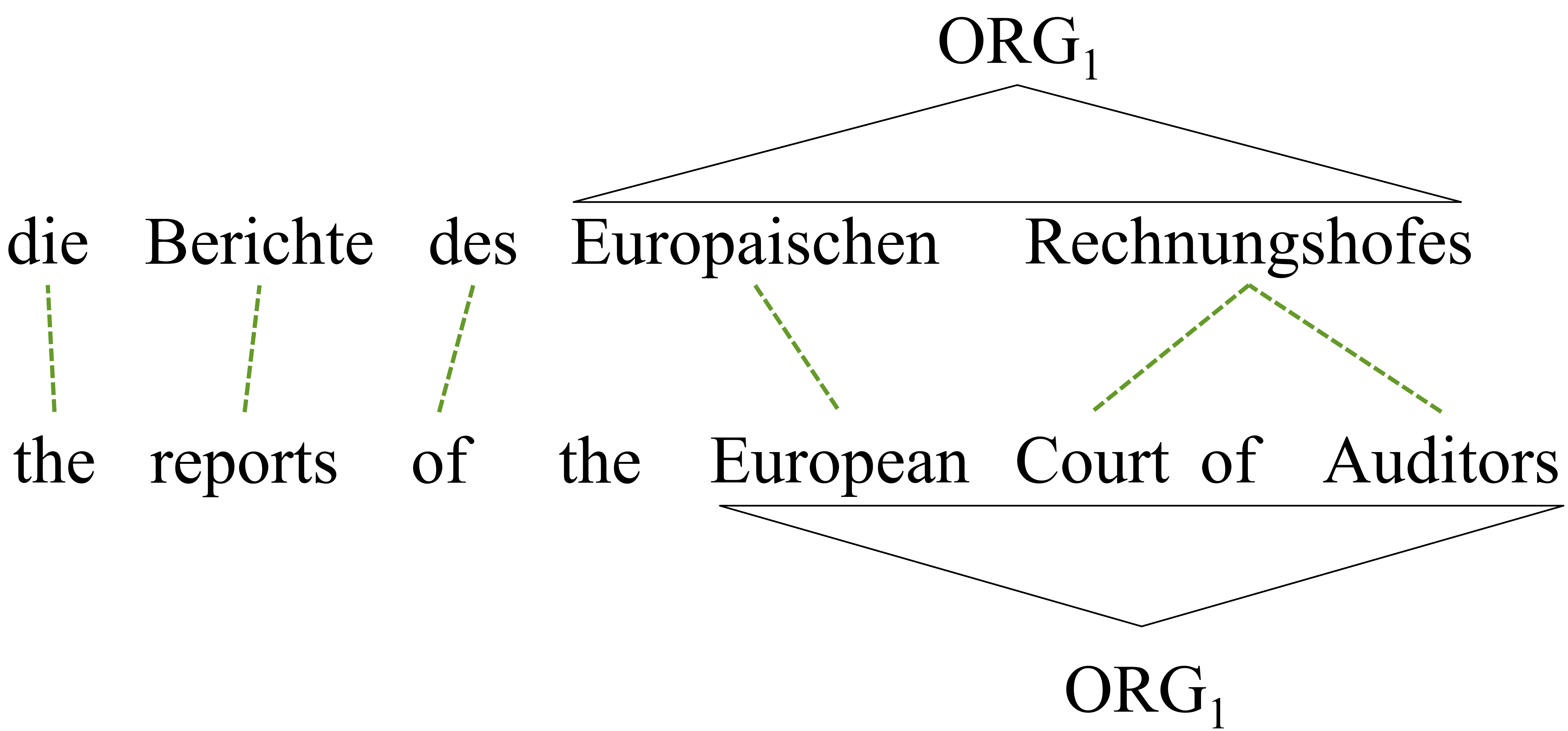
Learning Better Monolingual Models from Bilingual Data

John Blitzer

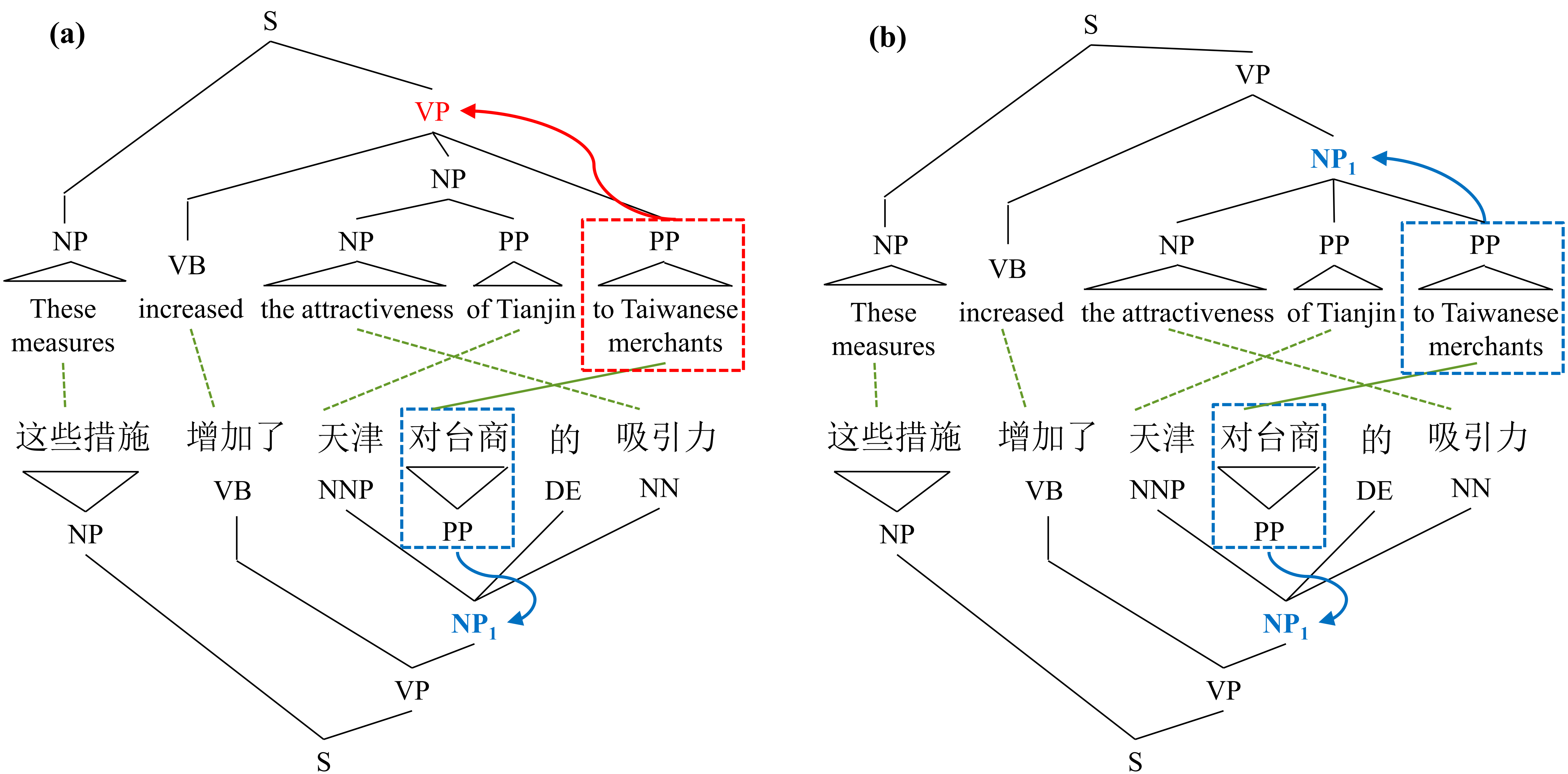
David Burkett, Wei Gao, Dan Klein, Slav Petrov, Ming Zhou

Improving Named Entity Recognition (NER)



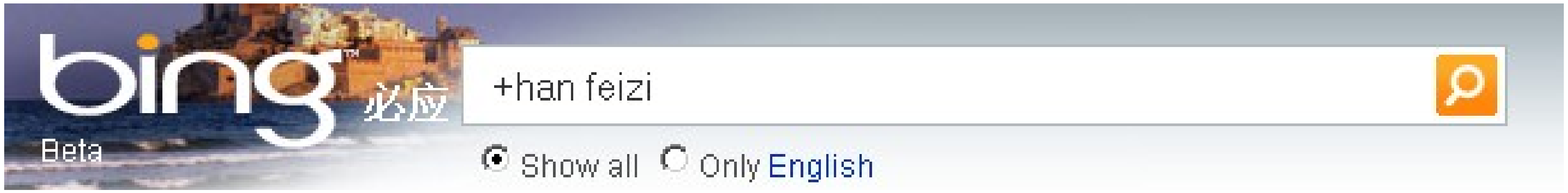
English parallel text can improve German NER

Improving Syntactic Parsing



Chinese parallel text can improve English PP attachment

Improving Web Search



Han Feizi - China culture
Han Fei, Li Si, Xunzi **Han Feizi**, Han Fei, Confucianism
General (but short) introduction to Han Feizi

韩非子 － 搜狐博客 Translate this page
韩非子 韩非子 ... 客服留言板 | 客服博客 | 客服邮箱 | 24小时客服热线:010-58511234
Chinese Spam

Han Feizi (book) - Wikipedia
The **Han Feizi** is a work written by **Han Feizi** at the end of the Warring States Period
On topic, but missing some information

韩非子_百度百科
是中国古代著名的哲学家、思想家，政论家和散文家，法家思想的集大成者，后世称“韩子”或“韩非子”。
Very complete biography of Han Feizi

韩非子
目录. • 初见秦第一 • 存韩第二 • 难言第三 • 爱臣第四 • 主道第五 • 有度第六 • 二柄第七 • 扬権第八
The complete works of Han Feizi

.... Lower ....
www.hawickert.de/**HanFeizi**.htm
A much more complete description of Han Feizi's work, with excerpts

Part 1: Sentence-Level Models

Input: Original Monolingual Models
Bilingual Data

Output: Bilingual Model
Improved Monolingual Models

Multi-view Training

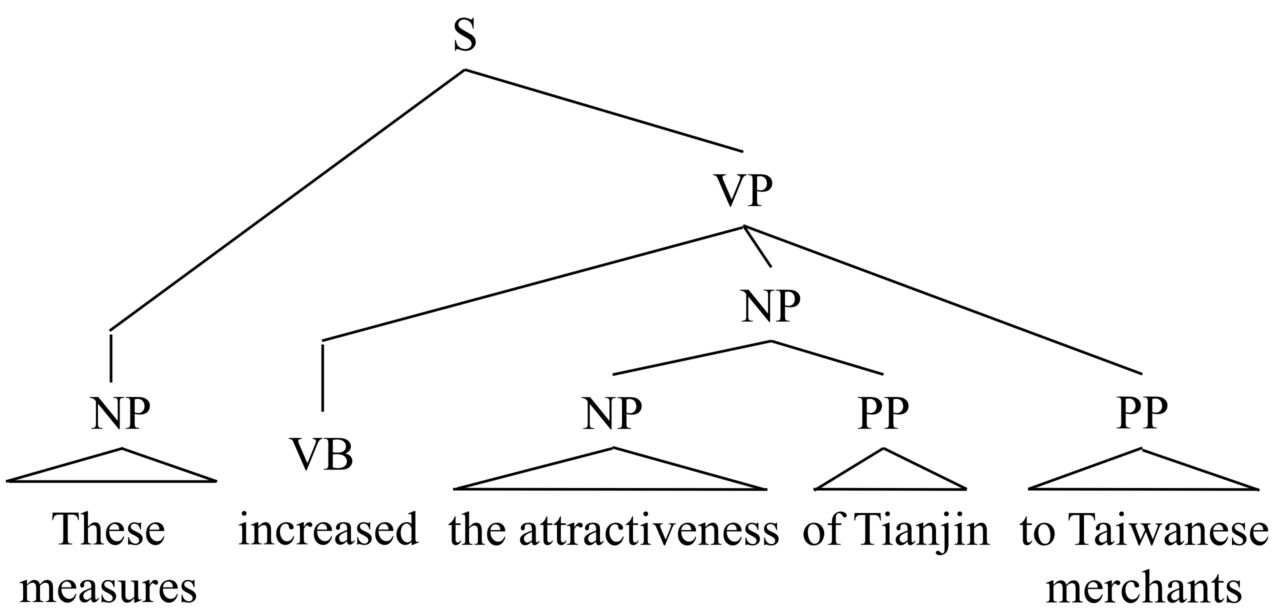
(1) Label bilingual data with original monolingual models

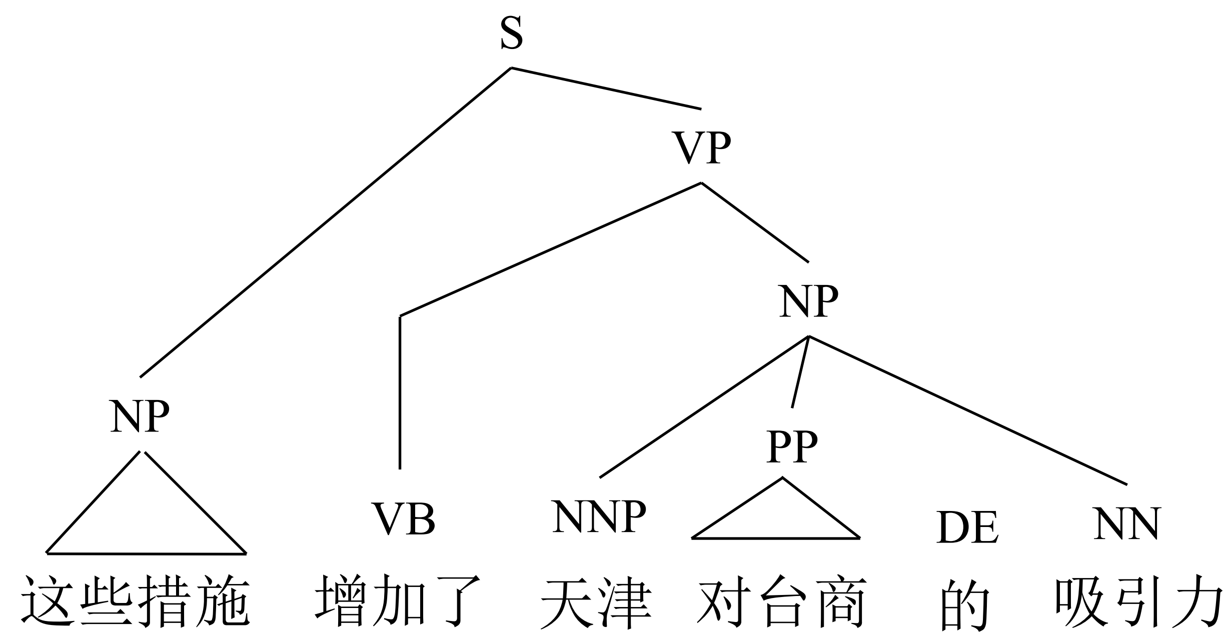
(2) Train bilingual model on the output of the monolingual models

(3) Combine bilingual and monolingual models

(4) Retrain improved monolingual models on combined output

Some Notation





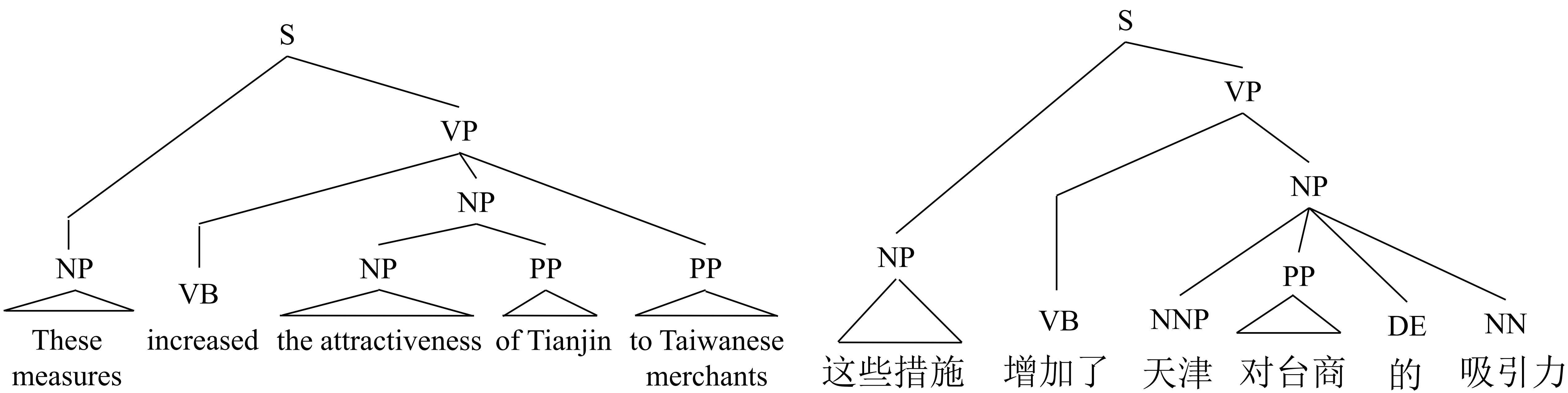
$$x = (x_1, x_2)$$

$$y = (y_1, y_2)$$

Bilingual label-label alignments

We want features that generalize (i.e. only use pairs of nodes in each tree)

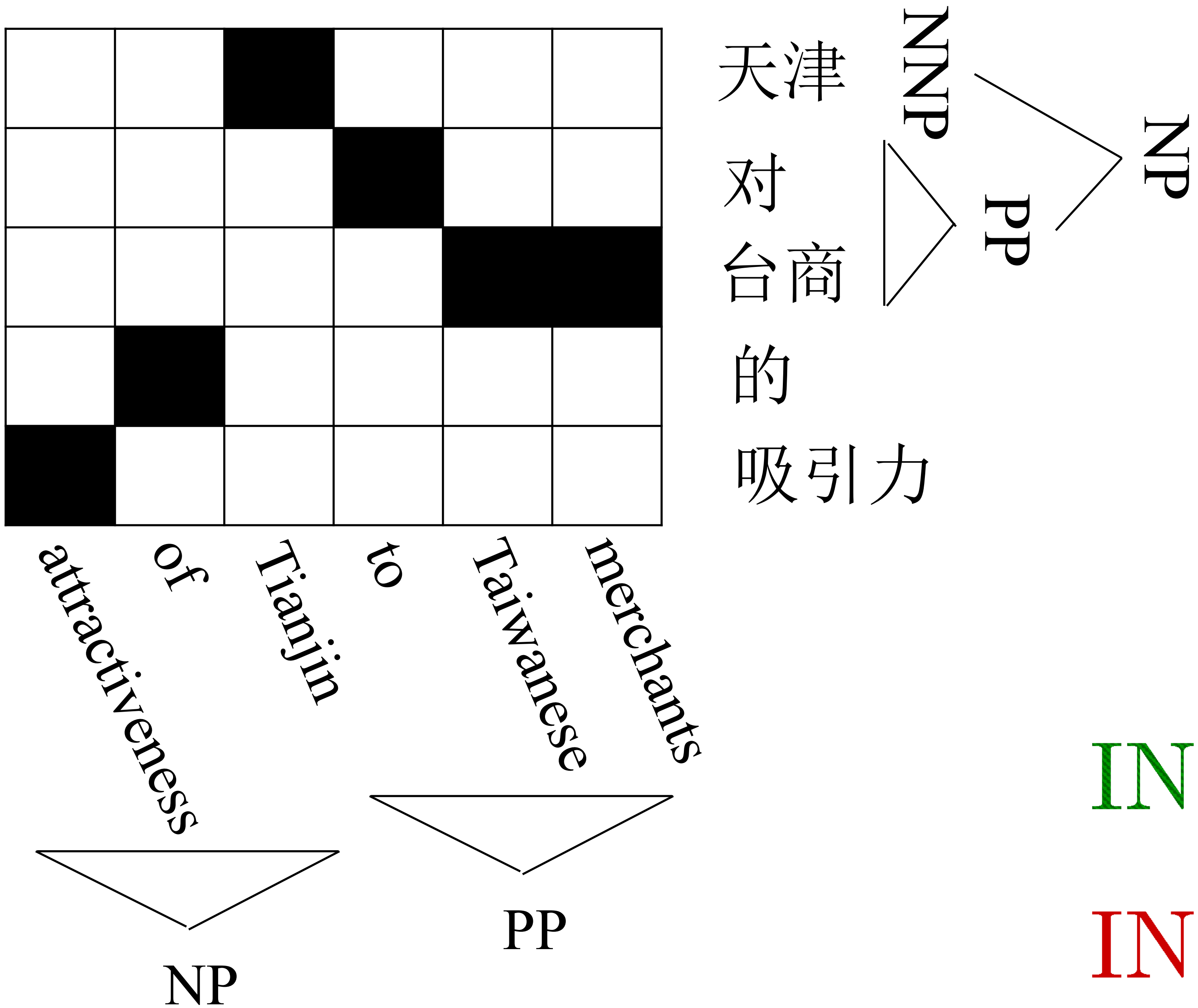
But we don't know how label pieces correspond



$$p_{\boldsymbol{\theta}}(y|x) = \sum_a p_{\boldsymbol{\theta}}(y, a|x) \qquad\qquad q_{\boldsymbol{\theta}}(y|x) = \max_a p_{\boldsymbol{\theta}}(y, a|x)$$

$$p_{\boldsymbol{\theta}}(y, a|x) = \exp\left[\boldsymbol{\theta}^\top \boldsymbol{\phi}(y_1, a, y_2) - A(\boldsymbol{\theta}; x)\right]$$

Bilingual Training: Inside Ratio Feature

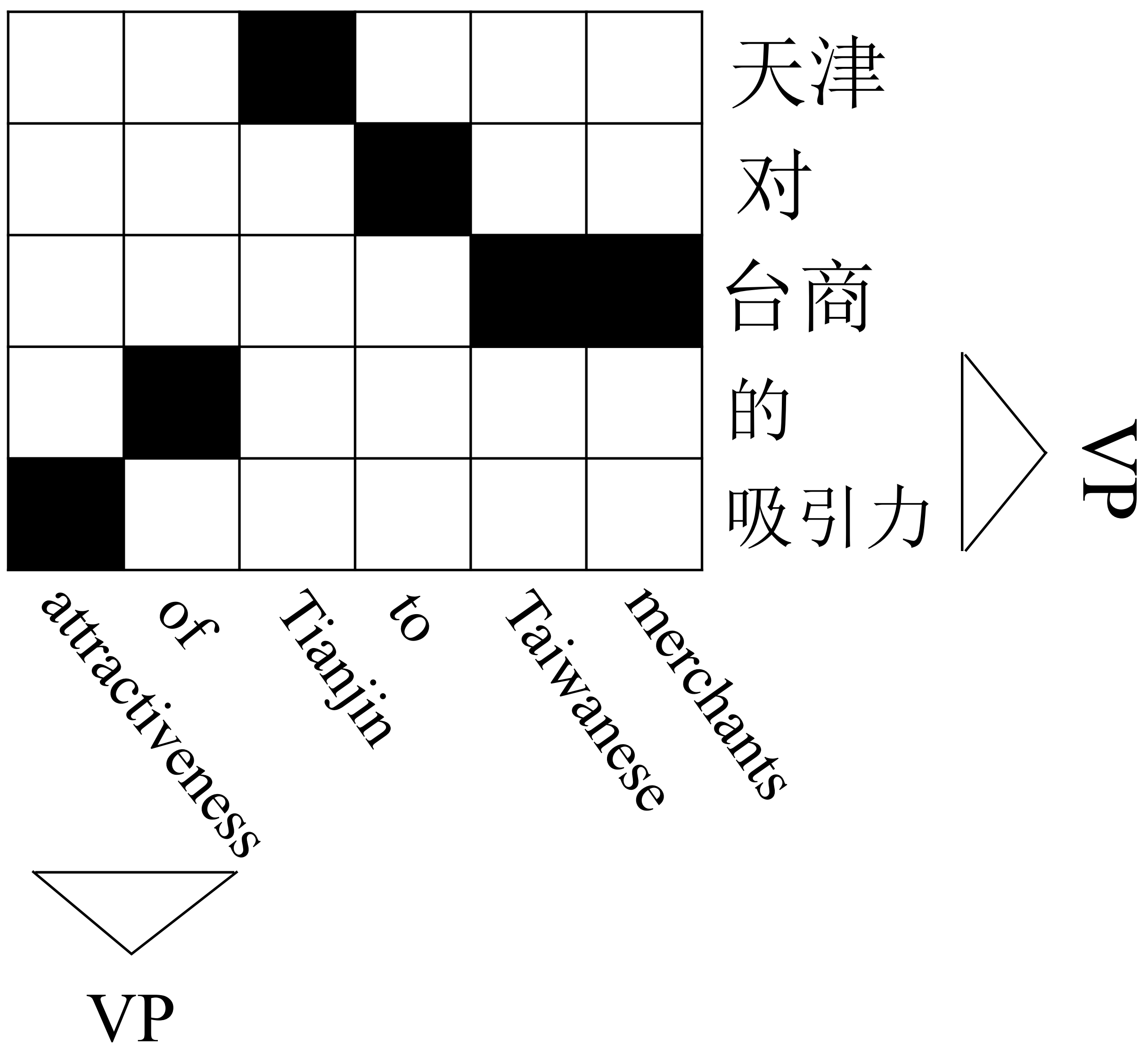


INS-RATE: $\frac{3}{3}$

INS-RATE: $\frac{1}{6}$

Monolingual Features in the Bilingual Model

We need some monolingual input for accurate modeling



But we can't include the full monolingual models as features

We include weakened versions of the monolingual models

$$q_{\lambda_1,\lambda_2,\boldsymbol{\theta}}(y|x) \stackrel{def}{=} \max_{a} \; \exp\left[\lambda_1 \ell_1^W + \lambda_2 \ell_2^W + \boldsymbol{\theta}^\top \boldsymbol{\phi}(y_1, a, y_2) - A(\lambda_1, \lambda_2, \boldsymbol{\theta}; x)\right] .$$

Final Training Procedure

Input: full and weakened monolingual models:

$p_1^F(y_1|x_1)$, $p_2^F(y_2|x_2)$, $p_1^W(y_1|x_1)$, $p_2^W(y_2|x_2)$

unannotated bilingual data

Output: bilingual parameters: $\hat{\boldsymbol{\theta}}, \hat{\lambda}_1, \hat{\lambda}_2$

1. Label U with full monolingual models:

   $\forall x \in U, \;\; \hat{y}_M = \operatorname{argmax}_y p_1(y_1|x_1) p_2(y_2|x_2)$.

2. Return $\operatorname{argmax}_{\lambda_1,\lambda_2,\boldsymbol{\theta}} \;\; \prod_{x\in U} q_{\boldsymbol{\theta},\lambda_1,\lambda_2}(\hat{y}_M|x)$

Combining Mono and Bilingual Models

How to do prediction? 3 Choices:

1) Bilingual model only:

$$\underset{y}{\operatorname{argmax}} \; \max_{a} \exp\left[\lambda_1 \ell_1^W + \lambda_2 \ell_2^W + \boldsymbol{\theta}^\top \boldsymbol{\phi}(y_1, a, y_2) - A(\lambda_1, \lambda_2, \boldsymbol{\theta}; x)\right]$$

2) Uniform combination:

$$\underset{y}{\operatorname{argmax}} \; \max_{a} \exp\left[\ell_1^F + \ell_2^F + \boldsymbol{\theta}^\top \boldsymbol{\phi}(y_1, a, y_2) - A(\lambda_1, \lambda_2, \boldsymbol{\theta}; x)\right]$$

3) Replace weakened with full monolingual model:

$$\underset{y}{\operatorname{argmax}} \; \max_{a} \exp\left[\lambda_1 \ell_1^F + \lambda_2 \ell_2^F + \boldsymbol{\theta}^\top \boldsymbol{\phi}(y_1, a, y_2) - A(\lambda_1, \lambda_2, \boldsymbol{\theta}; x)\right]$$

Parsing Data and Setup

- **Labeled Data:** Penn Treebank and Chinese Treebank
- **Parser:** Berkeley Parser (State-split, latent variable parser)
- **Weakened Models:** 3 iters of state splitting (5 for full model)
- **Unlabeled Data:** Parallel portion of Chinese treebank
- **Testing Data:** Parallel portion of Chinese treebank
- **Bilingual Features:** Burkett & Klein (2008)

Parsing Results – Chinese & English

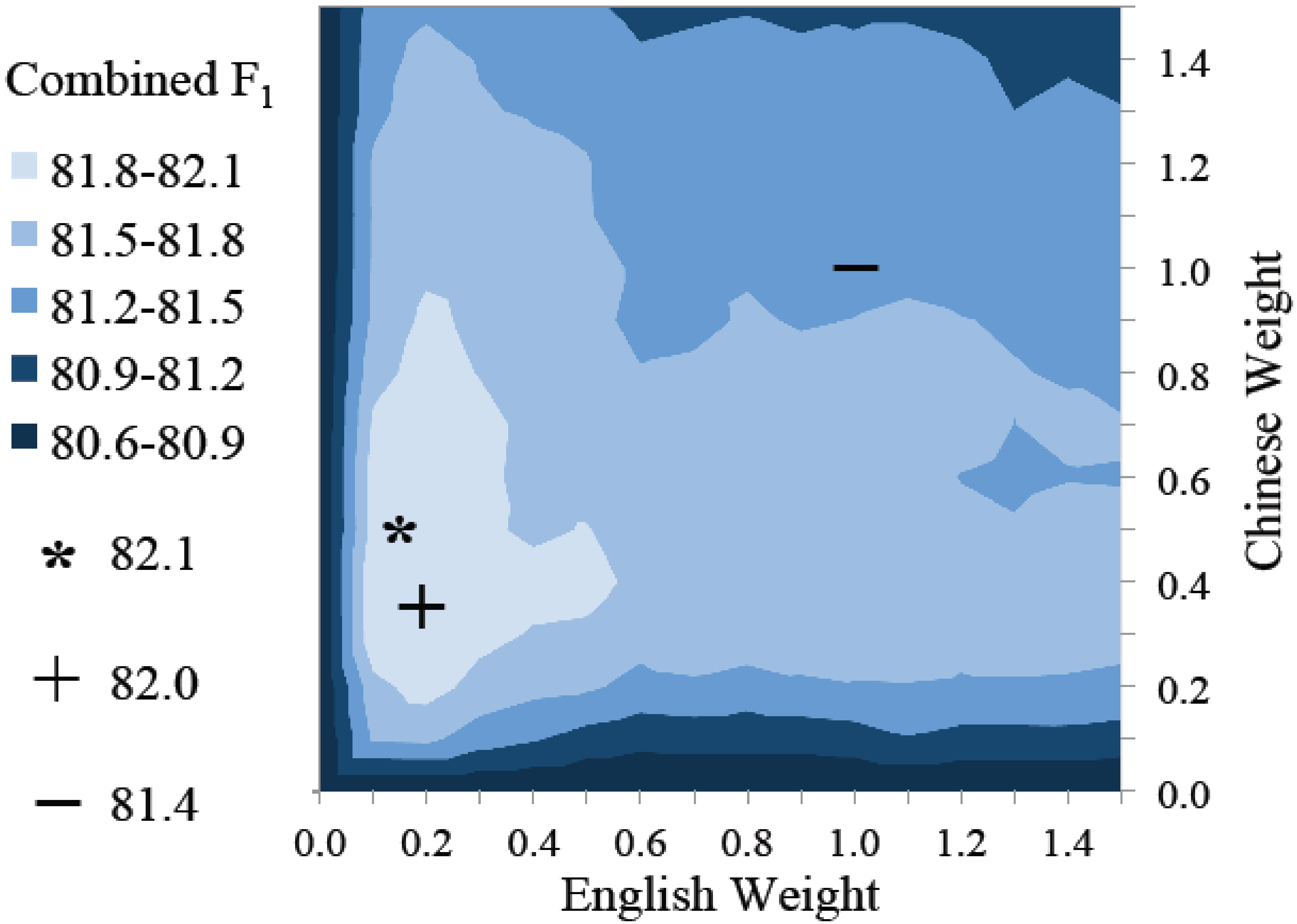
Model	**Chinese F1**	**English F1**
Monolingual Baselines		
Weakened Monolingual	78.3	67.6
Full Monolingual	84.2	75.4
Bilingual Models		
Bilingual only	80.4	70.8
Bilingual + Monolingual	**85.9**	**77.5**
Retrained Monolingual Models		
Self-Retrained	83.6	76.7
Bilingual Retrained	**83.9**	**77.4**

Syntactic MT Rule Extraction

- **Unlabeled Data:** 100,000 parallel Chinese-English sentences
- **Testing Data:** 1-reference test set from the same domain

Model	**BLEU**
Phrase-based	
Moses	18.8
Syntactic (Galley et al. 2006) Models	
Penn Treebank	18.7
Burkett & Klein (2008)	21.1
Bilingual	21.2

Comparing Prediction Methods



– Uniform combination **+ Weakened Weights** *** Optimal**

Part 2: Bilingual Web Search Ranking

Input: Bilingual query log, documents
Click-through statistics for each query-document pair

Output: Improved Ranking Model for Bilingual Queries

<u>Training</u>

(1) Create a bilingual ranking problem, where instances consist of pairs of similar web pages (one from each language)

(2) Train a ranking model that exploits bilingual and monolingual features

Prediction: Reconstruct monolingual ranking from bilingual ranking

We never need to show machine-translated pages to an end user!

Creating training data: From clickthrough rates to rankings

Input: Query pair, documents, & clickthroughs for each language

Bilingual query pair (***Mazda, 马自达***)		
doc	URL	Aggr. click #
e1	www.mazda.com	229
e2	www.mazdausa.com	185
e3	www.mazda.co.uk	5
e4	www.starmazda.com	2
e5	www.mazdamotorsports.com	2
	……	
c1	www.faw-mazda.com	50
c2	price.pcauto.com.cn/brand.jsp?bid=17	43
c3	auto.sina.com.cn/salon/FORD/MAZDA.shtm	20
c4	car.autohome.com.cn/brand/119/	18
c5	jsp.auto.sohu.com/view/brand-bid-263.html	9
	……	

→

e1>e2
e1>e3
…
e2>e3
e2>e4
…
e4>e5

→

c1>c2
c1>c3
…
c2>c3
c2>c4
…
c4>c5

From monolingual to bilingual rankings

2 natural conditions for constructing a bilingual ranking from monolingual rankings

$$\left(e_i^{(1)}, c_j^{(1)}\right) \succ \left(e_i^{(2)}, c_j^{(2)}\right) \quad \text{if and only if}$$

$$e_i^{(1)} > e_i^{(2)} \text{ and } c_j^{(1)} \geq c_j^{(2)}$$

or

$$e_j^{(1)} > e_j^{(2)} \text{ and } c_i^{(1)} \geq c_i^{(2)}$$

e1>e2, e1>e3, e1>e4

e2>e3, e2>e4

c1>c2, c1>c3, c1>c4

c2>e3, c2>c4

(e1,c1) > (e1, c2), (e1,c1) > (e2, c1)

(e1, c2) > (e1, c3), (e1, c2) > (e2, c2)

. . . .

(e2, c3) > (e2, c4), (e2, c3) > (e3, c3)

Learning a bilingual ranking function

Training data

(e1,c1) > (e1, c2),	(e1,c1) > (e2, c1)
(e1, c2) > (e1, c3),	(e1, c2) > (e2, c2)
. . . .	
(e2, c3) > (e2, c4),	(e2, c3) > (e3, c3)

$$f : (e_i, c_j) \longrightarrow \mathbb{R}$$

Score given by f allows us to reproduce the ranking

$$(e_i^1, c_j^1) \succ (e_i^2, c_j^2) \leftrightarrow f(e_i^1, c_j^1) > f(e_i^2, c_j^2)$$

We learn a linear function with RankSVM (Herbrich et al. 2000)

Features

- **Monolingual Features**
  - BM 25 features
  - language model ranking, pseudo-relevance feedback, etc.
  - PageRank (Brin and Page 1998) & HITS (Kleinberg 1999)

- **Bilingual Features**
  - Dictionary-based cosine similarity
  - Machine translation based similarities (forward & backward)
  - URL LCS ratio (URL)

    www.airbus.com vs. www.airbus.com.cn

Prediction – Construct monolingual ranking

- **Reconstructing a monolingual ranking is over-constrained**

(e1,c1): 0.4	(e1,c2): 0.3
(e2,c1): 0.1	(e2,c2): 0.5
(e3,c1): 0.6	(e3,c2): 0.2

- **Two heuristics (English):**

H-1 (max score)

$$s(e_i) = \max_j f\left(e_i, c_j\right)$$

e3: 0.6
e2: 0.5
e1: 0.4

H-2 (avg score)

$$s(e_i) = \frac{1}{n}\sum_j f\left(e_i, c_j\right)$$

e3: 0.4
e1: 0.35
e2: 0.3

How many queries are bilingual?

- Examples of local (monolingual) queries

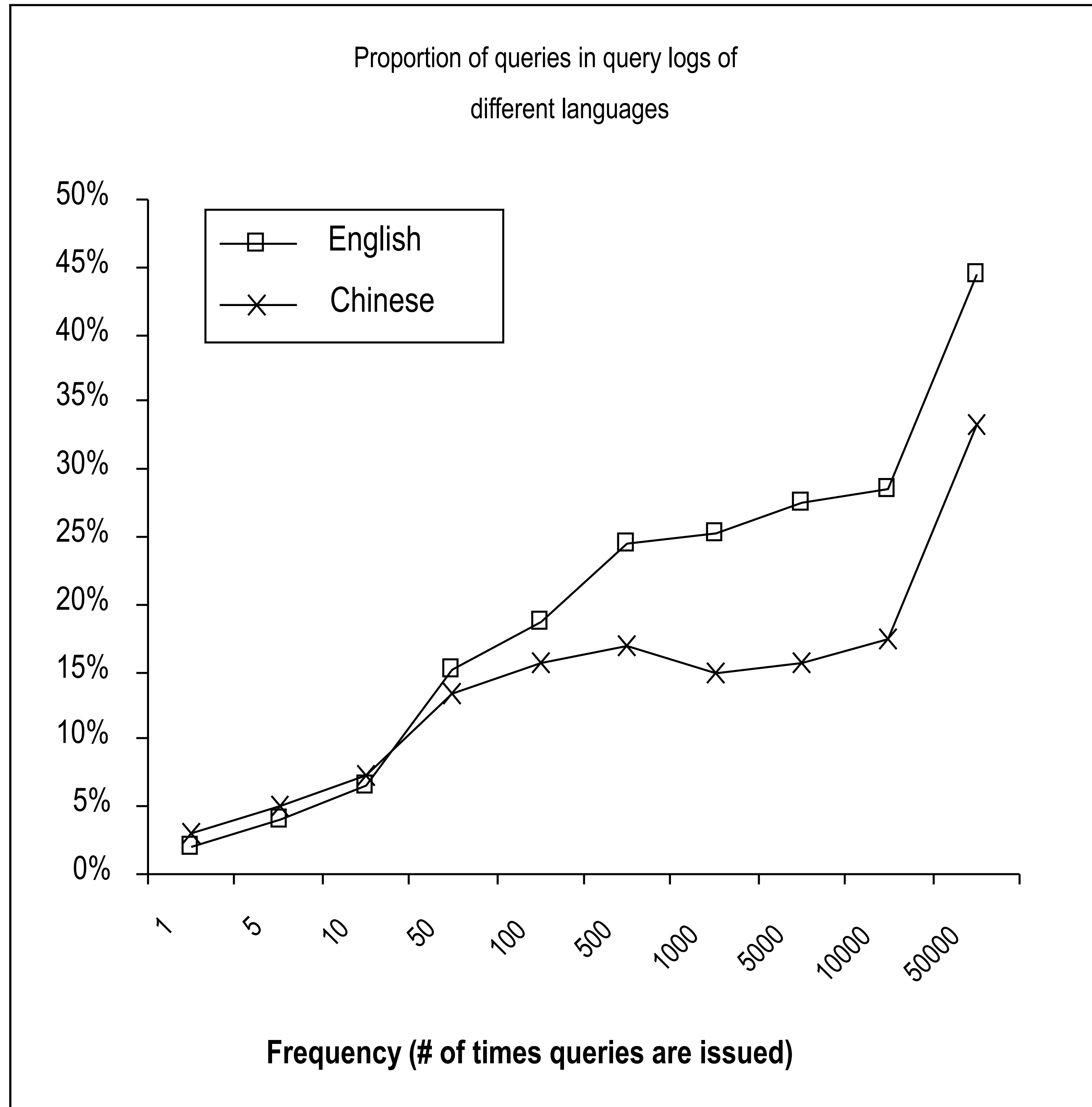
English: Map of Alabama
阿拉巴马地图

Chinese: 长虹电视机
Changhong TV set

- Statistics – Bilingual queries by token

English query log: 1.3%
Chinese query log: 2.3%

Bilingual queries by type



Evaluation Setup

- Query logs: AOL (English) and Sougou (Chinese)
- Total bilingual queries : 1,000 after discarding low click-through documents
- Total document count: 21,000 English, 28,000 Chinese
- Evaluation: Kendall's Tau on heldout click-through

Chinese Ranking Performance

	Pair	H-1 (max)	H-2 (mean)
Monolingual baseline	n/a	0.2935	0.2935
IR (no similarity)	0.3201	0.2938	0.2938
IR+DIC	0.3220	**0.2970*** (p=*0.0060*)	**0.2973*** (p=*0.0020*)
IR+MT	0.3299	**0.2992*** (p=*0.0034*)	**0.3008*** (p=*0.0003*)
IR+DIC+MT	0.3295	**0.2991*** (p=*0.0014*)	**0.3004*** (p=*0.0008*)
IR+DIC+MT+URL	0.2979	**0.2981*** (p=*0.0005*)	**0.3024*** (p=*1.5e-6*)

Top Improved Queries

Most improved CH queries	**Most improved EN queries**
沙门氏菌 (salmonella)	**free online tv (免费在线电视)**
苏格兰 (scotland)	weapons (武器)
咖啡因 (caffeine)	lily (百合)
墓志铭 (epitaph)	cable (电缆)
英国历史 (british history)	sunrider (仙妮蕾德)
政治漫画 (political cartoons)	**aniston (安妮斯顿)**

Conclusions

- Bilingual data is plentiful & covers many domains
- Monolingual models can be improved with bilingual data
- MT is useful as a backend, as well as a goal in itself

Thanks!

NER Data and Setup

- **Labeled Data:** CoNLL 2003 German and English corpora
- **Weakened Models:** Obtained by dropping features
- **Unlabeled Data:** European Parliamentary Proceedings
- **Testing Data:** Manually annotated parliamentary proceedings and parallel newswire text
- **Bilingual Features:** Typed and untyped bispan INS-OUT

NER Results – German Parliament

Model	Precision	Recall	F1
Monolingual Baselines			
Weakened Monolingual	71.3	36.4	48.2
Full Monolingual	69.4	44.4	54.0
Bilingual Models			
Bilingual only	**70.1**	66.3	68.2
Bilingual + Monolingual	**70.1**	**70.1**	**70.1**
Retrained Monolingual Models			
Self-Retrained	70.4	44.4	54.2
Bilingual Retrained	**74.5**	**63.6**	**68.6**