11,001 New Features for Statistical Machine Translation

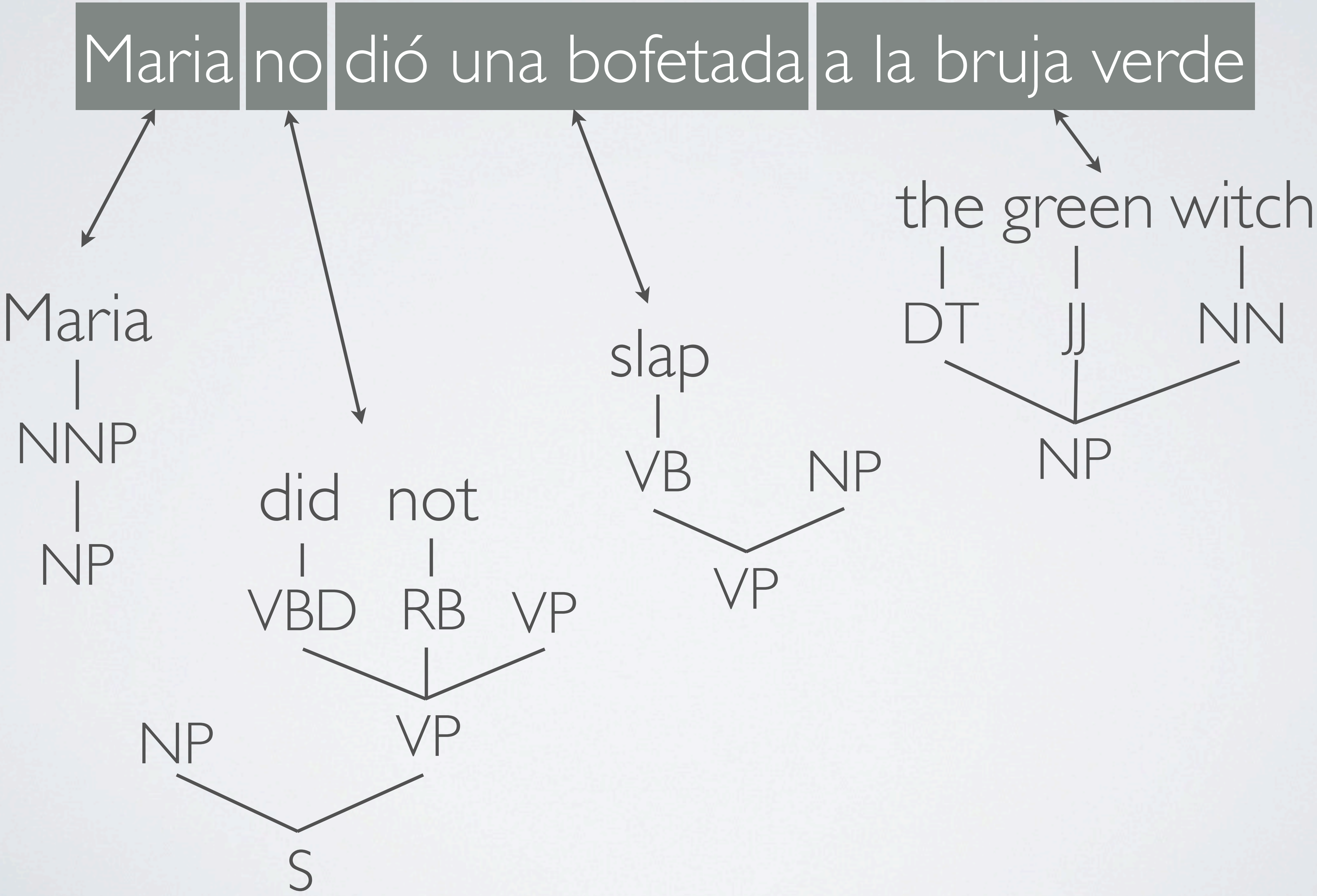
David Chiang
Kevin Knight
Wei Wang

ISI
Information Sciences Institute

LANGUAGE WEAVER

USC Viterbi
School of Engineering

MOTIVATION



Maria no dió una bofetada a la bruja verde

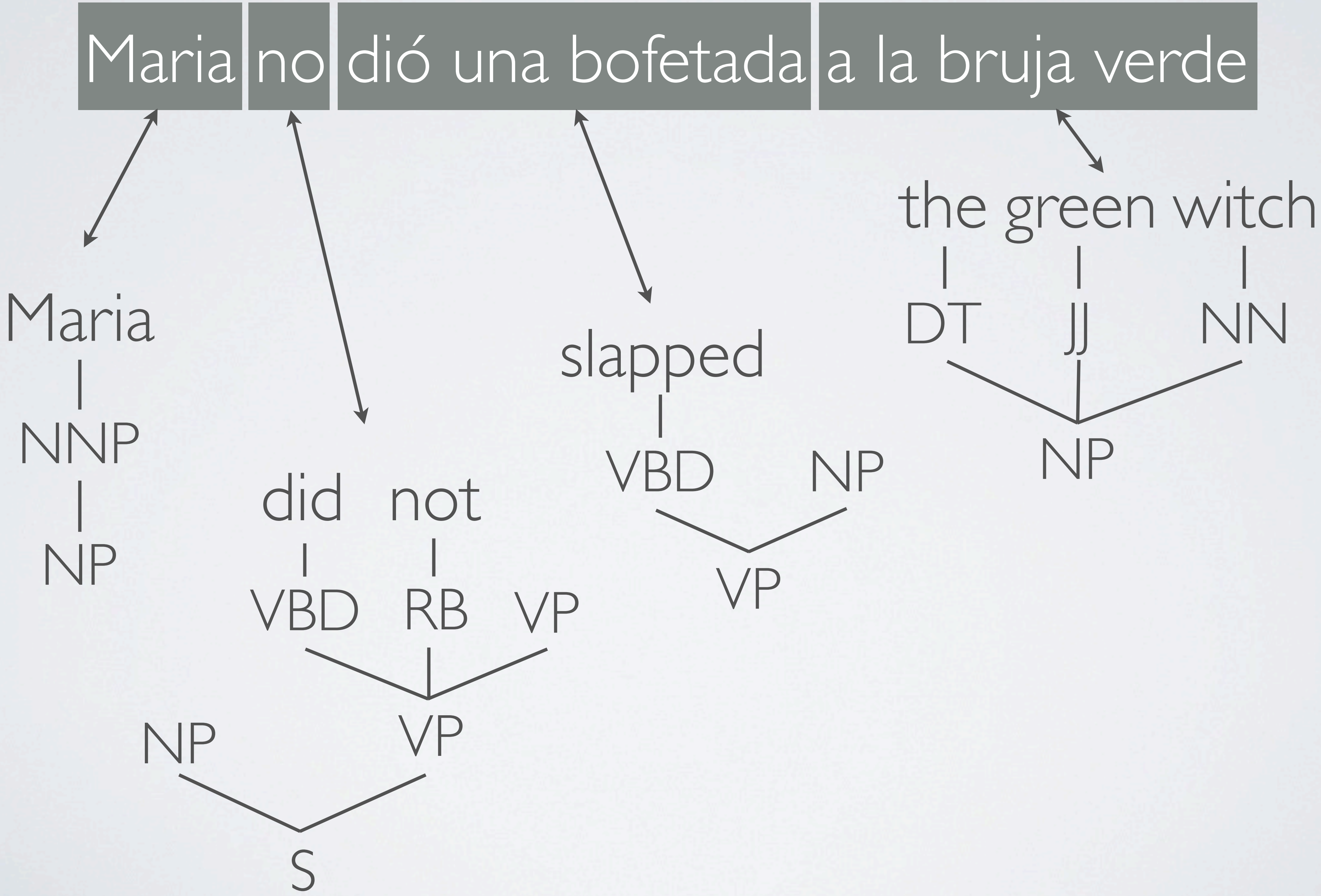
Maria
NNP
NP

did not
VBD RB VP
VP
NP VP
S

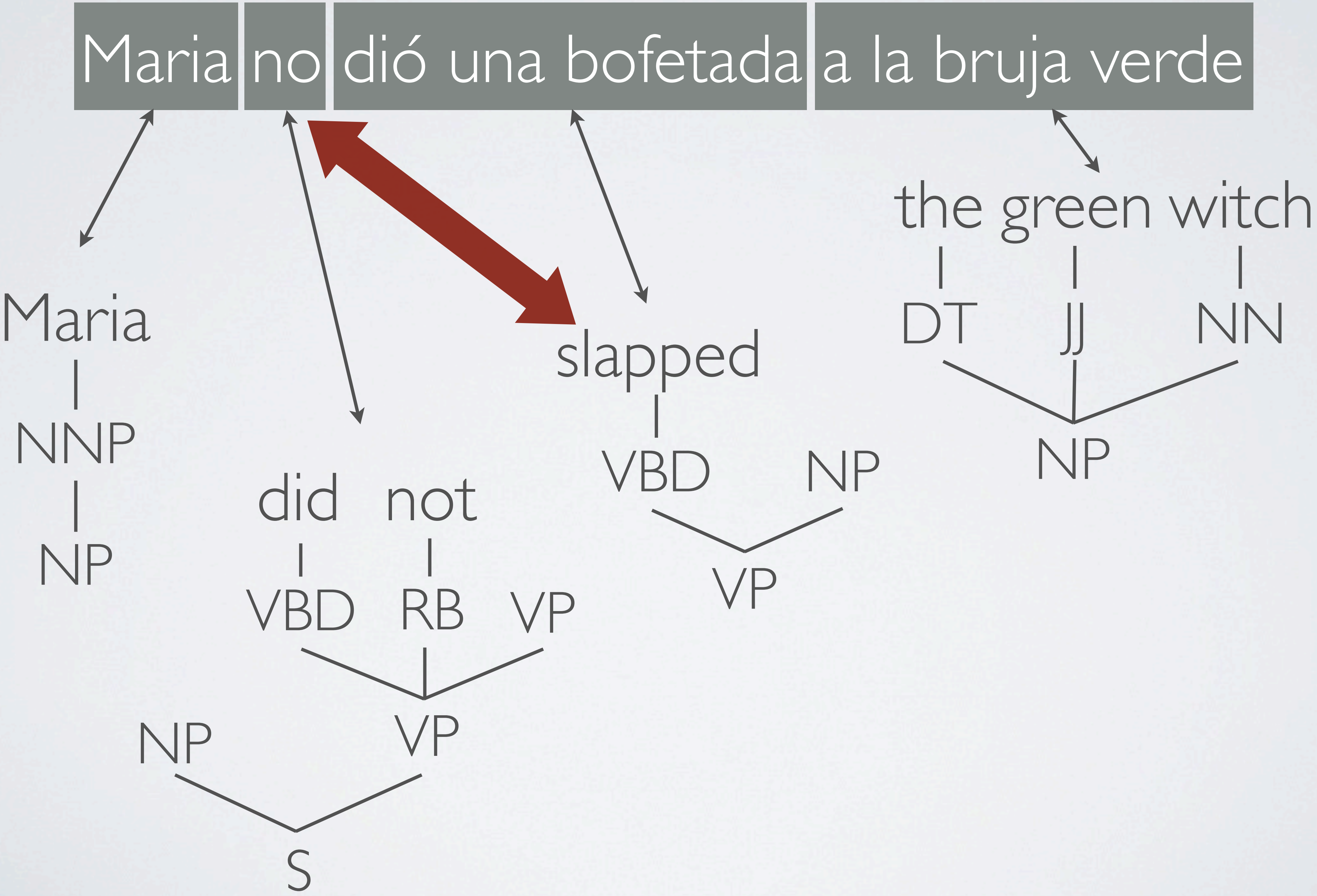
slap
VB NP
VP

the green witch
DT JJ NN
NP

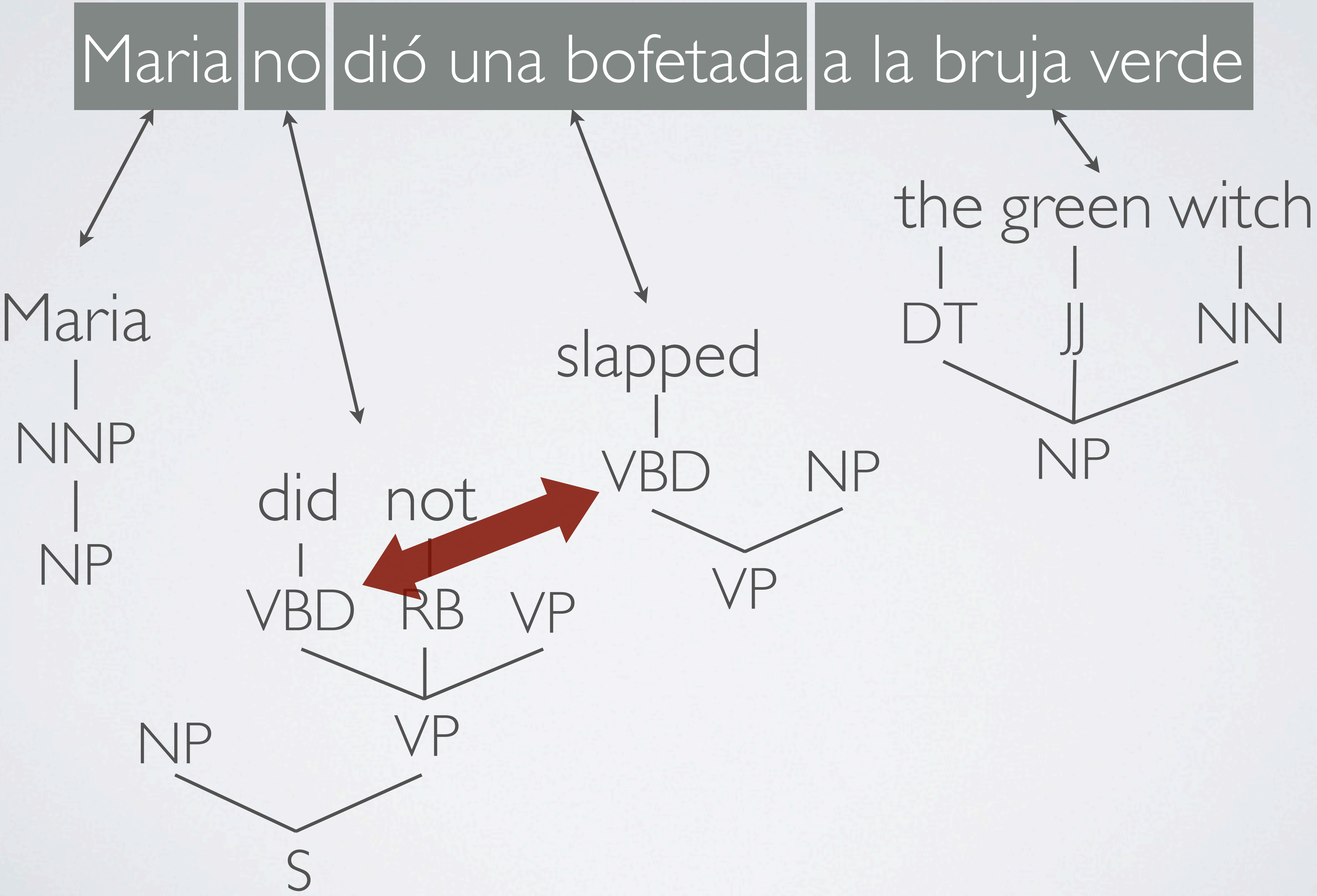
MOTIVATION


Maria
no
dió una bofetada
a la bruja verde
Maria
NNP
NP
did
not
VBD
RB
VP
VP
NP
S
slapped
VBD
NP
VP
the green witch
DT
JJ
NN
NP

MOTIVATION


Maria
no
dió una bofetada
a la bruja verde
Maria
NNP
NP
did
not
VBD
RB
VP
VP
NP
S
slapped
VBD
NP
VP
the green witch
DT
JJ
NN
NP

MOTIVATION



Maria | no | dió una bofetada | a la bruja verde

MOTIVATION

- Minimum error rate training (MERT) works for <30 features
- Margin infused relaxed algorithm (MIRA)
  - Online large-margin discriminative training
  - Scales better to large feature sets
  - Enables freer exploration of features

RESULTS

GALE 2008 Chinese-English data

System	Training	Features	BLEU
Hiero	MERT	11	36.1
	MIRA	10,990	37.6
Syntax	MERT	25	39.5
	MIRA	283	40.6

OVERVIEW

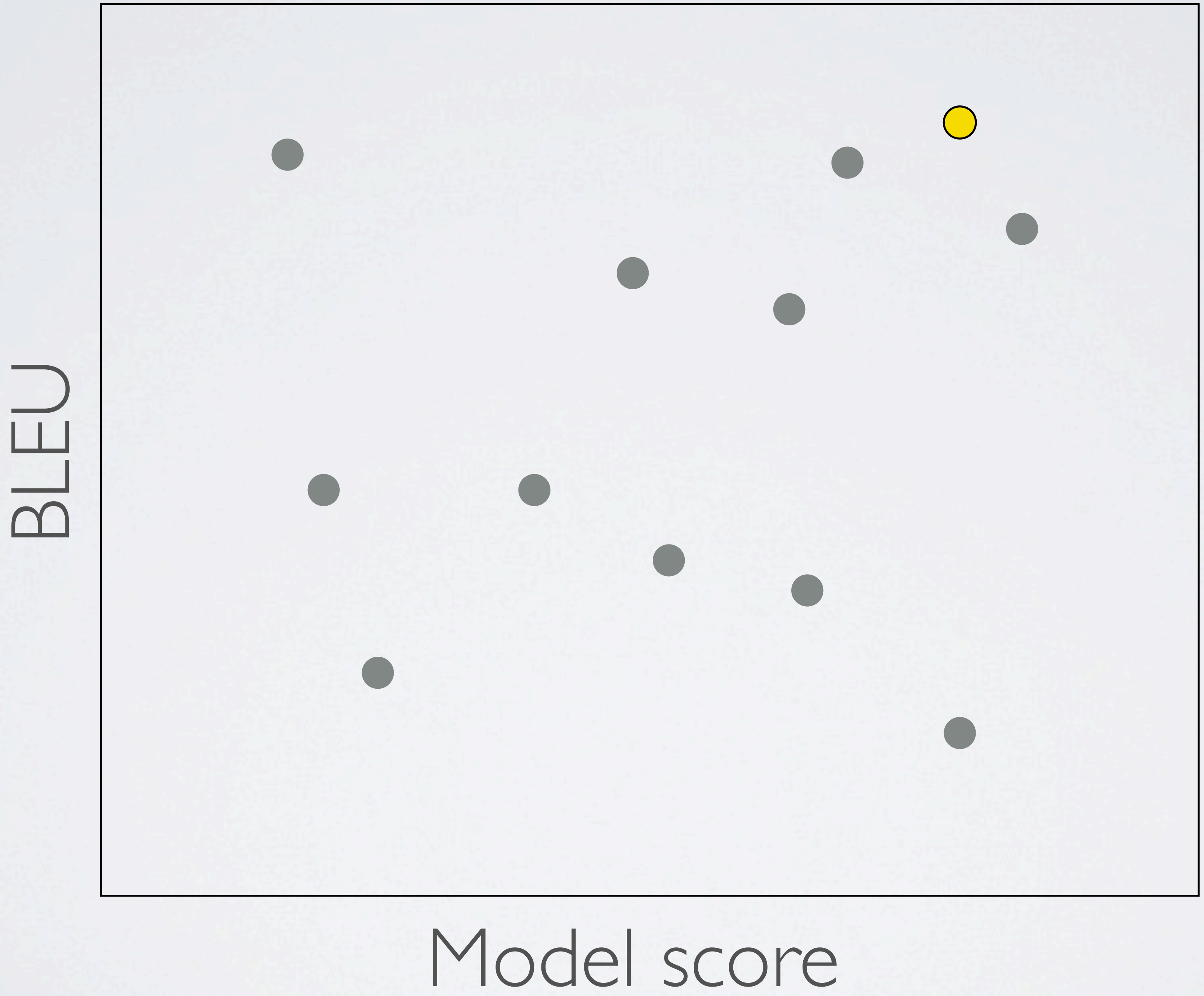
- Training
- Features
- Experiments

Training

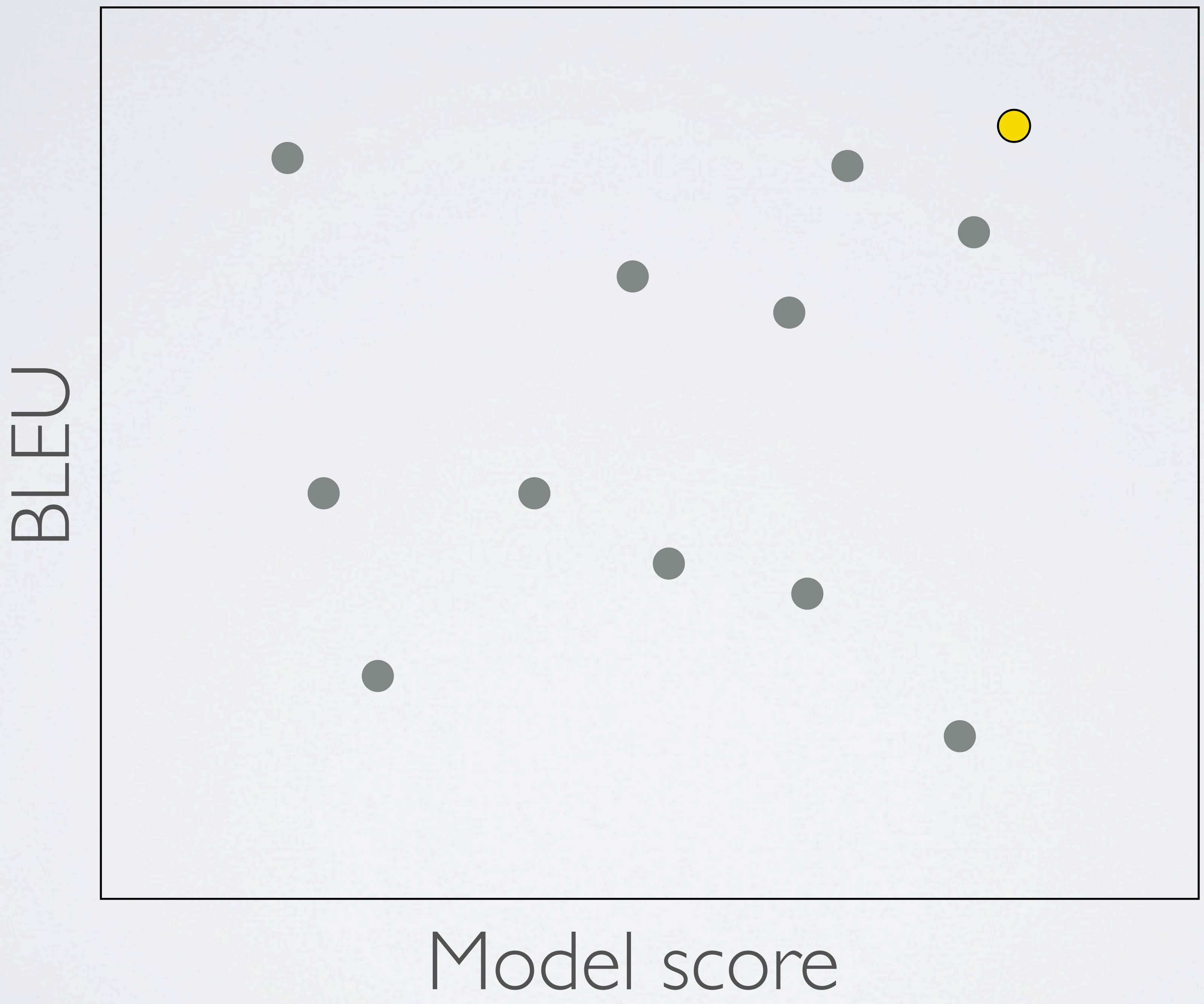
MIRA

- Crammer and Singer, 2003
- Applied to statistical MT by Watanabe et al., 2007
- Chiang, Marton, and Resnik, 2008:
  - use more of the forest
  - parallelize training

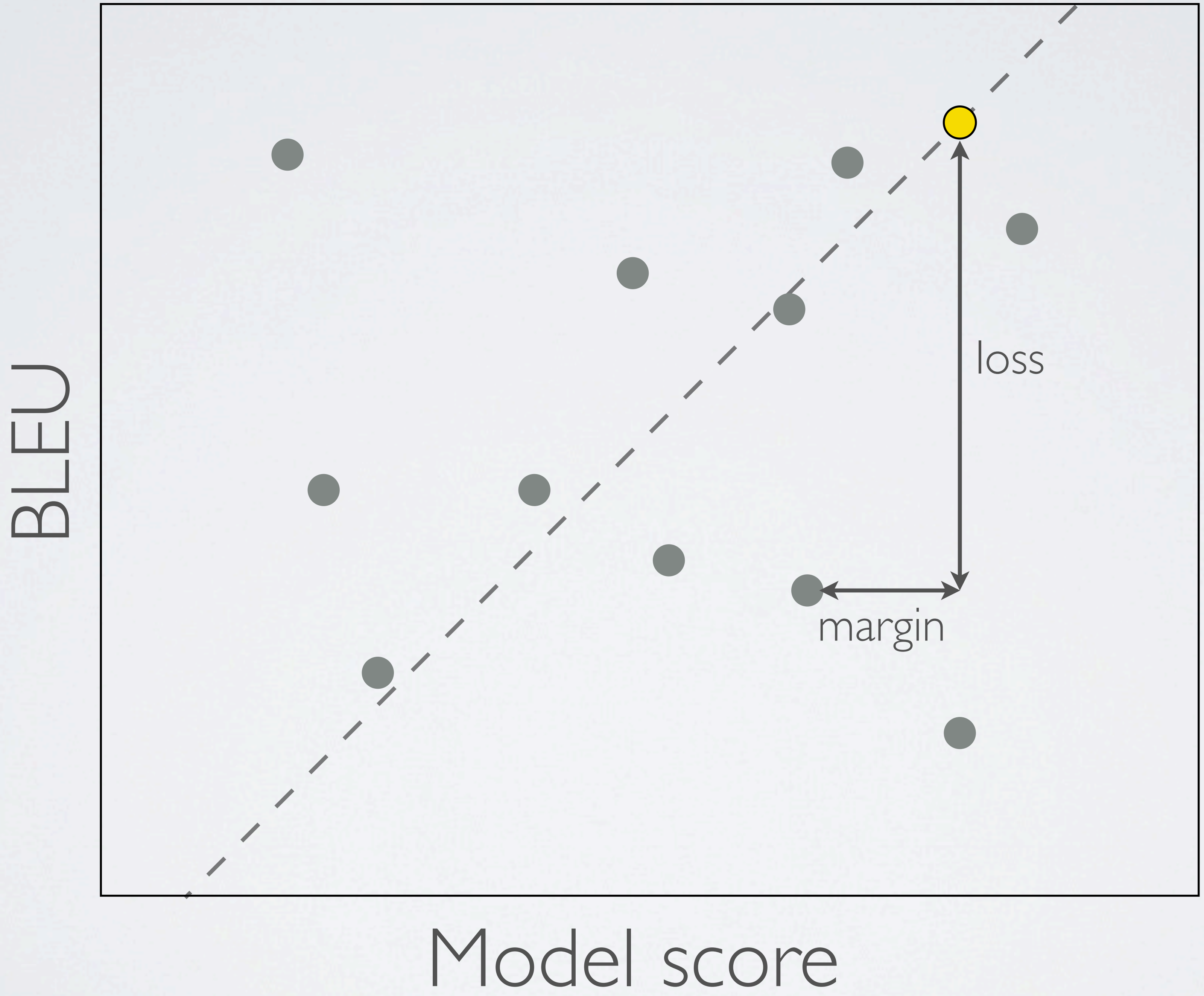
MERT



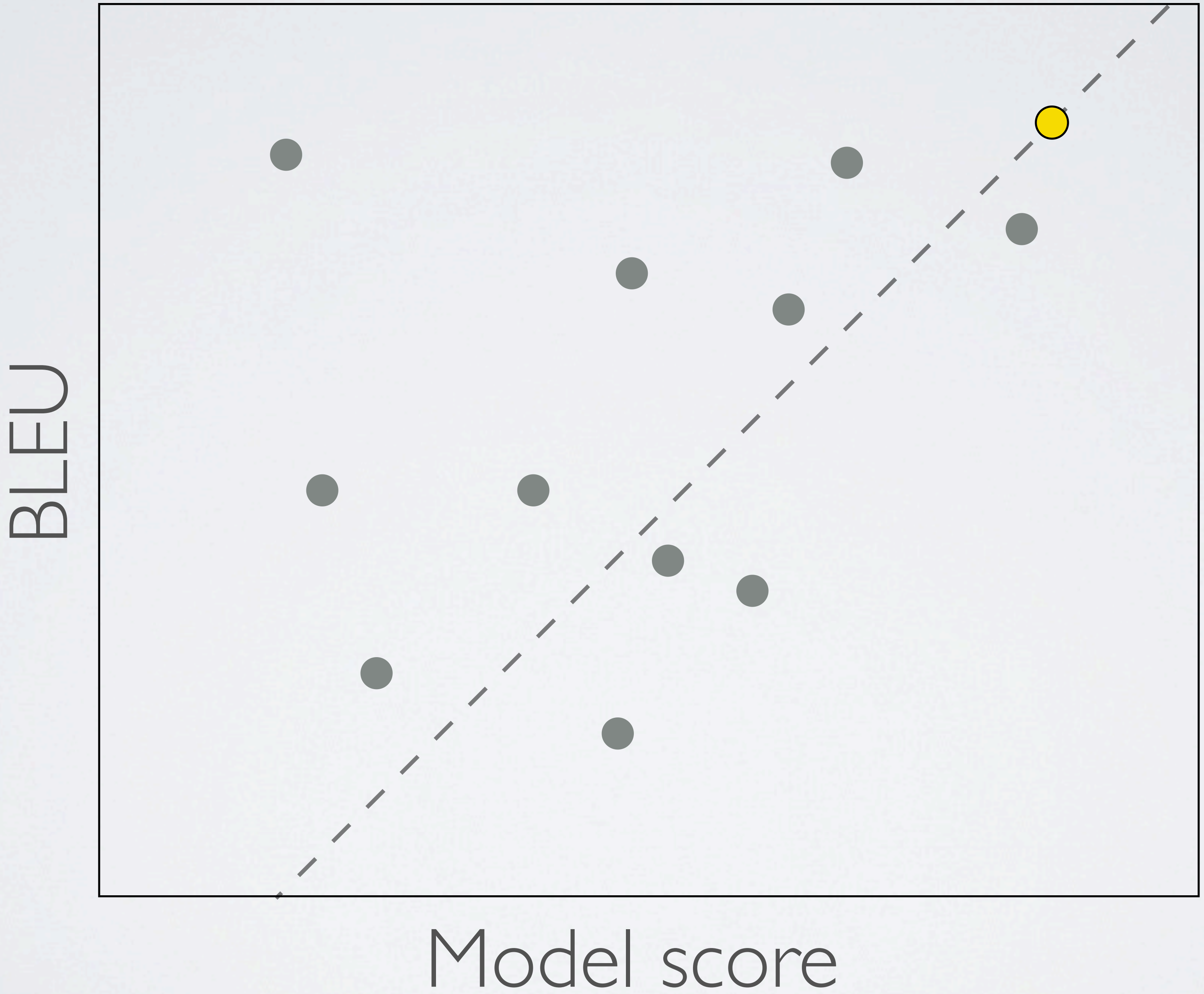
MERT



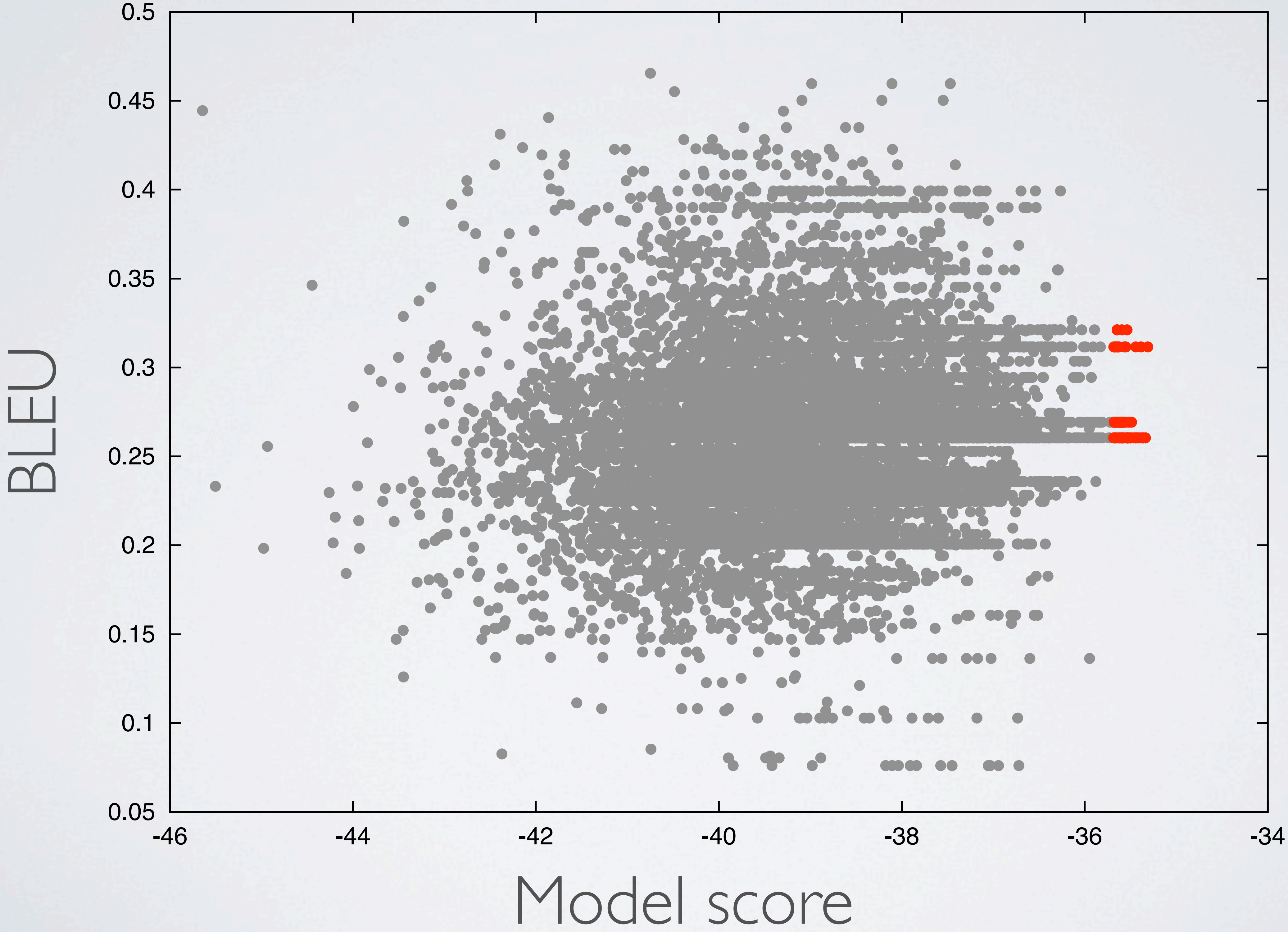
MIRA



MIRA



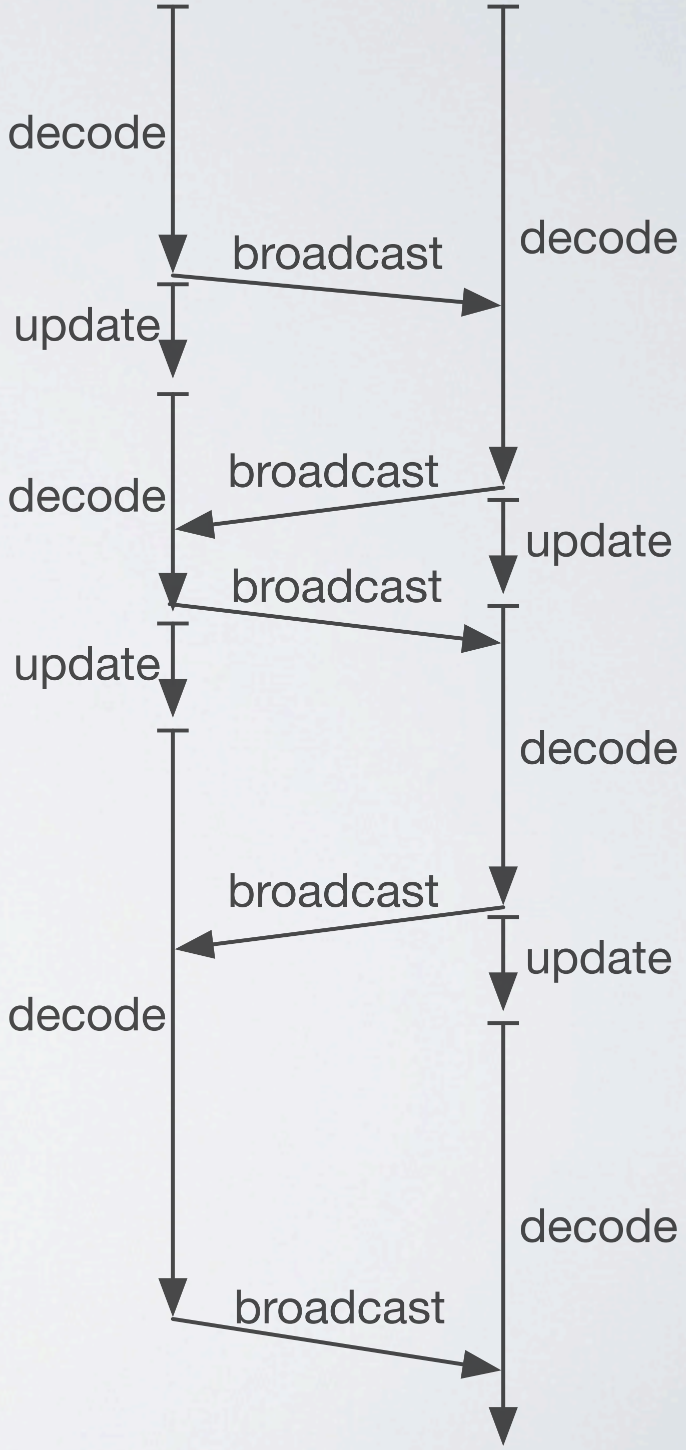
FOREST-BASED TRAINING



PARALLEL TRAINING

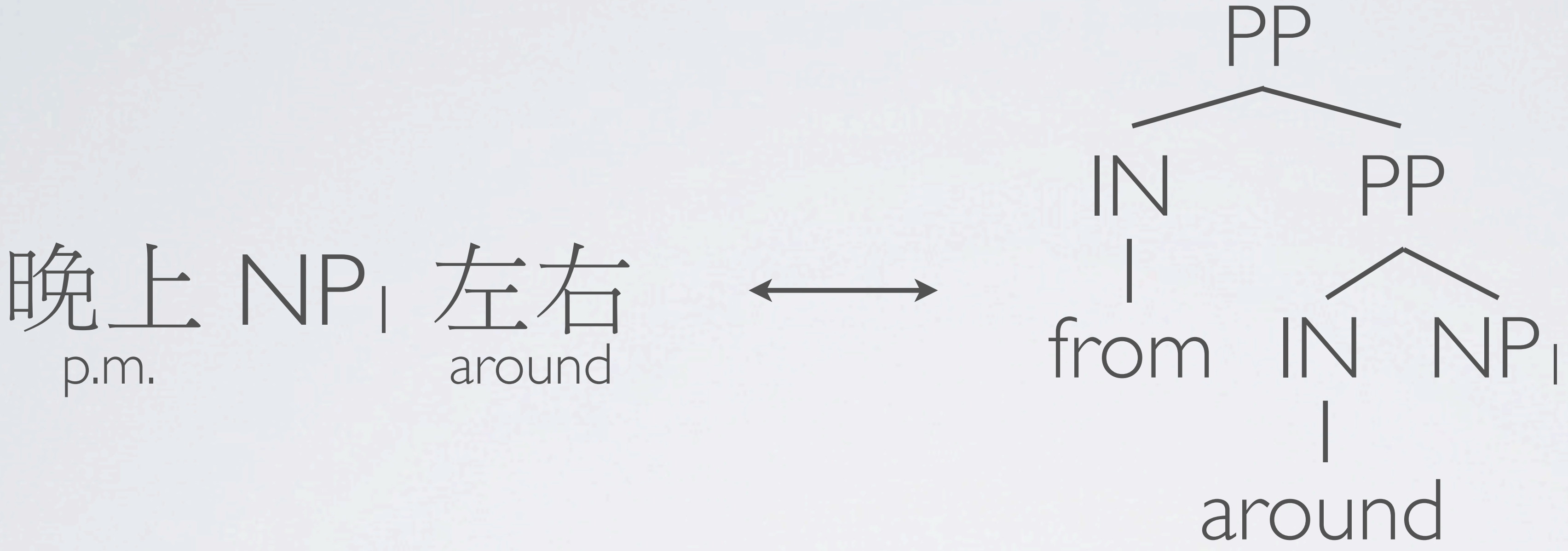
- Run *n* MIRA learners in parallel
- Share information among learners

Hiero	$n = 20$
Syntax	$n = 73$



Features

DISCOUNT FEATURES



晚上 NP_1 左右

p.m. around

⟷

PP

IN PP

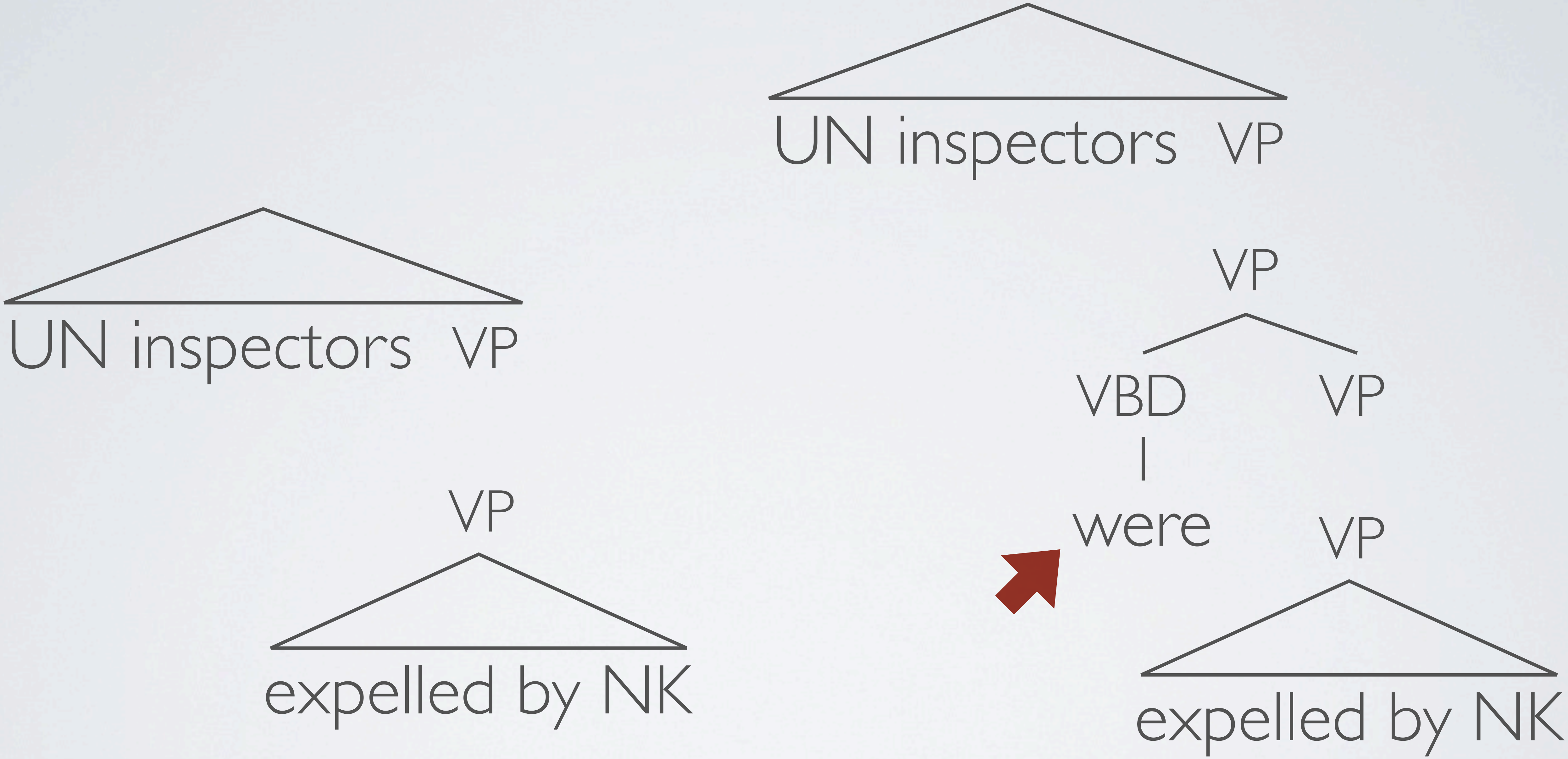
from IN NP_1

around

count=*1*

- Low counts are often overestimates
- Introduce a *count*=*1* feature that fires on 1-count rules, etc.

TARGET SYNTAX FEATURES



UN inspectors VP

VP

expelled by NK

UN inspectors VP

VP

VBD VP

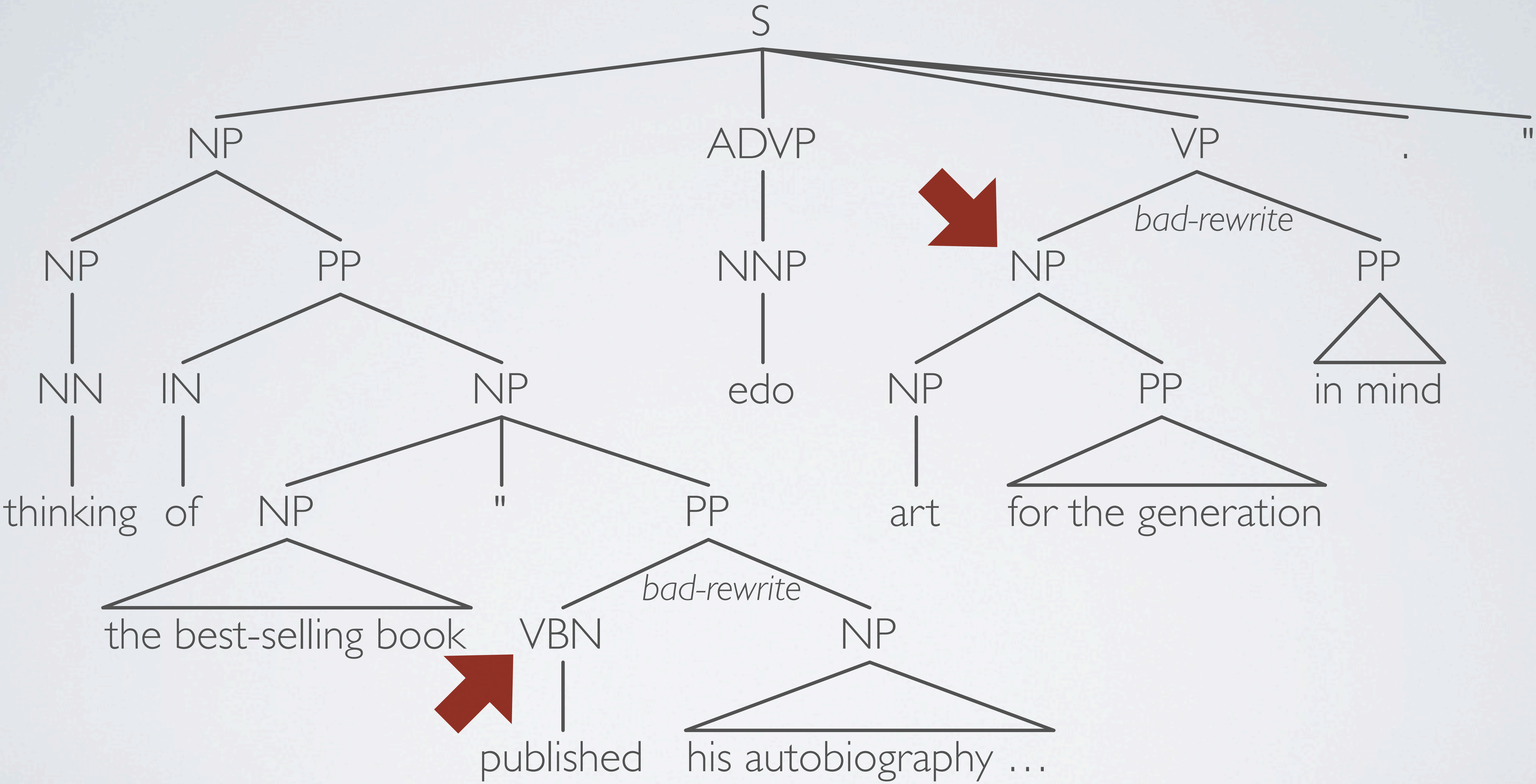
were

VP

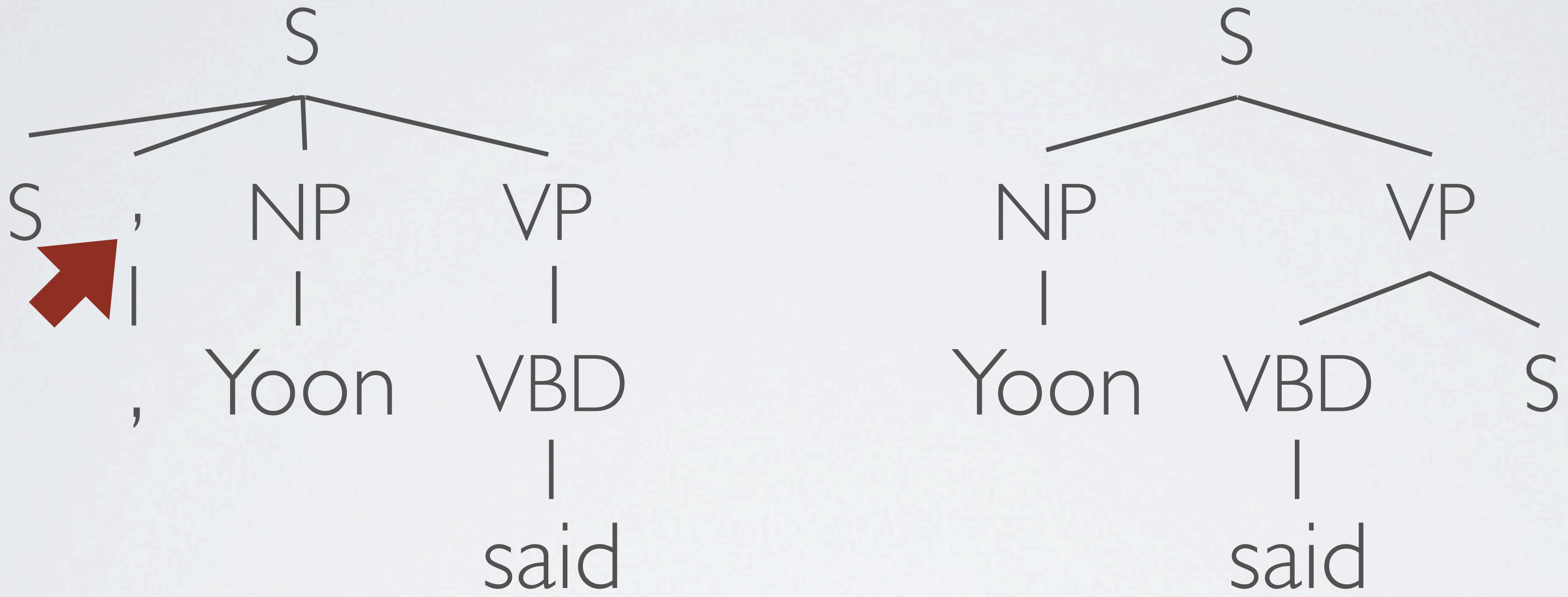
expelled by NK

insert-were

TARGET SYNTAX FEATURES



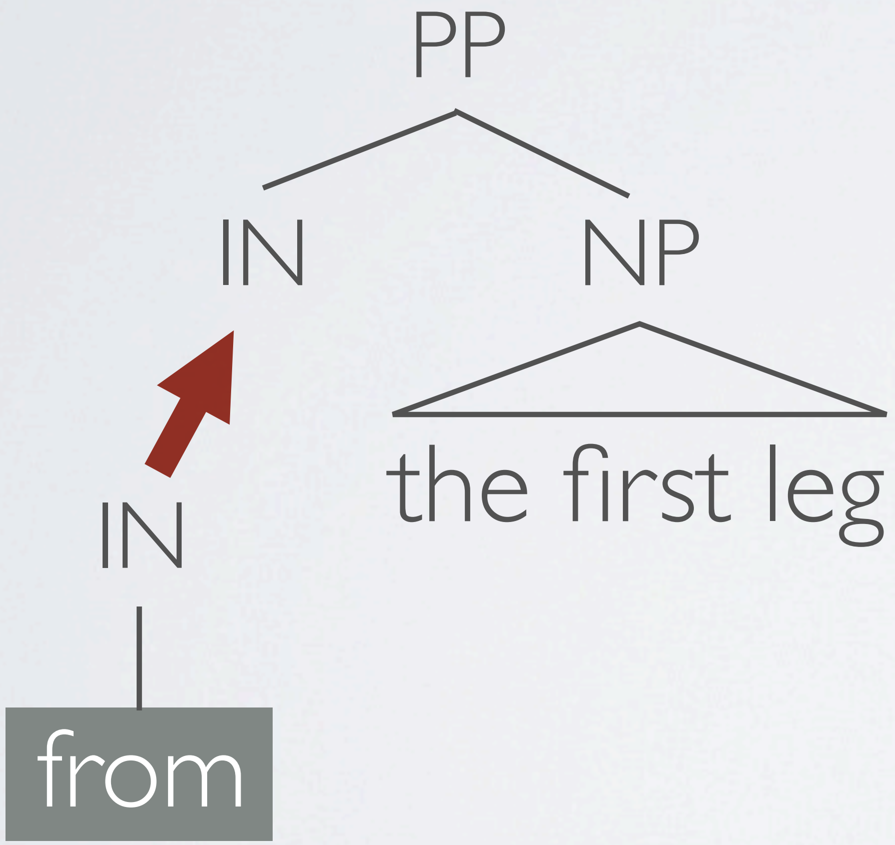
TARGET SYNTAX FEATURES



node=,

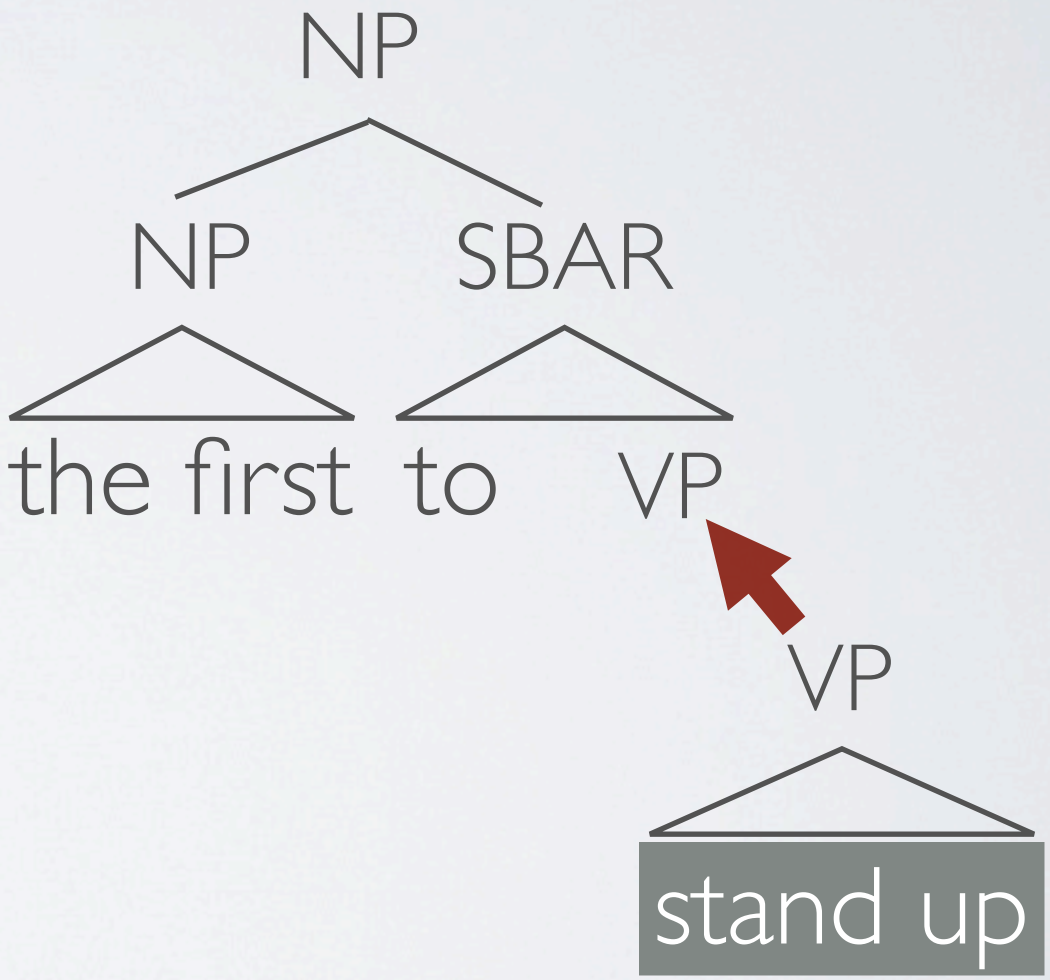
TARGET SYNTAX FEATURES

第一个 站 出来
first stand come out



root=IN

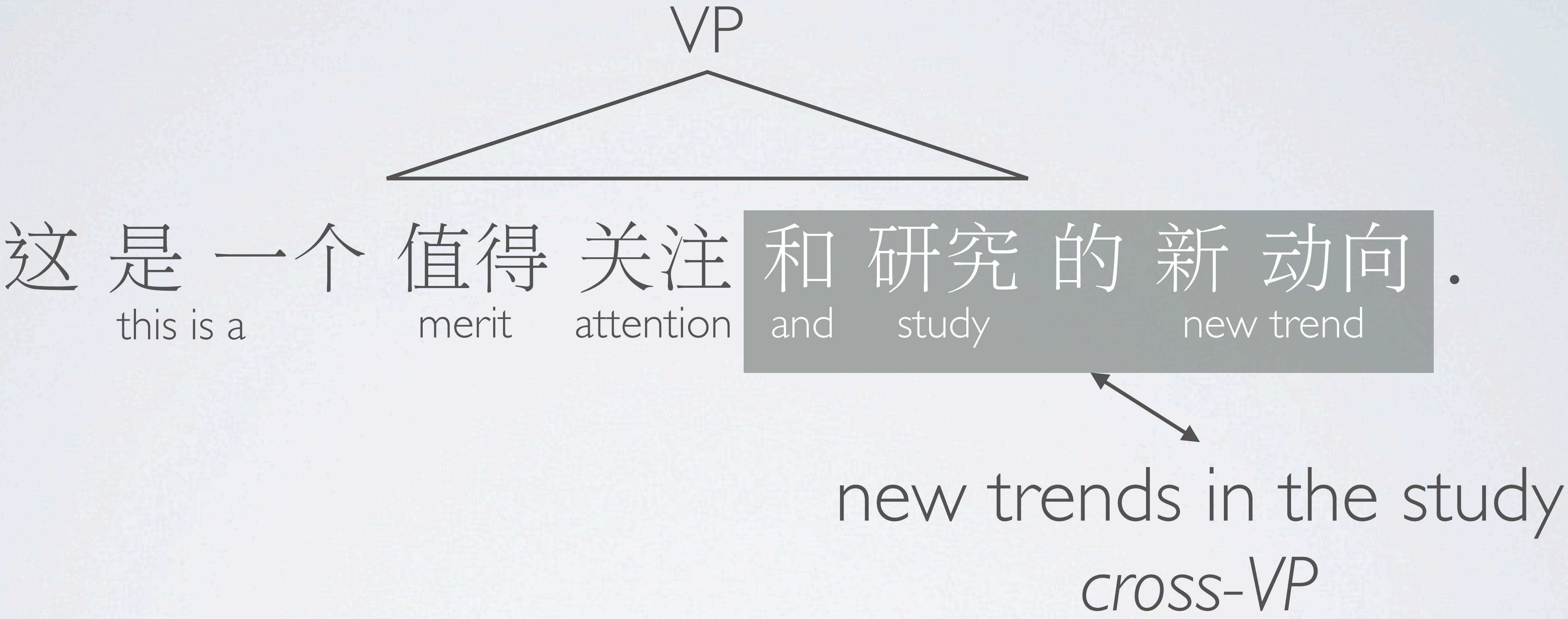
第一个 站 出来
first stand come out



root=VP

SOURCE CONTEXT FEATURES

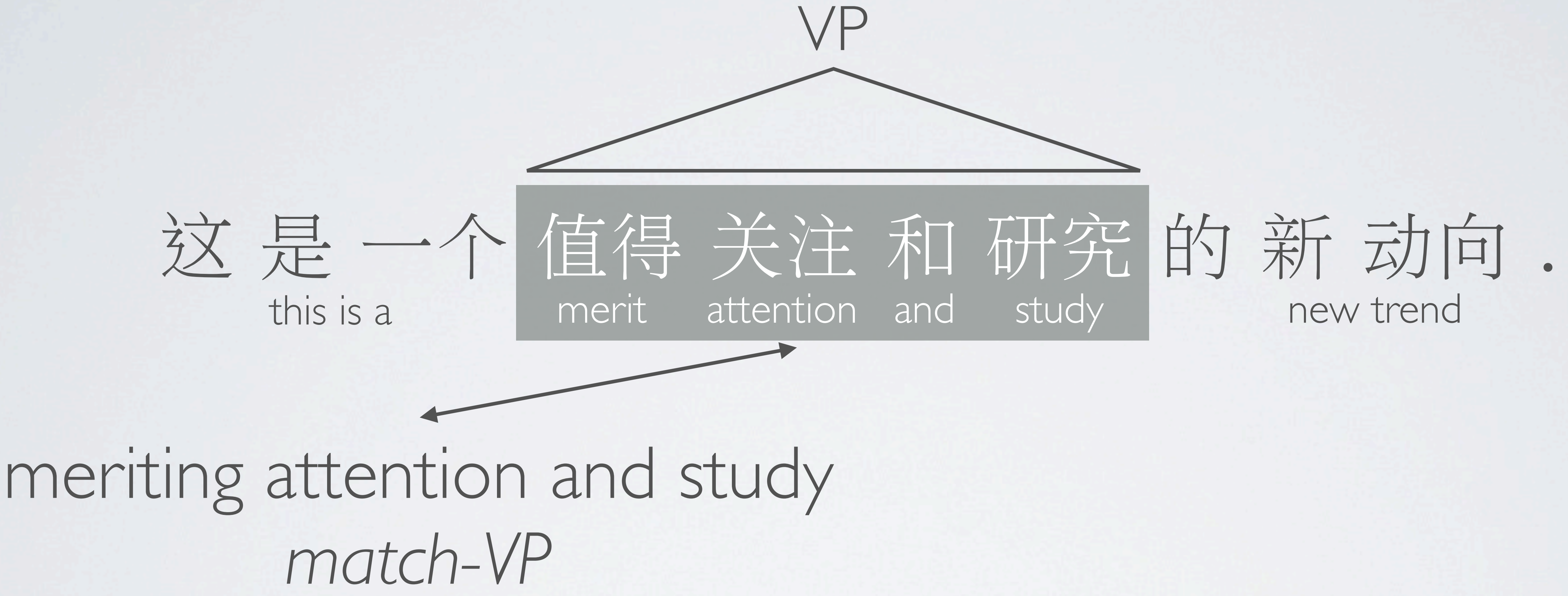
Marton & Resnik 2008; Chiang et al 2008



- Use external parser to infer source-side syntax
- Rewards and penalties for matching/crossing brackets

SOURCE CONTEXT FEATURES

Marton & Resnik 2008; Chiang et al 2008



VP

这 是 一个 值得 关注 和 研究 的 新 动向 .

this is a merit attention and study new trend

meriting attention and study

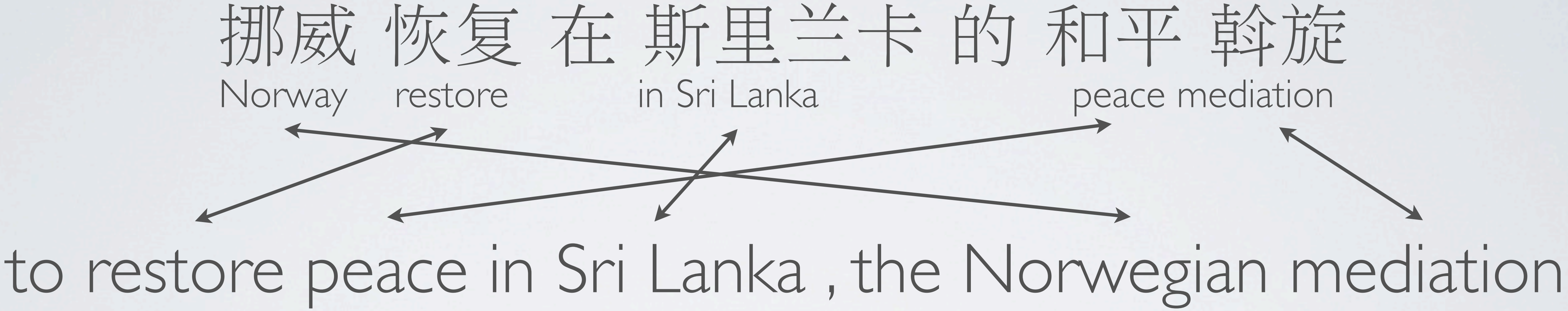
match-VP

- Use external parser to infer source-side syntax

- Rewards and penalties for matching/crossing brackets

SOURCE CONTEXT FEATURES

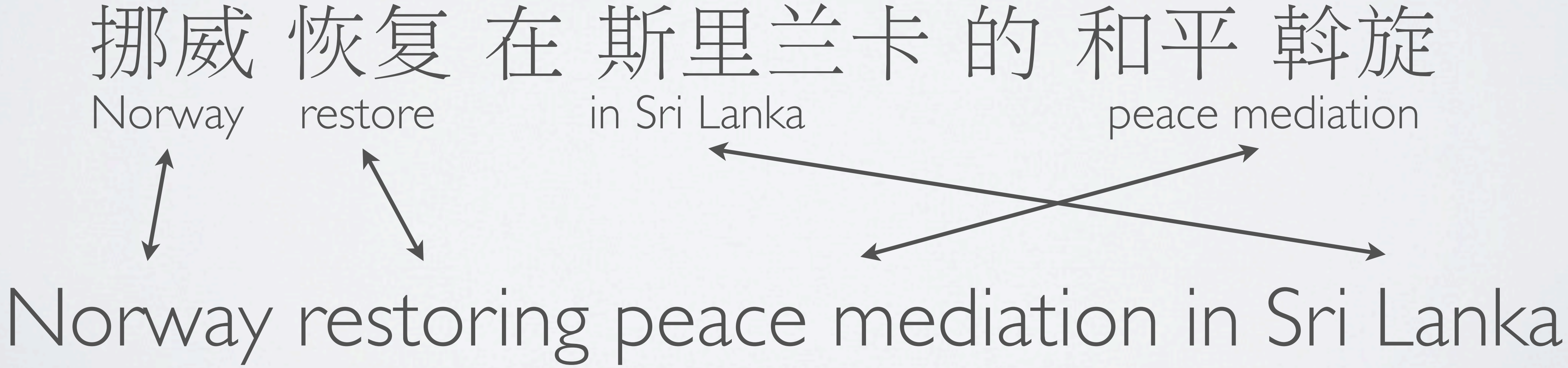
Chiang et al 2008



挪威 恢复 在 斯里兰卡 的 和平 斡旋

Norway restore in Sri Lanka peace mediation

to restore peace in Sri Lanka , the Norwegian mediation



挪威 恢复 在 斯里兰卡 的 和平 斡旋

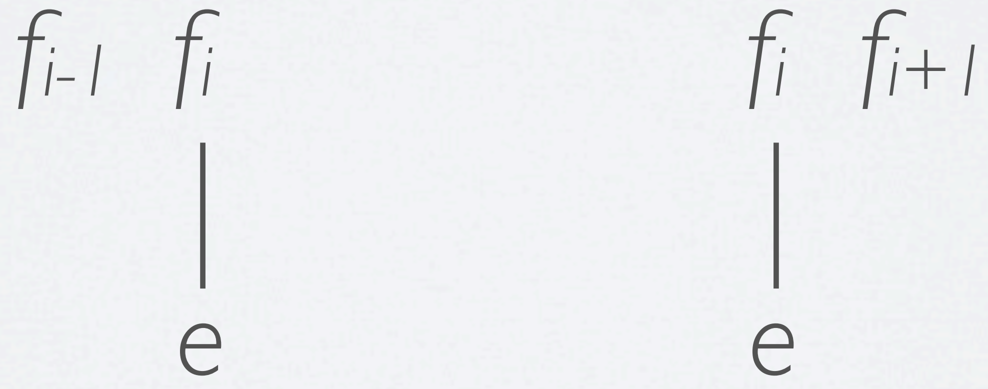
Norway restore in Sri Lanka peace mediation

Norway restoring peace mediation in Sri Lanka

SOURCE CONTEXT FEATURES

- Word context features: similar to Watanabe et al. 2007 and work on WSD in MT (Chan et al. 2007, Carpuat & Wu 2007)

- Relate a word’s translation with its left or right neighbor on the source side (just the 100 most frequent types)



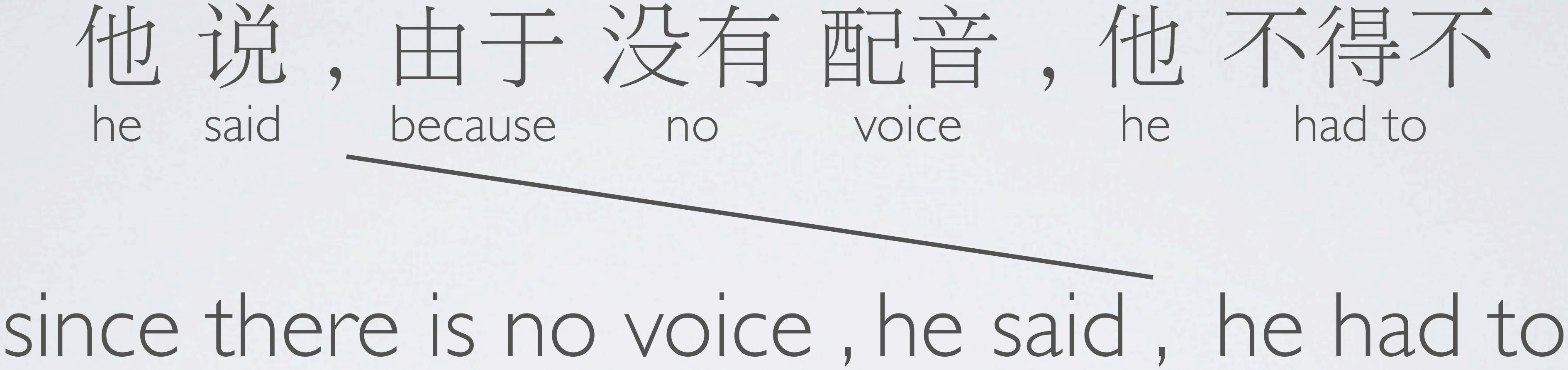
f_{i-1} f_i

e

f_i f_{i+1}

e

SOURCE CONTEXT FEATURES

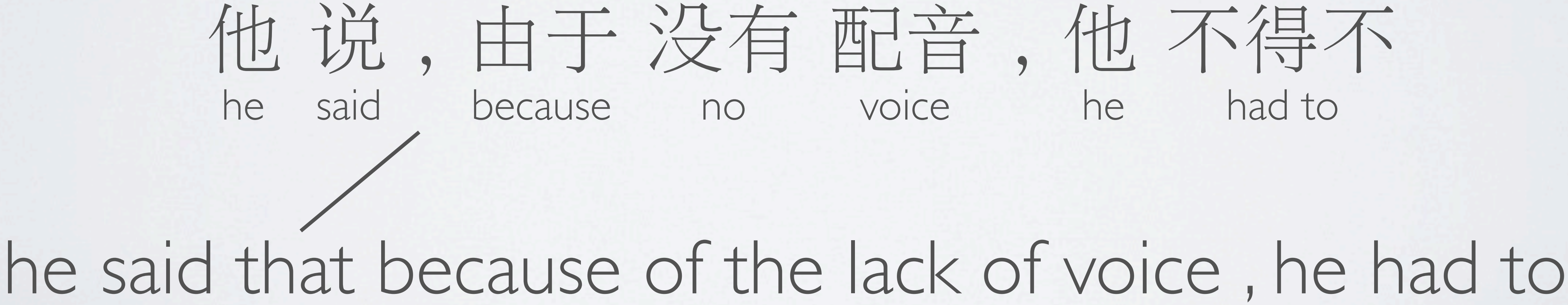


他 说 , 由于 没有 配音 , 他 不得不

he said because no voice he had to

since there is no voice , he said , he had to

f_i=, & f_{i-1}=说 & e=,



他 说 , 由于 没有 配音 , 他 不得不

he said because no voice he had to

he said that because of the lack of voice , he had to

f_i=, & f_{i-1}=说 & e=*that*

Experiments

TRAINING DATA

GALE 2008 Chinese-English data

	Hiero	Syntax
Parallel data	260M	65M
Language model	2G	1G
MERT/MIRA	58k	58k
Test	57k	57k

RESULTS (HIERO)

Chinese-English

Training	Features	#	BLEU
MERT	baseline	11	36.1
MIRA	+source-side syntax +distortion	56	36.9
	+discount	61	37.3
	+word context	10,990	37.6

RESULTS (SYNTAX)

Chinese-English

Training	Features	#	BLEU
MERT	baseline	25	39.5
MIRA	baseline	25	39.8
	rule overlap	132	39.9
	node count	136	40.0
	+discount +bad rewrite +insertion	283	40.6

CONCLUSIONS

- Using underutilized information for new features:
  - Source context is computationally efficient
  - Target syntax provides a rich structure
- MIRA is working well on new features, systems, languages