

Machine Translation 2: Phrase-Based Translation

Computational Linguistics

Alexander Koller

21 December 2018

slides contain material from mt-class.org

Translation quality

- We can measure quality of a translation in two dimensions:
 - ▶ *Adequacy*: How accurately does translation represent the meaning of the original?
 - ▶ *Fluency*: Is the translation a good string of the target language (“good English”)?
- How can we select a fluent translation?

Fluency

Israeli officials are responsible for airport security.

Israel is in charge of the security at this airport.

The security work for this airport is the responsibility of the Israel government.

Israeli side was in charge of the security of this airport.

Israel is responsible for the airport's security.

Israel is responsible for safety work at this airport.

Israel presides over the security of the airport.

Israel took charge of the airport security.

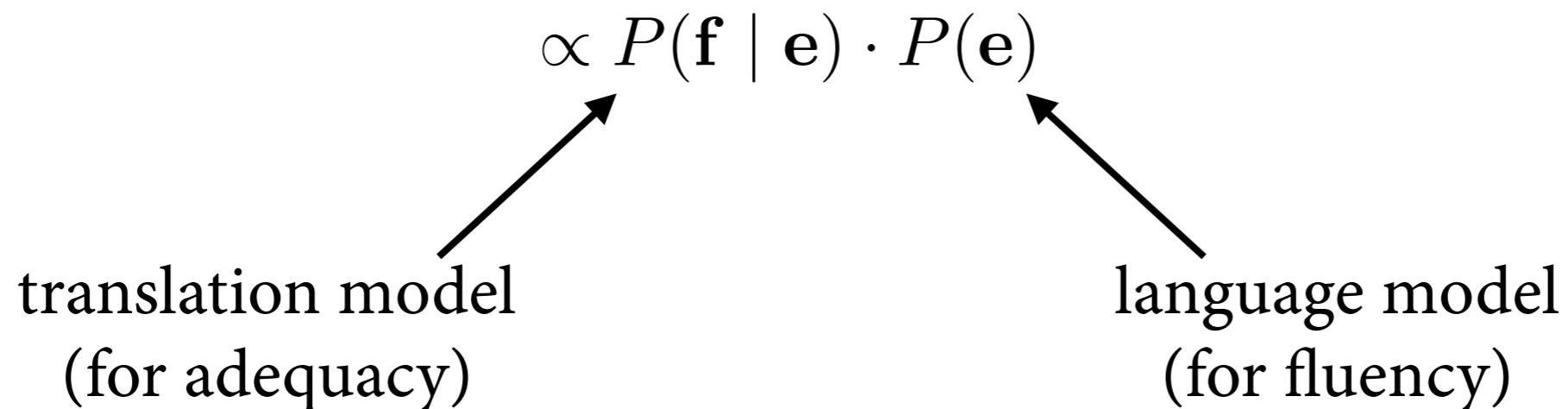
The safety of this airport is taken charge of by Israel.

This airport's security is the responsibility of the Israeli security officials.

Noisy Channel Model

- We can model fluency with a *language model* $P(\mathbf{e})$ of the target language.
 - ▶ can estimate from lots of monolingual data!
- Noisy Channel Model:

$$P(\mathbf{e} | \mathbf{f}) = \frac{P(\mathbf{f} | \mathbf{e}) \cdot P(\mathbf{e})}{P(\mathbf{f})}$$



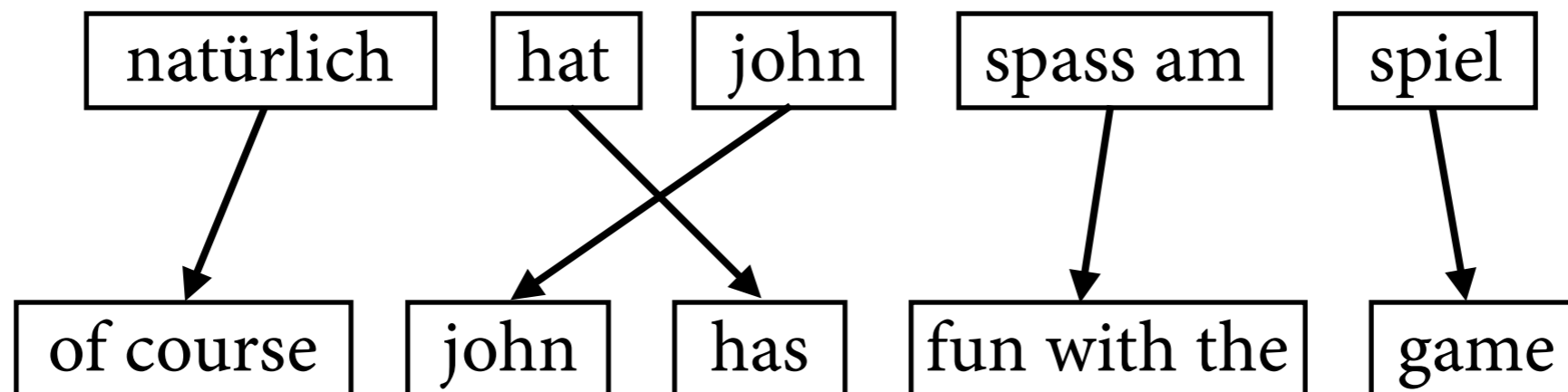
Word-based translation model

- Could derive model for word-by-word translation, e.g. from IBM Model 1:

$$P(\mathbf{f} | \mathbf{e}) = \sum_a P(\mathbf{f}, a | \mathbf{e})$$
$$\propto \prod_{j=1}^{l_f} \sum_{i=1}^{l_e} P(f_j | e_i)$$

Phrase-based translation

- But want to translate entire *phrases* (i.e. substrings):
 - ▶ translation of one word can consist of multiple words
 - ▶ context of word in phrase can help disambiguate



- Note: these “phrases” need not be linguistically meaningful constituents.

Phrase-based translation model

$$P(\mathbf{f} \mid \mathbf{e}) = \prod_{i=1}^I \phi(\bar{f}_i \mid \bar{e}_i) \cdot d(\text{start}_i - \text{end}_{i-1} - 1)$$

number of phrases

phrase translation probability

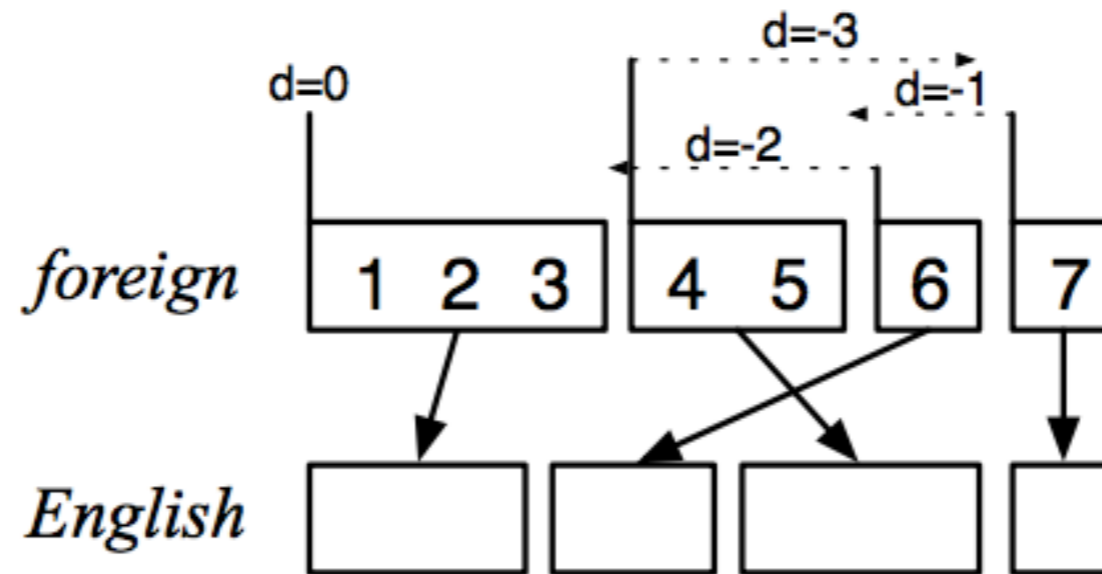
distance-based reordering model

The diagram illustrates the components of the phrase-based translation model equation. The equation is $P(\mathbf{f} \mid \mathbf{e}) = \prod_{i=1}^I \phi(\bar{f}_i \mid \bar{e}_i) \cdot d(\text{start}_i - \text{end}_{i-1} - 1)$. Three arrows point from text labels to parts of the equation: one from 'number of phrases' to the index I in the product; one from 'phrase translation probability' to the function $\phi(\bar{f}_i \mid \bar{e}_i)$; and one from 'distance-based reordering model' to the distance function $d(\text{start}_i - \text{end}_{i-1} - 1)$.

(the whole thing gets multiplied by $P(\mathbf{e})$ later)

Reordering Model

Let's assume a simple model for reordering for now.



phrase	translates	movement	distance
1	1-3	start at beginning	0
2	6	skip over 4-5	+2
3	4-5	move back over 4-6	-3
4	7	skip over 6	+1






Scoring function: $d(x) = \alpha^{|x|}$ — exponential with distance

Learning phrase translations




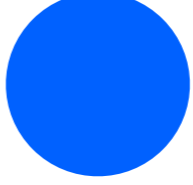

- Extend word alignments to phrase alignments.
- Collect all phrase pairs from the parallel corpus (both big and small — we want *all* phrase pairs).
- Estimate phrase translation probabilities $P(\mathbf{f} | \mathbf{e})$ using maximum likelihood estimation (plus smoothing).

Phrase Extraction

I open the box






watashi				
wa				
hako				
wo				
akemasu				

Phrase Extraction

	I	open	the	box
watashi				
wa				
hako				
wo				
akemasu				

akemasu / open

Phrase Extraction

	I	open	the	box
watashi				
wa				
hako				
wo				
akemasu				






watashi wa / I

Phrase Extraction

	I	open	the	box
watashi	●			
wa	●			
hako				●
wo				●
akemasu		●		






hako wo / box

Phrase Extraction

	I	open	the	box
watashi				
wa				
hako				
wo				
akemasu				

hako wo / the box

Phrase Extraction

	I	open	the	box
watashi				
wa				
hako				
wo				
akemasu				

hako wo akemasu / open the box

Decoding

- We now have:
 - ▶ noisy channel $P(\mathbf{e} | \mathbf{f}) \propto P(\mathbf{f} | \mathbf{e}) * P(\mathbf{e})$
 - ▶ language model $P(\mathbf{e})$
 - ▶ phrase-based translation model

$$P(\mathbf{f} | \mathbf{e}) = \prod_{i=1}^I \phi(\bar{f}_i | \bar{e}_i) \cdot d(\text{start}_i - \text{end}_{i-1} - 1)$$

- We need to solve the *decoding* problem:
for a given \mathbf{f} , compute $\text{argmax}_{\mathbf{e}} P(\mathbf{e} | \mathbf{f})$.

Basic idea

er geht ja nicht nach hause

Basic idea

er geht ja nicht nach hause

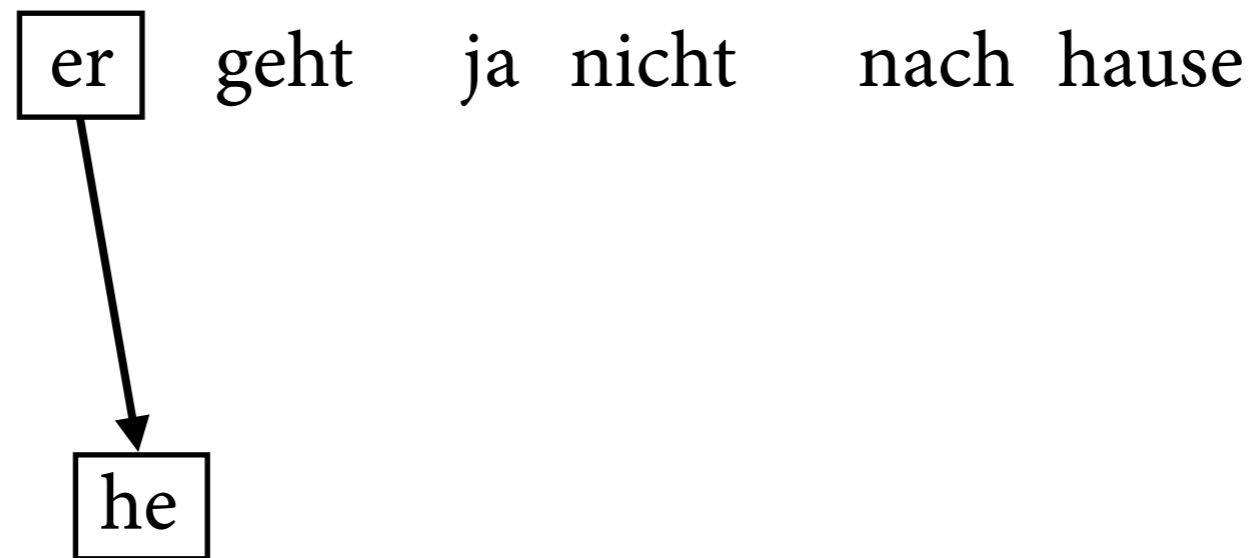


Basic idea

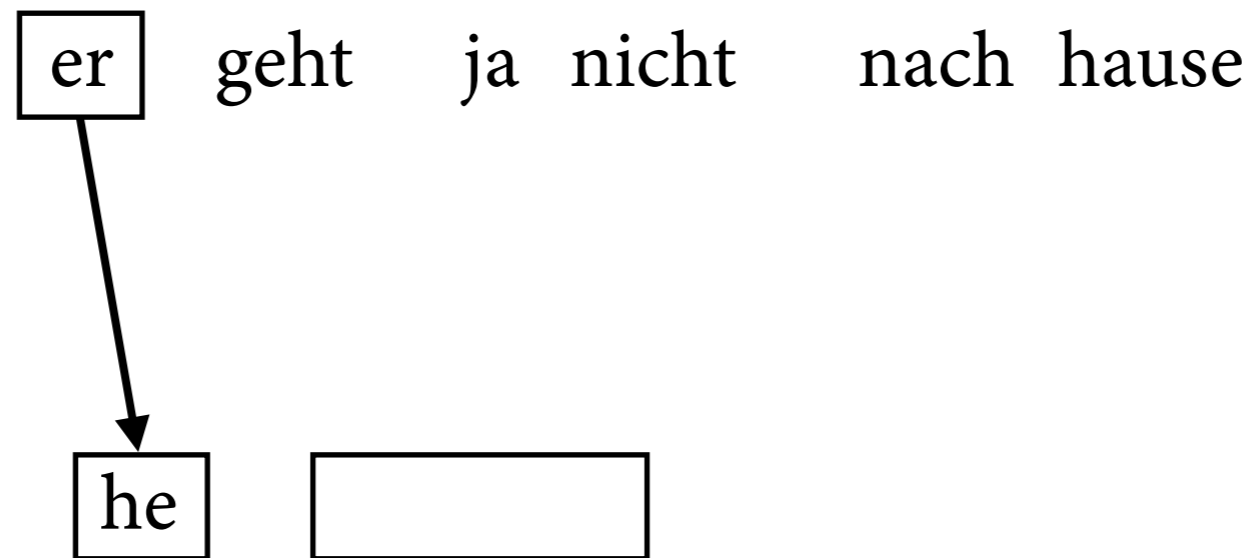
er geht ja nicht nach hause



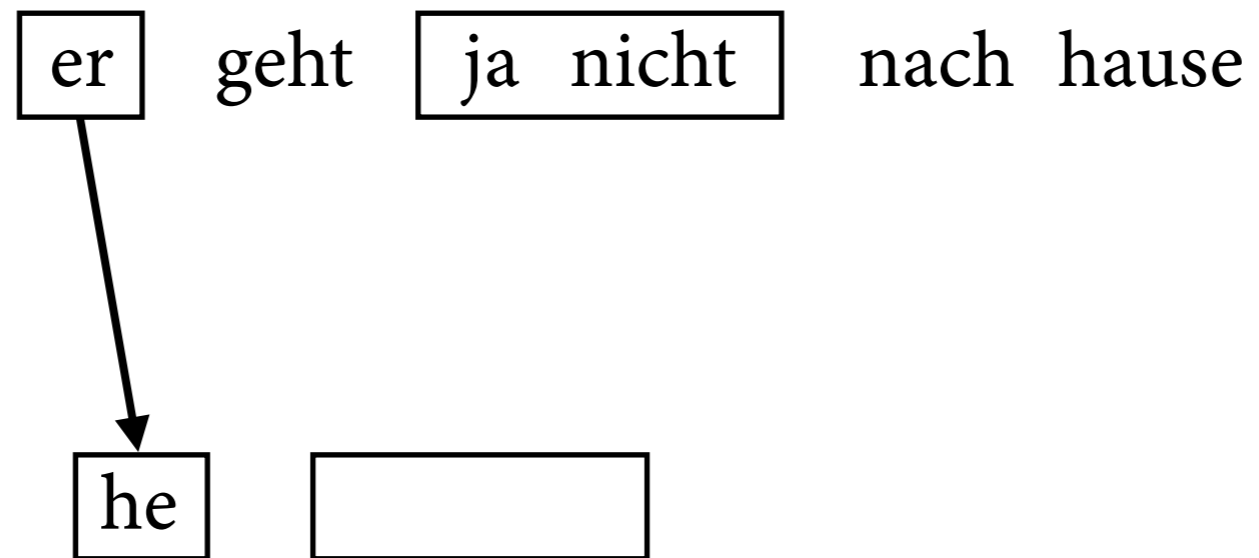
Basic idea



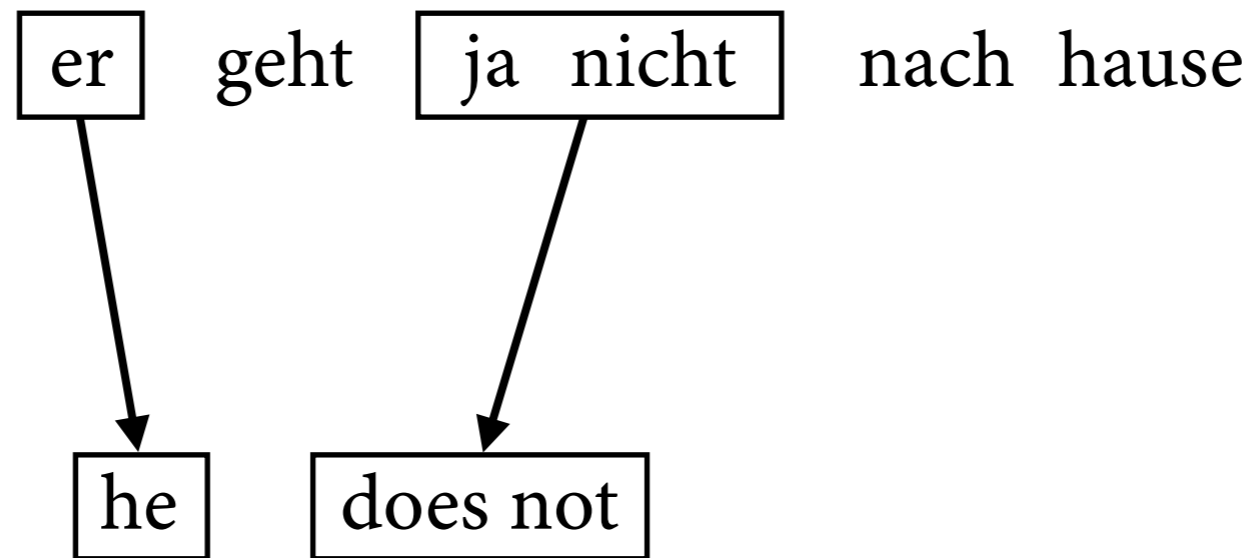
Basic idea



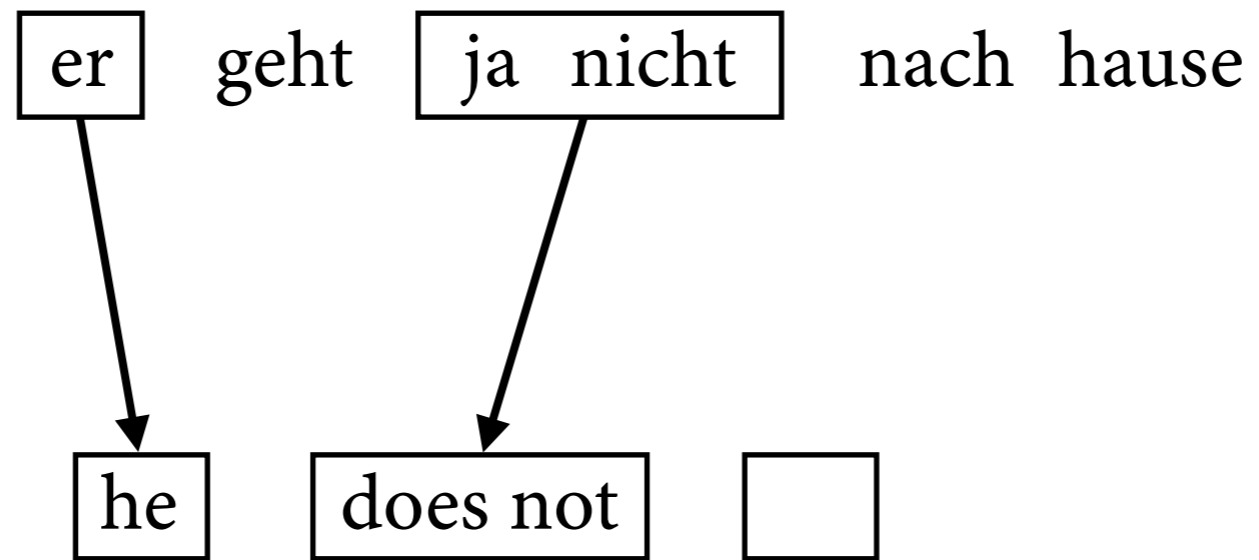
Basic idea



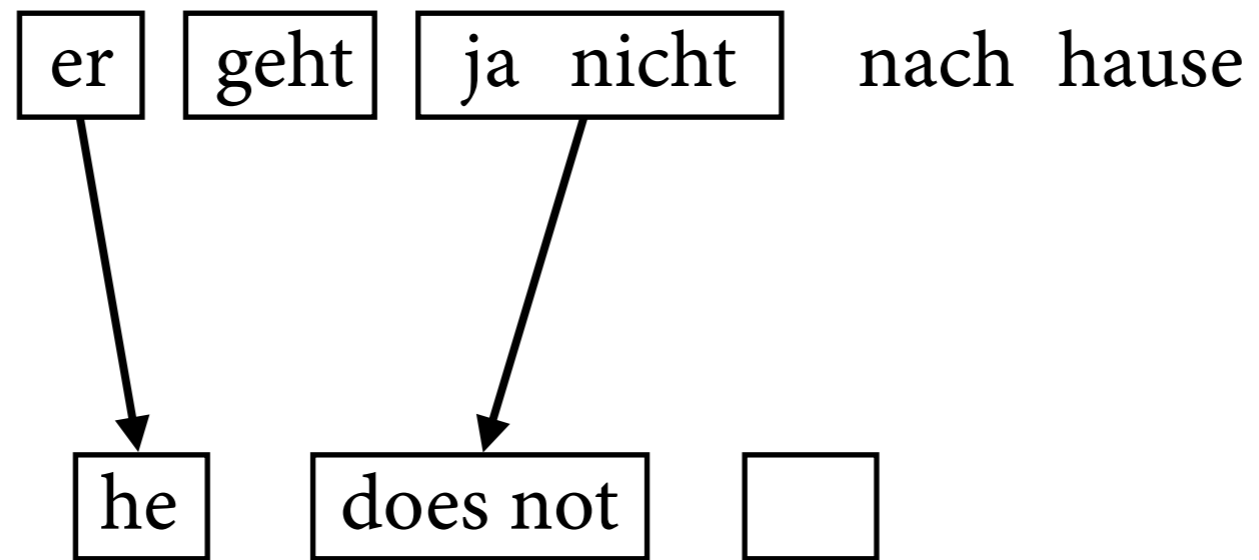
Basic idea



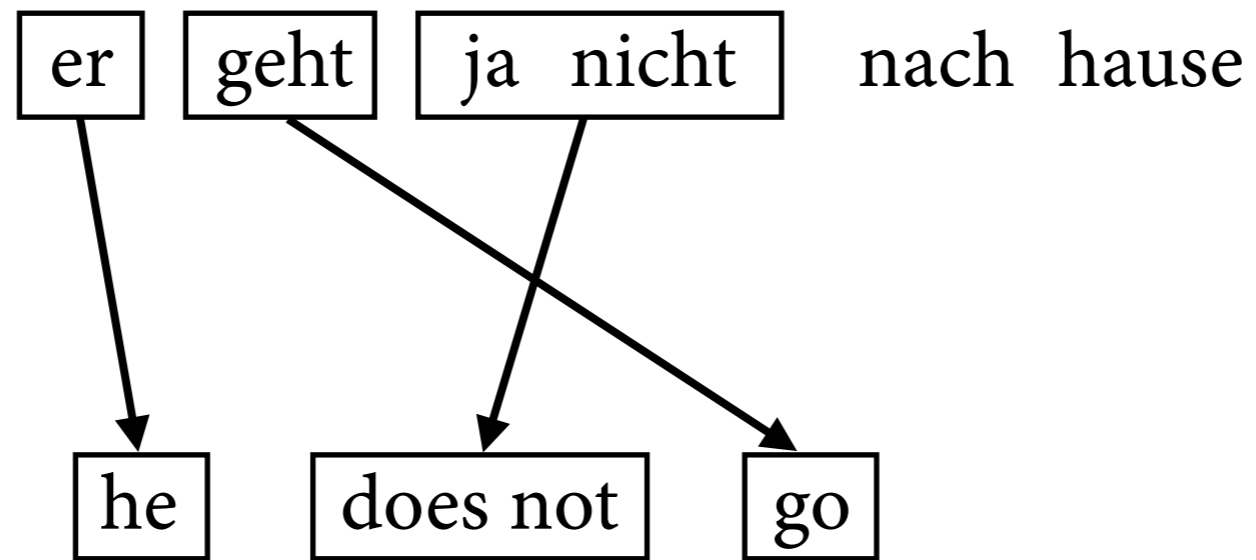
Basic idea



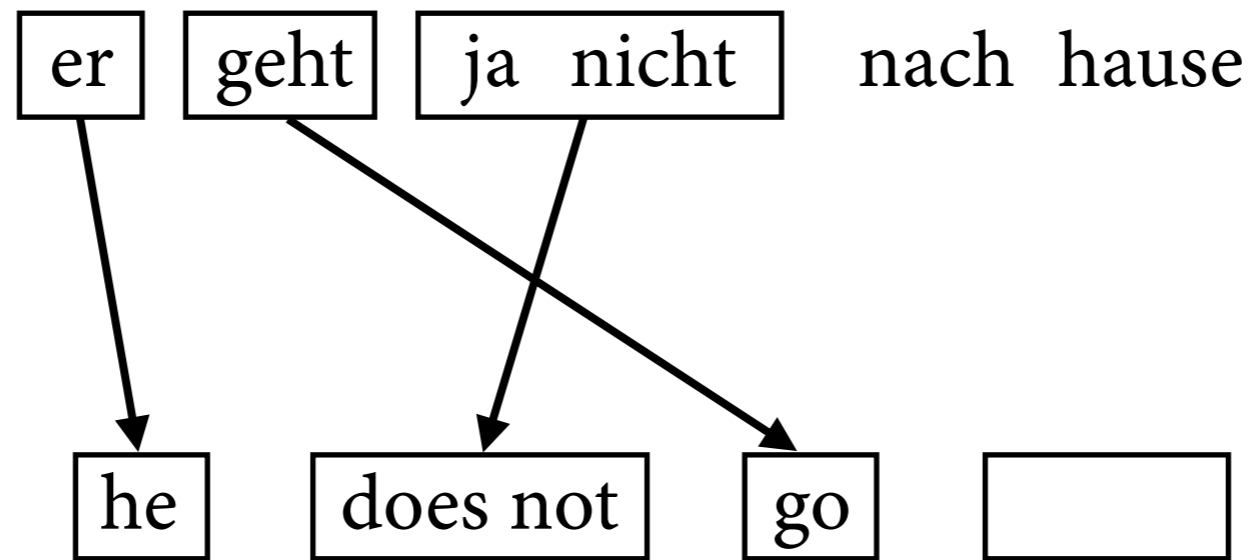
Basic idea



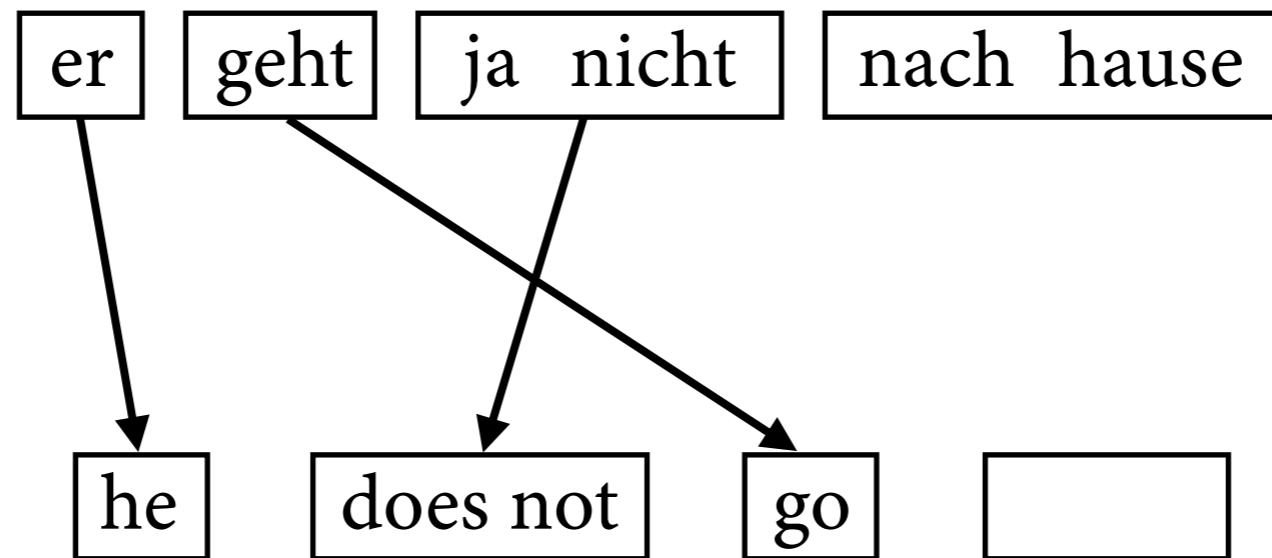
Basic idea



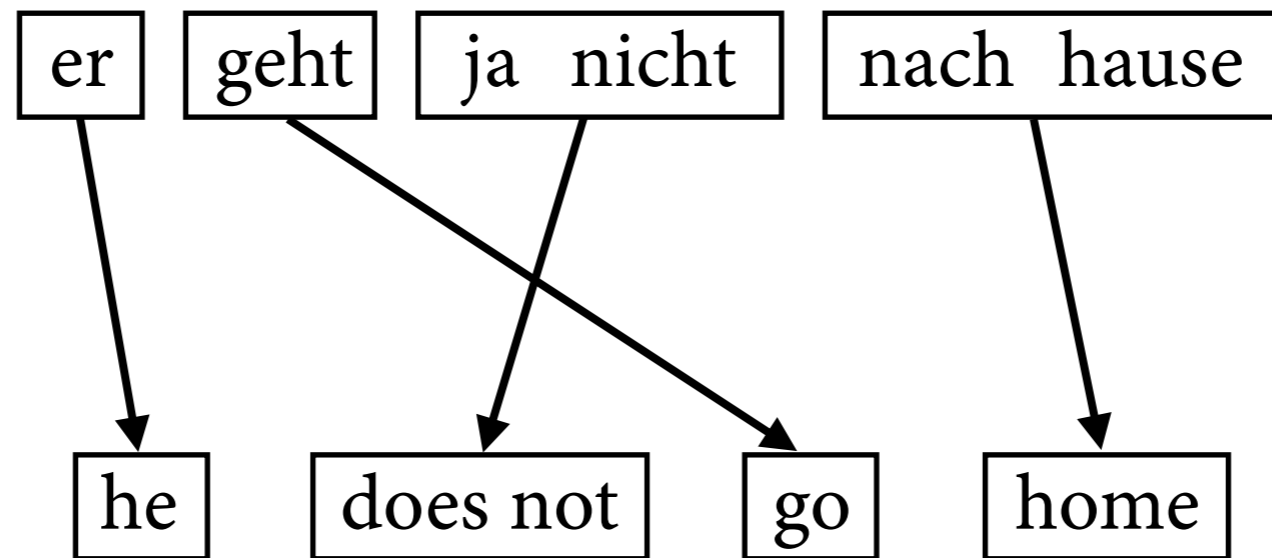
Basic idea



Basic idea



Basic idea



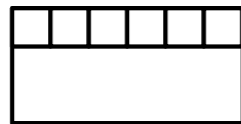
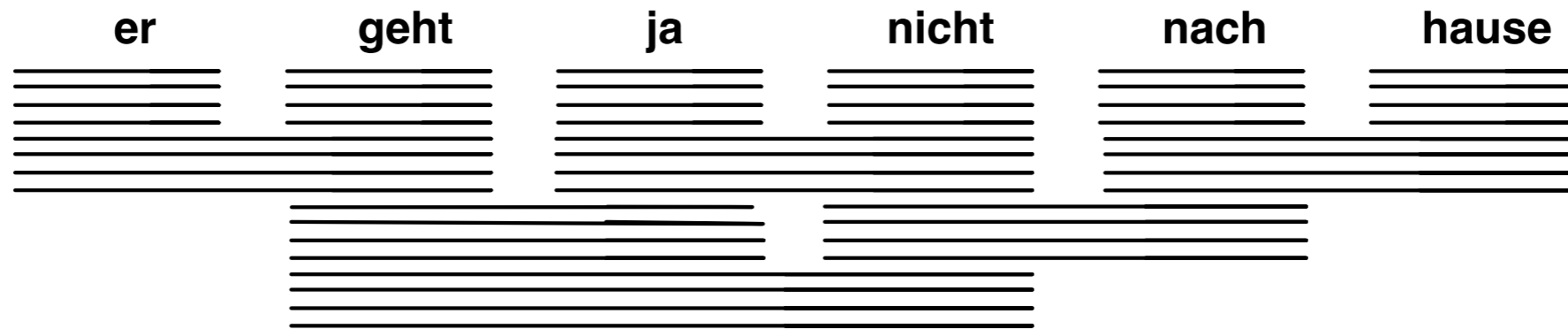
More realistically

er	geht	ja	nicht	nach	hause
he	is	yes	not	after	house
it	are	is	do not	to	home
, it	goes	, of course	does not	according to	chamber
, he	go	,	is not	in	at home
it is		not		home	
he will be		is not		under house	
it goes		does not		return home	
he goes		do not		do not	
	is		to		
	are		following		
	is after all		not after		
	does		not to		
	not				
	is not				
	are not				
	is not a				

- Many translation options to choose from
 - in Europarl phrase table: 2727 matching phrase pairs for this sentence
 - by pruning to the top 20 per phrase, 202 translation options remain

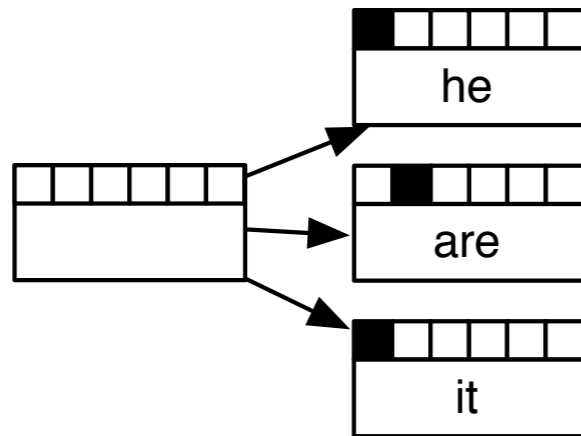
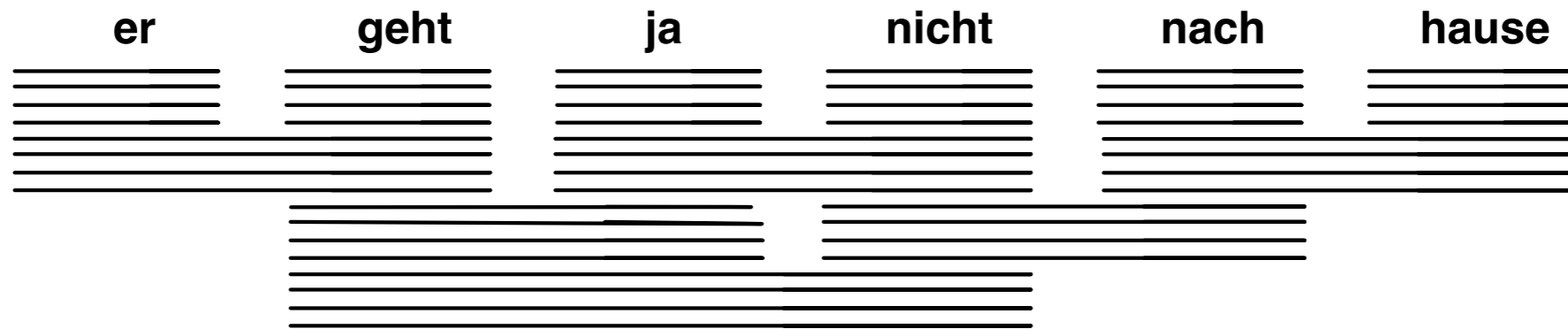
Decoding as Search

start with empty hypothesis (no words translated)



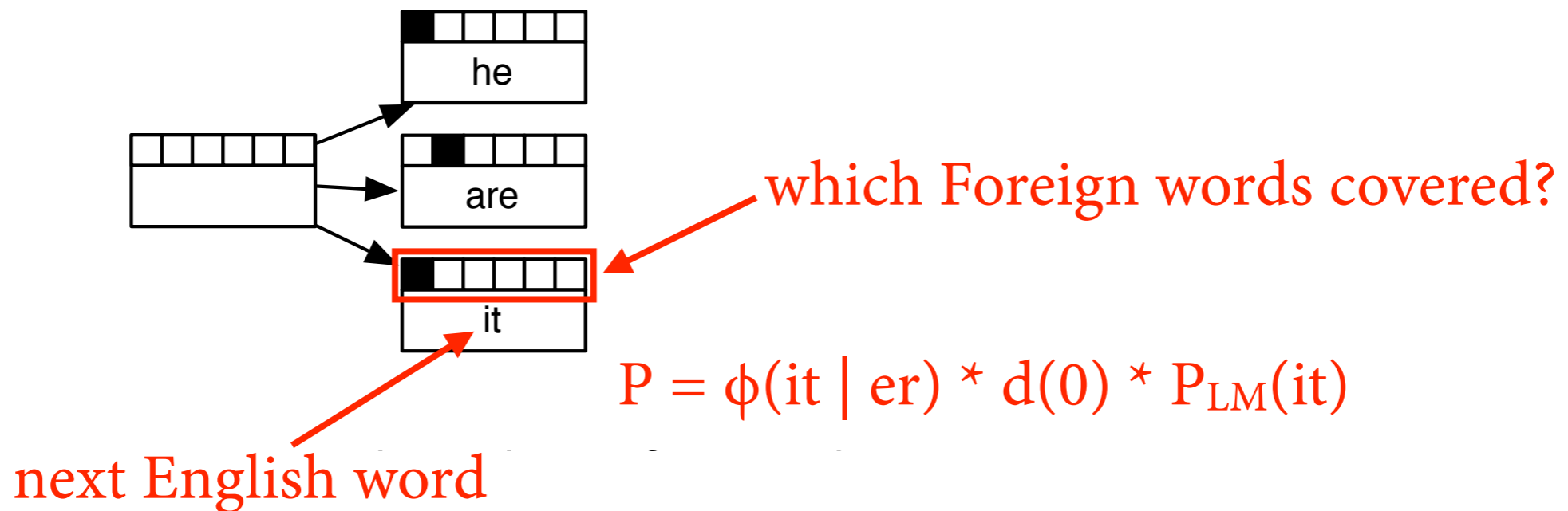
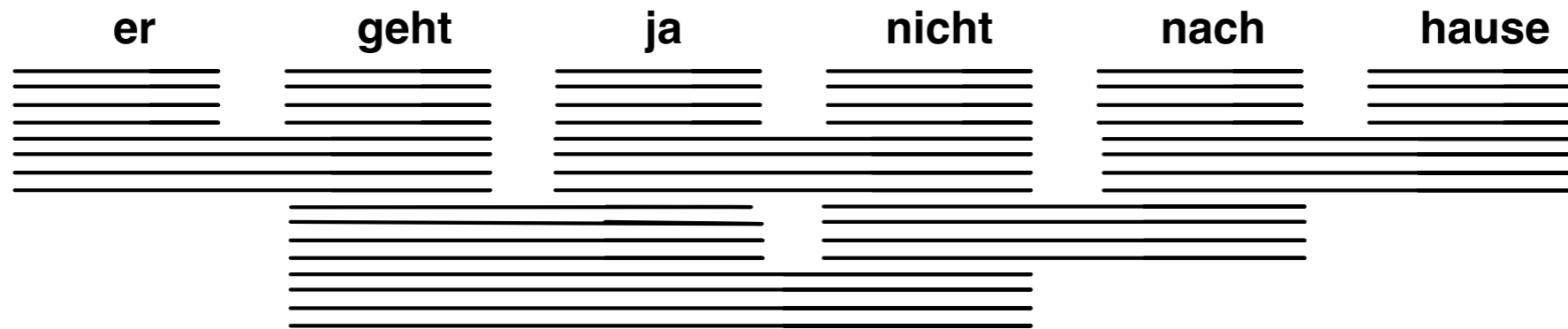
Decoding as Search

expand hypotheses by next English word



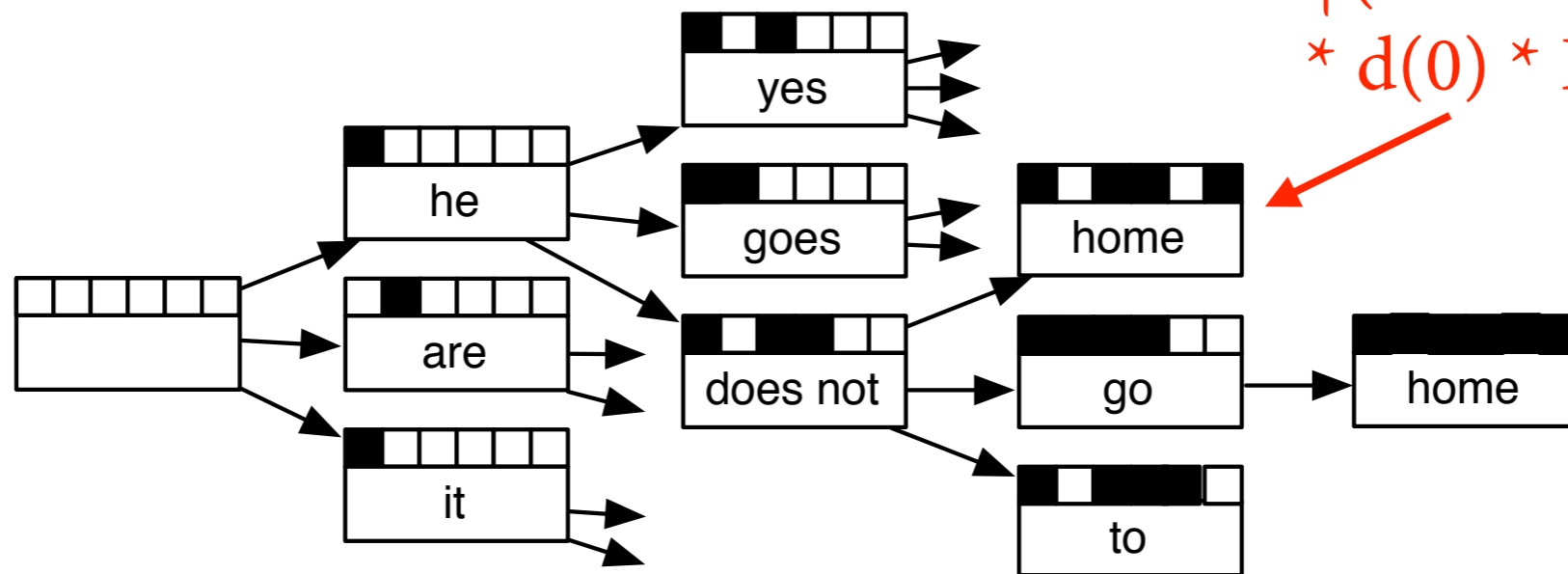
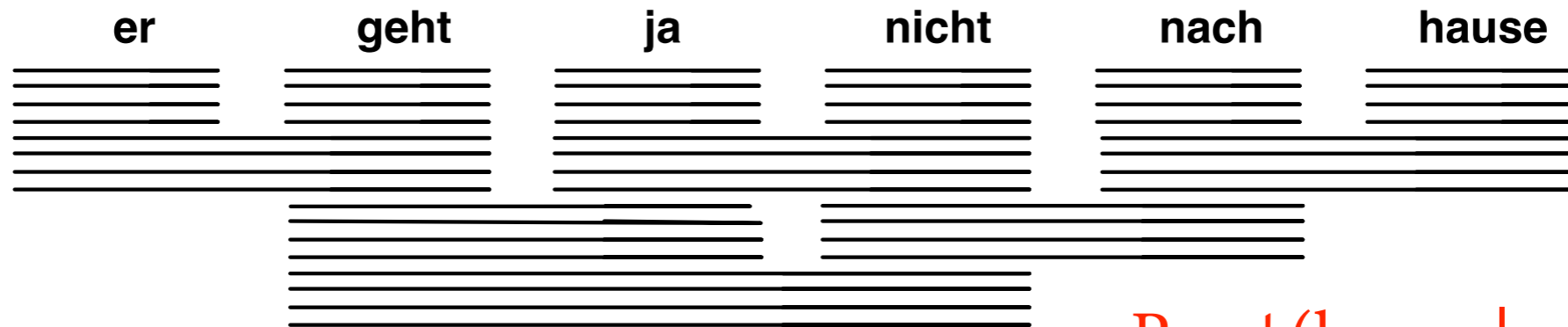
Decoding as Search

expand hypotheses by next English word



Decoding as Search

continue expanding hypotheses

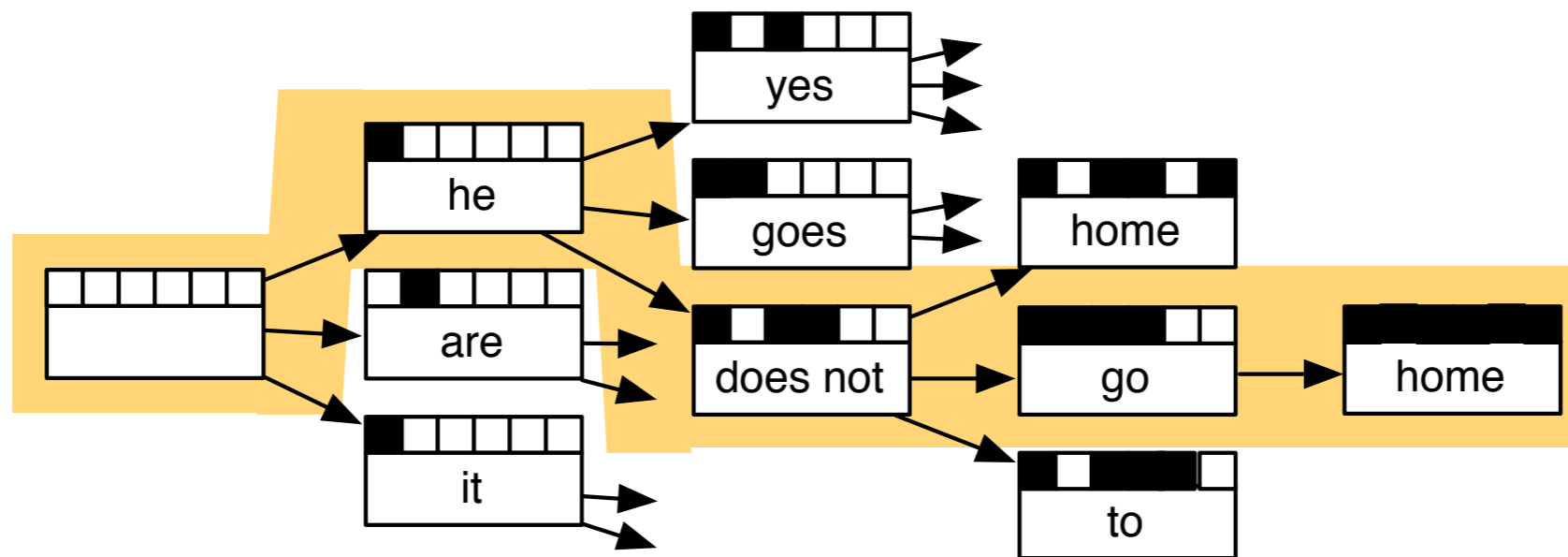
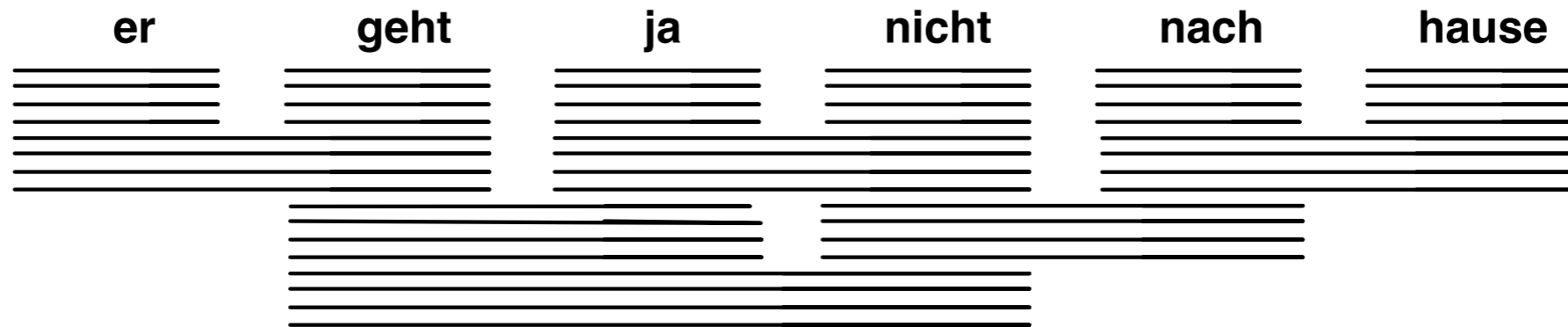


$$P = \phi(\text{home} \mid \text{nach Hause}) \\ * d(0) * P_{\text{LM}}(\text{home} \mid \text{not})$$



Decoding as Search

backtrack from highest-scoring complete hypothesis



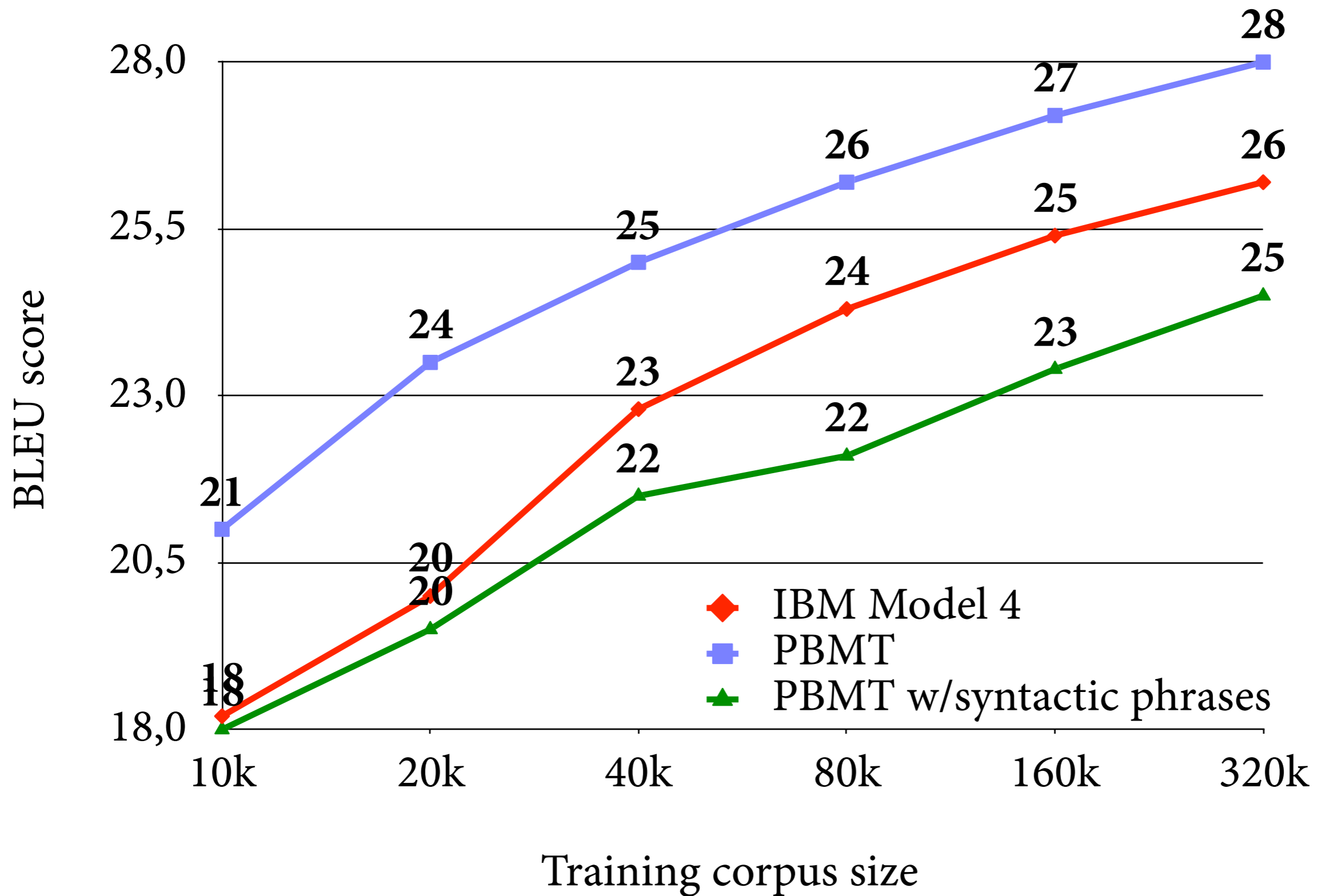
Computational issues

- Search space is huge.
 - ▶ exponential in sentence length (because of free reordering)
 - ▶ in fact, finding best translation is NP-complete
- Need heuristics to deal with complexity.
 - ▶ beam search: *stack decoding*
 - ▶ A* search

Putting linguistics in SMT

- Word-based, phrase-based SMT very naive from a linguistics perspective.
- Can we do better by putting linguistics into SMT? (At least a bit of syntax?)
- Received wisdom before 2005: phrase-based translation with lots of data much better; syntax hurts.

Syntax can hurt



Chinese-English reordering

(output of phrase-based system ATS)

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 .
Australia is with North Korea have dipl. rels. that few countries one of .

“Australia is one of the few countries that have diplomatic relations with North Korea.”

Chinese-English reordering

(output of phrase-based system ATS)

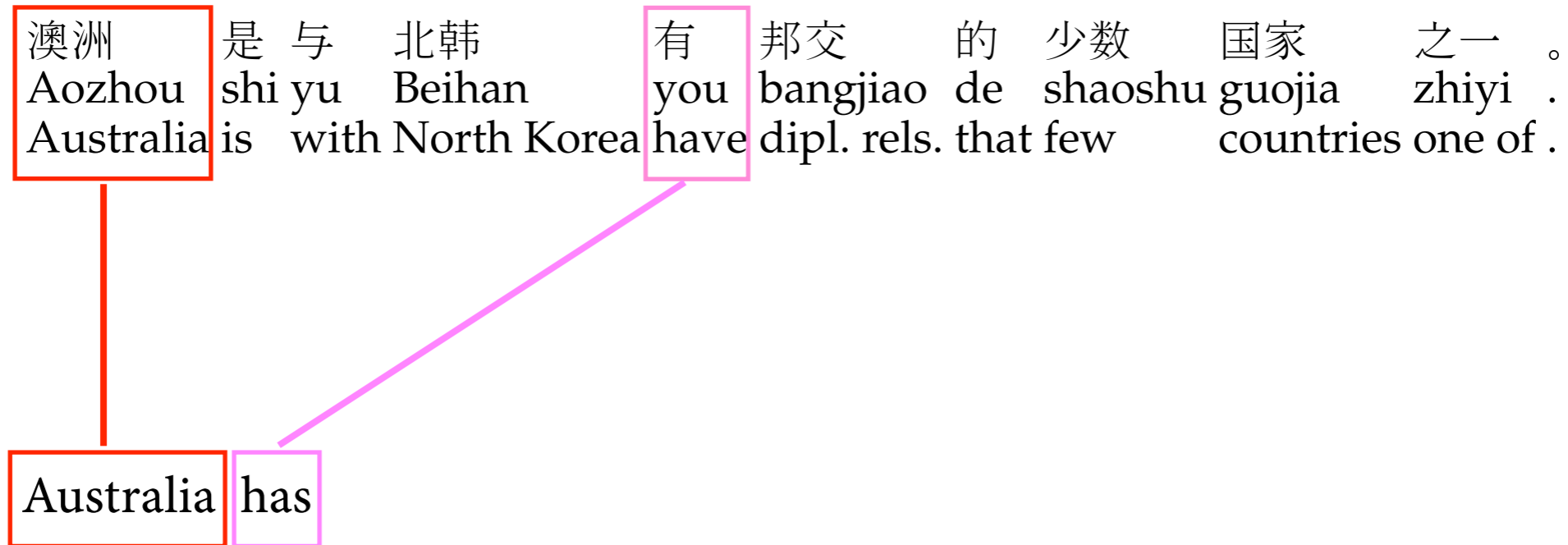
澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。
Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi .
Australia is with North Korea have dipl. rels. that few countries one of .

Australia

“Australia is one of the few countries that have diplomatic relations with North Korea.”

Chinese-English reordering

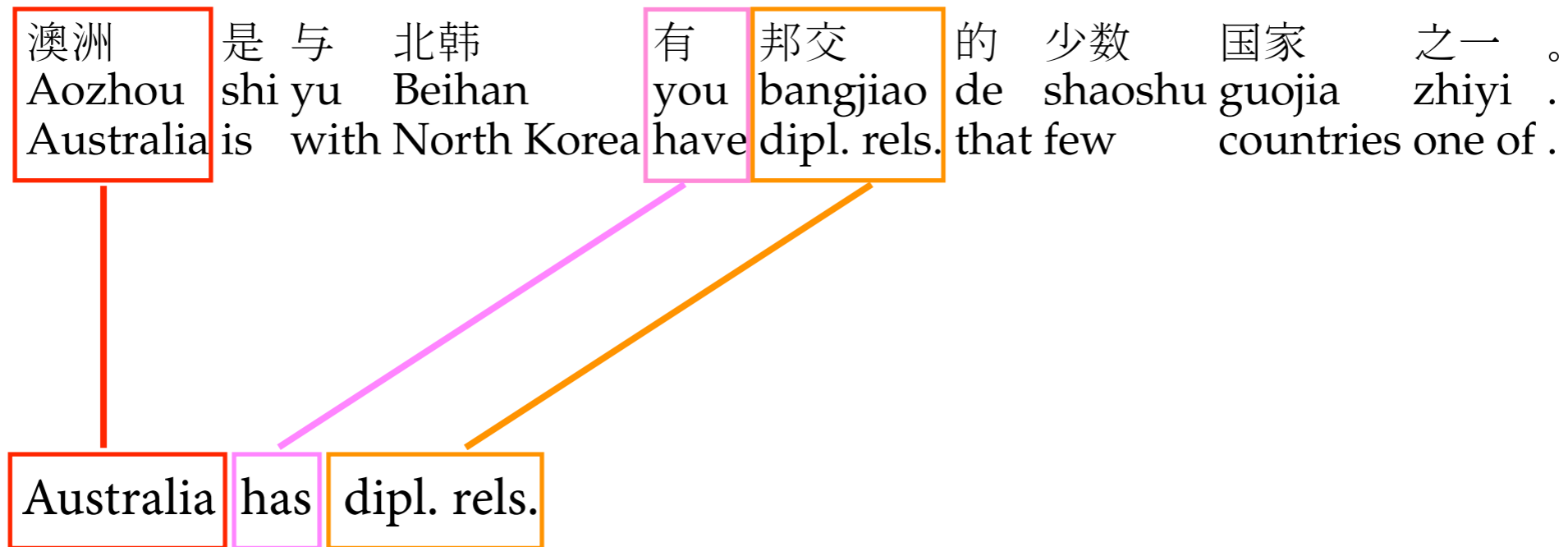
(output of phrase-based system ATS)



“Australia is one of the few countries that have diplomatic relations with North Korea.”

Chinese-English reordering

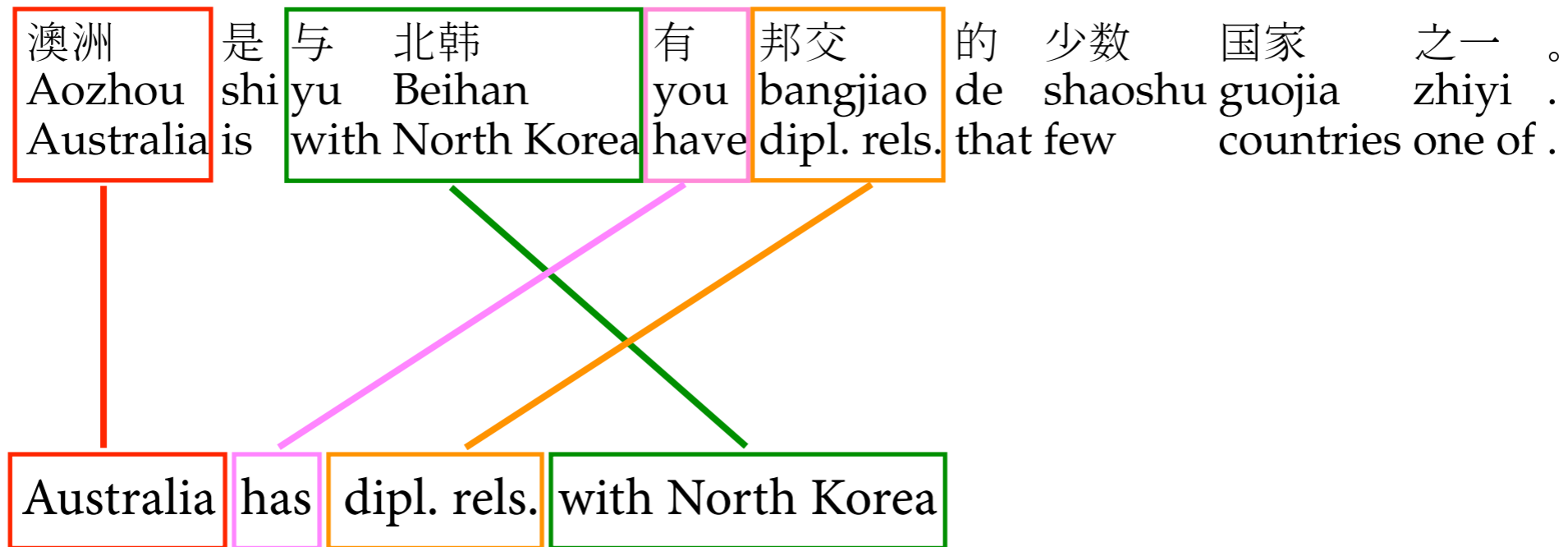
(output of phrase-based system ATS)



“Australia is one of the few countries that have diplomatic relations with North Korea.”

Chinese-English reordering

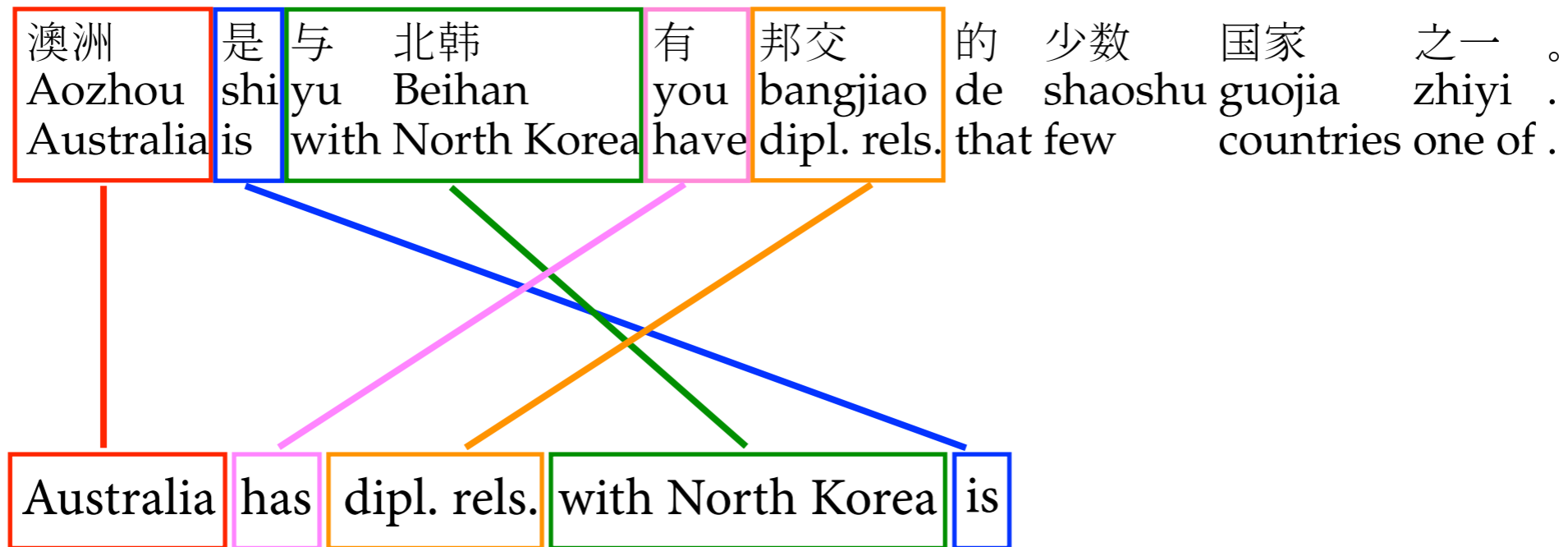
(output of phrase-based system ATS)



“Australia is one of the few countries that have diplomatic relations with North Korea.”

Chinese-English reordering

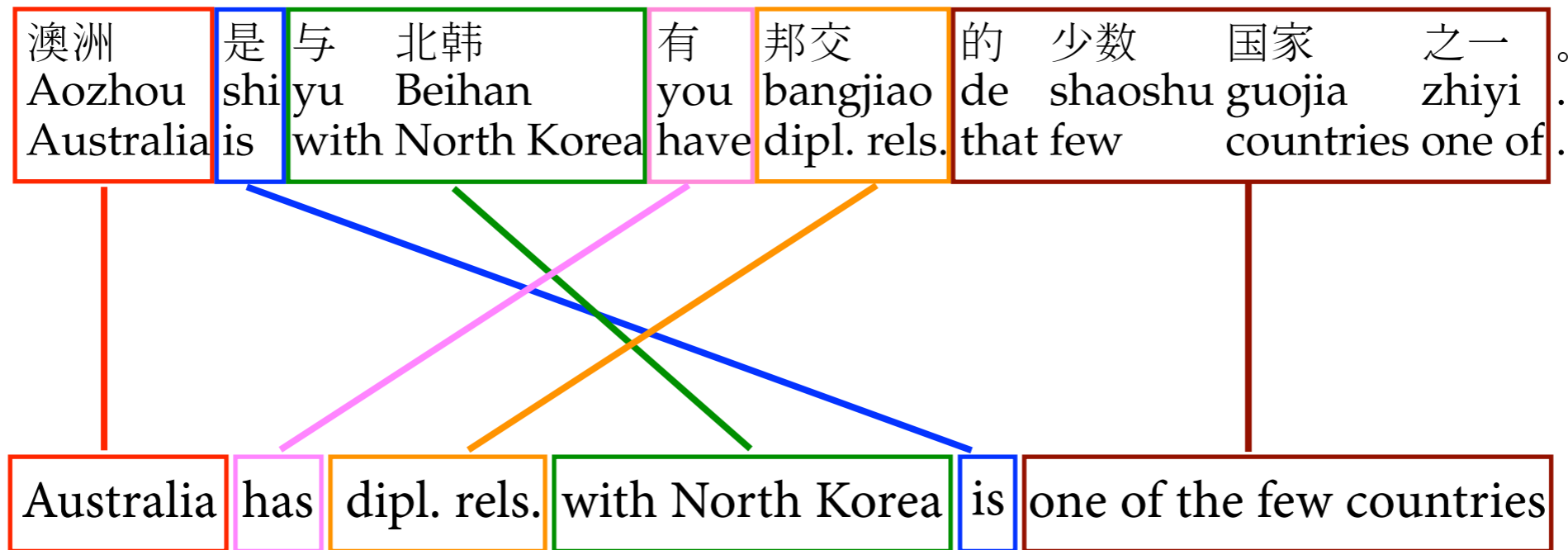
(output of phrase-based system ATS)



“Australia is one of the few countries that have diplomatic relations with North Korea.”

Chinese-English reordering

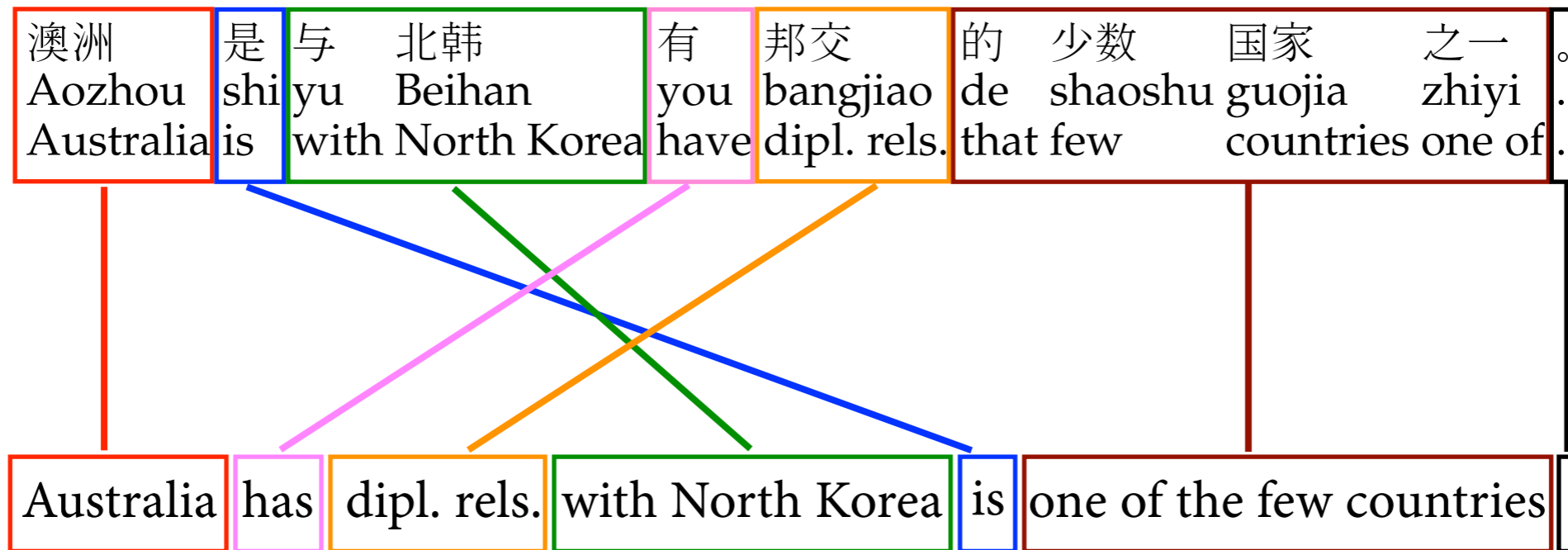
(output of phrase-based system ATS)



“Australia is one of the few countries that have diplomatic relations with North Korea.”

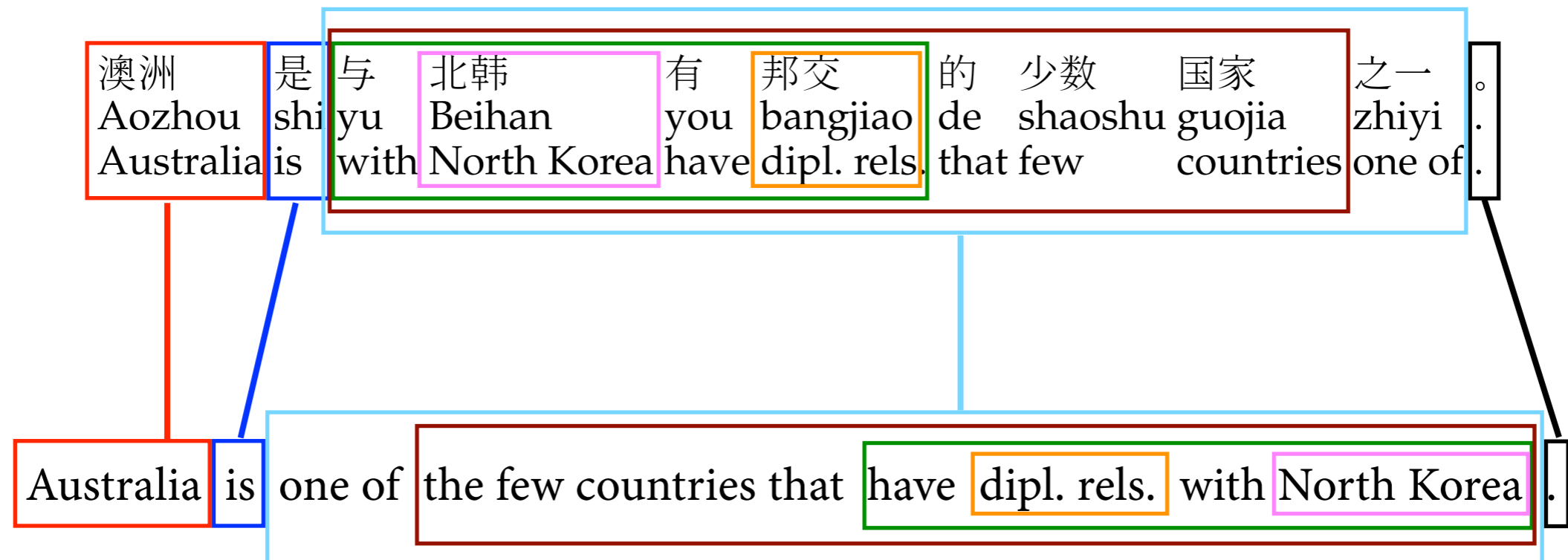
Chinese-English reordering

(output of phrase-based system ATS)



“Australia is one of the few countries that have diplomatic relations with North Korea.”

Syntax-based reordering



⟨yu [1] you [2], have [2] with [1]⟩

⟨[1] de [2], the [2] that [1]⟩

⟨[1] zhiyi, one of [1]⟩

“Australia is one of the few countries that have diplomatic relations with North Korea.”

Syntax-based translation



- Idea: Learn *synchronous* syntax rules that capture syntactic reordering between the two languages.
- Then much less unsystematic reordering necessary.
- We need to figure out:
 - ▶ how to represent translation rules
 - ▶ how to extract translation rules from data
 - ▶ how to define probability model (skipped here)
 - ▶ how to do decoding

Synchronous CFG

$S \rightarrow X^{(1)} / X^{(1)}$

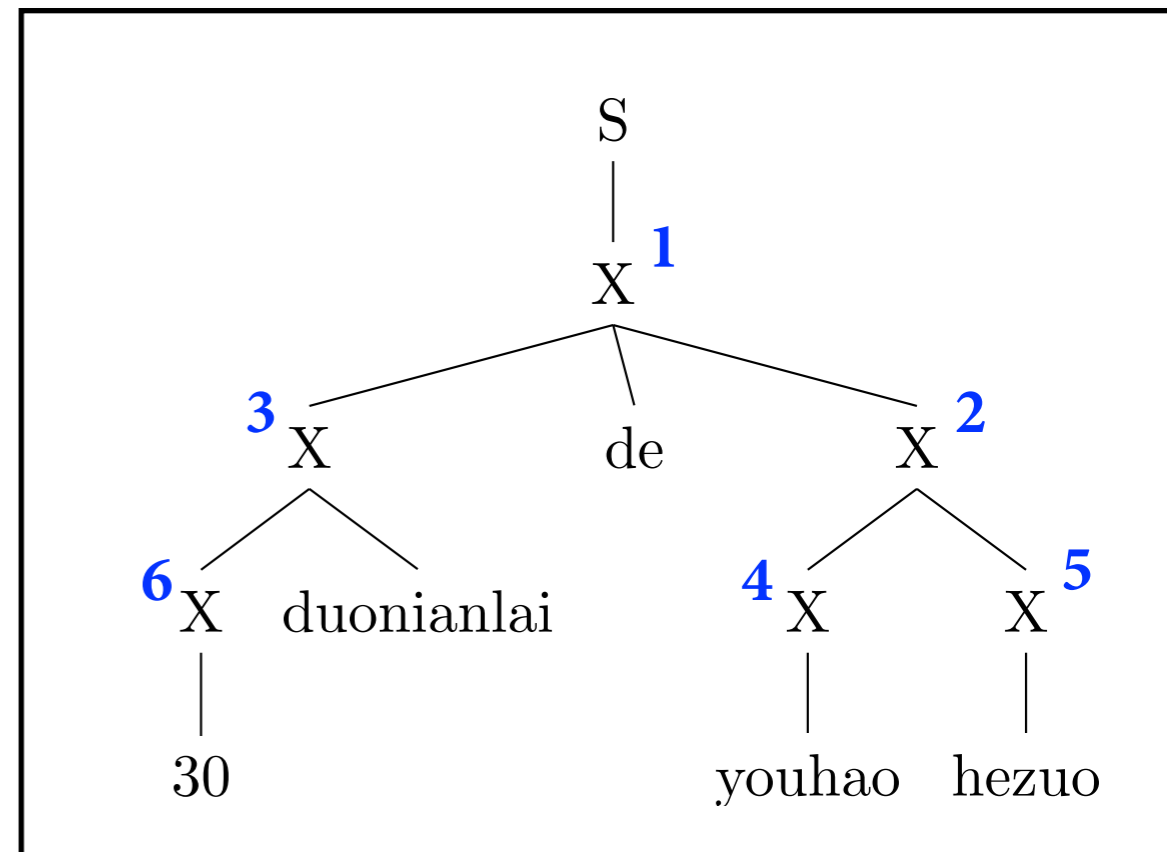
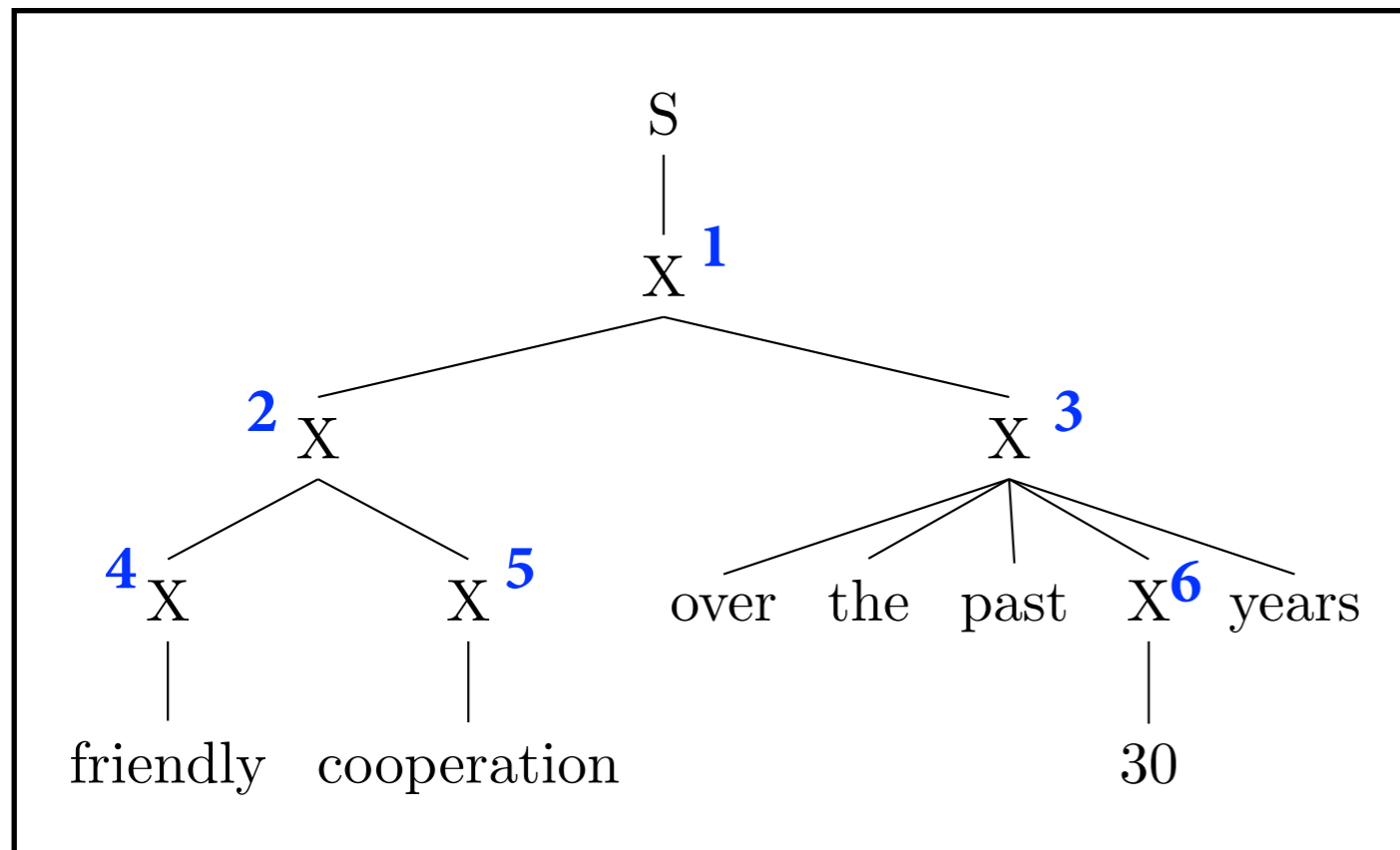
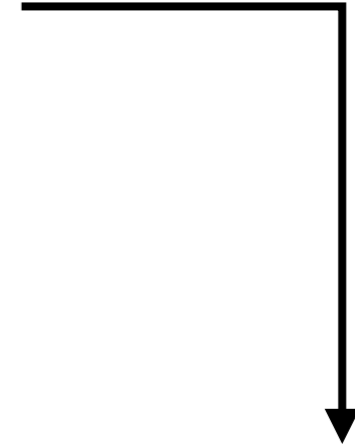
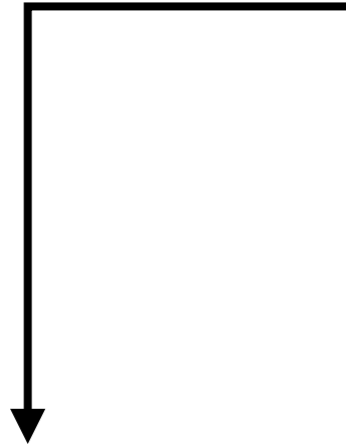
$X \rightarrow X^{(1)} \text{ de } X^{(2)} / X^{(2)} X^{(1)}$

$X \rightarrow X^{(1)} X^{(2)} / X^{(1)} X^{(2)}$

$X \rightarrow X^{(1)} \text{ duonianlai} / \text{over the last } X^{(1)} \text{ years}$

$X \rightarrow \text{yuohao} / \text{friendly}$

$X \rightarrow 30 / 30$



SCFG rule extraction

	friendly	cooperation	over	the	last	30	years
30						30	
duonianlai			over	the	last		years
de							
youhao	friendly						
hezuo		cooperation					

$X \rightarrow \text{youhao} / \text{friendly}$

$X \rightarrow 30 \text{ duonianlai} / \text{over the last 30 years}$

$X \rightarrow 30 / 30$

$X \rightarrow X^{(1)} \text{ duonianlai} / \text{over the last } X^{(1)} \text{ years}$

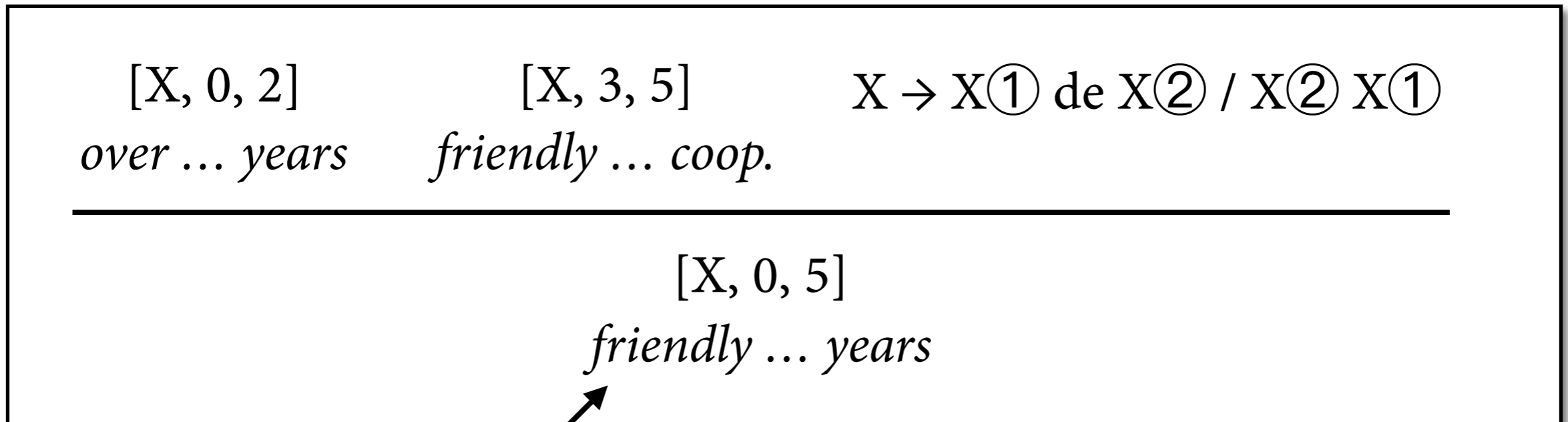
$X \rightarrow X^{(1)} X^{(2)} / X^{(1)} X^{(2)}$

$X \rightarrow X^{(1)} \text{ de } X^{(2)} / X^{(2)} X^{(1)}$

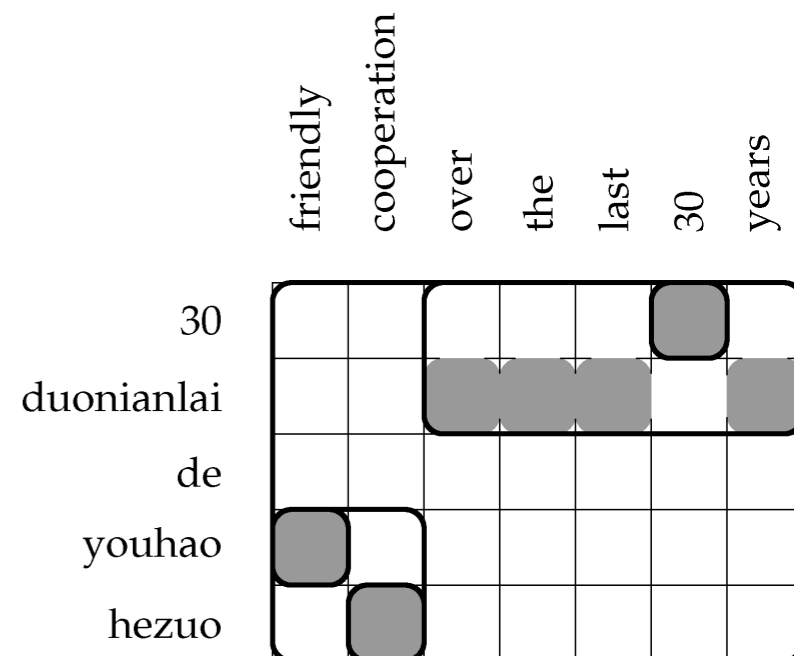
- Extract all phrase pairs as usual.
- Generate more rules by replacing sub-phrases by nonterminal X.
- Add “glue rules” $S \rightarrow S^{(1)} X^{(2)} / S^{(1)} X^{(2)}$ and $S \rightarrow X^{(1)} / X^{(1)}$ to start derivations.

Decoding schema

$f = \text{“30 duonianlai de youhao hezuo”}$



$\text{prob} = p_1 * p_2 * P(\text{rule}) * P_{\text{LM}}(\text{over} \mid \text{coop.})$



Pruning

- Problem: number of items blown up by factor of $|V|^{2m-2}$ for an m -gram language model.
- Need to use beam search: for each $[X, i, k]$ for Foreign positions i, k , keep only the best analyses.
- *Cube pruning*: improve runtime further by filling chart cell for $[A, i, k]$ from stream of rules $A \rightarrow B C$ and streams of items for cells $[B, i, j]$ and $[C, j, k]$ using n -best algorithm.

Cube Pruning

		[X, 6, 8; the scheme]	[X, 6, 8; the plan]	[X, 6, 8; the project]
		1	4	7
$X \rightarrow \langle \text{cong } X_{\square}, \text{from } X_{\square} \rangle$	1	2.1	5.1	
$X \rightarrow \langle \text{cong } X_{\square}, \text{from the } X_{\square} \rangle$	2	5.5		
$X \rightarrow \langle \text{cong } X_{\square}, \text{since } X_{\square} \rangle$	6			
$X \rightarrow \langle \text{cong } X_{\square}, \text{through } X_{\square} \rangle$	10			

		[X, 6, 8; the scheme]	[X, 6, 8; the plan]	[X, 6, 8; the project]
		1	4	7
	1	2.1	5.1	8.2
	2	5.5	8.5	
	6			
	10			

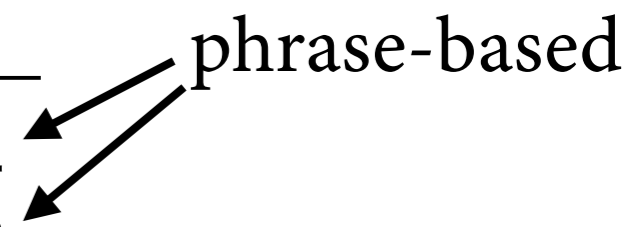
		[X, 6, 8; the scheme]	[X, 6, 8; the plan]	[X, 6, 8; the project]
		1	4	7
	1	2.1	5.1	8.2
	2	5.5	8.5	
	6	7.7		
	10			

Items for Chinese [X, 5, 8]

BLEU Comparison

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

phrase-based



Results: BLEU and Speed

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

time in seconds per sentence

Conclusion

- Noisy channel translation: combine translation model with language model.
- Phrase-based translation: Extract phrases (= arbitrary substrings) from word alignments.
 - ▶ different reordering models, e.g. with SCFGs
- Decoding algorithms must deal with huge search space. Need to do some clever form of beam search.
- Much current research uses neural networks instead.