A Hierarchical Phrase-Based Model for Statistical Machine Translation

David Chiang
University of Maryland
Institute for Advanced Computer Studies

In a nutshell

- Hiero: a new statistical translation model
- Significantly improves on the phrase-based model it generalizes (Pharaoh: Koehn et al)
- Synchronous context-free grammar allows more complex structural mappings
- Learnable without syntactic information

Example

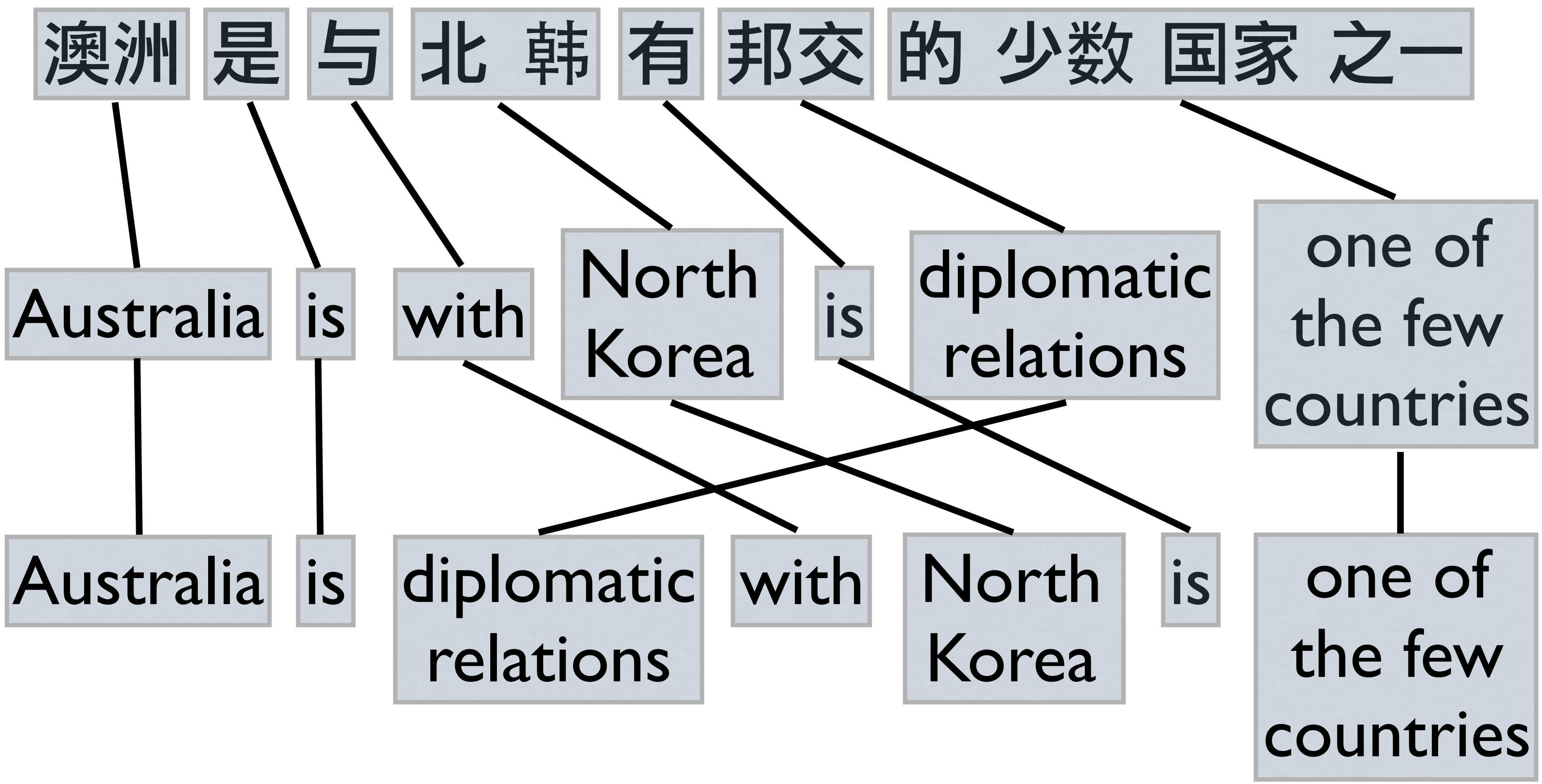
澳洲 是 与 北 韩 有 邦交 的 少数 国家 之一

Australia is with North Korea have diplomatic relations that few countries one of

Human translation:

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

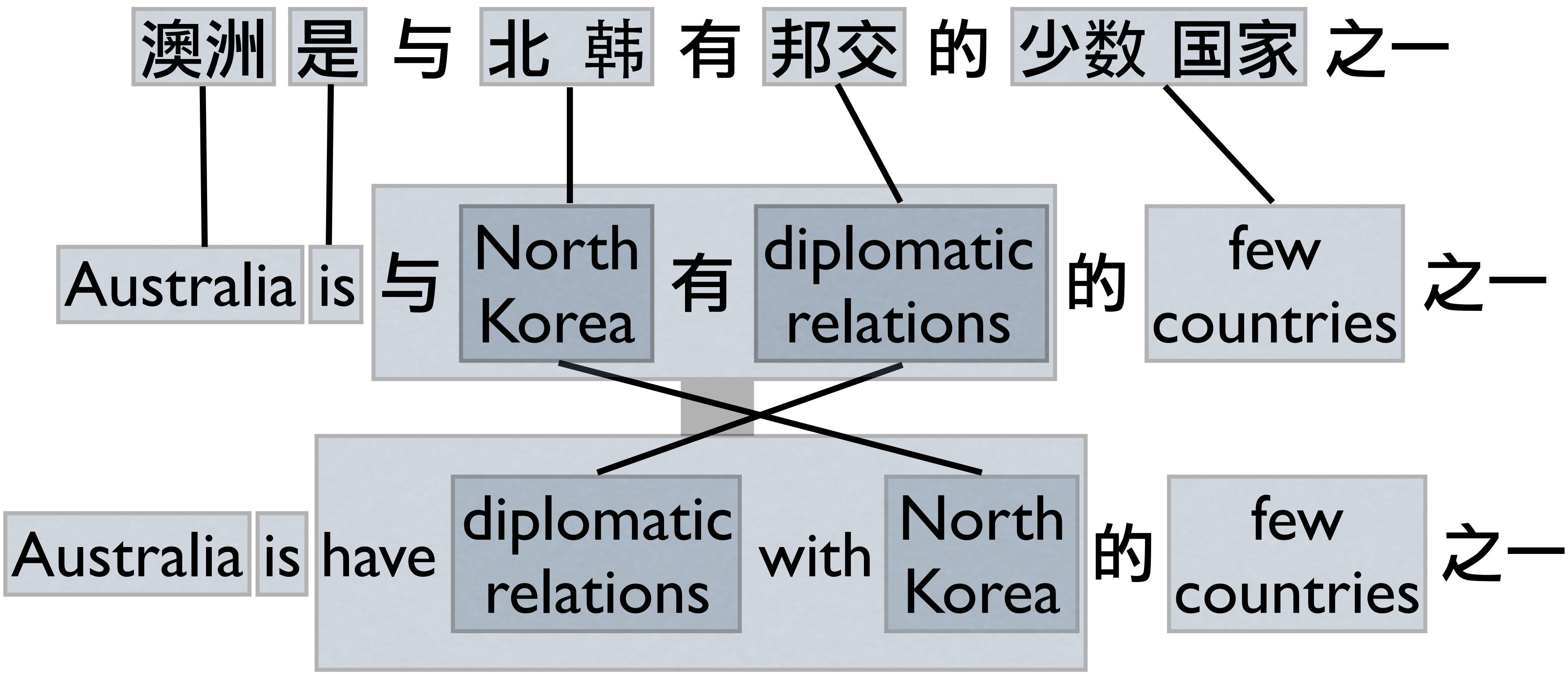
Phrase-based translation



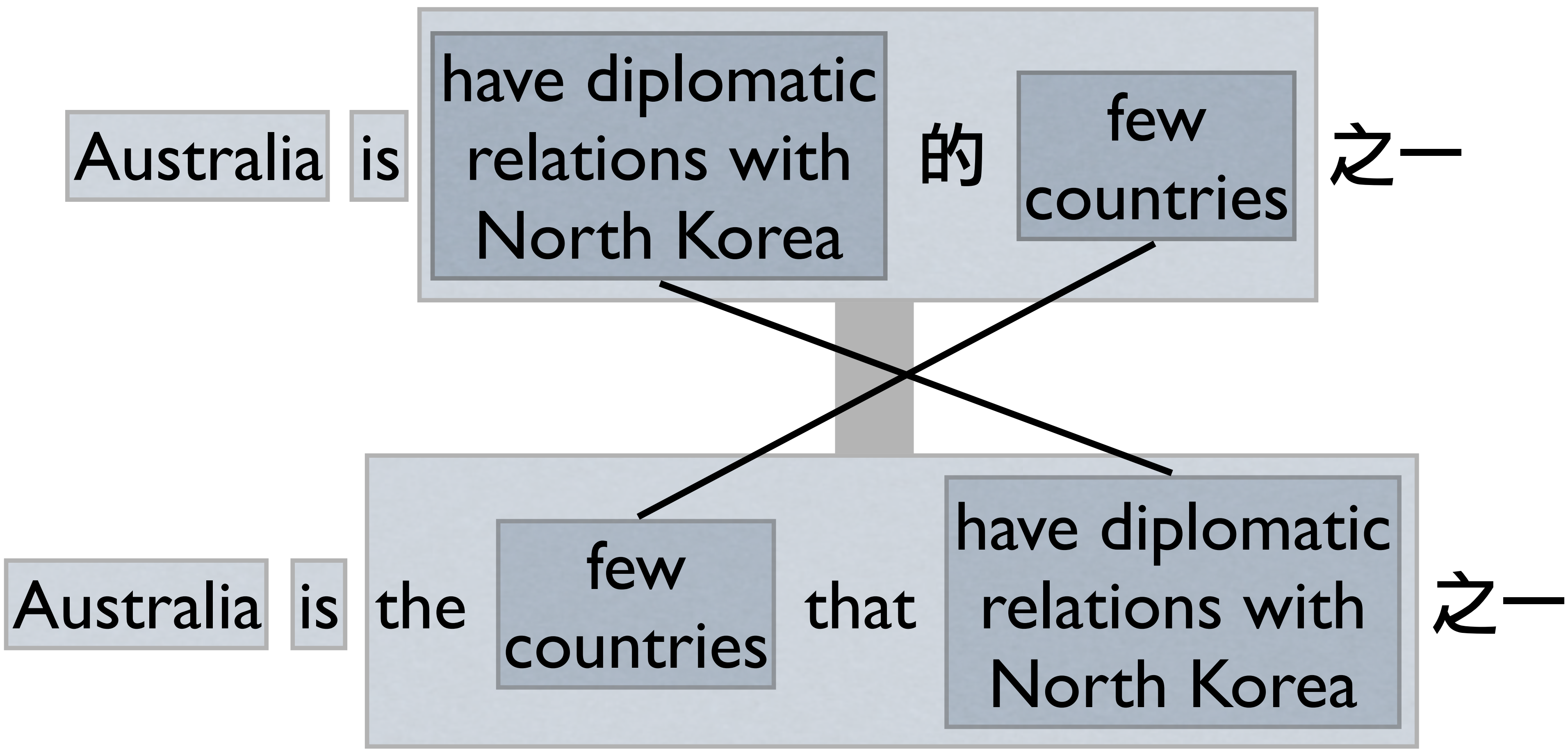
Phrase-based translation

- Phrase-based systems learn reorderings within phrases well
- Reordering of phrases, not so well: classically, no lexical sensitivity
- Why not use phrases to reorder phrases?

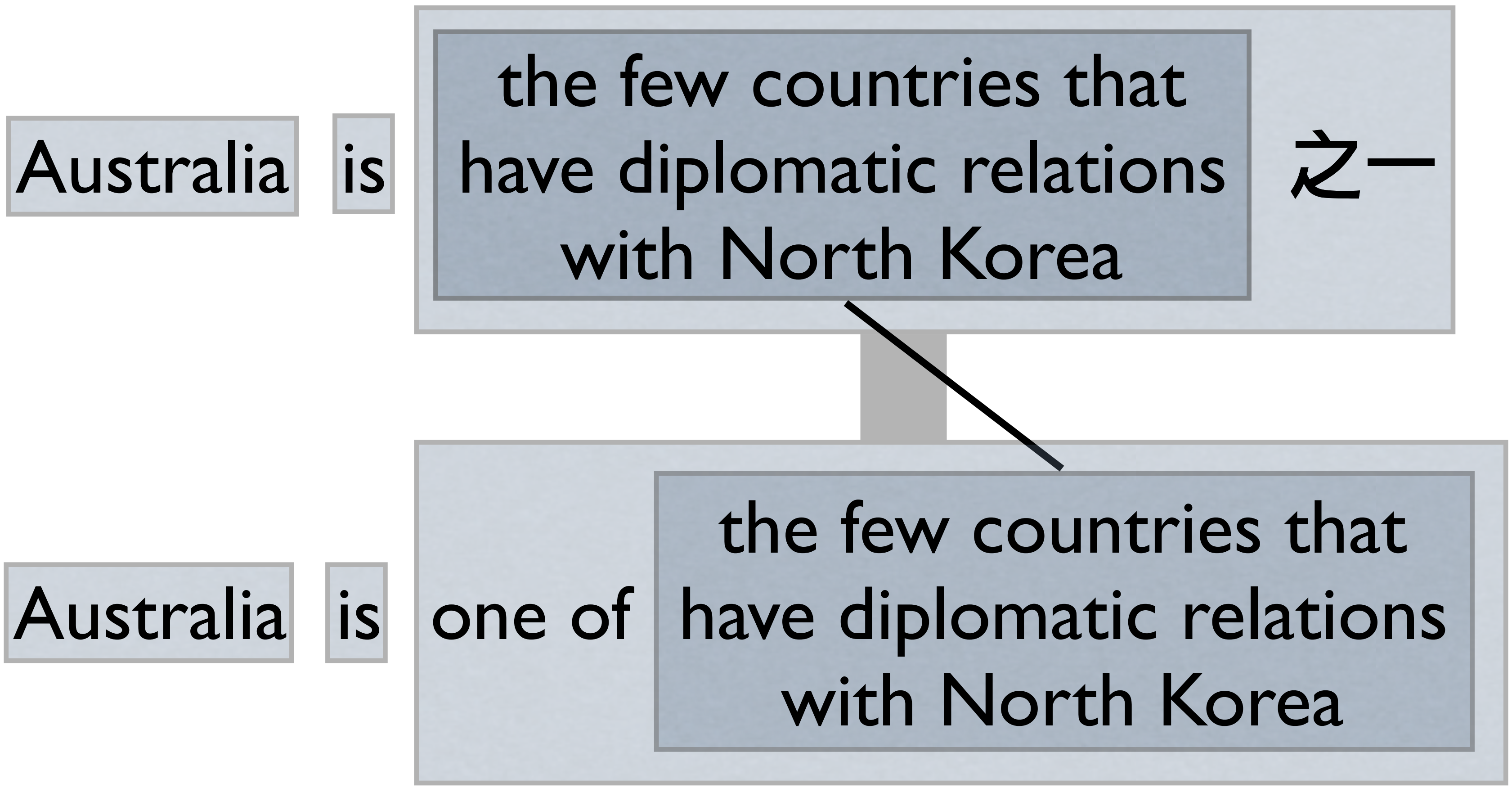
Hierarchical phrases



Hierarchical phrases



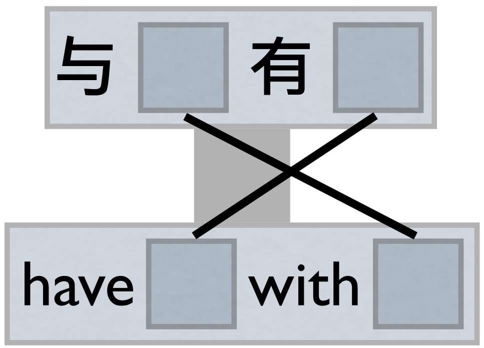
Hierarchical phrases



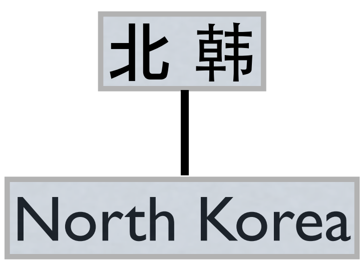
A new approach

- Formalize as productions of a synchronous CFG (aka syntax-directed translation schema, inversion transduction grammar)
- Learned without syntactic information (like Wu, Alshawi et al; unlike Yamada and Knight)
- Heavily lexicalized, as in phrase-based models

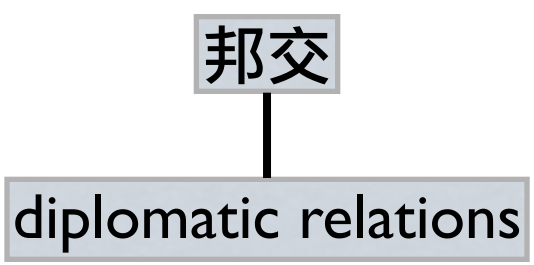
Synchronous CFG



(X → 与 X_1 有 X_2, X → have X_2 with X_1)

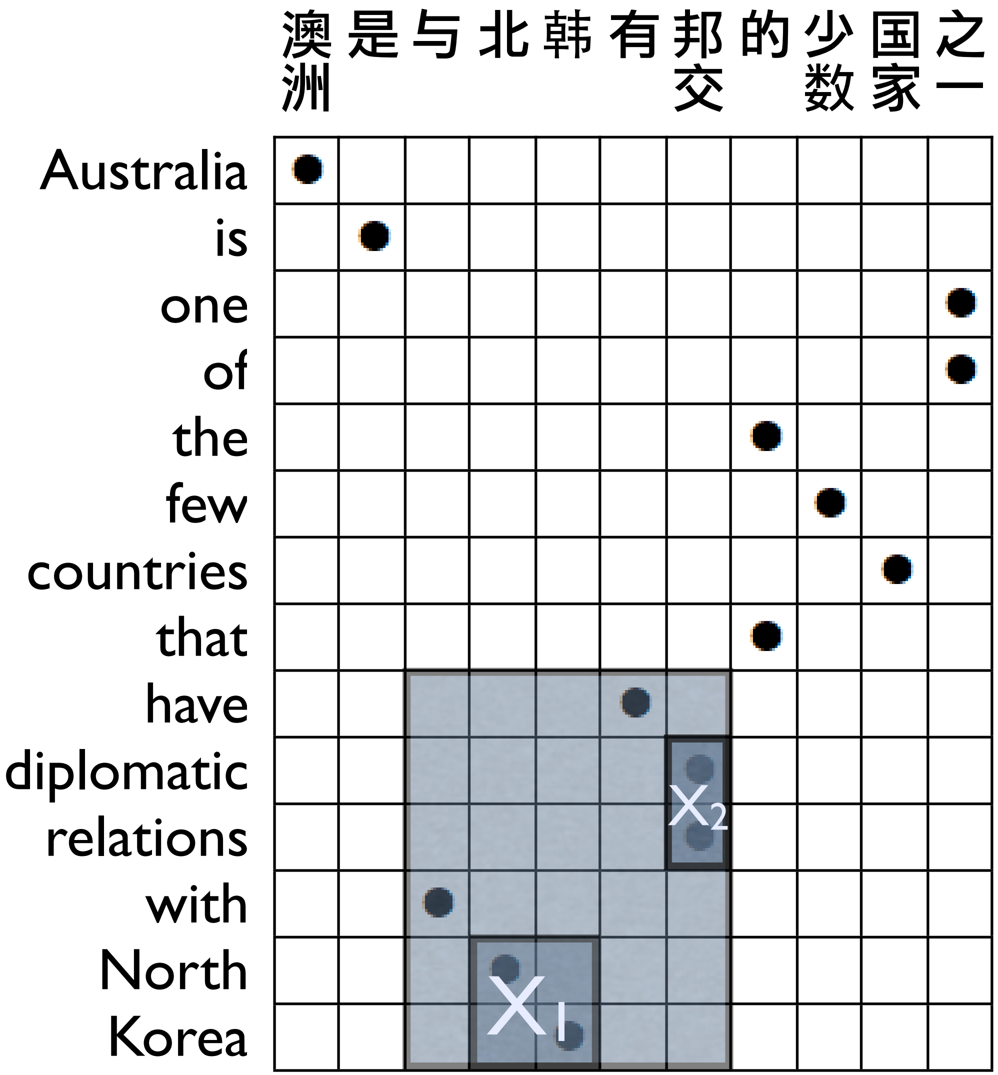


(X → 北 韩, X → North Korea)



(X → 邦交, X → diplomatic relations)

Grammar extraction



(与 北 韩 有 邦交, have diplomatic relations with North Korea)

(邦交, diplomatic relations)

(北 韩, North Korea)

($X \rightarrow$ 与 X_1 有 X_2, $X \rightarrow$ have X_2 with X_1)

Example rules

$X \rightarrow$ 的	$X \rightarrow$ ’s
$X \rightarrow X_1$ 的 X_2	$X \rightarrow$ the X_2 of X_1
$X \rightarrow X_1$ 的 X_2	$X \rightarrow$ the X_2 that X_1
$X \rightarrow$ 在	$X \rightarrow$ in
$X \rightarrow$ 在 X_1 下	$X \rightarrow$ under X_1
$X \rightarrow$ 在 X_1 前	$X \rightarrow$ before X_1
$X \rightarrow$ 今年 X_1	$X \rightarrow X_1$ this year
$X \rightarrow X_1$ 之一	$X \rightarrow$ one of X_1
$X \rightarrow X_1$ 总统	$X \rightarrow$ president X_1

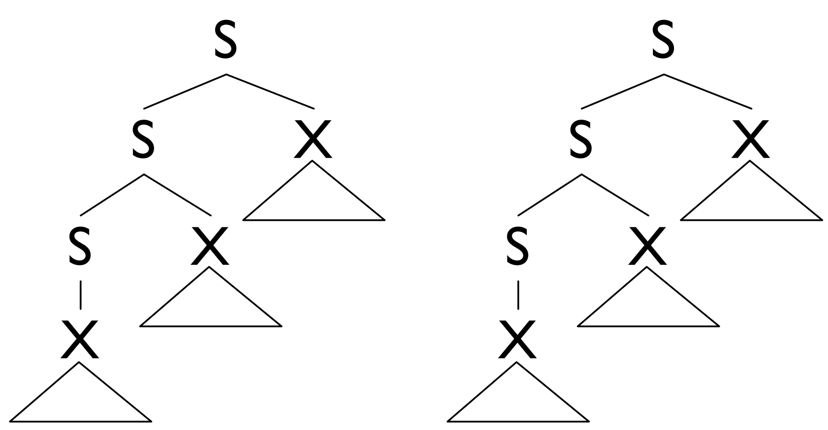
Glue rules

- Plus "glue" rules:

$$(S \rightarrow S_1 X_2, S \rightarrow S_1 X_2)$$

$$(S \rightarrow X_1, S \rightarrow X_1)$$

- Acts as fallback like phrase-based systems



Assumptions

- Per-sentence uniform distribution on initial phrases; per-phrase, on final rules
- Length limit on initial phrases (≤7–15) and final rules (≤5 term+nonterm)
- Limit to two nonterminals
- etc.

Probability model

Combine multiple features into a log-linear model (Och and Ney, 2002)

$$P(D) \propto \prod_{r \in D} \prod_{i} v_i(r)^{\lambda_i(r)}$$

D	derivation
r	rules in D
v_i	feature functions
λ_i	feature weights

Weights λ_i learned by maximum-BLEU training (Och 2003; Koehn implementation)

Model features

- Phrase translation:

$$p(X \rightarrow X_1 \text{之一} \mid X \rightarrow \text{one of } X_1)$$

$$p(X \rightarrow \text{one of } X_1 \mid X \rightarrow X_1 \text{之一})$$

- Lexical weighting (Koehn):

$$\tfrac{1}{2}[p(\text{之一} \mid \text{one}) + p(\text{之一} \mid \text{of})]$$

$$p(\text{one} \mid \text{之一}) \times p(\text{of} \mid \text{之一})$$

Rule scoring

- Problem: rules not actually observed
- Stipulate:
  - All the initial phrases extracted from a sentence get equal weight (Och)
  - All the rules extracted from an initial phrase get equal weight
- Then relative-frequency estimation

More model features

- But glue rule S $\rightarrow$ SX has dedicated feature
- Trigram language model:

  p(Australia | <*S*>) × p(is | <*S*> Australia) $\cdots$
- Number of English words
- Number of non-glue rules

Decoding

$_0$澳洲$_1$是$_2$与$_3$北$_4$韩$_5$有$_6$邦交$_7$的$_8$少数$_9$国家$_{10}$之一$_{11}$

- Parse Chinese side using CKY-like algorithm
- Thought of as deductive inference:

$$\frac{[X,3,5] \quad [X,6,7]}{[X,2,7]} \quad \text{because } X \rightarrow \text{与 } X \text{ 有 } X$$

Integrating the LM

$_0$澳洲$_1$是$_2$与$_3$北$_4$韩$_5$有$_6$邦交$_7$的$_8$少数$_9$国家$_{10}$之一$_{11}$

- Store English translations in hypotheses to calculate *n*-gram probabilities online:

$$\frac{[X,3,5,\text{North Korea}] \quad [X,6,7,\text{diplomatic relations}]}{[X,2,7,\text{have diplomatic}\ldots\text{North Korea}]}$$

- Elide all but outermost $n-1$ English words

Optimizations

- Prune search space to improve efficiency
  - For all (X, i, j), throw out hypotheses with score β worse than the best, or not in the top b
  - Extra optimization to reduce blowup due to language model
- Limit hierarchical phrase length (≤10 or 15)

Experiment: setup

- Training: FBIS (about 7M+9M words, Chinese-English newswire)
- Language model: 200M words English
- Max-BLEU training: MT Eval 2002
- Test: MT Eval 2003 (also newswire)
- Baseline: Pharaoh (Koehn et al), 2004 version, same training and features

Experiment: results

- MT Eval 2003 Chinese-English, case-*in*sensitive BLEU-4:

Pharaoh-2004	26.76
Hiero	28.77

- 7.5% relative improvement, statistically significant (bootstrapping)

Discussion

- Try adding a constituent-reward feature:

without	28.77
with	28.81

- Insignificant improvement
- Gets healthy weight, same as phrase penalty
- Bracket precision: 47% without, 76% with

Discussion

- Glue rule gets higher weight than any other rule
- Types of phrases used:

Glue	16%
Hierarchical	49%
Ordinary	35%

Conclusions

- Hiero’s structural mappings result in better performance than many phrase-based systems
- Learned from parallel text without syntactic information
- Future work: improve efficiency, induction of syntactic information