Forest MIRA

Forest rescoring in Joshua for MIRA training

Team Members

Matt Post (JHU)

Gideon Wenniger (Amsterdam)

Forest MIRA

Margin-infused relaxed algorithm

Crammer & Singer (2003, 2006)

Large scale discriminative tuner; maximizes model score difference between "hope" and "fear" translations

Chiang et al. (2008, 2009); Chiang (2012)

Batch implementation in Moses (`kbmira`)

Cherry & Foster (2012)

MIRA

Works by maximizing the margin between **hope** and **fear** items

hope items maximize

model score + λ BLEU score

fear items maximize

model score – λ BLEU score

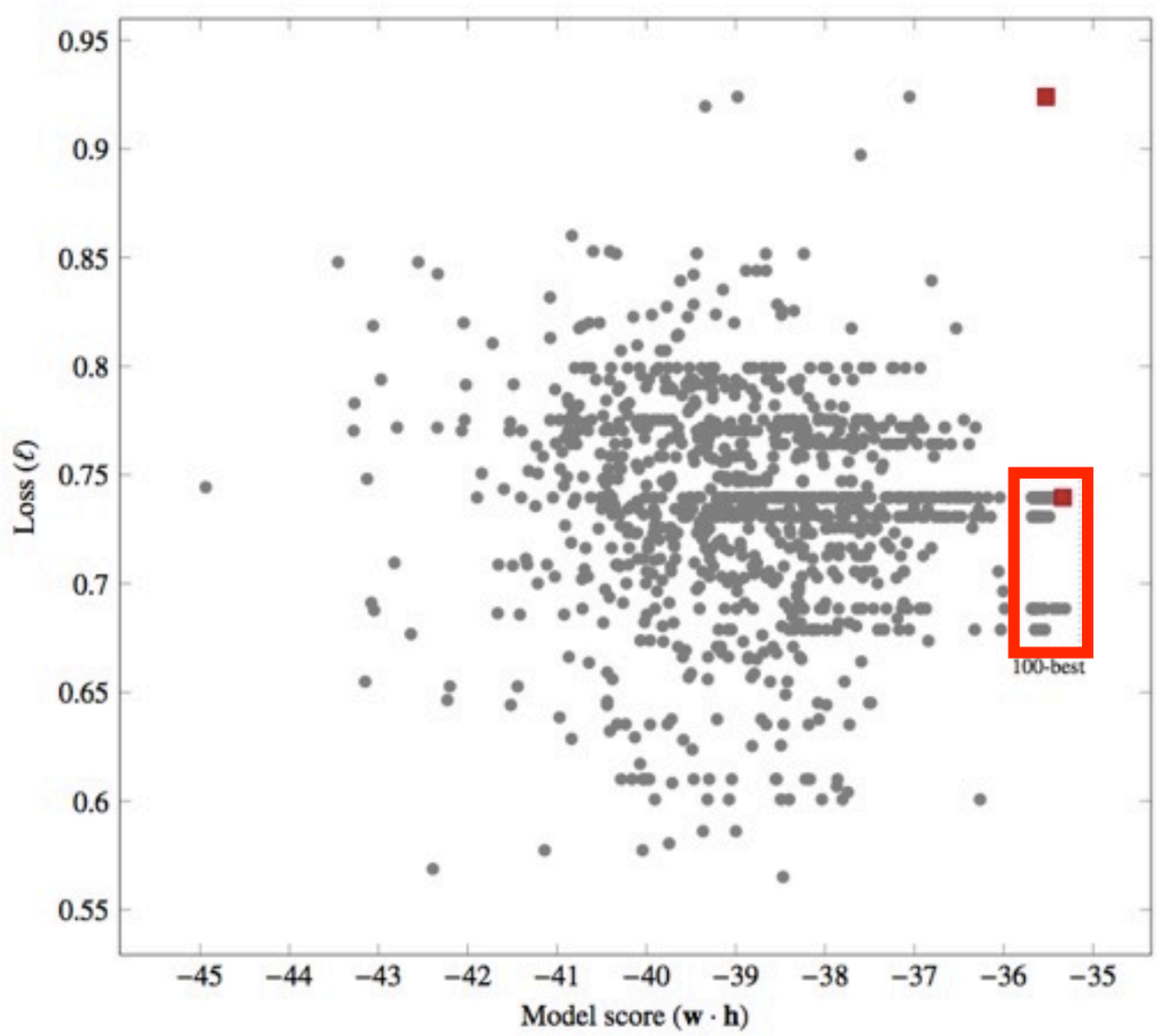
Problem

Moses' `kbmira` extracts hope and fear items from the k-best list

Visualization

(Chiang, 2012)

The k-best list is not representative of the model space!

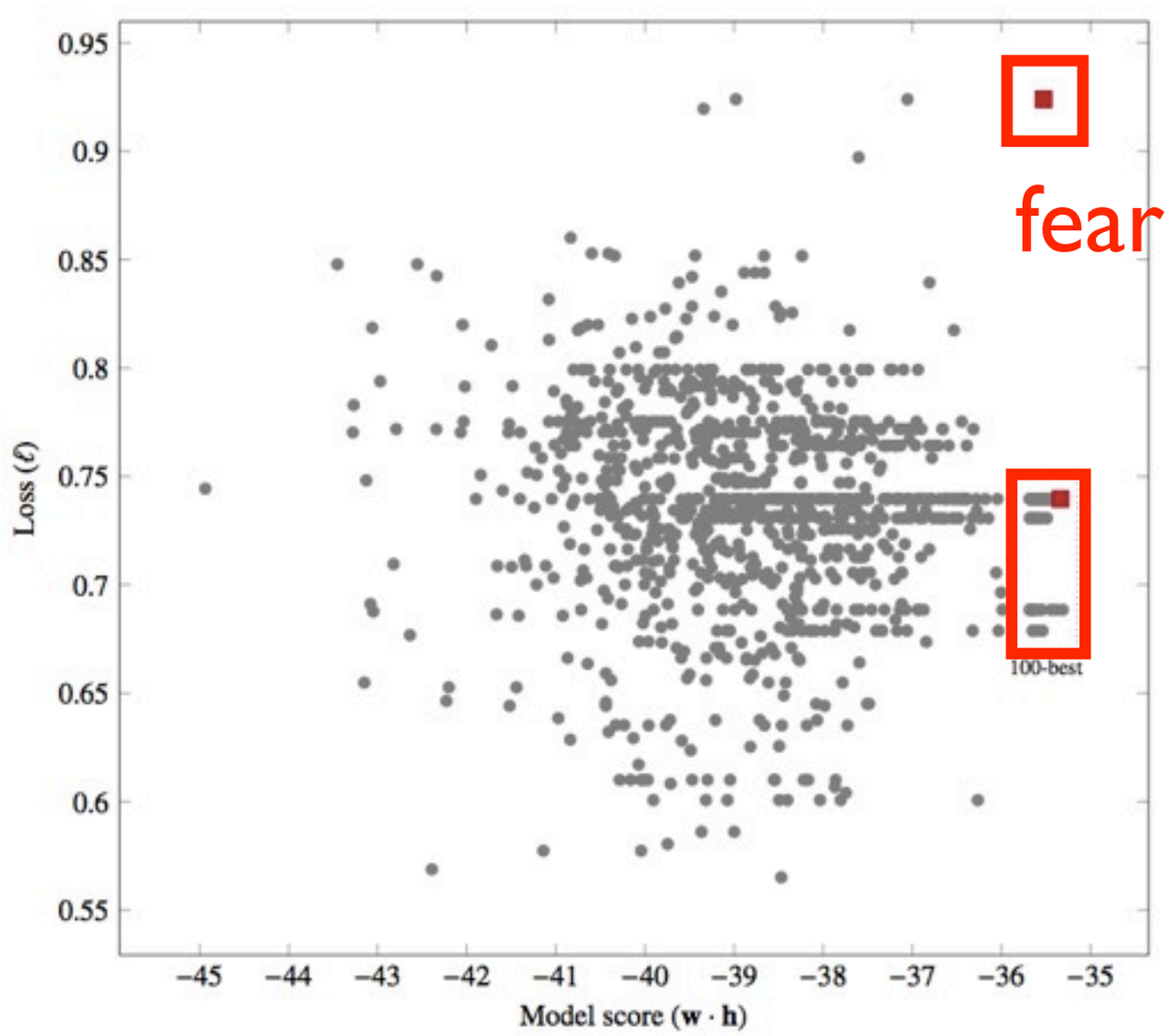


100 best

Visualization

(Chiang, 2012)

The k-best list is not representative of the model space!



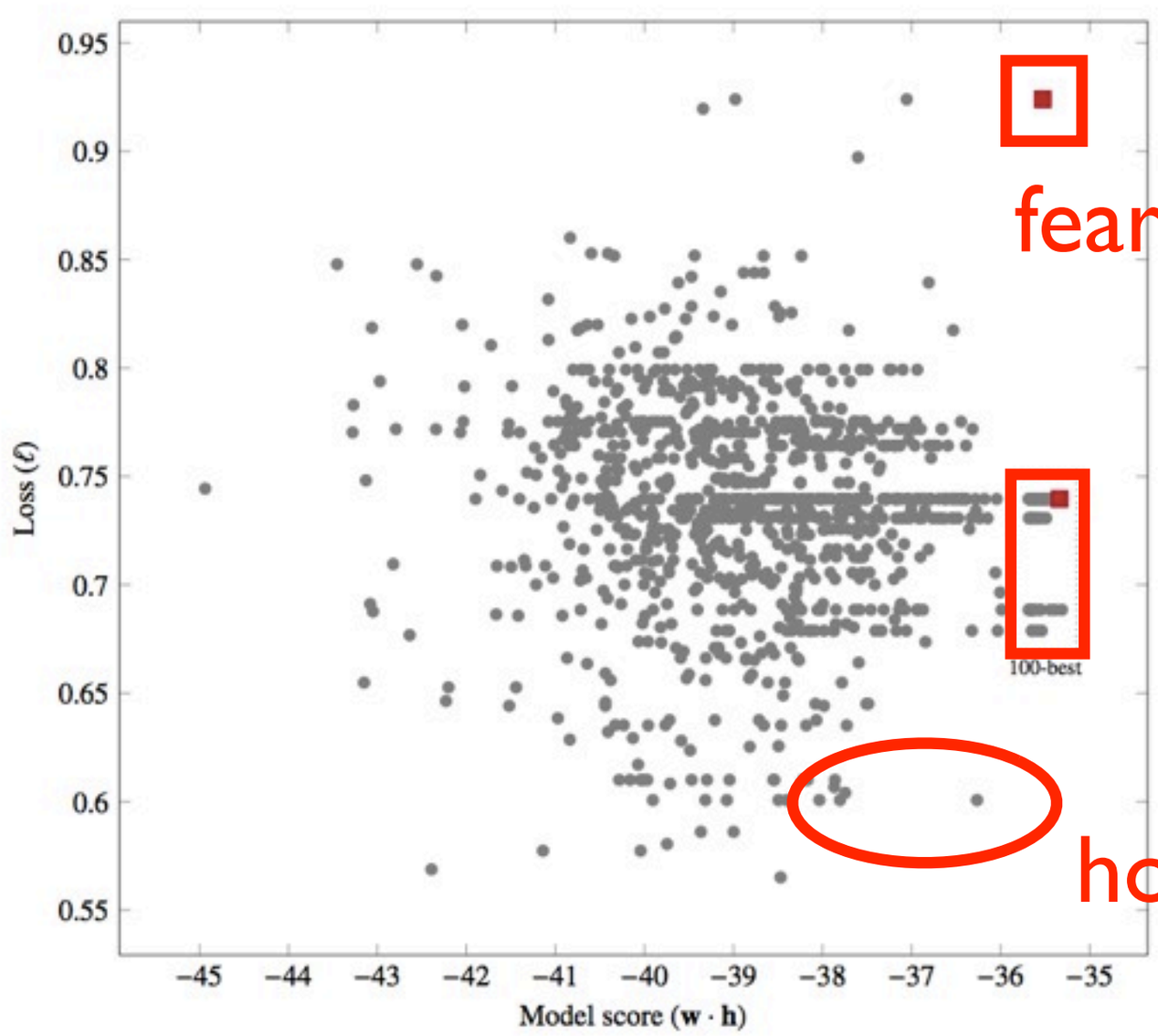
fearier

100 best

Visualization

(Chiang, 2012)

The k-best list is not representative of the model space!



fearier

100 best

hopier?

Goal

Implement forest rescoring in Joshua (Moses) for proper hope/fear updating in Cherry & Foster's kbmira implementation

Sub-goals

- BLEU approximation that factors over the hypergraph
- Forest rescoring and extraction
- Write many-featured feature function

Factorizable BLEU

$$\text{BP} \times \exp\left(\sum_{n=1}^{n} \frac{1}{4} \log p_n\right)$$

Don't clip counts for each p_n

Scale reference length to proportion of input consumed

$\lambda =$ (span width) / (sentence length)

BP = elen / λ • reflen

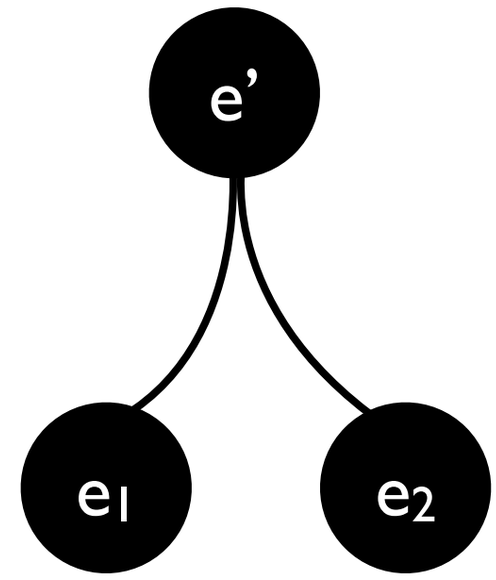
Scoring hyperedges

Decoder combines chart items by adding model scores (vectors of scores)

$$\text{score}(e') = \text{score}(e_1) + \text{score}(e_2)$$

Doesn't work with BLEU!

$$\text{BLEU}(e') \neq \text{BLEU}(e_1) + \text{BLEU}(e_2)$$



~BLEU

Instead, we have to store the sufficient statistics on each hyperedge

	n=1	n=2	n=3	n=4	len
counts	2	1	0	0	2

These we can sum together and then use to compute ~BLEU

Computing ~BLEU

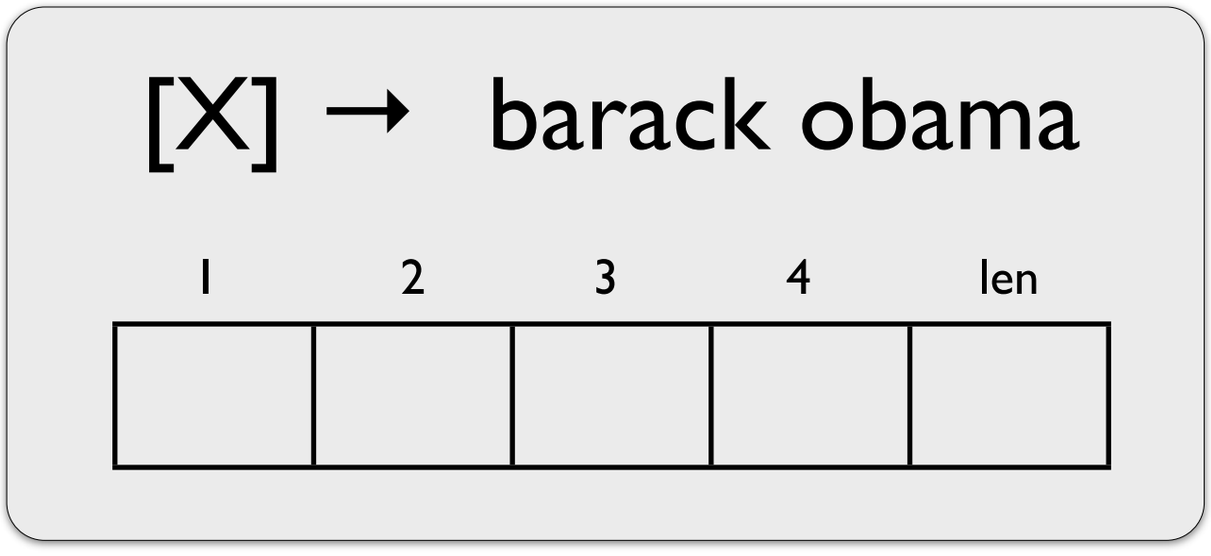
Terminal productions: accumulate all ngrams

... barack obama , president ...

Computing ~BLEU

Terminal productions: accumulate all ngrams

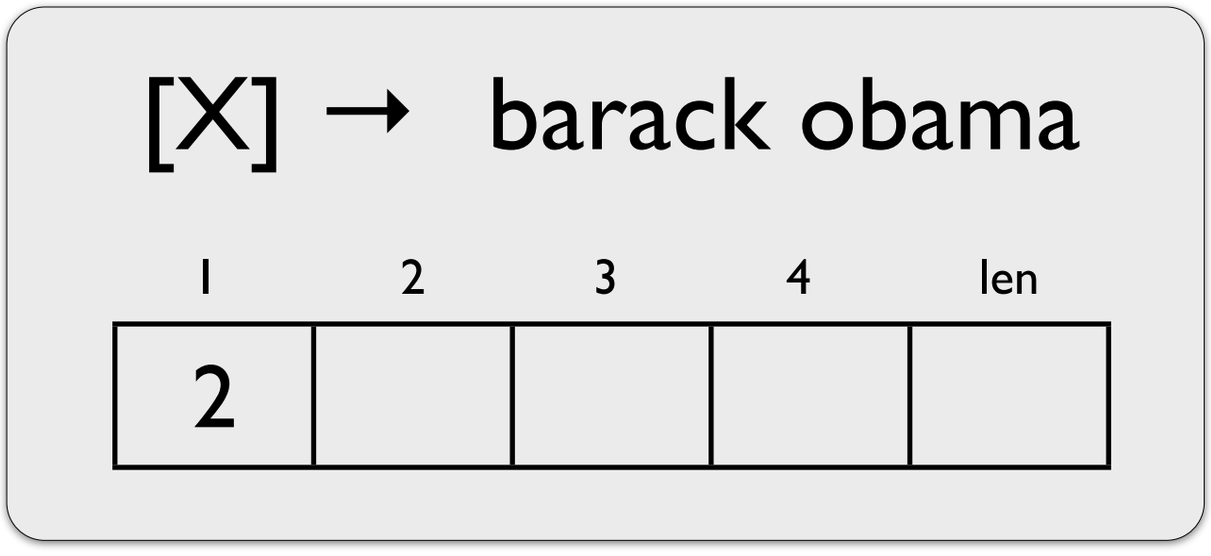
... barack obama , president ...



Computing ~BLEU

Terminal productions: accumulate all ngrams

... barack obama , president ...



Computing ~BLEU

Terminal productions: accumulate all ngrams

... barack obama , president ...

[X] → barack obama

1	2	3	4	len
2	1			

Computing ~BLEU

Terminal productions: accumulate all ngrams

... barack obama , president ...

[X] → barack obama

1	2	3	4	len
2	1	0		

Computing ~BLEU

Terminal productions: accumulate all ngrams

... barack obama , president ...

[X] → barack obama

1	2	3	4	len
2	1	0	0	

Computing ~BLEU

Terminal productions: accumulate all ngrams

... barack obama , president ...

[X] → barack obama

1	2	3	4	len
2	1	0	0	2

Computing ~BLEU

Mixed productions: accumulate all ngrams that overlap tail node boundaries

Computing ~BLEU

... barack obama , president ...

[X] → [X] , president

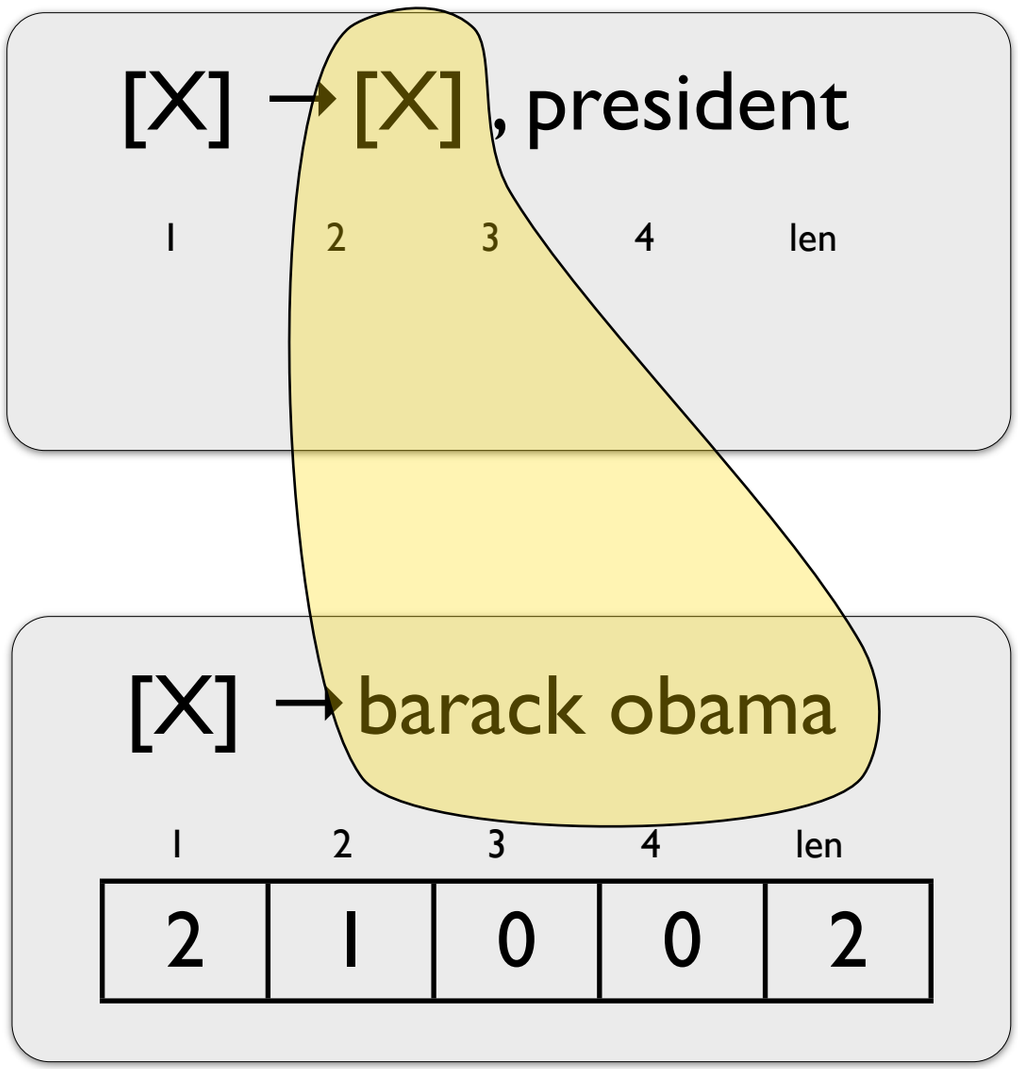
1	2	3	4	len

[X] → barack obama

1	2	3	4	len
2	1	0	0	2

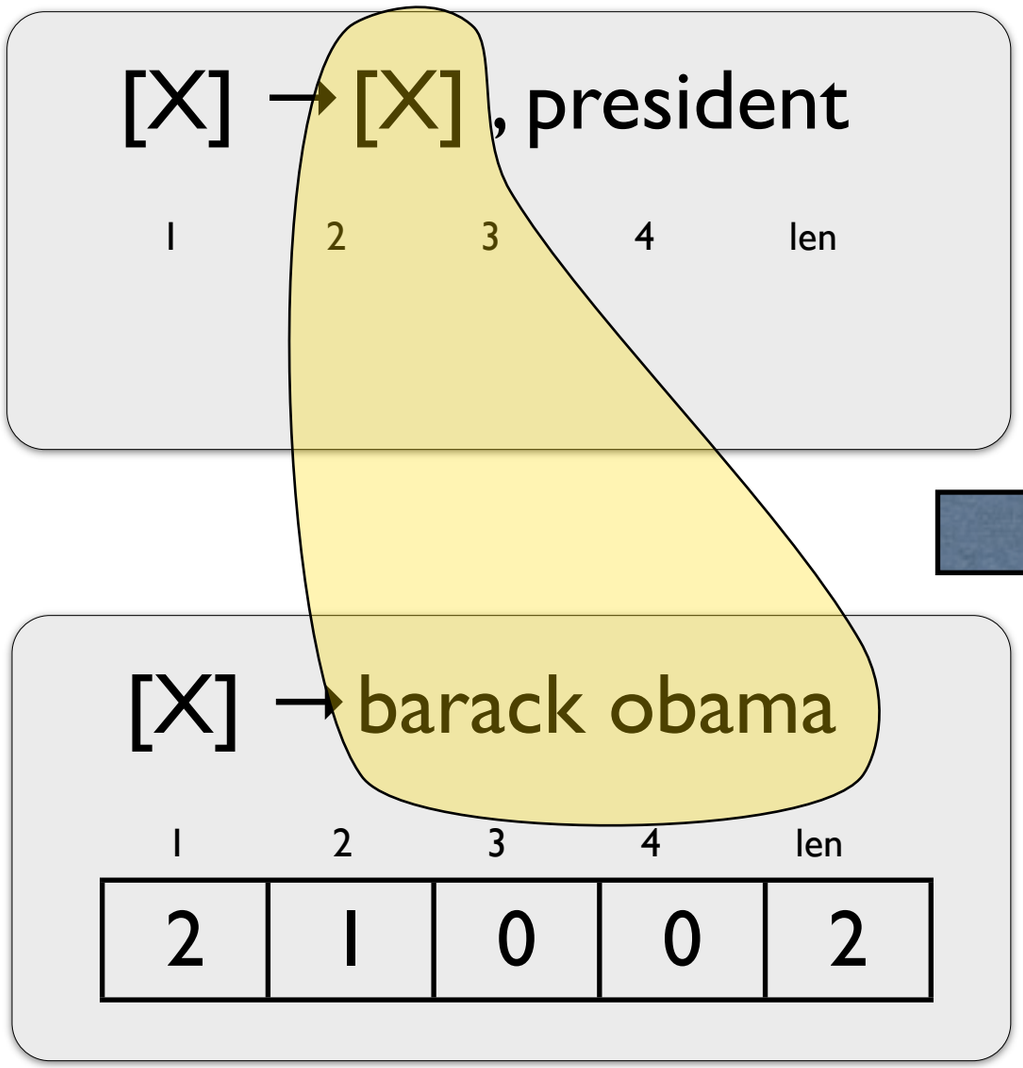
Computing ~BLEU

... barack obama , president ...



Computing ~BLEU

... barack obama , president ...

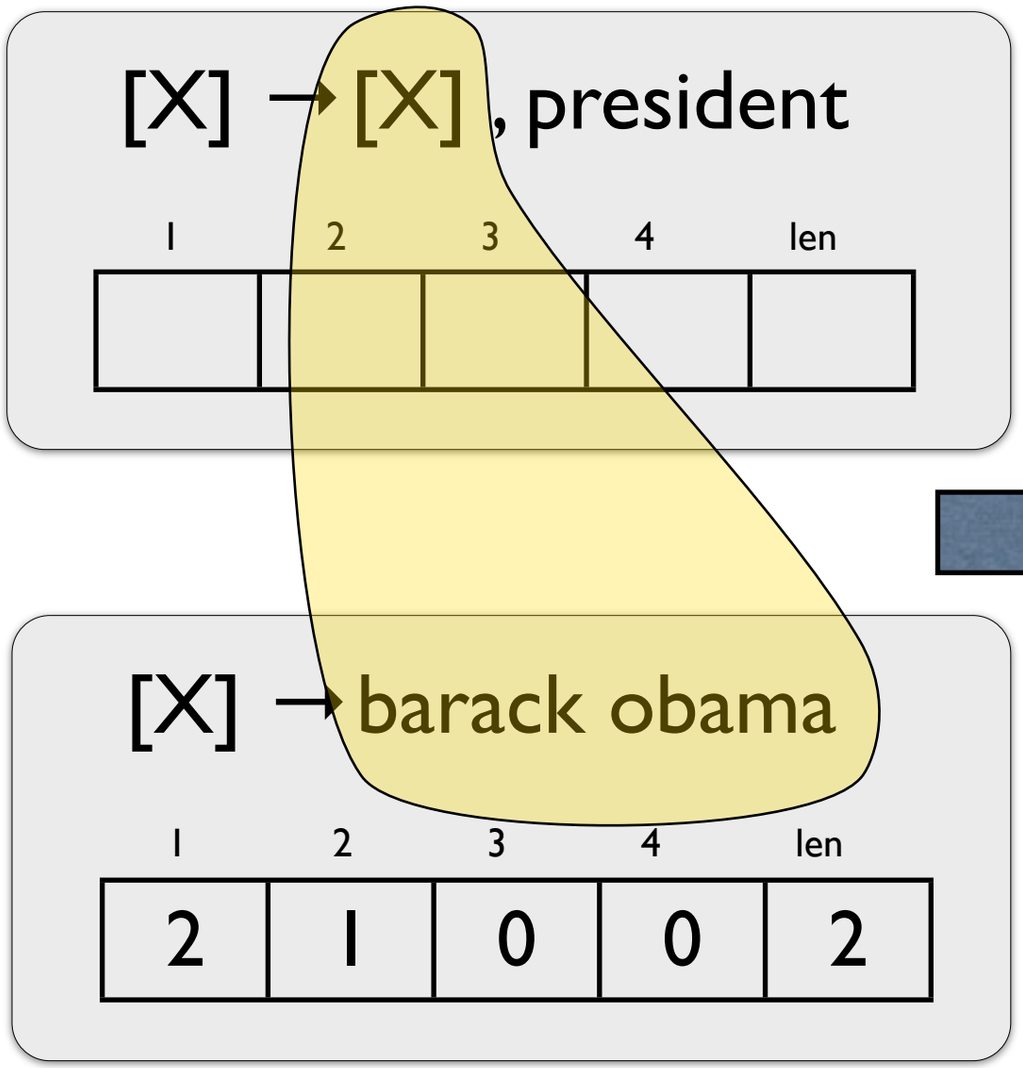


barack obama | , president

We count only ngrams in the new rule or crossing the | boundary

Computing ~BLEU

... barack obama , president ...

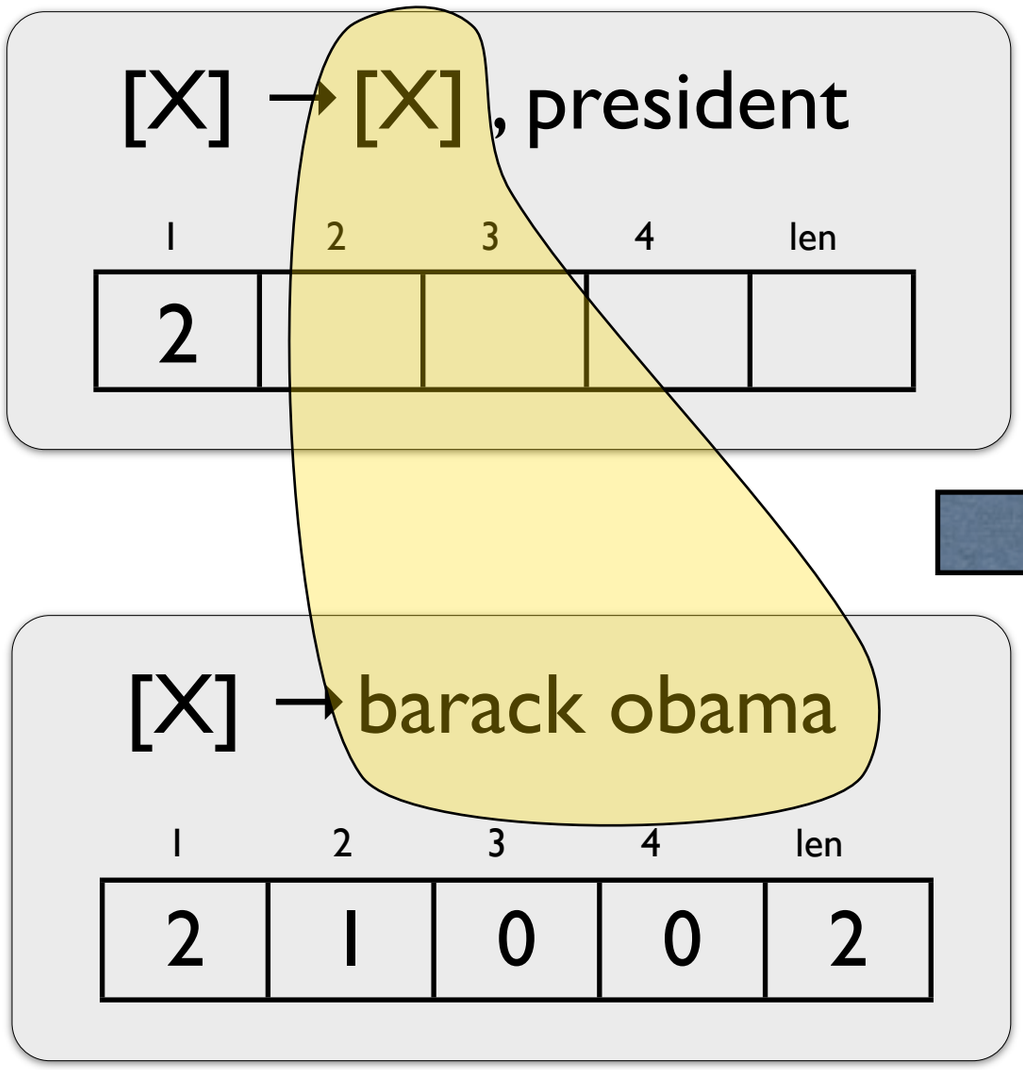


barack obama | , president

We count only ngrams in the new rule or crossing the | boundary

Computing ~BLEU

... barack obama , president ...

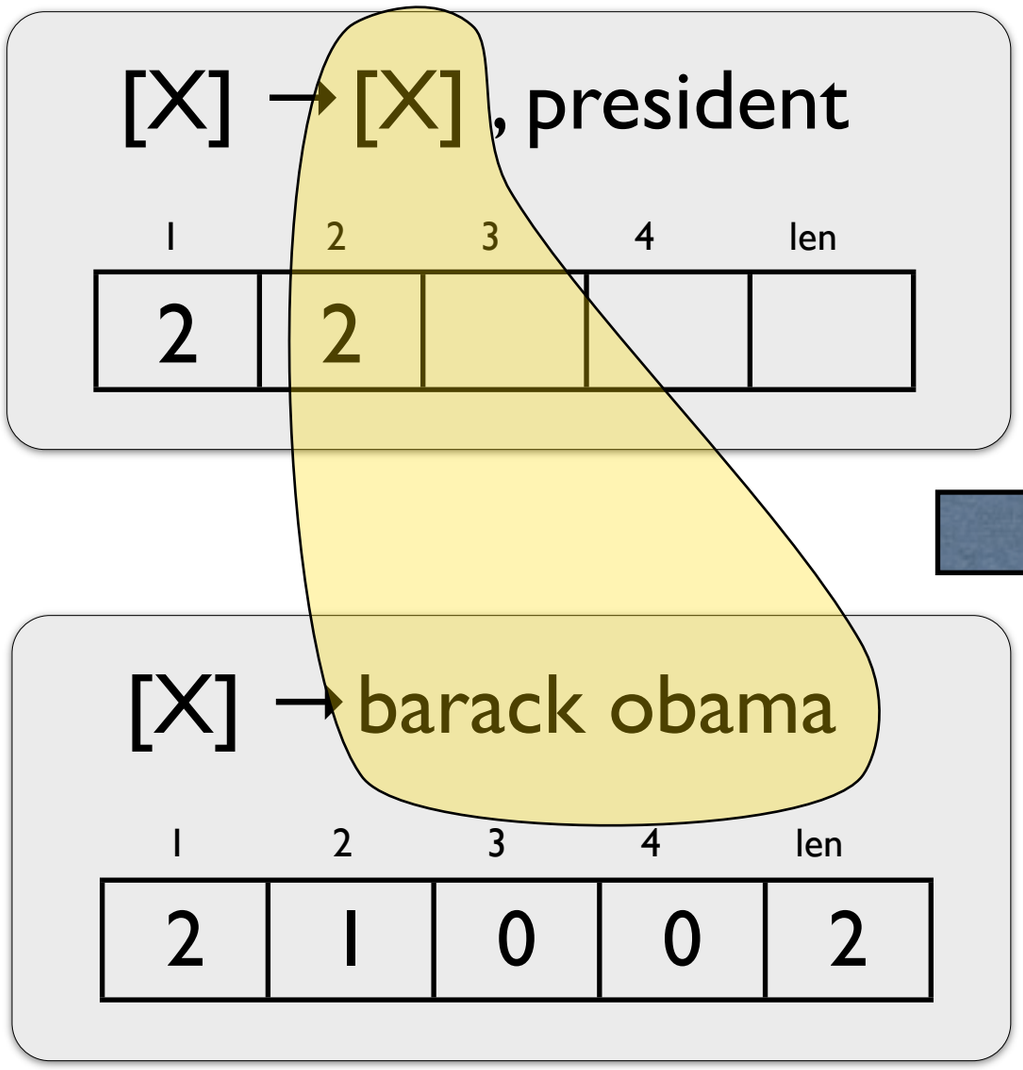


barack obama | , president

We count only ngrams in the new rule or crossing the | boundary

Computing ~BLEU

... barack obama , president ...

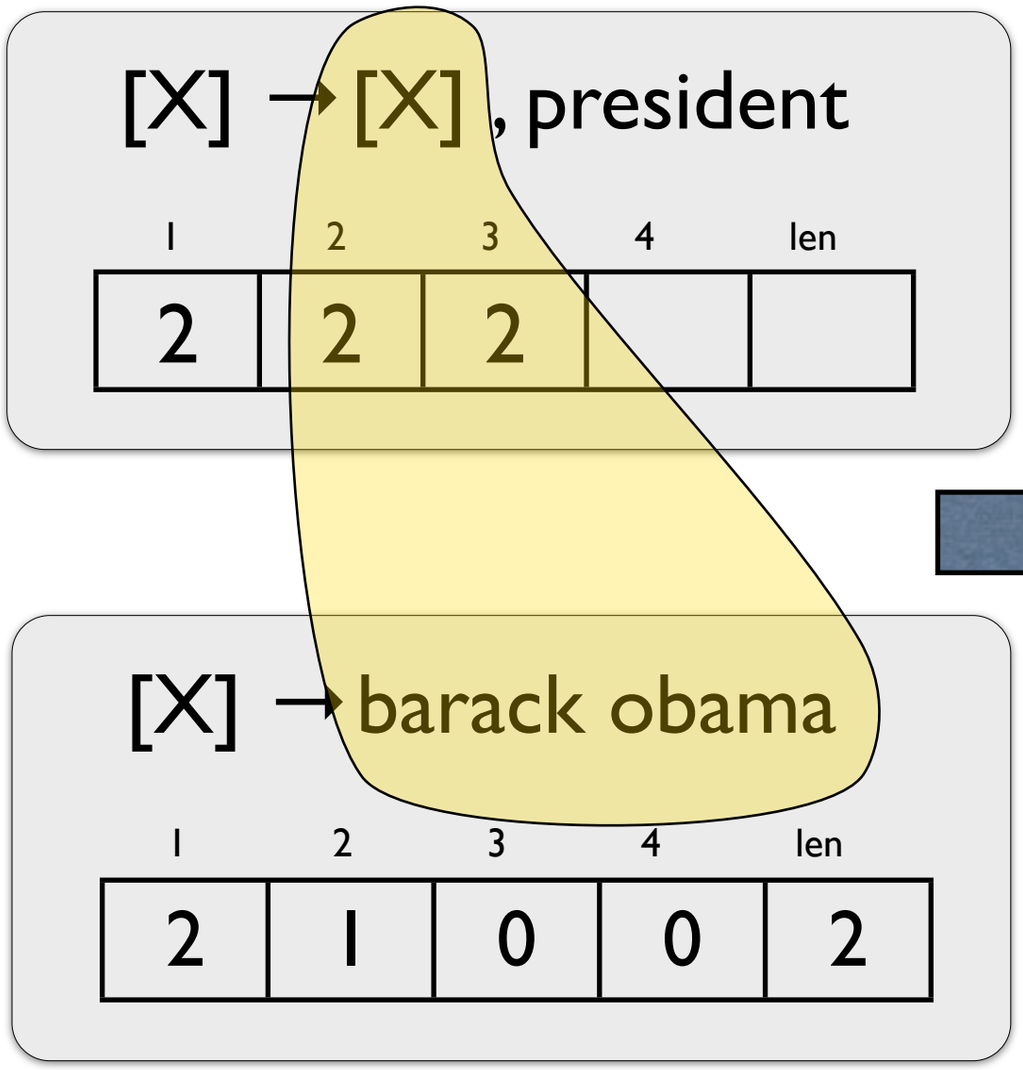


barack obama | , president

We count only ngrams in the new rule or crossing the | boundary

Computing ~BLEU

... barack obama , president ...

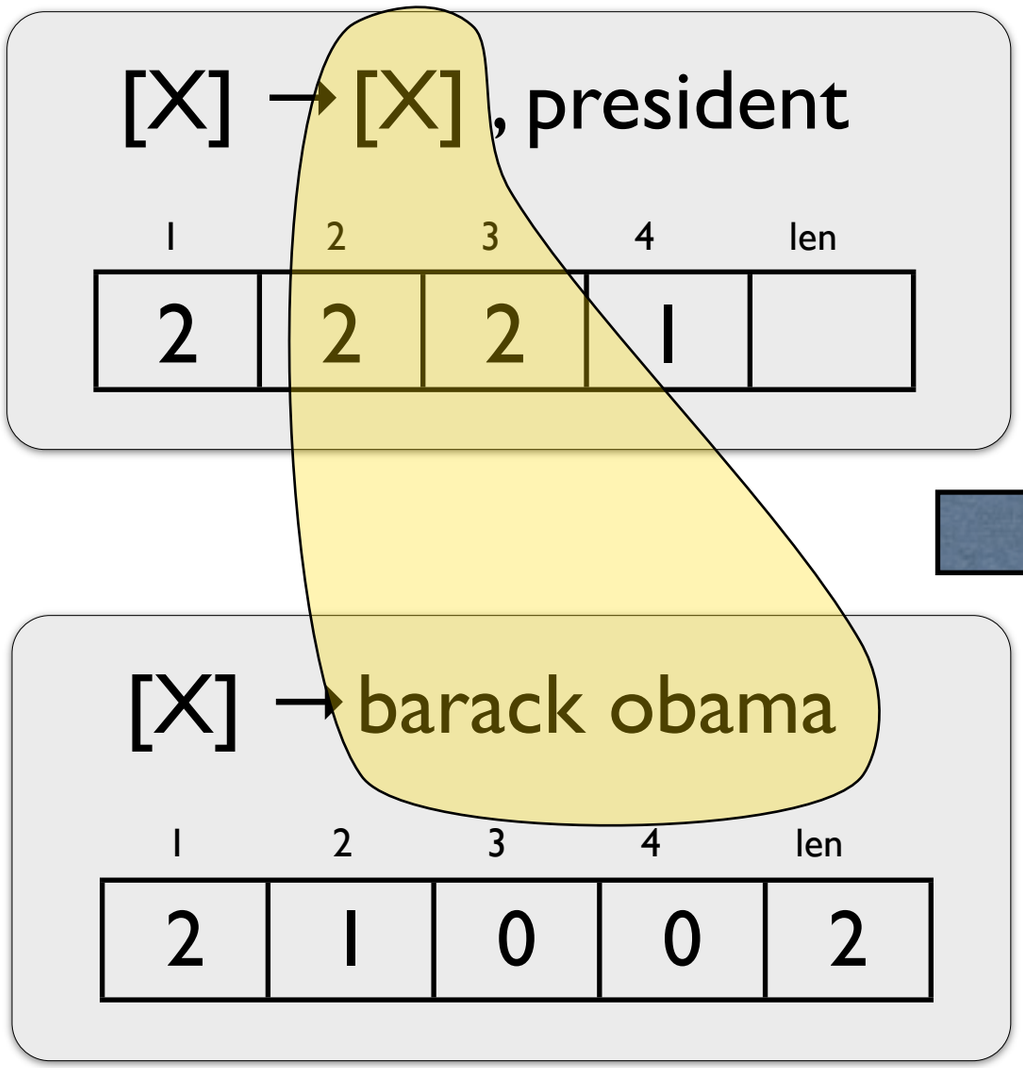


barack obama | , president

We count only ngrams in the new rule or crossing the | boundary

Computing ~BLEU

... barack obama , president ...

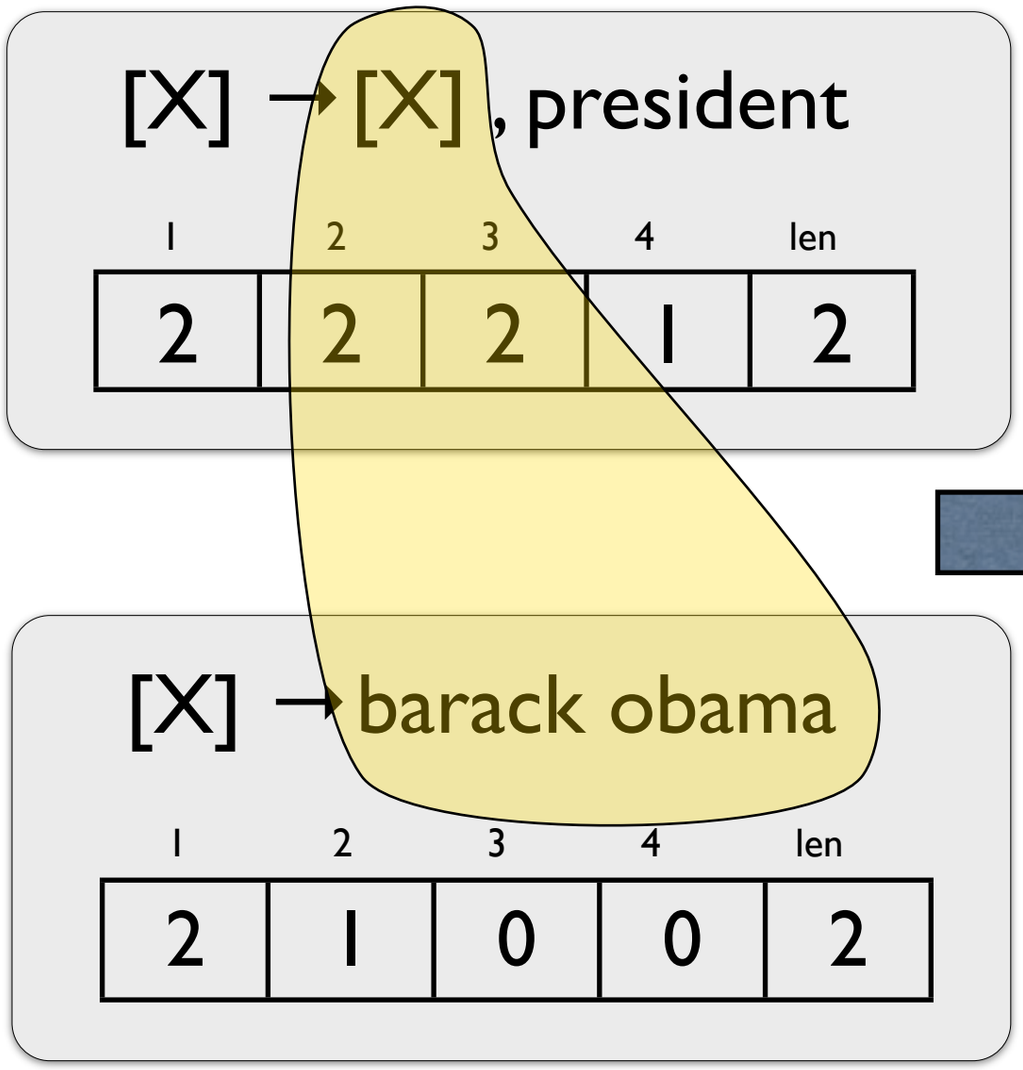


barack obama | , president

We count only ngrams in the new rule or crossing the | boundary

Computing ~BLEU

... barack obama , president ...

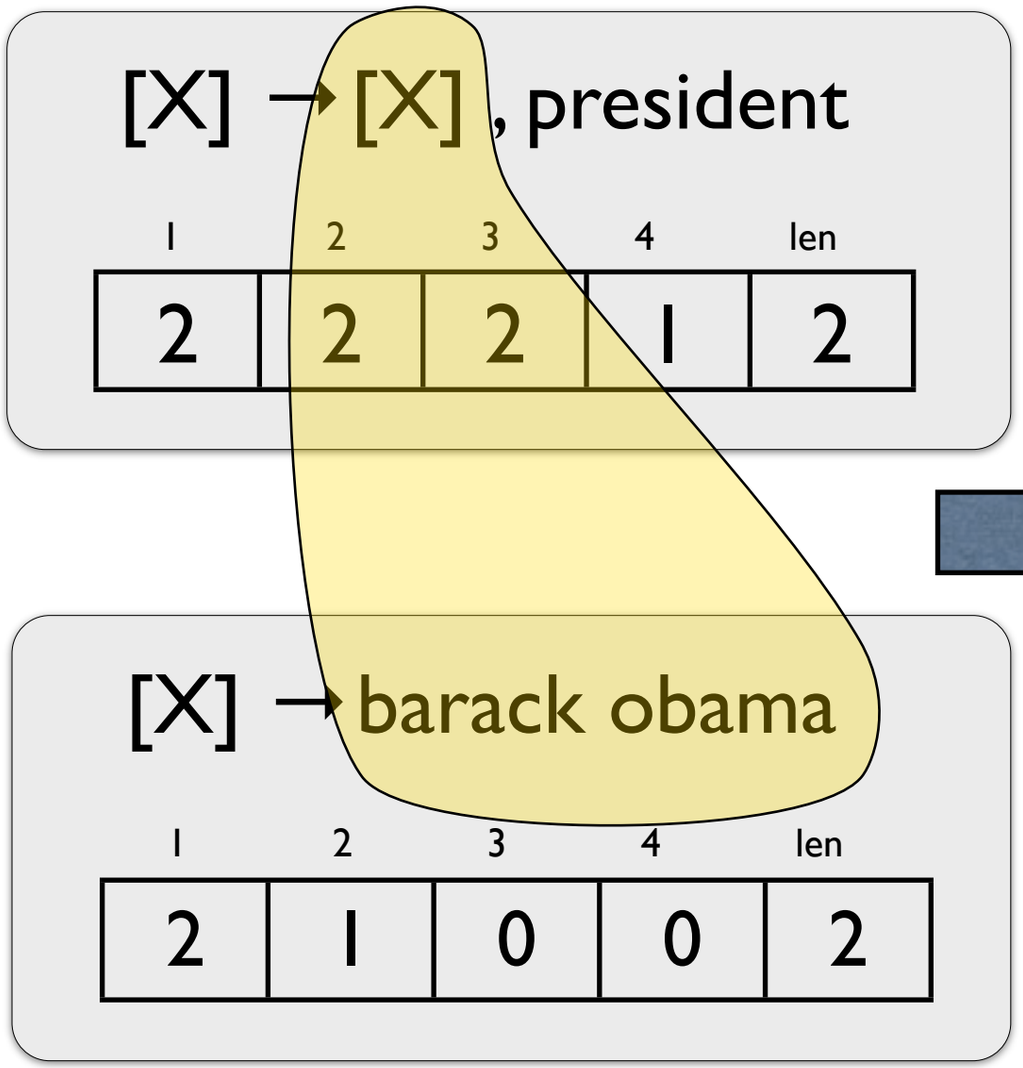


barack obama | , president

We count only ngrams in the new rule or crossing the | boundary

Computing ~BLEU

... barack obama , president ...



barack obama | , president

We count only ngrams in the new rule or crossing the | boundary

1	2	3	4	len
4	3	2	1	4

combined

Rescoring

Once we have the BLEU scores on all edges, we apply standard k-best three times

- Extract the model best (per usual)
- Rescore with model + ~BLEU (hope)
- Rescore with model – ~BLEU (fear)

Experiments

German-English Europarl + Common Crawl

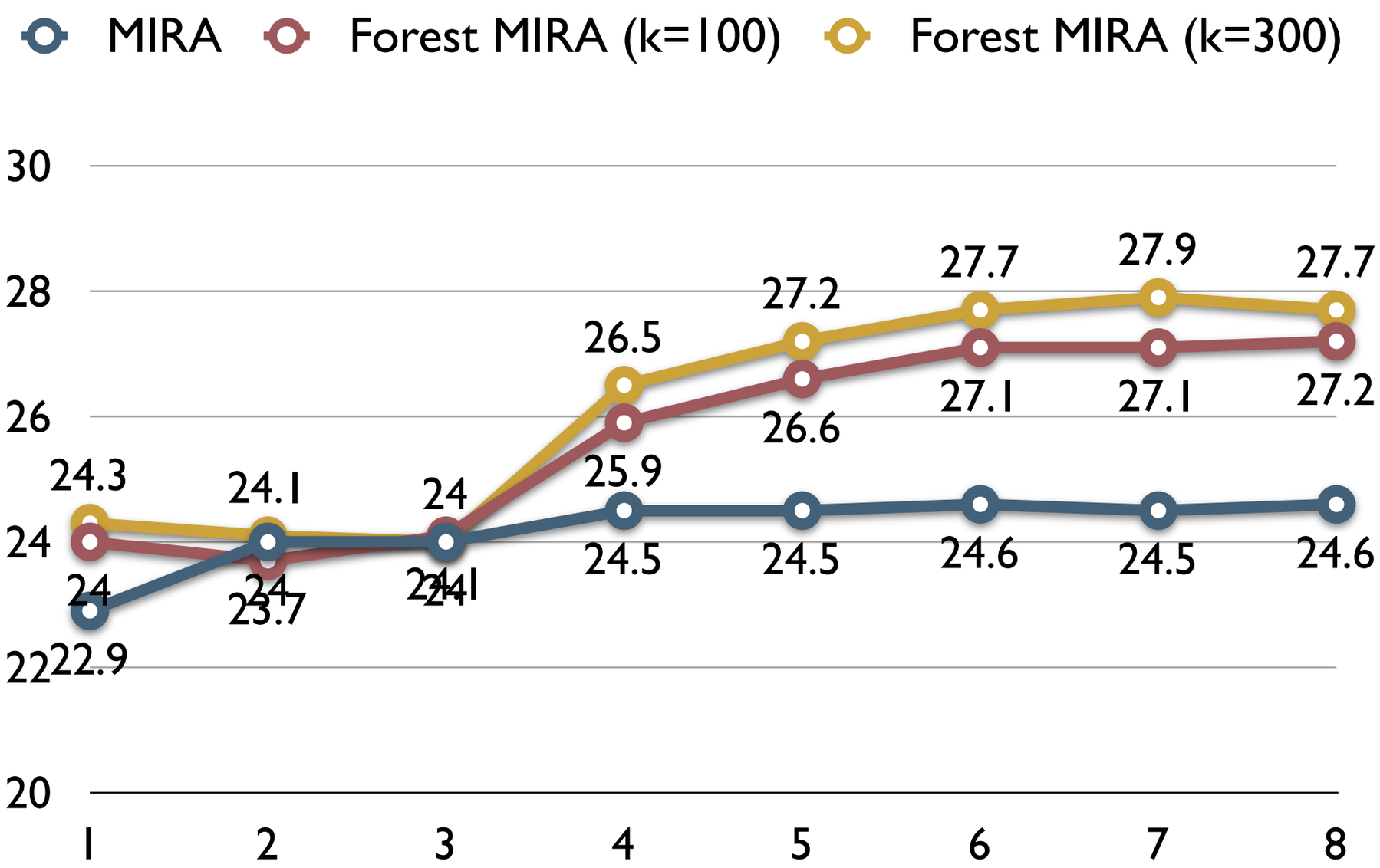
Tuning on newstest2010

Testing on newstest2012

Running MIRA for 8 iterations, test model with best tuning score

10 features

Tuning


MIRA
Forest MIRA (k=100)
Forest MIRA (k=300)
30
28
26
24
22
20
1
2
3
4
5
6
7
8
24.3
24.1
24
26.5
27.2
27.7
27.9
27.7
24
24
24
26.6
27.1
27.1
27.2
24
23.7
24
25.9
24.5
24.5
24.6
24.5
24.6
22.9

Test

Model	BLEU
MIRA	23.6
Forest MIRA (k=300)	23.1
Forest MIRA (k=100)	22.9

Problems

Can't use left-state minimization! We need the full state and don't want the complexity of state splitting

This means that we tune without left-state minimization, adding it back in for test, introducing a mismatch between models

Missing

Haven't included rolling / decaying BLEU "pseudo corpus" stats

Haven't tested on large feature sets (where we expect to see the most benefits, cf. Cherry (2013))

Summary

Shows promise

Implementation is a little messy but not difficult