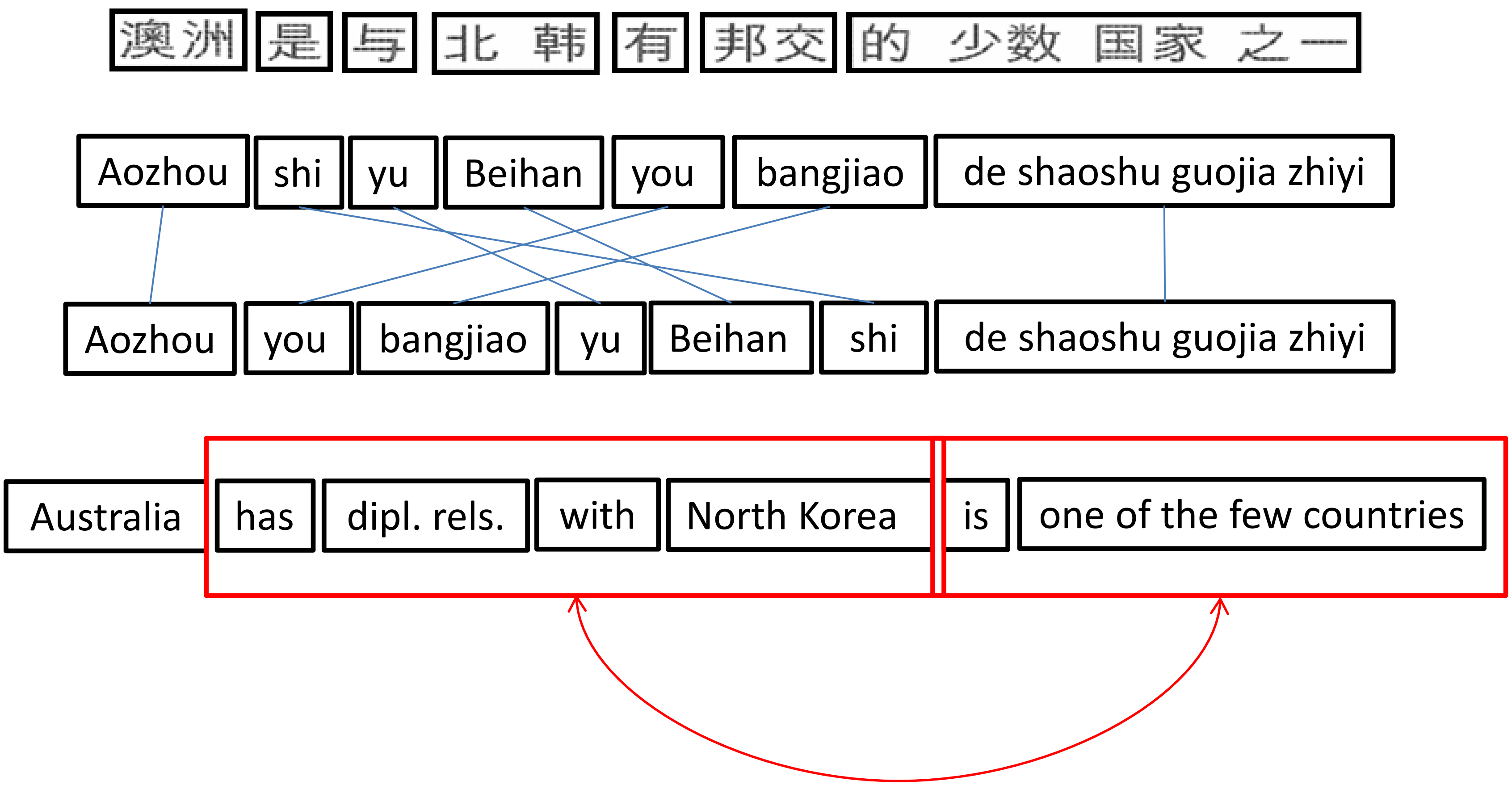
Hierarchical Phrase-based Translation

Marzieh Razavi
Maryam Siahbani
Ravikiran Vadlapudi

Introduction



澳洲 是 与 北 韩 有 邦交 的 少数 国家 之一

Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi

Aozhou shi have dipl. rels. with N. Korea de shaoshu guojia zhiyi

< yu $X_{\boxed{1}}$ you $X_{\boxed{2}}$, have $X_{\boxed{2}}$ with $X_{\boxed{1}}$ >

$$< X_{\boxed{1}} \textbf{ de } X_{\boxed{2}} \textbf{, the } X_{\boxed{2}} \textbf{ that } X_{\boxed{1}} >$$

Aozhou | shi | yu Beihan you bangjiao de shaoshu guojia | zhiyi

Aozhou | shi | one of | the few countries that have dep. rels. with North Korea

< $X_{\boxed{1}}$ zhiyi , one of $X_{\boxed{1}}$ >

Synchronous CFG

$$\boldsymbol{X} \rightarrow < \boldsymbol{\gamma},\ \boldsymbol{\alpha},\ \sim >$$

- X : non-terminal
- γ : strings of terminals and non-terminals for source
- α : strings of terminals and non-terminals for target
- $\sim$: 1-1 correspondence between non-terminals

$$\mathbf{X} \longrightarrow \textbf{<yu } \mathbf{X}_{\boxed{1}} \textbf{ you } \mathbf{X}_{\boxed{2}} \textbf{ , have } \mathbf{X}_{\boxed{2}} \textbf{ with } \mathbf{X}_{\boxed{1}}\textbf{>}$$

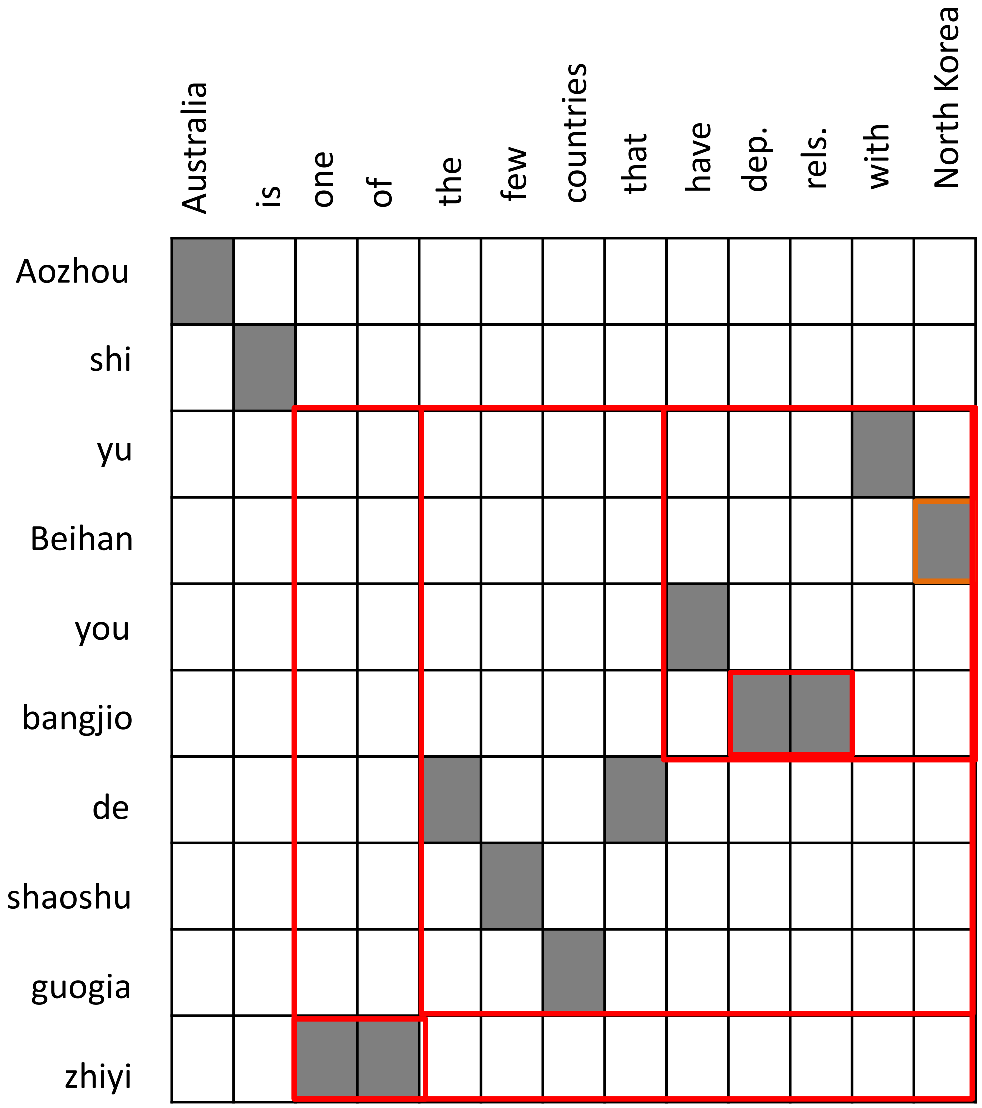
Rule Extraction

Rule Extraction

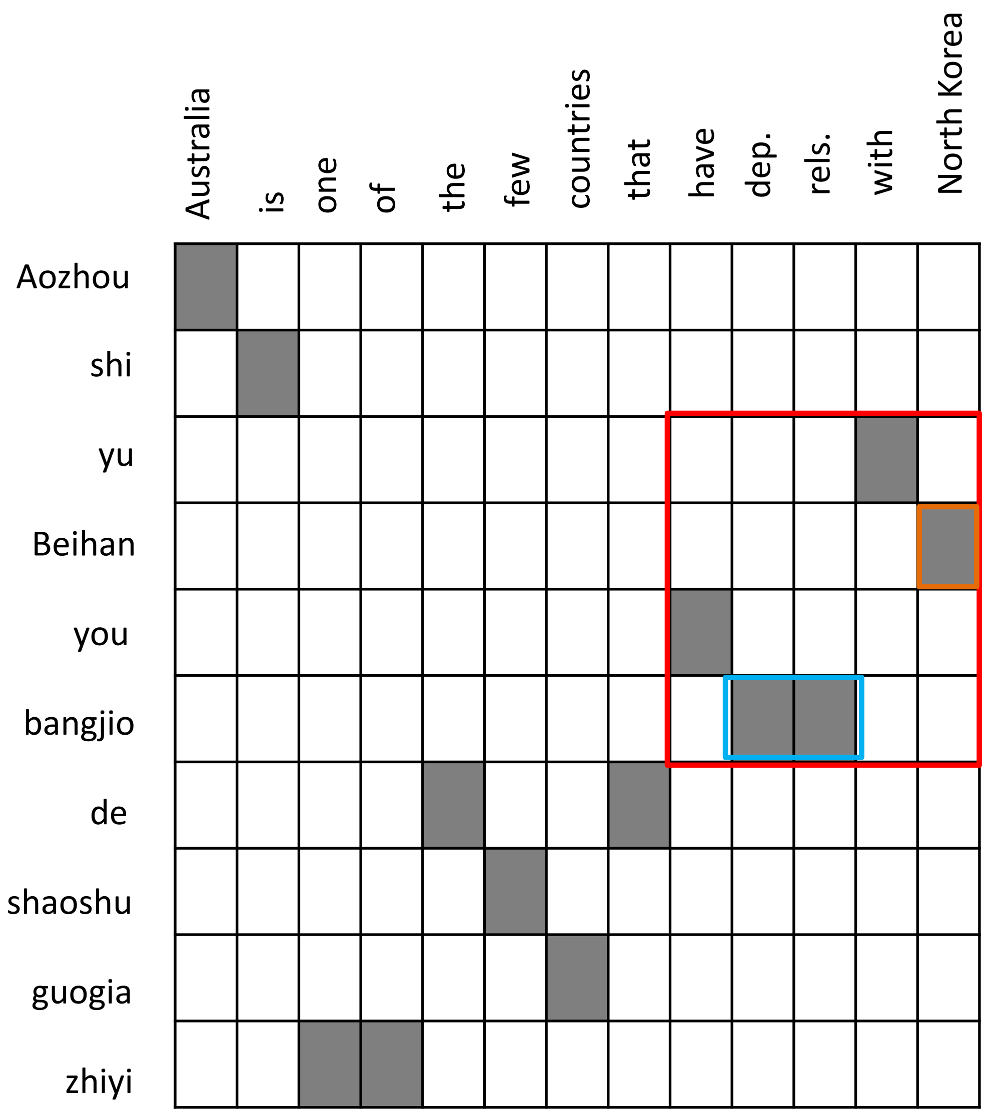
1. Identifying initial phrase pairs (similar to conventional phrase-based systems)

2. Extracting rules:
   a. Find phrases that contain other phrases
   b. Replace sub-phrases with non-terminals

Identifying Initial Phrases



Extracting Rules

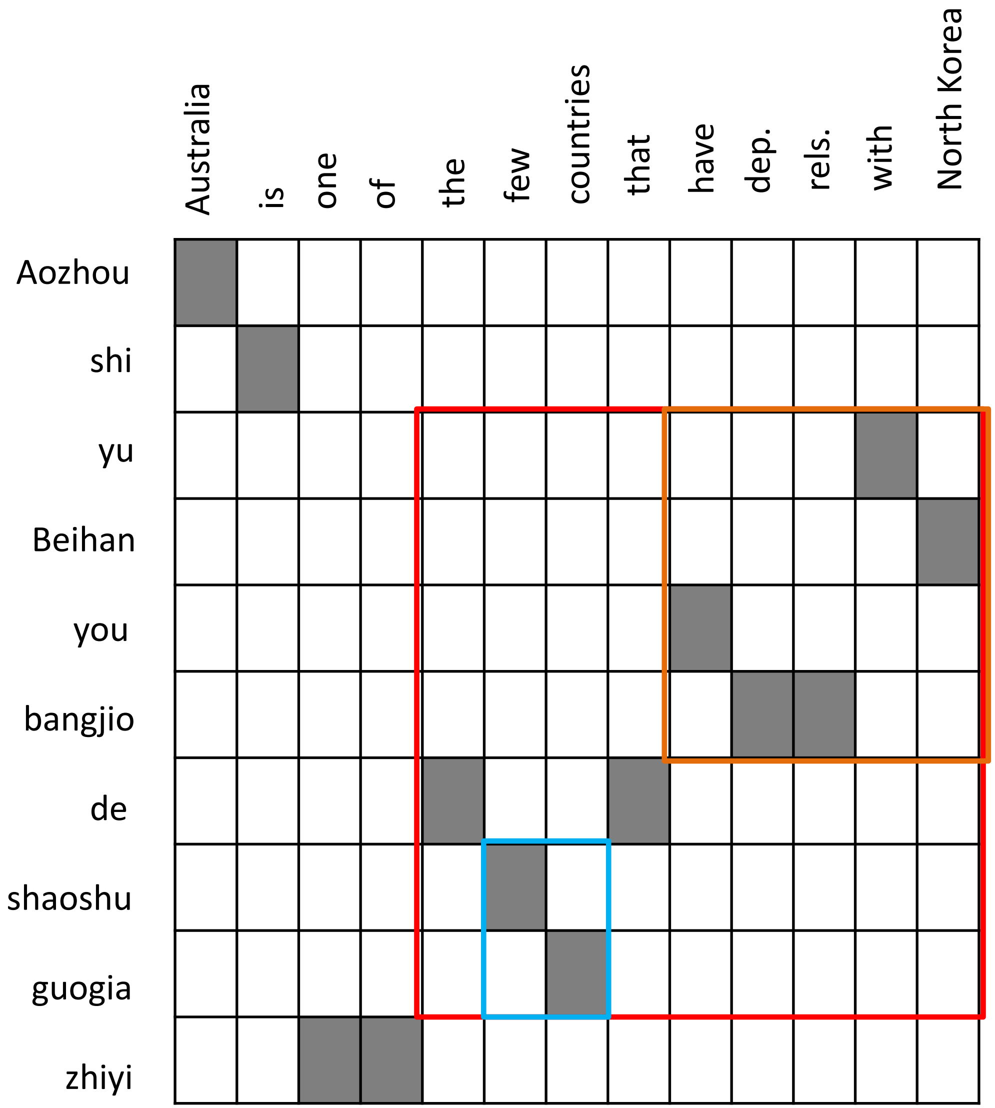


X $\longrightarrow$ <yu $X_{\boxed{1}}$ you $X_{\boxed{2}}$, have $X_{\boxed{2}}$ with $X_{\boxed{1}}$> (6)

X $\longrightarrow$ <Beihan, North Korea> (10)

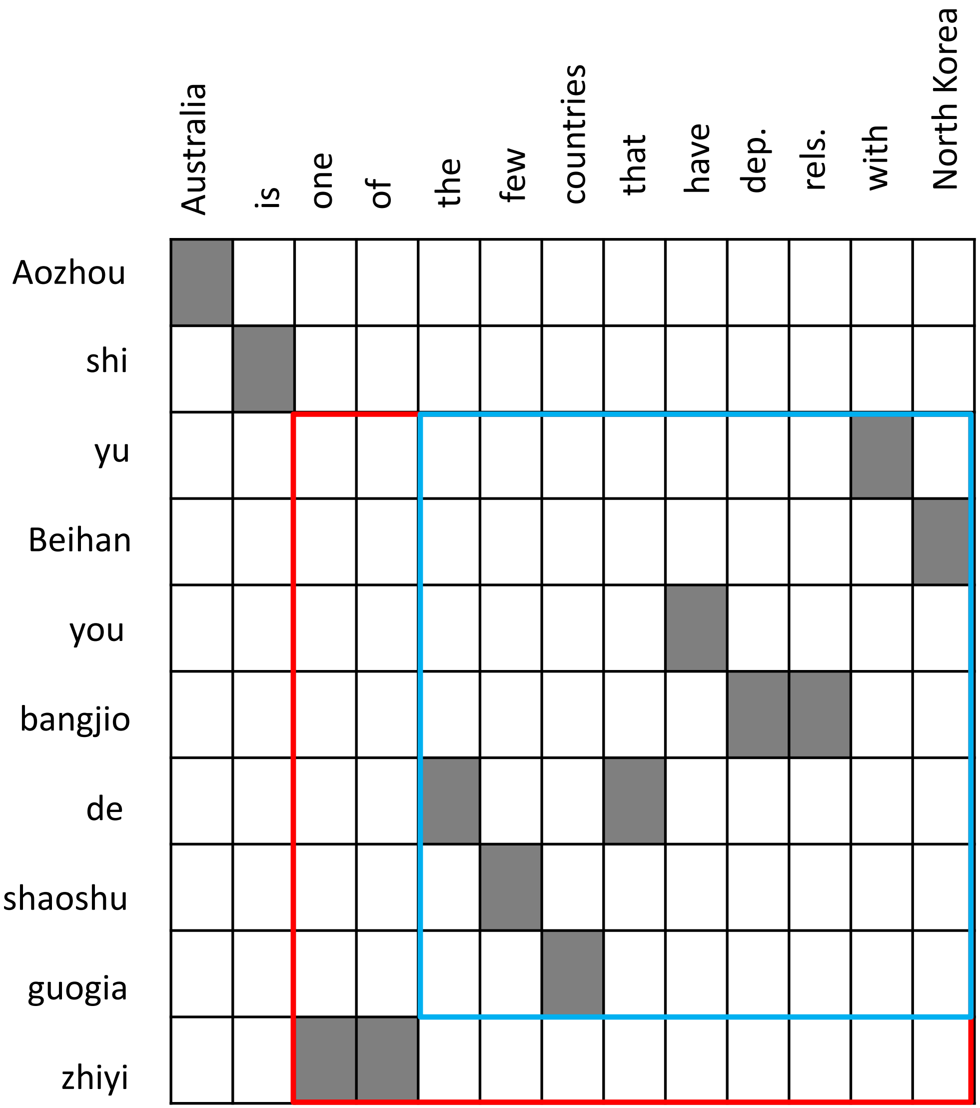
X $\longrightarrow$ <bangjio, dep. rels.> (12)

Extracting Rules



$$X \rightarrow \langle X_{\boxed{1}}\ \text{de}\ X_{\boxed{2}}\ ,\ \text{the}\ X_{\boxed{2}}\ \text{that}\ X_{\boxed{1}} \rangle \quad (7)$$

Extracting Rules



$$X \longrightarrow \langle X_{\boxed{1}}\ \textbf{zhiyi}\ ,\ \textbf{one of}\ X_{\boxed{1}} \rangle \quad \textbf{(8)}$$

X ⟶ <shaoshu guogia, few countries>

Filtering the Grammar

- Limit the length of initial phrases to 10 words on either side.
- Limit the rules to five nonterminals plus terminals on the French side
- Rules can have at most two nonterminals
  - simplifies the decoder implementation.
- It is prohibited for nonterminals to be adjacent on the French side
  - major cause of spurious ambiguity

Other Rules

- Glue Rules: for dividing source side to chunks and translating one chunk at a time

$$S \rightarrow < S_{\boxed{1}} X_{\boxed{2}}, S_{\boxed{1}} X_{\boxed{2}} > \quad (14)$$

$$S \rightarrow < X_{\boxed{1}}, X_{\boxed{1}} > \quad (15)$$

- Entity Rules: for translating numbers, dates, …

$$X \rightarrow < X_{\boxed{1}} dunianlai,\ over\ the\ last\ X_{\boxed{1}}\ years >$$

$$
\begin{aligned}
&\langle S_{\boxed{1}}, S_{\boxed{1}} \rangle \\
&\quad \stackrel{(14)}{\Longrightarrow} \langle S_{\boxed{2}} X_{\boxed{3}}, S_{\boxed{2}} X_{\boxed{3}} \rangle \\
&\quad \stackrel{(14)}{\Longrightarrow} \langle S_{\boxed{4}} X_{\boxed{5}} X_{\boxed{3}}, S_{\boxed{4}} X_{\boxed{5}} X_{\boxed{3}} \rangle \\
&\quad \stackrel{(15)}{\Longrightarrow} \langle X_{\boxed{6}} X_{\boxed{5}} X_{\boxed{3}}, X_{\boxed{6}} X_{\boxed{5}} X_{\boxed{3}} \rangle \\
&\quad \stackrel{(9)}{\Longrightarrow} \langle \text{Aozhou } X_{\boxed{5}} X_{\boxed{3}}, \text{Australia } X_{\boxed{5}} X_{\boxed{3}} \rangle \\
&\quad \stackrel{(11)}{\Longrightarrow} \langle \text{Aozhou shi } X_{\boxed{3}}, \text{Australia is } X_{\boxed{3}} \rangle \\
&\quad \stackrel{(8)}{\Longrightarrow} \langle \text{Aozhou shi } X_{\boxed{7}} \text{ zhiyi, Australia is one of } X_{\boxed{7}} \rangle \\
&\quad \stackrel{(7)}{\Longrightarrow} \langle \text{Aozhou shi } X_{\boxed{8}} \text{ de } X_{\boxed{9}} \text{ zhiyi, Australia is one of the } X_{\boxed{9}} \text{ that } X_{\boxed{8}} \rangle \\
&\quad \stackrel{(6)}{\Longrightarrow} \langle \text{Aozhou shi yu } X_{\boxed{1}} \text{ you } X_{\boxed{2}} \text{ de } X_{\boxed{9}} \text{ zhiyi,} \\
&\qquad \text{Australia is one of the } X_{\boxed{9}} \text{ that have } X_{\boxed{2}} \text{ with } X_{\boxed{1}} \rangle \\
&\quad \stackrel{(10)}{\Longrightarrow} \langle \text{Aozhou shi yu Beihan you } X_{\boxed{2}} \text{ de } X_{\boxed{9}} \text{ zhiyi,} \\
&\qquad \text{Australia is one of the } X_{\boxed{9}} \text{ that have } X_{\boxed{2}} \text{ with North Korea} \rangle \\
&\quad \stackrel{(12)}{\Longrightarrow} \langle \text{Aozhou shi yu Beihan you bangjiao de } X_{\boxed{9}} \text{ zhiyi,} \\
&\qquad \text{Australia is one of the } X_{\boxed{9}} \text{ that have diplomatic relations with North Korea} \rangle \\
&\quad \stackrel{(13)}{\Longrightarrow} \langle \text{Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi,} \\
&\qquad \text{Australia is one of the few countries that have diplomatic relations with North Korea} \rangle
\end{aligned}
$$

Model

Model

- General log-linear model over derivations D:

$$P(D) \propto \prod_i \varphi_i(D)^{\lambda_i}$$

$$\varphi_i(D) = \prod_{(X \rightarrow <\gamma, \alpha>) \in D} \varphi_i(X \rightarrow <\gamma, \alpha>)$$

Weighted Synchronous CFG

- Weights function over derivations D:

$$w(D) = \prod_{(X \rightarrow < \gamma, \alpha >) \in D} w(X \rightarrow < \gamma, \alpha >)$$

- Weight for the rules :

$$w(X \rightarrow < \gamma, \alpha >) = \prod_{i \neq LM} \varphi_i (X \rightarrow < \gamma, \alpha >)^{\lambda_i}$$

- Probability model:

$$P(D) \propto P_{LM}(e)^{\lambda_{LM}} \times w(D)$$

Features

Features

- $P(\gamma|\alpha), P(\alpha|\gamma)$
- Lexical weight $P_w(\gamma|\alpha), P_w(\alpha|\gamma)$
  - How well the words in α translate the words in γ
- Language Model
- Extracted rules (with penalty exp(-1))
- Glue rules (with penalty exp(-1))
- Word penalty
- Dates, numbers, ….

Training

Training

- Estimate the parameters of phrase translation and lexical weighting:
  - Give a count 1 to each initial phrase pair occurrence
  - Distribute its weight uniformly among the rules obtained by subtracting sub-phrases from it
  - This distribution is considered as observed data
  - Use relative-frequency estimation to obtain $P(\gamma|\alpha), P(\alpha|\gamma)$
- Learn the parameters λ_i of log-linear model:
  - MERT

Decoding

Basic Algorithm

- Objective

$$\hat{e} = e\left(\underset{D \text{ s.t. } f(D)=f}{\arg\max}\; P(D)\right)$$

- Inference Rules

$$\frac{Z \rightarrow f_{i+1} : w}{[Z, i, i+1] : w}$$

$$\frac{Z \rightarrow XY : w \quad [X, i, k] : w_1 \quad [Y, k, j] : w_2}{[Z, i, j] : w_1 w_2 w}$$

Basic Algorithm

Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi

0 1 2 3 4 5 6 7 8 9

$$\frac{X \rightarrow Aozhou}{[X, 0, 1] : w_1}$$

Basic Algorithm

Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi

0 1 2 3 4 5 6 7 8 9

$$\frac{X \rightarrow Aozhou}{[X, 0, 1] : w_1}$$

$$\frac{X \rightarrow shi}{[X, 1, 2] : w_3}$$

Basic Algorithm

Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi

0 1 2 3 4 5 6 7 8 9

$$\frac{X \rightarrow Aozhou}{[X, 0, 1] : w_1}$$

$$\frac{X \rightarrow shi}{[X, 1, 2] : w_3}$$

$$\frac{X \rightarrow Beihon}{[X, 3, 4] : w_2}$$

$$\frac{X \rightarrow bangjiao}{[X, 5, 6] : w_4}$$

$$\frac{X \rightarrow shaoshu \;\; guojio}{[X, 7, 9] : w_5}$$

Basic Algorithm

Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi

0 1 2 3 4 5 6 7 8 9

$$\frac{X \rightarrow Aozhou}{[X, 0, 1] : w_1}$$

$$\frac{X \rightarrow shi}{[X, 1, 2] : w_3}$$

$$\frac{X \rightarrow Beihon}{[X, 3, 4] : w_2}$$

$$\frac{X \rightarrow bangjiao}{[X, 5, 6] : w_4}$$

$$\frac{X \rightarrow shaoshu\ \ guojio}{[X, 7, 9] : w_5}$$

$$\frac{Z \rightarrow X\, zhiyi : w_6 \quad [X 7, 9] : w_5}{[Z, 7, 11] : w_5 w_6}$$

Basic Algorithm

Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi

0 1 2 3 4 5 6 7 8 9

$$\frac{X \rightarrow Aozhou}{[X, 0, 1] : w_1}$$

$$\frac{X \rightarrow shi}{[X, 1, 2] : w_3}$$

$$\frac{X \rightarrow Beihon}{[X, 3, 4] : w_2}$$

$$\frac{X \rightarrow bangjiao}{[X, 5, 6] : w_4}$$

$$\frac{Z \rightarrow X\, zhiyi : w_6 \quad [X 7, 9] : w_5}{[Z, 7, 11] : w_5 w_6}$$

Basic Algorithm

Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi

0 1 2 3 4 5 6 7 8 9

$$\frac{X \to Aozhou}{[X, 0, 1] : w_1}$$

$$\frac{X \to shi}{[X, 1, 2] : w_3}$$

$$\frac{Z \to yu\ X_1\ you\ X_2 : w_7 \quad [X_1, 3, 4] : w_2 \quad [X_1, 5, 6] : w_4}{[Z, 2, 6] : w_7 w_2 w_4}$$

$$\frac{X \to Beihon}{[X, 3, 4] : w_2}$$

$$\frac{X \to bangjiao}{[X, 5, 6] : w_4}$$

$$\frac{Z \to X\ zhiyi : w_6 \quad [X 7, 9] : w_5}{[Z, 7, 11] : w_5 w_6}$$

Basic Algorithm

Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi

0 1 2 3 4 5 6 7 8 9

$$\frac{X \rightarrow Aozhou}{[X, 0, 1] : w_1}$$

$$\frac{X \rightarrow shi}{[X, 1, 2] : w_3}$$

$$\frac{Z \rightarrow yu\ X_1\ you\ X_2 : w_7 \quad [X_1, 3, 4] : w_2 \quad [X_1, 5, 6] : w_4}{[Z, 2, 6] : w_7 w_2 w_4}$$

$$\frac{Z \rightarrow X\, zhiyi : w_6 \quad [X 7, 9] : w_5}{[Z, 7, 11] : w_5 w_6}$$

Goal : [S,0,n]

K-best Lists

- Identify k-best derivations
- Used for Minimum error rate training
- Example: L1 = {1,2,6,10} and L2 = {1,4,7}

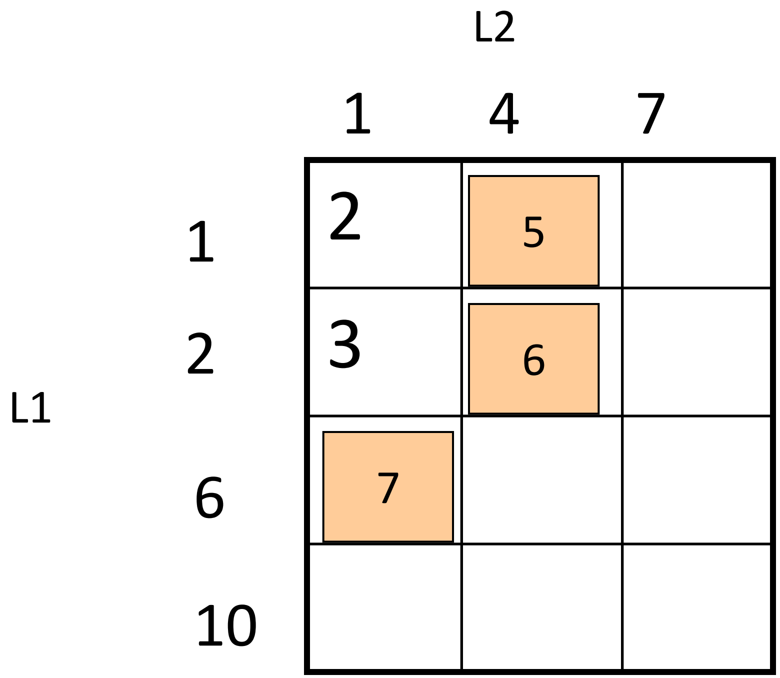
$$\frac{Z \rightarrow XY : w \quad \overset{\text{L1}}{\boxed{[X, i, k] : w_1}} \quad \overset{\text{L2}}{\boxed{[Y, k, j] : w_2}}}{[Z, i, j] : w_1 w_2 w}$$

		L2		
		1	4	7
	1			
L1	2			
	6			
	10			

$$\frac{Z \to XY : w \quad \overset{\text{L1}}{\boxed{[X, i, k] : w_1}} \quad \overset{\text{L2}}{\boxed{[Y, k, j] : w_2}}}{[Z, i, j] : w_1 w_2 w}$$

		L2		
		1	4	7
L1	1	2	5	
	2	3		
	6			
	10			

$$\frac{Z \to XY : w \qquad \overset{\text{L1}}{\boxed{[X, i, k] : w_1}} \qquad \overset{\text{L2}}{\boxed{[Y, k, j] : w_2}}}{[Z, i, j] : w_1 w_2 w}$$



$$\frac{Z \rightarrow XY : w \quad \overset{\text{L1}}{\boxed{[X, i, k] : w_1}} \quad \overset{\text{L2}}{\boxed{[Y, k, j] : w_2}}}{[Z, i, j] : w_1 w_2 w}$$

L2

L1		1	4	7
	1	2	5	8
	2	3	6	
	6	7		
	10			

$$\frac{Z \to XY : w \quad \overset{\text{L1}}{\boxed{[X, i, k] : w_1}} \quad \overset{\text{L2}}{\boxed{[Y, k, j] : w_2}}}{[Z, i, j] : w_1 w_2 w}$$

Adding the Language Model

- Rescoring
  - Finding the k-best list using –LM parser
  - Rescoring the k-best list using LM
  - Linear in k
  - We may need to set k to be extremely high
- Intersection
- Cube Pruning

Intersection

$$\frac{X \to \langle f_{i+1}^{j}, \alpha \rangle : w}{[X, i, j; q(\alpha)] : wp(\alpha)}$$

$$\frac{Z \to \langle f_{i+1}^{i_1} X f_{j_1+1}^{j}, \alpha \rangle : w \quad [X, i_1, j_1; e_1] : w_1}{[Z, i, j; q(\alpha')] : ww_1 p(\alpha')}$$

$$\alpha' = \alpha[e_1 / X]$$

$$\frac{Z \to \langle f_{i+1}^{i_1} X_{\boxed{1}} f_{j_1+1}^{i_2} Y_{\boxed{2}} f_{j_2+1}^{j}, \alpha \rangle : w \quad [X, i_1, j_1; e_1] : w_1 \quad [Y, i_2, j_2; e_2] : w_2}{[Z, i, j; q(\alpha')] : ww_1 w_2 p(\alpha')}$$

$$\alpha' = \alpha[e_1 / X_{\boxed{1}}, e_2 / Y_{\boxed{2}}]$$

Intersection

- Two function to correctly calculate the LM score of a sentence piecemeal

$$p(a_1 \dots a_l) = \prod_{\substack{m \le i \le l \\ \star \notin \{a_{i-m+1} \dots a_{i-1}\}}} P_{LM}(a_i | a_{i-m+1} \dots a_{i-1})$$

$$\mathrm{q}(a_1 \dots a_l) = \begin{cases} a_1 \dots a_{m-1} \star a_{l-m+1} \dots a_l & \mathit{if}\ l \ge m \\ a_1 \dots a_l & \mathit{otherwise} \end{cases}$$

- p calculates LM probabilities for all the complete m grams
- q keeps the last and first m-1 words of a string

Intersection

Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi

0 1 2 3 4 5 6 7 8 9

$$\frac{X \rightarrow \langle Aozhou, Australia \rangle : w_1}{[X, 0, 1] : w_1 p(Australia)}$$

Intersection

Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi

0 1 2 3 4 5 6 7 8 9

$$\frac{X \rightarrow \langle Aozhou, Australia \rangle : w_1}{[X, 0, 1] : w_1 p(Australia)} \qquad \frac{X \rightarrow \langle shi, is \rangle : w_3}{[X, 1, 2] : w_3 p(is)}$$

Intersection

Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi

0 1 2 3 4 5 6 7 8 9

$$\frac{X \rightarrow \langle Aozhou, Australia \rangle : w_1}{[X, 0, 1] : w_1 p(Australia)} \quad \frac{X \rightarrow \langle shi, is \rangle : w_3}{[X, 1, 2] : w_3 p(is)} \quad \frac{X \rightarrow \langle shaoshu\ \ guojio, few\ countries \rangle : w_5}{[X, 7, 9] : w_5 p(few\ countries)}$$

$$\frac{X \rightarrow \langle Beihon, North\ Korea \rangle : w_2}{[X, 3, 4] : w_2 p(North\ Korea)} \quad \frac{X \rightarrow \langle bangjiao, diplomatic\ relations \rangle : w_4}{[X, 5, 6] : w_4 p(diplomatic\ relations)}$$

Intersection

Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi

0 1 2 3 4 5 6 7 8 9

$$\frac{X \rightarrow \langle Aozhou, Australia \rangle : w_1}{[X, 0, 1] : w_1 p(Australia)} \quad \frac{X \rightarrow \langle shi, is \rangle : w_3}{[X, 1, 2] : w_3 p(is)} \quad \frac{X \rightarrow \langle shaoshu\ \ guojio, few\ countries \rangle : w_5}{[X, 7, 9] : w_5 p(few\ countries)}$$

$$\boxed{\frac{X \rightarrow \langle Beihon, North\ Korea \rangle : w_2}{[X, 3, 4] : w_2 p(North\ Korea)}} \quad \boxed{\frac{X \rightarrow \langle bangjiao, diplomatic\ relations \rangle : w_4}{[X, 5, 6] : w_4 p(diplomatic\ relations)}}$$

$$\frac{Z \rightarrow \langle yu\ X_1\ you\ X_2, yuX_1youX_2 \rangle : w_7 \quad [X_1, 3, 4, North\ Korea] : w_2' \quad [X_2, 5, 6, diplmatic\ relations] : w_4'}{[Z, 2, 6, have\ dipl * with\ NK] : w_7 w_2' w_4' p(have\ dipl\ rels\ with\ NK)}$$

Intersection

- Too slow in practice

- Pruning : for each span throw out items with score worse than:
  - the score of *b*th best item for that span
  - β + the score of best item for that span

Cube Pruning

		[X, 6, 8; the scheme]	[X, 6, 8; the plan]	[X, 6, 8; the project]
		1	4	7
$X \rightarrow \langle \text{cong } X_{\boxed{1}}, \text{from } X_{\boxed{1}} \rangle$	1	2.1	5.1	8.2
$X \rightarrow \langle \text{cong } X_{\boxed{1}}, \text{from the } X_{\boxed{1}} \rangle$	2	5.5	8.5	11.5
$X \rightarrow \langle \text{cong } X_{\boxed{1}}, \text{since } X_{\boxed{1}} \rangle$	6	7.7	10.6	13.1
$X \rightarrow \langle \text{cong } X_{\boxed{1}}, \text{through } X_{\boxed{1}} \rangle$	10	11.1	14.3	17.3

$\Rightarrow$

[X, 5, 8; from the $\star$ the scheme] : 2.1

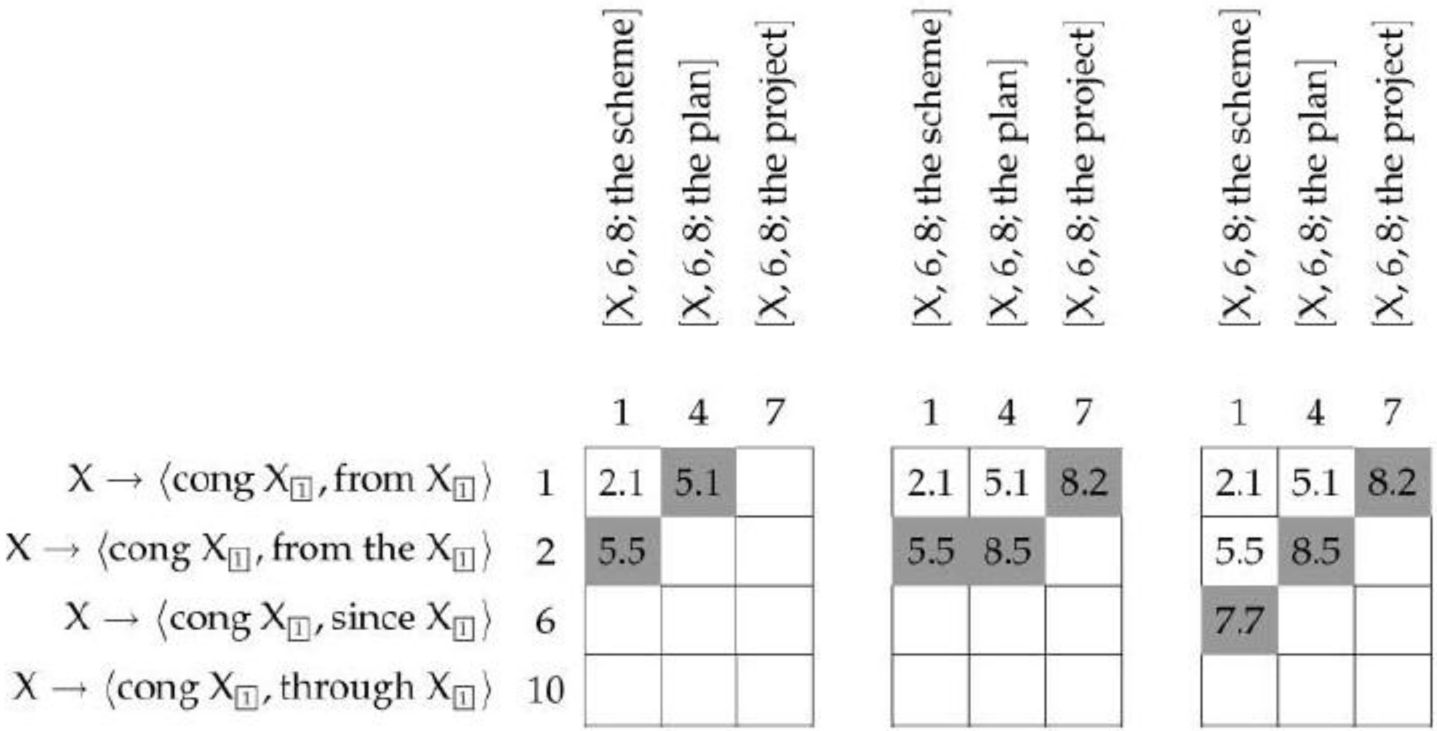
[X, 5, 8; from the $\star$ the plan] : 5.1

[X, 5, 8; from the $\star$ the scheme] : 5.5

[X, 5, 8; since the $\star$ the scheme] : 7.7

$\vdots$

Cube Pruning



Experiments

Experimental Results

- Comparing performances of decoding methods

Method	Settings	Time	BLEU
rescore	$k = 10^4$	16	33.31
rescore	$k = 10^5$	139	33.33
intersect*		1455	37.09
cube prune	$\varepsilon = 0$	23	36.14
cube prune	$\varepsilon = 0.1$	35	36.77
cube prune	$\varepsilon = 0.2$	111	36.91

Experimental Results

- 2 baselines :
  - ATS
  - Hiero Monotone : same as Hiero except without any non-terminals on right hand side

System	MT03	MT04	MT05
Hiero Monotone	28.27 ± 1.03	28.83 ± 0.74	26.35 ± 0.92
ATS	30.84 ± 0.99	31.74 ± 0.73	30.50 ± 0.95
Hiero	33.72 ± 1.12	34.57 ± 0.82	31.79 ± 0.91

Questions ??