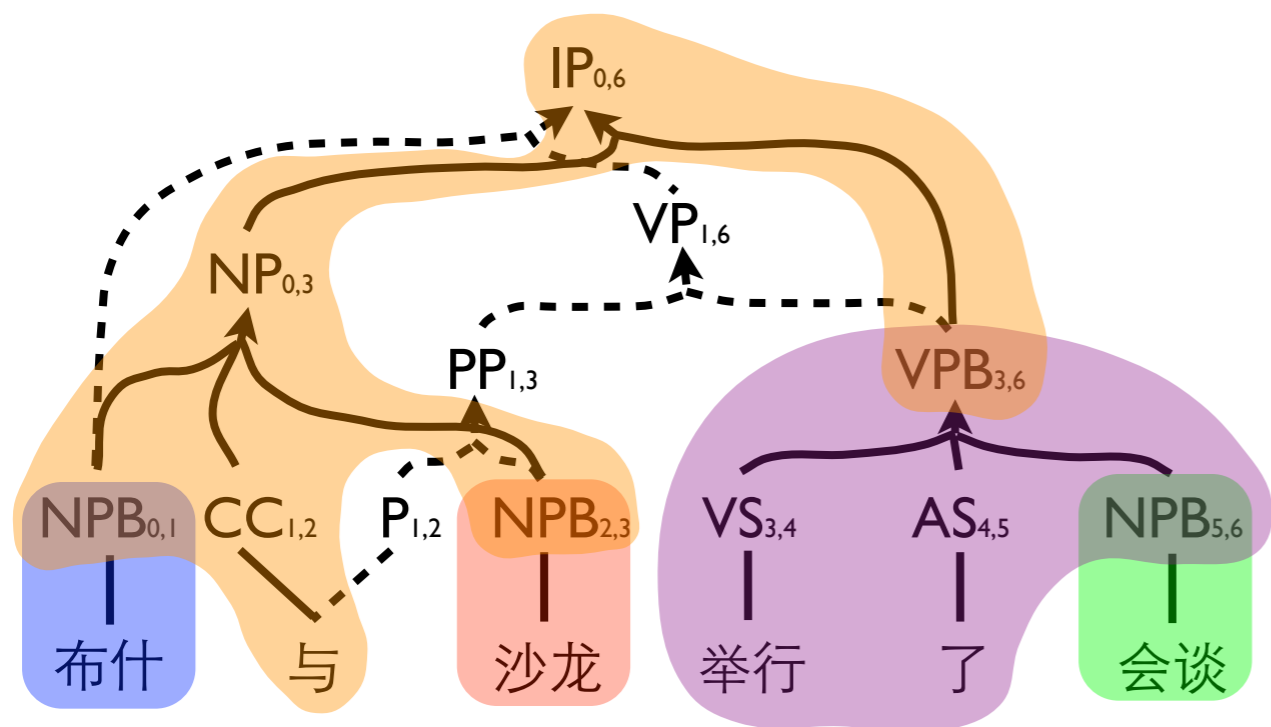
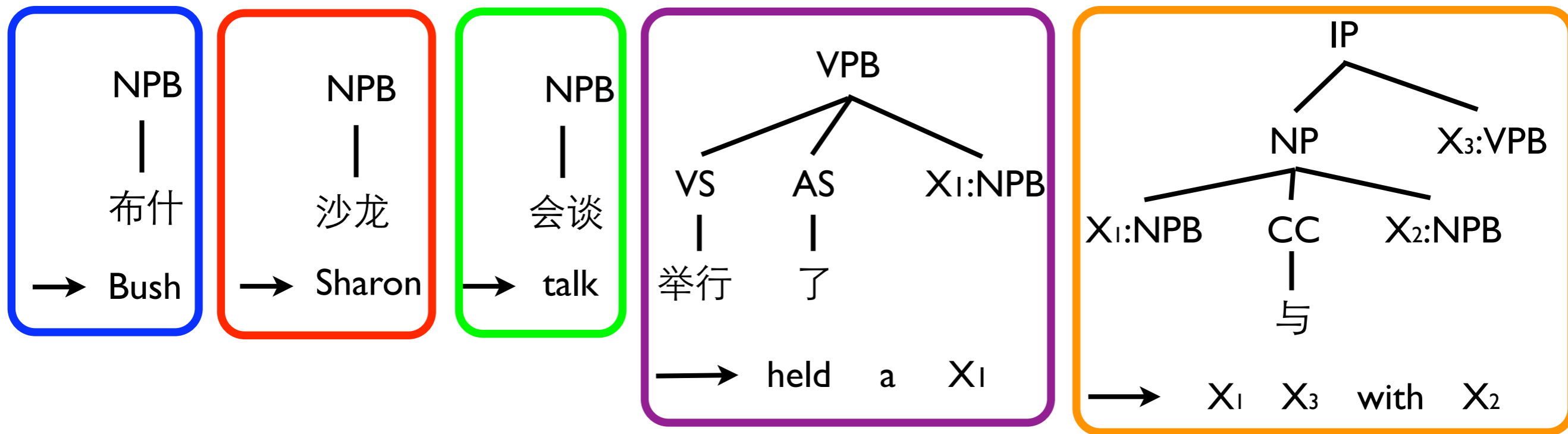
Translation Forest



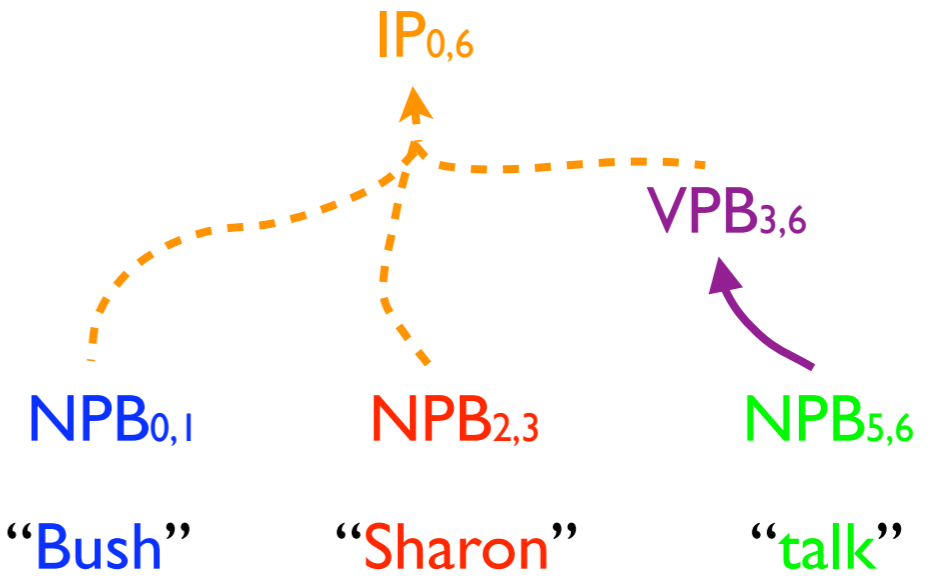
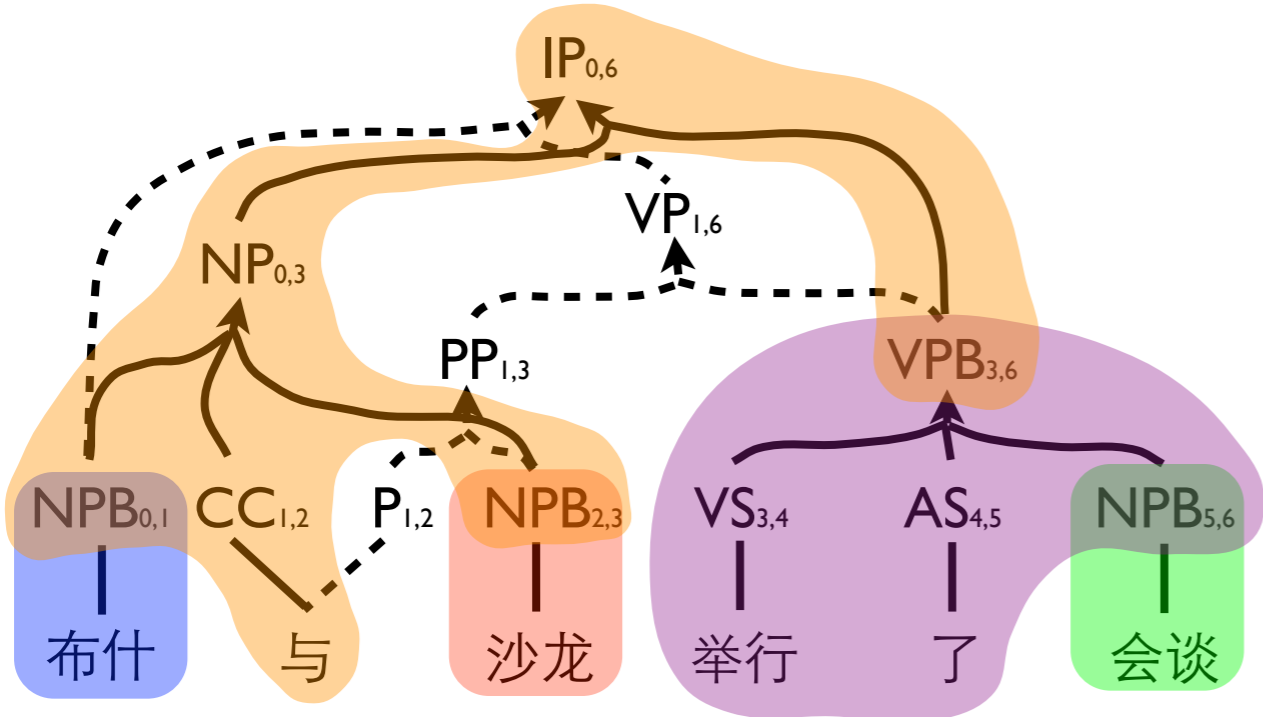
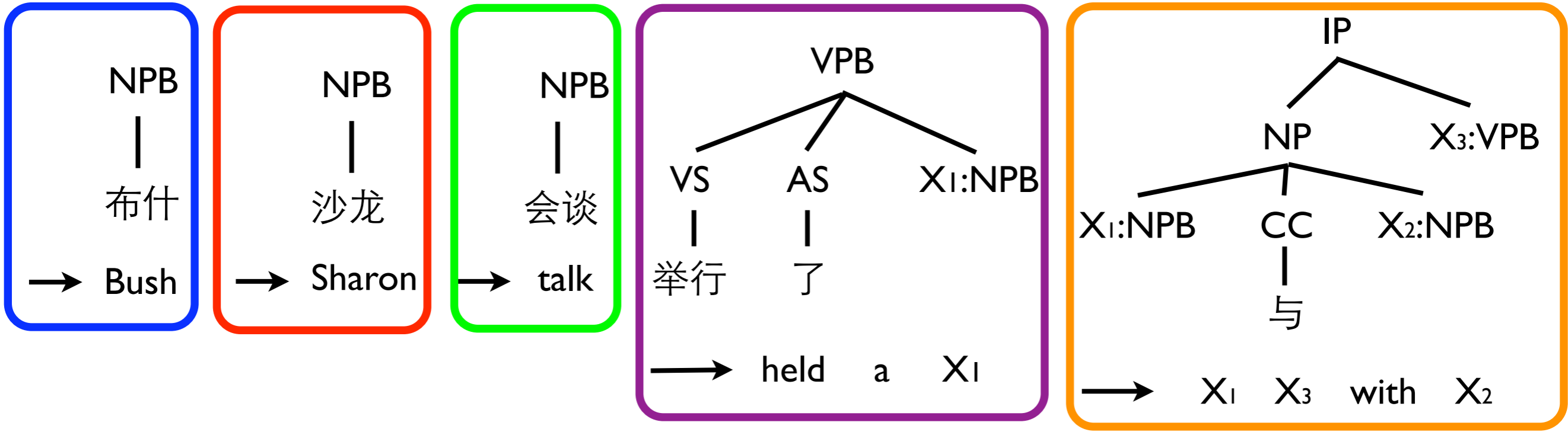
Translation Forest



Translation Forest

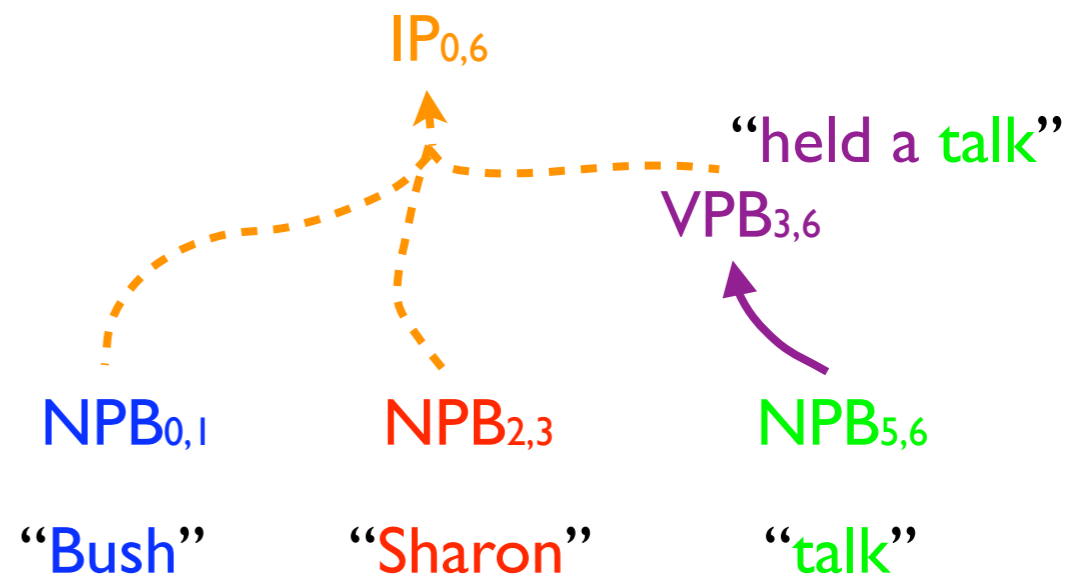
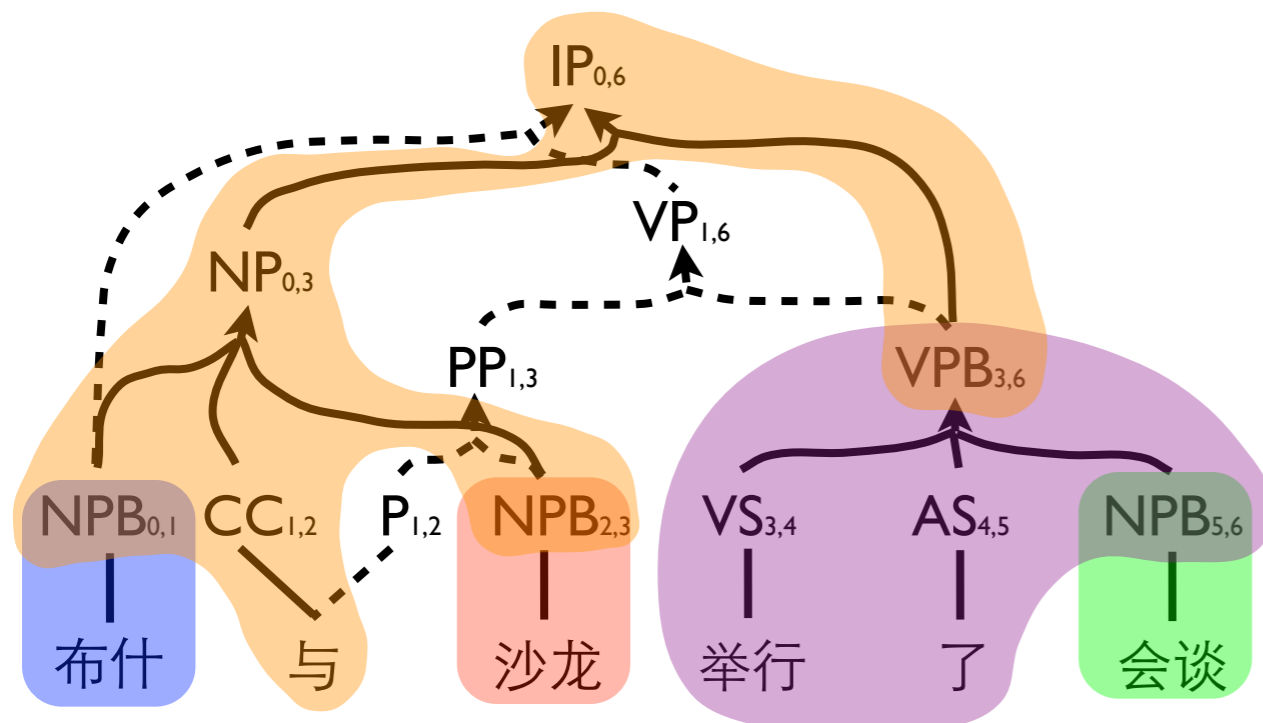
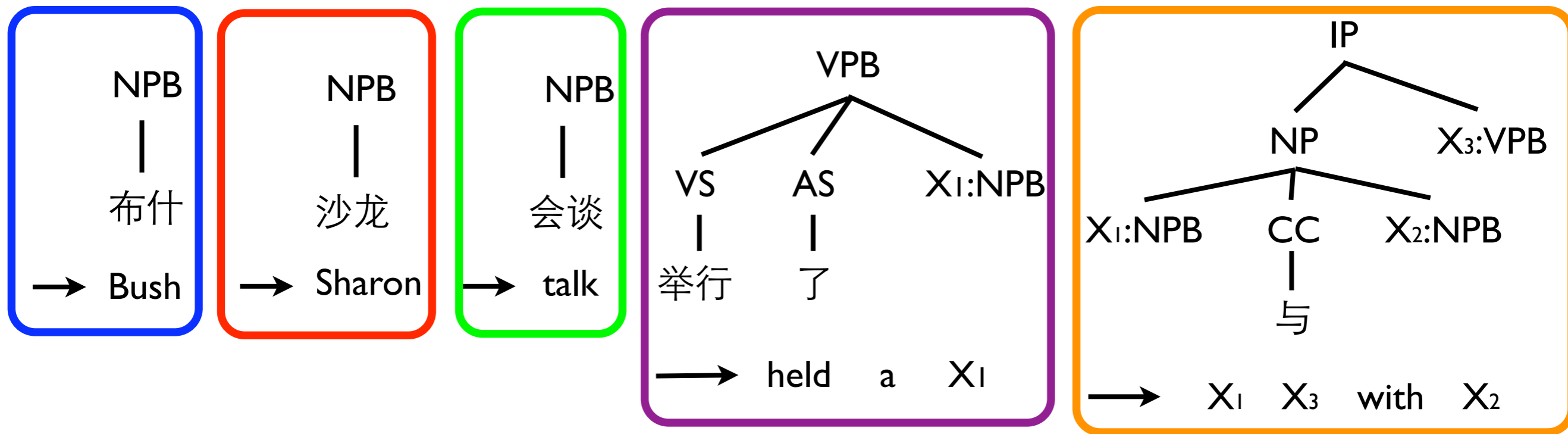


Translation Forest

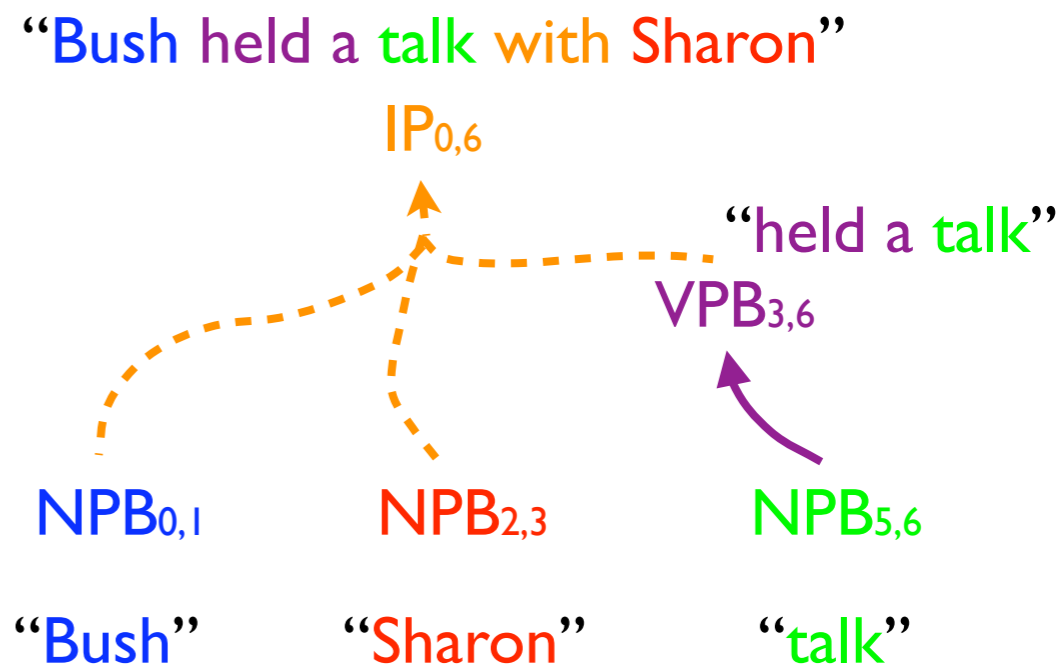
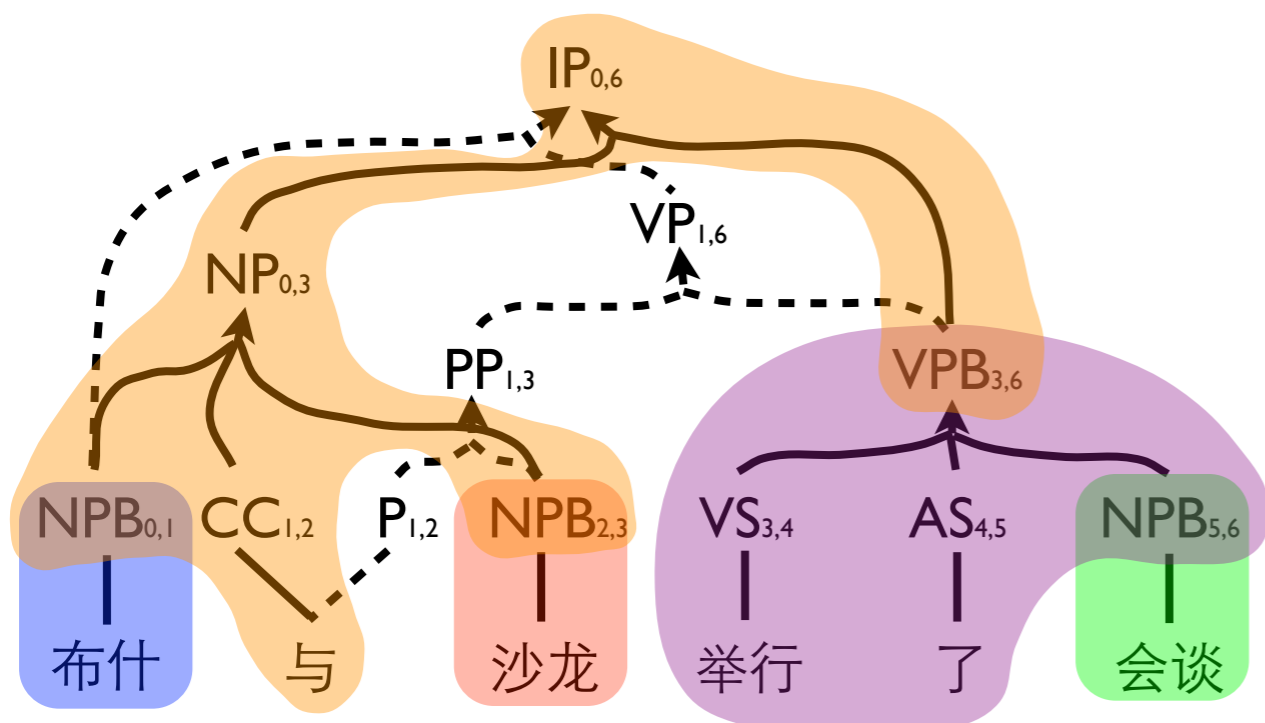
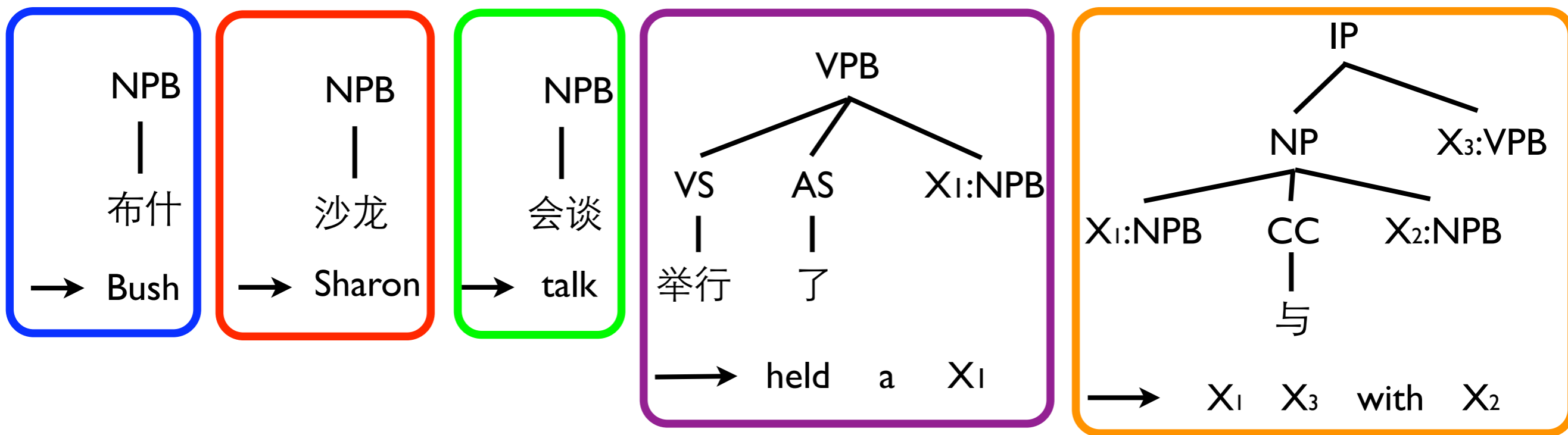


(Mi et al., 2008)

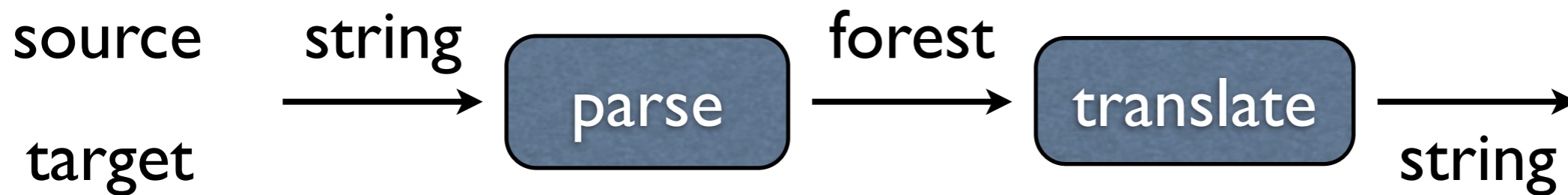
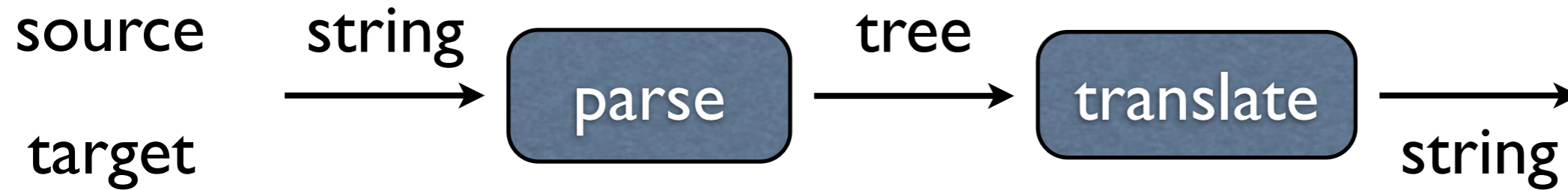
Translation Forest



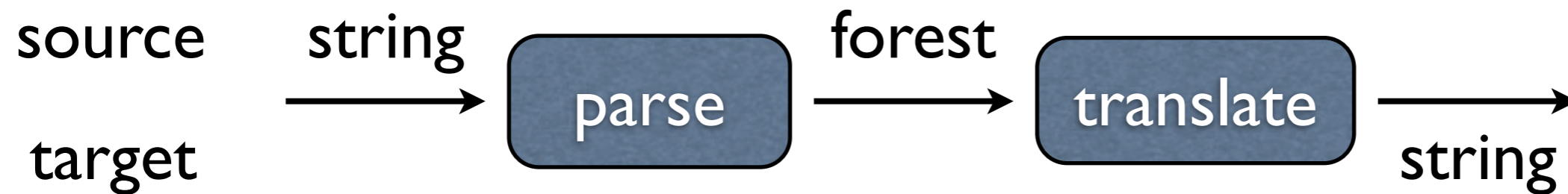
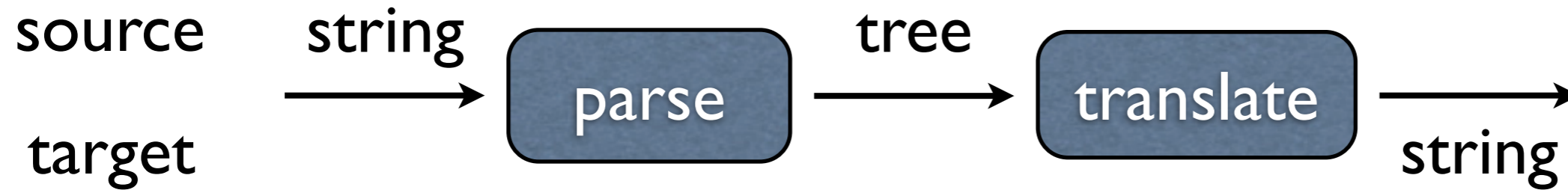
Translation Forest



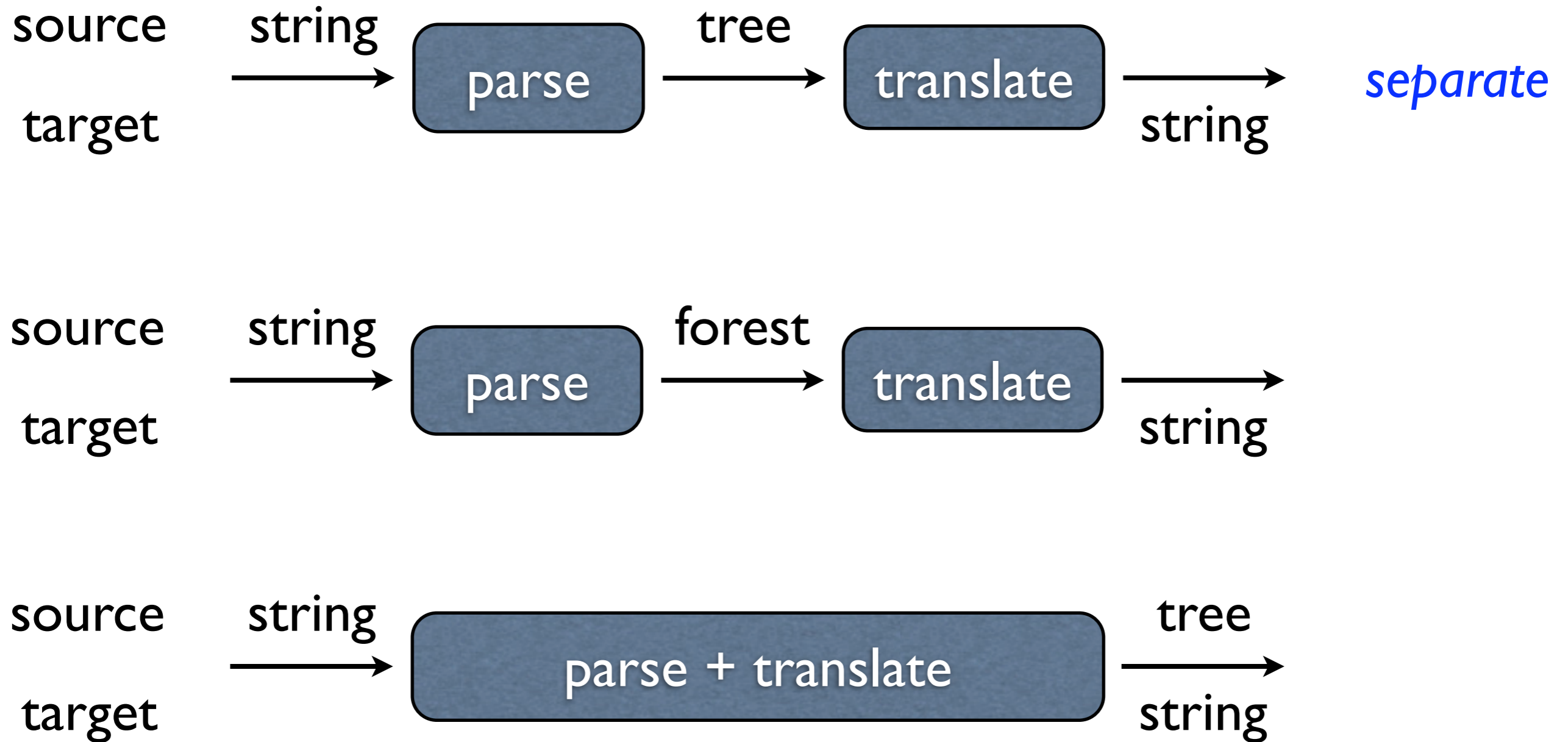
Parsing and Translation: **Separate** Vs. **Joint**



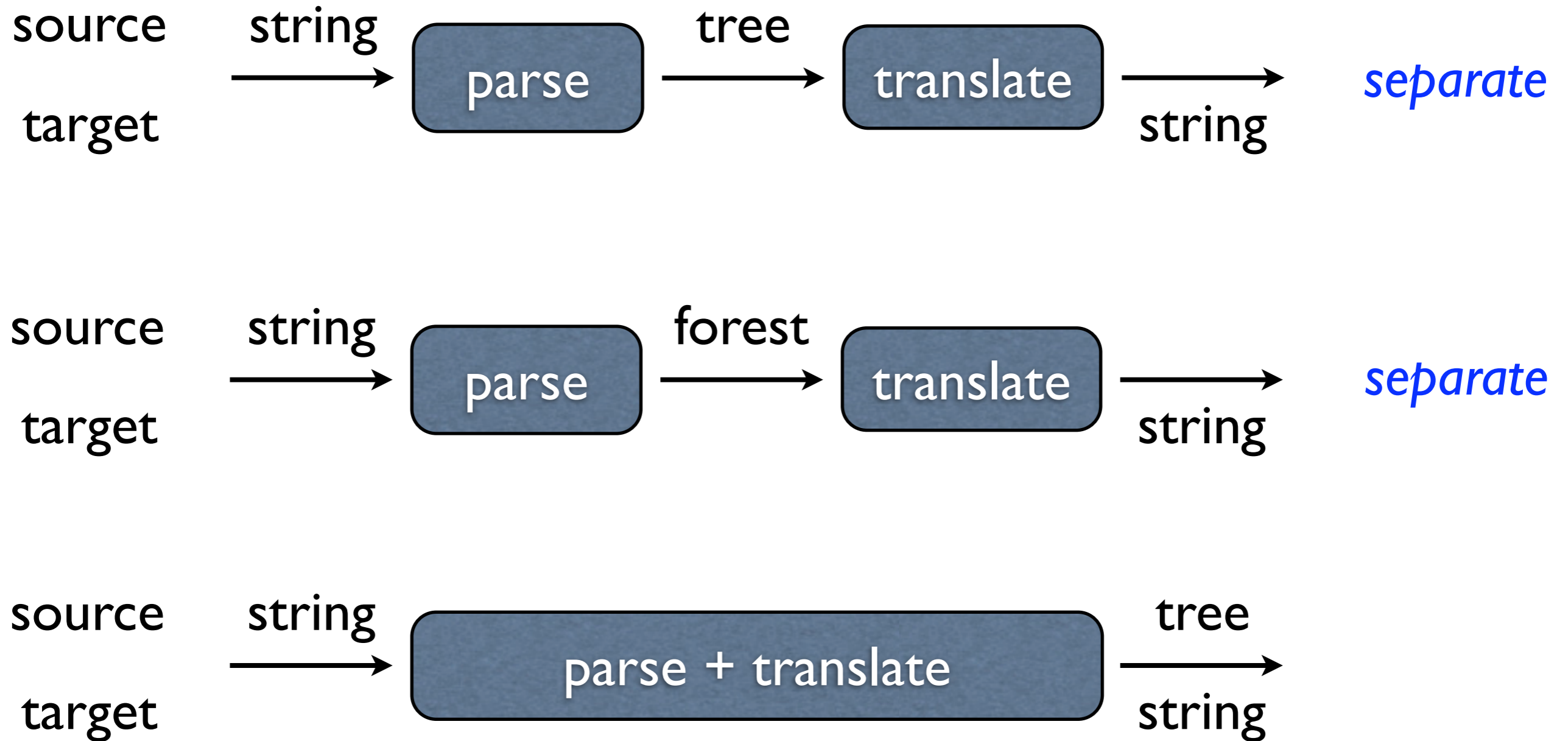
Parsing and Translation: **Separate** Vs. **Joint**



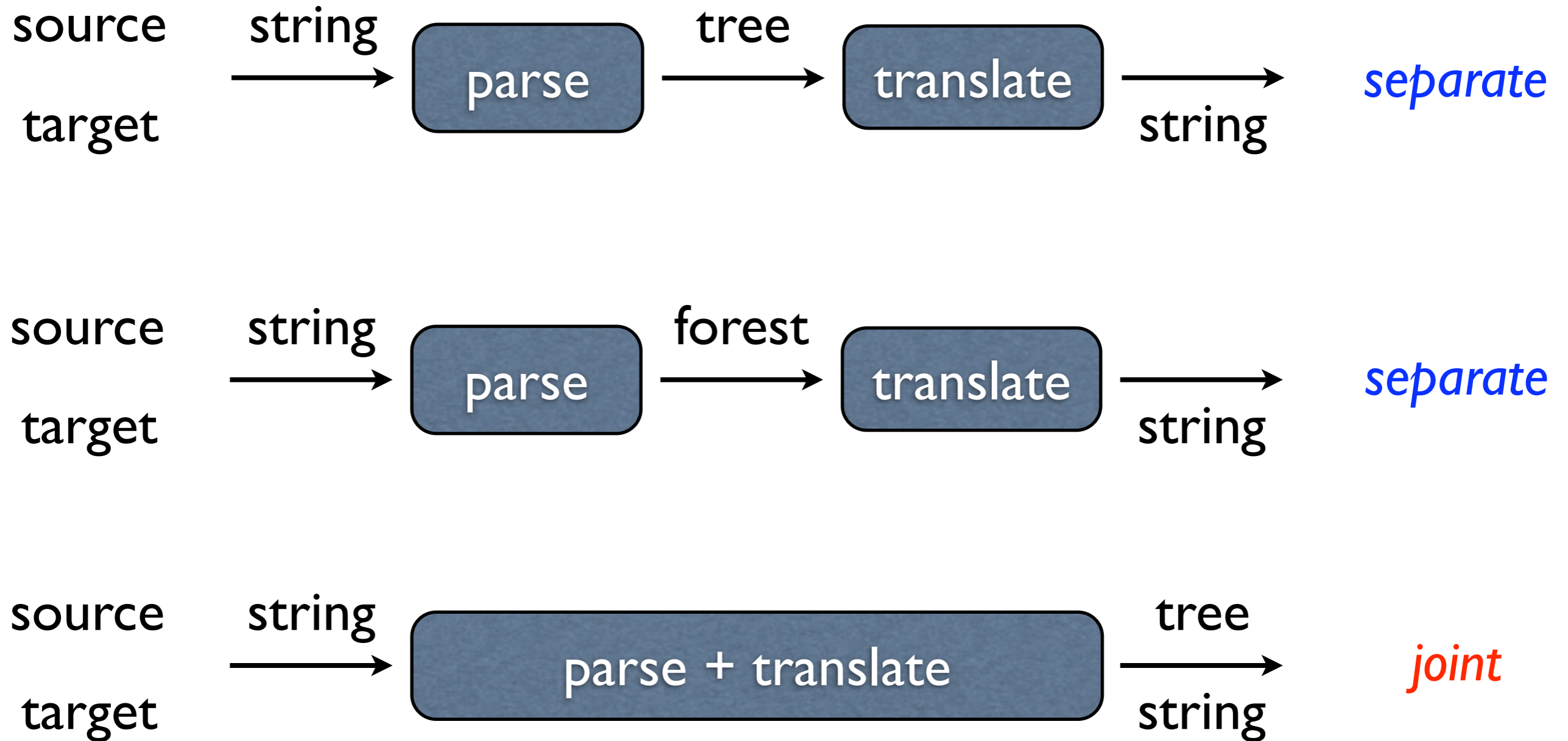
Parsing and Translation: **Separate** Vs. **Joint**



Parsing and Translation: **Separate** Vs. **Joint**



Parsing and Translation: **Separate** Vs. **Joint**



Joint Parsing and Translation

布什 与 沙龙 举行 了 会谈

(Liu and Liu, 2010)

Joint Parsing and Translation

NPB
|
布什
→ Bush

布什 与 沙龙 举行 了 会谈

Joint Parsing and Translation

NPB

|

布什

→

Bush

NPB

|

布什

与

沙龙

举行

了

会谈

Joint Parsing and Translation

NPB
|
布什
→ Bush

NPB
|
布什 与 沙龙 举行 了 会谈

Bush

Joint Parsing and Translation

NPB

|

布什

与

沙龙

举行

了

会谈

Bush

Joint Parsing and Translation

NPB
|
沙龙
→ Sharon

NPB
|
布什 与 沙龙 举行 了 会谈

Bush

Joint Parsing and Translation

NPB
|
沙龙
→ Sharon

NPB NPB
| |
布什 沙龙 举行 了 会谈

Bush

Joint Parsing and Translation

NPB
|
沙龙
→ Sharon

NPB NPB
| |
布什 沙龙 举行 了 会谈

Bush

Sharon

(Liu and Liu, 2010)

Joint Parsing and Translation

NPB

|

布什

Bush

与

NPB

|

沙龙

Sharon

举行

了

会谈

Joint Parsing and Translation

NPB
|
会谈
→ talk

NPB
|
布什

与

NPB
|
沙龙

举行

了

会谈

Bush

Sharon

Joint Parsing and Translation

NPB
|
会谈
→ talk

NPB
|
布什

与

NPB
|
沙龙

举行

了

NPB
|
会谈

Bush

Sharon

Joint Parsing and Translation

NPB
|
会谈
→ talk

NPB
|
布什

与

NPB
|
沙龙

举行

了

NPB
|
会谈

Bush

Sharon

talk

(Liu and Liu, 2010)

Joint Parsing and Translation

NPB
|
布什

Bush

与

NPB
|
沙龙

Sharon

举行

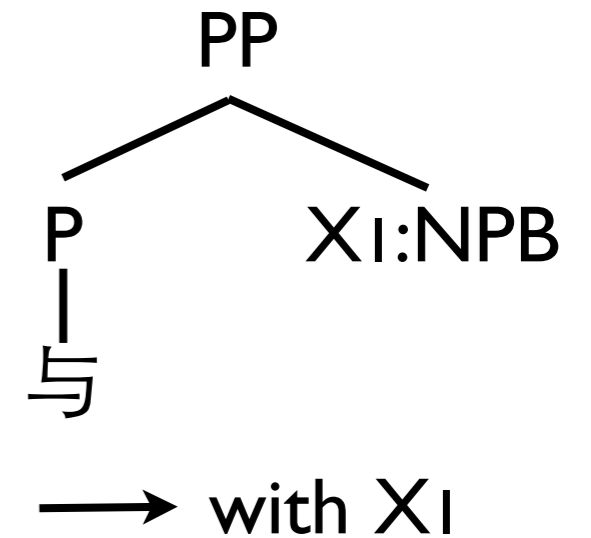
了

NPB
|
会谈

talk

(Liu and Liu, 2010)

Joint Parsing and Translation



NPB
|
布什

Bush

与

NPB
|
沙龙

Sharon

举行

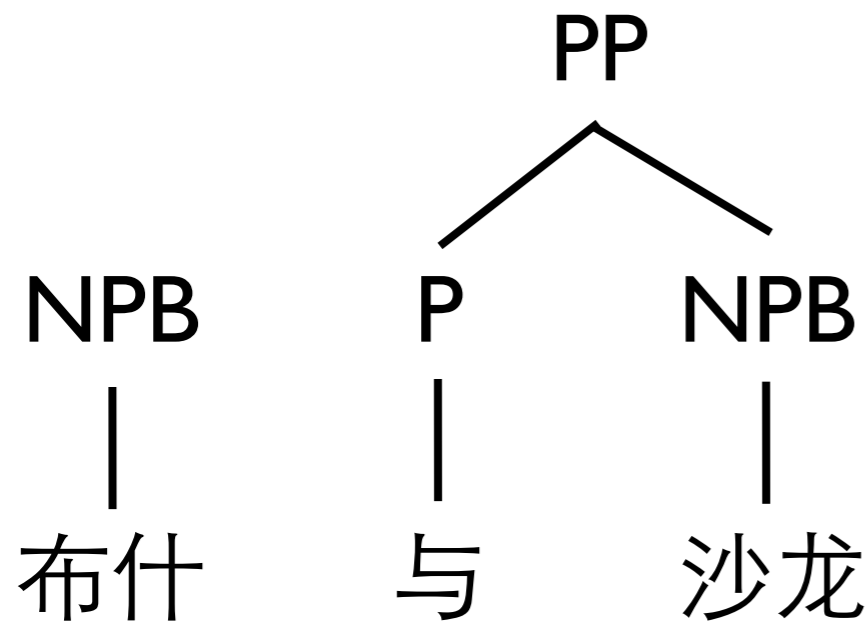
了

NPB
|
会谈

talk

(Liu and Liu, 2010)

Joint Parsing and Translation

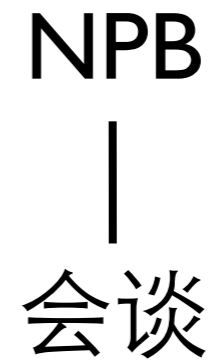


Bush

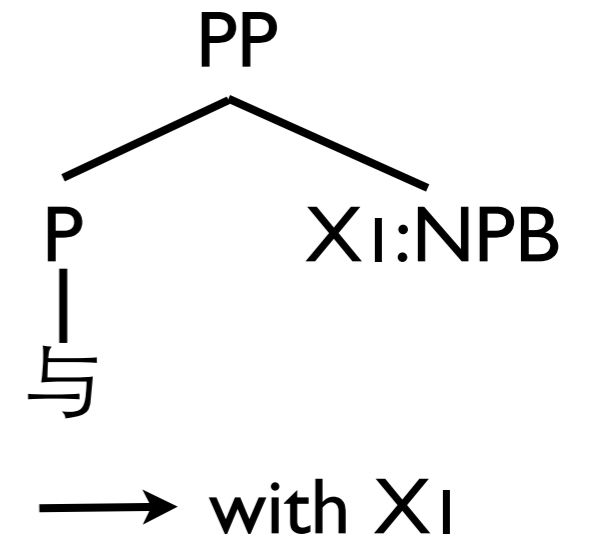
Sharon

举行

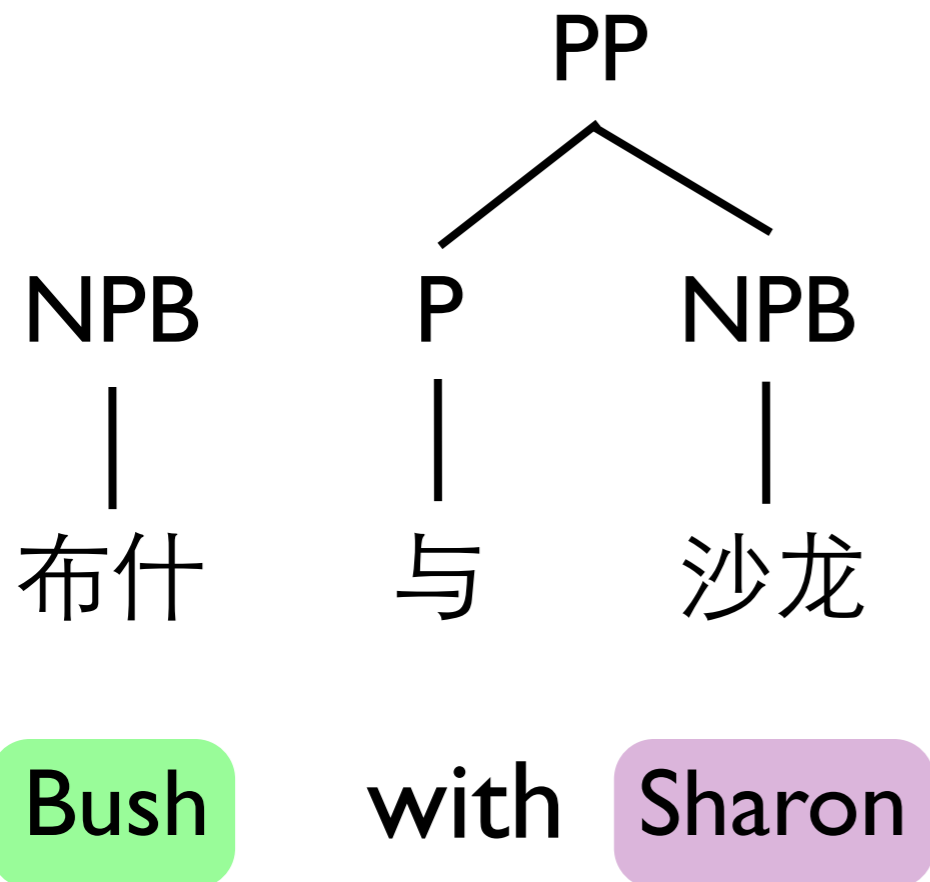
了



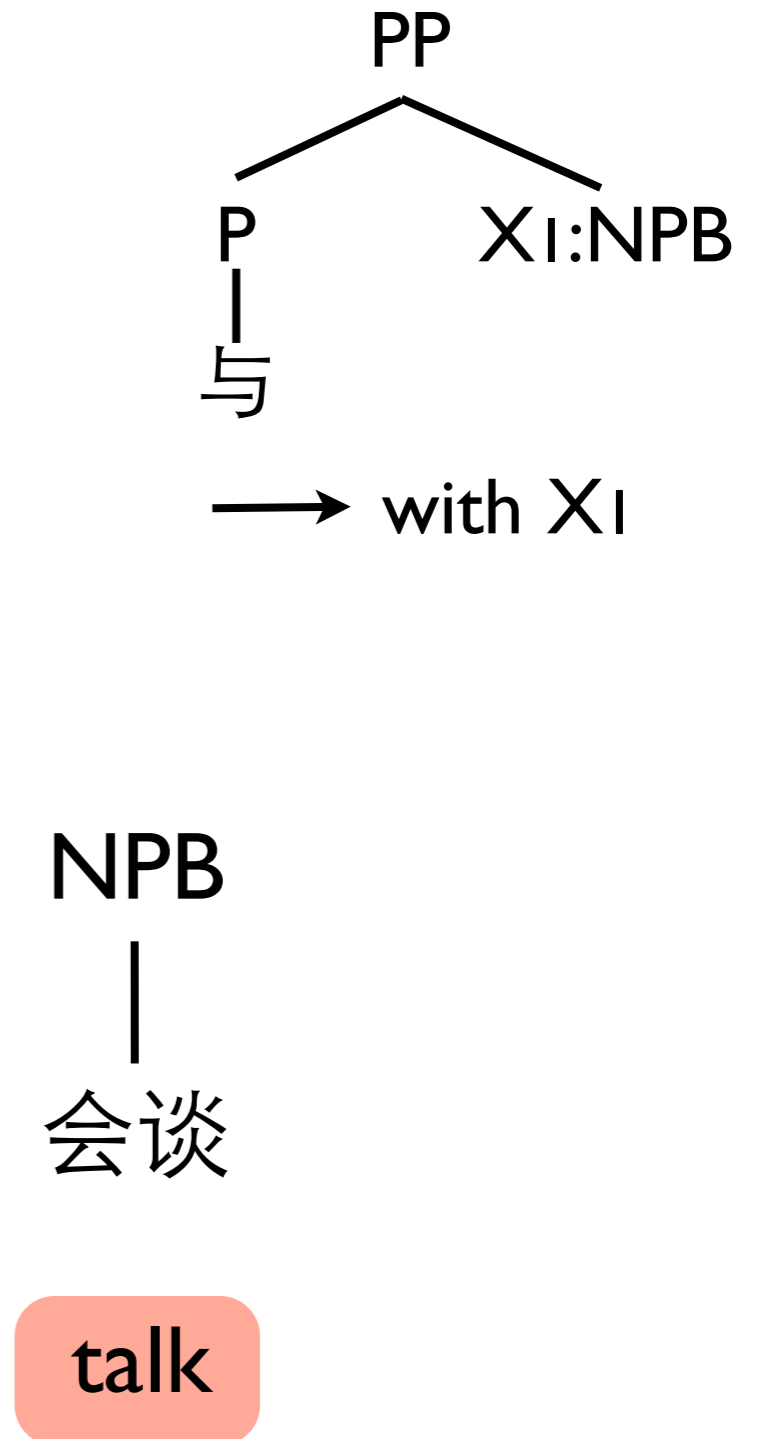
talk



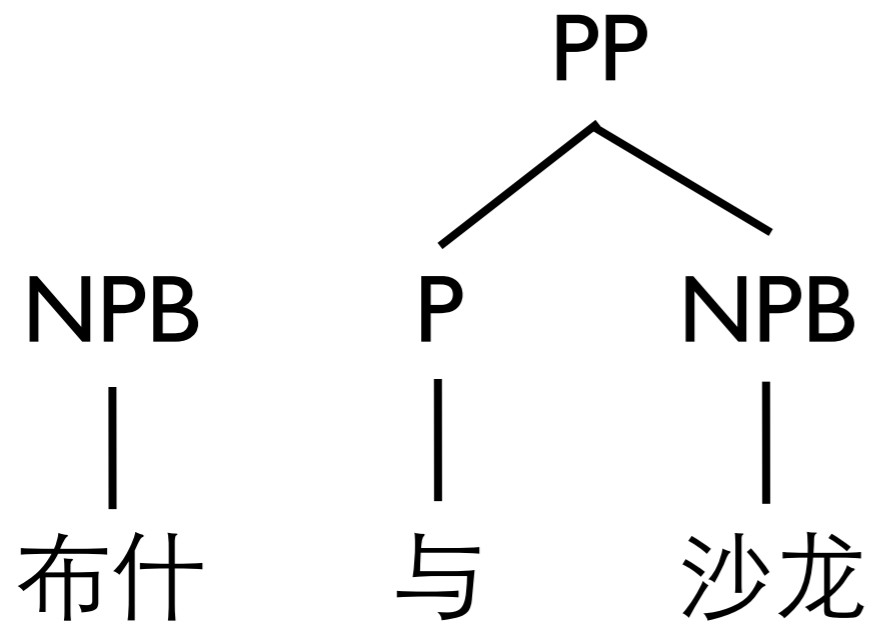
Joint Parsing and Translation



举行 了



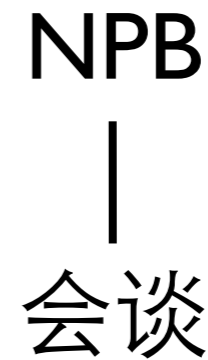
Joint Parsing and Translation



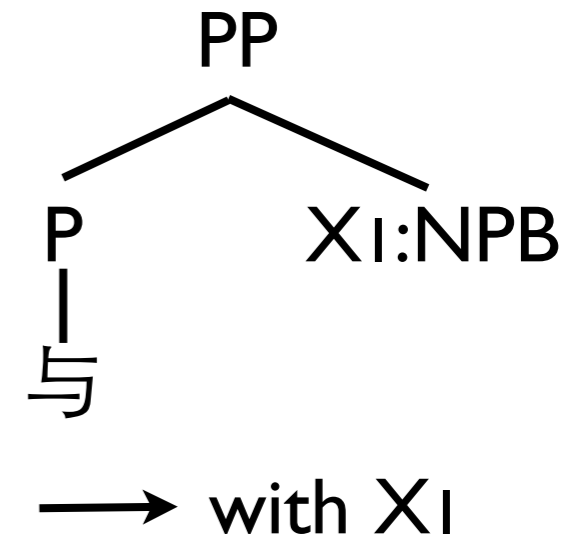
Bush

with Sharon

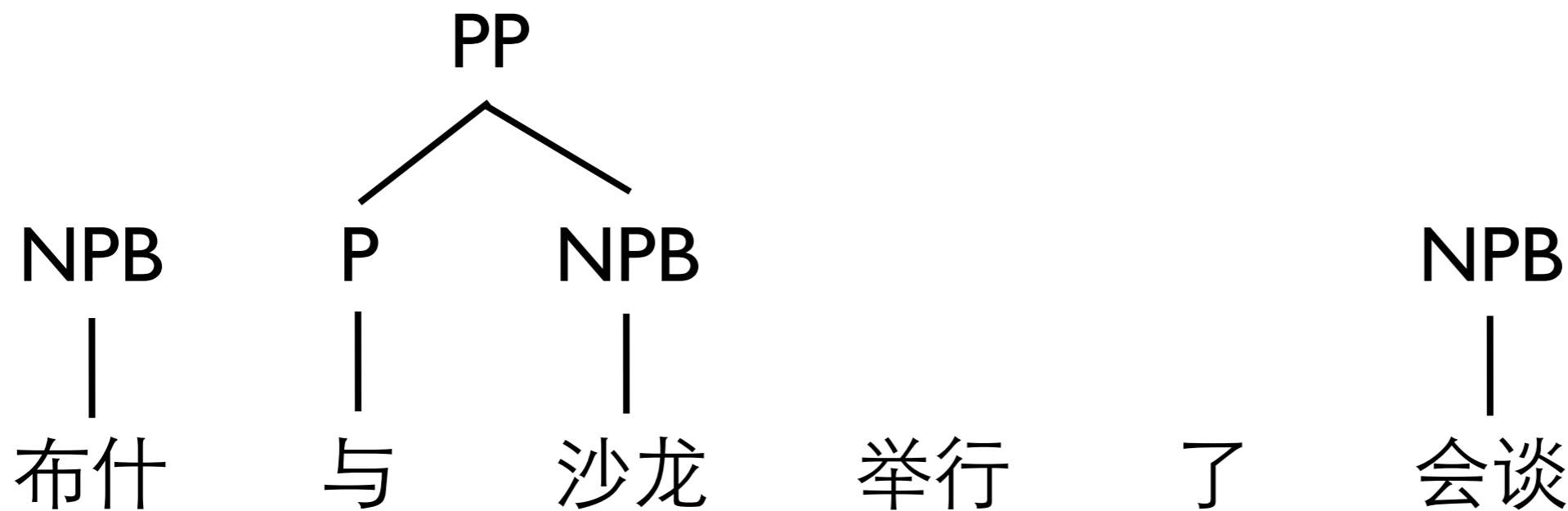
举行 了



talk



Joint Parsing and Translation



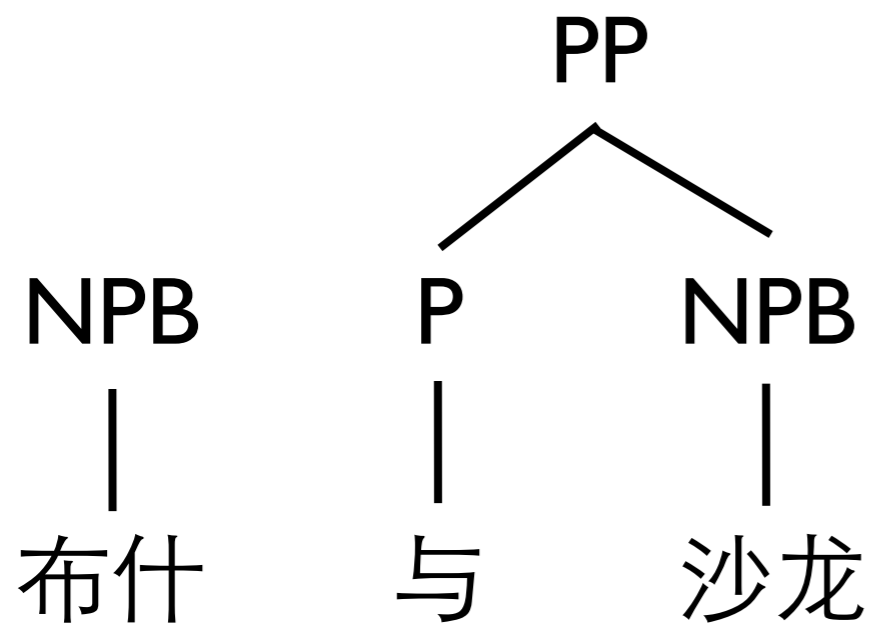
Bush

with Sharon

talk

(Liu and Liu, 2010)

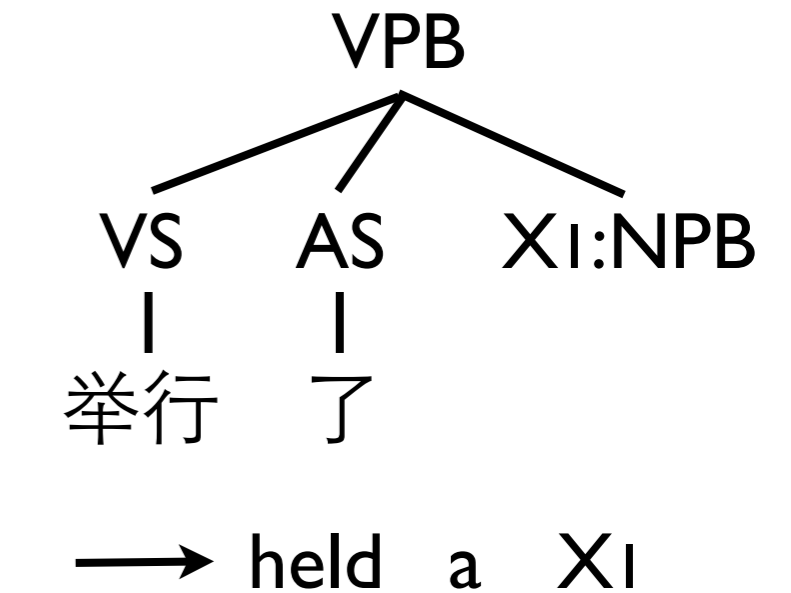
Joint Parsing and Translation



Bush

with Sharon

举行 了

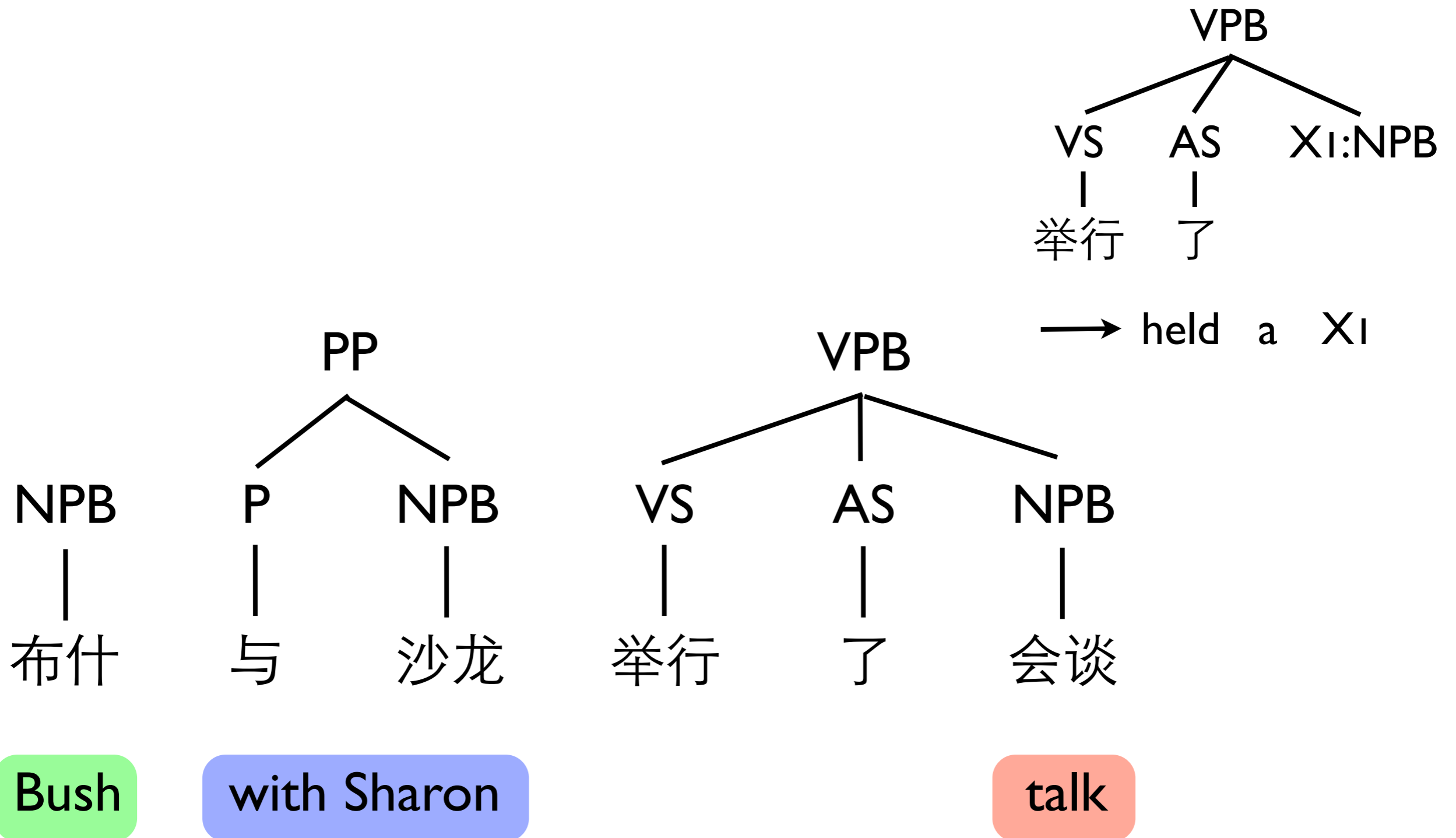


NPB
|
会谈

talk

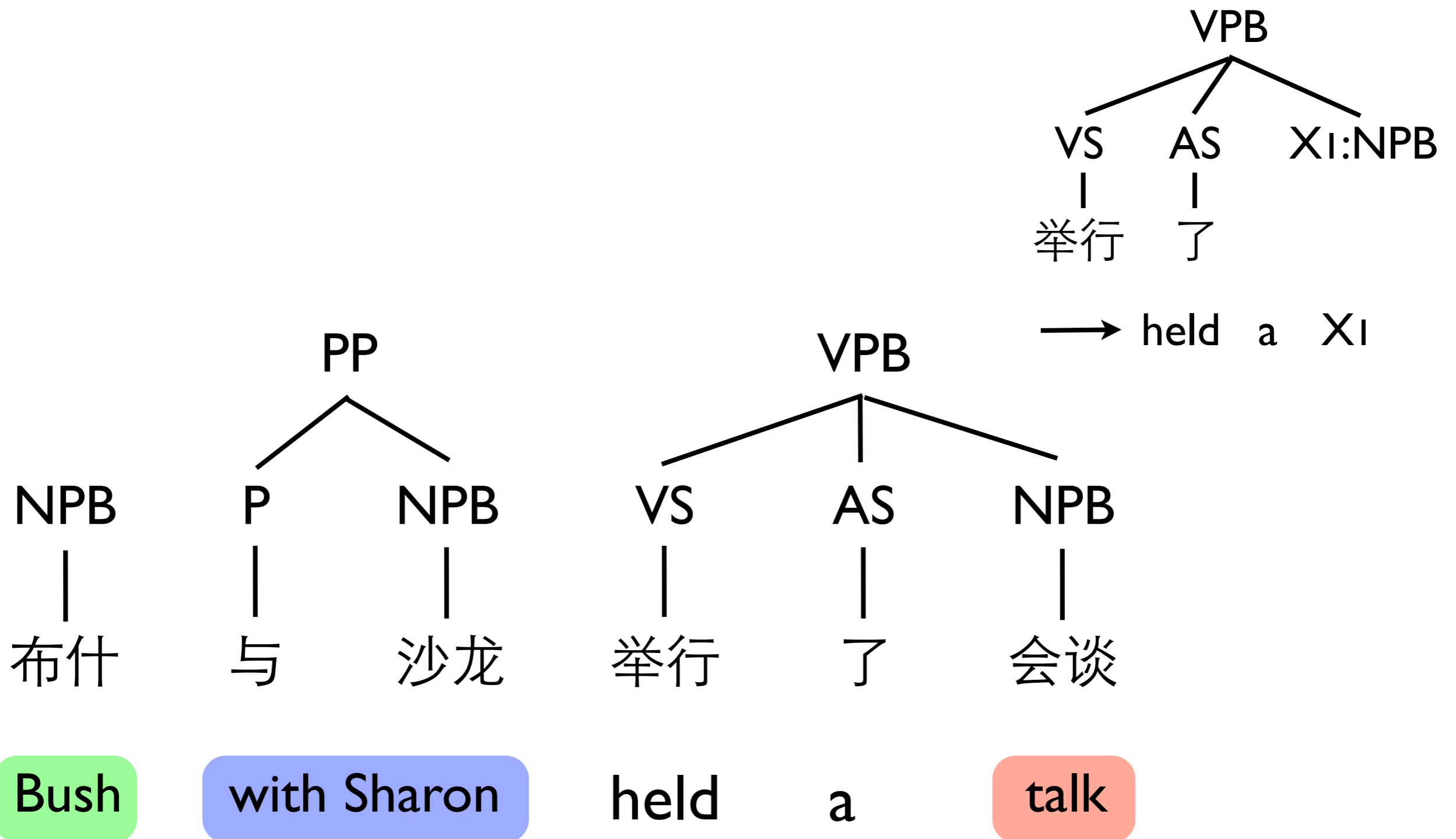
(Liu and Liu, 2010)

Joint Parsing and Translation



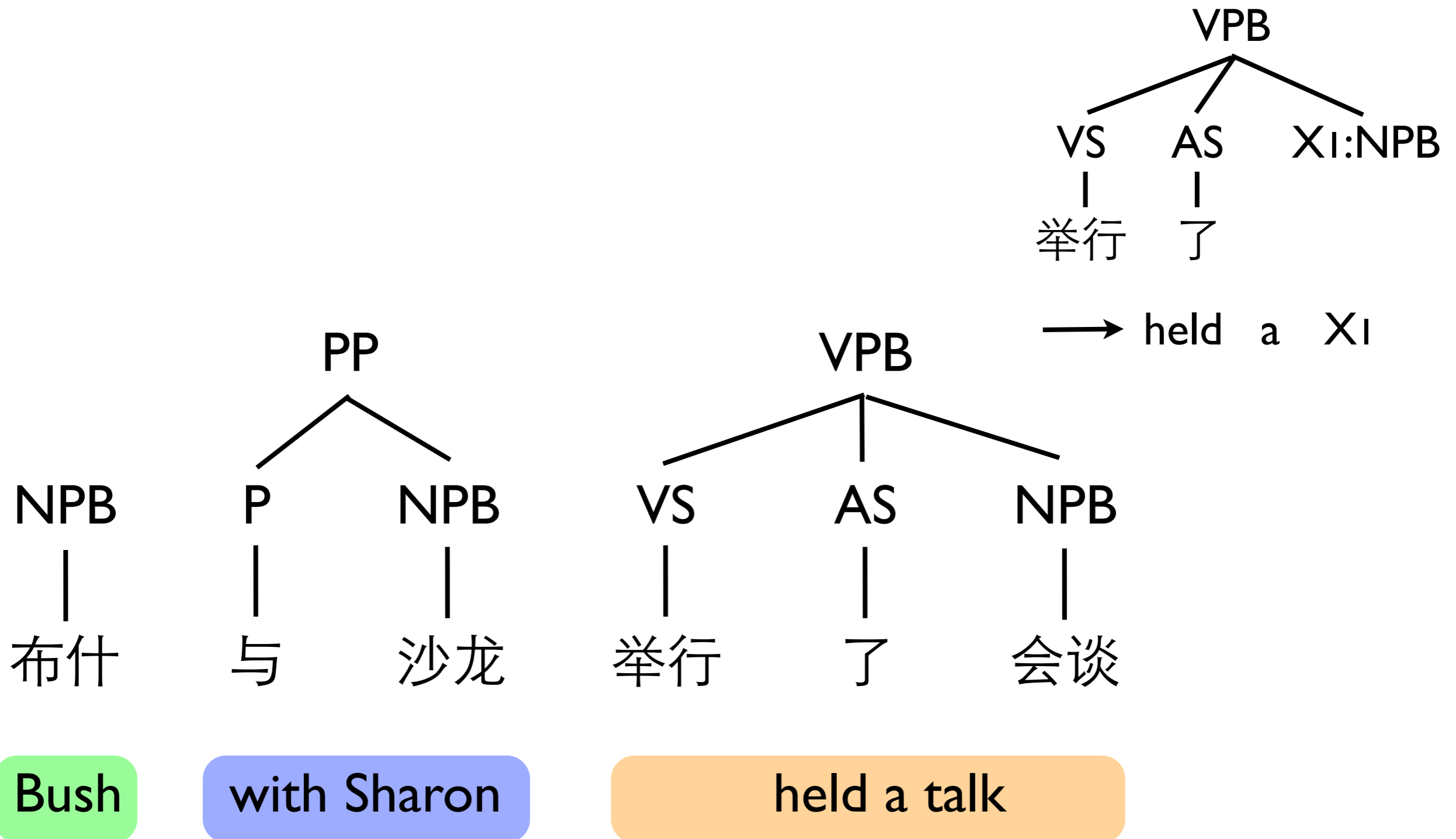
(Liu and Liu, 2010)

Joint Parsing and Translation

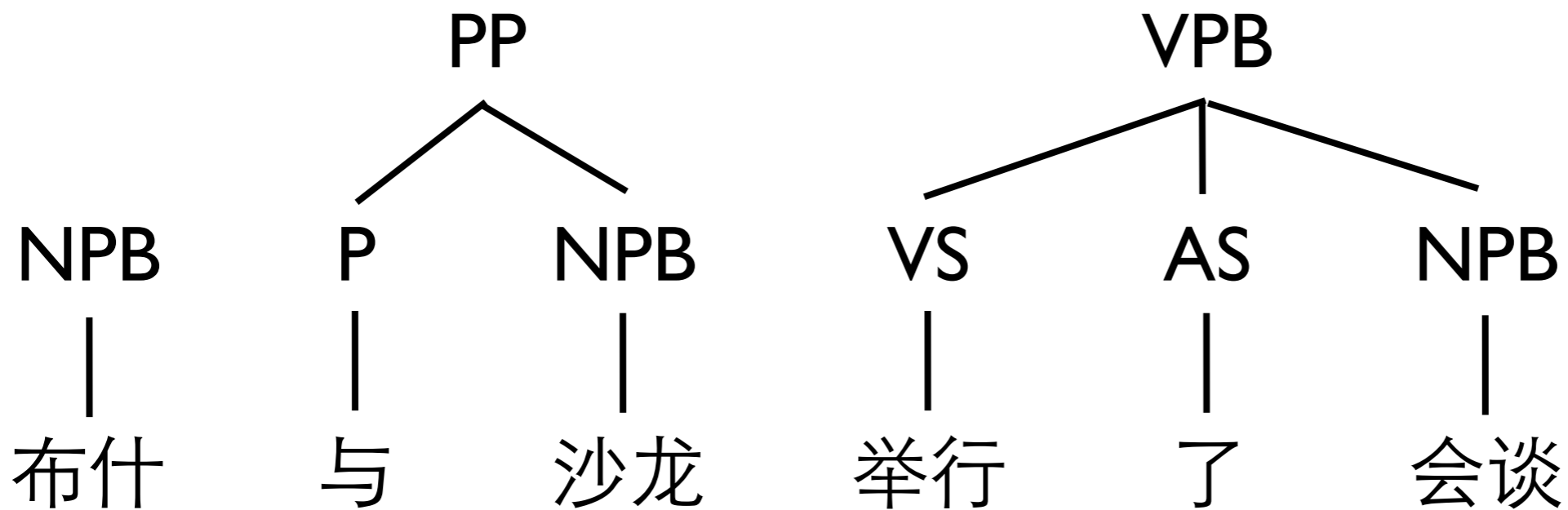


(Liu and Liu, 2010)

Joint Parsing and Translation



Joint Parsing and Translation

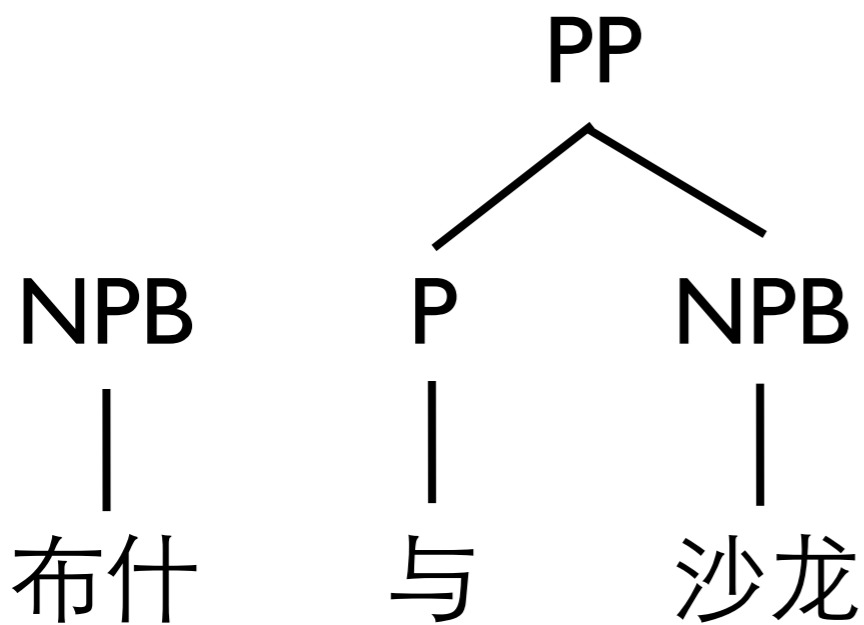
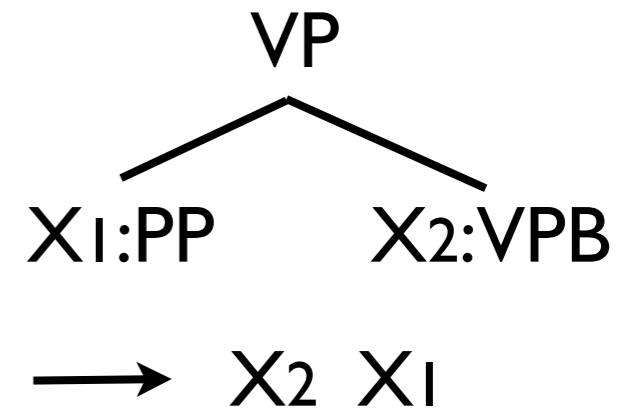


Bush

with Sharon

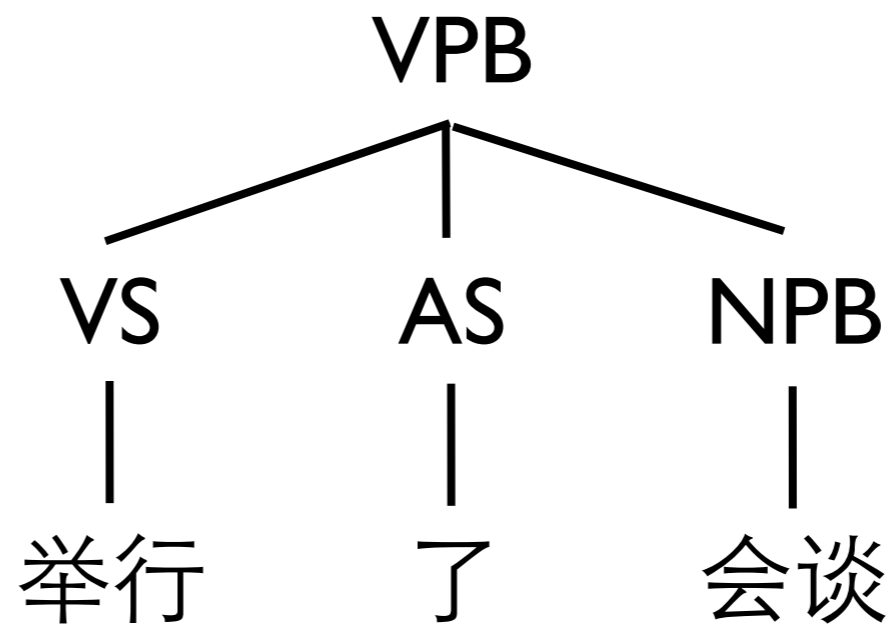
held a talk

Joint Parsing and Translation



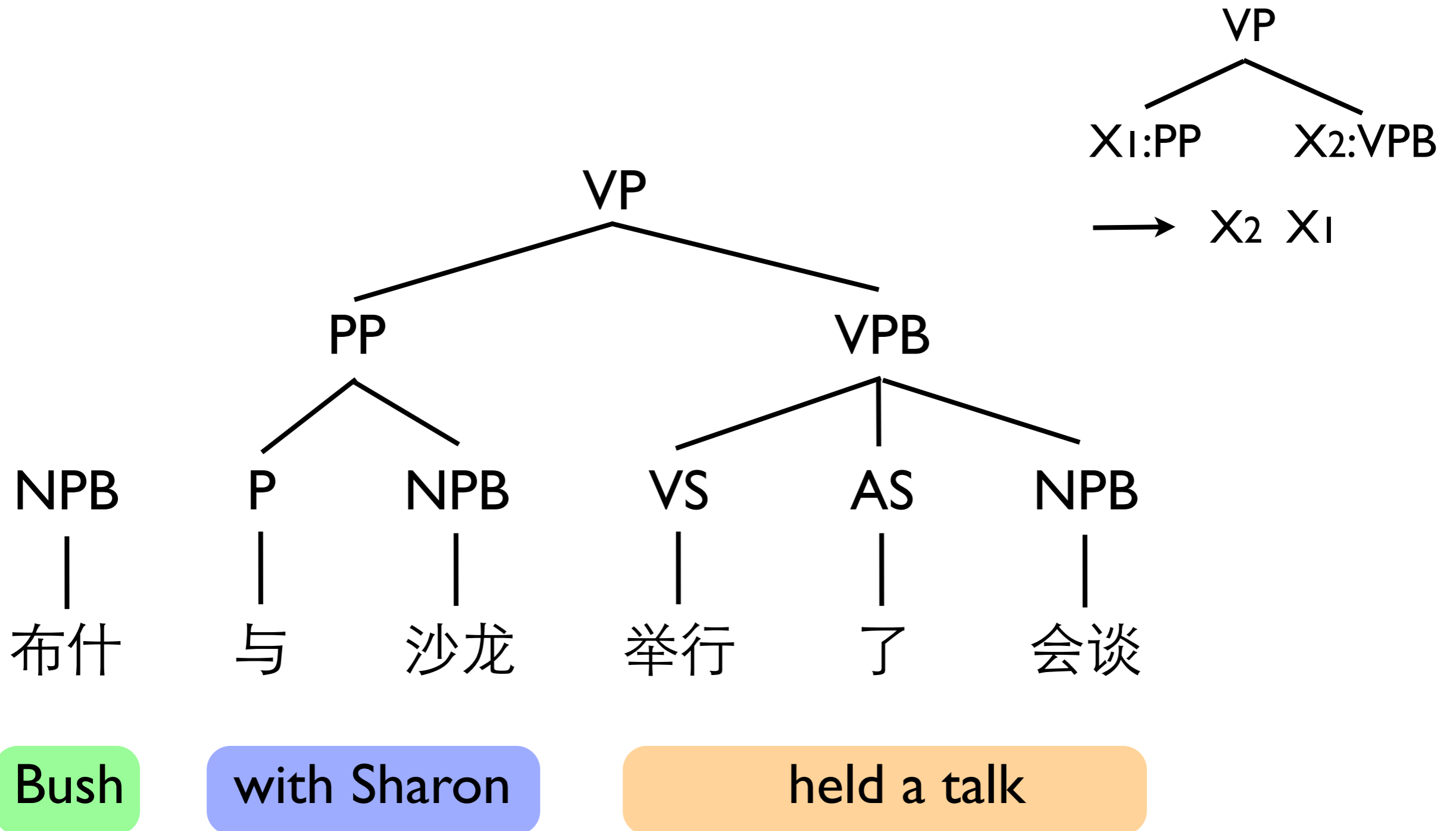
Bush

with Sharon

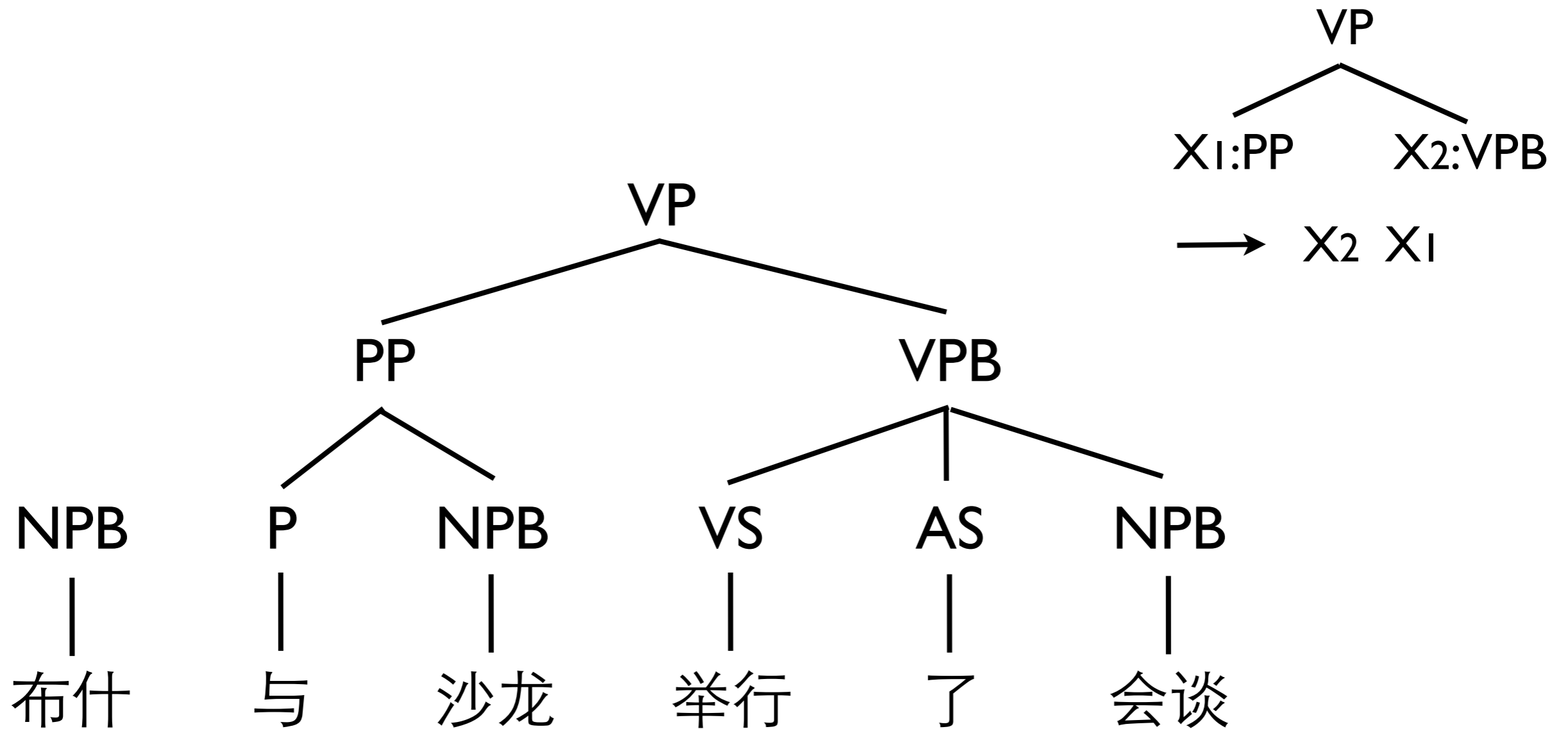


held a talk

Joint Parsing and Translation

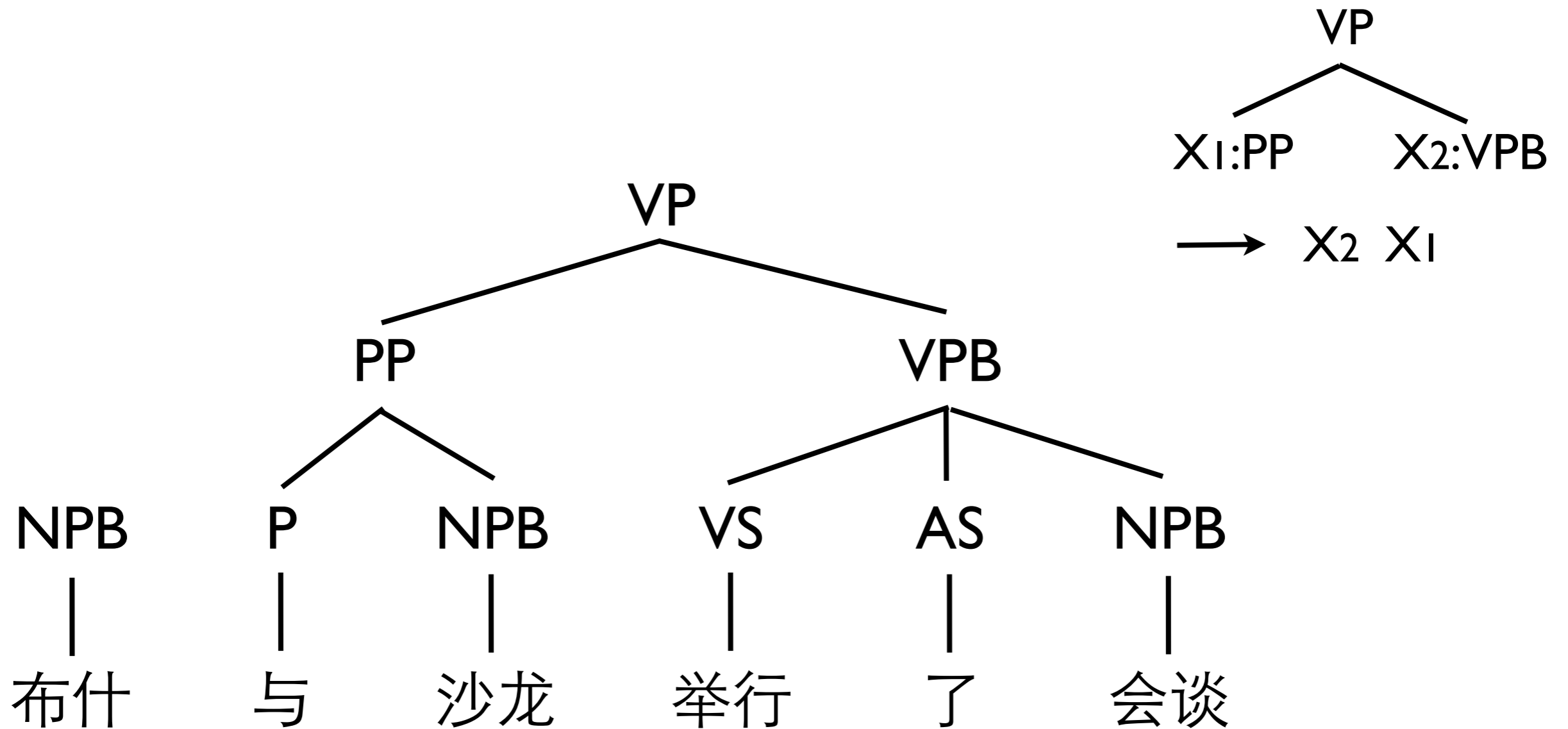


Joint Parsing and Translation



Bush

Joint Parsing and Translation

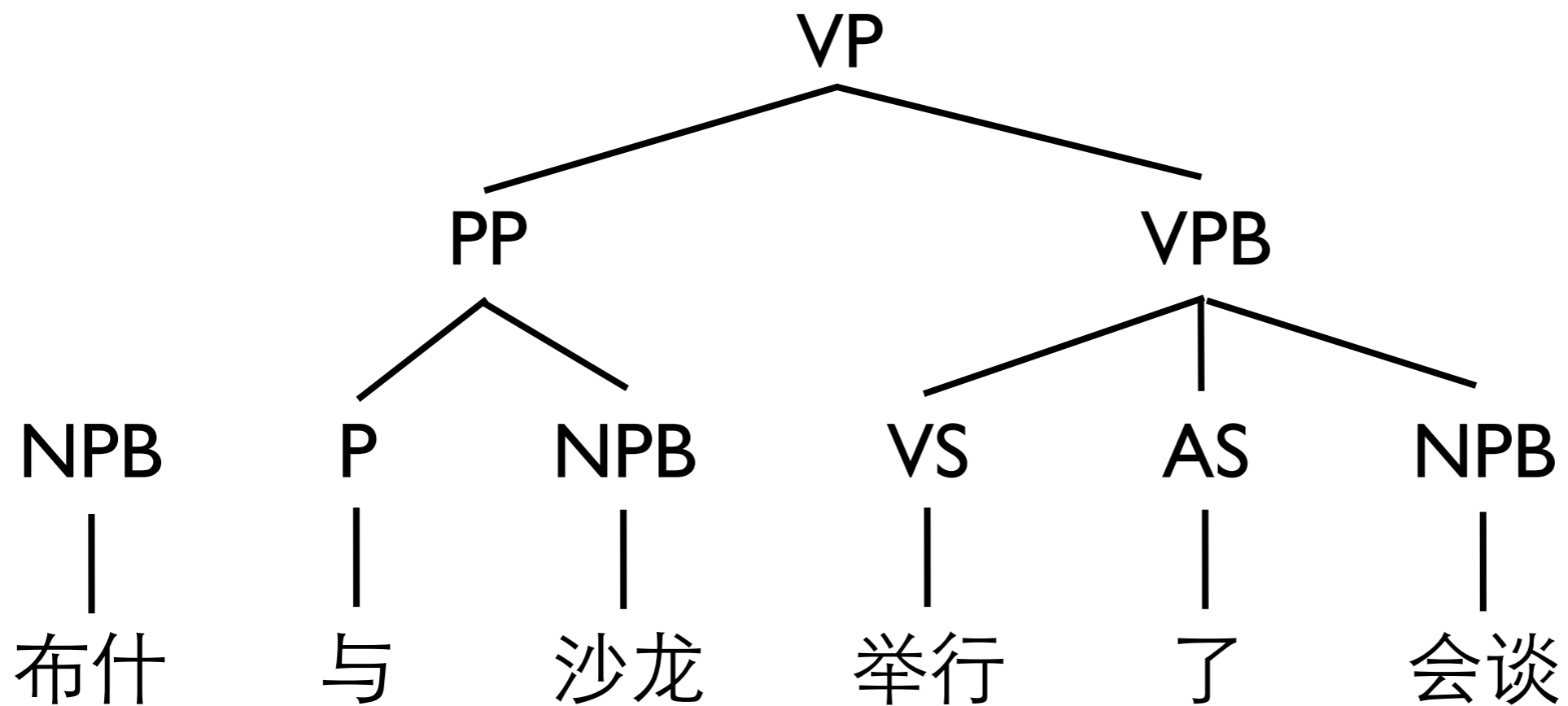


Bush

held a talk with Sharon

(Liu and Liu, 2010)

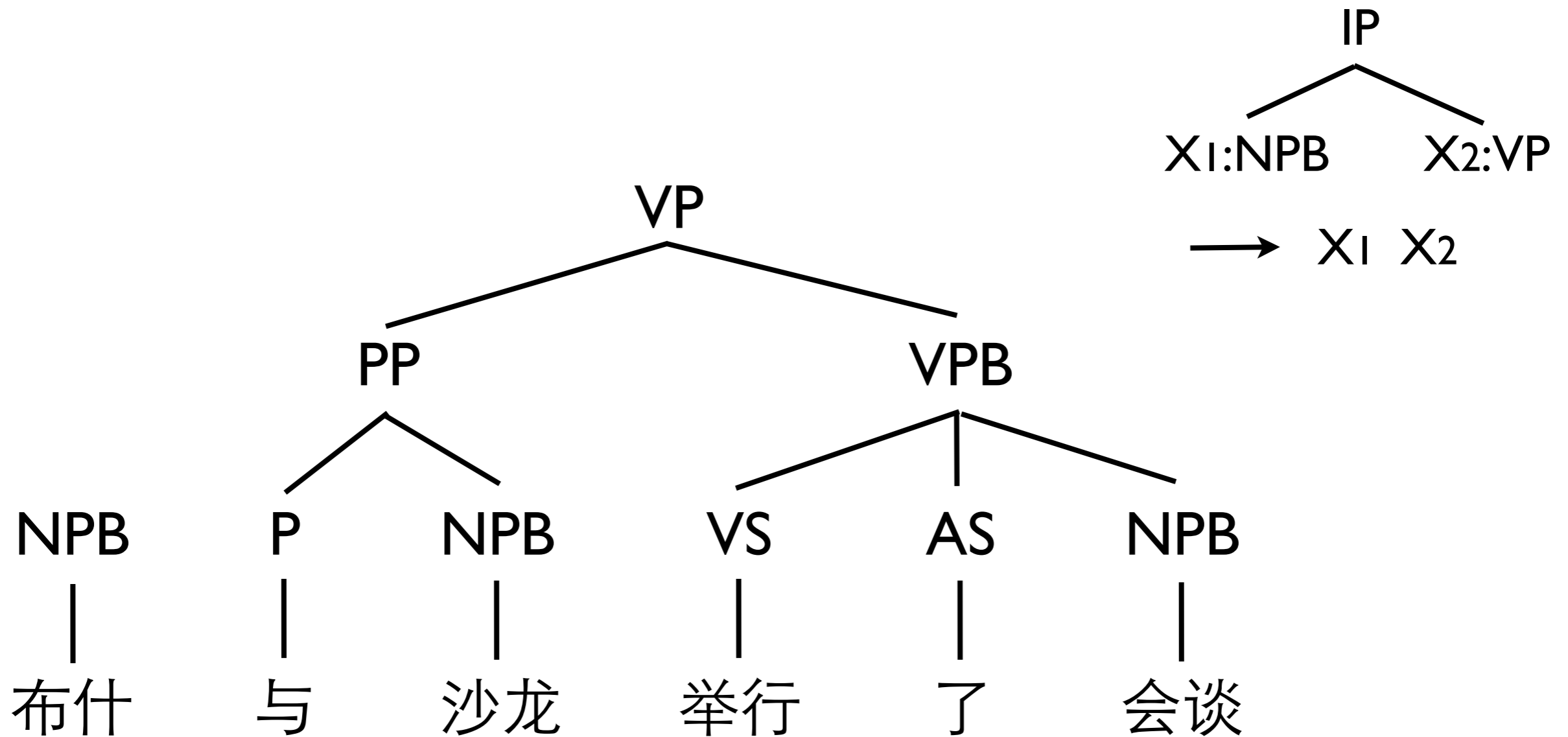
Joint Parsing and Translation



Bush

held a talk with Sharon

Joint Parsing and Translation

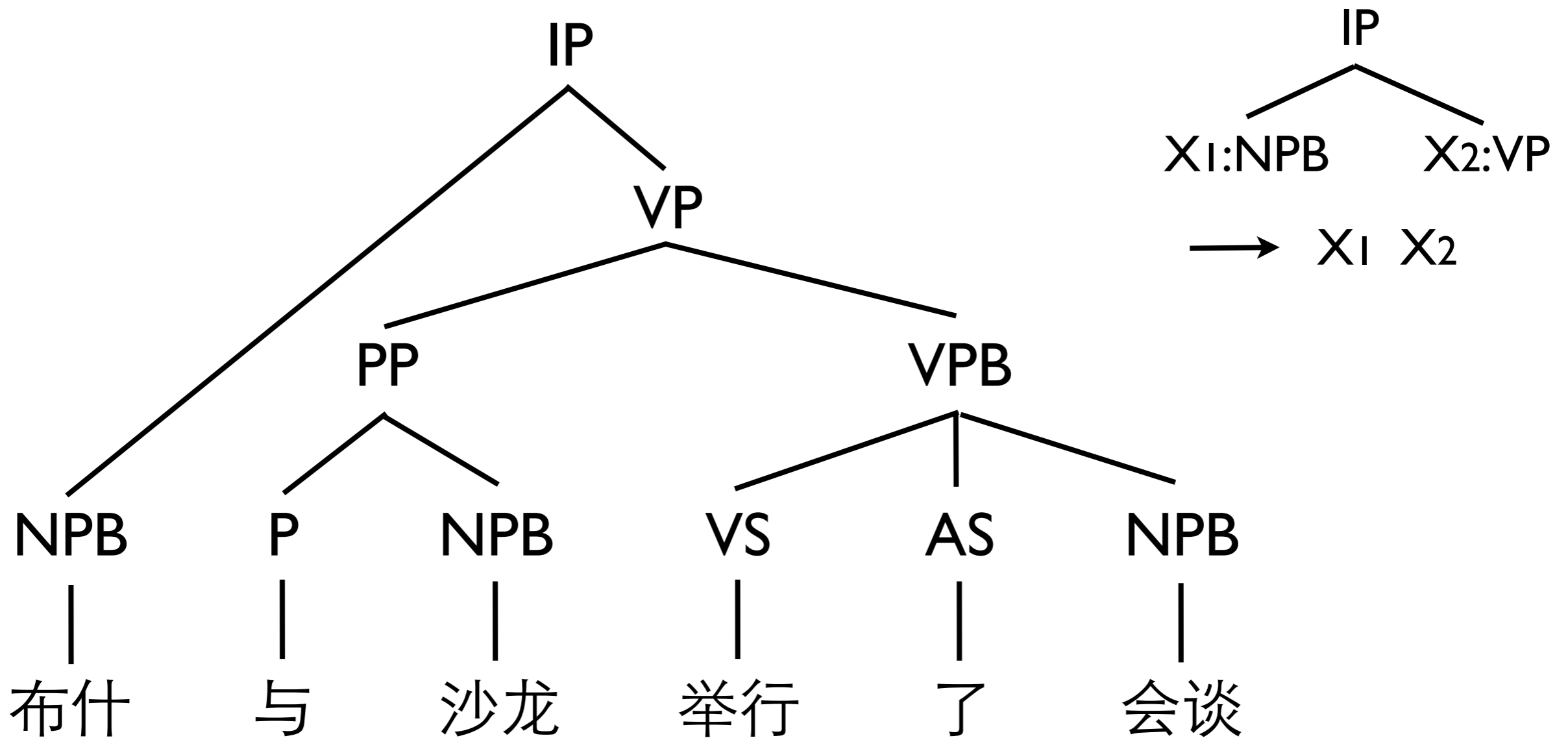


Bush

held a talk with Sharon

(Liu and Liu, 2010)

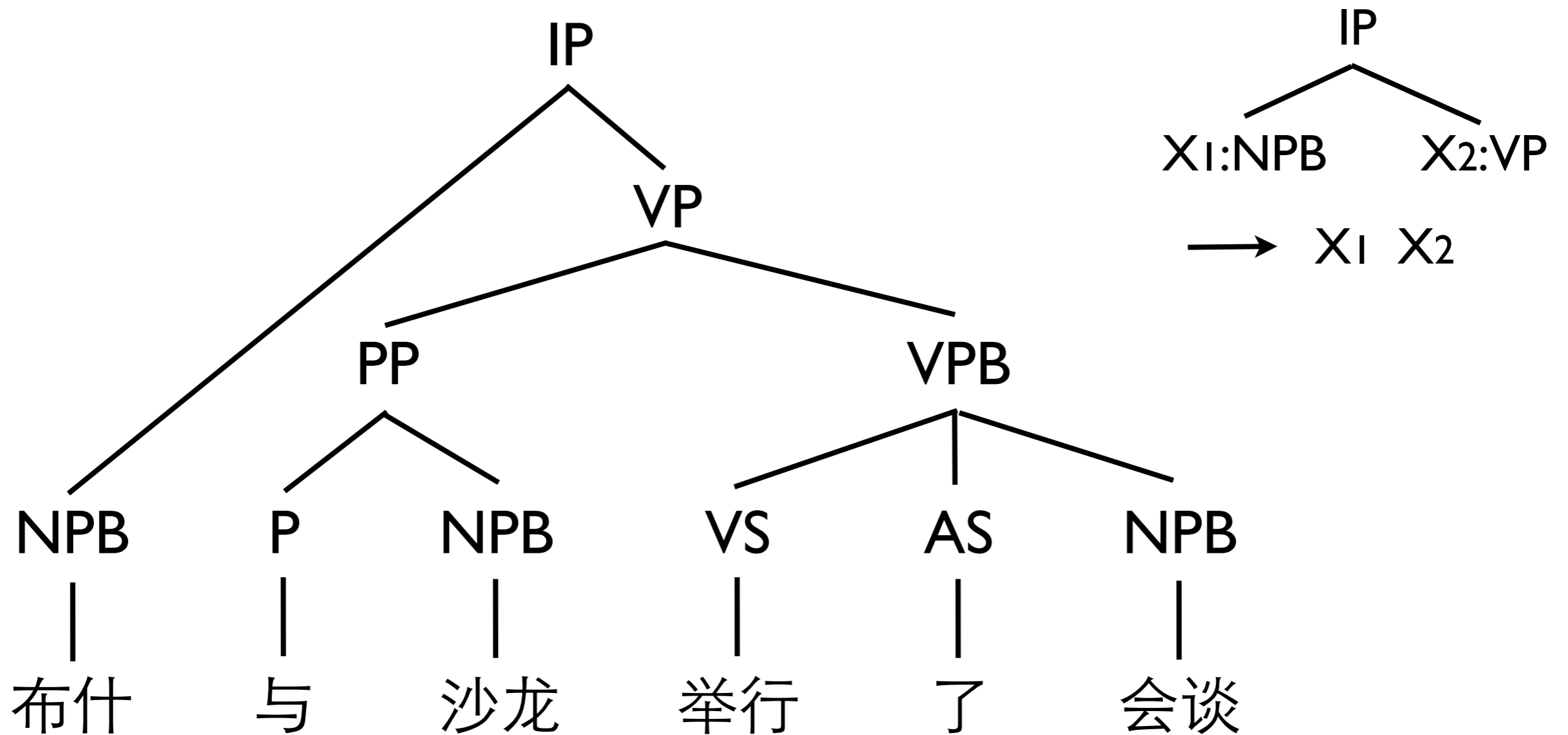
Joint Parsing and Translation



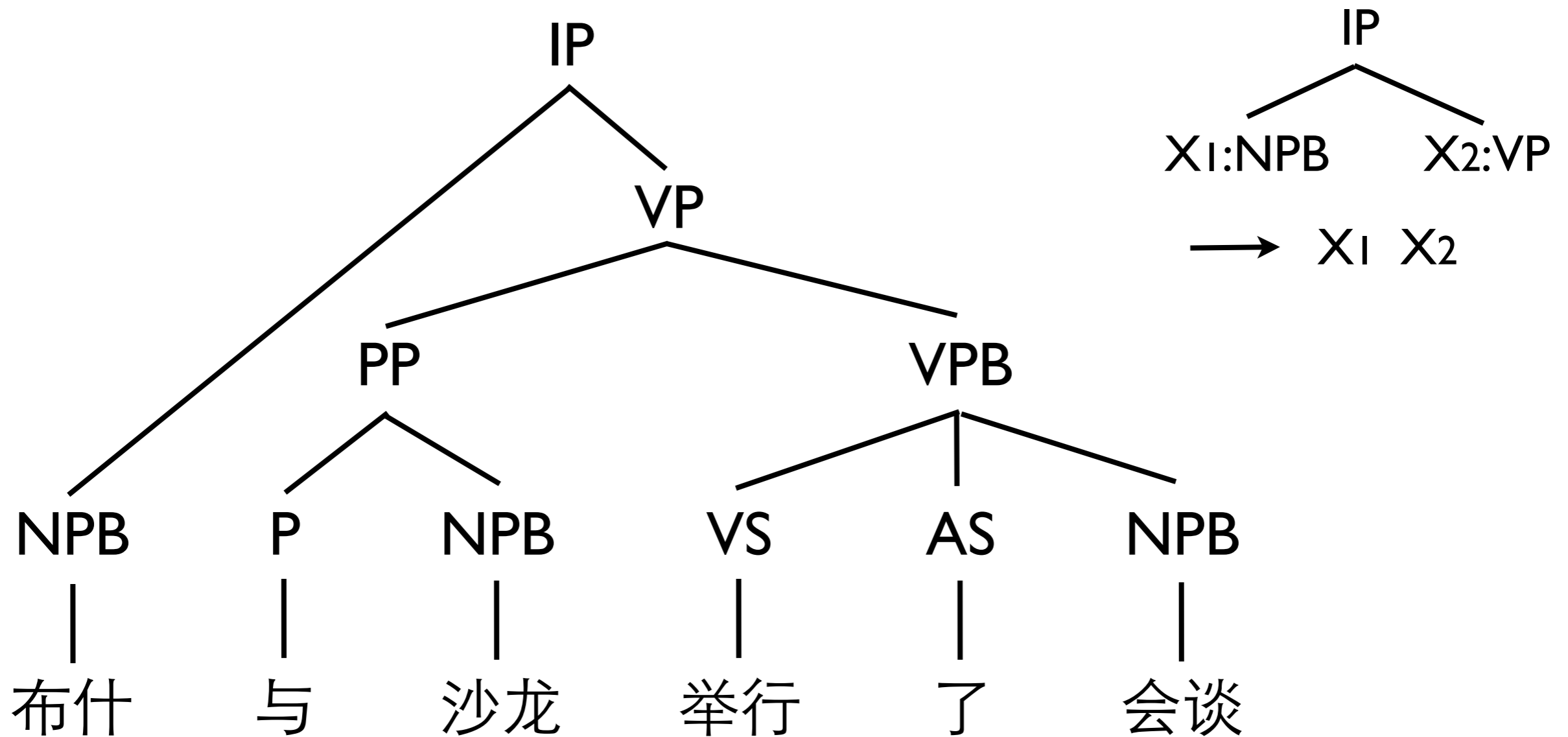
Bush

held a talk with Sharon

Joint Parsing and Translation



Joint Parsing and Translation



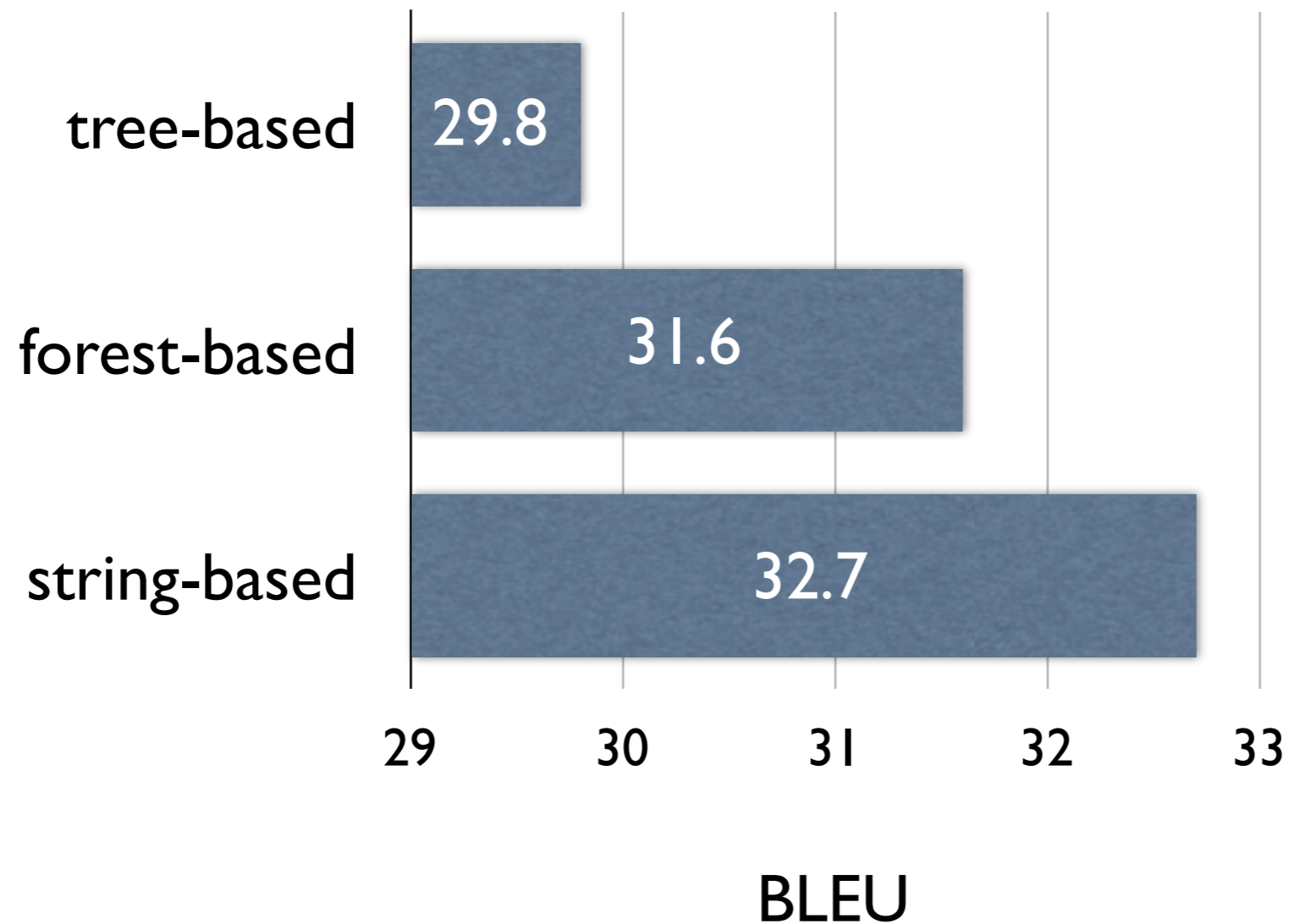
Bush held a talk with Sharon

Model and Training

$$score(\mathbf{f}, \pi, \mathbf{e}) = \sum_k \lambda_k h_k(\mathbf{f}, \pi, \mathbf{e})$$

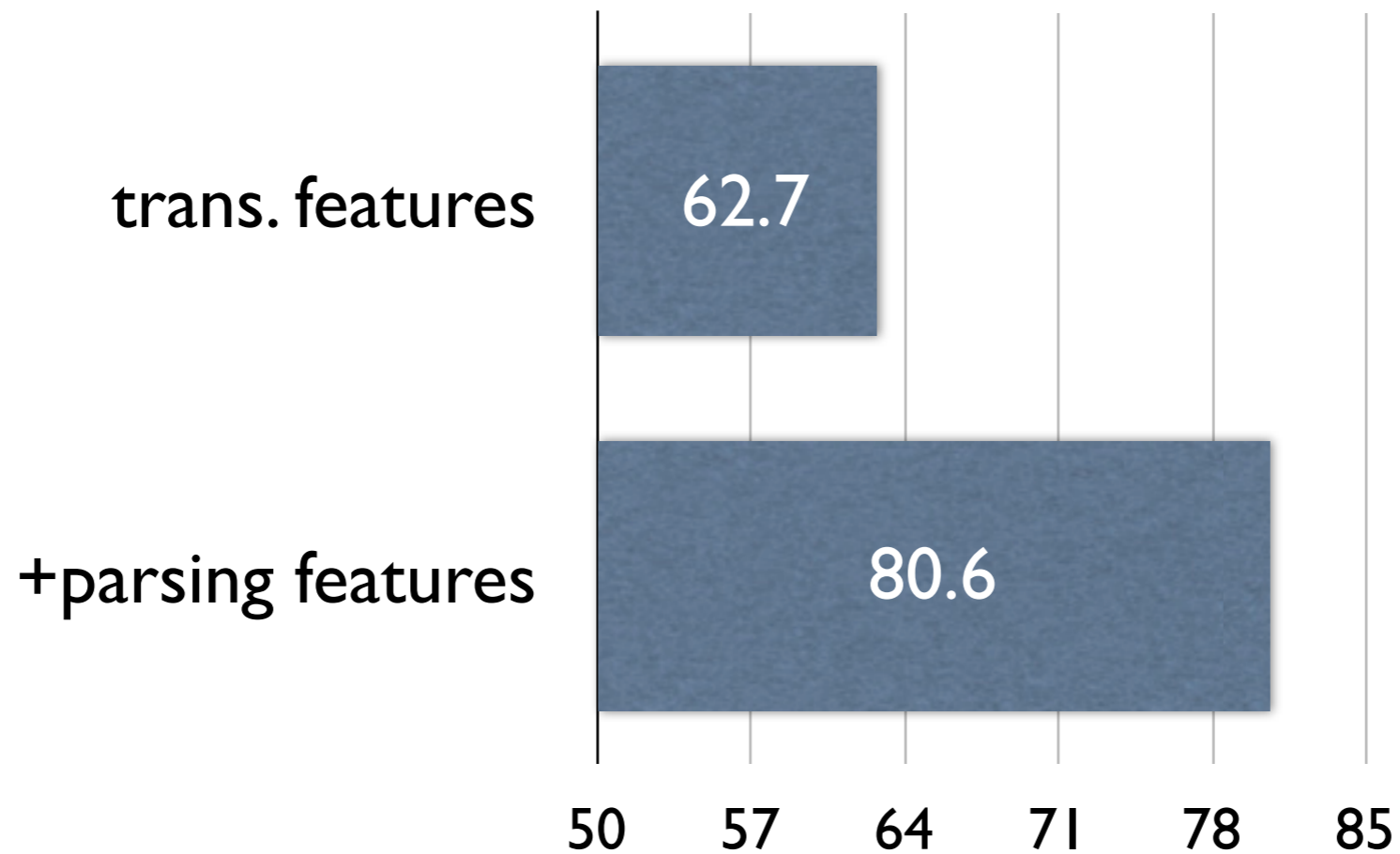
- We use a linear model to combine **parsing** and **translation** models as features
- The minimum-error-rate training algorithm is used for training feature weights

Translation Evaluation



(Liu and Liu, 2010)

Parsing Evaluation



(Liu and Liu, 2010)

Search Space Comparison

Search Space Comparison

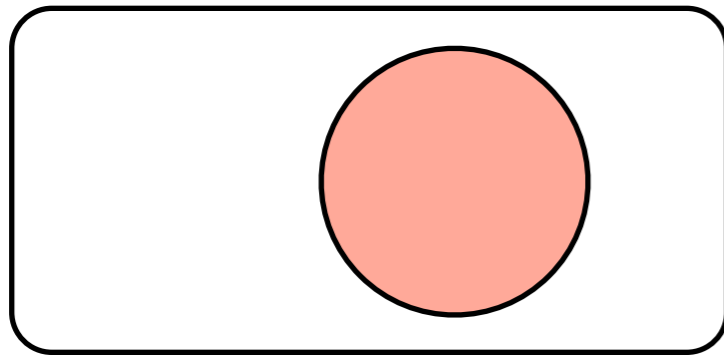
tree-based

Search Space Comparison



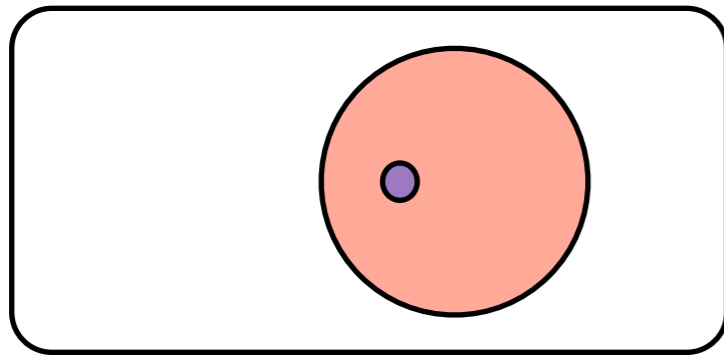
tree-based

Search Space Comparison



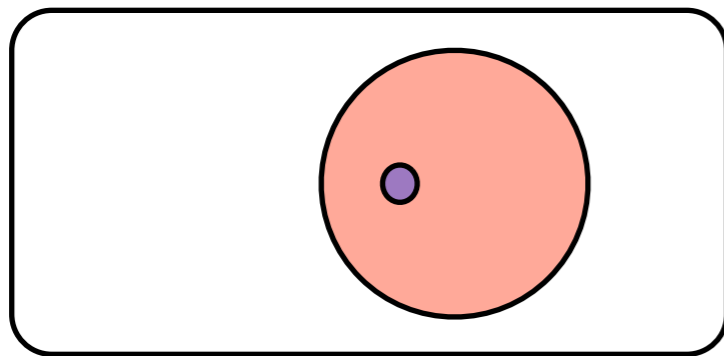
tree-based

Search Space Comparison

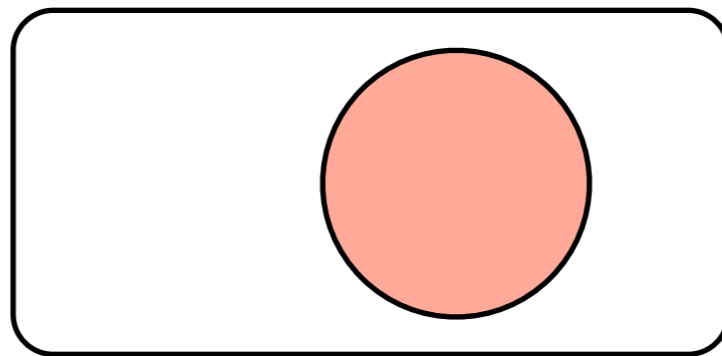


tree-based

Search Space Comparison

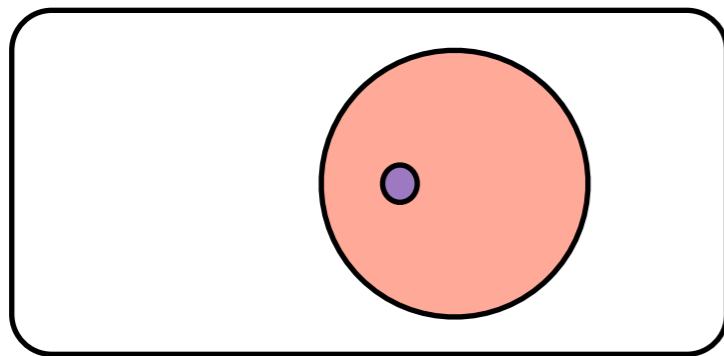


tree-based

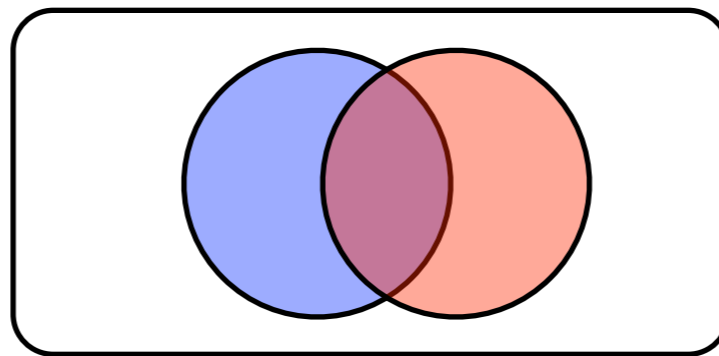


forest-based

Search Space Comparison

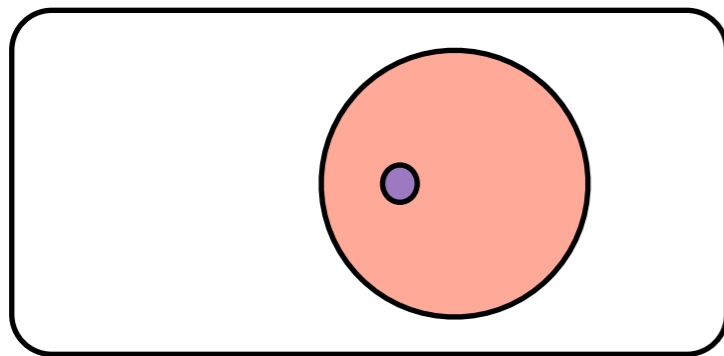


tree-based

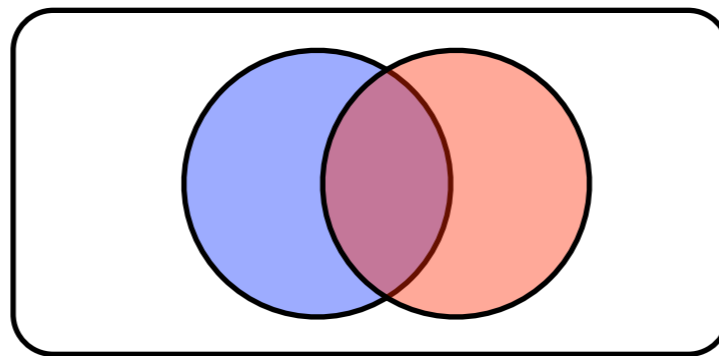


forest-based

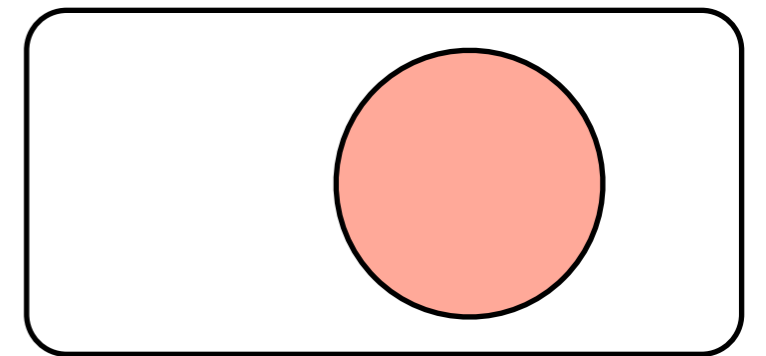
Search Space Comparison



tree-based

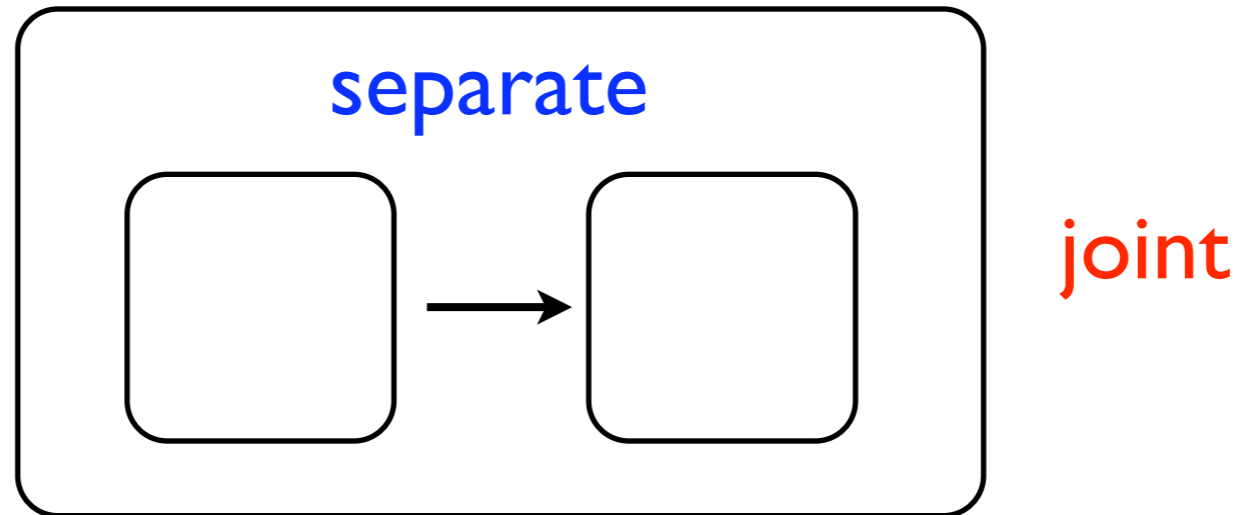


forest-based



string-based

Conclusion



- Joint decoding integrates separate steps in a pipeline to eliminate the mistake propagation problem
- Tokenization, parsing, and translation can interact with each other in a discriminative framework

Future Work



Thanks

Backup Slides for Q&A

Better Tokenization = Better Translation?

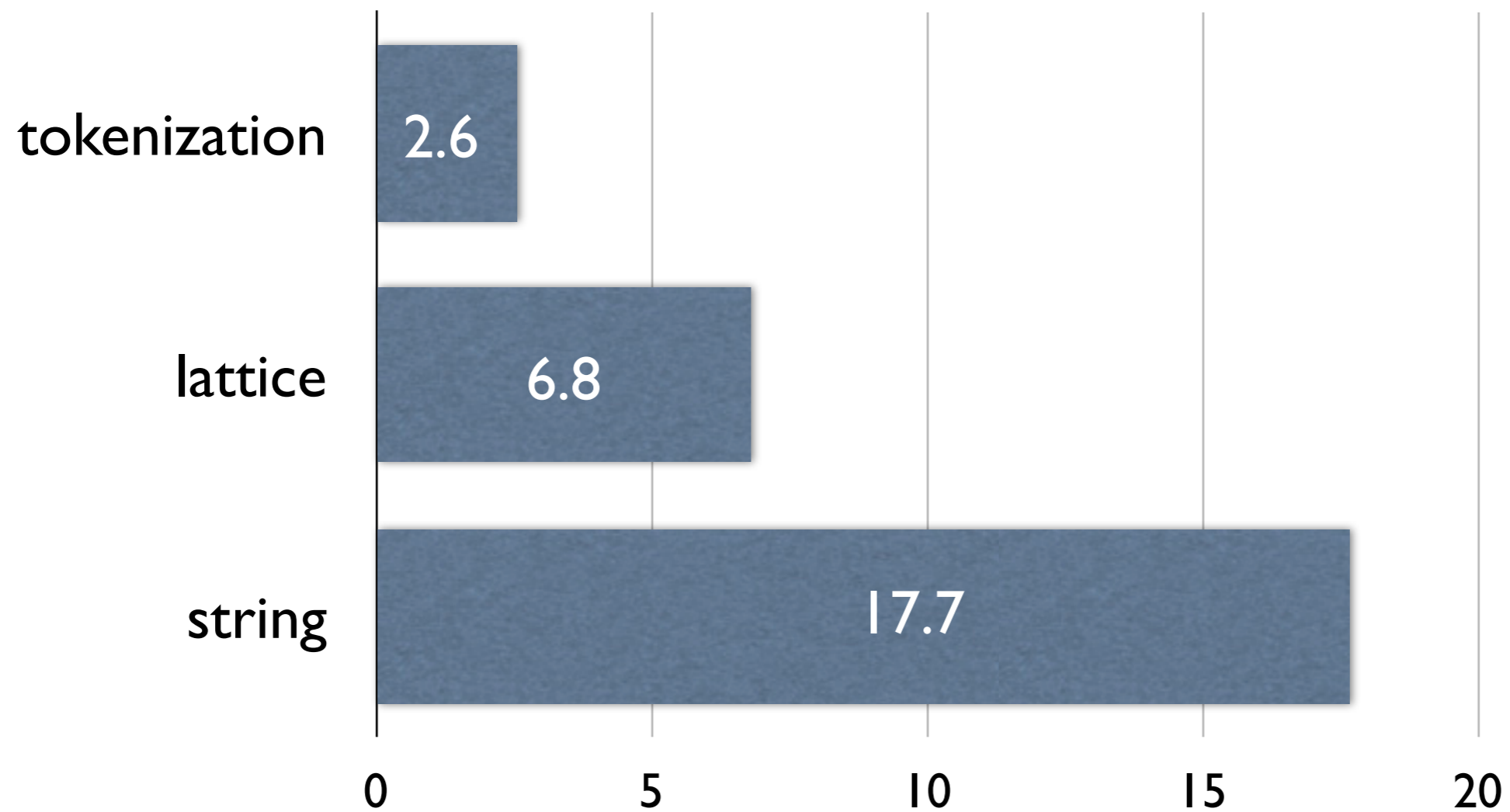
decoding

training

	F	BLEU
Max F	97.37	27.43
Max BLEU	92.49	34.88

(Xiao et al., 2010)

Speed Comparison



Speed Comparison (cont.)

