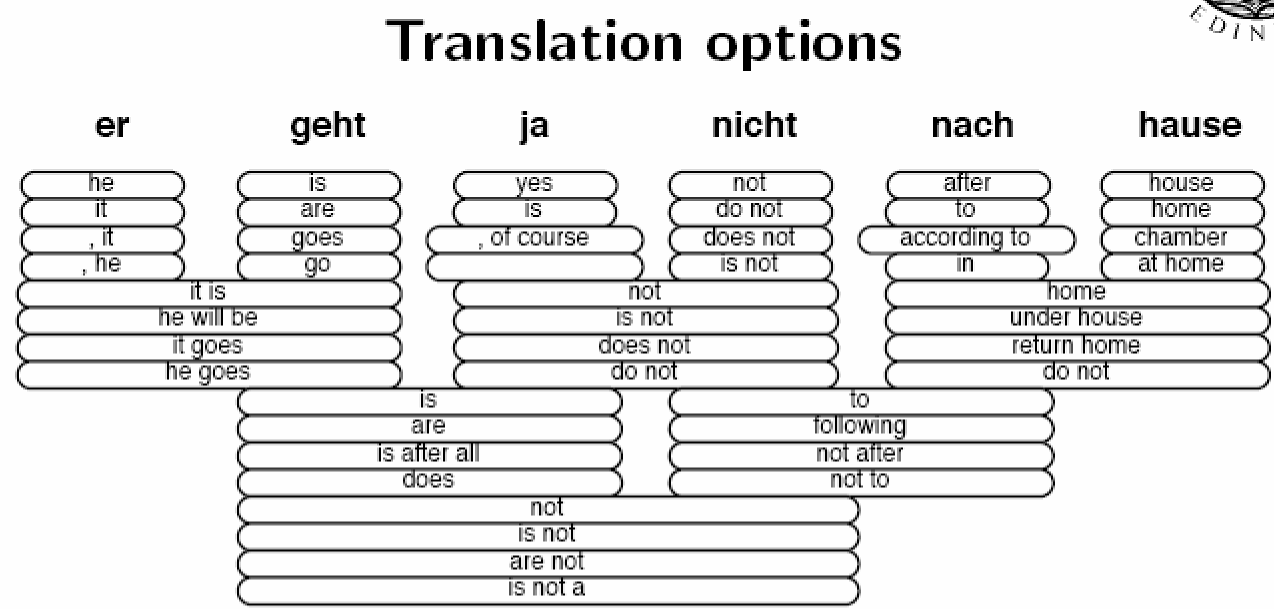
Developments in Hierarchical Phrase-based Translation

Philip Resnik
University of Maryland

Work done with
David Chiang, Chris Dyer, Nitin Madnani, and Adam Lopez

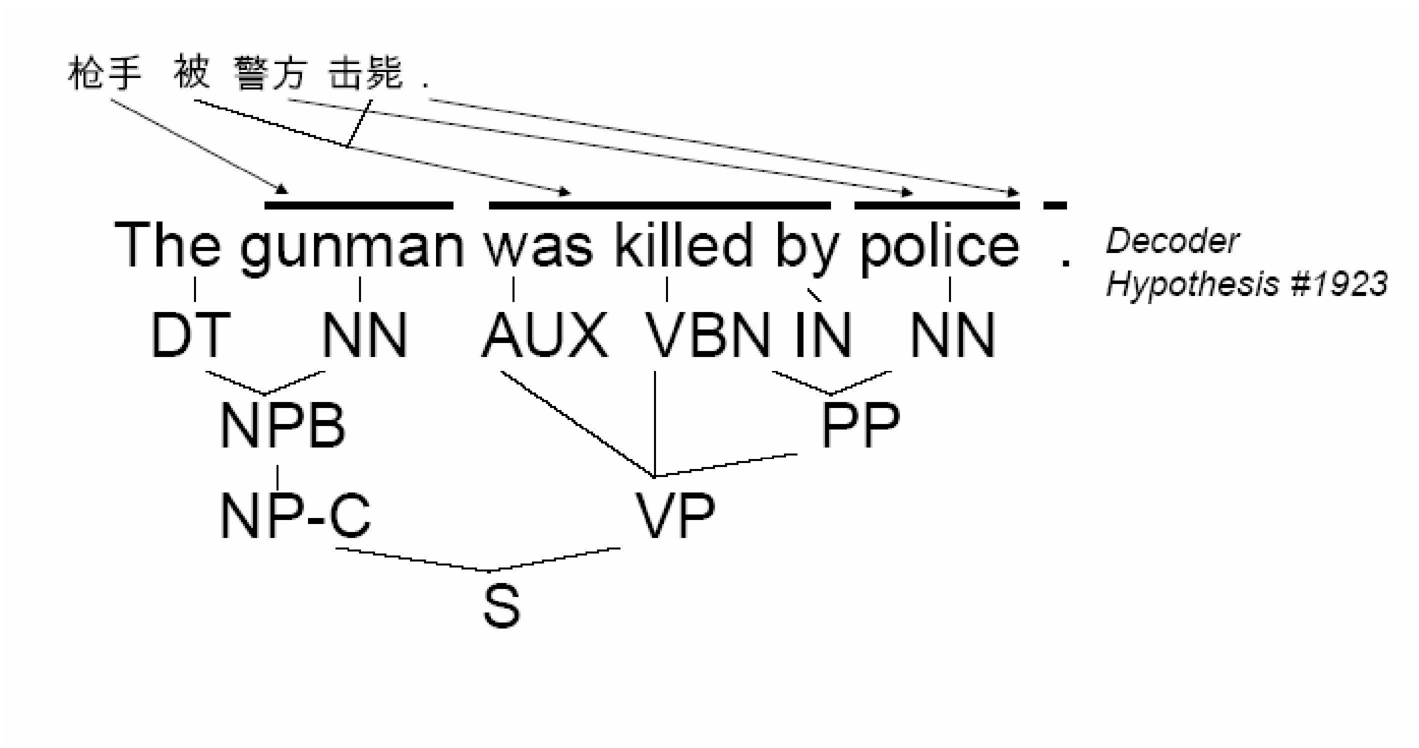
Some things you've seen recently…

Shamelessly stolen from Philipp Koehn

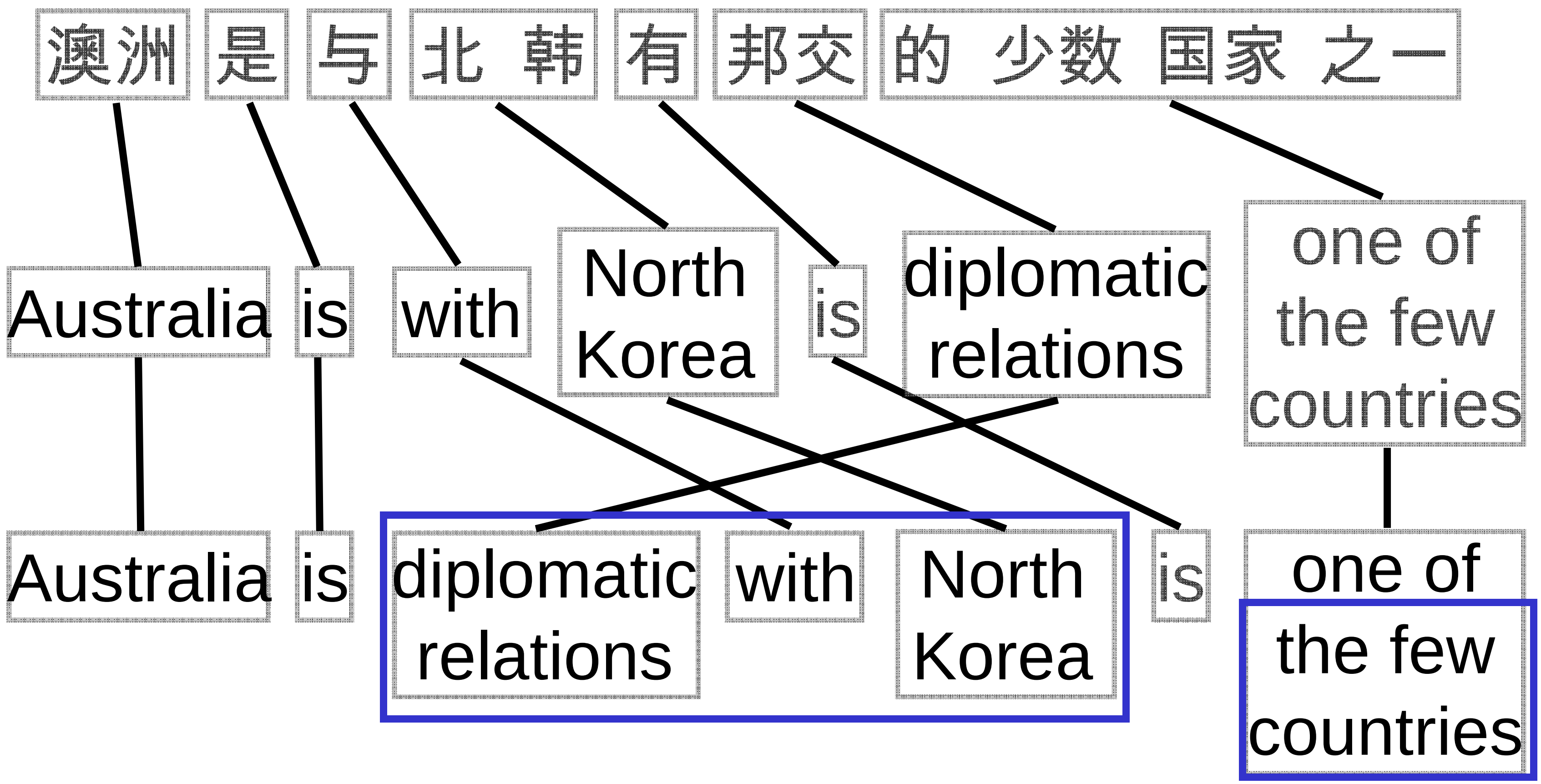


Some things you've seen recently…

Shamelessly stolen from Kevin Knight

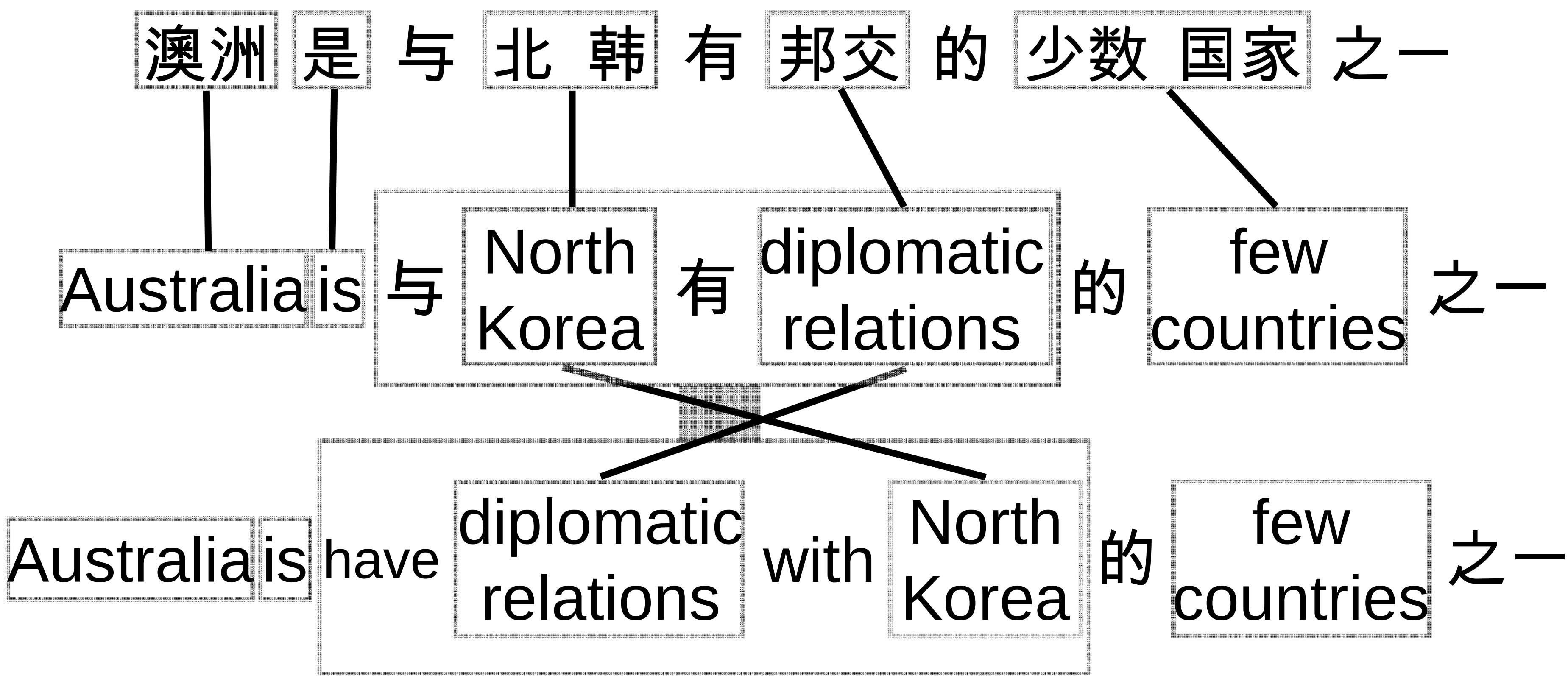


Flat Phrases

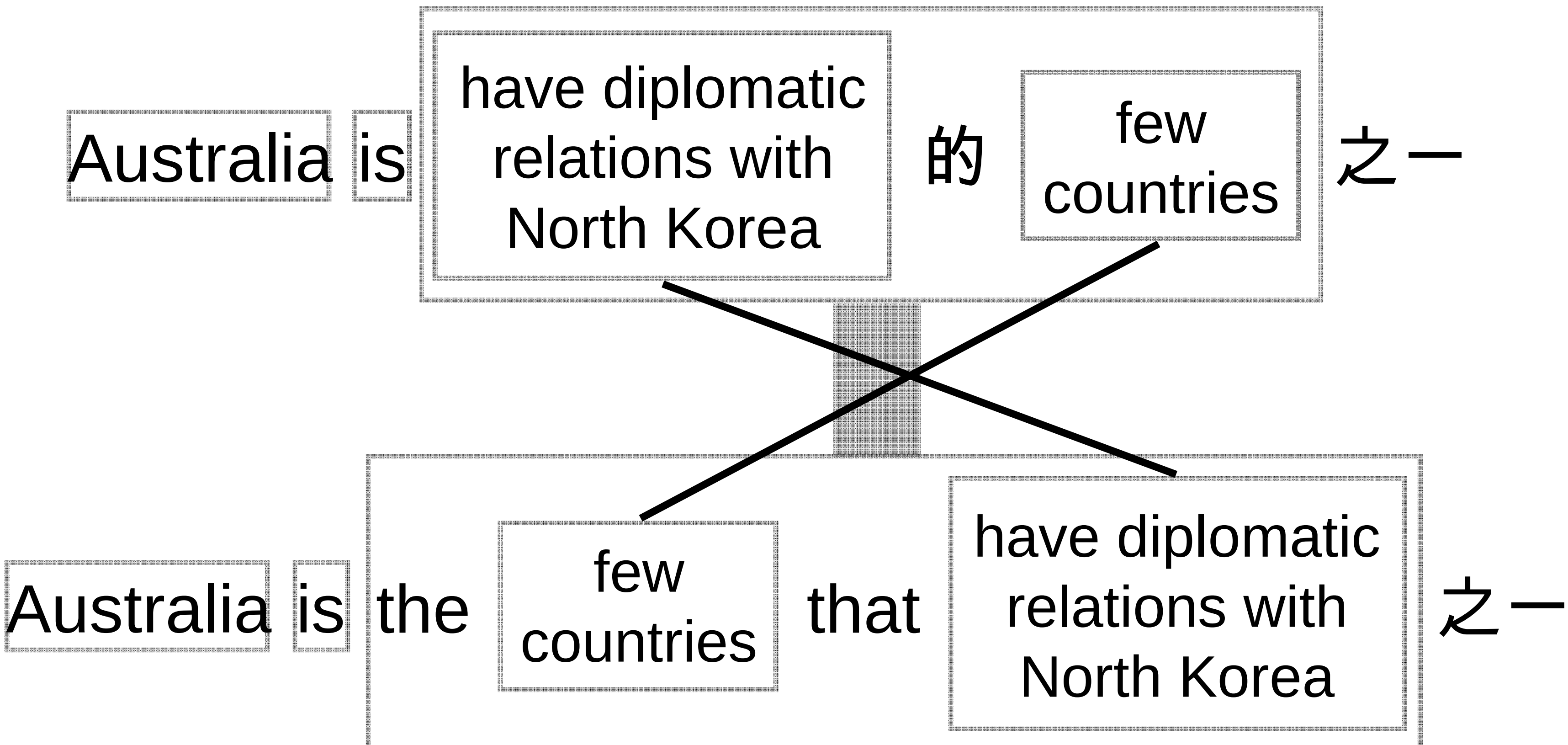


Can we capture this modification relationship without ISI-style syntactic modeling?

Hierarchical phrases



Hierarchical phrases

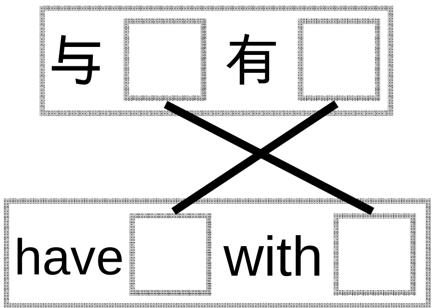


Hierarchical phrases

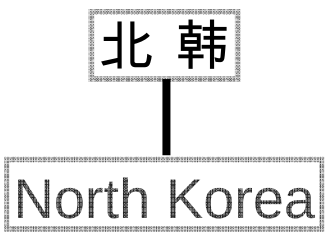
Australia is the few countries that have diplomatic relations with North Korea 之一

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

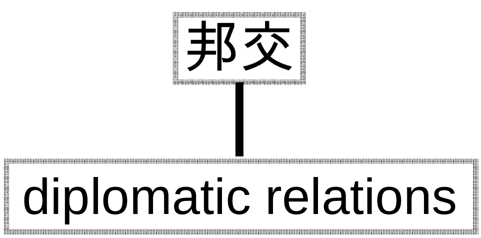
Synchronous CFG



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

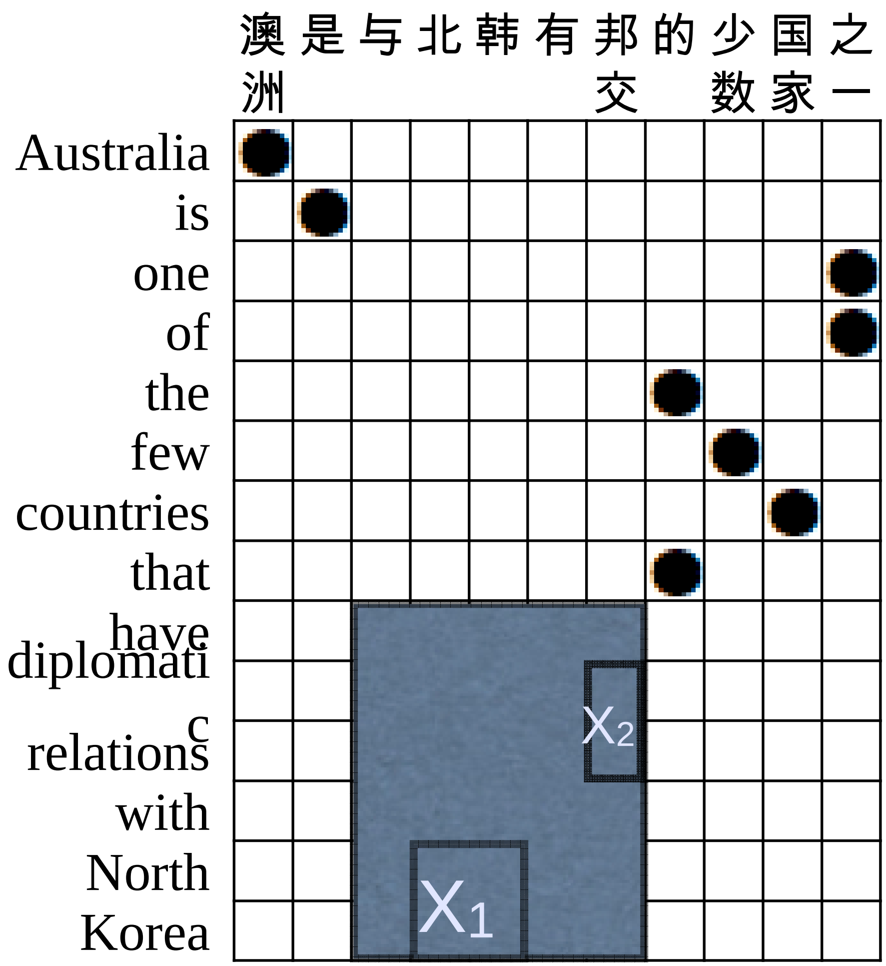


($X \rightarrow$ 北 韩, $X \rightarrow$ North Korea)



($X \rightarrow$ 邦交, $X \rightarrow$ diplomatic relations)

Grammar extraction



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

(邦交, diplomatic relations)

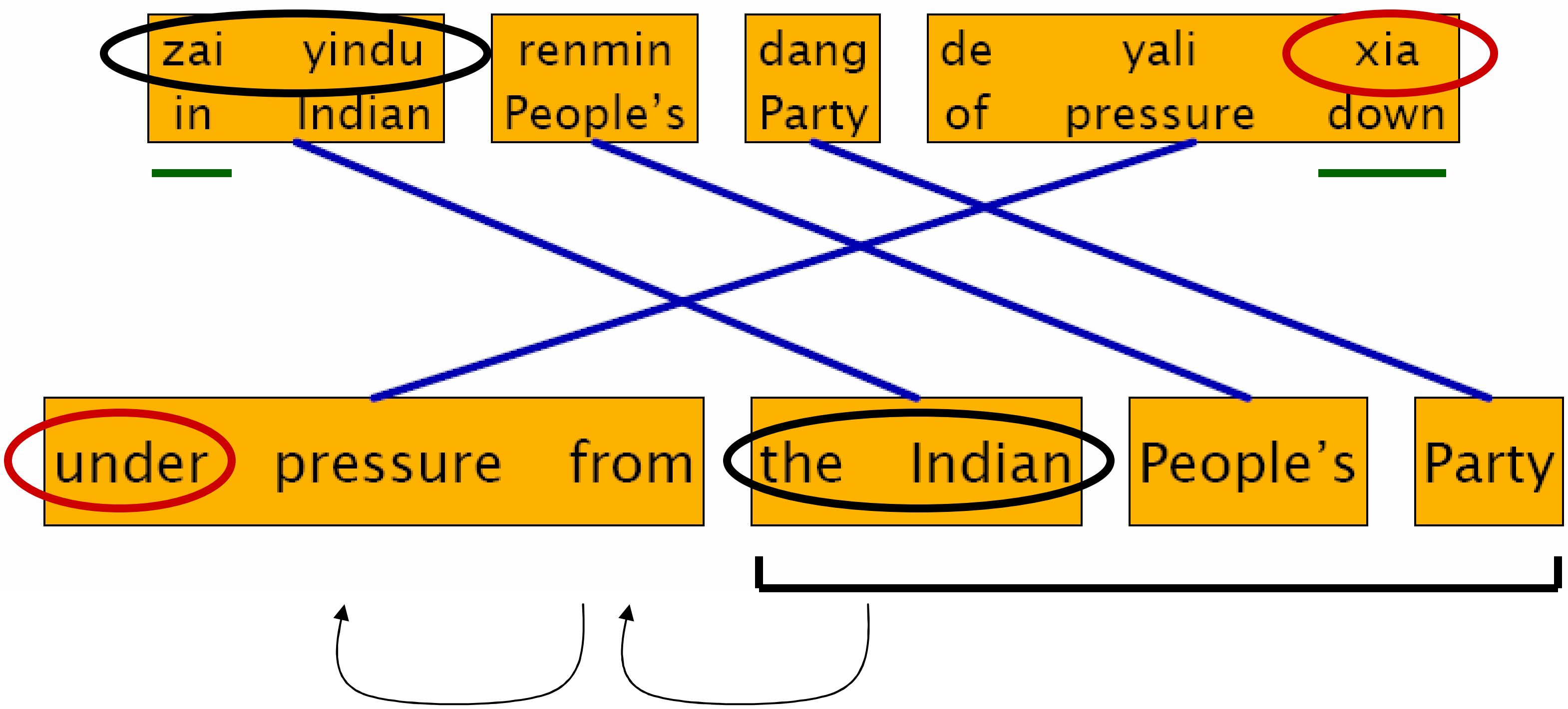
(北 韩, North Korea)

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

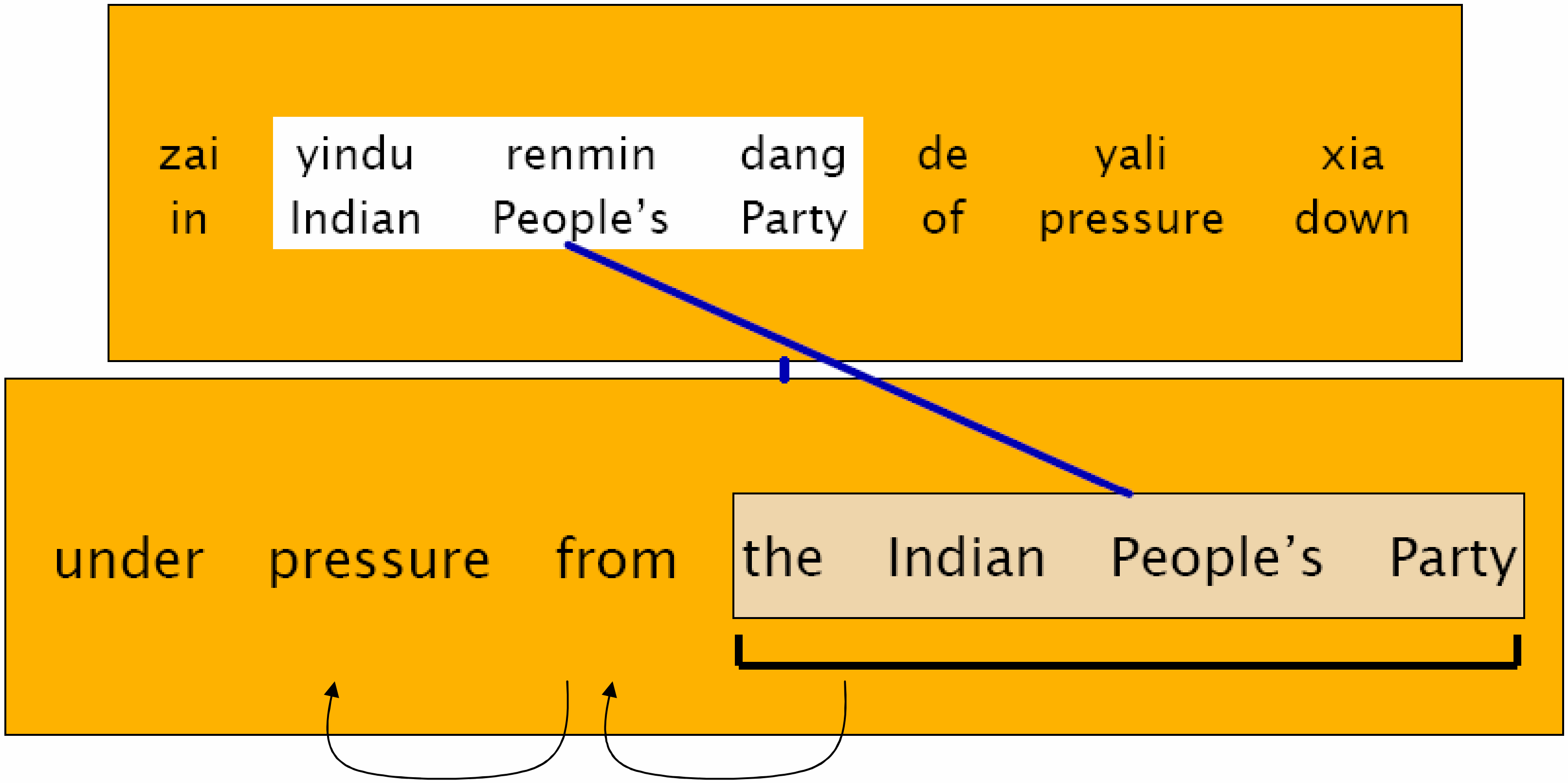
Rank	Chinese	English
1	,	,
2	.	.
3	"	"
4	de	the
5	,	and
1710	X zongtong	president X
2097	$X_{\boxed{1}}$ de $X_{\boxed{2}}$	the $X_{\boxed{2}}$ of $X_{\boxed{1}}$
2850	jingnian X	X this year
10781	zai X xia	under X
32738	zai X nei	within X
218421	X de yali	pressure from X
300091	zai X yali xia	under pressure from X

Permits dependencies over long distances without memorizing intervening material (sparseness!)

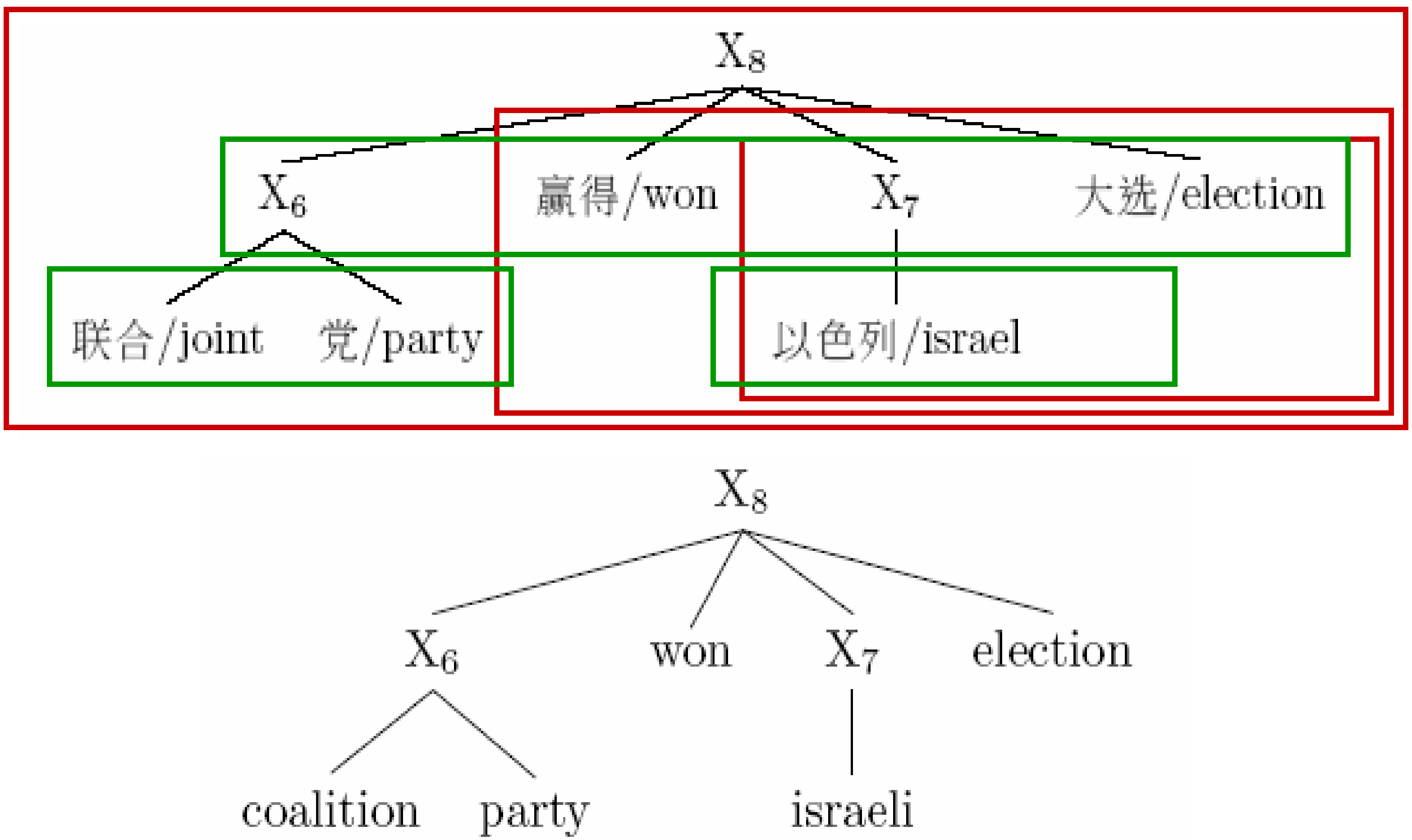
Non-Hierarchical Phrases



Hierarchical Modeling



Structures Useful for MT



Hiero: Hierarchical Phrase-Based Translation

- Introduced by Chiang (2005, 2007)
- Moves from phrase-based models toward syntax
  - Phrase table → Synchronous CFG
    - Learn reordering rules together with phrases

      X → < 与 X1 有 X2, have X2 with X1 >

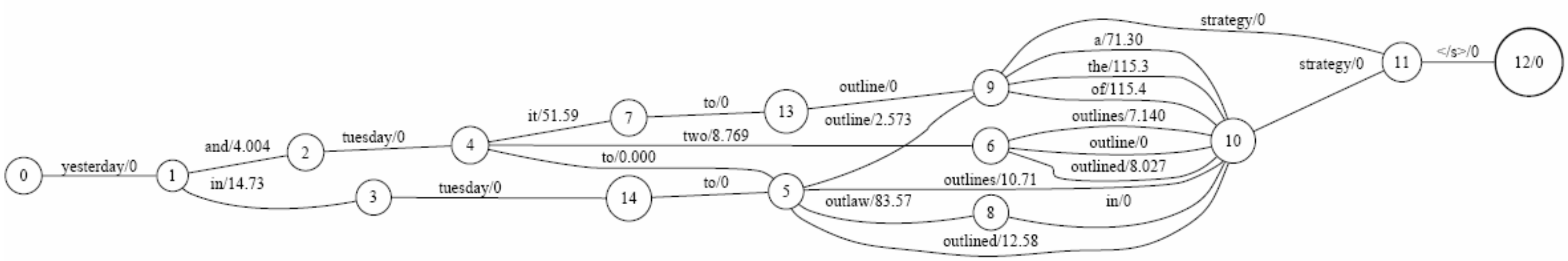
      X → < 北 韩, North Korea>
  - Decoder → Parser
    - CKY parser
    - Target side of grammar intersected with finite state LM
    - Log-linear model tuned to optimize objective (BLEU, TER, …)

Roadmap

- Brief review of Hiero
- New developments
  - Confusion network decoding (Dyer)
  - Suffix arrays for richer features (Lopez)
  - Paraphrase to improve parameter tuning (Madnani)
- Summary and conclusions

Confusion Network Decoding for Translating ASR Output

- ASR systems produce word graphs:

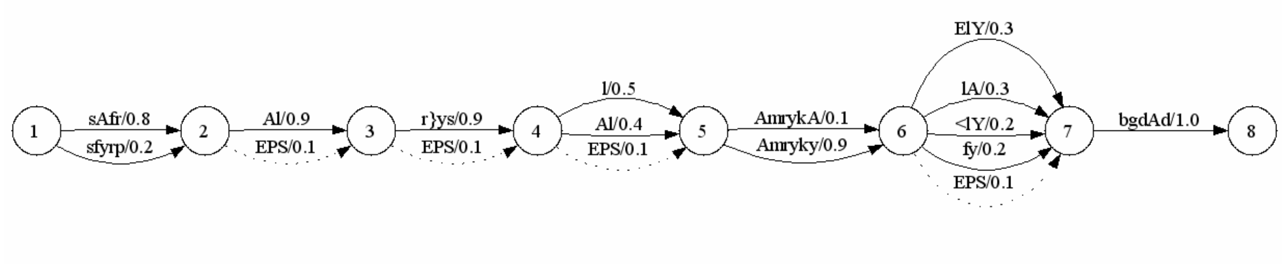


- Equivalent to weighted FSA

- However, Hiero assumes 1-best input

Confusion networks (a.k.a. pinched lattices, meshes, sausages)

- Approximation of a word lattice (Mangu, et al., 2000)
  - Every path through the network hits every node
  - Probability distribution over words at a given position



  - Special symbol ε (epsilon) represents a skip.

Translating from Confusion Networks

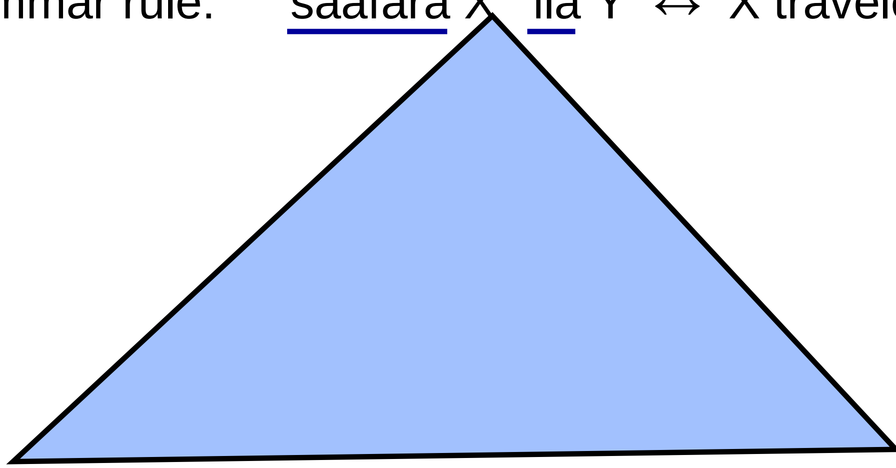
- Confusion networks for MT
  - Many more paths than in the source lattice
  - Nice properties for dynamic programming

- Decoding confusion networks beats 1-best hypothesis with a phrase-based model
  - Bertoldi, et al. 2005
- Decoding confusion networks is highly efficient with a phrase-based model
  - Hopkins Summer Workshop
    - Moses decoder accepts input as a confusion network
  - Bertoldi, et al. 2007

The value of hierarchy in the face of ambiguity

Input:	saafara	al-ra’iisu	^{c}ala ‘ila	Baghdad

Grammar rule: saafara X ‘ila Y ↔ X traveled to Y



al-rajulu al-manfiyu allathiy laa yuħibbu al-ṭayaraana

Parsing Confusion Networks

- Efficient CKY parsing available
  - Insight: except for the initialization pass (processing terminal symbols), standard CKY already operates on “confusion networks”.

Parsing Confusion Networks

	Text	Confusion Networks
• Axioms:	$\dfrac{}{[X \rightarrow \bullet\gamma, i, i] : w} \quad (X \xrightarrow{w} \langle \gamma,\, \alpha \rangle) \in G$	$\dfrac{}{[X \rightarrow \bullet\gamma, i, i] : w} \quad (X \xrightarrow{w} \langle \gamma,\, \alpha \rangle) \in G$
• Inferences:	$\dfrac{[X \rightarrow \alpha \bullet f_{j+1}\beta, i, j] : w}{[X \rightarrow \alpha f_{j+1} \bullet \beta, i, j+1] : w}$ $\dfrac{[Z \rightarrow \alpha \bullet X\beta, i, k] : w_1 \quad [X \rightarrow \gamma\bullet, k, j] : w_2}{[Z \rightarrow \alpha X \bullet \beta, i, j] : w_1 \times w_2}$	$\dfrac{[X \rightarrow \alpha \bullet \mathbf{F}_{j+1,k}\beta, i, j] : w}{[X \rightarrow \alpha\mathbf{F}_{j+1,k} \bullet \beta, i, j+1] : w \times \mathbf{p}_{j+1,k}}$ $\dfrac{[X \rightarrow \alpha \bullet \beta, i, j] : w}{[X \rightarrow \alpha \bullet \beta, i, j+1] : w \times \mathbf{p}_{j+1,k}} \quad \mathbf{F}_{j+1,k} = \epsilon$ $\dfrac{[Z \rightarrow \alpha \bullet X\beta, i, k] : w_1 \quad [X \rightarrow \gamma\bullet, k, j] : w_2}{[Z \rightarrow \alpha X \bullet \beta, i, j] : w_1 \times w_2}$
• Goal:	$[S \rightarrow \gamma\bullet, 0, n]$	$[S \rightarrow \gamma\bullet, 0, n]$

Model features

	Hierarchical	Non-hierarchical
	P_{LM}	P_{LM}
	$P(\gamma\|\alpha)$	$P(\bar{f}\|\bar{e})$
	$P(\alpha\|\gamma)$	$P(\bar{e}\|\bar{f})$
	$P_w(\gamma\|\alpha)$	$P_w(\bar{f}\|\bar{e})$
	$P_w(\alpha\|\gamma)$	$P_w(\bar{e}\|\bar{f})$
λ_{CN}	$P(I_k\|\mathcal{G})$	$P(\bar{f}\|\mathcal{G})$
	word penalty	word penalty
	1 non-terminal penalty	distortion
	2 non-terminal penalty	

Application: spoken language translation

- Experiments
  - Chinese – English (IWSLT 2006)
    - Small standard training bitext (<1M words)
    - Trigram LM from English side of bitext only
    - Spontaneous and read speech from the travel domain
    - Text only development data! ($\lambda_{CN}=\lambda_{LM}$)
  - Arabic – English (BNAT05)
    - UMD training bitext (6.7M words)
    - Trigram LM from bitext and portions of Gigaword
    - Broadcast news and broadcast conversations
    - ASR output development data. (λ_{CN} tuned by MERT)

Chinese-English (IWSLT 2006)

Input	WER	Hiero*	Moses*
verbatim	**0.0**	**19.63**	**18.40**
read, 1-best (CN)	**24.9**	**16.37**	**15.69**
read, full CN	**16.8**	**16.51**	**15.59**
spont., 1-best (CN)	**32.5**	**14.96**	**13.57**
spont., full CN	**23.1**	**15.61**	**14.26**

$p<0.05$

Noisier signal → more improvement

*** BLEU, 7 references**

Performance impact

- The impact on decoding time is minimal
  - Roughly the average depth of the confusion network
  - Similar to the impact in a phrase-based system
    - Moses: 3.8x slower over 1-best baseline
    - Hiero: 4.3x slower over 1-best baseline

- Both systems have efficient disk-based formats available to them
  - Adaptation of Zens & Ney (2007)

Arabic-English (BNAT05)

Input	WER	Hiero*	Moses*
Verbatim	**0.0**	**26.46**	**25.13**
1-best	**12.2**	**23.64**	**22.64**
Full CN	**7.5**	**24.58**	**22.61**

p<0.01

n.s.

p<0.05

Extremely low WER (audio was part of recognizer training data).

Hiero appears to make better use of ambiguity.

p<0.05

*** BLEU, 1 reference**

Another Application: Decoder-Guided Morphological Backoff

- Morphological complexity makes the sparse data problem even more acute

- Example: Czech → English
  - Hypothesis:

    From the US side of the Atlantic all such odůvodnění appears to be a totally bizarre.
  - Target:

    From the American side of the Atlantic, all of these rationales seem utterly bizarre.

Solving the morphology dilemma with confusion networks

- Conventional solution: reduce morphological complexity by removing morphemes
  - Lemmatize (Goldwater & McCloskey 2005)
  - Truncate (Och)
  - Collapse meaningless distinctions (Talbot and Osborne, 2006)
  - Backoff for words you don't know how to translate (Yang and Kirchhoff)
  - Problem: the removed morphemes contain important translation information
- Surface only:

  From the US side of the Atlantic all such odůvodnění appears to be a totally bizarre.

- Lemma only:

  From the [US] side of the Atlantic with any such justification seem completely bizarre.

Solving the morphology dilemma with confusion networks

atlantiku

atlantik

- Use confusion ... cess to *both* representations ...

z	amerického	břehu	atlantiku	se	veskerá	taková	odůvodnění	jeví	jako	naprosto	bizarní	.
	americký	břeh	atlantik	s		takový		jevit				

- Use surface forms if it makes sense to do so, otherwise back off to lemmas, with individual choices *guided by the model*.
- Create single grammar by combining the rules from both grammars
- Variety of cost assignment strategies available.

Czech-English results

Input	BLEU*
Surface forms only	22.74
Backoff (~ Yang & Kirchhoff 2006)	**23.94**
Lemmas only	22.50
Surface+Lemma (CN)	**25.01**

- Improvements for using CNs are significant at $p<.05$, CN > surface at $p < .01$
- WMT07 training data (2.6M words), trigram LM

• Best system on Czech-English task at WMT'07 on all evaluation measures.

*** 1 reference translation**

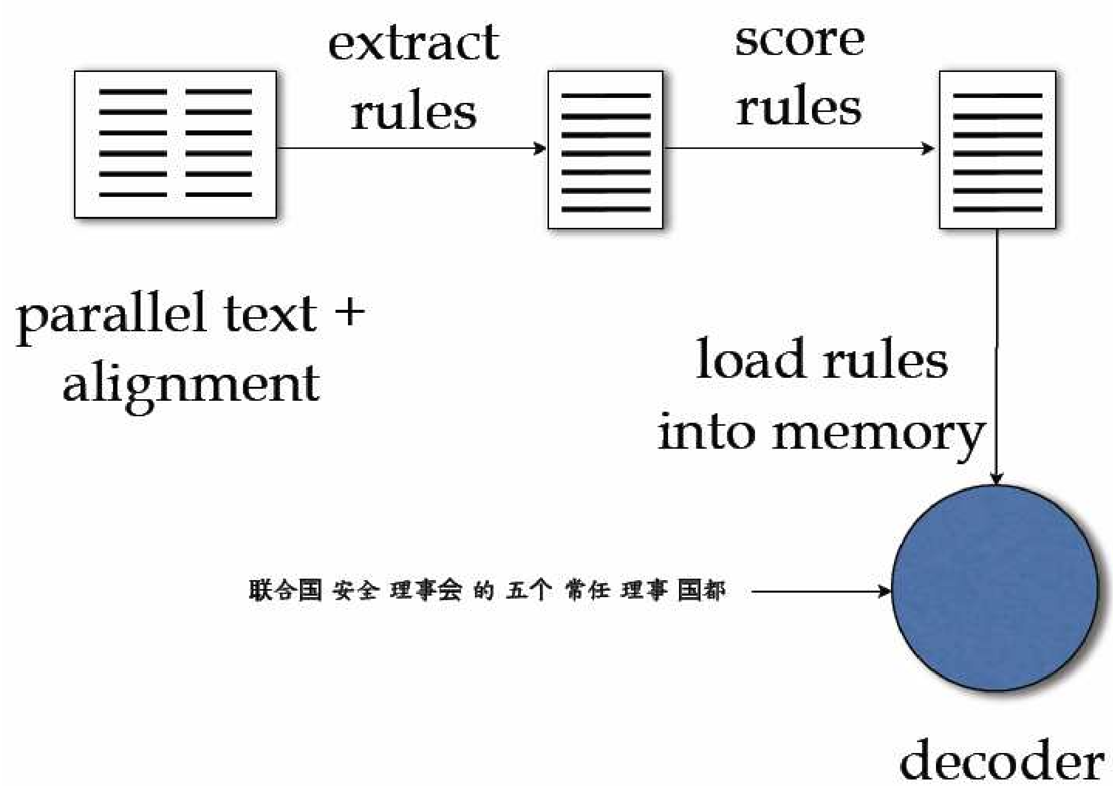
Confusion Networks Summary

- Keeping as much information as possible is a good idea.
  - Alternative transcription hypotheses from ASR
  - Full morphological information
- Hierarchical phrase-based models outperform conventional models
  - Higher absolute baseline
  - Better utilization of ambiguity in the signal (cf. Arabic results)
- Decoding ambiguous input can be done efficiently
- Current work: Arabic morphological backoff

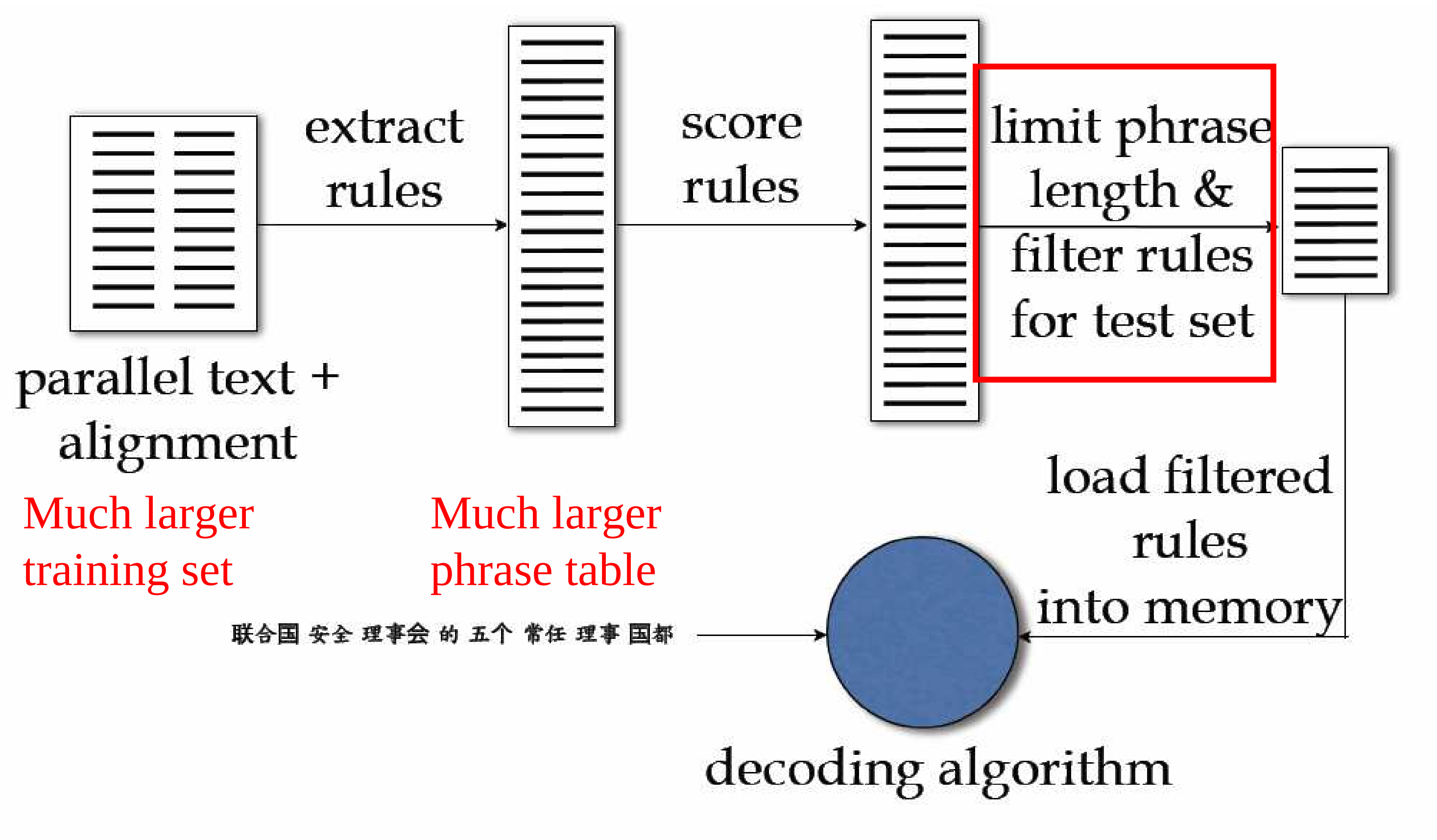
Roadmap

- Brief review of Hiero
- New developments
  - Confusion network decoding (Dyer)
  - Suffix arrays for richer features (Lopez)
  - Paraphrase to improve parameter tuning (Madnani)
- Summary and conclusions

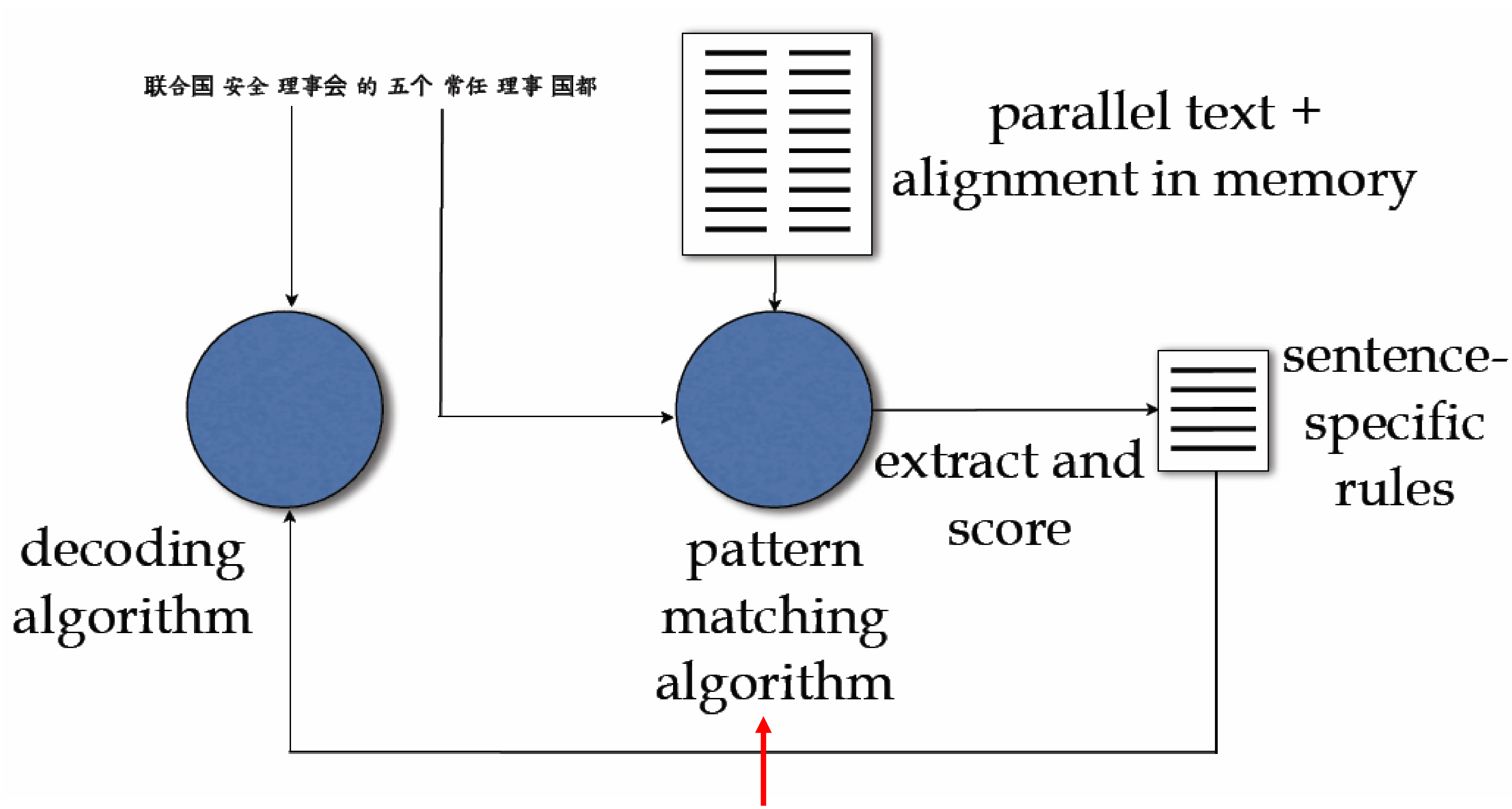
Standard Decoder Architecture



Standard Decoder Architecture



Alternative Decoder Architecture
(Callison-Burch et al., Zhang and Vogel et al.)



Look up (or sample from) all e for substring f

Hierarchical Phrase Based Translation with Suffix Arrays

- Key idea: instead of pre-tabulating information to support features like p(e|f), look up instances of f in the training bitext, <u>on the fly</u>
- Facilitates:
  - Scaling to large training corpora
  - Use of arbitrary length phrases
  - Ability to decode without test set specific filtering
  - Features that use broader context
  - Features that use corpus annotations

Example

(using English as source language for readability)

Input Pattern it persuades him and it disheartens him

Query Patterns

it
persuades
him
and
disheartens
it persuades
persuades him
him and
and it
it disheartens
disheartens him

it persuades him
persuades him and
him and it
and it disheartens
it disheartens him
it persuades him and
persuades him and it
him and it disheartens
and it disheartens him
it persuades him and it
persuades him and it disheartens
him and it disheartens him

…
and it || y él
and it || y ella
and it || pero él
…

Looking source patterns up on the fly

subj

… y él parece que …

… discussed the issue with her and it seems as if …

subj

… mejor pero él …

… offered the organization a better alternative and it …

… y el otro .   …

… built between the new building and it . After proposing …

Efficient Pattern Matching

- If the F side of the bitext is indexed using a *suffix array*, lookup of all matches can be done very quickly.

Example (using English as source language for readability)

it makes him and it mars him . it sets him on and it takes him off . #

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

3 and it mars him . it sets him on and it takes him off . #

12 and it takes him off . #

2 him and it mars him . it sets him on and it takes him off . #

15 him off . #

10 him on and it takes him off . #

6 him . it sets him on and it takes him off . #

0 it makes him and it mars him . it sets him on and it takes him ...

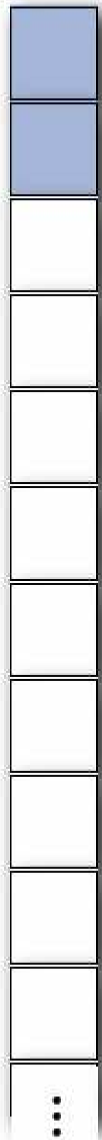
4 it mars him . it sets him on and it takes him off . #

⋮ ...

it persuades him and it disheartens him

Query pattern w

and it



and it mars him . it sets him ...
and it takes him off . #
him and it mars him . it sets ...
him off . #
him on and it takes him off . #
him . it sets him on and it ...
it makes him and it mars ...
it mars him . it sets him on ...
it sets him on and it takes ...
it takes him off . #
makes him and it mars him ...

Problem: patterns with gaps

(using English as source language for readability)

Input Pattern it persuades him and it disheartens him

Query Patterns

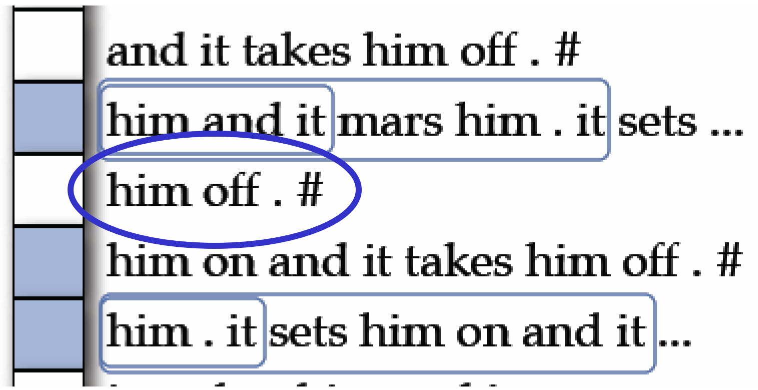
it X and	it X disheartens him
it X it	it X and X him
it X disheartens	persuades him X disheartens
it X him	persuades him X him
persuades X it	persuades X it disheartens
persuades X disheartens	persuades X disheartens him
persuades X him	him and X him
it persuades X it	him X disheartens him
it persuades X disheartens	it persuades him X disheartens
it persuades X him	it persuades him X him
it X and it	it persuades X it disheartens
it X it disheartens	it persuades X disheartens him

…

- Instances of pattern are no longer contiguous in suffix array

- Naïve approaches (e.g. using intersection of subpatterns) are <u>very</u> inefficient – baseline timing result is that decoding takes **2241 seconds** per sentence!

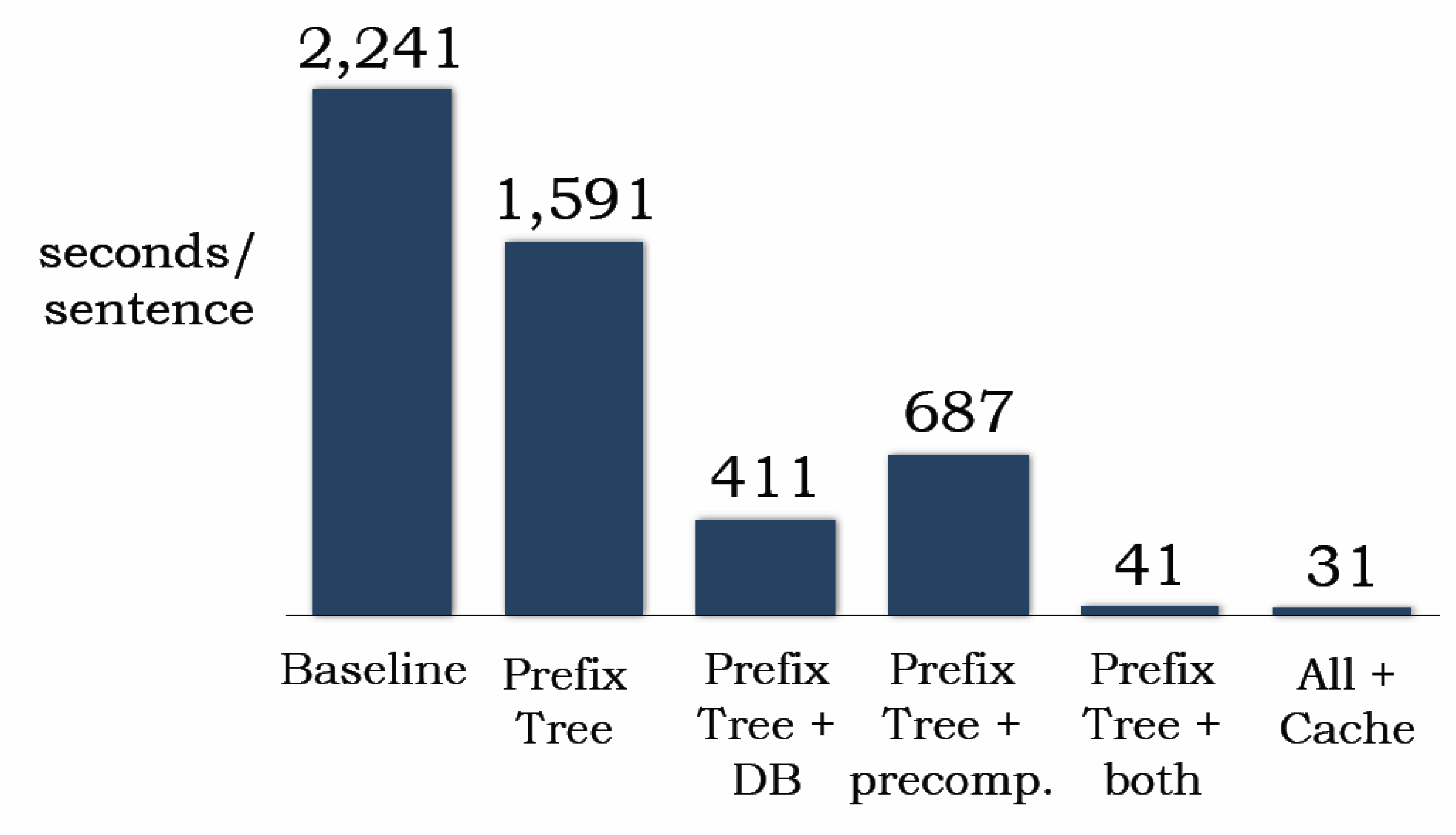
Query pattern: him X it



Algorithmic extensions

- Exploiting redundancy using prefix tree with suffix links (Zhang and Vogel 2005)
- Double binary search (Baeza-Yates 2004) for cases where there is an infrequent subpattern
- Precomputation for cases where there are multiple frequent subpatterns
- Caching

Timing Results



Applications

- Sampling for feature value estimation
- Features based on context
- Features based on annotations

- Take-home message: the suffix array framework allows very rapid exploration of a larger feature space.

Roadmap

- Brief review of Hiero
- New developments
  - Confusion network decoding (Dyer)
  - Suffix arrays for richer features (Lopez)
  - Paraphrase to improve parameter tuning (Madnani)
- Summary and conclusions

Using paraphrases to improve parameter tuning

- Virtually all SMT systems tune model parameters by optimizing an objective function that compares decoder output to reference translations (e.g. BLEU).
- It's widely accepted that multiple references per translation are better.
- But references are expensive to obtain.
- Could we exloit a quantity/quality tradeoff by increasing the number of references artificially?

Example

H: *the cat was devoured by the canine*

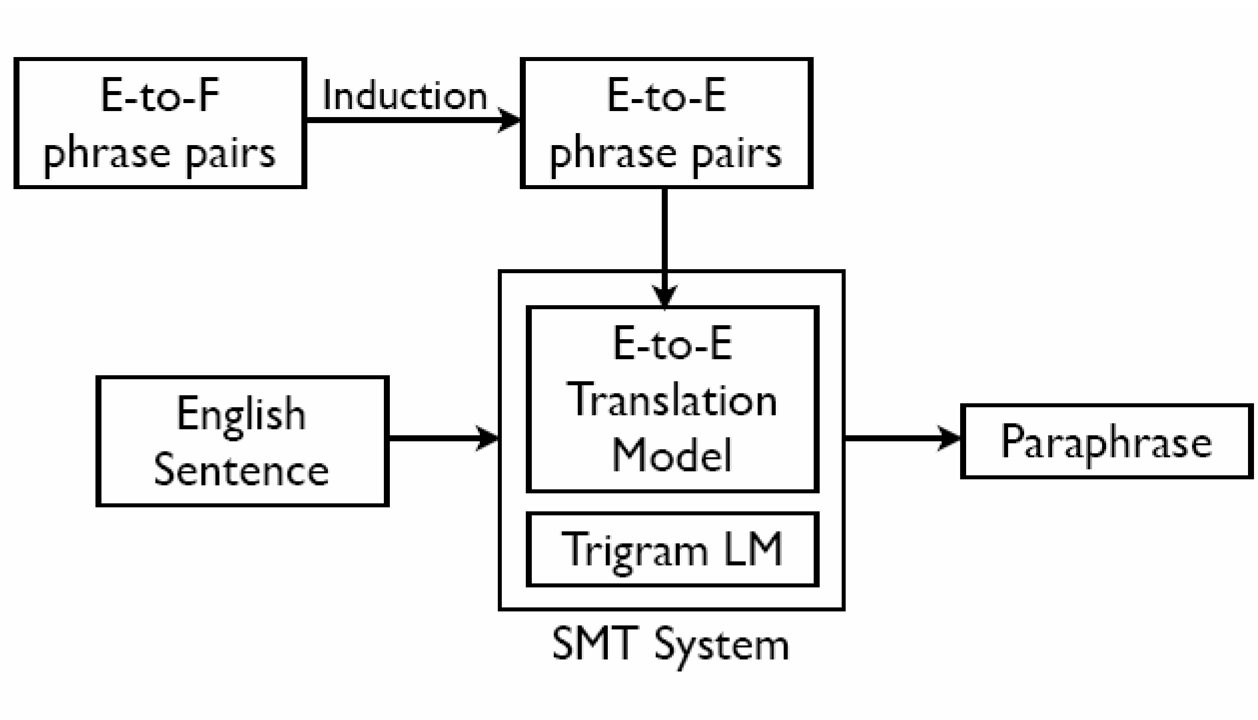
R_1: *the dog ate the cat*

R_2: *the cat was devoured by the dog*

R_3: *the dog devoured the cat*

R_4: *the feline was eaten by the canine*

Paraphrase as English-to-English translation



Examples
(Europarl, using French as pivot)

we must bear in mind the community as a whole .
we must remember the wider community .

they should be better coordinated and more effective .
they should improve the coordination and efficacy .

women are still one of the most vulnerable sections of society , whose rights are rudely trampled underfoot by the current social and economic system .
they remain one of the weakest in society , whose duties are abruptly scorned by the present social and economic order .

that is what we are waiting to hear from the european commission .
that is what we expected from the meeting .

this occurred not far away and not very long ago .
this substances not far behind and very recently .

Examples
(NIST’03 test set using Chinese as pivot)

the copy of the ultimatum has been sent to un security council .
the text of the ultimatum was rushed to the security council .

france circulated its proposal in the form of a " non-official paper " .
french transmits its recommendations to serve as a " non-official document " .

(hong kong , macao and taiwan) macao passes bill to avoid double taxation
(hong kong , macau and taiwan) macau adopts draft avoidance of double taxation

however , people know little about the cause of the disease so far .
however , persons are not sure present cause .

however , many experts said that technically speaking alone , the time for the deployment of a missile defense system was not ripe .
however , many experts believe that the new site alone , the duration of deploy a missile defense system immature .

Experiment

- Source Language: Chinese
- Training: newswire parallel text (850000 sentences)
- Dev set: NIST MT'03 (919 sentences)
- Test set: NIST MT'05 (1082 sentences)
- LM: SRILM trigram model with modified Kneser-Ney smoothing, 155M words
- Metrics: BLEU-4 and TER (lowercased)

Results

Tuning References	*BLEU*	*TER*
2H	30.43	59.82
2H + 2P	31.10*	58.79
4H	31.26	58.66
4H + 4P	31.68	58.24

- Score tuning on four human references is matched (statistically) with only two human references needed.
- "Standard" (for NIST) four references can still improve.

Tuning References	*BLEU*	*TER*
1H	29.39	62.37
1H + 1P	31.06*	59.39

- Potentially more interesting scenario, since any bitext provides one human reference translation per source sentence.
- Raises the possibility of topic and genre-specific parameter tuning.

Conclusions

- Hiero is both a framework and a strategy for bringing more linguistically relevant properties into statistical MT
  - Start with hierarchy, lexically anchored reordering
  - Be driven by parallel data, not by monolingual analysis
  - Embrace and extend phrase-based ideas that work well
  - Tackle cross-cutting challenges (e.g. more ref translations)

Thanks and acknowledgements

- The work presented would not have been possible without the many good ideas and generous assistance from the following people:

  Nicola Bertoldi
  David Chiang
  Marcello Federico
  Ian Lane
  Lidia Mangu
  Smaranda Muresan
  Daniel Zeman
  Richard Zens

And thank <u>you</u>!