Tree-to-String Alignment Template for Statistical Machine Translation

Yang Liu, Qun Liu, and Shouxun Lin
Institute of Computing Technology
Chinese Academy of Sciences

INSTITUTE OF COMPUTING TECHNOLOGY 中科院计算所

Outline

- Introduction
- Tree-to-String Alignment Template
- Training
- Decoding
- Experiments
- Recent Advances
- Conclusion and Future Work

Current Situation of SMT

INSTITUTE OF COMPUTING TECHNOLOGY

- Phrase-based models are state-of-the-art
  - Och and Ney, 2004
- Syntax-based models are in rapid development
  - Wu, 1997
  - Alshawi et al., 2000
  - Yamada and Knight, 2001
  - Melamed, 2004
  - Galley et al., 2004
  - Graehl and Knight, 2004
  - Chiang, 2005
  - Quirk et al., 2005
  - Ding and Palmer, 2005

Challenges to Syntax-based Models

INSTITUTE OF COMPUTING TECHNOLOGY

- Most syntax-based models do not show improvement over phrase-based ones. Why?
  - complexity
  - non-isomorphism
  - non-perfect training data
  - …

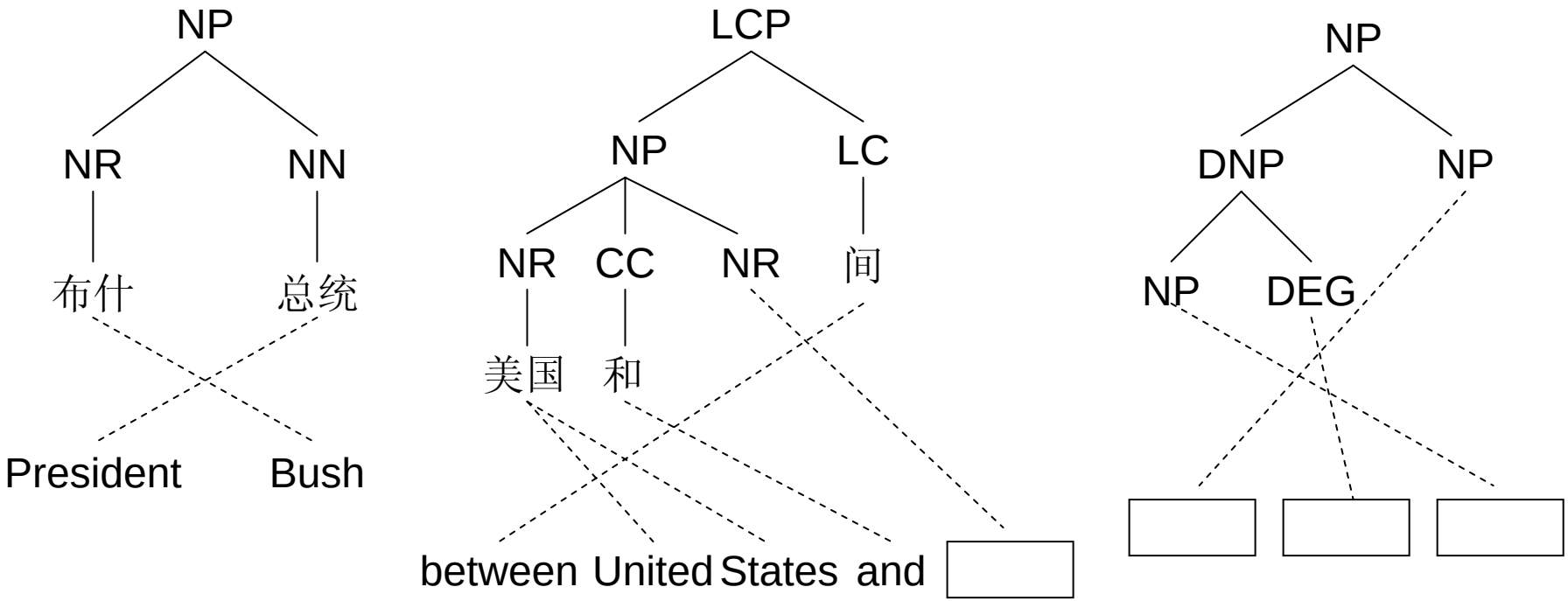
Our Work

- Goal: simple and powerful
- Key: tree-to-string alignment template (TAT)
- Distinctions from previous work:
  - Model the syntax of the source language
  - Exploit bilingual phrases to strengthen the TAT-based model

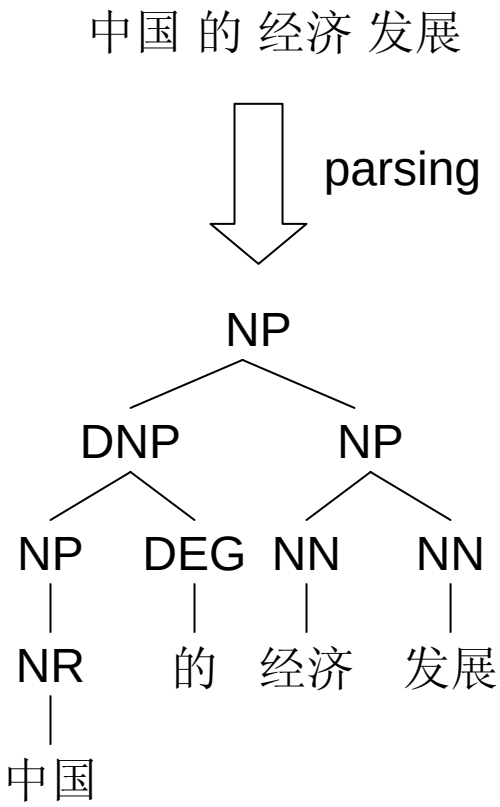
Outline

- *Introduction*
- Tree-to-String Alignment Template
- Training
- Decoding
- Experiments
- Recent Advances
- Conclusion and Future Work

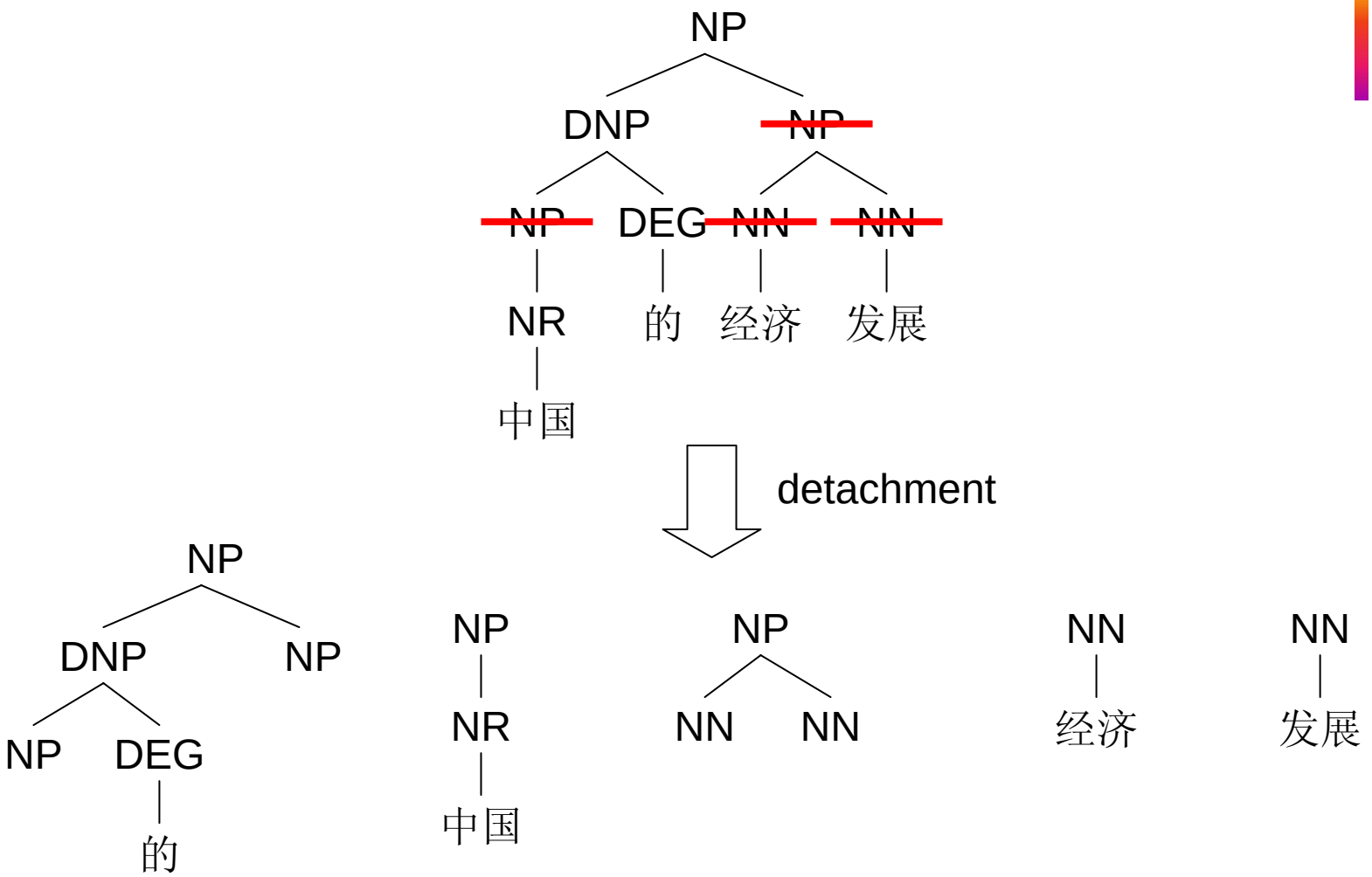
Tree-to-String Alignment Template



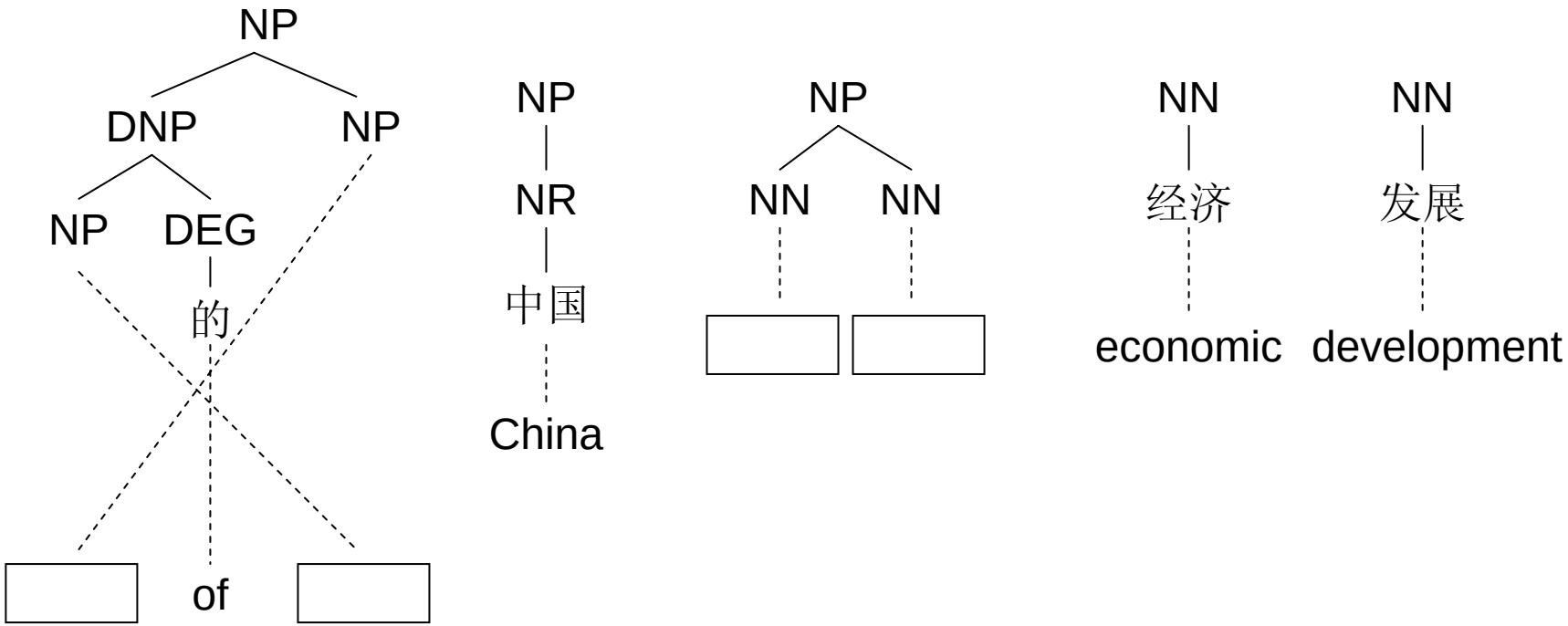
Translation Process : Parsing



Translation Process : Detachment

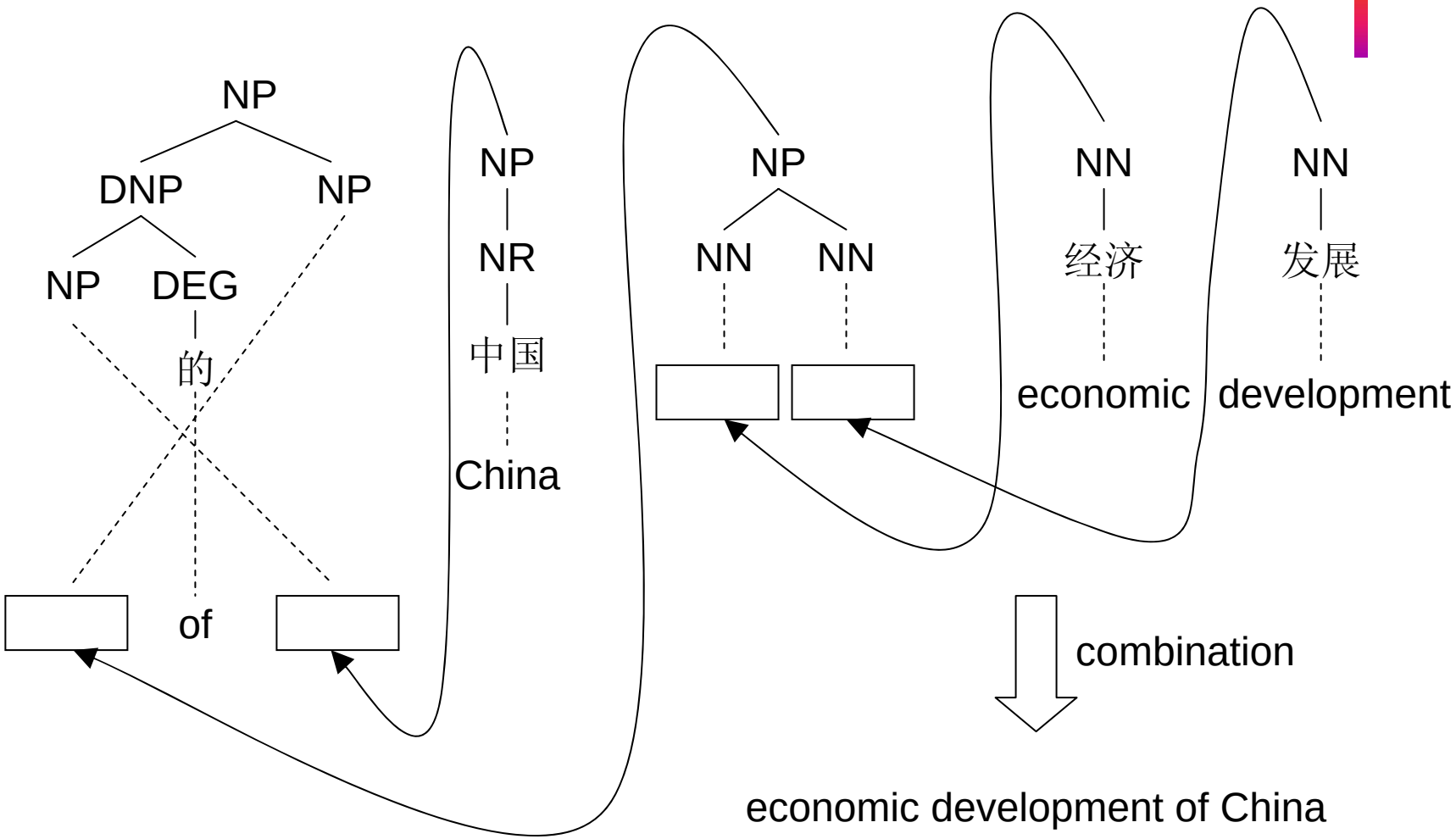


Translation Process : Production



Go to page 27

Translation Process : Combination



Log-linear Model

$$Pr(e_1^I, z_1^K | f_1^J)$$

$$= \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J, z_1^K)]}{\sum_{e'^I_1, z'^K_1} \exp[\sum_{m=1}^{M} \lambda_m h_m(e'^I_1, f_1^J, z'^K_1)]}$$

$$\hat{e}_1^I = \underset{e_1^I, z_1^K}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J, z_1^K) \right\}$$

Feature Functions

p(e|f)

$$h_1(e_1^I, f_1^J) = \log \prod_{k=1}^{K} \frac{N(z) \cdot \delta(T(z), \tilde{T}_k)}{N(T(z))}$$

p(f|e)

$$h_2(e_1^I, f_1^J) = \log \prod_{k=1}^{K} \frac{N(z) \cdot \delta(T(z), \tilde{T}_k)}{N(S(z))}$$

lex(f|e)

$$h_3(e_1^I, f_1^J) = \log \prod_{k=1}^{K} lex(T(z)|S(z)) \cdot \delta(T(z), \tilde{T}_k)$$

lex(e|f)

$$h_4(e_1^I, f_1^J) = \log \prod_{k=1}^{K} lex(S(z)|T(z)) \cdot \delta(T(z), \tilde{T}_k)$$

TAT penalty

$$h_5(e_1^I, f_1^J) = K$$

trigram LM

$$h_6(e_1^I, f_1^J) = \log \prod_{i=1}^{I} p(e_i|e_{i-2}, e_{i-1})$$

word penalty

$$h_7(e_1^I, f_1^J) = I$$

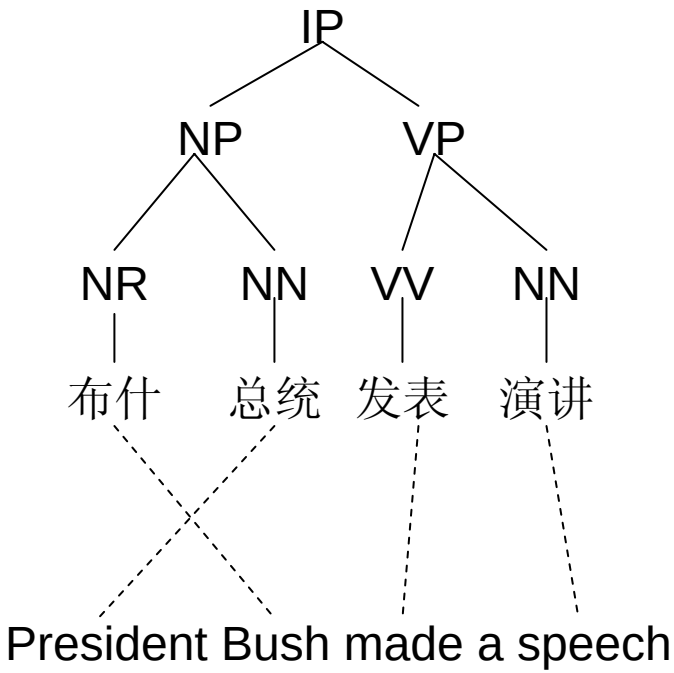
Outline

- *Introduction*
- *Tree-to-String Alignment Template*
- Training
- Decoding
- Experiments
- Recent Advances
- Conclusion and Future Work

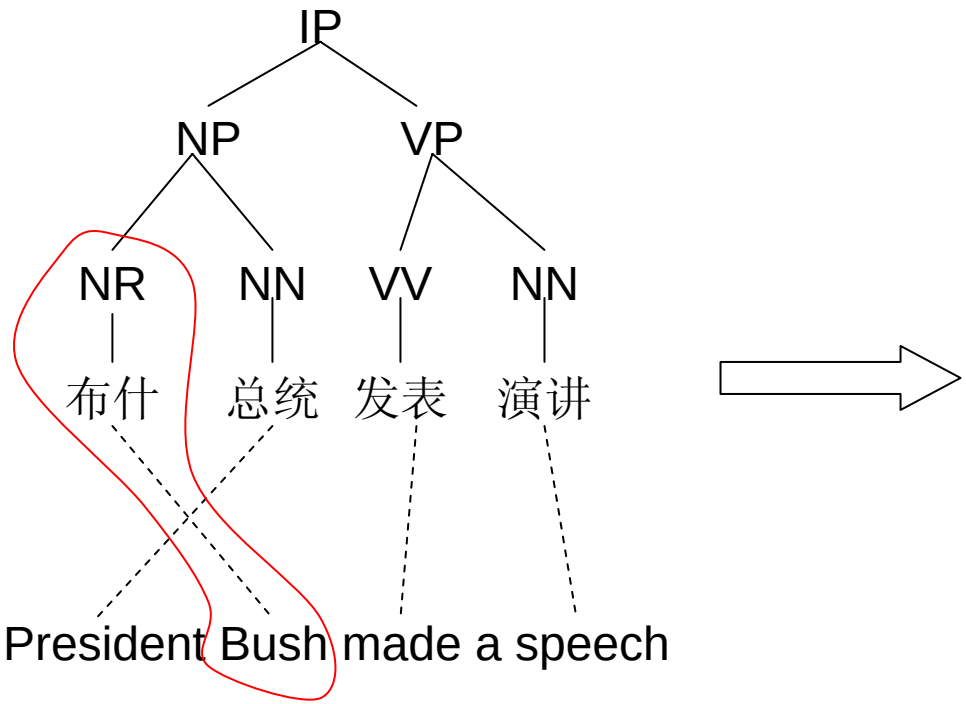
Extract TATs

- TSA (Tree-String-Alignment)
  - bilingual phrase with tree over the source string
- Bottom-up strategy
- Impose several restrictions to reduce the magnitude
  - maximal height of the tree
  - maximal number of children of a node
  - both the head and tail of target string must be aligned

An Example

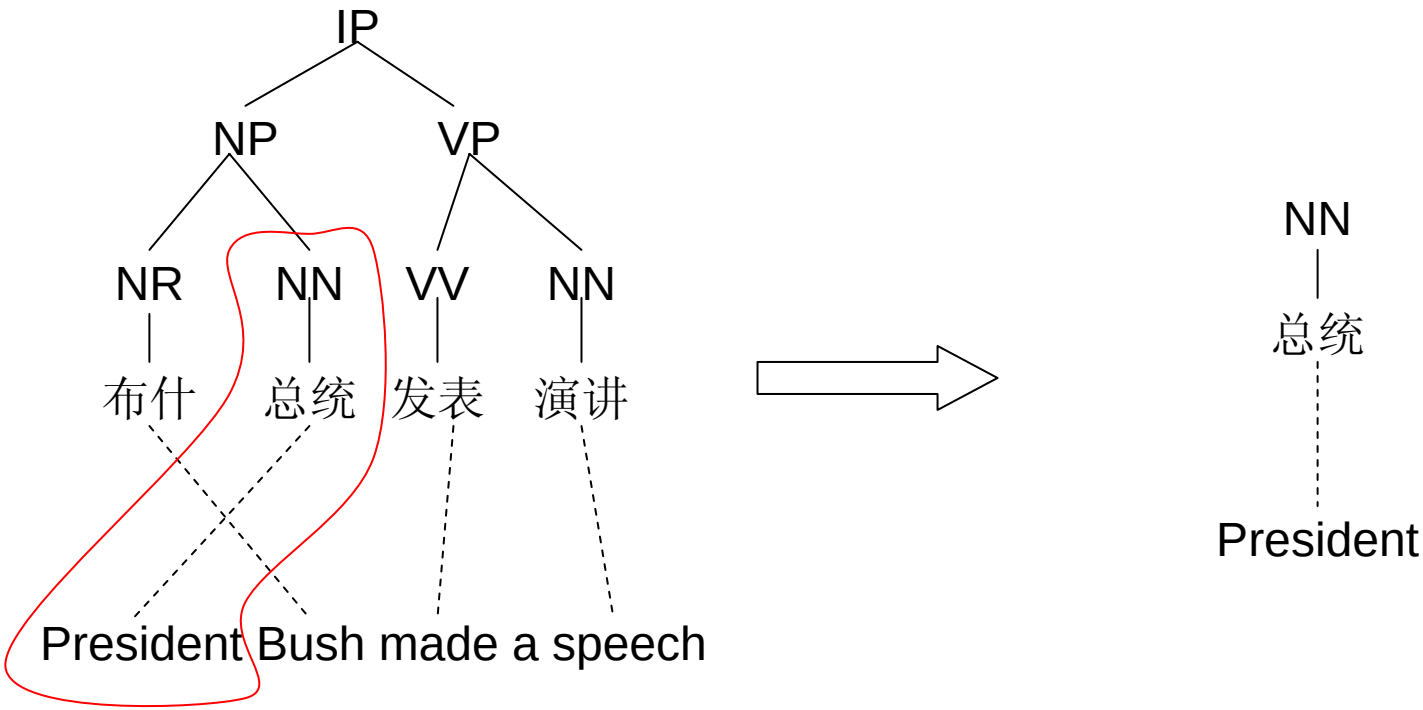


An Example

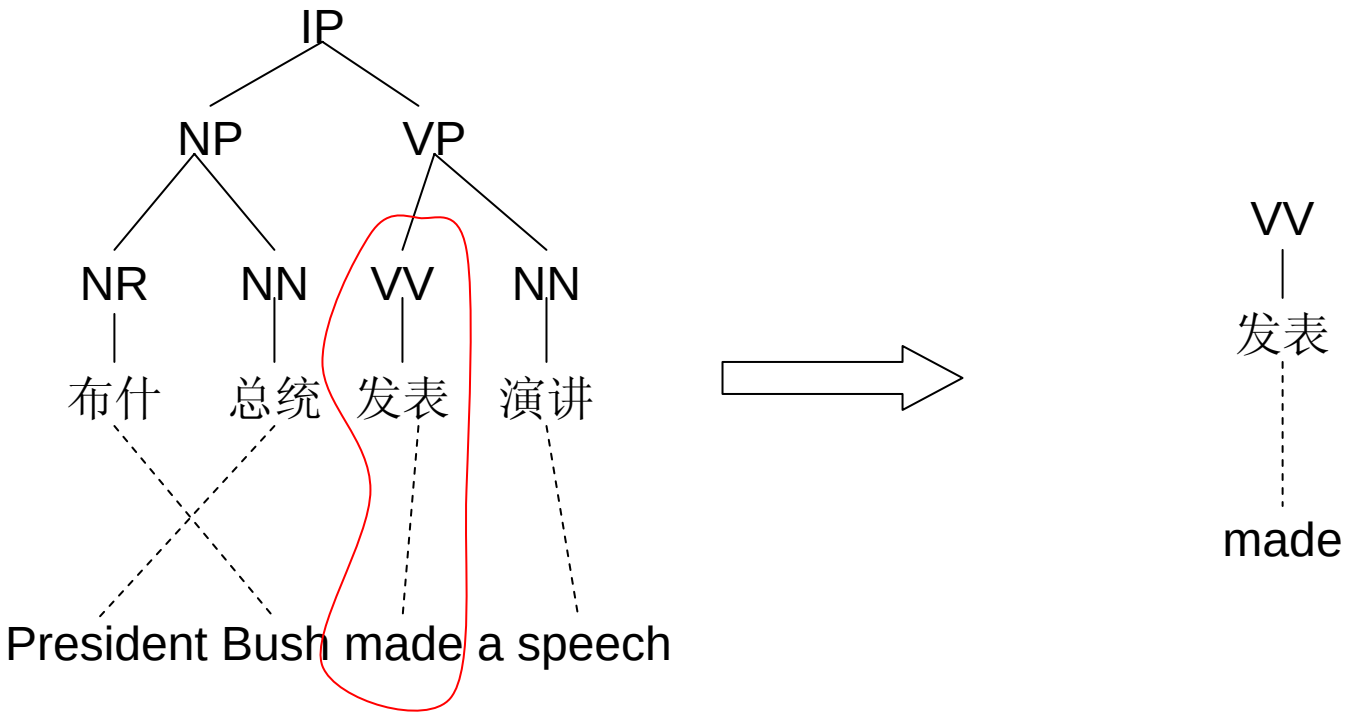

IP
NP
VP
NR
NN
VV
NN
布什
总统
发表
演讲
President Bush made a speech


NR
布什
Bush

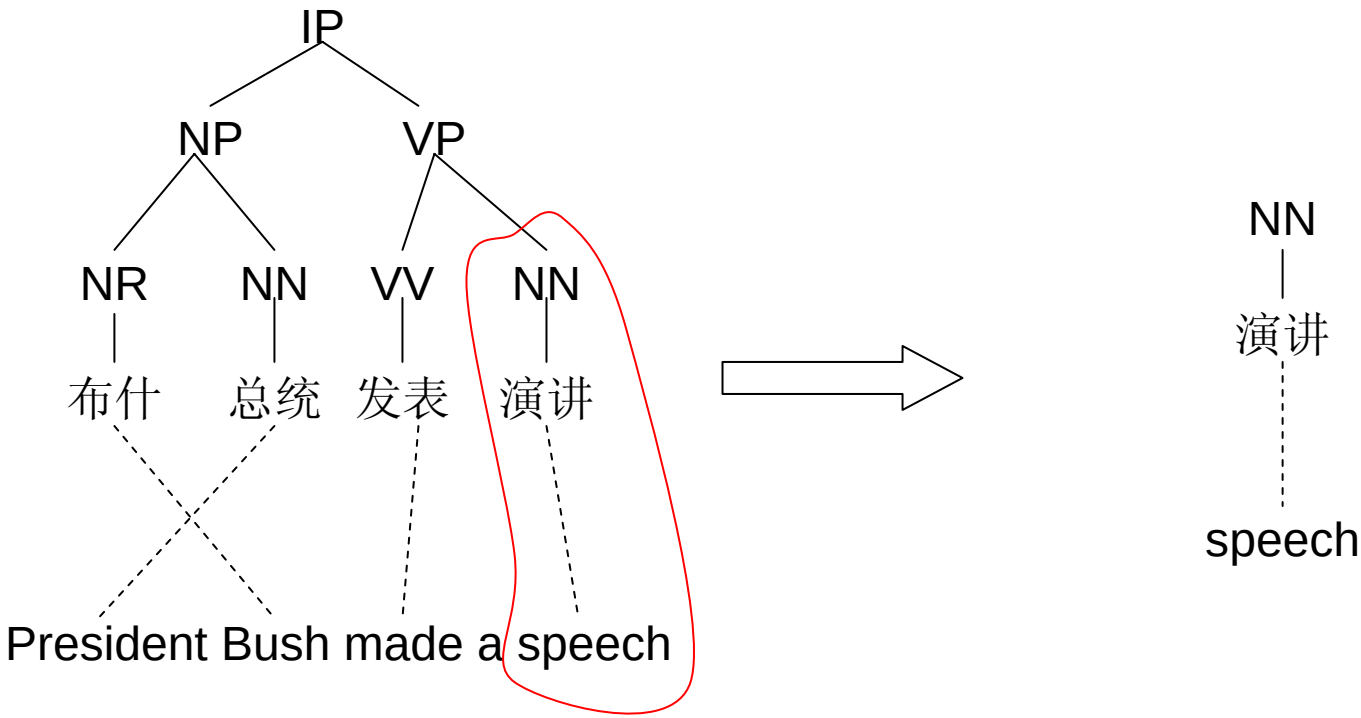
An Example


IP
NP
VP
NR
NN
VV
NN
布什
总统
发表
演讲
President Bush made a speech
NN
总统
President

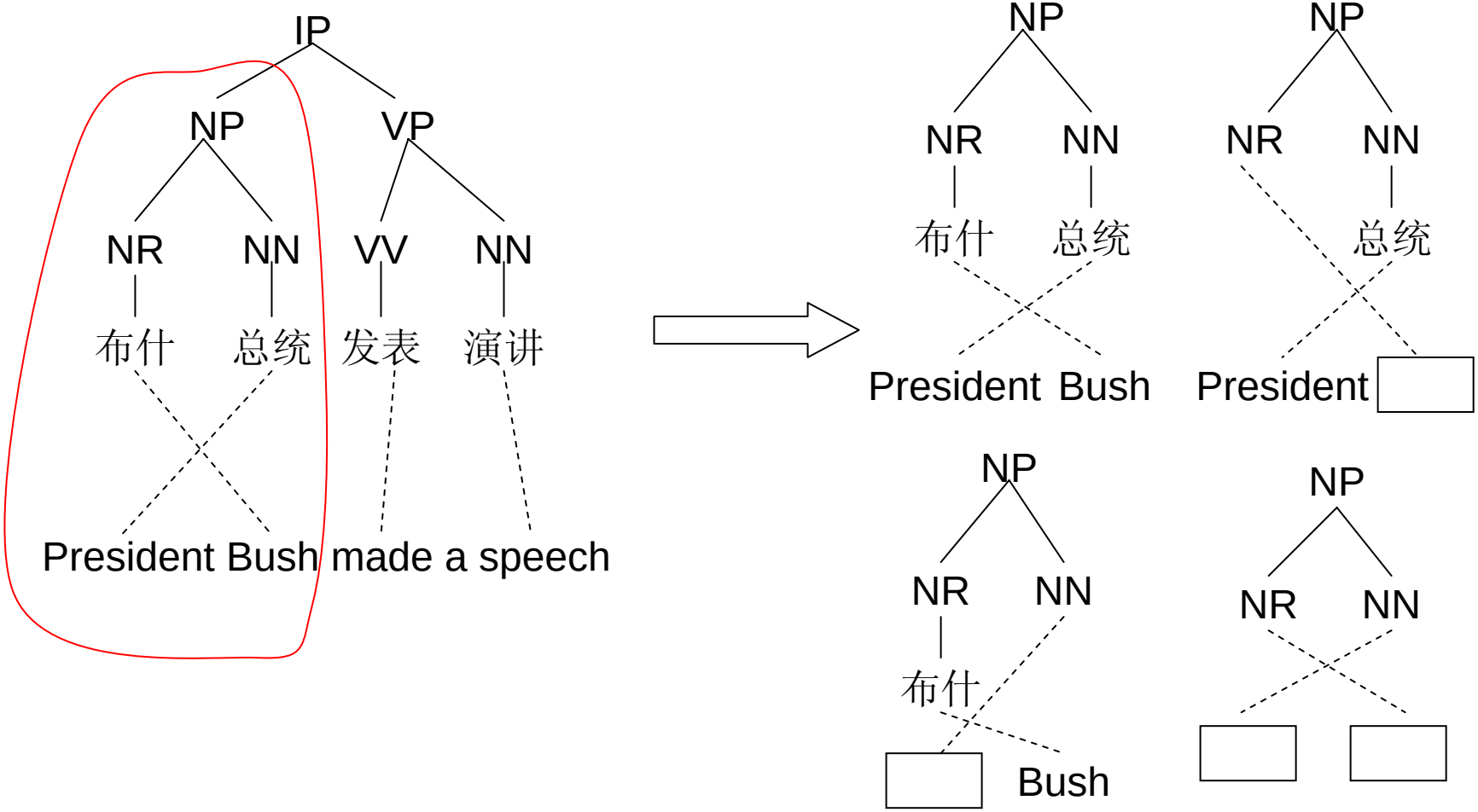
An Example


IP
NP
VP
NR
NN
VV
NN
布什
总统
发表
演讲
President Bush made a speech
VV
发表
made

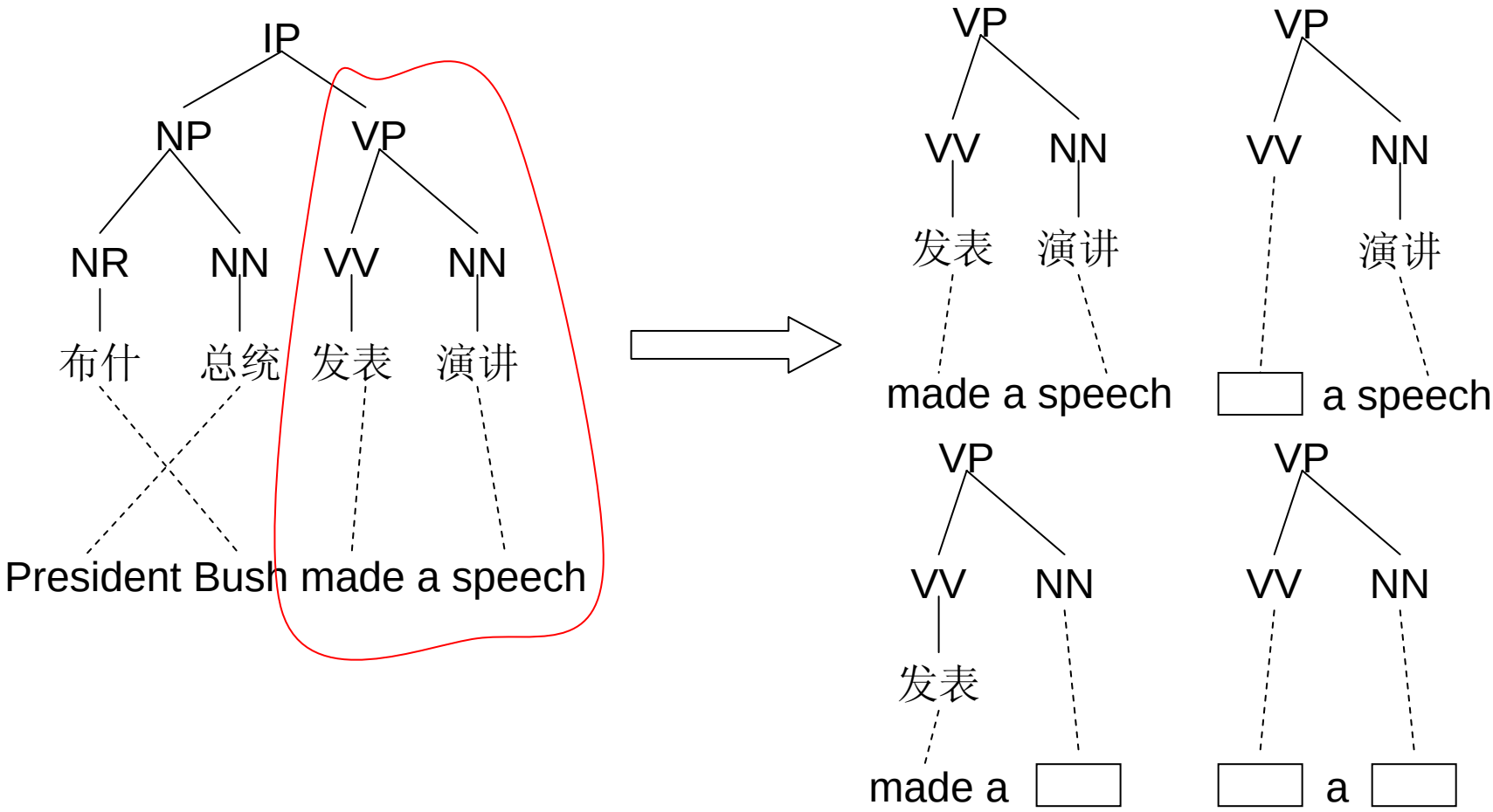
An Example


IP
NP
VP
NR
NN
VV
NN
布什
总统
发表
演讲
President Bush made a speech
NN
演讲
speech

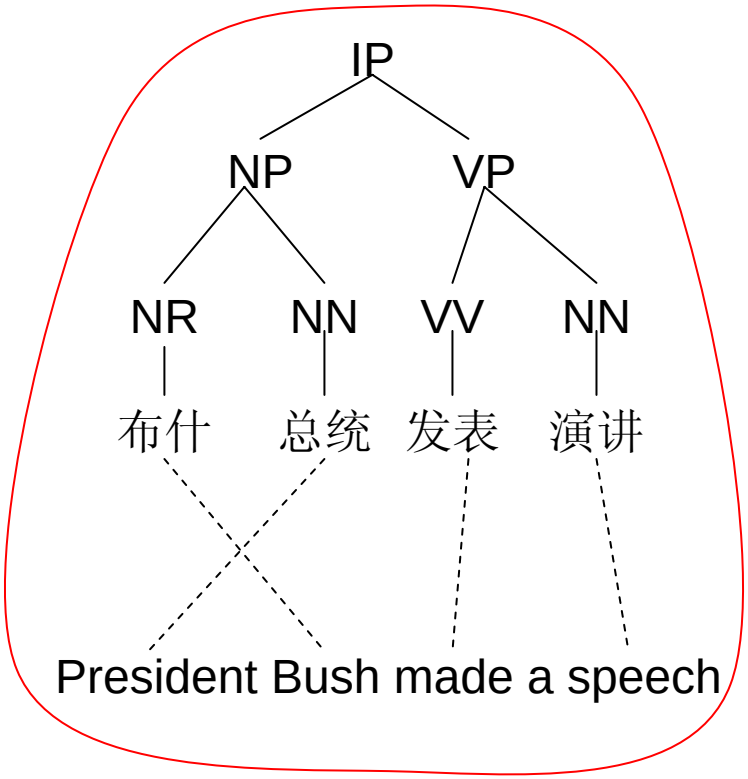
An Example

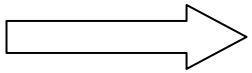

IP
NP
VP
NR
NN
VV
NN
布什
总统
发表
演讲
President Bush made a speech
NP
NR
NN
布什
总统
President Bush
NP
NR
NN
总统
President
NP
NR
NN
布什
Bush
NP
NR
NN

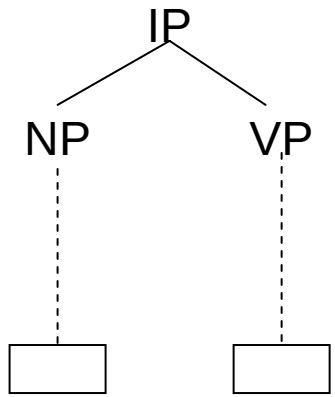
An Example


IP
NP
VP
NR
NN
VV
NN
布什
总统
发表
演讲
President Bush made a speech
VP
VV
NN
发表
演讲
made a speech
VP
VV
NN
演讲
a speech
VP
VV
NN
发表
made a
VP
VV
NN
a

An Example







h=2, c=2

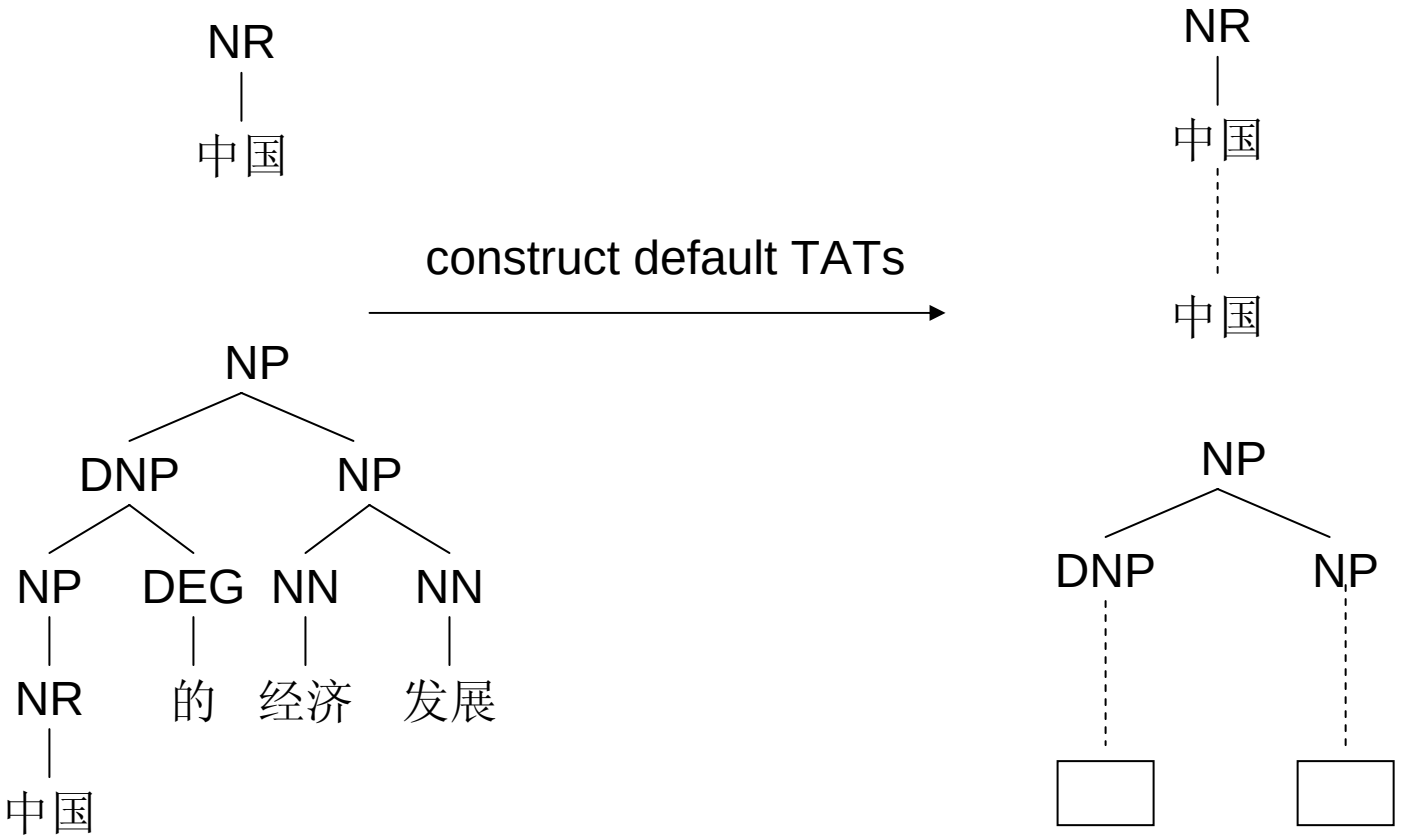
Outline

- *Introduction*
- *Tree-to-String Alignment Template*
- *Training*
- Decoding
- Experiments
- Recent Advances
- Conclusion and Future Work

Decoding

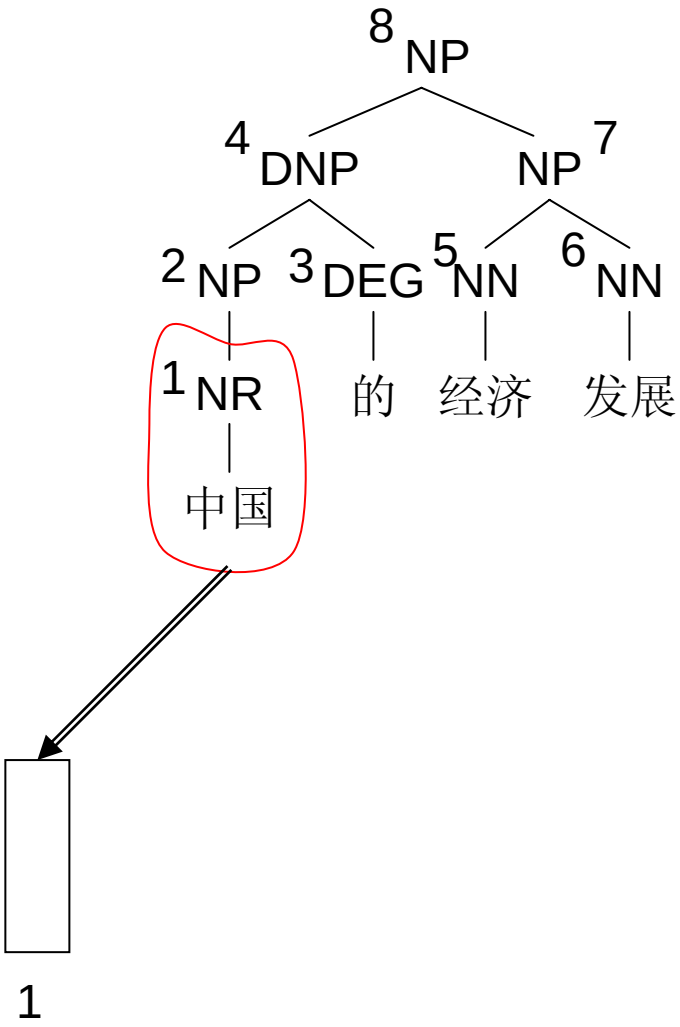
- Bottom-up beam search
- For each sub tree, we compute a list of candidate translations (derivations)
- A candidate translation contains the following information:
  - TAT sequence
  - partial translation
  - accumulated feature values
  - accumulated probability
  - …

Default TAT


NR
中国
construct default TATs
NP
DNP
NP
NP
DEG
NN
NN
NR
的
经济
发展
中国
NR
中国
中国
NP
DNP
NP

An Example

Go to page 10



Usable TAT

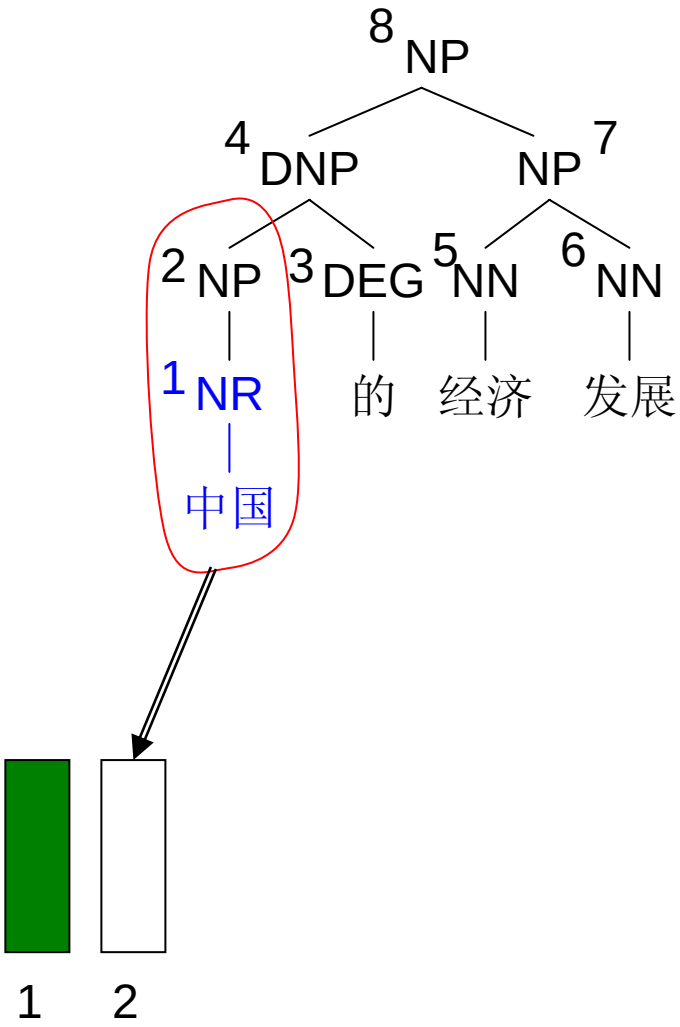
Derivation

(NR 中国)	China	1:1

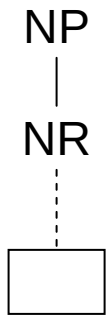
Translation

China

An Example



Usable TAT (default)



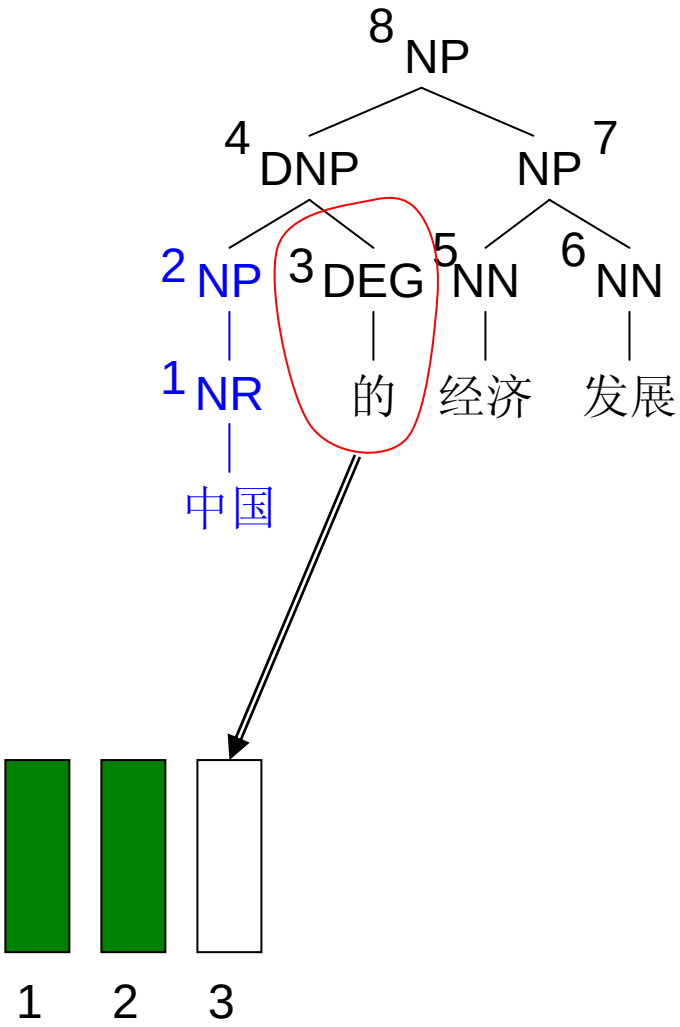
Derivation

(NP (NR))	X	1:1
(NR 中国)	China	1:1

Translation

China

An Example



Usable TAT (default)

DEG
|
的
⋮
的

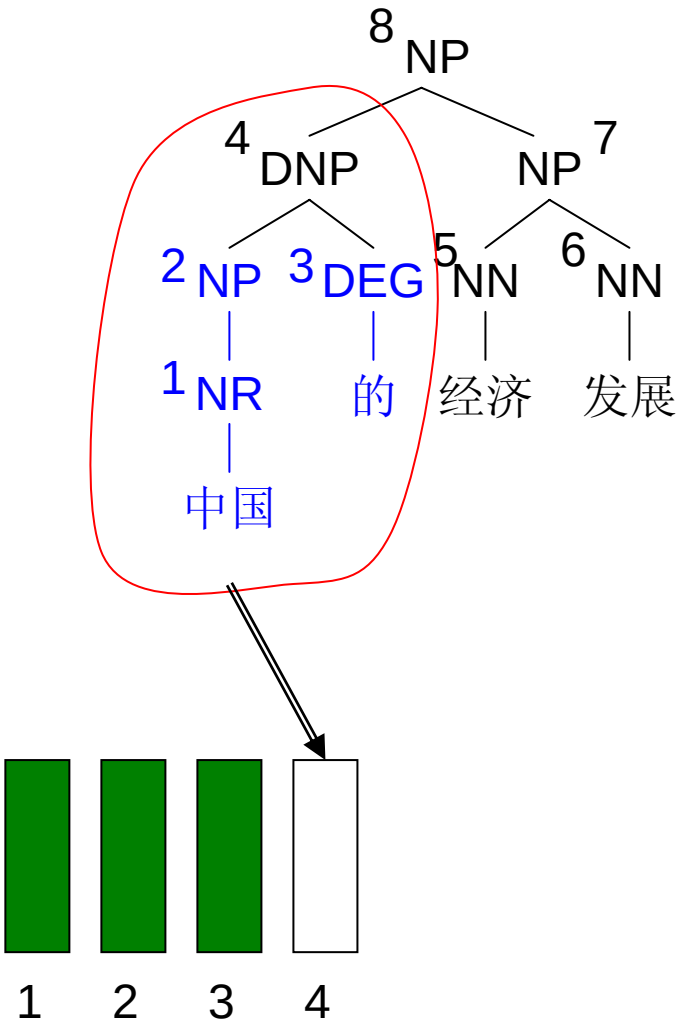
Derivation

(DEG 的)	的	1:1

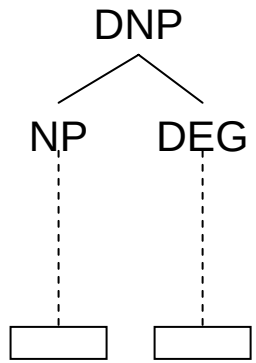
Translation

的

An Example



Usable TAT (default)



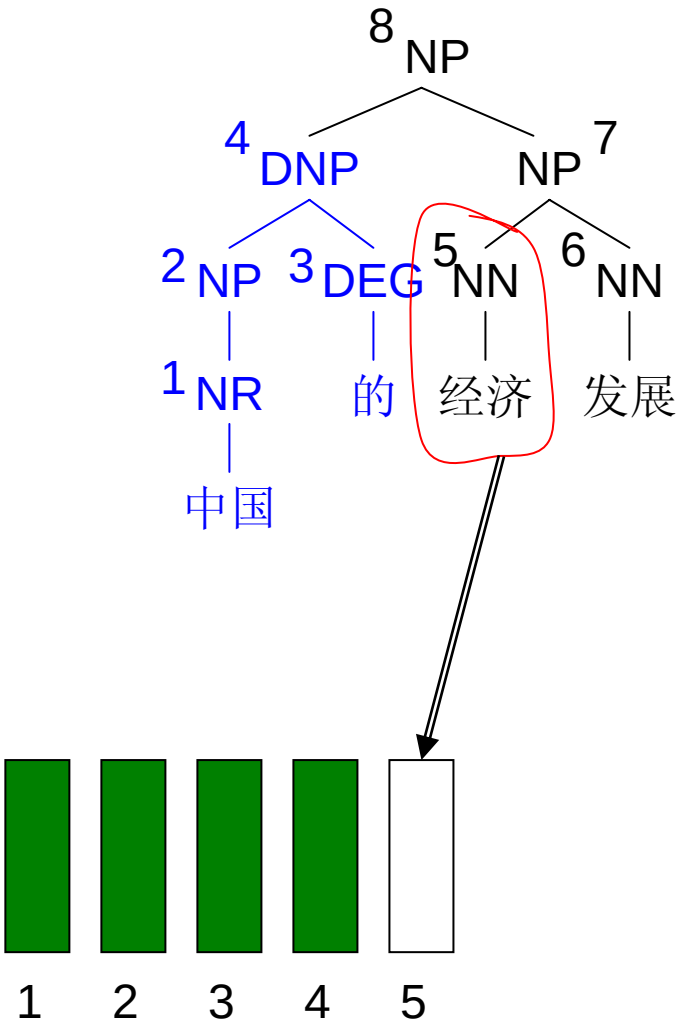
Derivation

(DNP (NP) (DEG))	X1 | X2	1:1 2:2
(NP (NR 中国))	China	1:1
(DEG 的)	的	1:1

Translation

China 的

An Example



Usable TAT

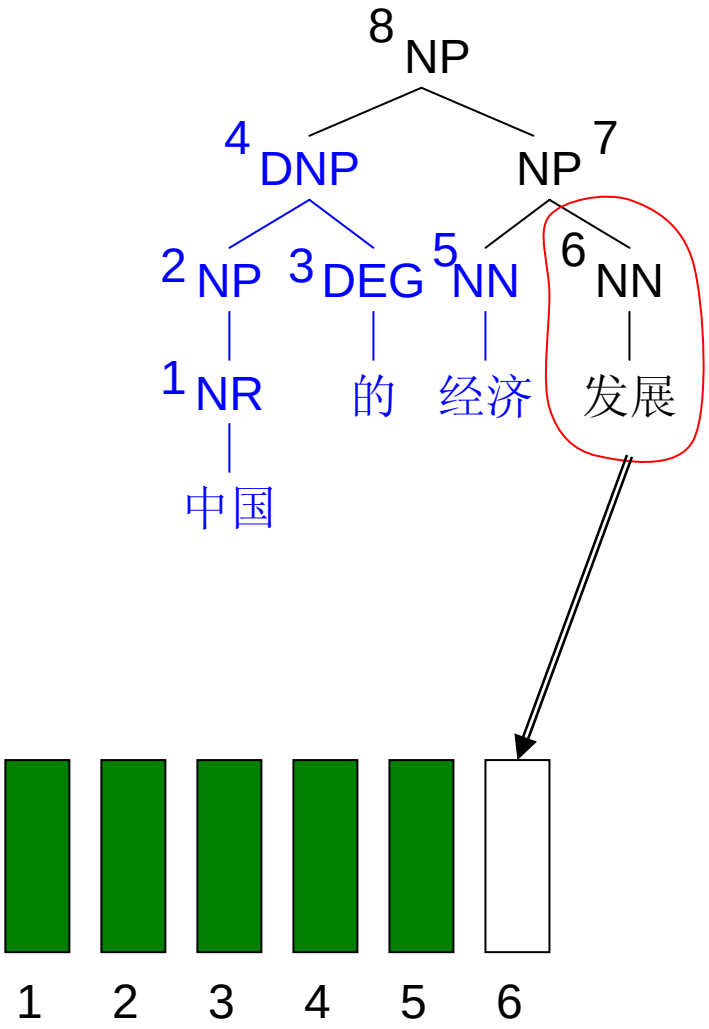
Derivation

(NN 经济)	economic	1:1

Translation

economic

An Example



Usable TAT

NN
发展
development

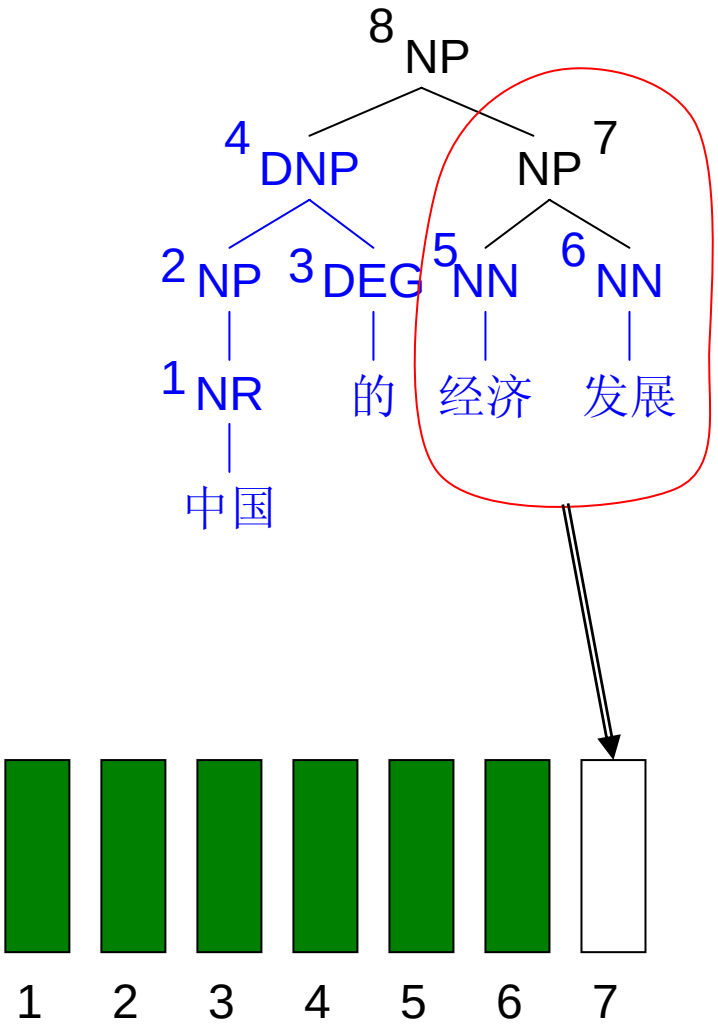
Derivation

(NN 发展)	development	1:1

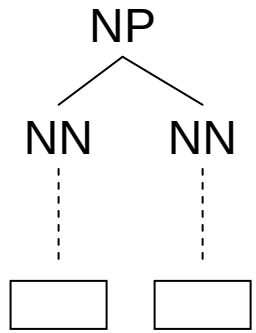
Translation

development

An Example



Usable TAT



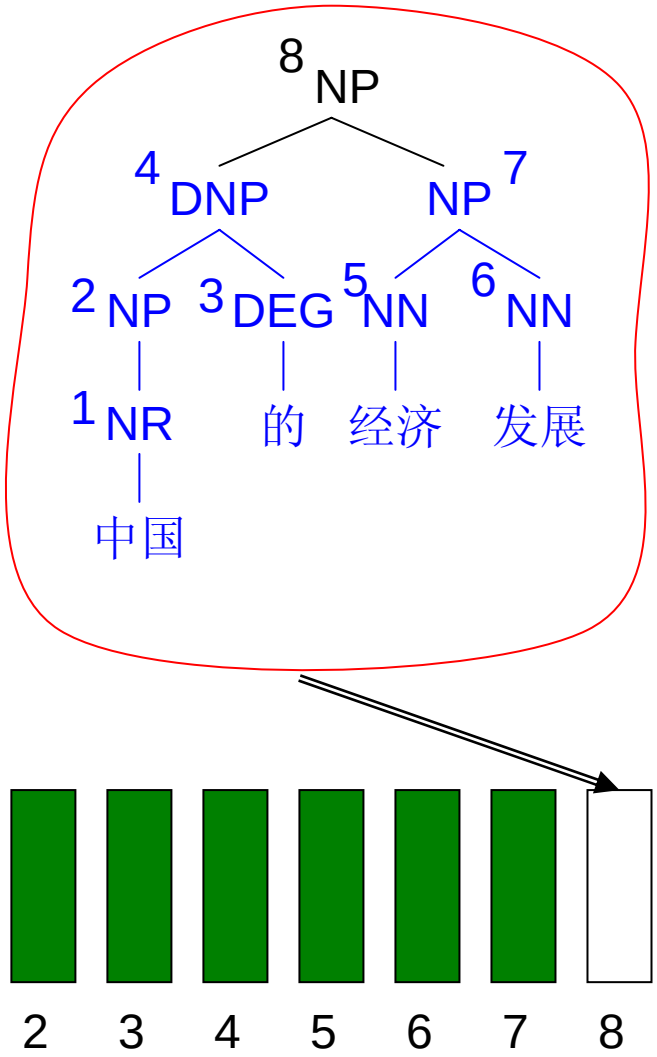
Derivation

(NP (NN) (NN))	X1 | X2	1:1 2:2
(NN 经济)	economic	1:1
(DEG 发展)	development	1:1

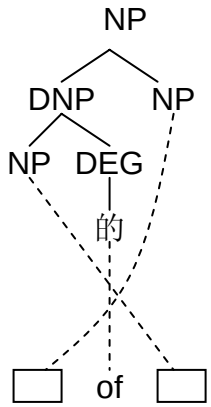
Translation

development

An Example



Usable TAT



Derivation

(NP (DNP (NP) (DEG (的))) (NP))	X1 | of | X2	1:3 2:2 3:1
(NP (NR))	X	1:1
(NR 中国)	China	1:!
(NP (NN) (NN))	X1 | X2	1:1 2:2
(NN 经济)	economic	1:1
(NN 发展)	development	1:1

Translation

economic development of China

Recombination

INSTITUTE OF COMPUTING TECHNOLOGY

The economic development of China is very rapid .

The economic develop of China is quite rapid .

The economic developing of Chinese is rapid .

The economic development of Chinese are quite rapid .

To perform recombination, we combine candidate translations that share the same leading and trailing bigrams (for trigram language model) in each stack.

Pseudo Code

```
initialize derivationStackVec[1 .. nodeCount]
for i=1 to nodeCount
    for each TAT usable to the subtree
        compute derivations
        add the derivations to derivationStackVec[i]
        prune derivationStackVec[i]
find the best derivation in derivationStackVec[nodeCount]
```

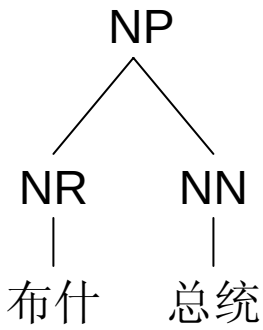
Note that we have not developed n-best list generation algorithm yet. To perform minimum error rate training, we just use the translations in the final stack.

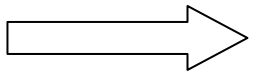
Treat BPs as TATs

INSTITUTE OF COMPUTING TECHNOLOGY

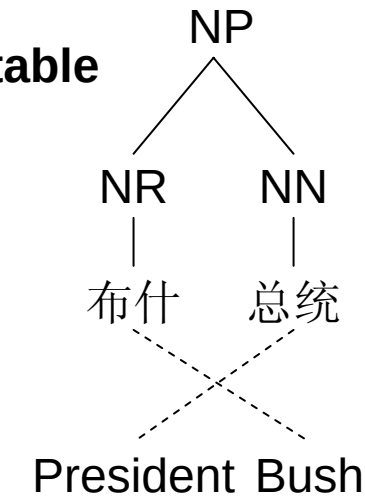
- Why?
  - bilingual phrases are “cheaper” than TATs
  - syntactic analysis is not reliable
  - lose useful non-syntactic phrase pairs due to strict restrictions
- How?
  - treat bilingual phrases as special TATs without tree over the source side

An Example



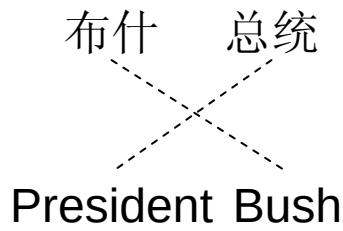


TAT table



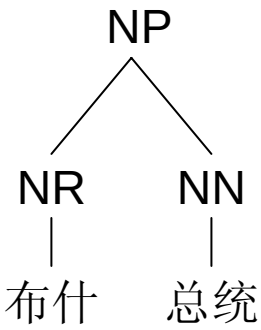
0.4 0.2 0.3 0.5

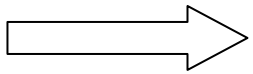
BP table



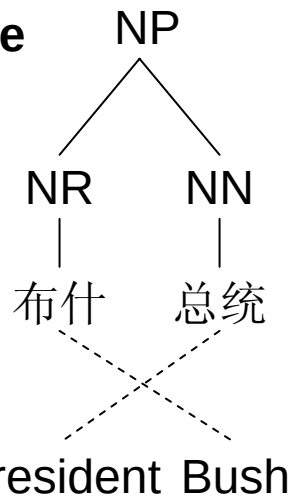
0.3 0.6 0.2 0.4

An Example



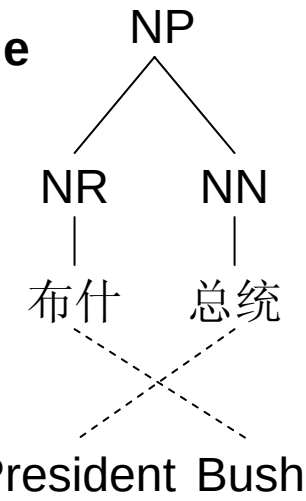


TAT table



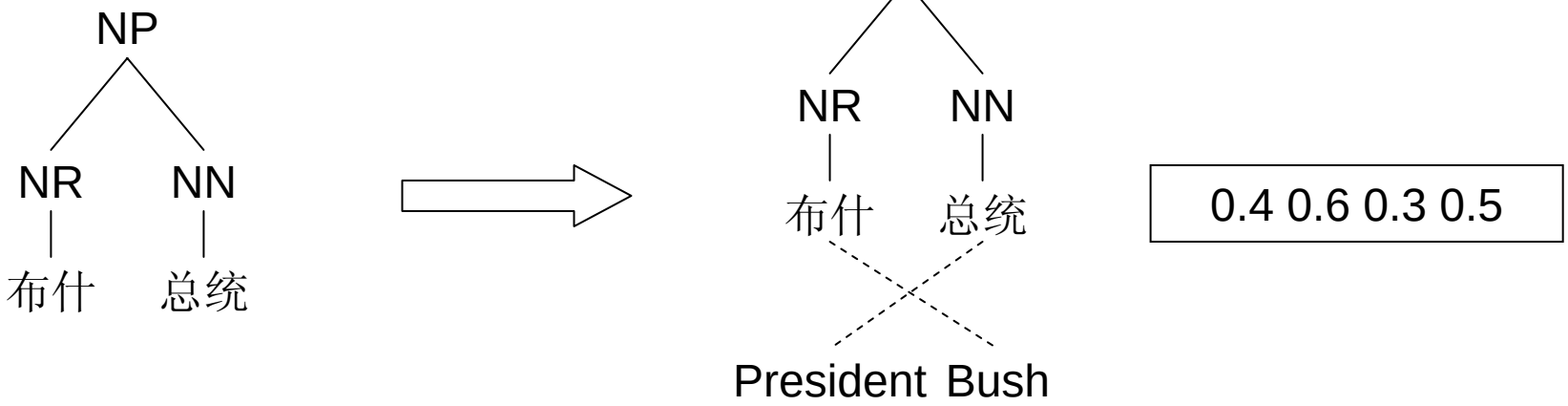
0.4 0.2 **0.3 0.5**

BP table



0.3 **0.6** 0.2 0.4

An Example



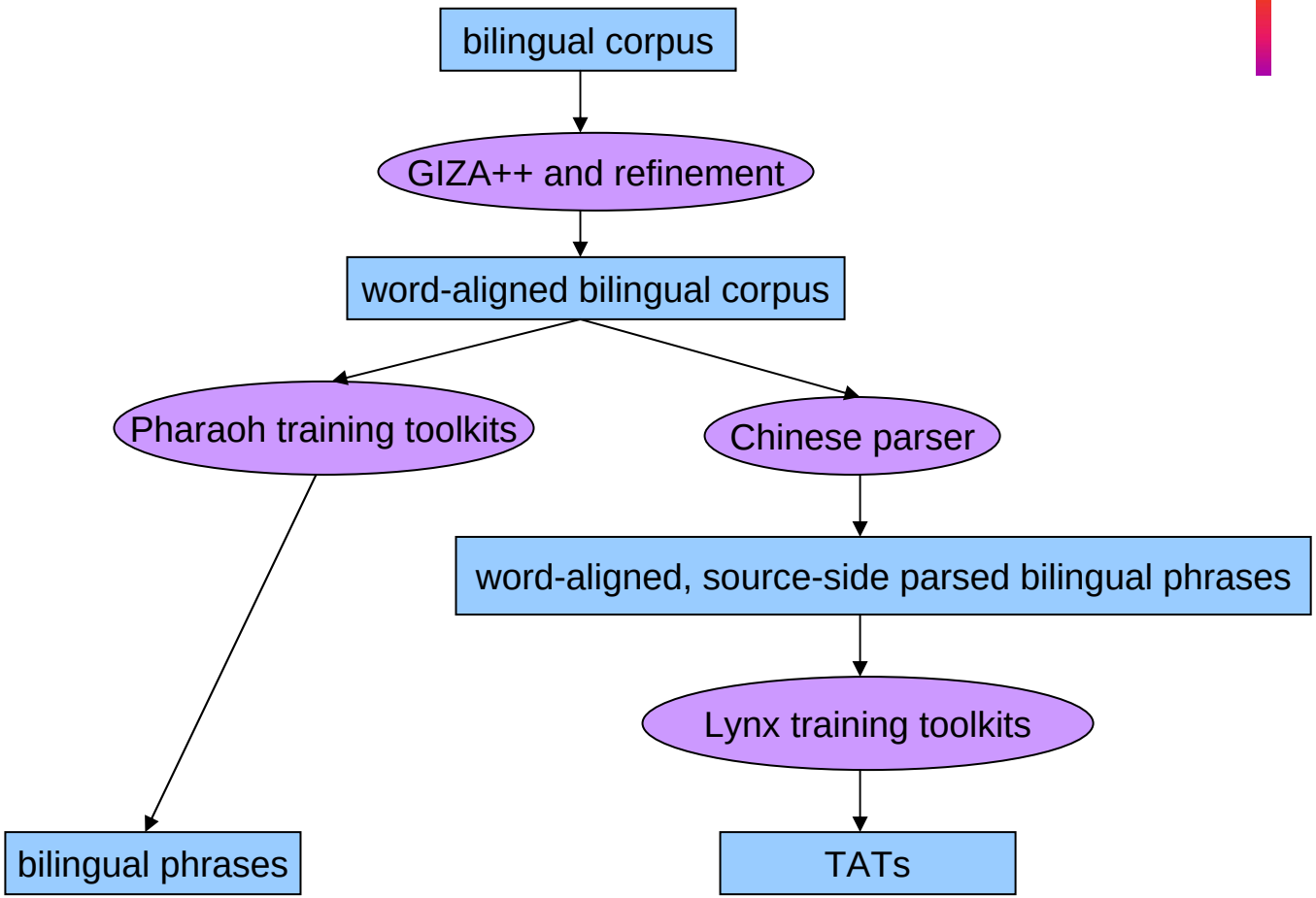
Outline

- *Introduction*
- *Tree-to-String Alignment Template*
- *Training*
- *Decoding*
- Experiments
- Recent Advances
- Conclusion and Future Work

Experiments

- Baseline: Pharaoh (Koehn et al., 2004)
- Training corpus: 31,149 sentence pairs with 843K Chinese words and 949K English words
- Development set: 2002 NIST Chinese-to-English test set (571 of 878 sentences)
- Test set: 2005 NIST Chinese-to-English test set (1,082 sentences)

Data Process



Some Tools

- Evaluation: mteval-v11b.pl
- Language Model: SRI Language Modeling Toolkits (Stolcke, 2002)
- Significance test: Zhang et al., 2004
- Parser: Xiong et al., 2005
- Minimum error rate training: optimizeV5IBMBLEU.m (Venugopal and Vogel, 2005)

Results

System	Features	BLEU4
Pharaoh	$\mathrm{d} + \phi(e\|f)$	0.0573 ± 0.0033
	$\mathrm{d} + \mathrm{lm} + \phi(e\|f) + \mathrm{wp}$	0.2019 ± 0.0083
	$\mathrm{d} + \mathrm{lm} + \phi(f\|e) + \mathrm{lex}(f\|e) + \phi(e\|f) + \mathrm{lex}(e\|f) + \mathrm{pp} + \mathrm{wp}$	0.2089 ± 0.0089
Lynx	h_1	0.1639 ± 0.0077
	$h_1 + h_6 + h_7$	0.2100 ± 0.0089
	$h_1 + h_2 + h_3 + h_4 + h_5 + h_6 + h_7$	0.2178 ± 0.0080

Comparison of Pharaoh and Lynx with different feature settings

Lynx achieves an absolute improvement of 0.9% (4.3% relative) over Pharaoh in terms of BLEU score. This difference is statistically significant ($p < 0.01$).

Effect of Using BPs

	BLEU4
tat	0.2178 ± 0.0080
tat + bp	0.2240 ± 0.0083

Effect of Using Bilingual Phrases for Lynx

Using bilingual phrases brings an absolute improvement of 0.6% in terms of BLEU score

Outline

- *Introduction*
- *Tree-to-String Alignment Template*
- *Training*
- *Decoding*
- *Experiments*
- Recent Advances
- Conclusion and Future Work

Recent Advances

- Scaling to large data
- Use BPs to improve fluency

Scaling to Large Data

- Bilingual corpus (train BPs and TATs)
  - 2.6M sentence pairs (68.1M Chinese words and 73.8M English words)
  - Use all the data to obtain BPs and a portion of 800K pairs to obtain TATs
- Monolingual corpora (train LM)
  - English side of the bilingual corpus (73.8M words)
  - Xinhua portion of Gigaword corpus (181M words)

Using BPs to Improve Fluency

INSTITUTE OF COMPUTING TECHNOLOGY

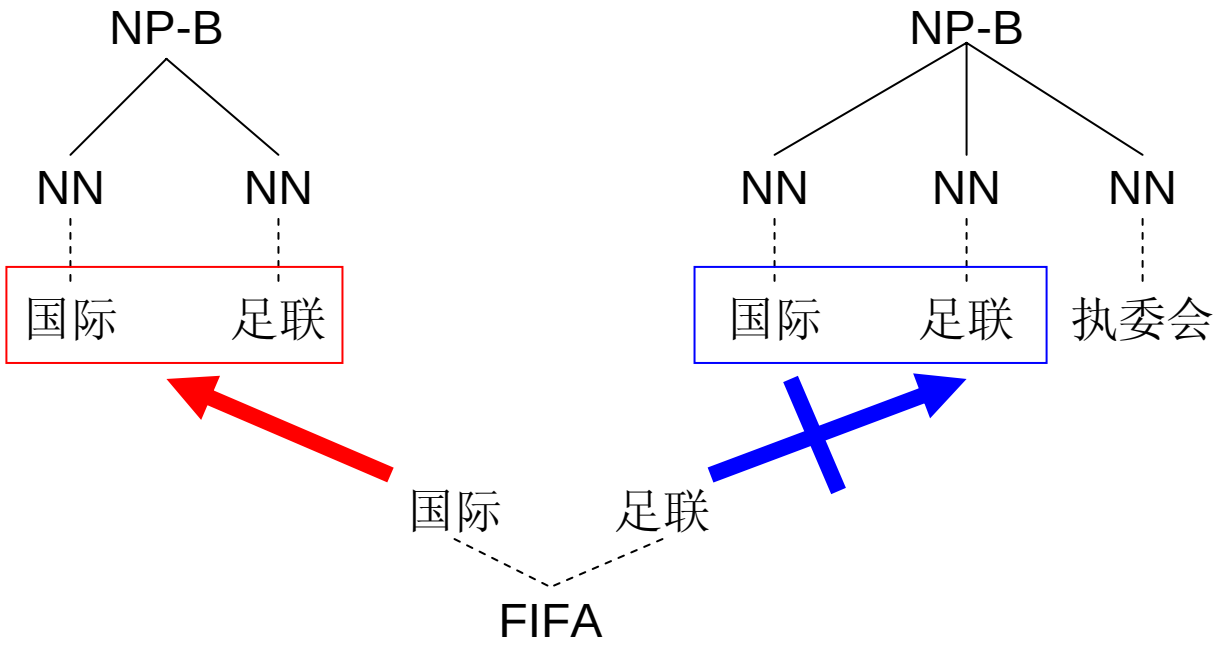
Problem with Lynx:

国际足联将严惩足球场上的欺骗行为
FIFA will severely punish cheat behaviour on the football field

国际足联执委会还宣布了一些改革措施。
international 足联 Executive Committee also announces that some reform measures.

How could this happen?

Two Parse Trees

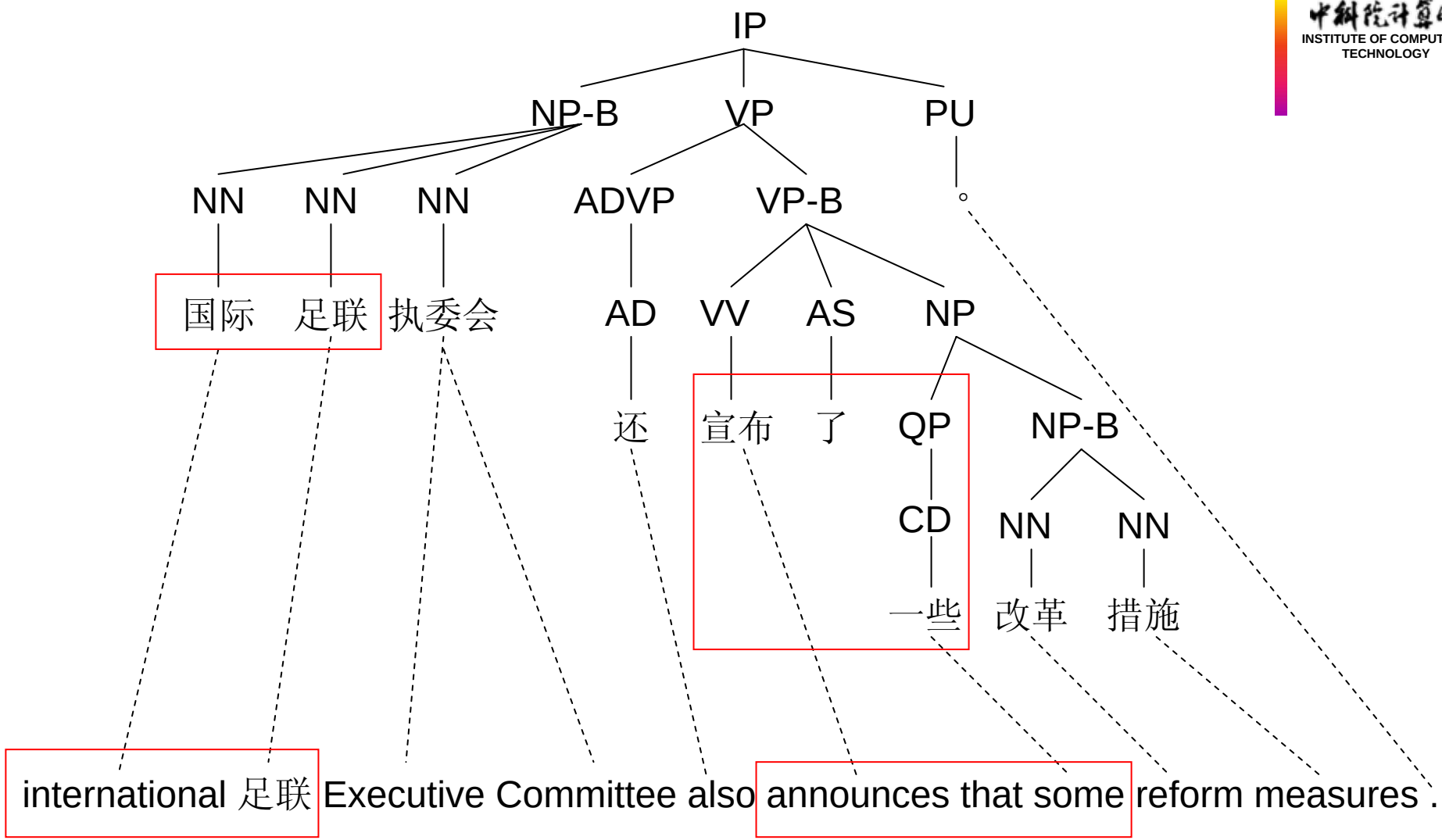


the strength of BPs is restricted!

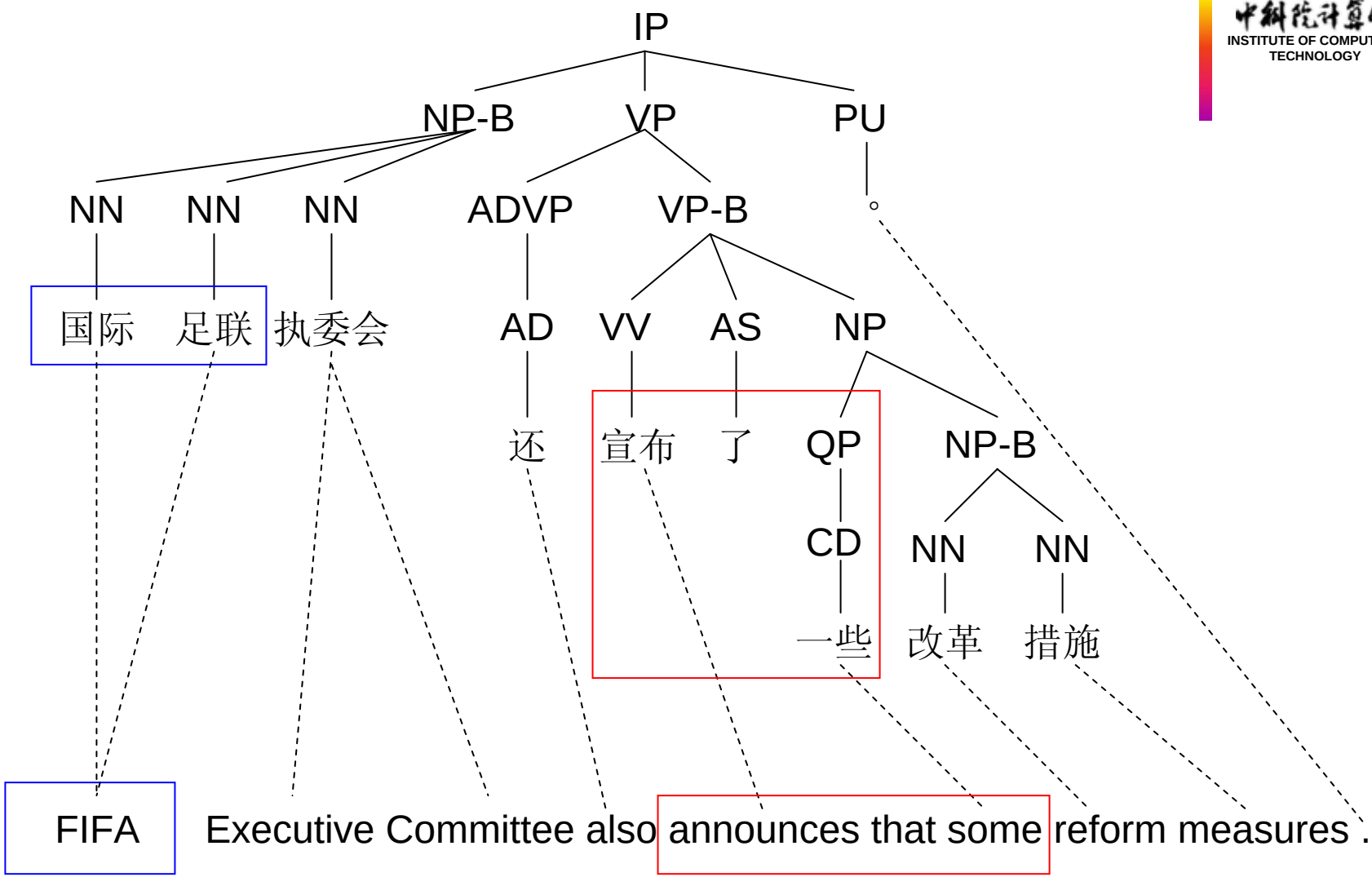
Solution

- When the search ends, refine the output by replacing strings with more ***fluent*** ones with the help of alignment.
- Use language model to measure fluency
- If there are more than one candidates, choose the one with highest score (take translation probabilities into account)

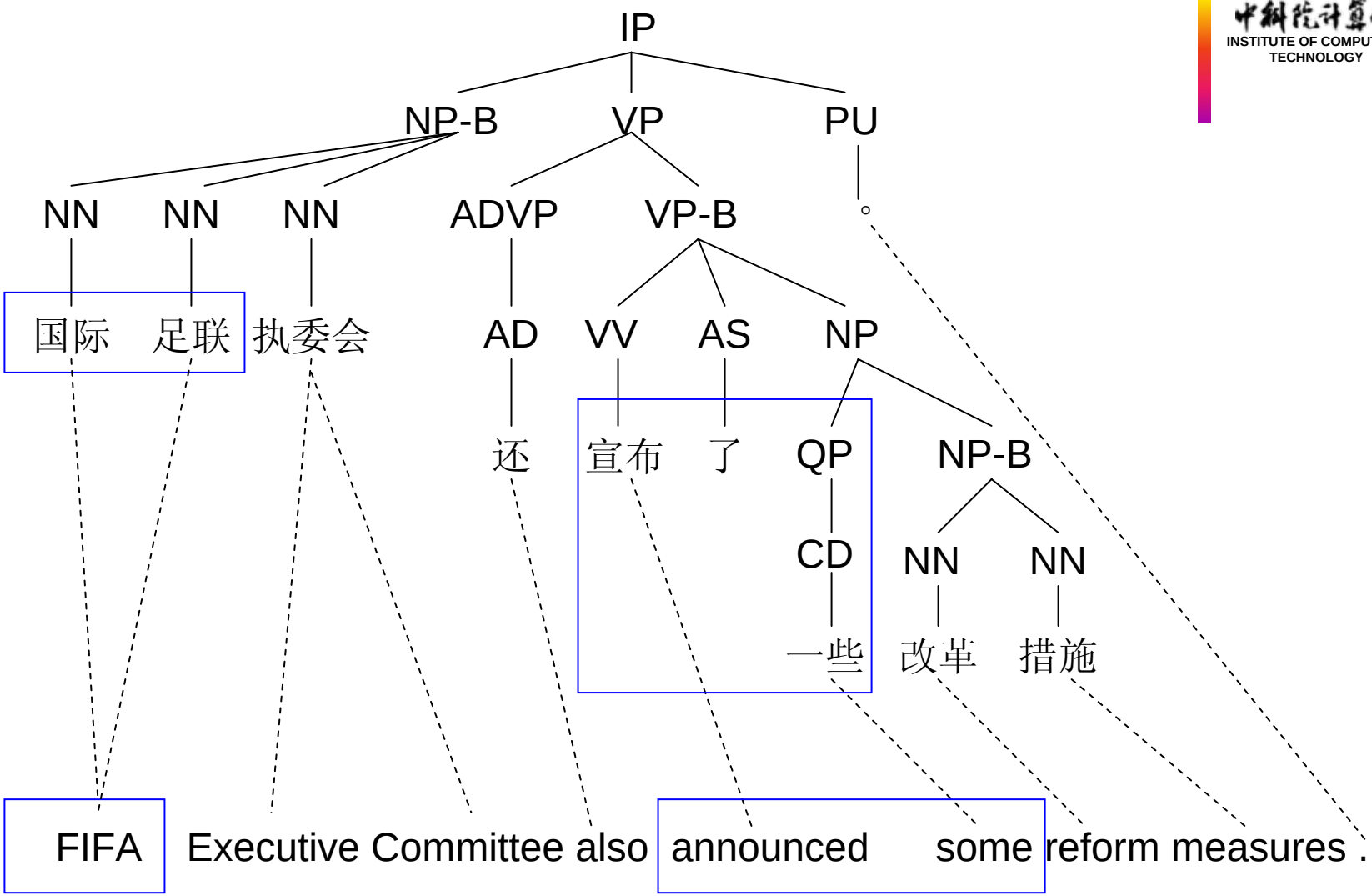
An Example


IP
NP-B
VP
PU
NN
NN
NN
ADVP
VP-B
国际
足联
执委会
AD
VV
AS
NP
还
宣布
了
QP
NP-B
CD
NN
NN
一些
改革
措施
international 足联 Executive Committee also announces that some reform measures .

An Example


IP
NP-B
VP
PU
NN
NN
NN
ADVP
VP-B
国际
足联
执委会
AD
VV
AS
NP
还
宣布
了
QP
NP-B
CD
NN
NN
一些
改革
措施
FIFA
Executive Committee also announces that some reform measures .

An Example


IP
NP-B
VP
PU
NN
NN
NN
ADVP
VP-B
国际
足联
执委会
AD
VV
AS
NP
还
宣布
了
QP
NP-B
CD
NN
NN
一些
改革
措施
FIFA
Executive Committee also
announced
some
reform measures .

Results

Training Data (pairs)		Language Model		Improve Fluency	BLEU4
TAT	BP	Data (words)	Order		
31K	-	949K	one 3-gram	No	0.2178
31K	31K	949K	one 3-gram	No	0.2240
31K	800K	73M	one 3-gram	No	0.2431
800K	2.6M	73M	one 3-gram	No	0.2692
800K	2.6M	73M ∣ 181M	two 3-gram	No	0.2934
800K	2.6M	73M ∣ 181M	two 4-gram	No	0.3047
800K	2.6M	73M ∣ 181M	two 4-gram	Yes	**0.3184**

Results of Lynx on test set with various settings.

Outline

- *Introduction*
- *Tree-to-String Alignment Template*
- *Training*
- *Decoding*
- *Experiments*
- *Recent Advances*
- Conclusion and Future Work

Conclusion

- The TAT-based translation model is simple and powerful
- Bilingual phrases can be used to strengthen the TAT-based model:
  - Treat them as special TATs
  - Use them to improve the fluency of the output

Future Work

- N-best list generation
- Better training methods for TATs
- Upgrading Tree-to-String to Forest-to-String (allow forest instead of tree over source string)

Thanks!