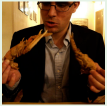
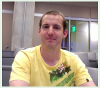
Outline


Vlad Eidelman

Ziyuan Wang

Adam Lopez

Jon Graehl

ThuyLinh Nguyen

- 3:20pm Parametric models: posterior regularisation. Desai
- 3:35pm Training models with rich features spaces. Vlad
- 3:50pm Decoding with complex grammars. Adam
- 4:20pm Closing remarks. Phil
- 4:25pm Finish.

Phrase Clustering with Posterior Regularization

CLSP Summer Workshop 2010
SMT Team
Desai Chen
joint work with Trevor Cohn

Outline

- clustering problem
- EM with posterior regularization
- results and future experiments

Phrase clustering

Phrases are defined as contiguous spans aligned with each other

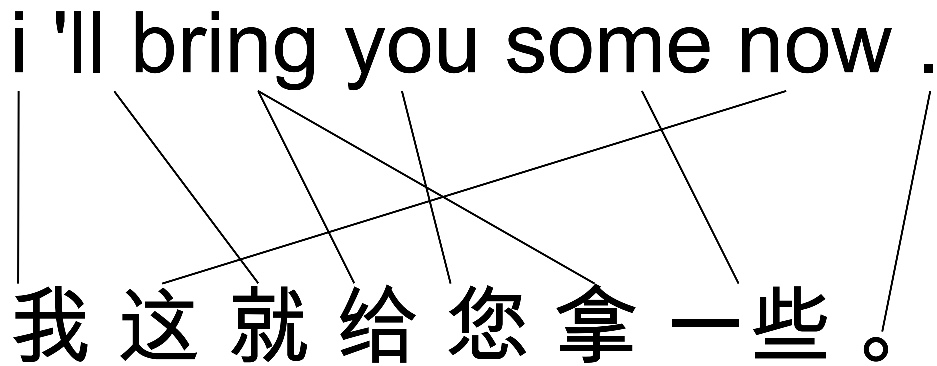
i 'll bring you some now .

我 这 就 给 您 拿 一些 。

Example from btec

Phrase clustering

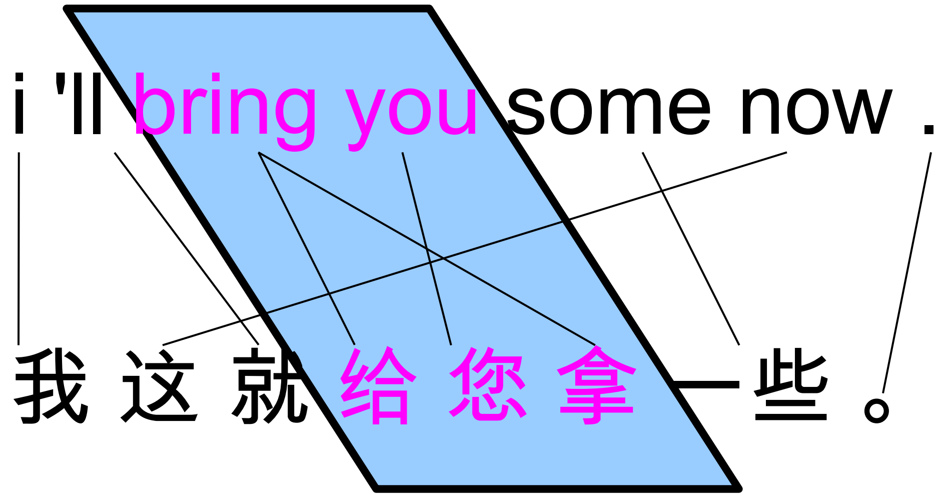
Phrases are defined as contiguous spans aligned with each other



Example from btec

Phrase clustering

Phrases are defined as contiguous spans aligned with each other



Example from btec

Phrase clustering

Phrases are defined as contiguous spans aligned with each other

i 'll bring you some now .

我 这 就 给 您 拿 一些 。

Phrase clustering

Contexts are words before or after the phrase:

target side context

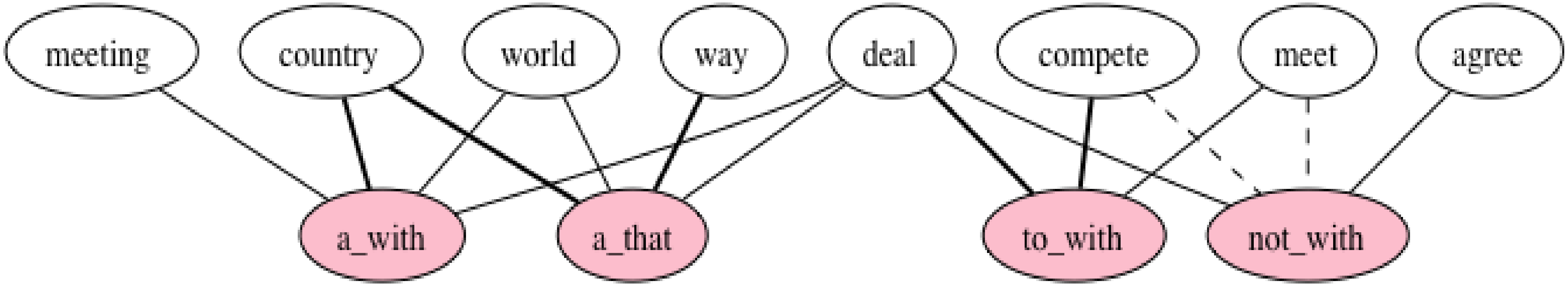
i 'll bring you some now .

我 这 就 给 您 拿 一些 。

source side context

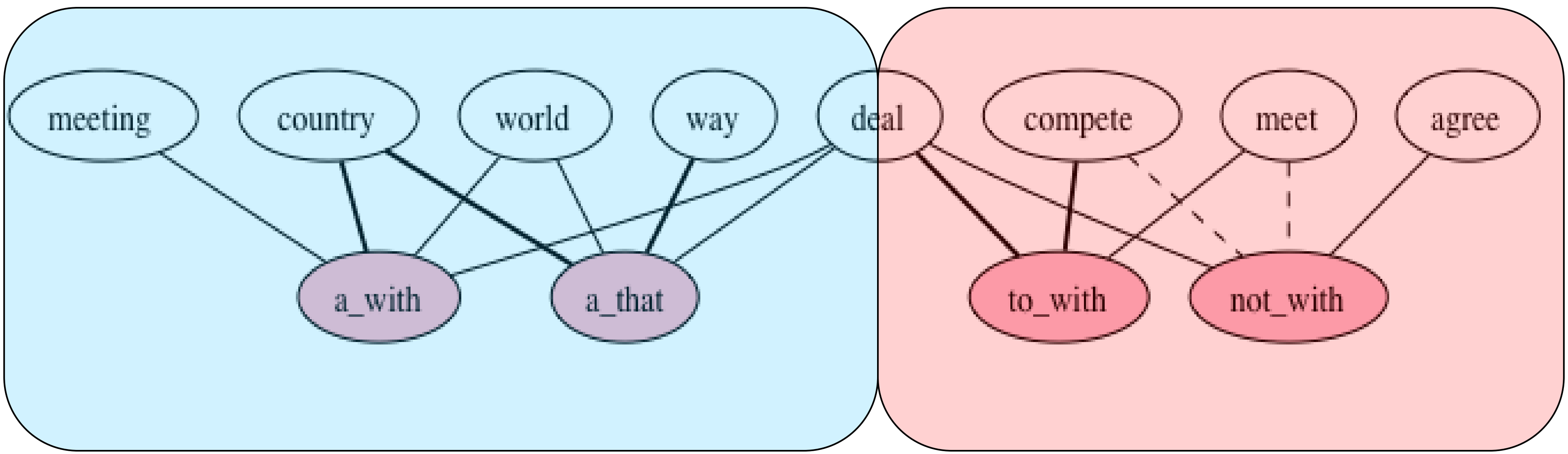
Objective

Put all phrase-context pairs into categories



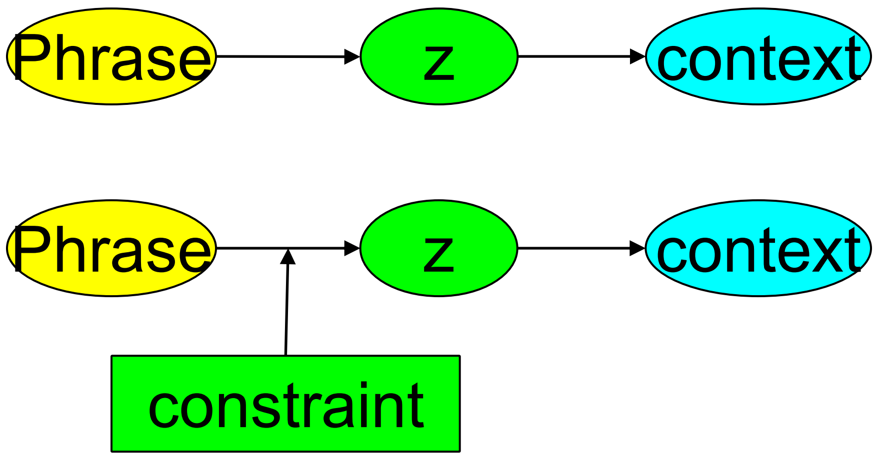
Objective

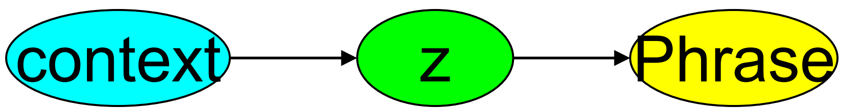
Put all phrase-context pairs into categories



Outline

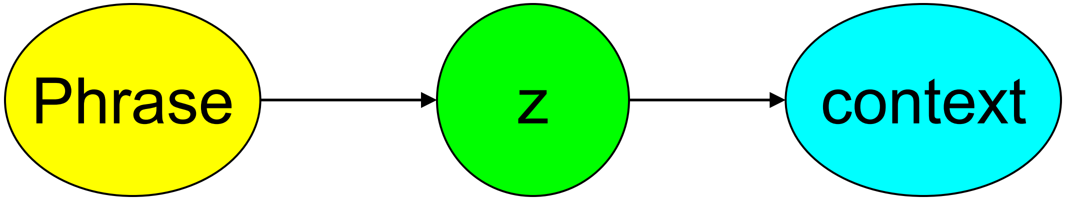
- Where do phrases come from?
- EM with posterior regularization
- results and future experiment





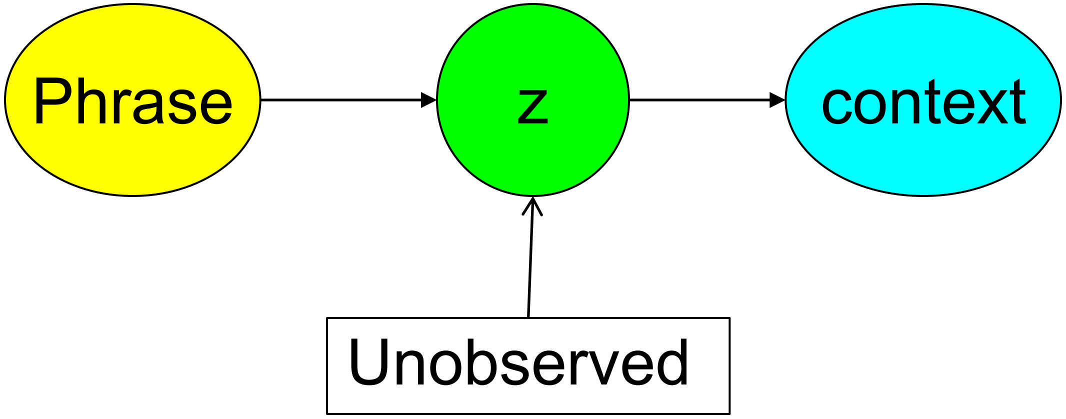
Expectation-Maximization

- naïve Bayes model for phrase labeling



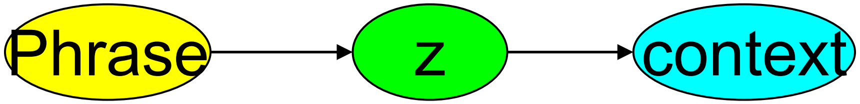
EM clustering

- naïve Bayes model for phrase labeling



EM clustering

- naïve Bayes model for phrase labeling

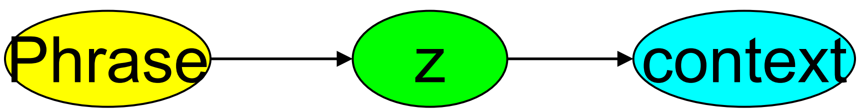


E-step

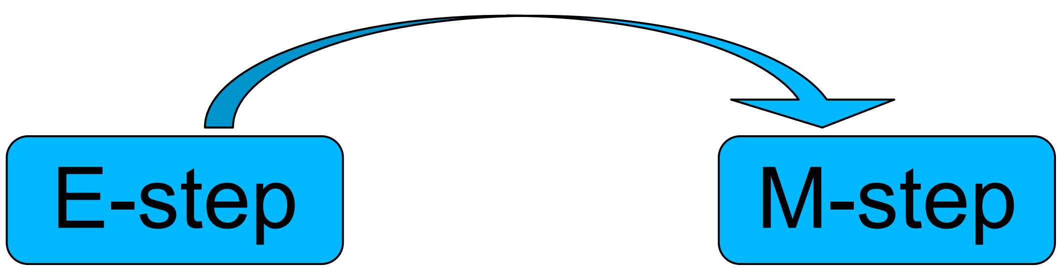
M-step

EM clustering

- naïve Bayes model for phrase labeling

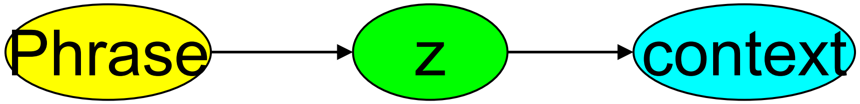


$$q(z|p,c) = P_\theta(z|p,c)$$

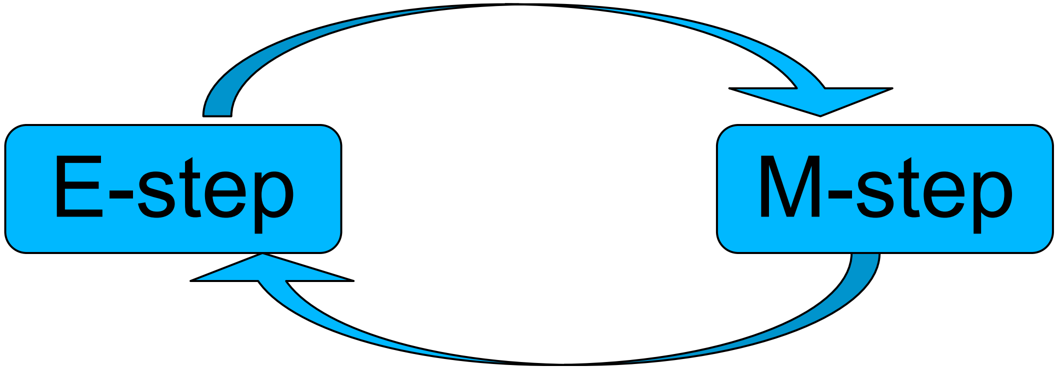


EM clustering

- naïve Bayes model for phrase labeling



$$q(z|p,c) = P_\theta(z|p,c)$$



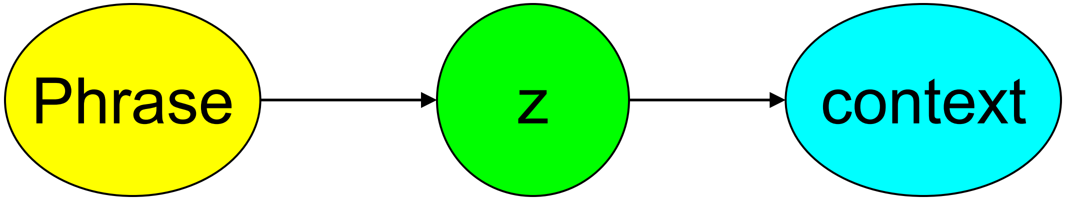
$$\theta = MLE\ q(z|p,c)$$

Problem with EM

- Problem: EM uses as many categories as it wants for each phrase.
- We want to limit the number of categories associated with each phrase.

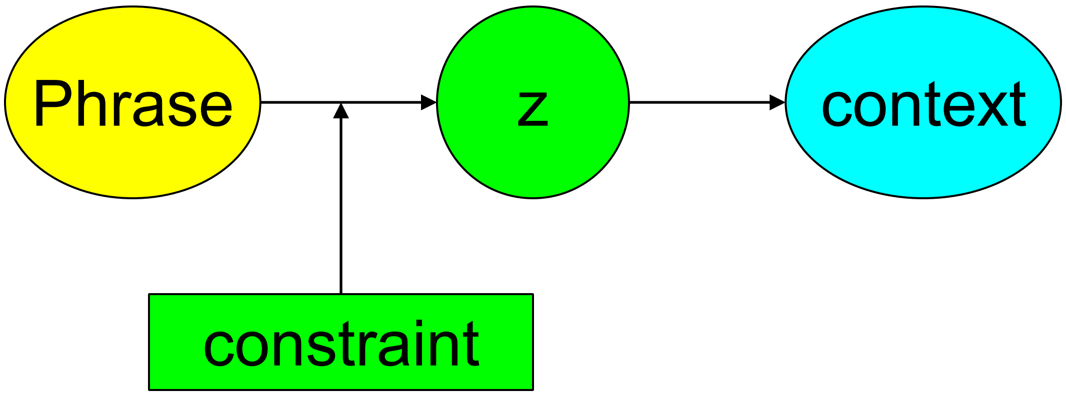
Sparsity constraints

- Sparsity:Each phrase/context should be labeled with fewer kinds of labels.



Sparsity constraints

- Sparsity:Each phrase/context should be labeled with fewer kinds of labels.



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

Contexts:

i understand there are some sightseeing bus tours here , is that right ?

there are only a few seats left in the dress circle .

well , of course there are fine restaurants .

your hotel brochure shows there are some tennis counts at your hotel .

Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

Contexts:
i understand there are some sightseeing bus tours here , is that right ?

there are only a few seats left in the dress circle .

well , of course there are fine restaurants .

your hotel brochure shows there are some tennis counts at your hotel .

Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants
brochure shows _ some tennis

Sparsity constraints

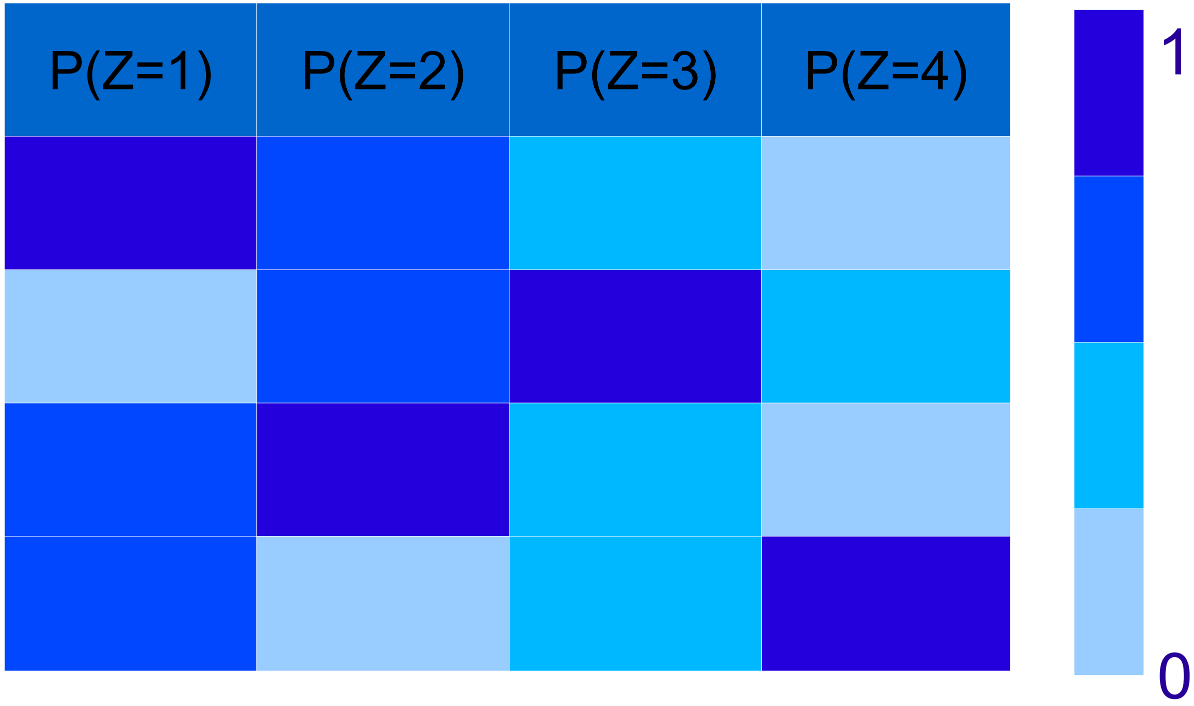
Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants
brochure shows _ some tennis



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

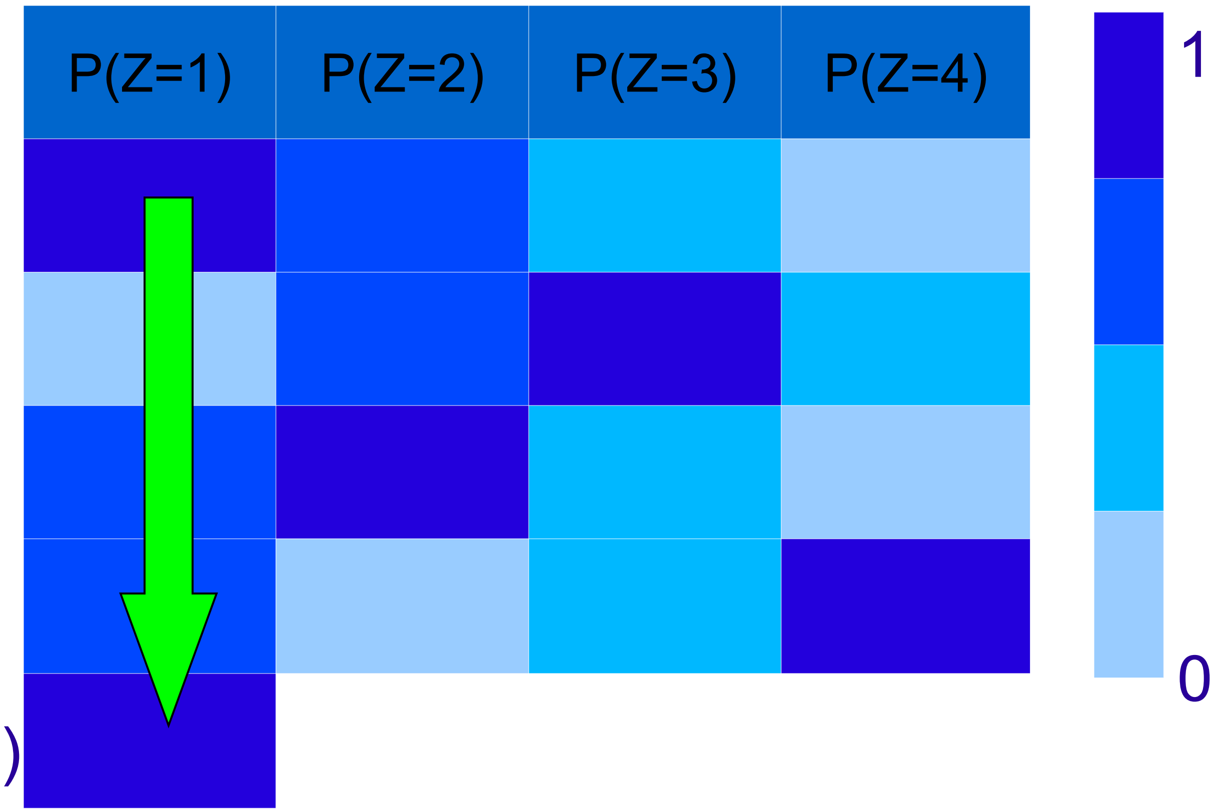
Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

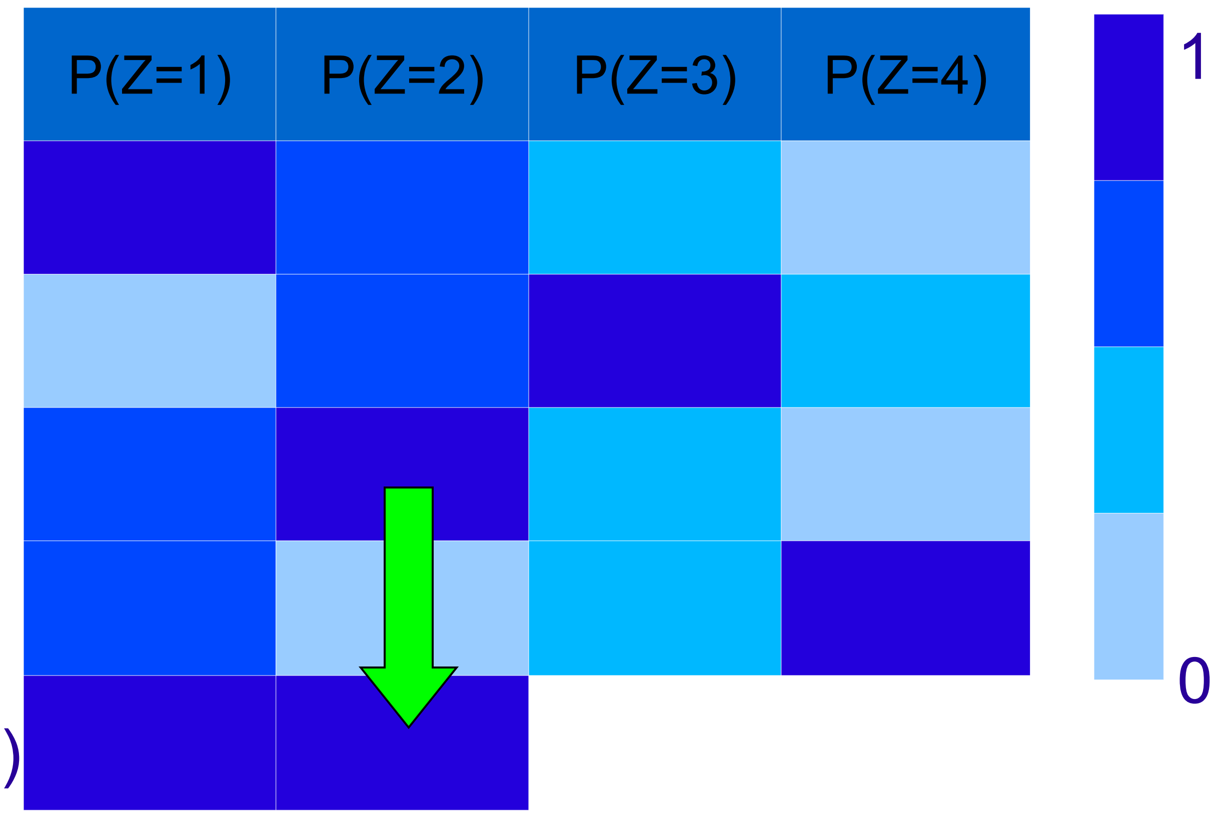
Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

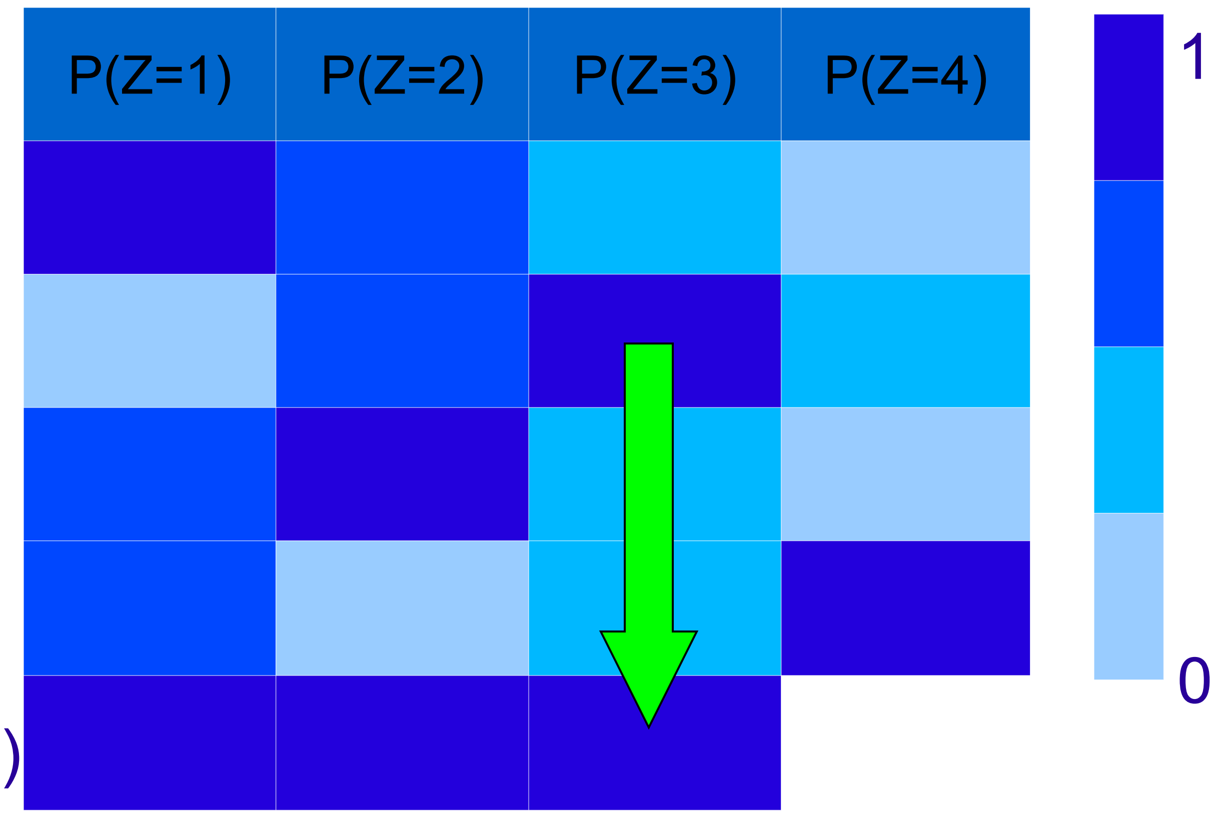
Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

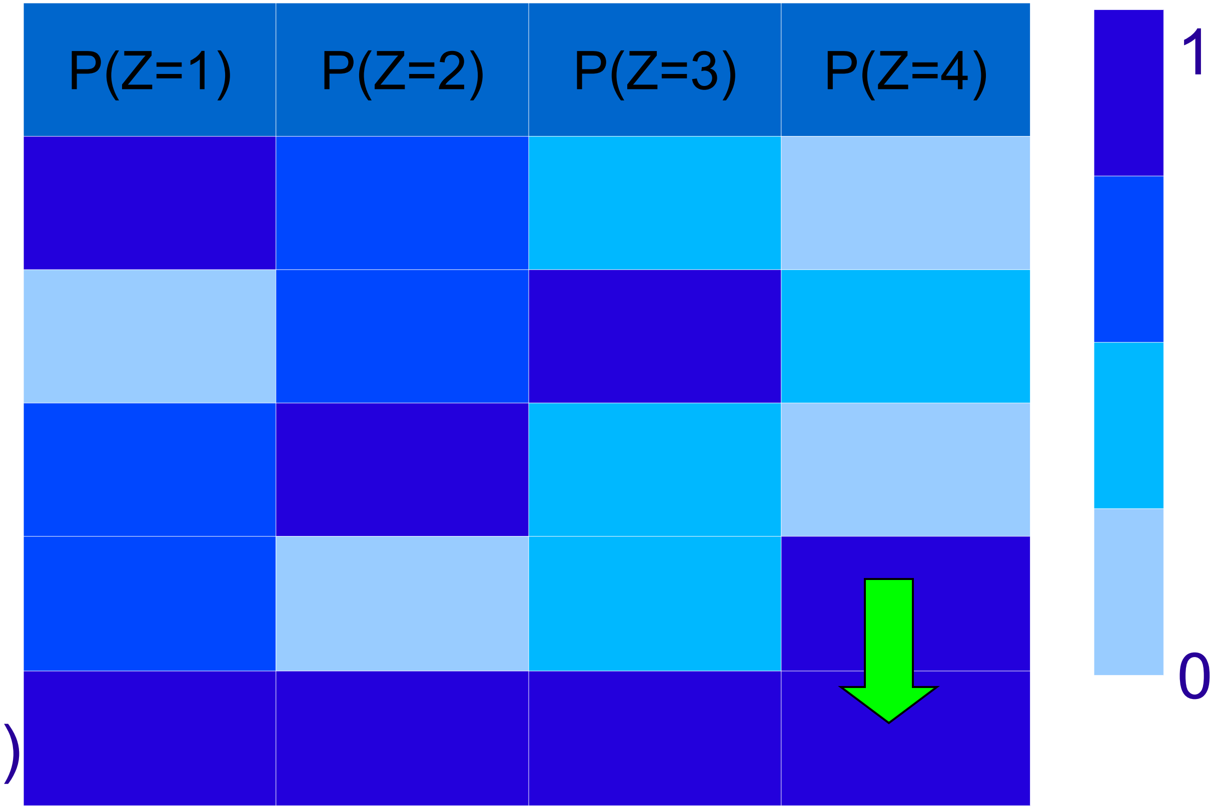
Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

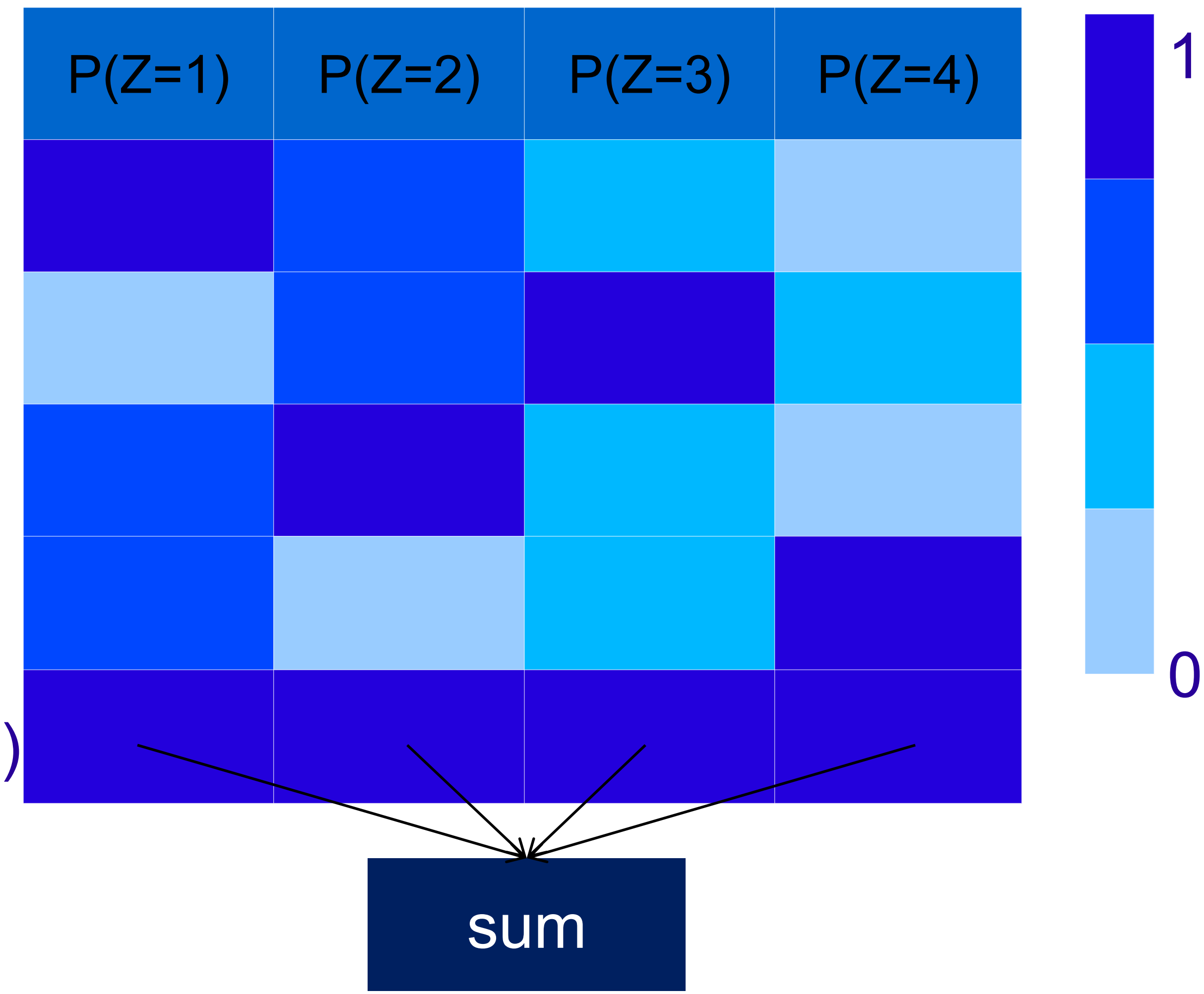
Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

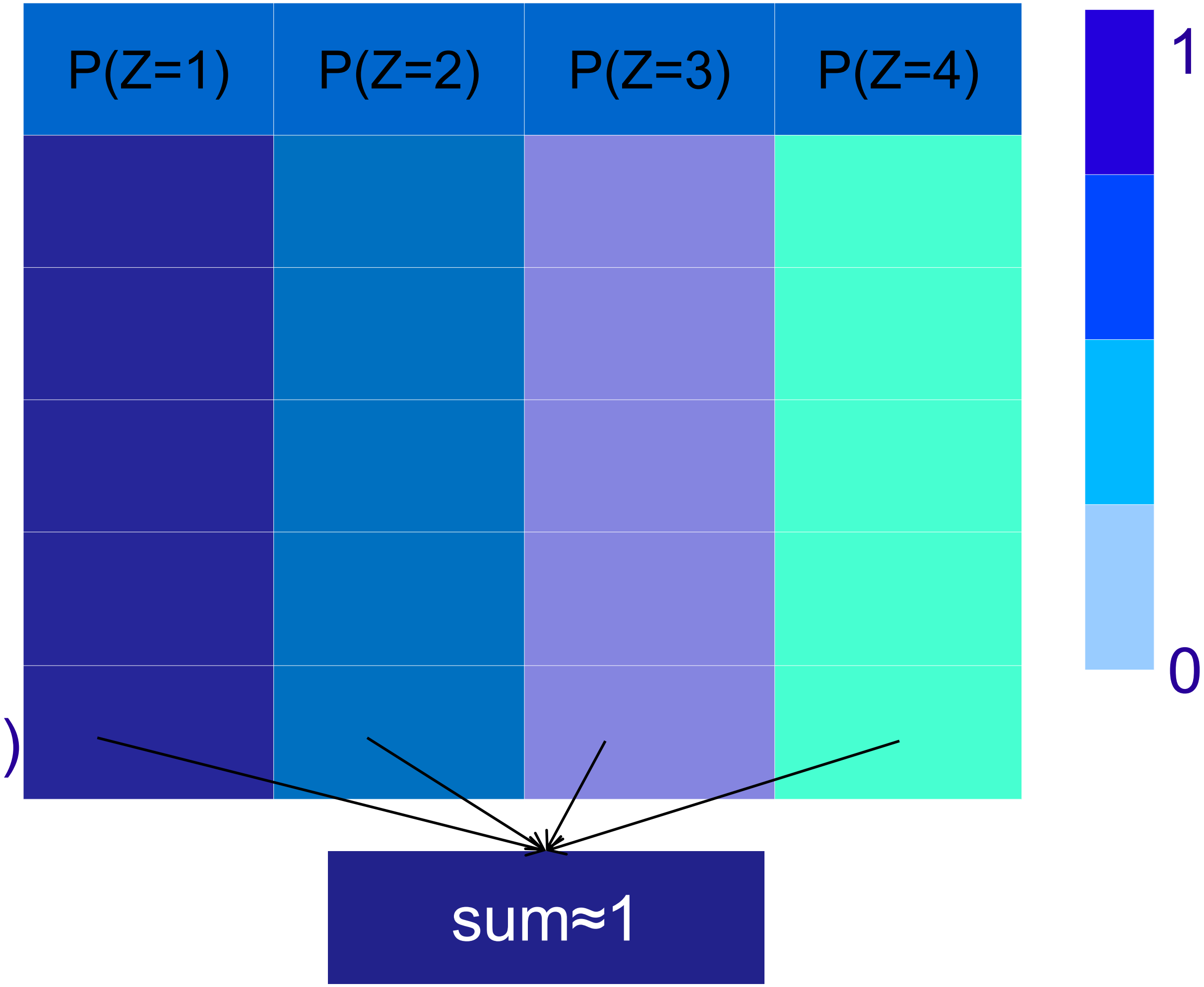
Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)

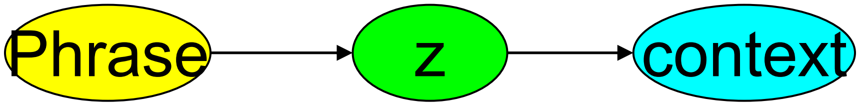


Posterior Regularization

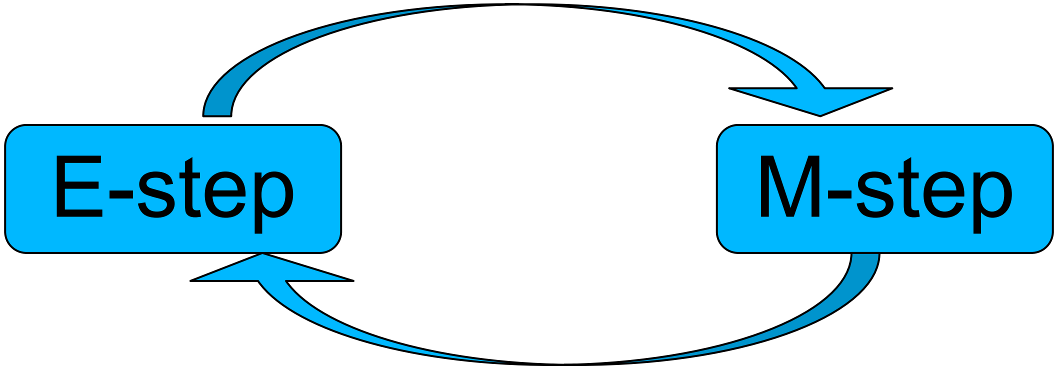
- Follows *Posterior Regularization for Structured Latent Variable Models,* Ganchev et al., 2009
- During E-step, impose constraints on the posterior q to guide the search

Posterior Regularization

- impose constraints on the posterior q



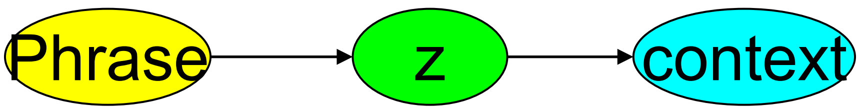
$$q(z|p,c) = \arg\min_{q \in Q} KL(q||P_\theta)$$

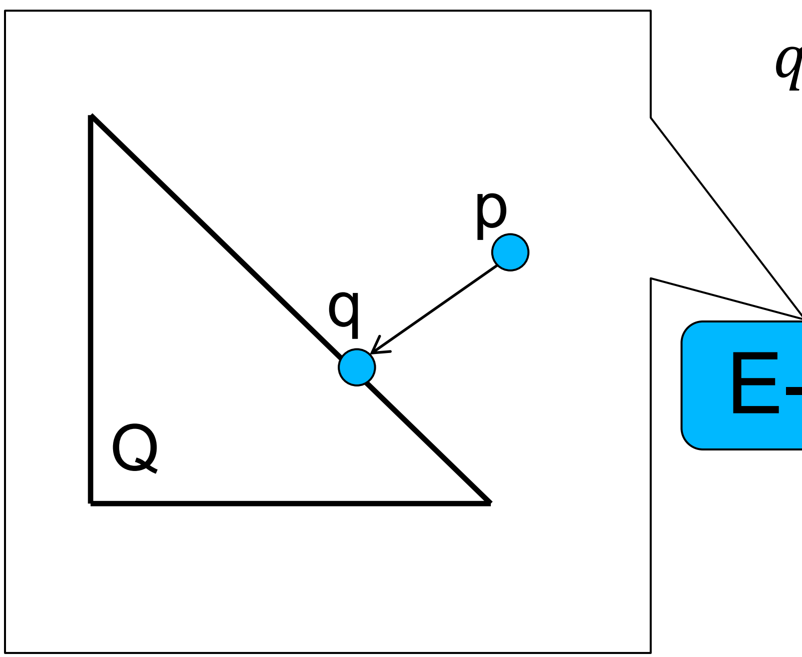


$$\theta = MLE\ q(z|p,c)$$

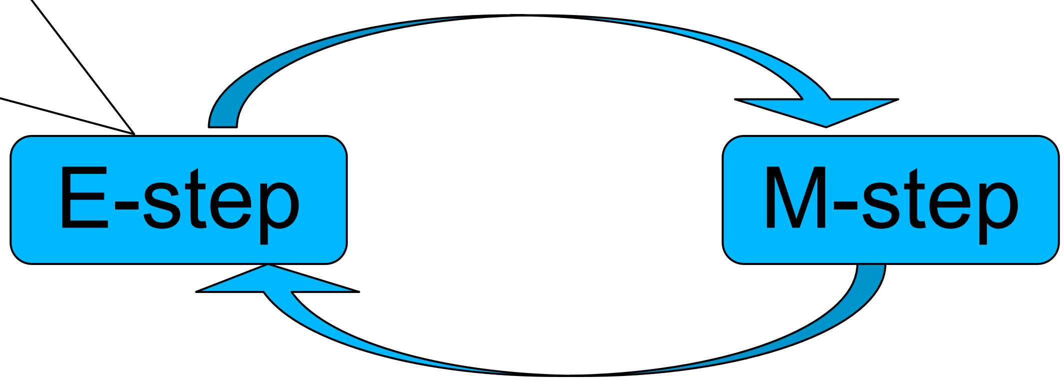
Posterior Regularization

- impose constraints on the posterior q





$$q(z|p,c) = \arg\min KL(q||P_\theta)$$
$$q \in Q$$



$$\theta = MLE\ q(z|p,c)$$

Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: like this

Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants
brochure shows _ some tennis

Define feature functions:

$$\phi_{i,j}(p,z) = \begin{cases} 1 \; if \; p = i \; and \; z = j \\ 0 \qquad\qquad otherwise \end{cases}$$

Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

- Soft constraint. Softness controlled by σ.
- During E-step, find q distribution:

$$\min_{q,c_{p,z}} KL(q||P_\theta) + \sigma \sum_{p,z} c_{p,z}$$

$$\text{s.t. } E_q\left[\phi_{p,z}\right] \le c_{p,z}$$

where “c”s are maximums of expectation for each word tag pair by definition.

Primitive results

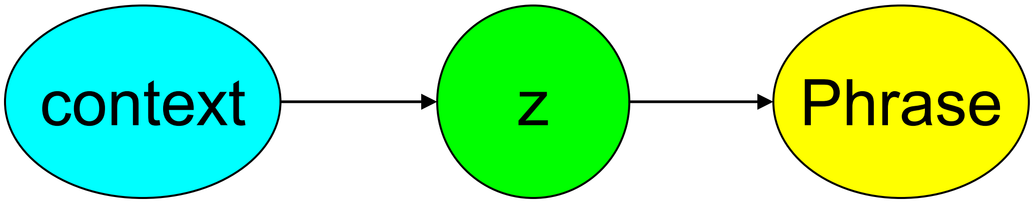
- Constrained model gives clustering that's more sparse
- Clustering for a few phrases with 25 tags on BTEC ZH-EN

Phrase/Word	Count of the most used tag		Number of tags used	
the	1194	**1571**	11	**4**
there is	**53**	50	5	**4**
'd like	723	**873**	5	**2**

More experiments

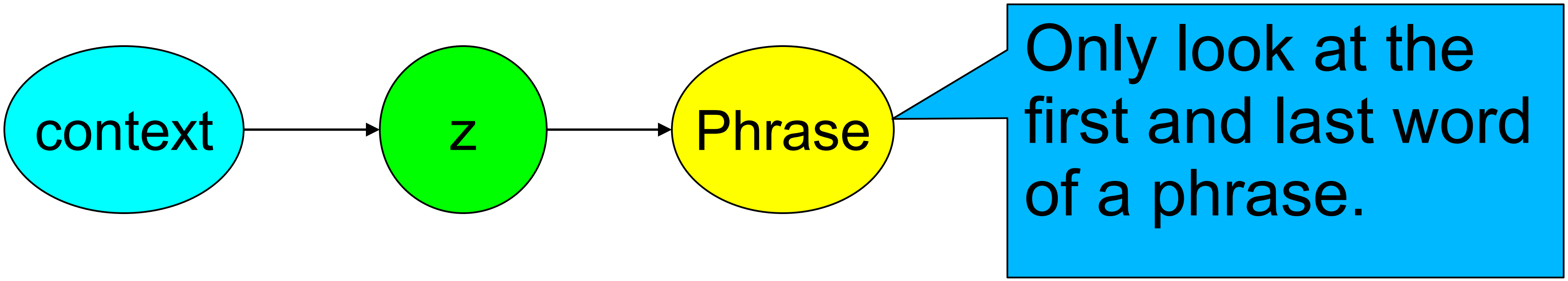
- agreement constraint: different “good” models should agree on posterior distribution
- what model to agree with: another naïve Bayes model in the reverse direction or in the other language.

Agreement model



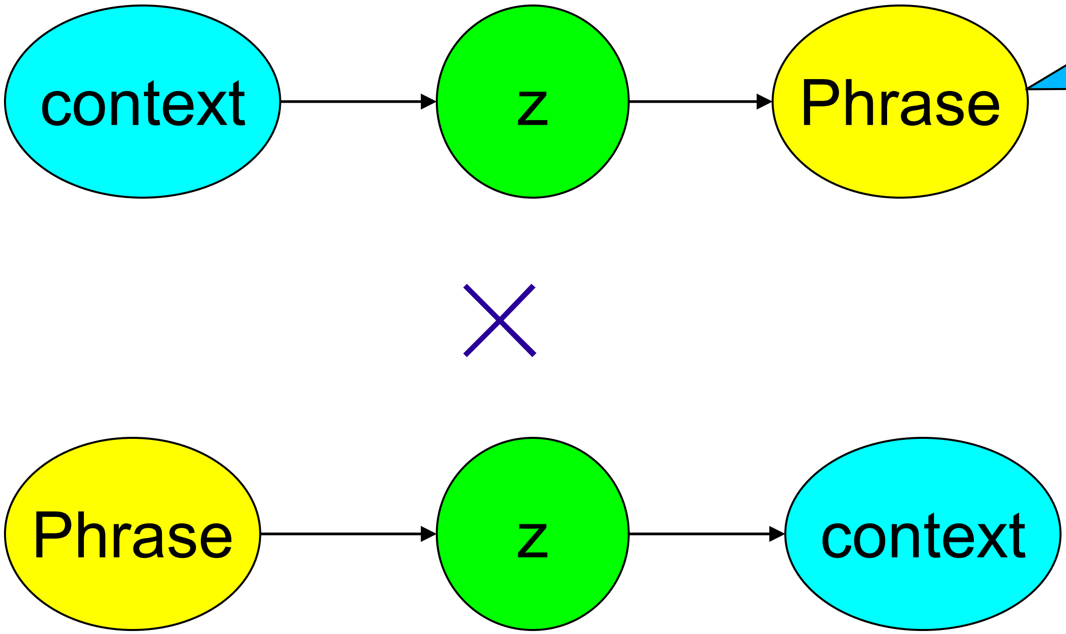
- implementation: multiply posteriors of two models together.

Agreement model



- implementation: multiply posteriors of two models together.

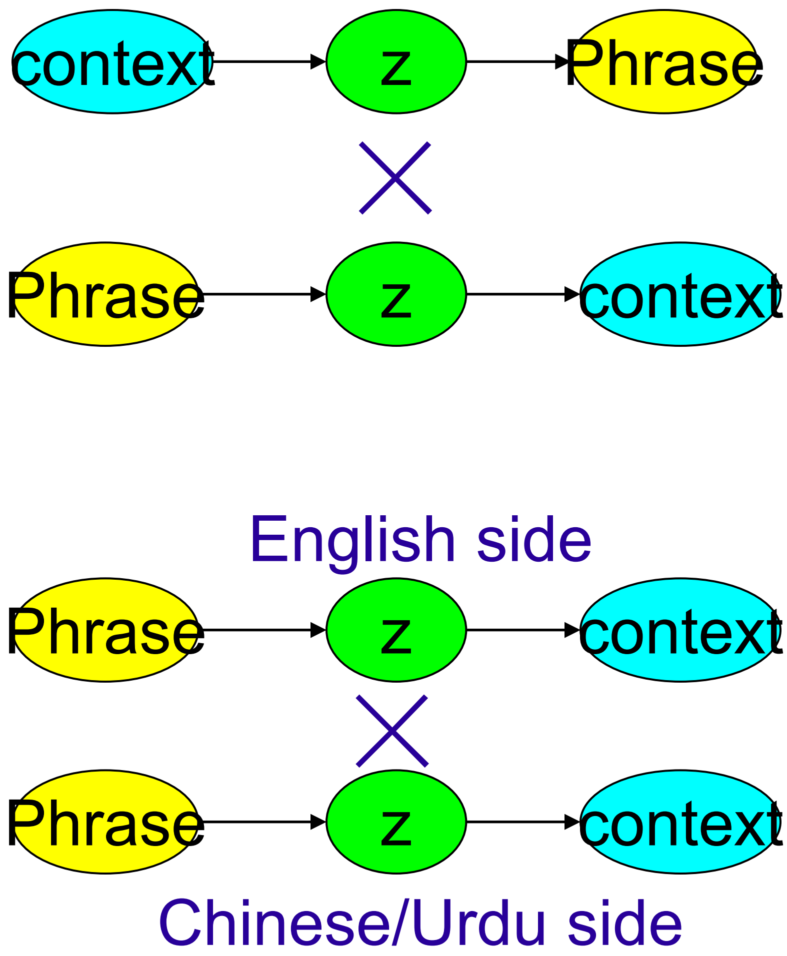
Agreement model



Only look at the first and last word of a phrase.

- implementation: multiply posteriors of two models together.

Agreement model

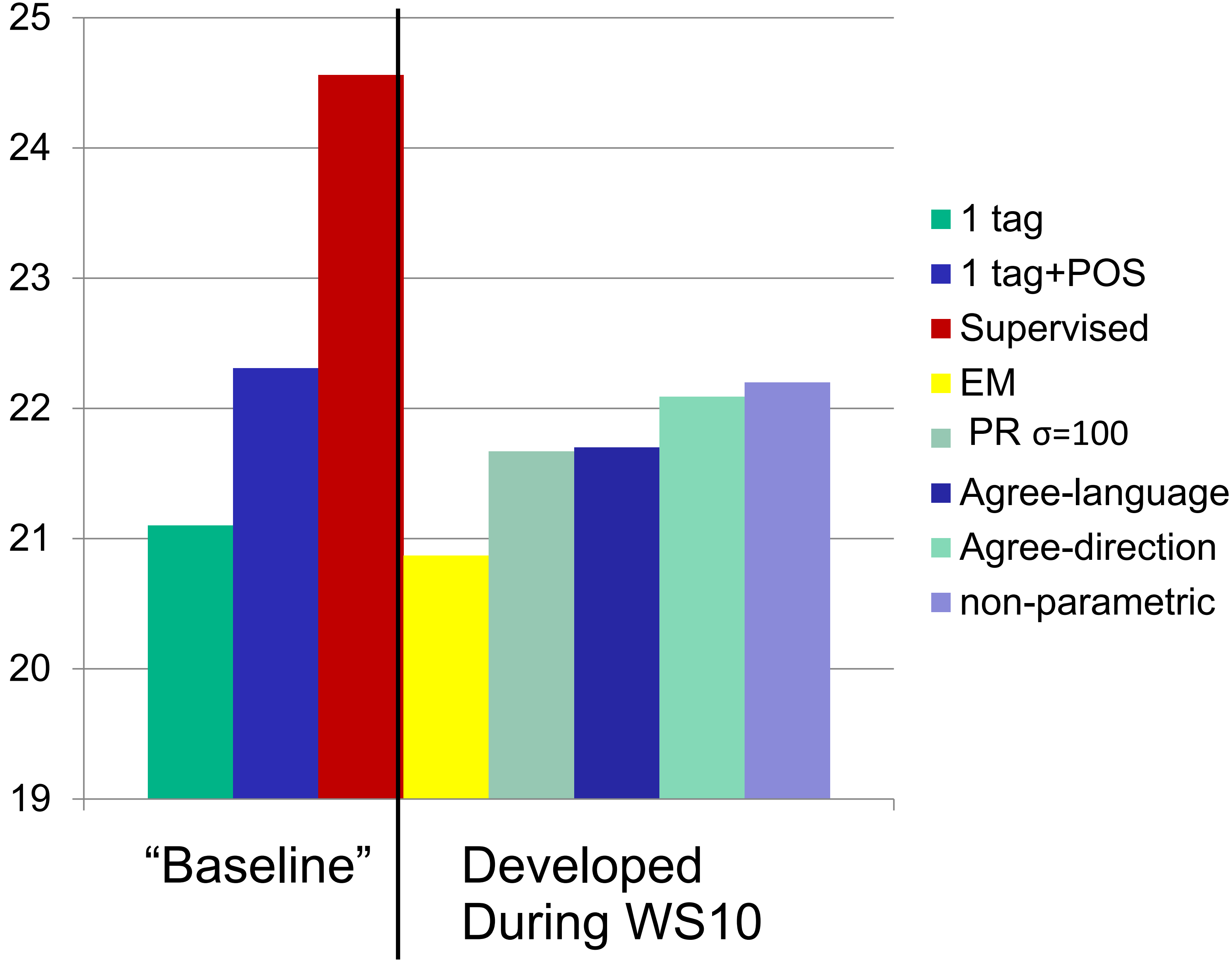


- implementation: multiply posteriors of two models together.

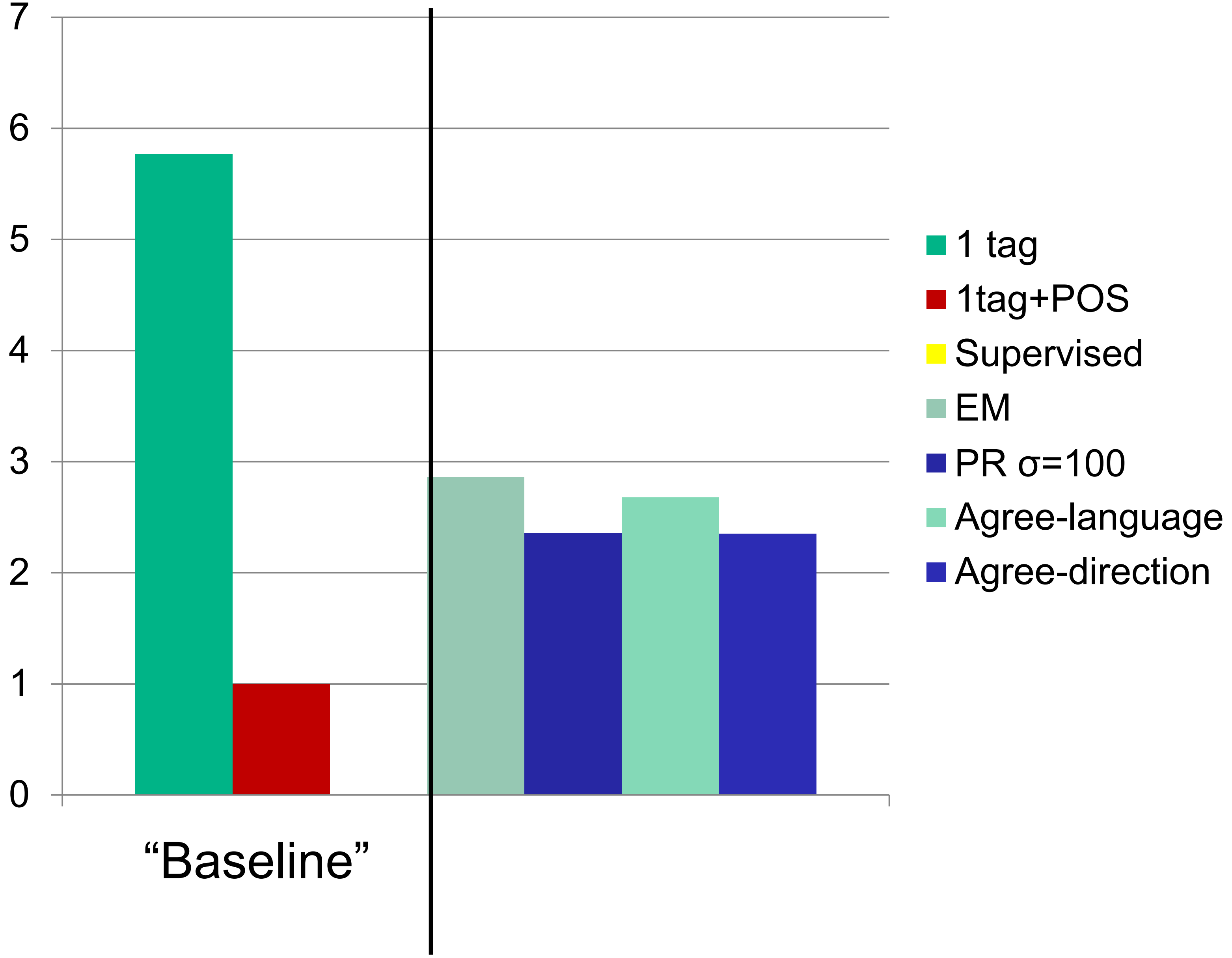
Outline

- Where do phrases come from?
- EM with posterior regularization
- results and future experiments

Evaluation through the translation pipeline on Urdu-English data
BLEU score, higher is better

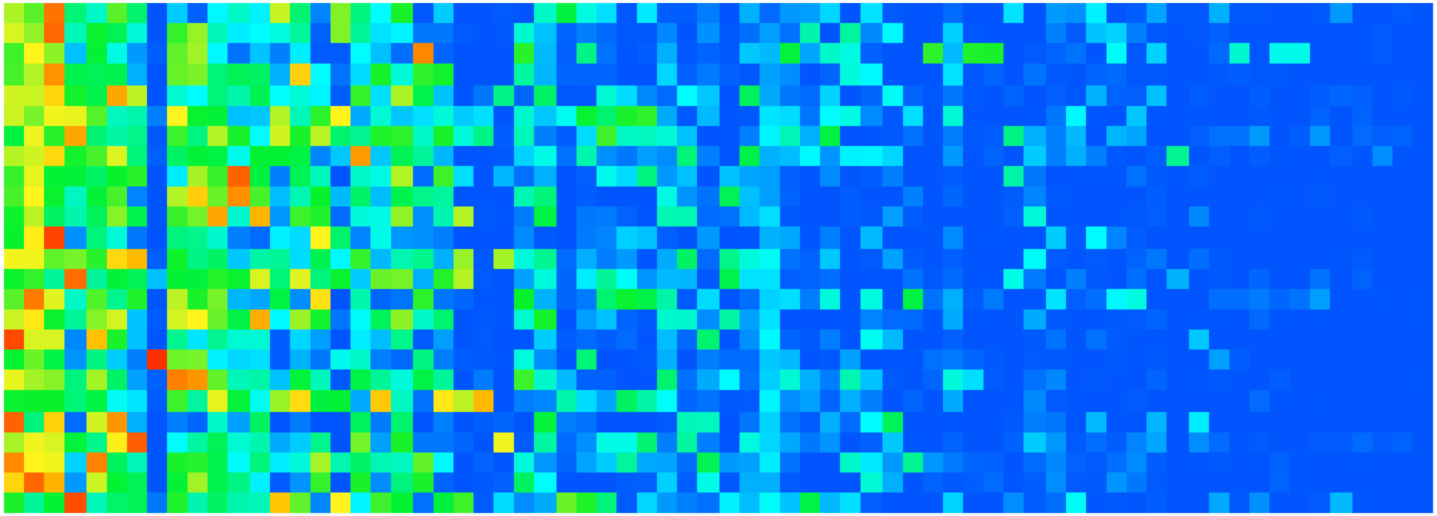


Evaluation against supervised grammar (Conditional Entropy, lower is better)

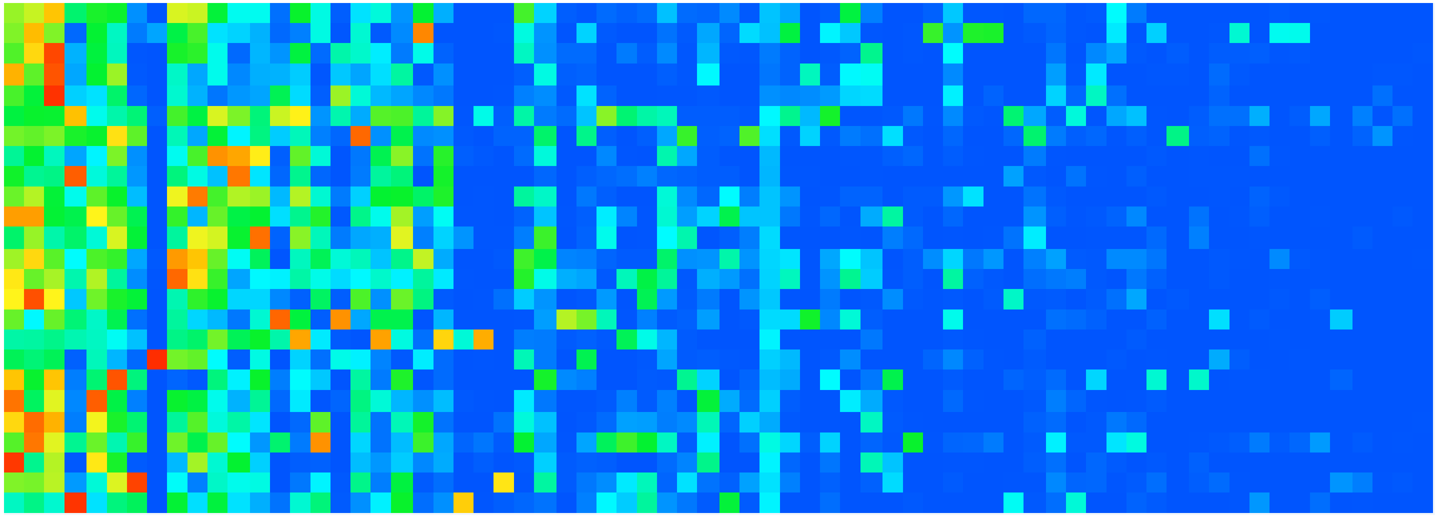


Confusion matrix against supervised labeling

EM



Agreement model between languages



Things we didn’t have time to get working

- Semi-supervised training with POS tags.
- Label single-word phrases with their POS tags.

Things we didn't have time to get working

Bayesian Bayesian Bayesian

- variational Bayes inference

Bayesian Bayesian Bayesian

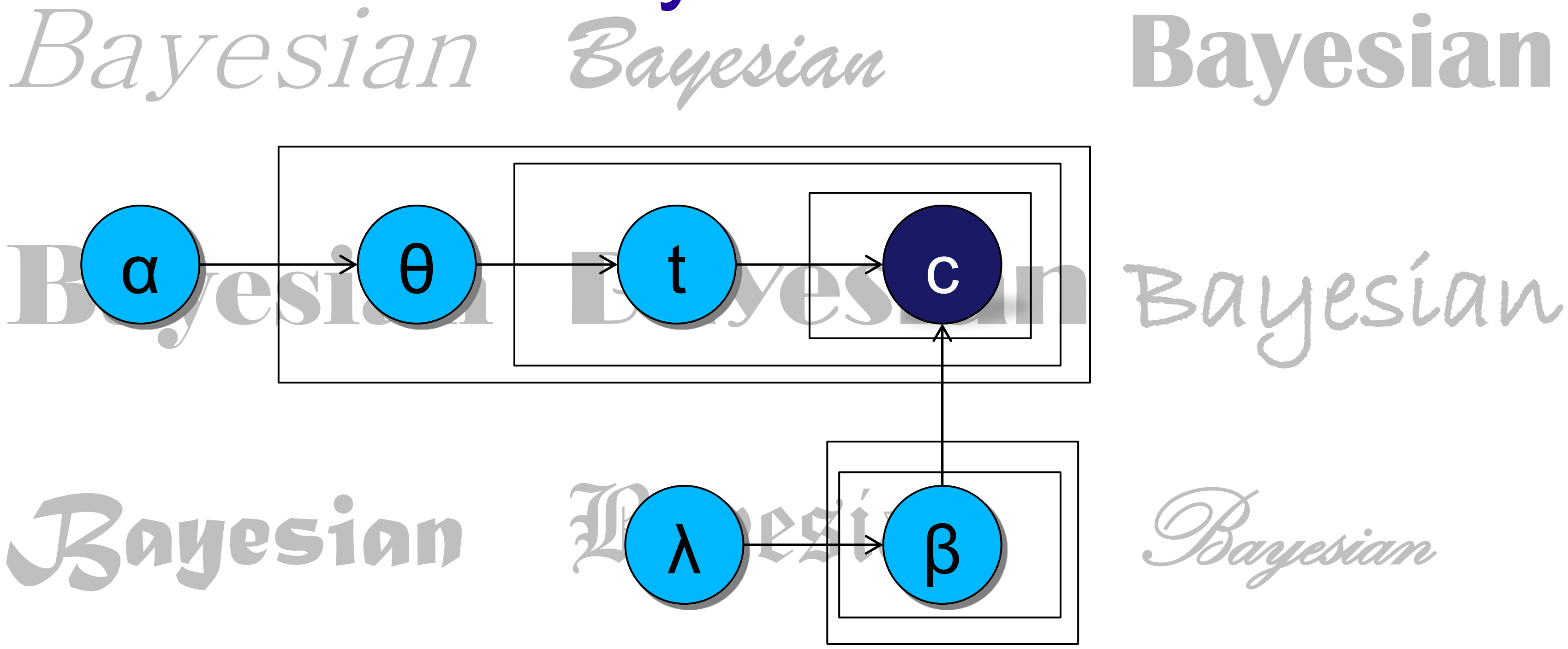
Bayesian Bayesian Bayesian

Bayesian Bayesian Bayesian

Things we didn't have time to get working

Bayesian Bayesian Bayesian

- variational Bayes inference



Outline

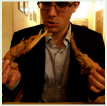
- Where do phrases come from?
- EM with posterior regularization
- results and future experiments

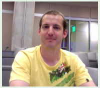
Thanks!

Outline


Vlad Eidelman


Ziyuan Wang


Adam Lopez


Jon Graehl


ThuyLinh Nguyen

- 3:20pm Parametric models: posterior regularisation. Desai
- 3:35pm Training models with rich features spaces. Vlad
- 3:50pm Decoding with complex grammars. Adam
- 4:20pm Closing remarks. Phil
- 4:25pm Finish.

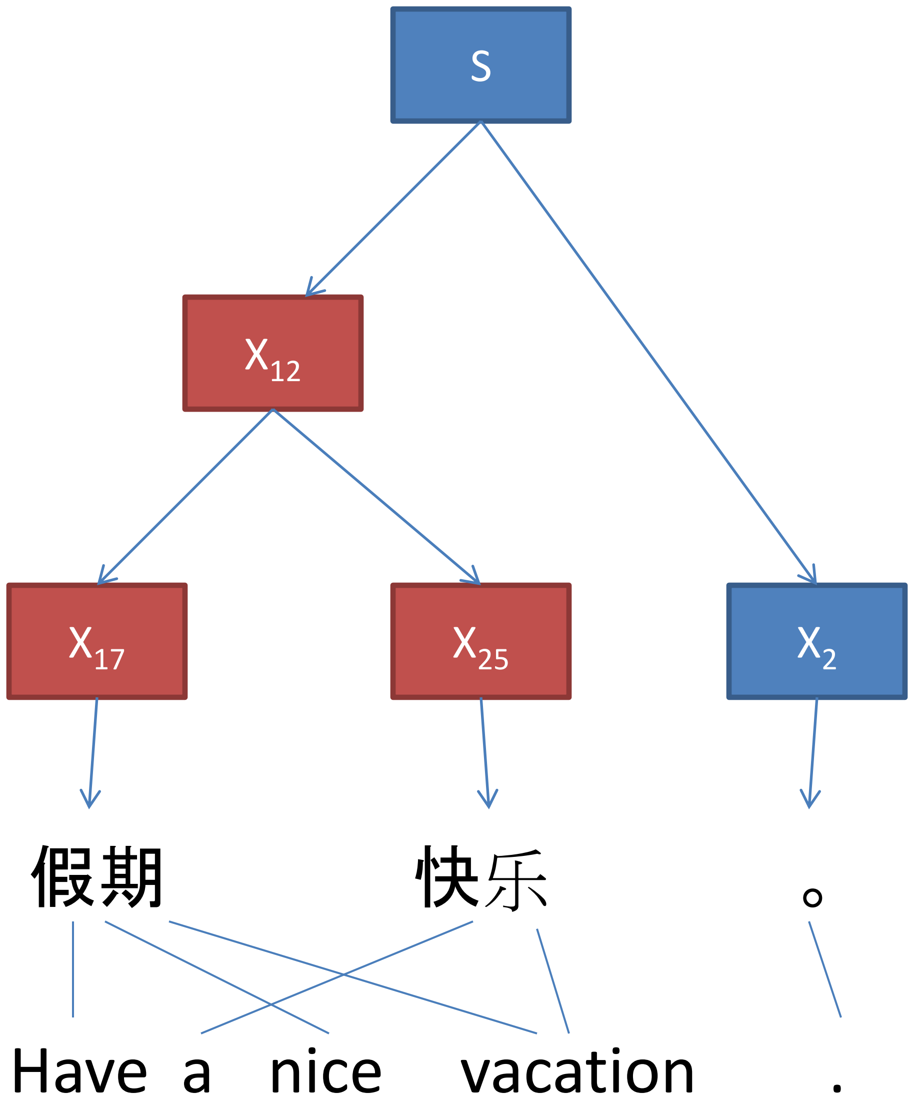
Discriminative Training

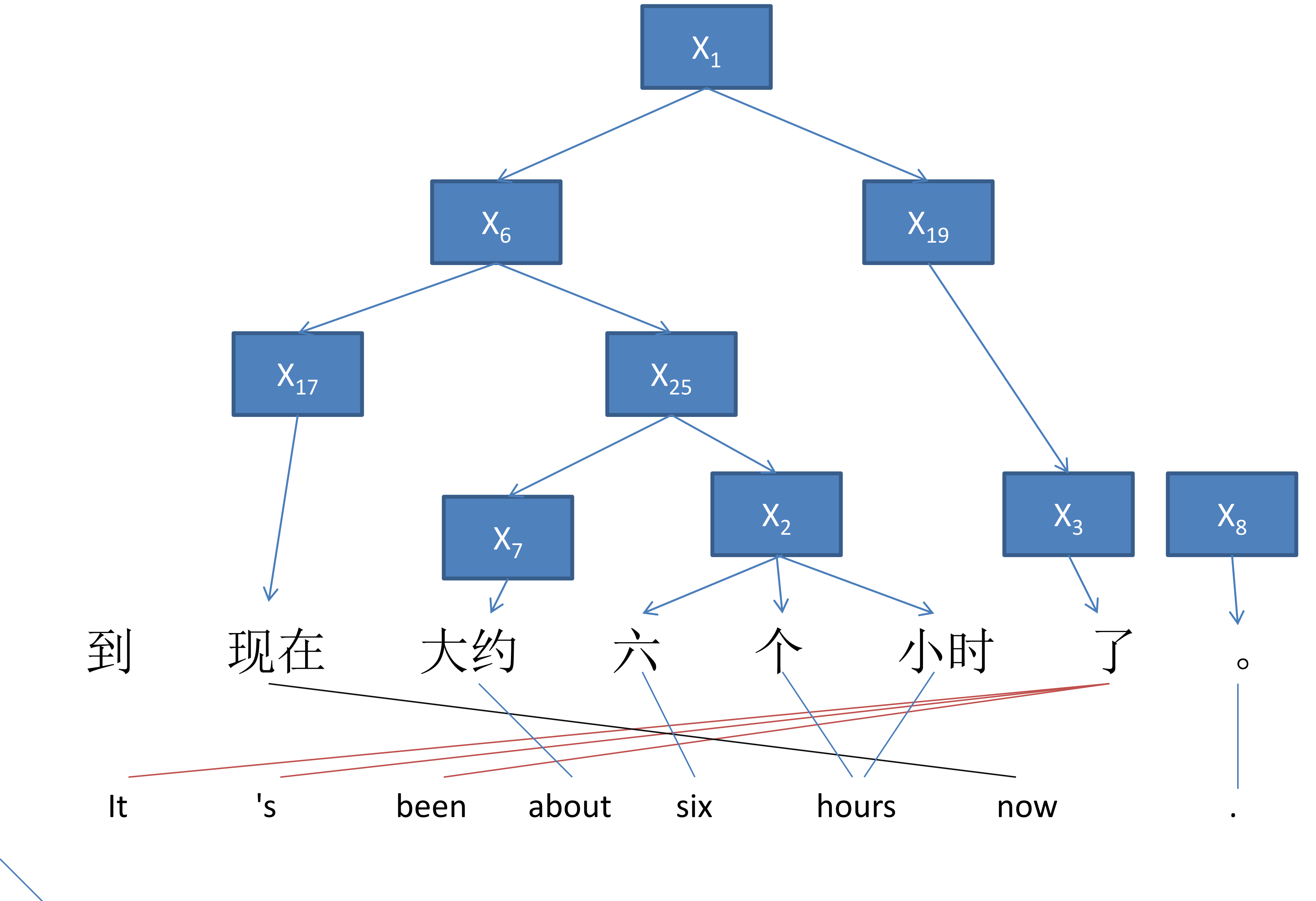
Vladimir Eidelman
Ziyuan Wang

Motivation

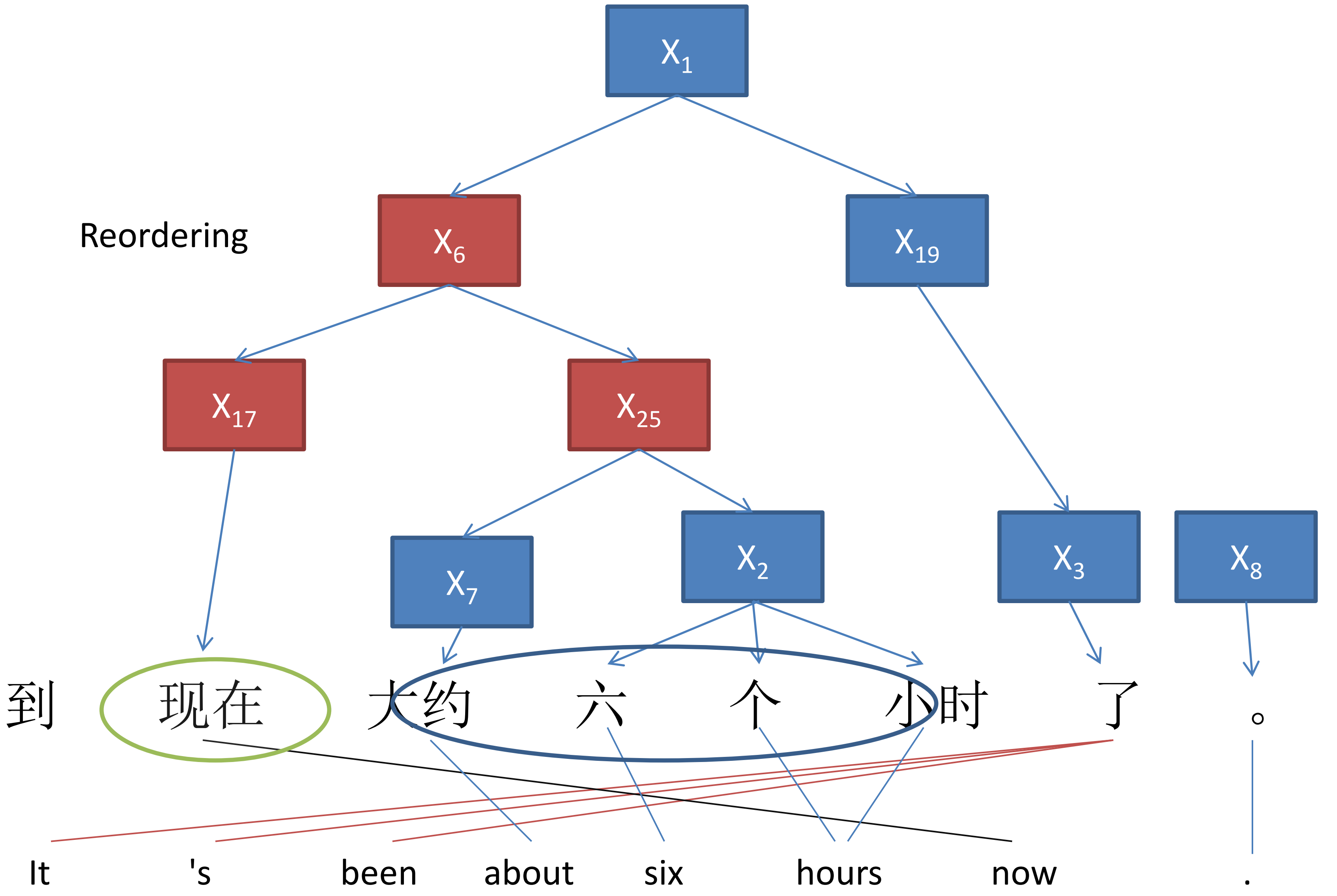
- Extract sparser features from grammars
  - Source Syntax
  - Target Syntax
  - Source Context
  - Glue Features
  - OOV
  - Backoff Rule
  - Morpheme construction
- …

Source Syntactic Features

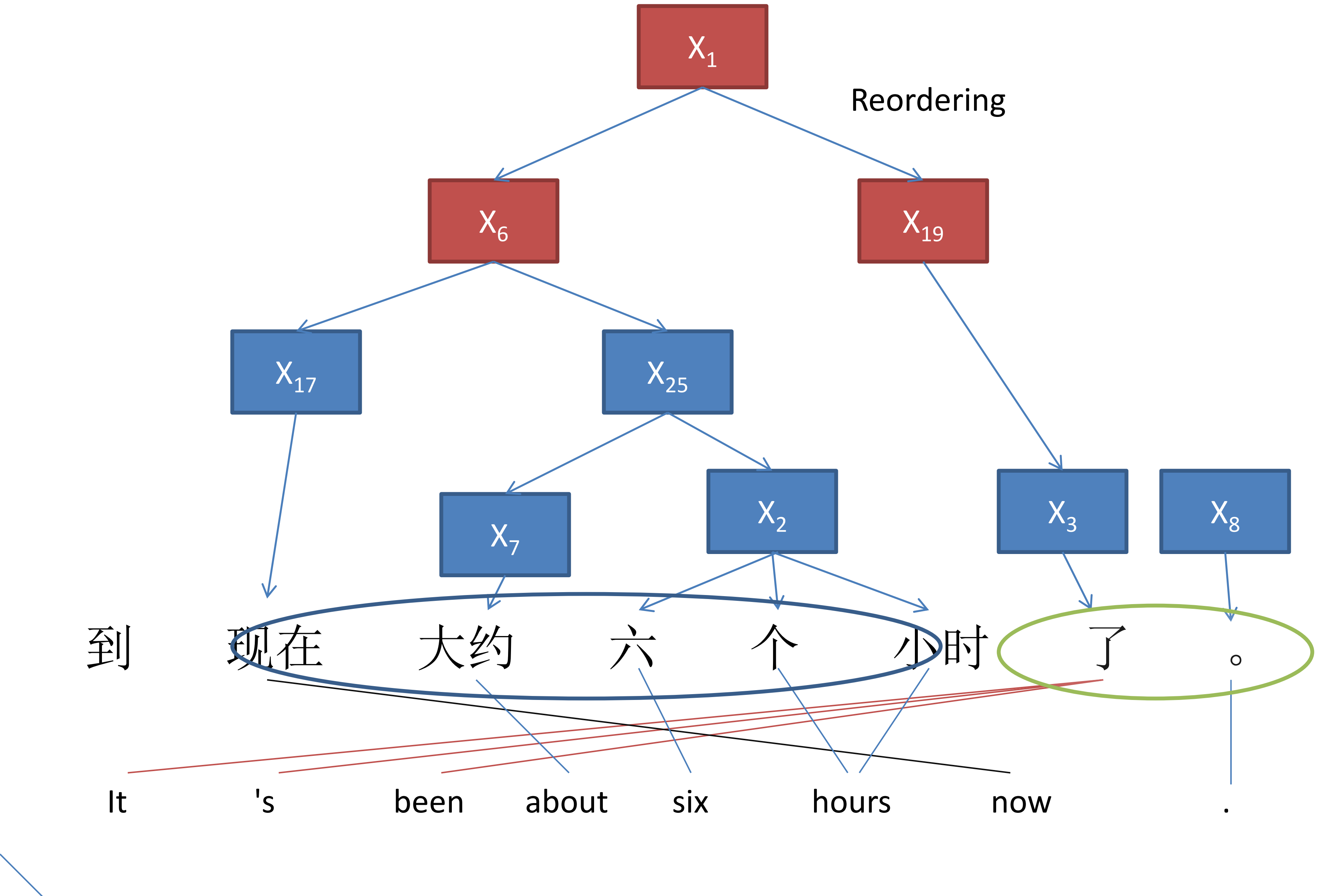




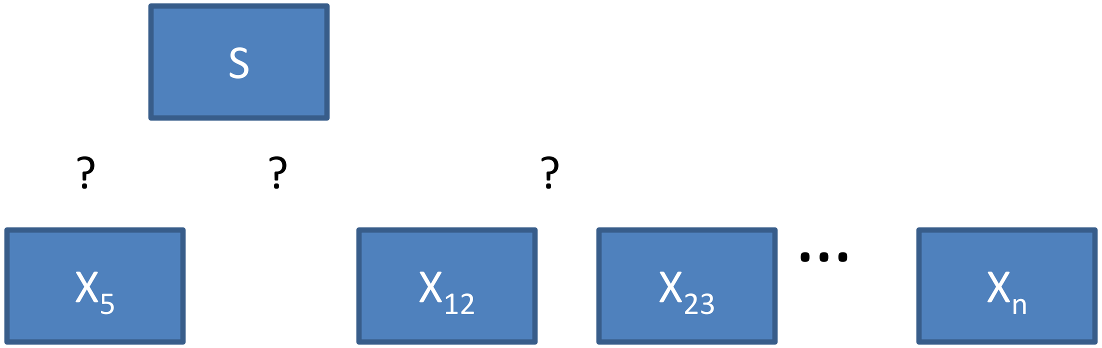
X1
X6
X19
X17
X25
X7
X2
X3
X8
到
现在
大约
六
个
小时
了
。
It
's
been
about
six
hours
now
.



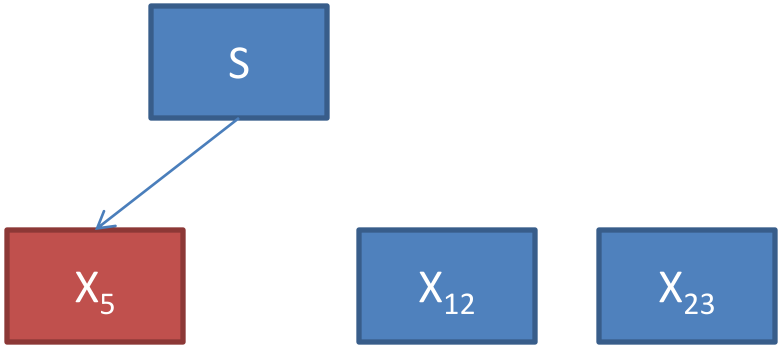
Reordering
X1
X6
X19
X17
X25
X7
X2
X3
X8
到
现在
大约
六
个
小时
了
。
It
's
been
about
six
hours
now
.


X_1
Reordering
X_6
X_{19}
X_{17}
X_{25}
X_7
X_2
X_3
X_8
到 现在 大约 六 个 小时 了 。
It 's been about six hours now .

Glue Feature

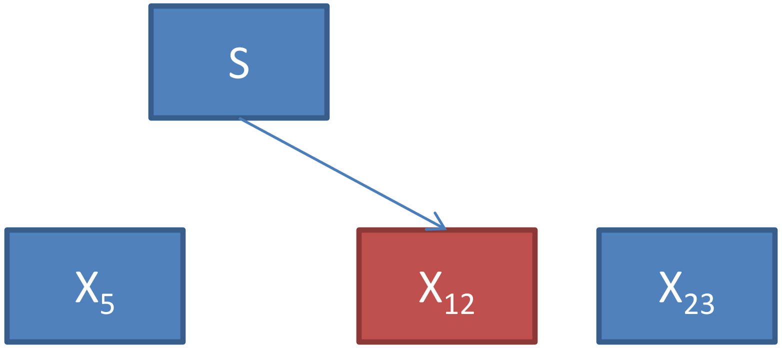


Glue Feature



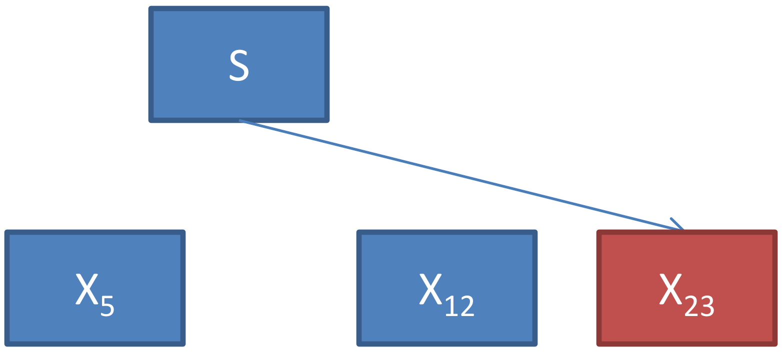
Feature: $S\_X_5 = 1$

Glue Feature



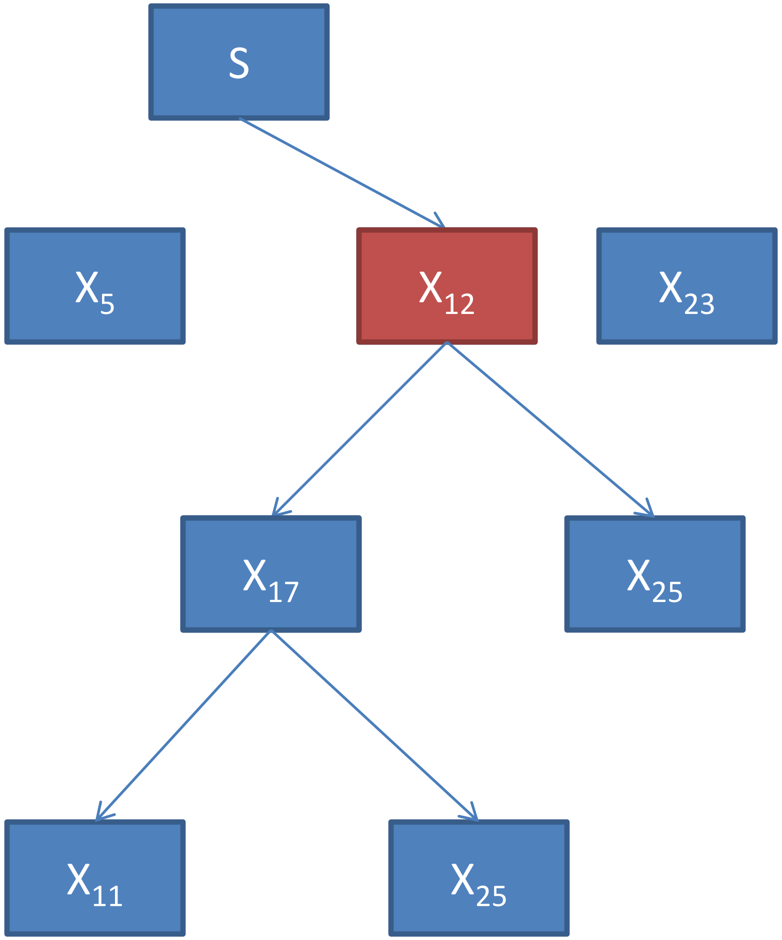
Feature: $S\_X_{12} = 1$

Glue Feature

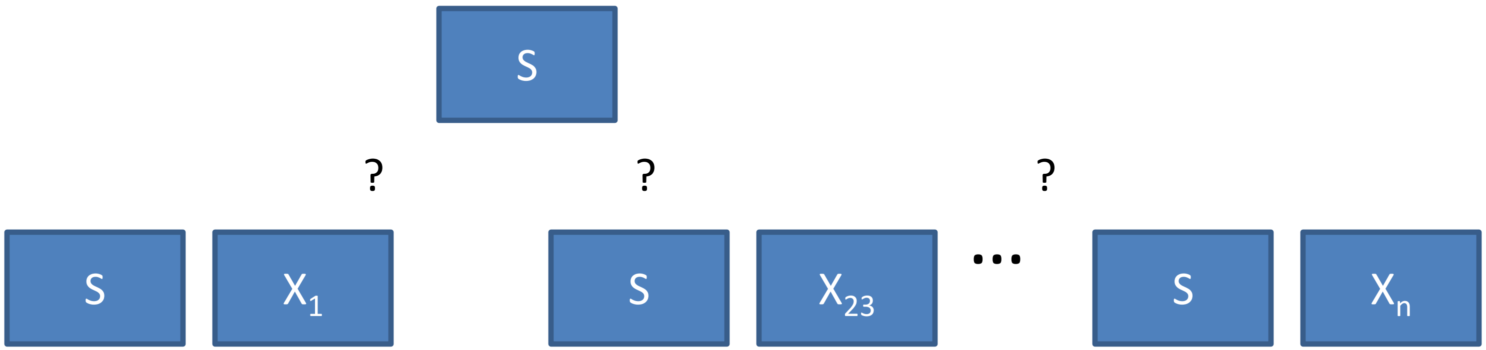


Feature: $S\_X_{23} = 1$

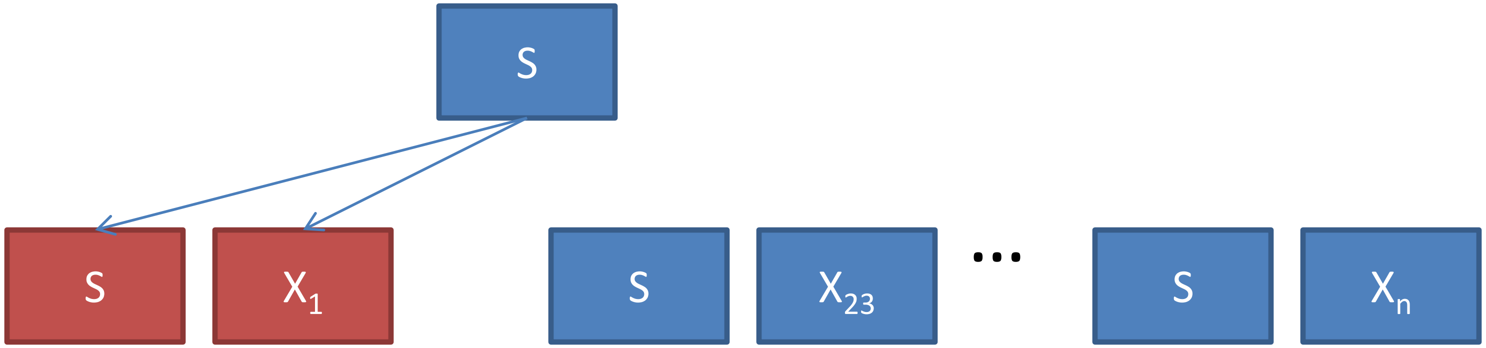
Glue Feature



Glue Feature

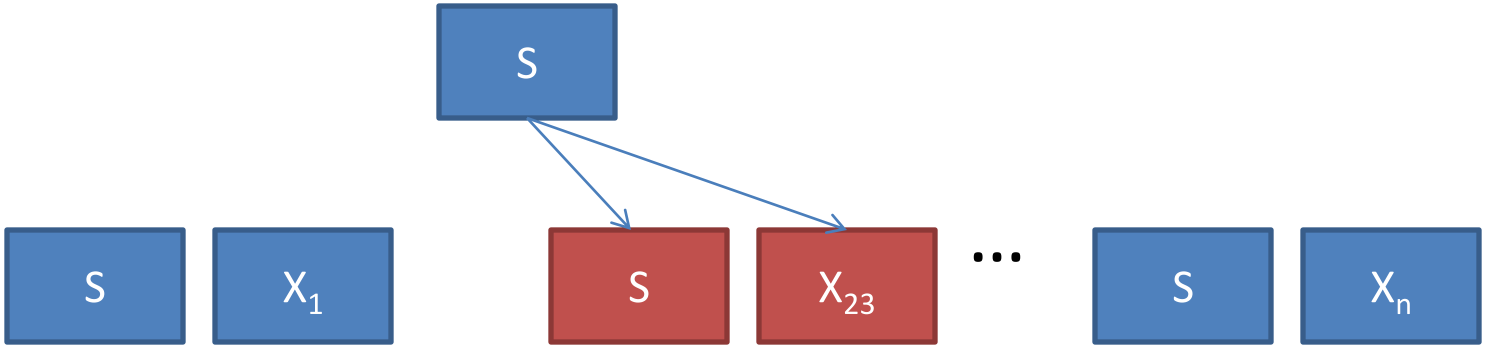


Glue Feature



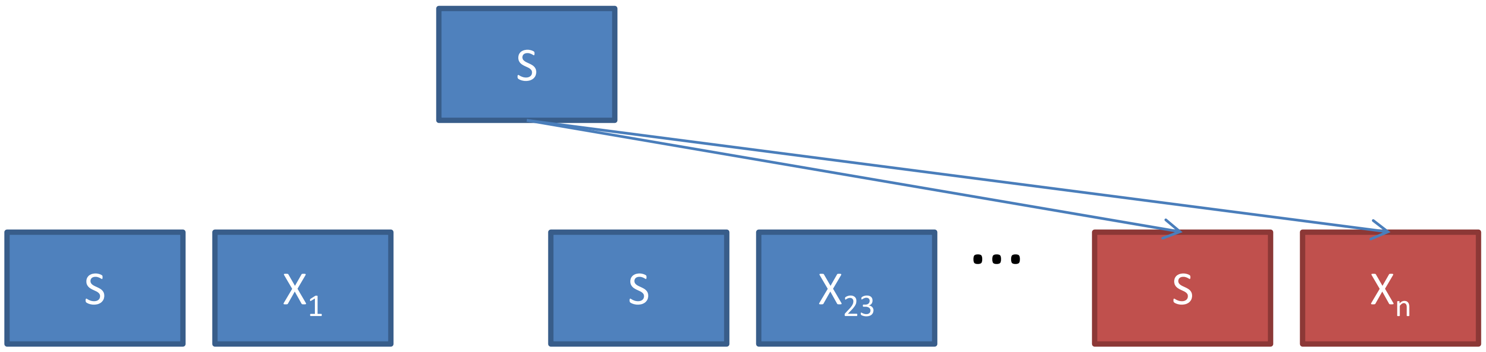
Feature: Glue_SX_1 = 1

Glue Feature



Feature: $Glue\_SX_{23} = 1$

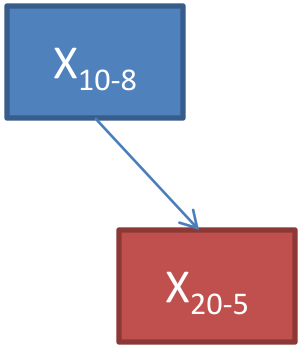
Glue Feature



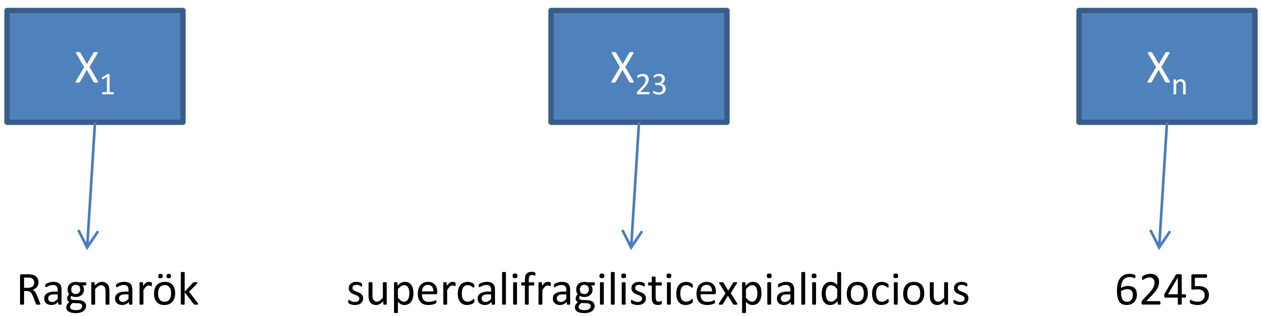
Feature: Glue_SX$_n$= 1

Backoff

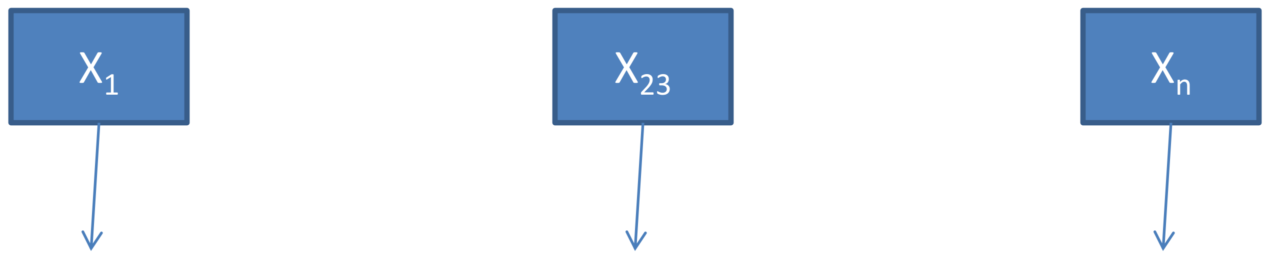
- In place of or combination with current dense feature



OOV


X_1
X_{23}
X_n
Ragnarök
supercalifragilisticexpialidocious
6245

OOV



Ragnarök	supercalifragilisticexpialidocious	6245
Noun?	Adjective?	Number?

We want to…

- optimize model parameters to maximize translation quality on some metric (BLEU)
- do discriminative training so we can have features that directly help translation
- have thousands++ features

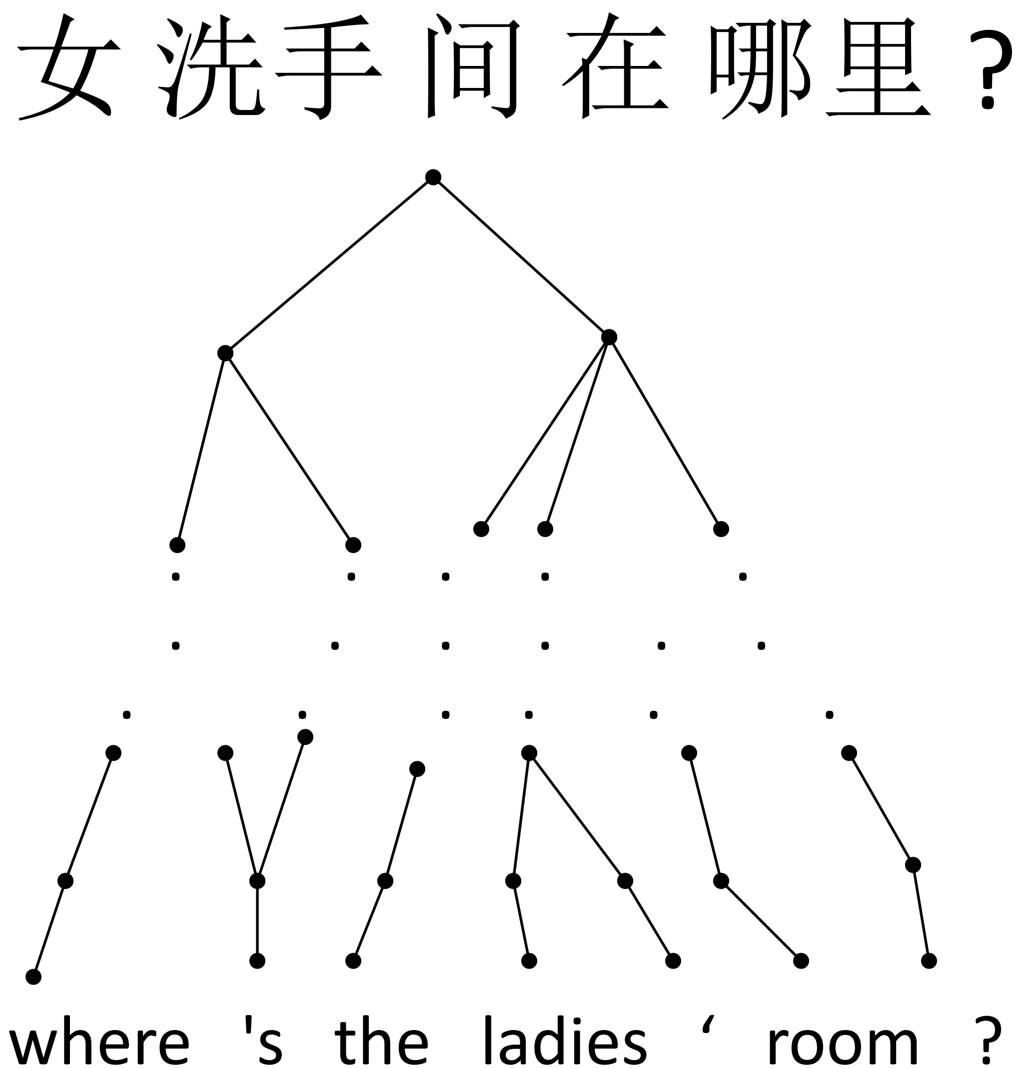
Motivation

- Minimum Error Rate Training
  - Does not scale well to more than handful of features
    - P(e) – Language Model
    - P(f|e) – Translation Rule
    - Pass through penalty
- Alternative approaches
  - Expected BLEU training
  - MIRA
- Evaluation
  - Language invariability (parameters, iterations, etc)
  - Standardizes comparison

Training Comparison

	MERT	**MIRA**	**Expected BLEU**
Type	1-best	Margin-based	Probabilistic
Objective	Minimize error	Minimize loss augmented score	Minimize expected error
Optimization	Line search	QP	Gradient based
Limitations	Direction of search unknown	Approximation of reference	Approximate expectation

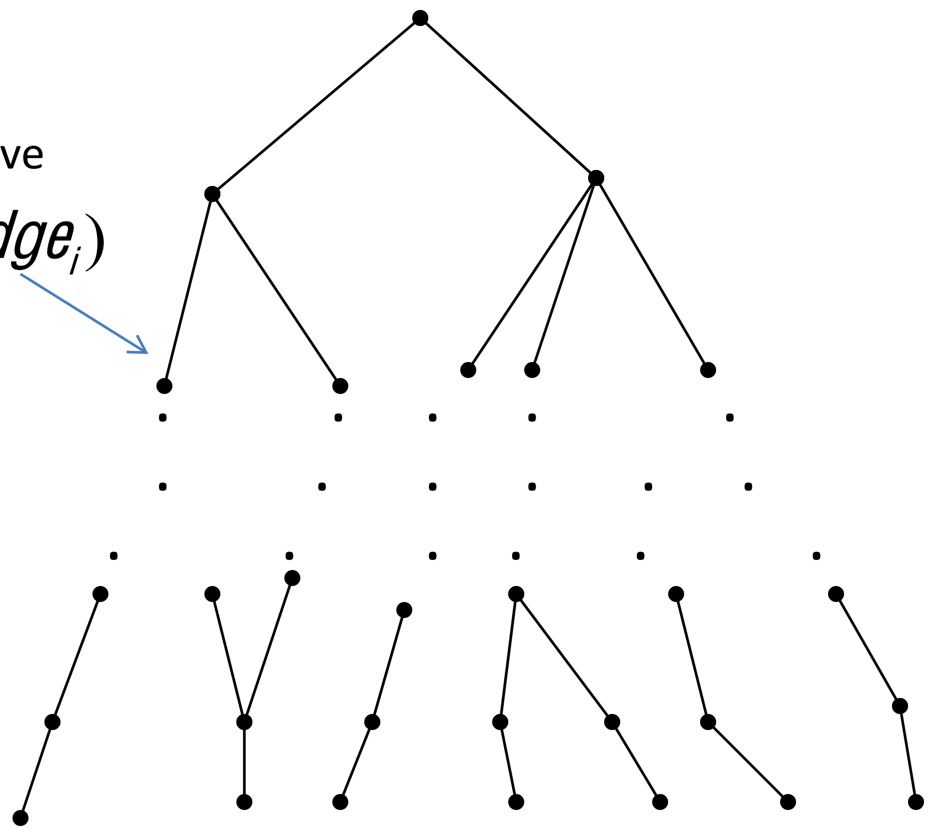
MIRA and Expected BLEU



女 洗手 间 在 哪里 ?

Decomposable Objective

$$s(edge_i) = w \cdot f(edge_i)$$

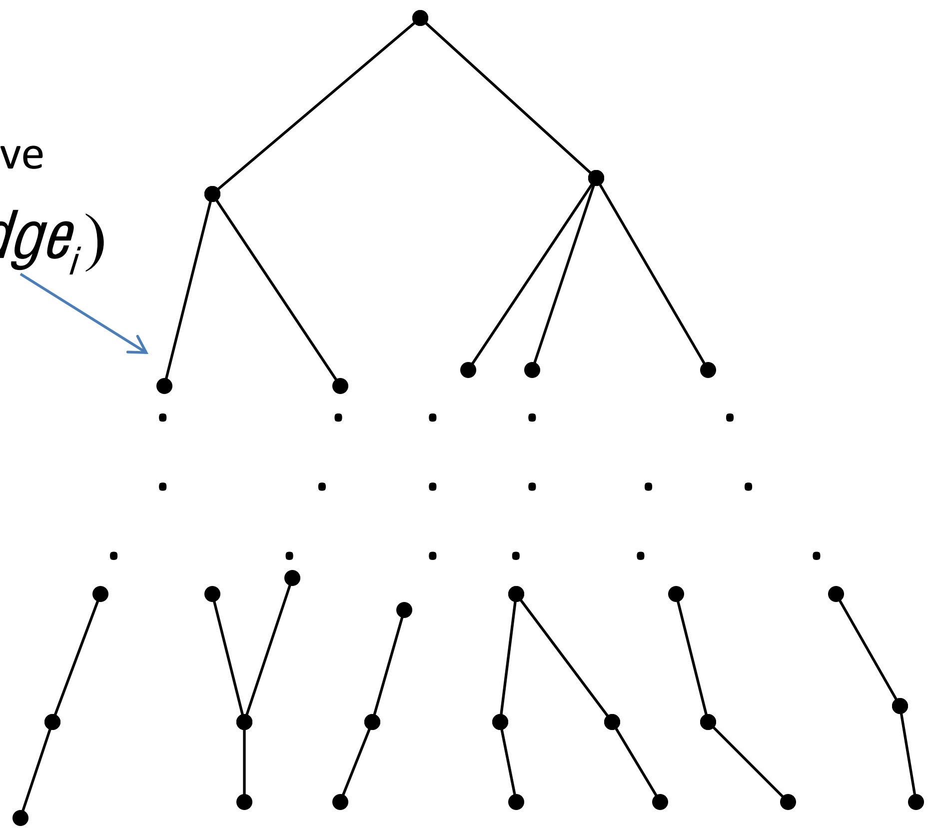


where 's the ladies ‘ room ?

女 洗手 间 在 哪里 ?

Decomposable Objective

$$s(edge_i) = w \cdot f(edge_i)$$



where 's the ladies ‘ room ?

$$\sum_{edge \in derivation} s(edge) = S(e)$$

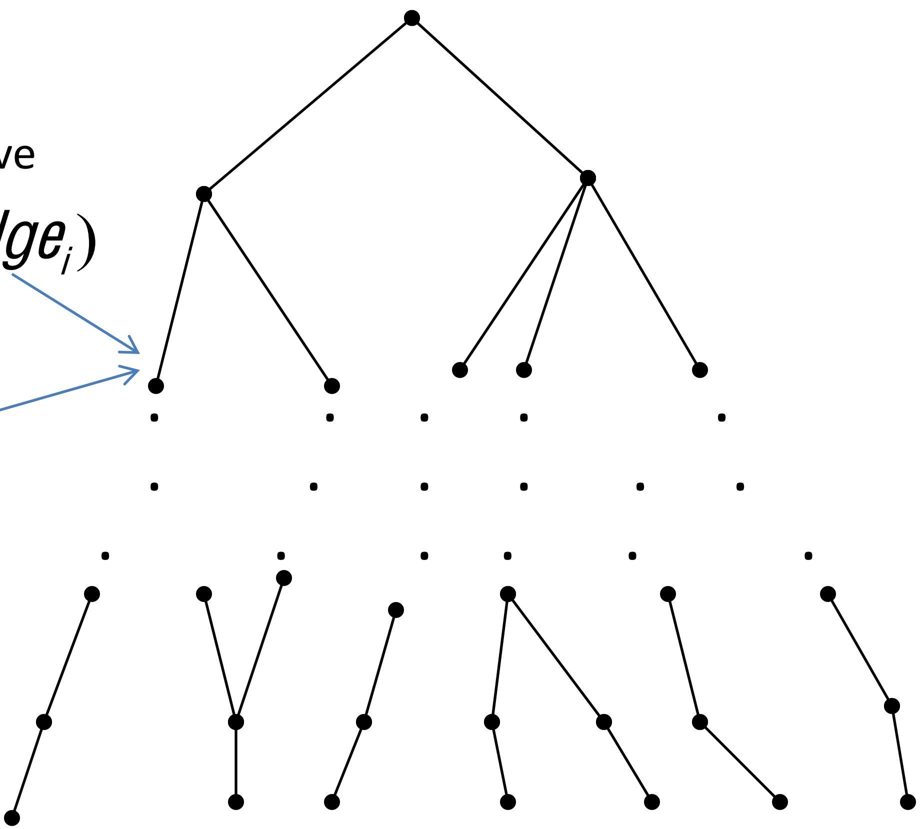
女洗手间在哪里？

Decomposable Objective

$$s(edge_i) = w \cdot f(edge_i)$$

Loss

$$approxBLEU$$



where 's the ladies ' room ?

$$\sum_{edge \in derivation} s(edge) = S(e)$$

女 洗 手 间 在 哪 里 ？

Sentence level loss

Decomposable Objective

$p(edge_i) = s(edge_i) = w \cdot f(edge_i)$

Expected BLEU

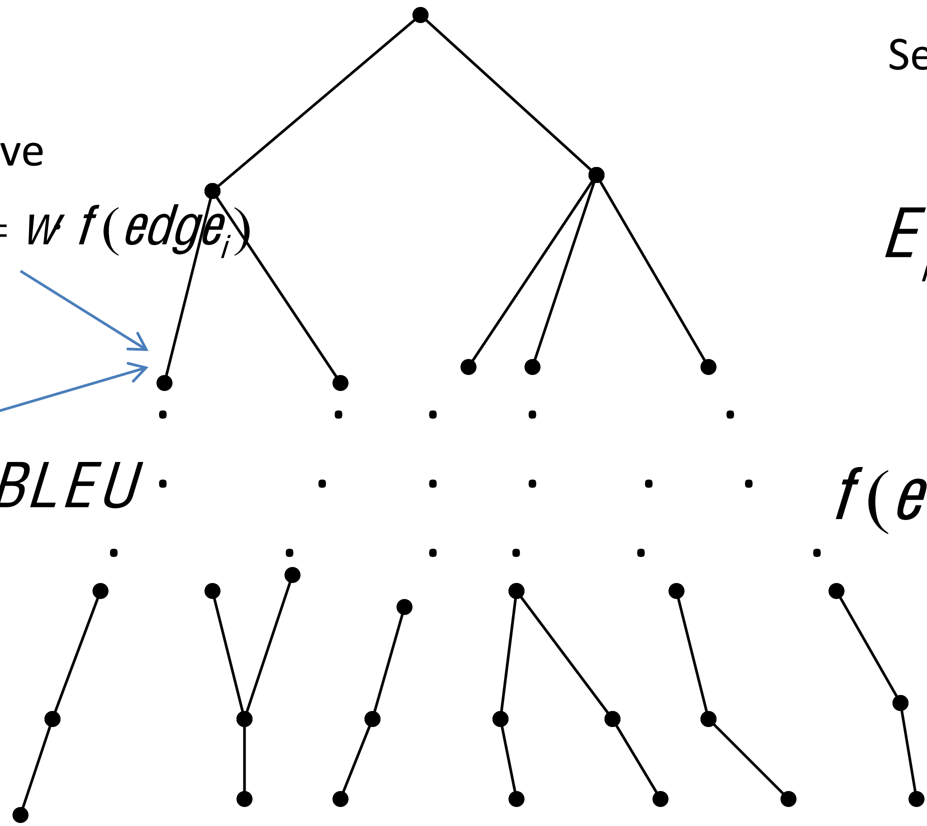
$E_p[\, f(edge_i)]$

Loss

$f(edge_i) = approxBLEU$

MIRA

$f(edge_i) + p(edge_i)$



where 's the ladies ‘ room ?

$$\sum_{edge \in derivation} s(edge) = S(e)$$

Margin Infused Relaxed Algorithm (MIRA)

- Online Large-Margin Learning

- Crammer and Singer (2003)
  - Multi-class classification
- Taskar (2003)
  - Extension to structured value prediction
- Watanabe (2007), Chiang (2009)
  - Application to MT

Basic Learning Algorithm

Training data: D = (**x**,**y**)

$$weight_0 = 0, total = 0, c = 0$$

$$for \ \ iteration \ 1 \rightarrow n$$

$$\quad for \ \ d = (x, y) \in D$$

$$\qquad weight_{c+1} = update \ \ weight_c \ \ with \, d$$

$$\qquad total = total + weight_{c+1}$$

$$\qquad c = c + 1$$

$$weight = \frac{total}{n \times size(D)}$$

Update

$$s(x, y) = w \cdot f(x, y)$$

- Learn w so that correct outputs are given higher score than incorrect ones

$$\min \| w_{i+1} - w_i \|$$

  – Keep the norm of the change to the weights as small as possible

- Subject to margin constraints:

$$s(x, y) - s(x, z) \geq Loss(y, z)$$

  – Create margin between correct instance y, and incorrect instance z at least as large as the Loss of z
  – for all z which are possible labels of x

k-best MIRA

$$Training\ data: D = (e, f)$$

$$weight_0 = 0, total = 0, c = 0$$

$$for\ iteration\ 1 \rightarrow n$$

$$for\ d = (x, y) \in D$$

$$Generate\ kbest(f) = \{e \ldots e_k\}$$

$$Generate\ margin\ constraints\ \forall e \in kbest(f)$$

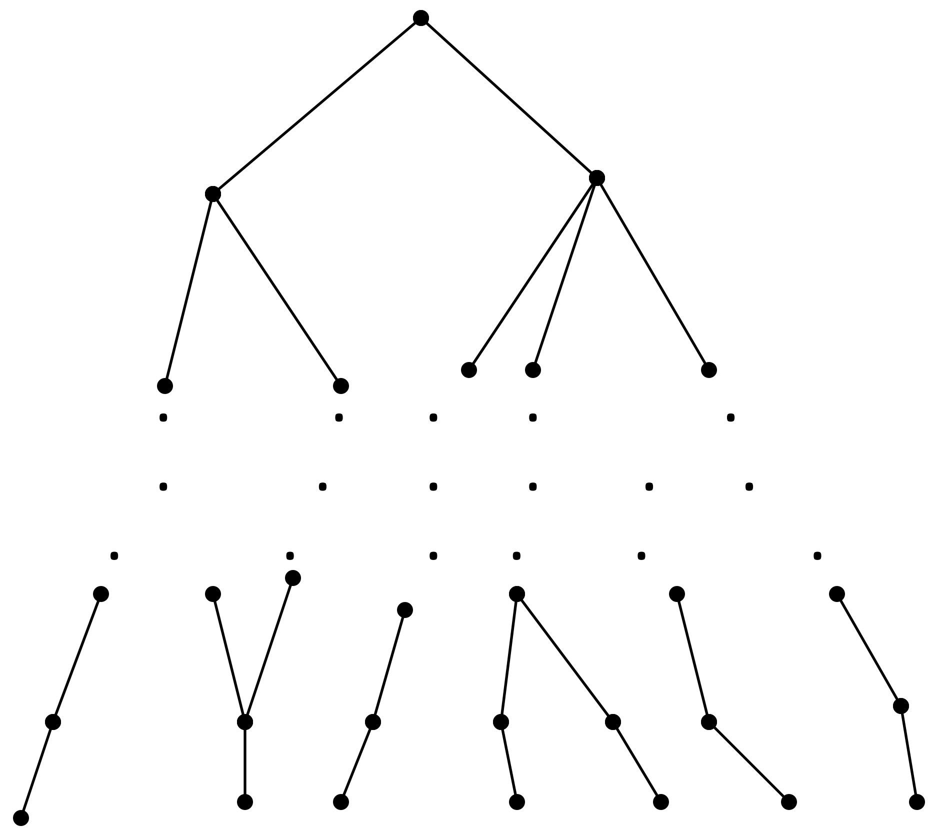
$$weight_{c+1} = update\ weight_c\ with\ d$$

$$total = total + weight_{c+1}$$

$$c = c + 1$$

$$weight = \frac{total}{n \times size(D)}$$

女 洗手 间 在 哪里 ?

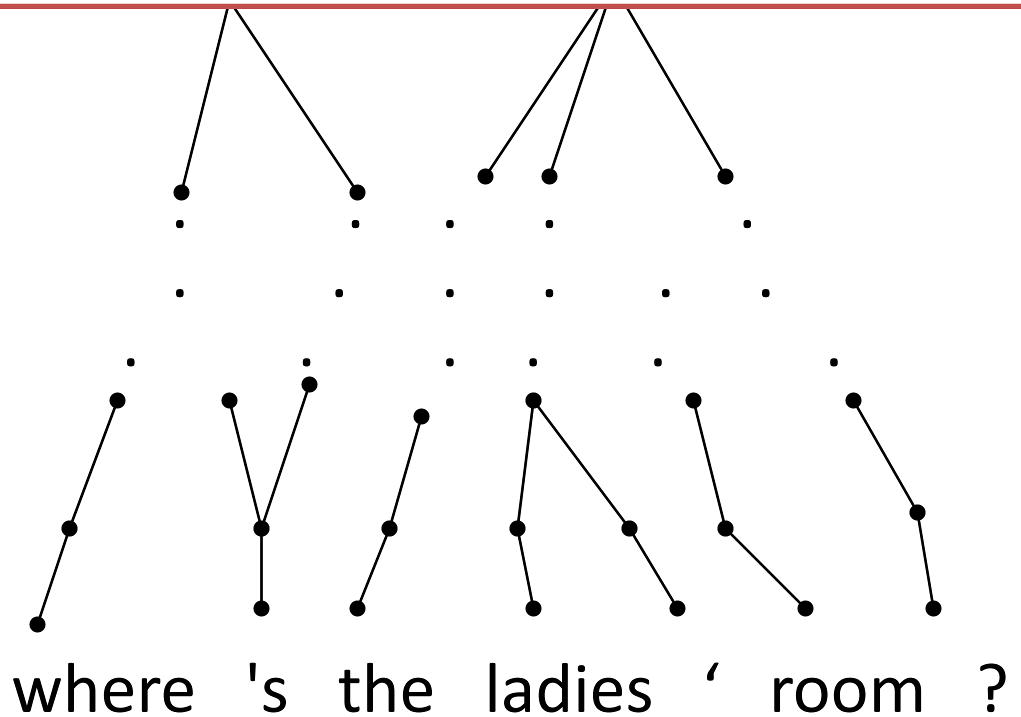


where 's the ladies ‘ room ?

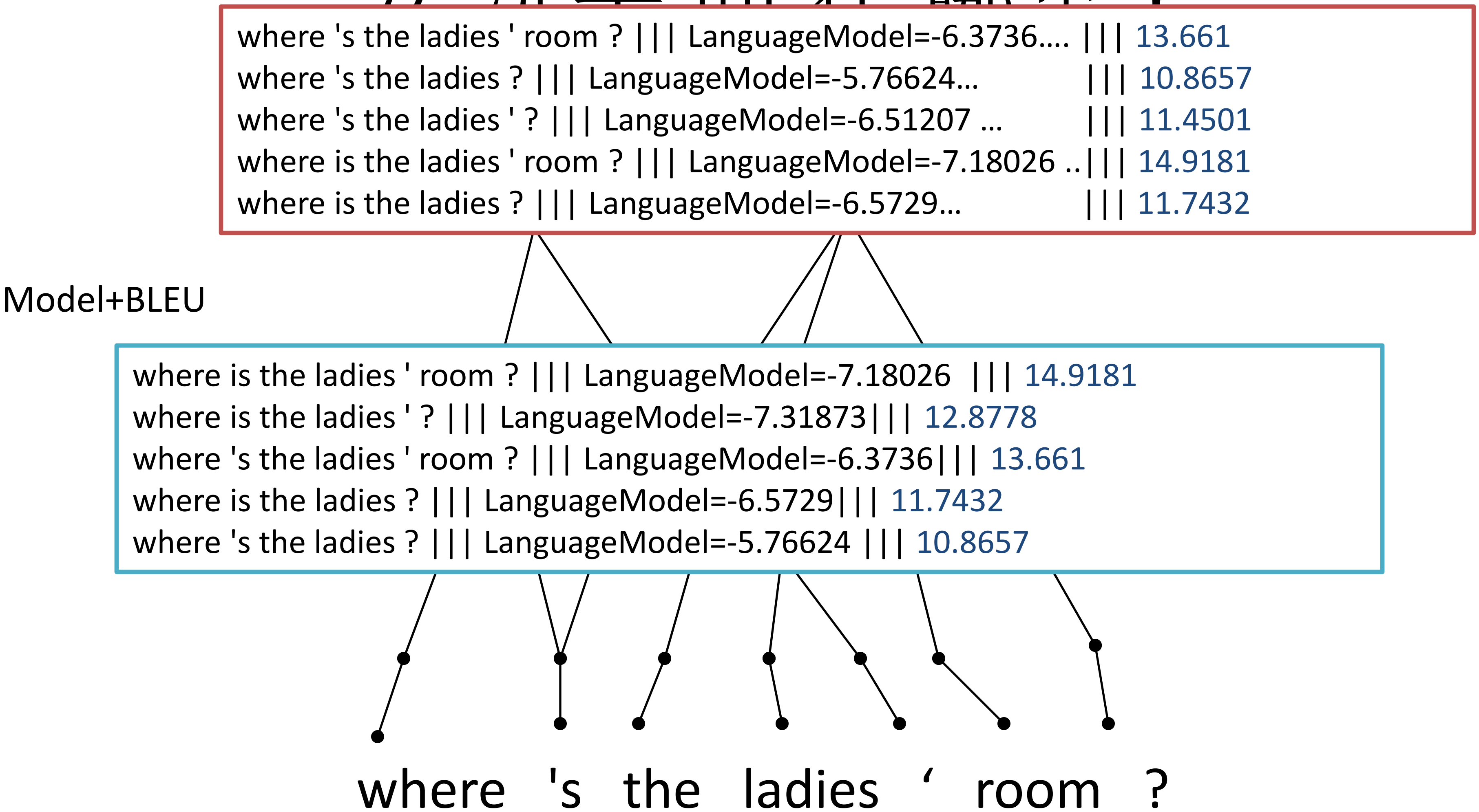
女洗手间在哪里？

Model

where 's the ladies ' room ? ||| LanguageModel=-6.3736.... ||| 13.661
where 's the ladies ? ||| LanguageModel=-5.76624... ||| 10.8657
where 's the ladies ' ? ||| LanguageModel=-6.51207 ... ||| 11.4501
where is the ladies ' room ? ||| LanguageModel=-7.18026 ..||| 14.9181
where is the ladies ? ||| LanguageModel=-6.5729... ||| 11.7432



女洗手间在哪里？



where 's the ladies ' room ? ||| LanguageModel=-6.3736.... ||| 13.661
where 's the ladies ? ||| LanguageModel=-5.76624... ||| 10.8657
where 's the ladies ' ? ||| LanguageModel=-6.51207 ... ||| 11.4501
where is the ladies ' room ? ||| LanguageModel=-7.18026 ..||| 14.9181
where is the ladies ? ||| LanguageModel=-6.5729... ||| 11.7432
Model+BLEU
where is the ladies ' room ? ||| LanguageModel=-7.18026 ||| 14.9181
where is the ladies ' ? ||| LanguageModel=-7.31873||| 12.8778
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the ladies ? ||| LanguageModel=-6.5729||| 11.7432
where 's the ladies ? ||| LanguageModel=-5.76624 ||| 10.8657
where 's the ladies ‘ room ?

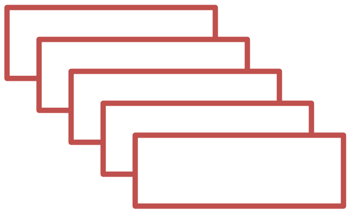
女洗手间在哪里？

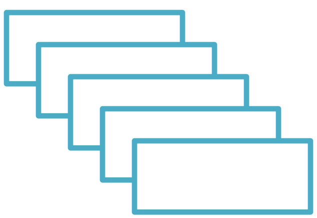
where 's the ladies ' room ? ||| LanguageModel=-6.3736.... ||| 13.661
where 's the ladies ? ||| LanguageModel=-5.76624... ||| 10.8657
where 's the ladies ' ? ||| LanguageModel=-6.51207 ... ||| 11.4501
where is the ladies ' room ? ||| LanguageModel=-7.18026 ..||| 14.9181
where is the ladies ? ||| LanguageModel=-6.5729... ||| 11.7432

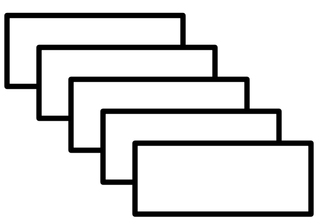
where is the ladies ' room ? ||| LanguageModel=-7.18026 ||| 14.9181
where is the ladies ' ? ||| LanguageModel=-7.31873||| 12.8778
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the ladies ? ||| LanguageModel=-6.5729||| 11.7432
where 's the ladies ? ||| LanguageModel=-5.76624 ||| 10.8657

Model - BLEU

where is the bus depot for the ladies ' room ? ||| LanguageModel=-10.7635 ||| 11.7463
where is the bus depot for the ladies ? ||| LanguageModel=-10.1561||| 10.0082
where is the bus depot for the ladies ' ? ||| LanguageModel=-10.902...||| 10.1763
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the bus depot for the ladies ' room . ||| LanguageModel=-11.1228 ||| 10.8613







Model	**47.96**
Model+ BLEU	54.08
Model- BLEU	24.10

where 's the ladies ' room ? ||| LanguageModel=-6.3736....||| 13.661
where 's the ladies ? ||| LanguageModel=-5.76624... ||| 10.8657
where 's the ladies ' ? ||| LanguageModel=-6.51207 ... ||| 11.4501
where is the ladies ' room ? ||| LanguageModel=-7.18026 .. ||| 14.9181
where is the ladies ? ||| LanguageModel=-6.5729... ||| 11.7432

Oracle Translation

where is the ladies ' room ? ||| LanguageModel=-7.18026 ||| 14.9181
where is the ladies ' ? ||| LanguageModel=-7.31873||| 12.8778
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the ladies ? ||| LanguageModel=-6.5729||| 11.7432
where 's the ladies ? ||| LanguageModel=-5.76624 ||| 10.8657

where is the bus depot for the ladies ' room ? ||| LanguageModel=-10.7635 ||| 11.7463
where is the bus depot for the ladies ? ||| LanguageModel=-10.1561||| 10.0082
where is the bus depot for the ladies ' ? ||| LanguageModel=-10.902...||| 10.1763
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the bus depot for the ladies ' room . ||| LanguageModel=-11.1228 ||| 10.8613

where 's the ladies ' room ? ||| LanguageModel=-6.3736....||| 13.661
where 's the ladies ? ||| LanguageModel=-5.76624... ||| 10.8657
where 's the ladies ' ? ||| LanguageModel=-6.51207 ... ||| 11.4501
where is the ladies ' room ? ||| LanguageModel=-7.18026 .. ||| 14.9181
where is the ladies ? ||| LanguageModel=-6.5729... ||| 11.7432

Δf(e) *Loss*

where is the ladies ' room ? ||| LanguageModel=-7.18026 ||| 14.9181
where is the ladies ' ? ||| LanguageModel=-7.31873||| 12.8778
where 's the ladies ' room ? ||| LanguageModel=-6.3736 ||| 13.661
where is the ladies ? ||| LanguageModel=-6.5729||| 11.7432
where 's the ladies ? ||| LanguageModel=-5.76624 ||| 10.8657

where is the bus depot for the ladies ' room ? ||| LanguageModel=-10.7635 ||| 11.7463
where is the bus depot for the ladies ? ||| LanguageModel=-10.1561||| 10.0082
where is the bus depot for the ladies ' ? ||| LanguageModel=-10.902...||| 10.1763
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the bus depot for the ladies ' room . ||| LanguageModel=-11.1228 ||| 10.8613

where 's the ladies ' room ? ||| LanguageModel=-6.3736....||| 13.661
where 's the ladies ? ||| LanguageModel=-5.76624... ||| 10.8657
where 's the ladies ' ? ||| LanguageModel=-6.51207 ... ||| 11.4501
where is the ladies ' room ? ||| LanguageModel=-7.18026 .. ||| 14.9181
where is the ladies ? ||| LanguageModel=-6.5729... ||| 11.7432

Δf(e)

Loss

where is the ladies ' room ? ||| LanguageModel=-7.18026 ||| 14.9181
where is the ladies ' ? ||| LanguageModel=-7.31873||| 12.8778
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the ladies ? ||| LanguageModel=-6.5729||| 11.7432
where 's the ladies ? ||| LanguageModel=-5.76624 ||| 10.8657

where is the bus depot for the ladies ' room ? ||| LanguageModel=-10.7635 ||| 11.7463
where is the bus depot for the ladies ? ||| LanguageModel=-10.1561||| 10.0082
where is the bus depot for the ladies ' ? ||| LanguageModel=-10.902...||| 10.1763
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the bus depot for the ladies ' room . ||| LanguageModel=-11.1228 ||| 10.8613

where 's the ladies ' room ? ||| LanguageModel=-6.3736....||| 13.661
where 's the ladies ? ||| LanguageModel=-5.76624... ||| 10.8657
where 's the ladies ' ? ||| LanguageModel=-6.51207 ... ||| 11.4501
where is the ladies ' room ? ||| LanguageModel=-7.18026 .. ||| 14.9181
where is the ladies ? ||| LanguageModel=-6.5729... ||| 11.7432

Δf(e) *Loss*

where is the ladies ' room ? ||| LanguageModel=-7.18026 ||| 14.9181
where is the ladies ' ? ||| LanguageModel=-7.31873||| 12.8778
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the ladies ? ||| LanguageModel=-6.5729||| 11.7432
where 's the ladies ? ||| LanguageModel=-5.76624 ||| 10.8657

where is the bus depot for the ladies ' room ? ||| LanguageModel=-10.7635 ||| 11.7463
where is the bus depot for the ladies ? ||| LanguageModel=-10.1561||| 10.0082
where is the bus depot for the ladies ' ? ||| LanguageModel=-10.902...||| 10.1763
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the bus depot for the ladies ' room . ||| LanguageModel=-11.1228 ||| 10.8613

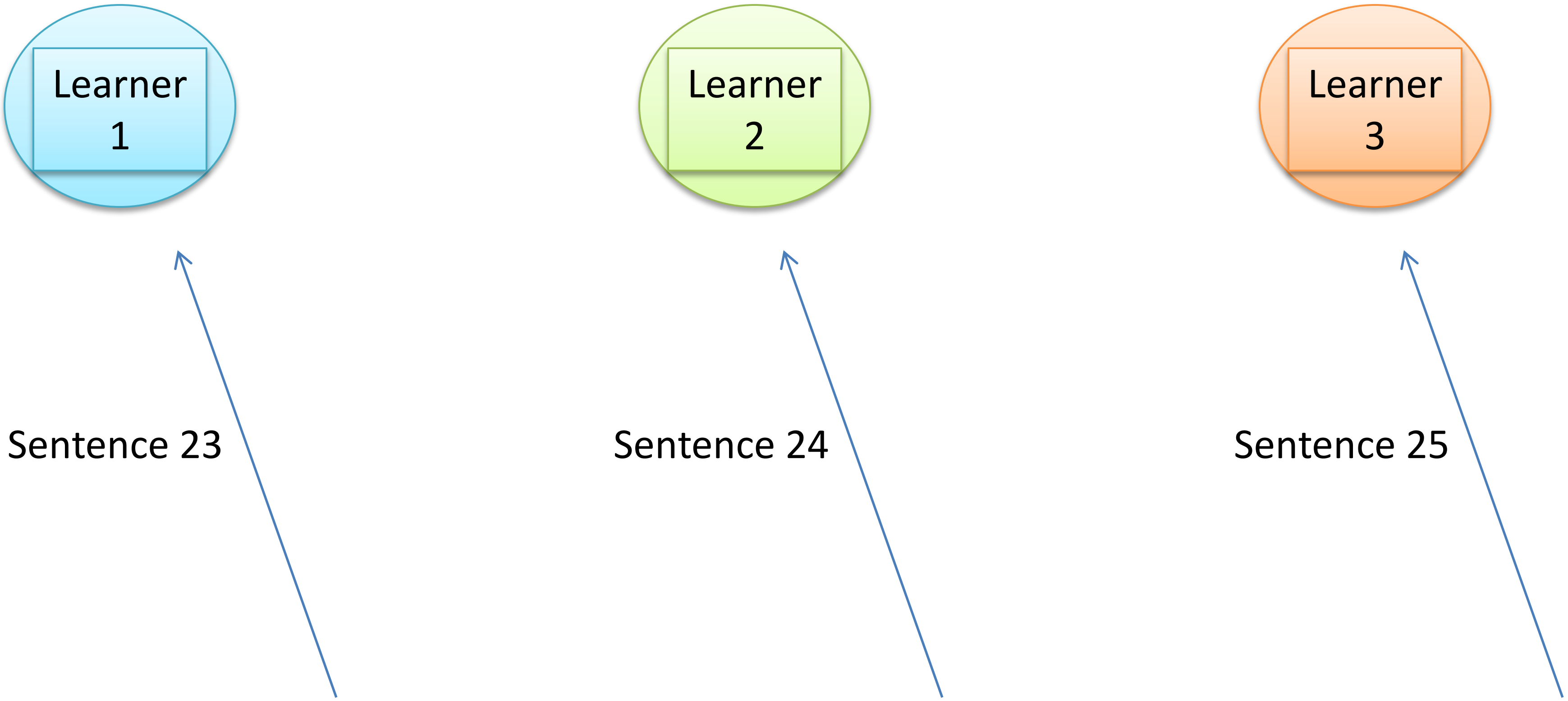
where 's the ladies ' room ? ||| LanguageModel=-6.3736....||| 13.661
where 's the ladies ? ||| LanguageModel=-5.76624... ||| 10.8657
where 's the ladies ' ? ||| LanguageModel=-6.51207 ... ||| 11.4501
where is the ladies ' room ? ||| LanguageModel=-7.18026 .. ||| 14.9181
where is the ladies ? ||| LanguageModel=-6.5729... ||| 11.7432

Δf(e) *Loss*

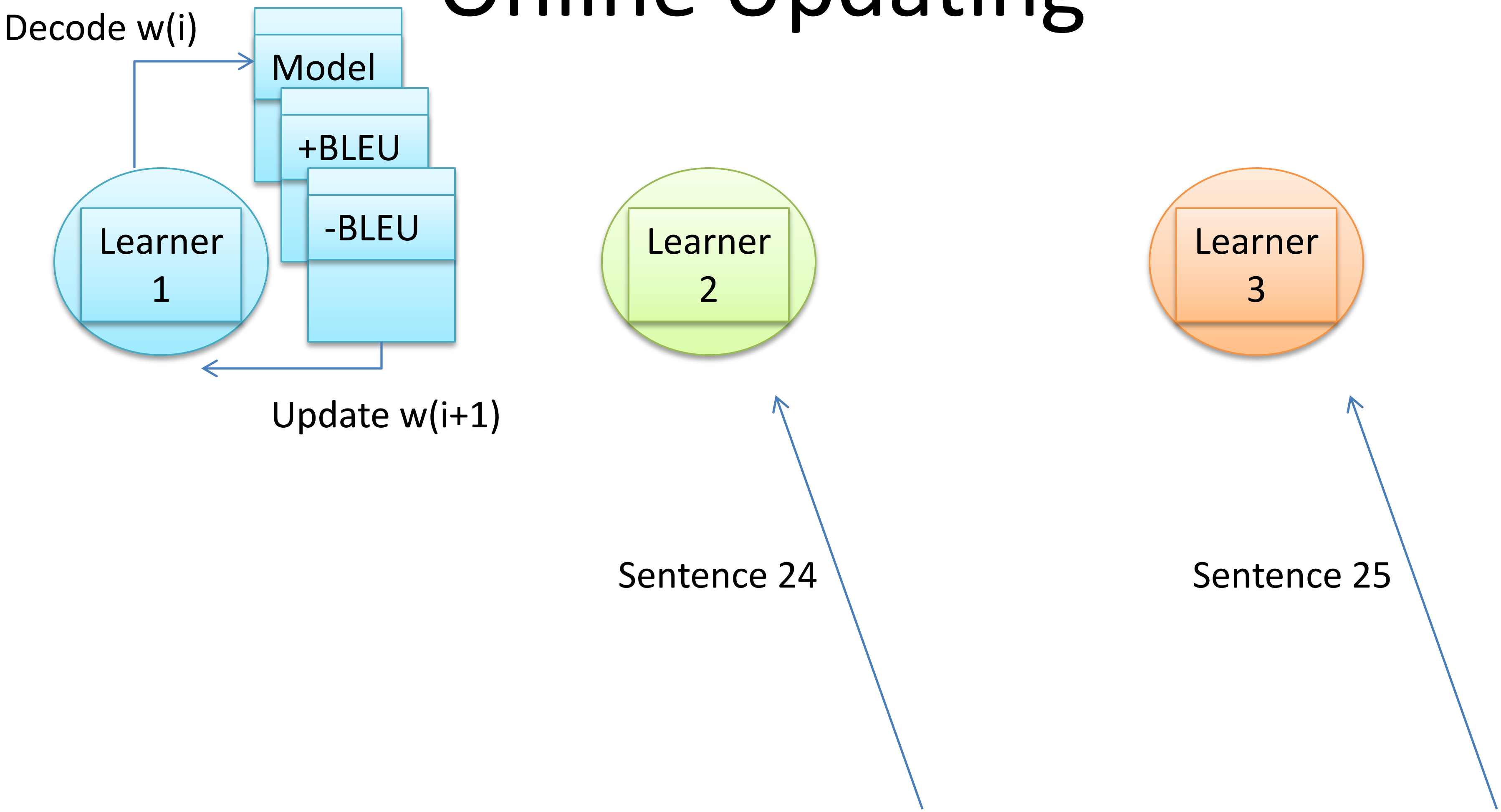
where is the ladies ' room ? ||| LanguageModel=-7.18026 ||| 14.9181
where is the ladies ' ? ||| LanguageModel=-7.31873||| 12.8778
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the ladies ? ||| LanguageModel=-6.5729||| 11.7432
where 's the ladies ? ||| LanguageModel=-5.76624 ||| 10.8657

where is the bus depot for the ladies ' room ? ||| LanguageModel=-10.7635 ||| 11.7463
where is the bus depot for the ladies ? ||| LanguageModel=-10.1561||| 10.0082
where is the bus depot for the ladies ' ? ||| LanguageModel=-10.902...||| 10.1763
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the bus depot for the ladies ' room . ||| LanguageModel=-11.1228 ||| 10.8613

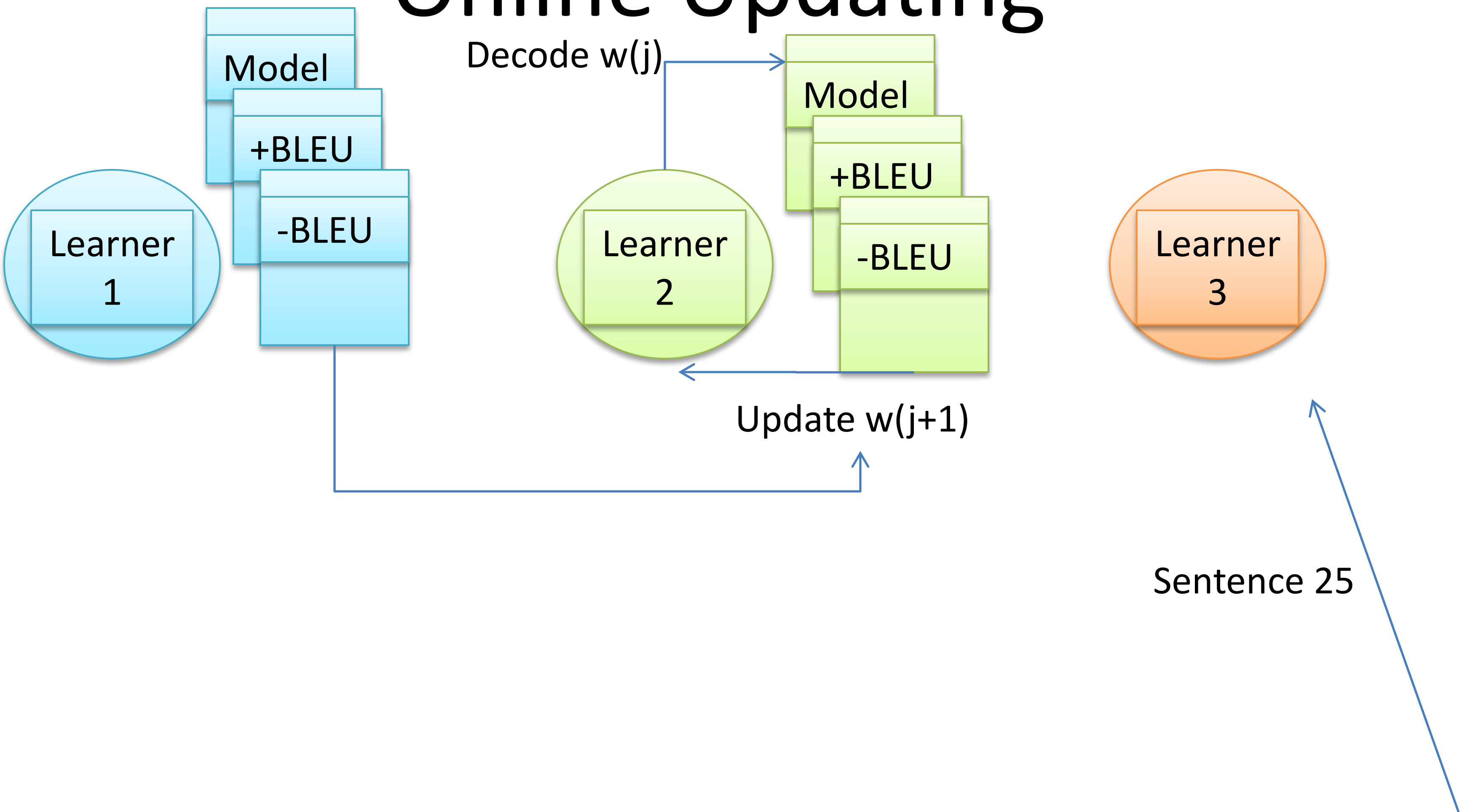
Online Updating



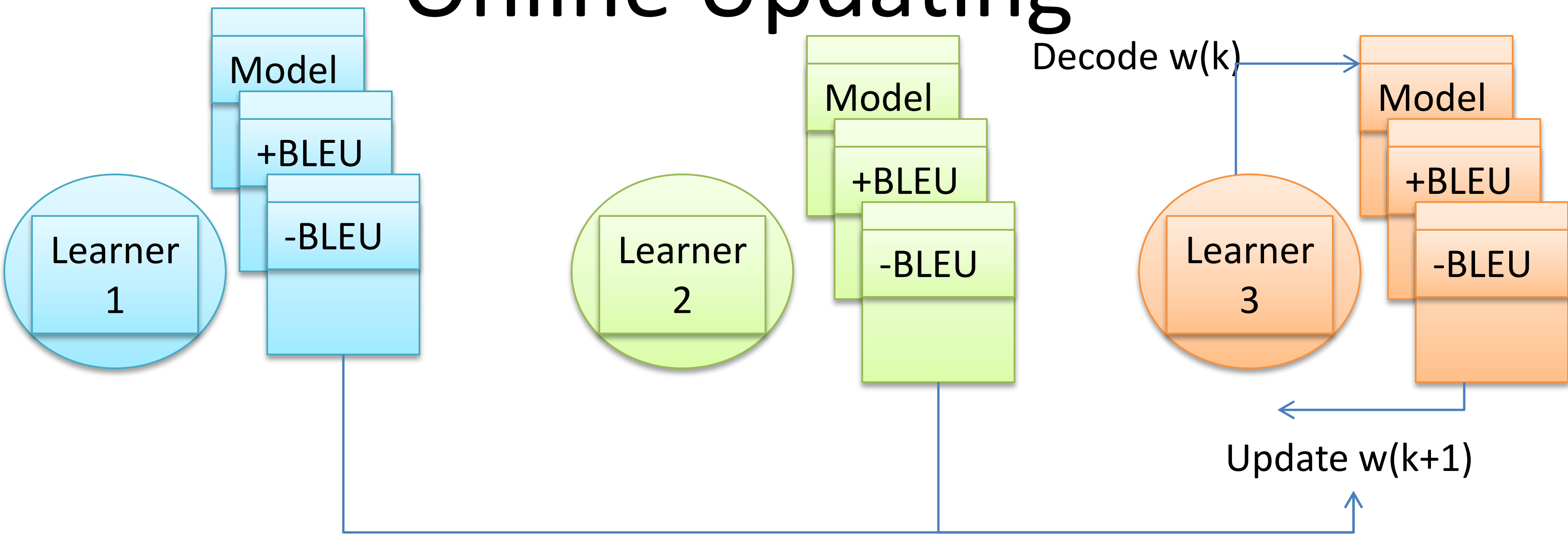
Online Updating



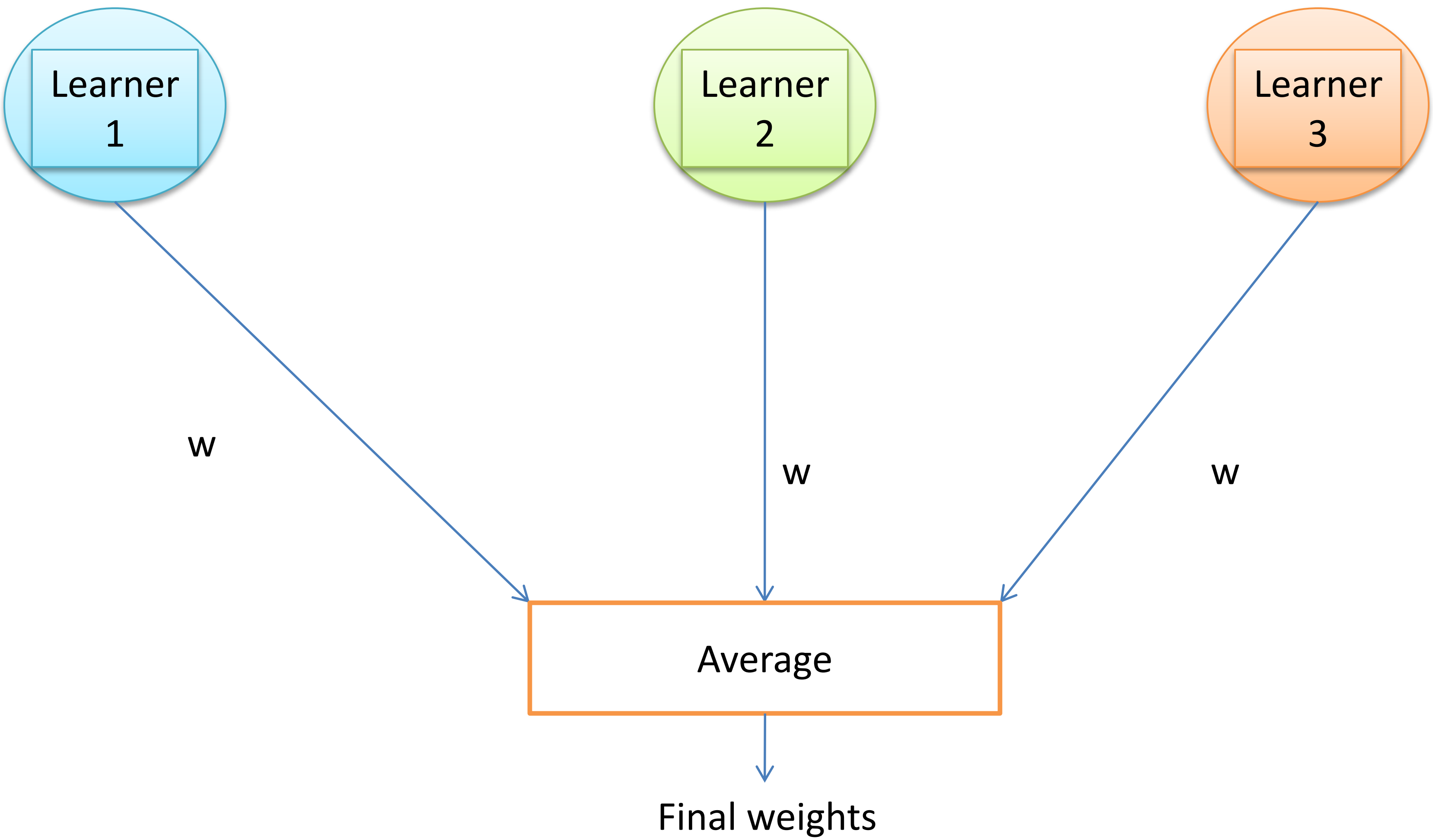
Online Updating



Online Updating



Online Updating



Expected BLEU

- BLEU is just a geometric mean of ngram precisions

Hypothesis
A B C D E F G
A C F B C D E F
C A F B D E F G

Reference
A B C D E F G
A C B D E G
A C D E F G

Expected BLEU

- Brevity penalty when hypothesis < reference

Expected BLEU

- Usually perform 1-best BLEU
  - argmax
- Expected BLEU replaces it argmax with sum
  - Function becomes continuous w.r.t weights

- Use approximate brevity penalty
  - Replace argmax with sum

- Differentiable

Expected BLEU

- Procedure:

1)LBFGS tuning for several iterations until convergence on the hypergraph

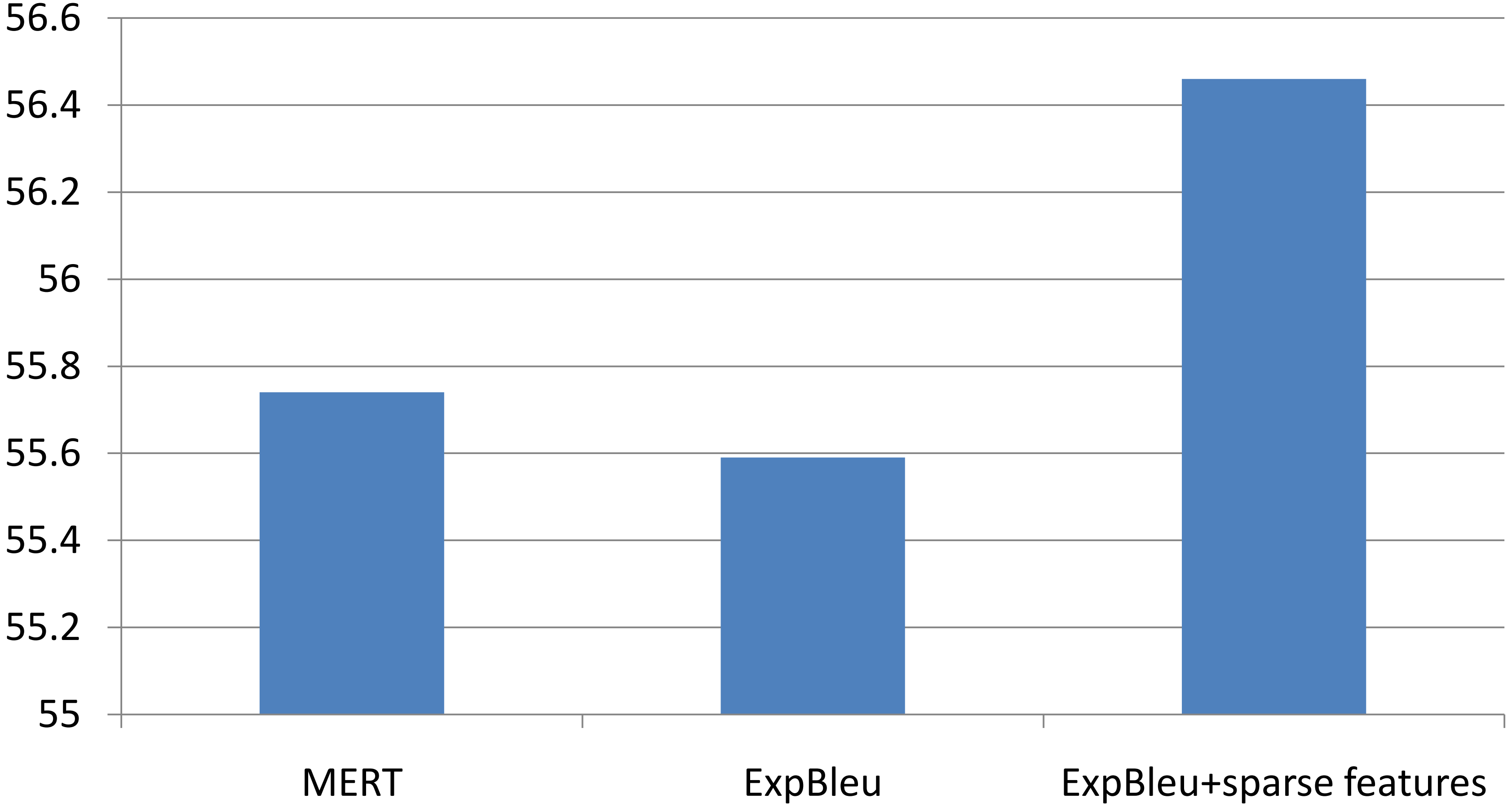
2)Re-decode the source data, generate updated hypergraph

3) Repeat

Preliminary Experiments

- Compare Expected BLEU with MIRA with equivalent grammar on same test set
- Incorporate fine-grained features
  - Source Syntax
  - Target Syntax
  - Source Context
  - Glue Features
  - OOV
  - Backoff Rule

Preliminary Results



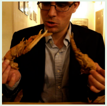
Coming Soon…

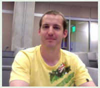
Decoding with Complex Grammars

Outline


Vlad Eidelman


Ziyuan Wang


Adam Lopez


Jon Graehl


ThuyLinh Nguyen

- 3:20pm Parametric models: posterior regularisation. Desai
- 3:35pm Training models with rich features spaces. Vlad
- 3:50pm Decoding with complex grammars. Adam
- 4:20pm Closing remarks. Phil
- 4:25pm Finish.

Efficient Decoding for Synchronous Context-Free Grammars

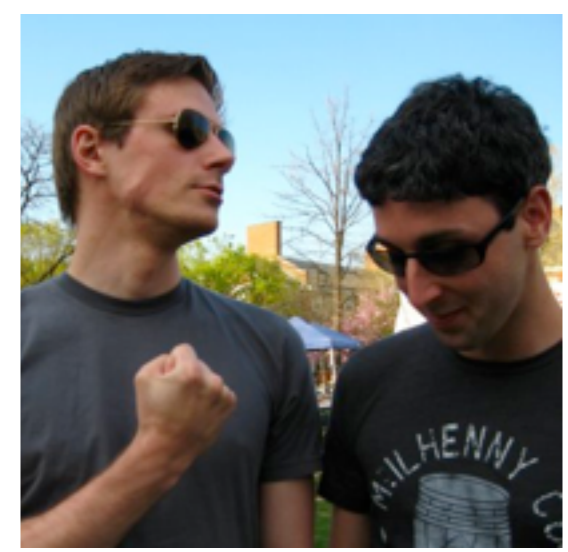
Adam Lopez (Edinburgh)
Jonathan Graehl (ISI)
Chris Dyer (CMU)

with thanks to:
the whole workshop team,
Juri Ganitkevitch (JHU) & Jonny Weese (JHU)

Efficient Decoding for Synchronous Context-Free Grammars

Adam Lopez (Edinburgh)
Jonathan Graehl (ISI)
Chris Dyer (CMU)

with thanks to:
the whole workshop team,
Juri Ganitkevitch (JHU) & Jonny Weese (JHU)



The Story So Far

The Story So Far

- Induce a grammar.

The Story So Far

- Induce a grammar.
- Tune some model parameters.

The Story So Far

- Induce a grammar.
- Tune some model parameters.

- Get a BLEU score.

The Story So Far

- Induce a grammar.
- Tune some model parameters.
- *Decode a test set.*
- Get a BLEU score.

The Story So Far

- Induce a grammar.
- Tune some model parameters.
  - *Decode a tuning set.*
- *Decode a test set.*
- Get a BLEU score.

The Story So Far

- Induce a grammar.
  - ***Decode the training data.***
- Tune some model parameters.
  - ***Decode a tuning set.***
- ***Decode a test set.***
- Get a BLEU score.

The Price of Performance

- 1 Category (baseline): **20.8**
- 25 Categories: **21.7**

The Price of Performance

- 1 Category (baseline): 3.0 sec/sentence
- 25 Categories: 52 sec/sentence

Some Questions

Some Questions

- Why is it so slow?

Some Questions

- Why is it so slow?
- How can we speed it up?

Some Questions

- Why is it so slow?
- How can we speed it up?
- What's the big idea?

Context-Free Grammar

$$X^1 \rightarrow dianzi\ shang$$
$$X^2 \rightarrow dianzi\ shang$$
$$X^3 \rightarrow mao$$
$$X^4 \rightarrow X^1\ de\ X^3$$
$$X^4 \rightarrow X^2\ de\ X^3$$
$$X^5 \rightarrow X^1\ de\ X^3$$
$$S \rightarrow X^4$$
$$S \rightarrow X^5$$

Context-Free Grammar

$X^1 \rightarrow dianzi\ shang$

$X^2 \rightarrow dianzi\ shang$

$X^3 \rightarrow mao$

$X^4 \rightarrow X^1\ de\ X^3$

$X^4 \rightarrow X^2\ de\ X^3$

$X^5 \rightarrow X^1\ de\ X^3$

$S \rightarrow X^4$

$S \rightarrow X^5$

S

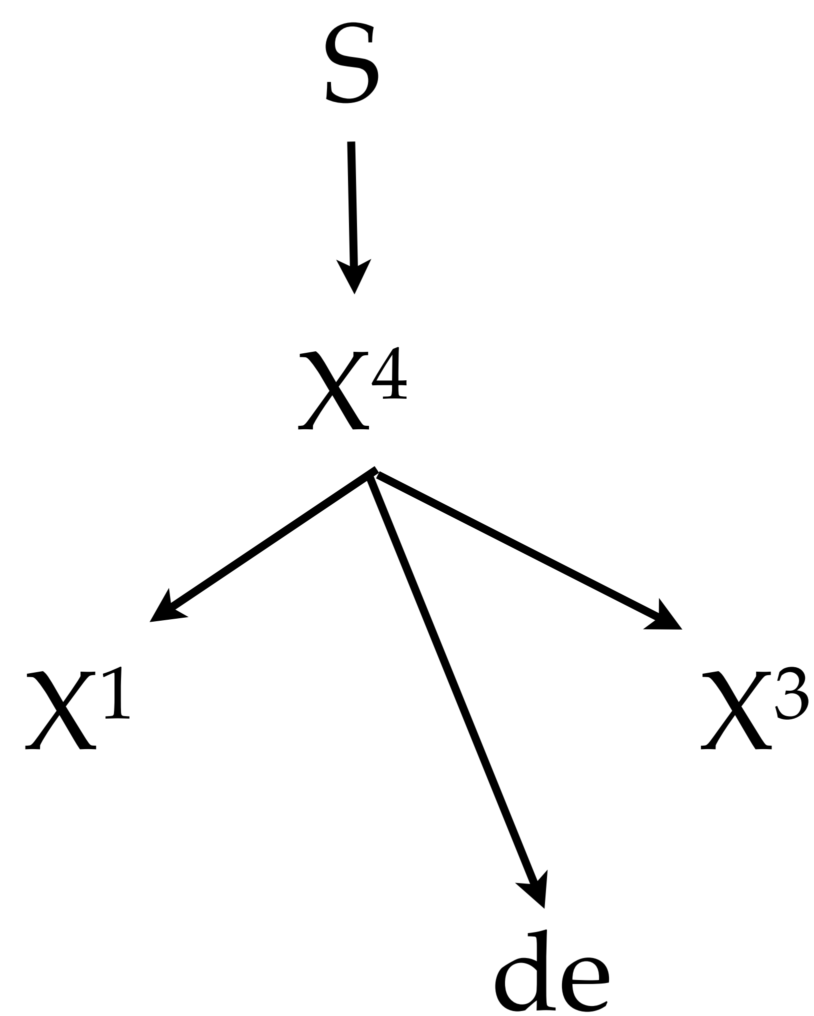
Context-Free Grammar

$$X^1 \rightarrow dianzi\ shang$$
$$X^2 \rightarrow dianzi\ shang$$
$$X^3 \rightarrow mao$$
$$X^4 \rightarrow X^1\ de\ X^3$$
$$X^4 \rightarrow X^2\ de\ X^3$$
$$X^5 \rightarrow X^1\ de\ X^3$$
$$S \rightarrow X^4$$
$$S \rightarrow X^5$$

$$\begin{tikzcd} S \arrow[d] \\ X^4 \end{tikzcd}$$

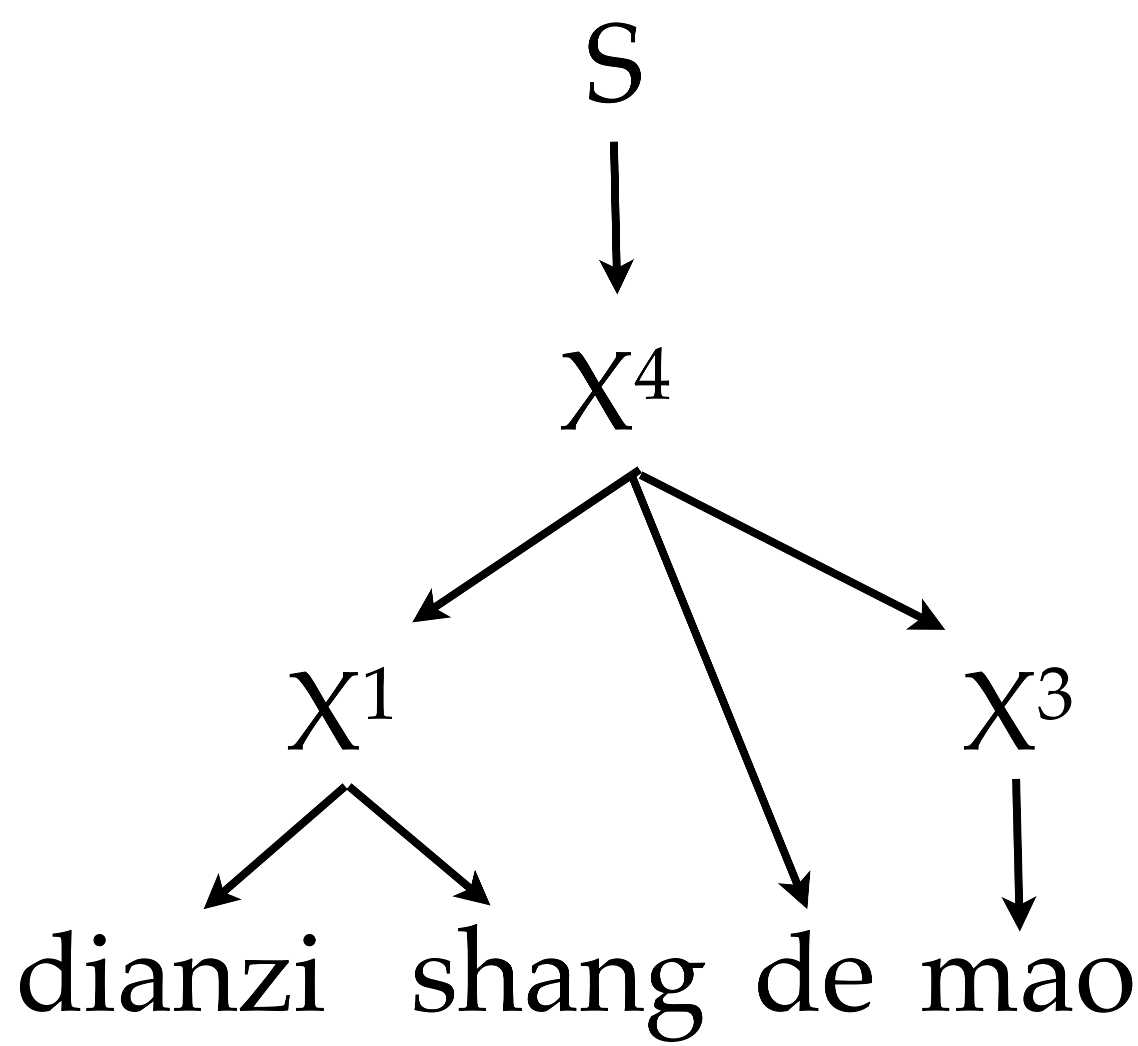
Context-Free Grammar

$$X^1 \rightarrow dianzi\ shang$$
$$X^2 \rightarrow dianzi\ shang$$
$$X^3 \rightarrow mao$$
$$X^4 \rightarrow X^1\ de\ X^3$$
$$X^4 \rightarrow X^2\ de\ X^3$$
$$X^5 \rightarrow X^1\ de\ X^3$$
$$S \rightarrow X^4$$
$$S \rightarrow X^5$$



Context-Free Grammar

$$X^1 \rightarrow dianzi\ shang$$
$$X^2 \rightarrow dianzi\ shang$$
$$X^3 \rightarrow mao$$
$$X^4 \rightarrow X^1\ de\ X^3$$
$$X^4 \rightarrow X^2\ de\ X^3$$
$$X^5 \rightarrow X^1\ de\ X^3$$
$$S \rightarrow X^4$$
$$S \rightarrow X^5$$



Synchronous Context-Free Grammar

$$X^1 \rightarrow dianzi\ shang/the\ mat$$
$$X^2 \rightarrow dianzi\ shang/mat$$
$$X^3 \rightarrow mao/the\ cat$$
$$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$$
$$X^4 \rightarrow X^2\ de\ X^3/X^3\ of\ X^2$$
$$X^5 \rightarrow X^1\ de\ X^3/X^1\ 's\ X^3$$
$$S \rightarrow X^4/X^4$$
$$S \rightarrow X^5/X^5$$

Synchronous Context-Free Grammar

$X^1 \rightarrow dianzi\ shang/the\ mat$

$X^3 \rightarrow mao/the\ cat$

$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$

$S \rightarrow X^5/X^5$

Synchronous Context-Free Grammar

$X^1 \rightarrow dianzi\ shang/the\ mat$
$X^3 \rightarrow mao/the\ cat$
$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$
$S \rightarrow X^5/X^5$

S S

Synchronous Context-Free Grammar

$X^1 \rightarrow dianzi\ shang/the\ mat$

$X^3 \rightarrow mao/the\ cat$

$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$

$S \rightarrow X^5/X^5$

S ······································ S

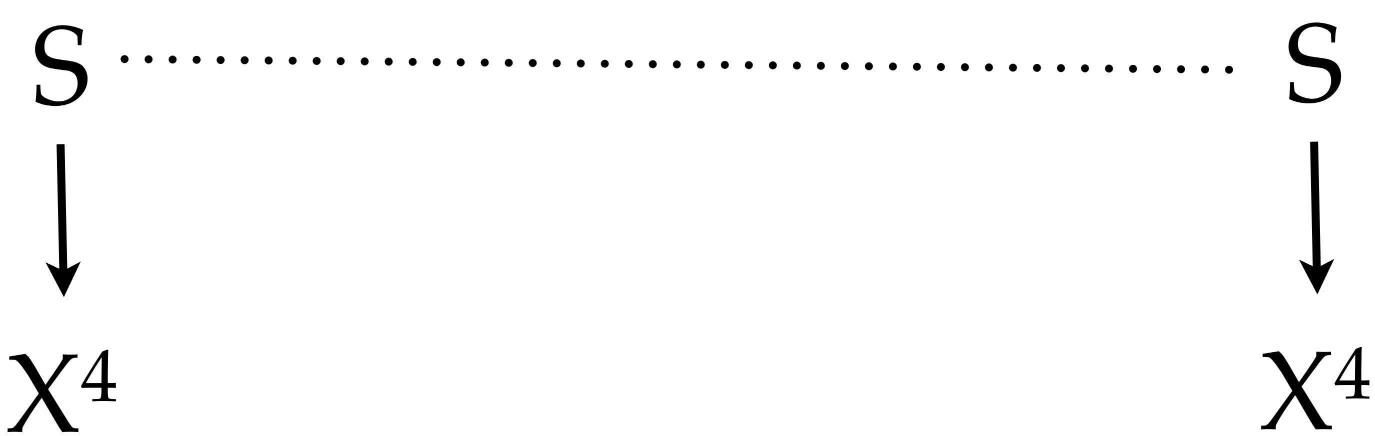
Synchronous Context-Free Grammar

$X^1 \rightarrow dianzi\ shang/the\ mat$

$X^3 \rightarrow mao/the\ cat$

$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$

$S \rightarrow X^5/X^5$



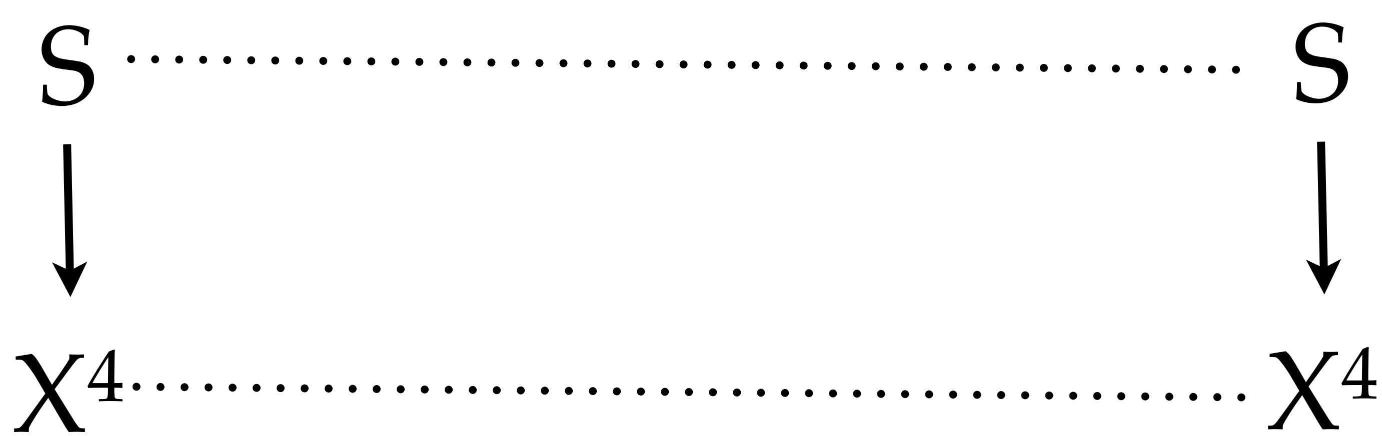
Synchronous Context-Free Grammar

$X^1 \rightarrow \textit{dianzi shang}/\textit{the mat}$

$X^3 \rightarrow \textit{mao}/\textit{the cat}$

$X^4 \rightarrow X^1 \; \textit{de} \; X^3 / X^3 \; \textit{on} \; X^1$

$S \rightarrow X^5 / X^5$



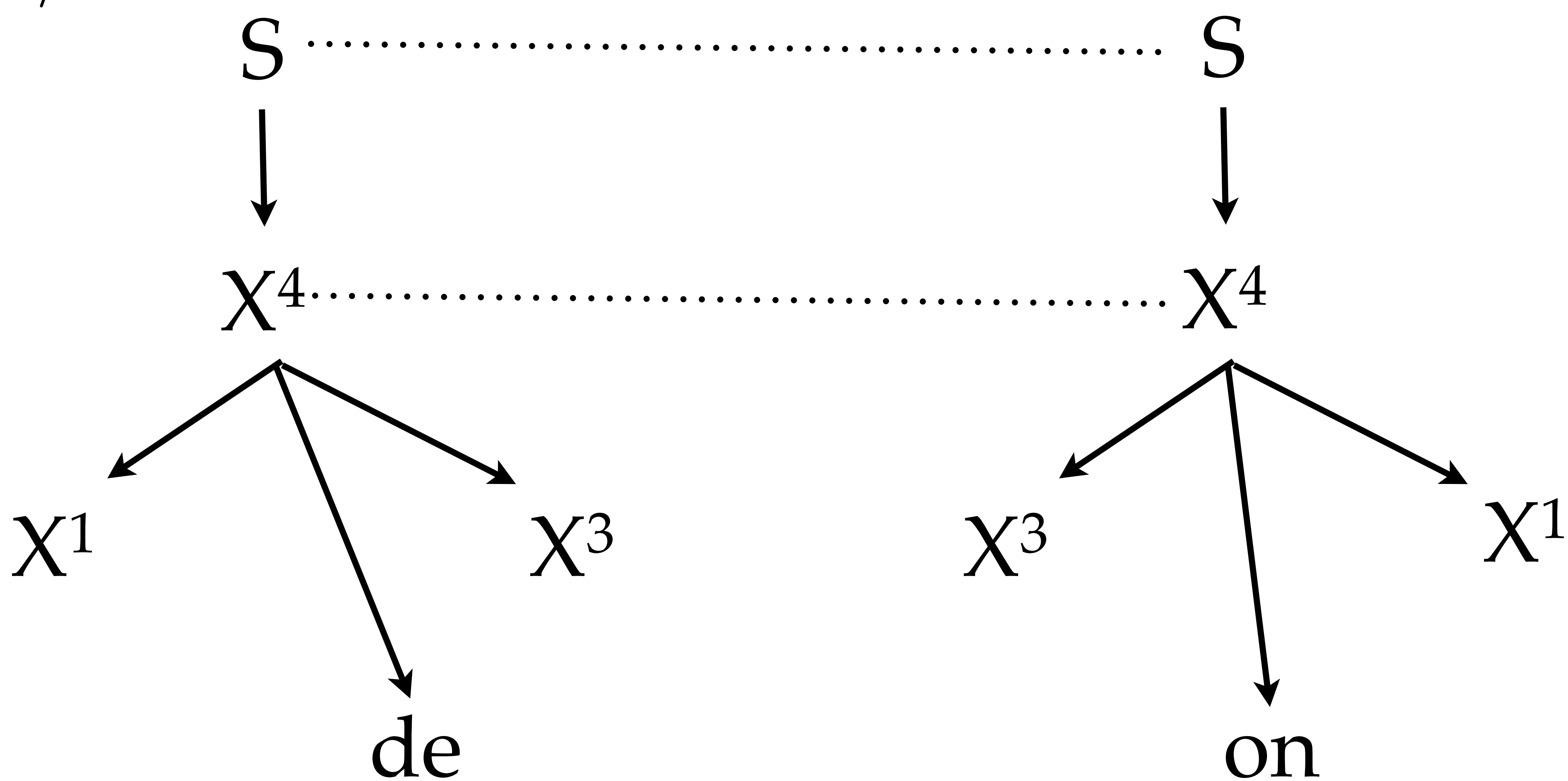
Synchronous Context-Free Grammar

$X^1 \rightarrow dianzi\ shang/the\ mat$

$X^3 \rightarrow mao/the\ cat$

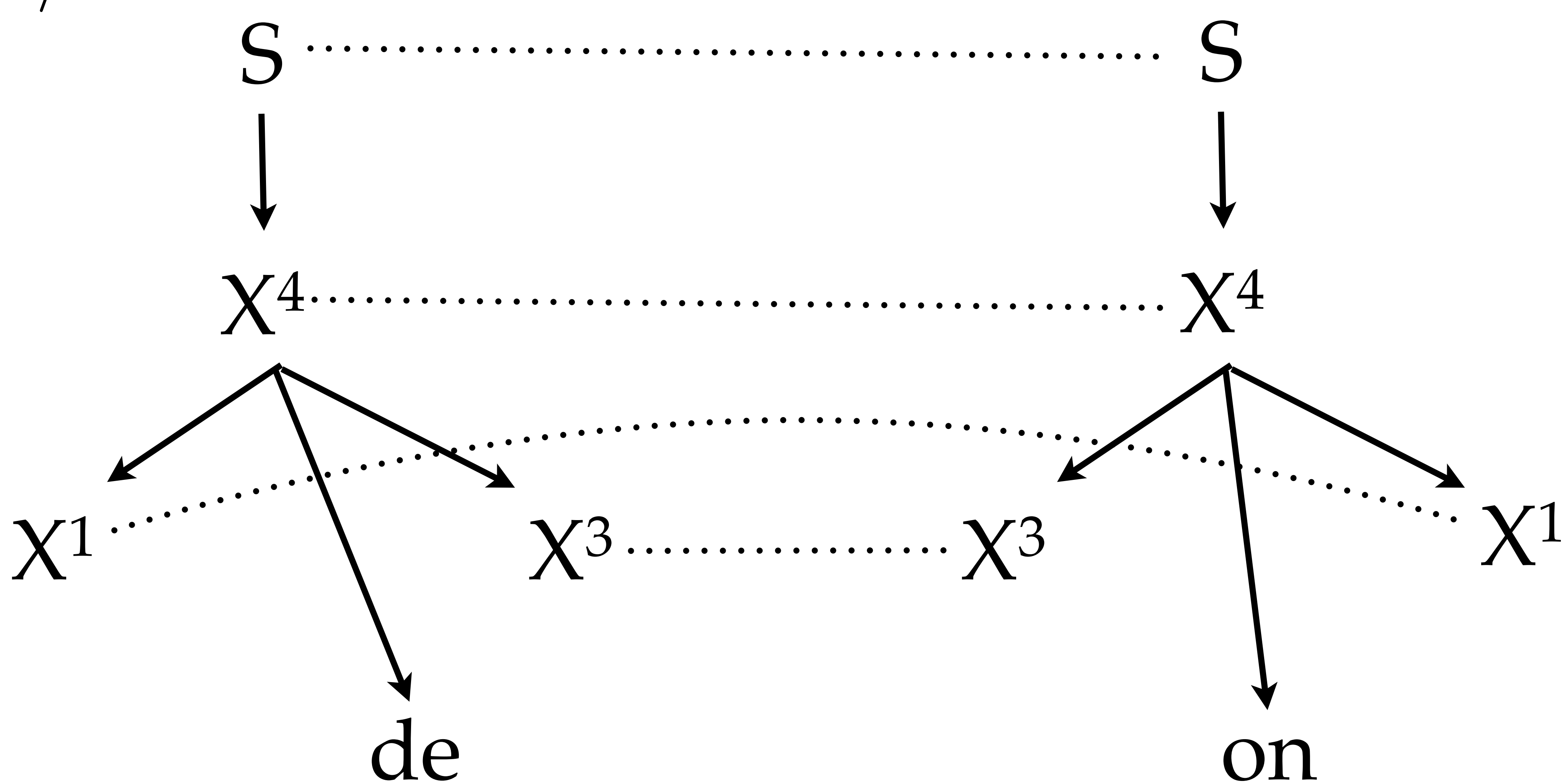
$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$

$S \rightarrow X^5/X^5$



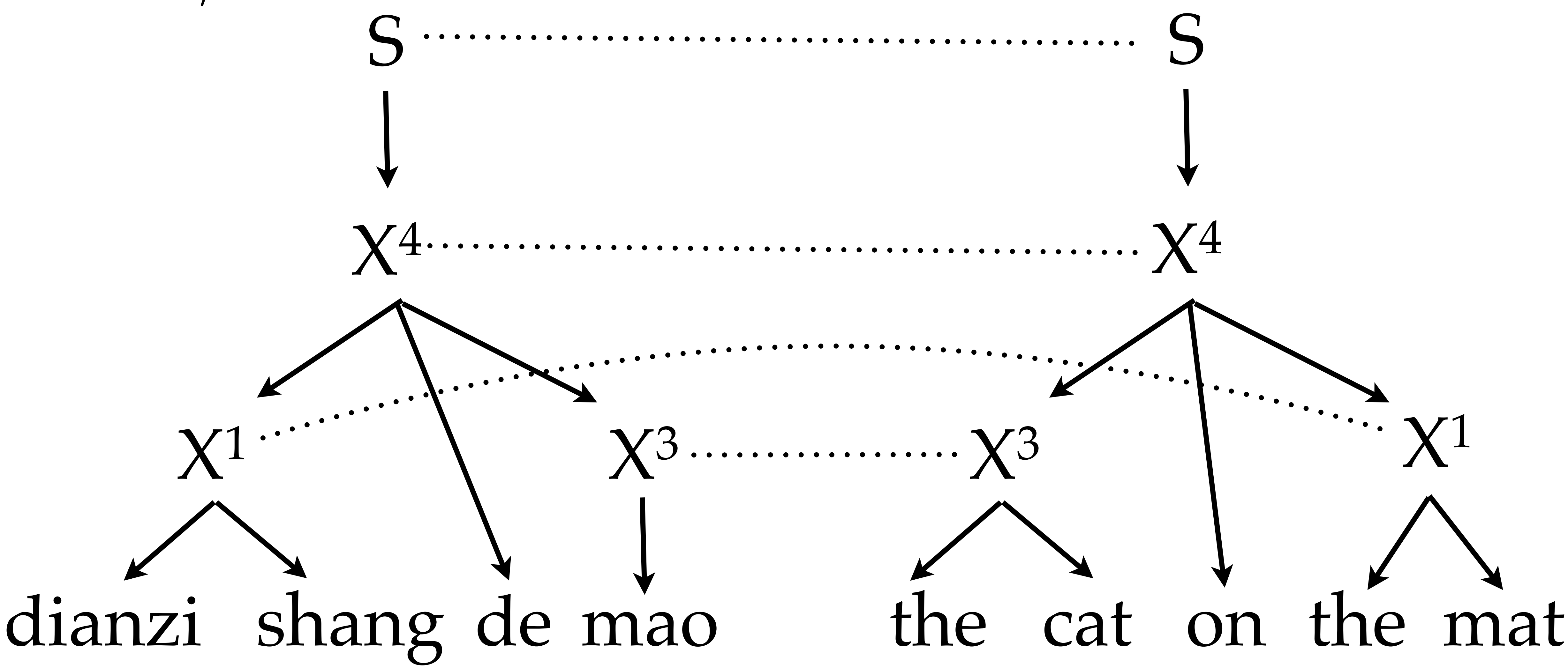
Synchronous Context-Free Grammar

$X^1 \rightarrow dianzi\ shang/the\ mat$
$X^3 \rightarrow mao/the\ cat$
$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$
$S \rightarrow X^5/X^5$



Synchronous Context-Free Grammar

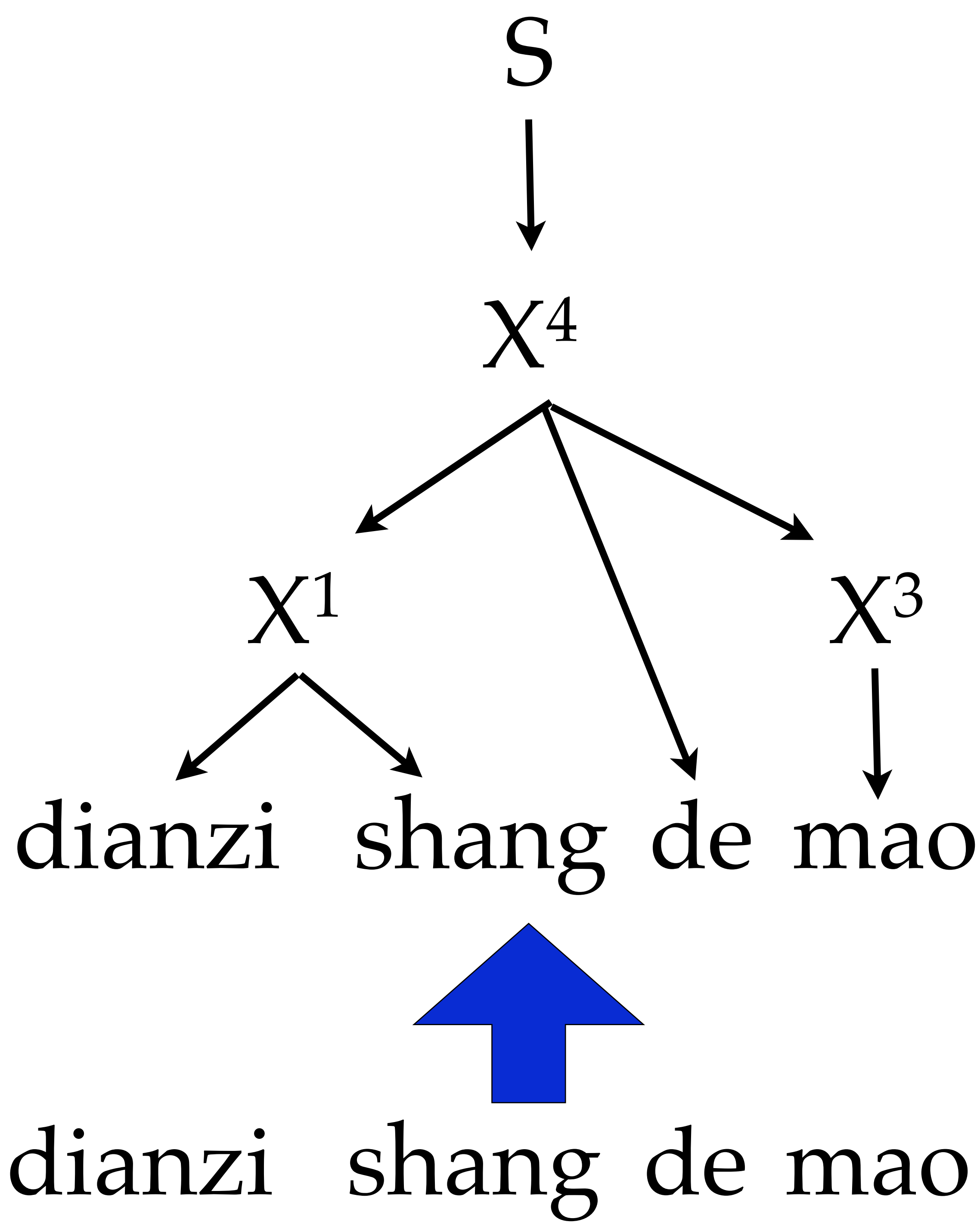
$X^1 \rightarrow dianzi\ shang/the\ mat$
$X^3 \rightarrow mao/the\ cat$
$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$
$S \rightarrow X^5/X^5$



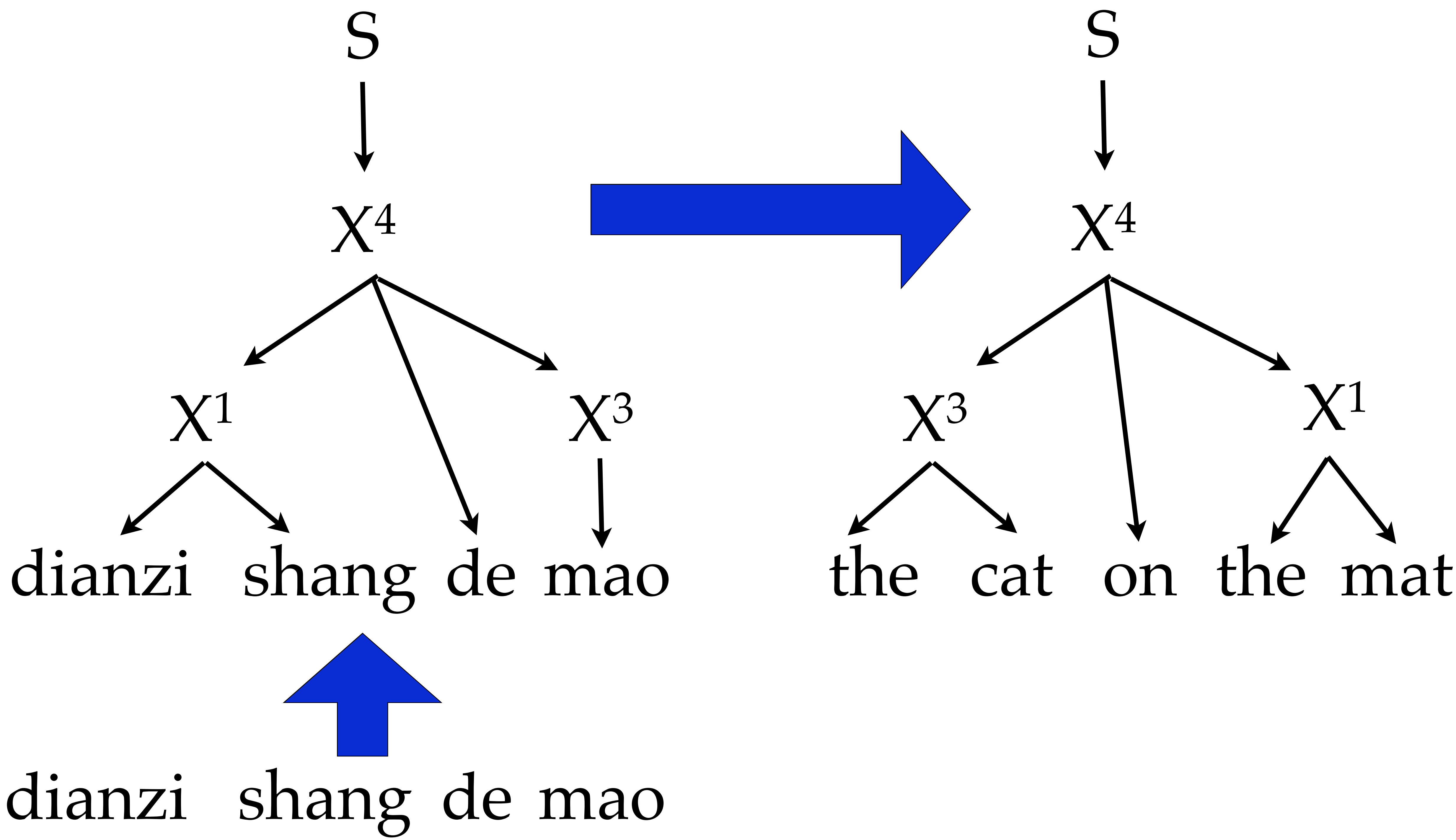
Synchronous Context-Free Grammar

dianzi shang de mao

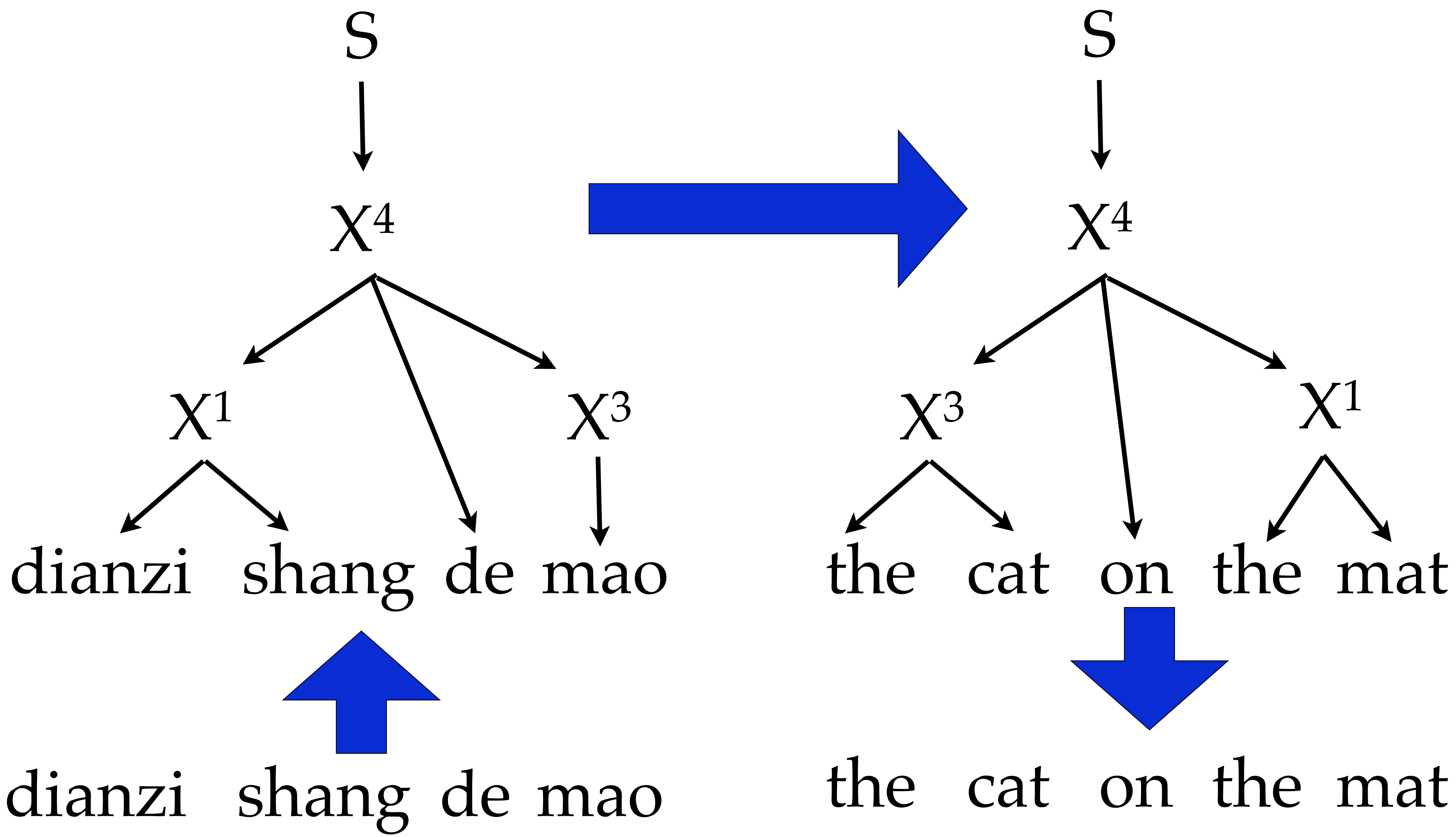
Synchronous Context-Free Grammar



Synchronous Context-Free Grammar



Synchronous Context-Free Grammar



Parsing

$X^1 \rightarrow dianzi\ shang/the\ mat$

$X^2 \rightarrow dianzi\ shang/mat$

$X^3 \rightarrow mao/the\ cat$

$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$

$X^4 \rightarrow X^2\ de\ X^3/X^3\ of\ X^2$

$X^5 \rightarrow X^1\ de\ X^3/X^1\ 's\ X^3$

$S \rightarrow X^4/X^4$

$S \rightarrow X^5/X^5$

dianzi shang de mao

Parsing

$X^1 \rightarrow dianzi\ shang$

$X^2 \rightarrow dianzi\ shang$

$X^3 \rightarrow mao$

$X^4 \rightarrow X^1\ de\ X^3$

$X^4 \rightarrow X^2\ de\ X^3$

$X^5 \rightarrow X^1\ de\ X^3$

$S \rightarrow X^4$

$S \rightarrow X^5$

dianzi shang de mao

Parsing

$X^1 \rightarrow dianzi\ shang$

$X^2 \rightarrow dianzi\ shang$

$X^3 \rightarrow mao$

$X^4 \rightarrow X^1\ de\ X^3$

$X^4 \rightarrow X^2\ de\ X^3$

$X^5 \rightarrow X^1\ de\ X^3$

$S \rightarrow X^4$

$S \rightarrow X^5$

$_0$ dianzi $_1$ shang $_2$ de $_3$ mao $_4$

Parsing

$X^1 \rightarrow dianzi\ shang$

$X^2 \rightarrow dianzi\ shang$

$X^3 \rightarrow mao$

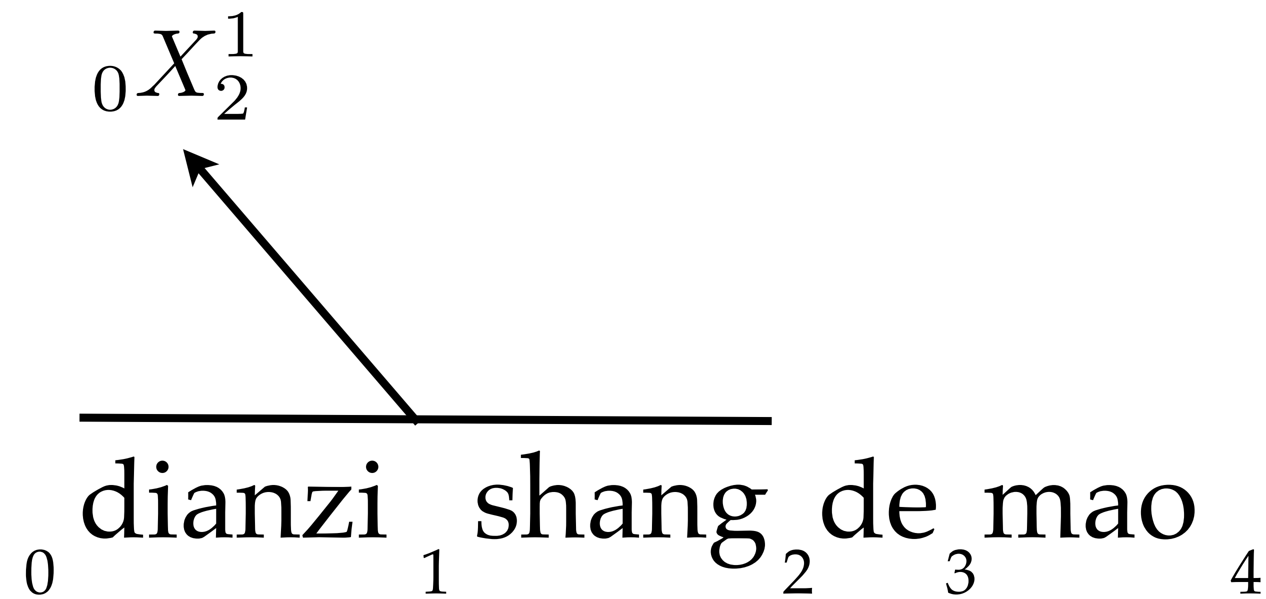
$X^4 \rightarrow X^1\ de\ X^3$

$X^4 \rightarrow X^2\ de\ X^3$

$X^5 \rightarrow X^1\ de\ X^3$

$S \rightarrow X^4$

$S \rightarrow X^5$



Parsing

$X^1 \rightarrow dianzi\ shang$

$X^2 \rightarrow dianzi\ shang$

$X^3 \rightarrow mao$

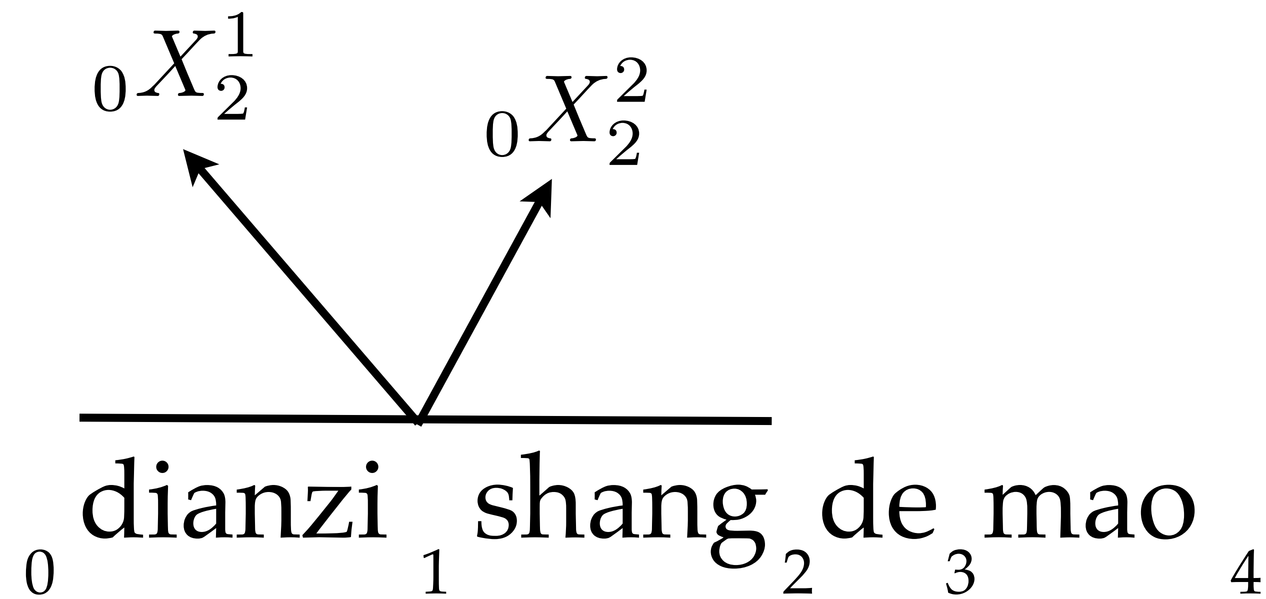
$X^4 \rightarrow X^1\ de\ X^3$

$X^4 \rightarrow X^2\ de\ X^3$

$X^5 \rightarrow X^1\ de\ X^3$

$S \rightarrow X^4$

$S \rightarrow X^5$



Parsing

$X^1 \rightarrow dianzi\ shang$

$X^2 \rightarrow dianzi\ shang$

$X^3 \rightarrow mao$

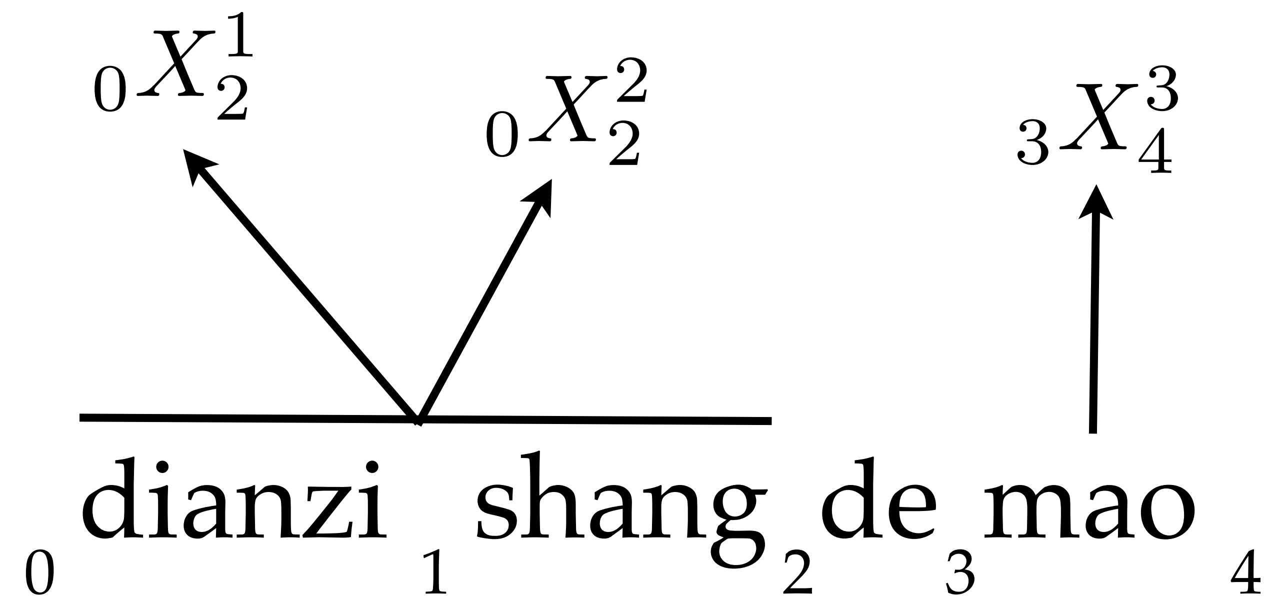
$X^4 \rightarrow X^1\ de\ X^3$

$X^4 \rightarrow X^2\ de\ X^3$

$X^5 \rightarrow X^1\ de\ X^3$

$S \rightarrow X^4$

$S \rightarrow X^5$



Parsing

$X^1 \rightarrow dianzi\ shang$

$X^2 \rightarrow dianzi\ shang$

$X^3 \rightarrow mao$

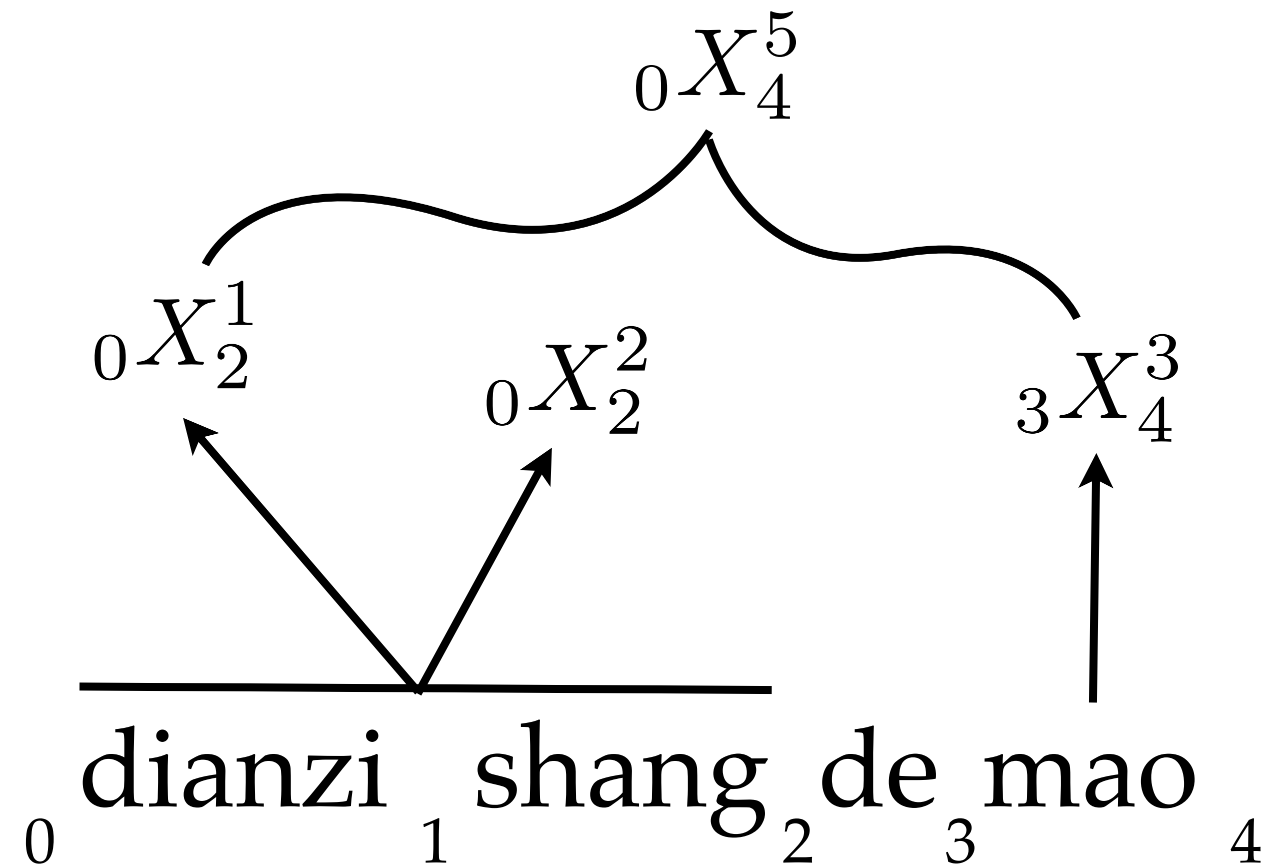
$X^4 \rightarrow X^1\ de\ X^3$

$X^4 \rightarrow X^2\ de\ X^3$

$X^5 \rightarrow X^1\ de\ X^3$

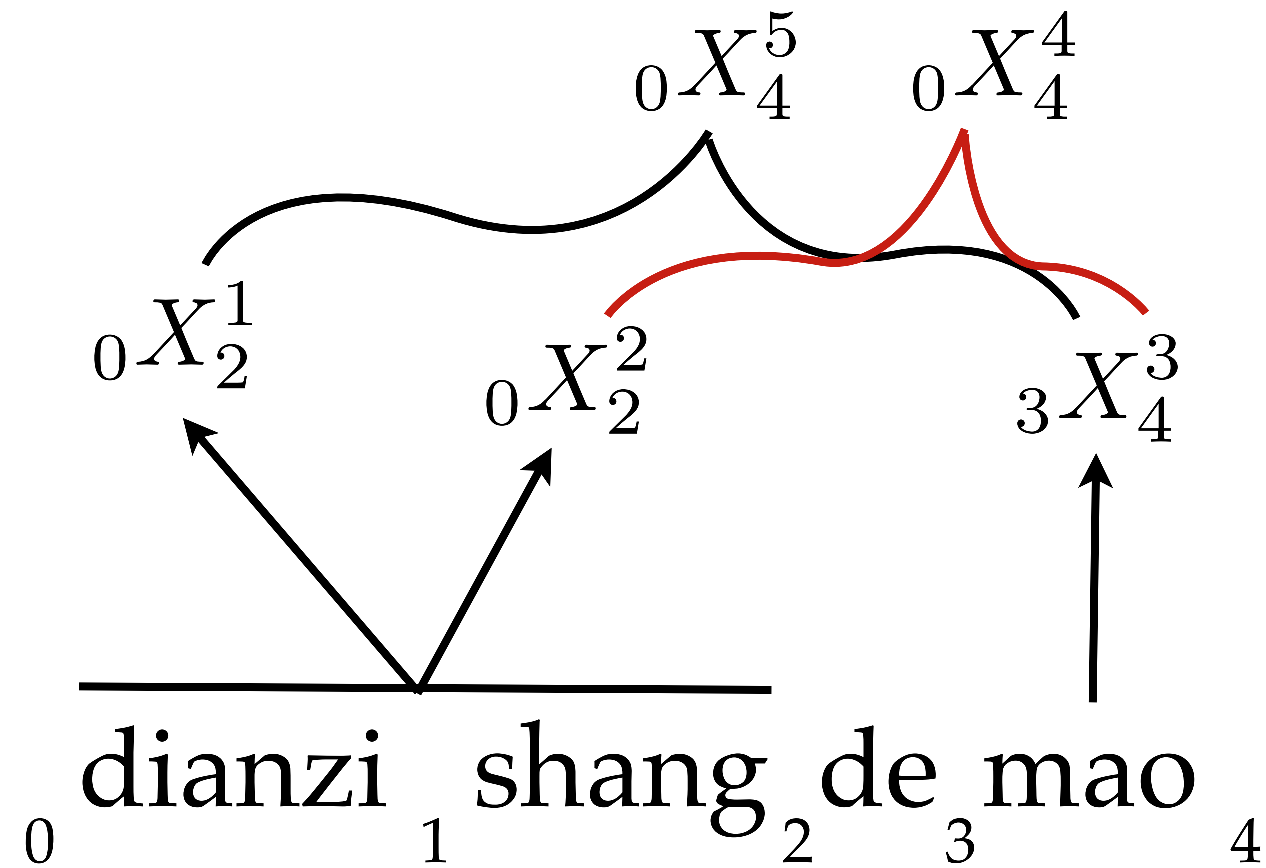
$S \rightarrow X^4$

$S \rightarrow X^5$



Parsing

$$X^1 \rightarrow \textit{dianzi shang}$$
$$X^2 \rightarrow \textit{dianzi shang}$$
$$X^3 \rightarrow \textit{mao}$$
$$X^4 \rightarrow X^1 \; de \; X^3$$
$$X^4 \rightarrow X^2 \; de \; X^3$$
$$X^5 \rightarrow X^1 \; de \; X^3$$
$$S \rightarrow X^4$$
$$S \rightarrow X^5$$



Parsing

$X^1 \rightarrow dianzi\ shang$

$X^2 \rightarrow dianzi\ shang$

$X^3 \rightarrow mao$

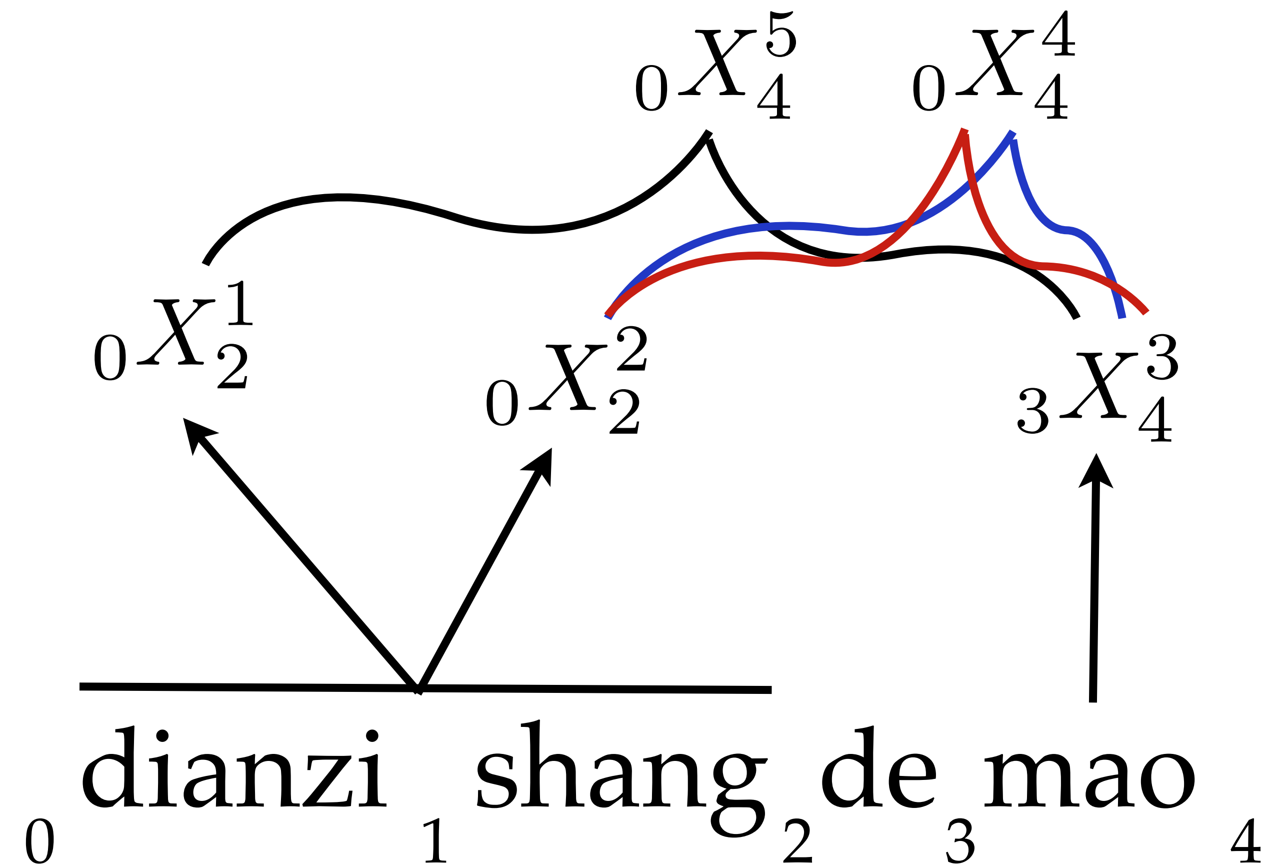
$X^4 \rightarrow X^1\ de\ X^3$

$X^4 \rightarrow X^2\ de\ X^3$

$X^5 \rightarrow X^1\ de\ X^3$

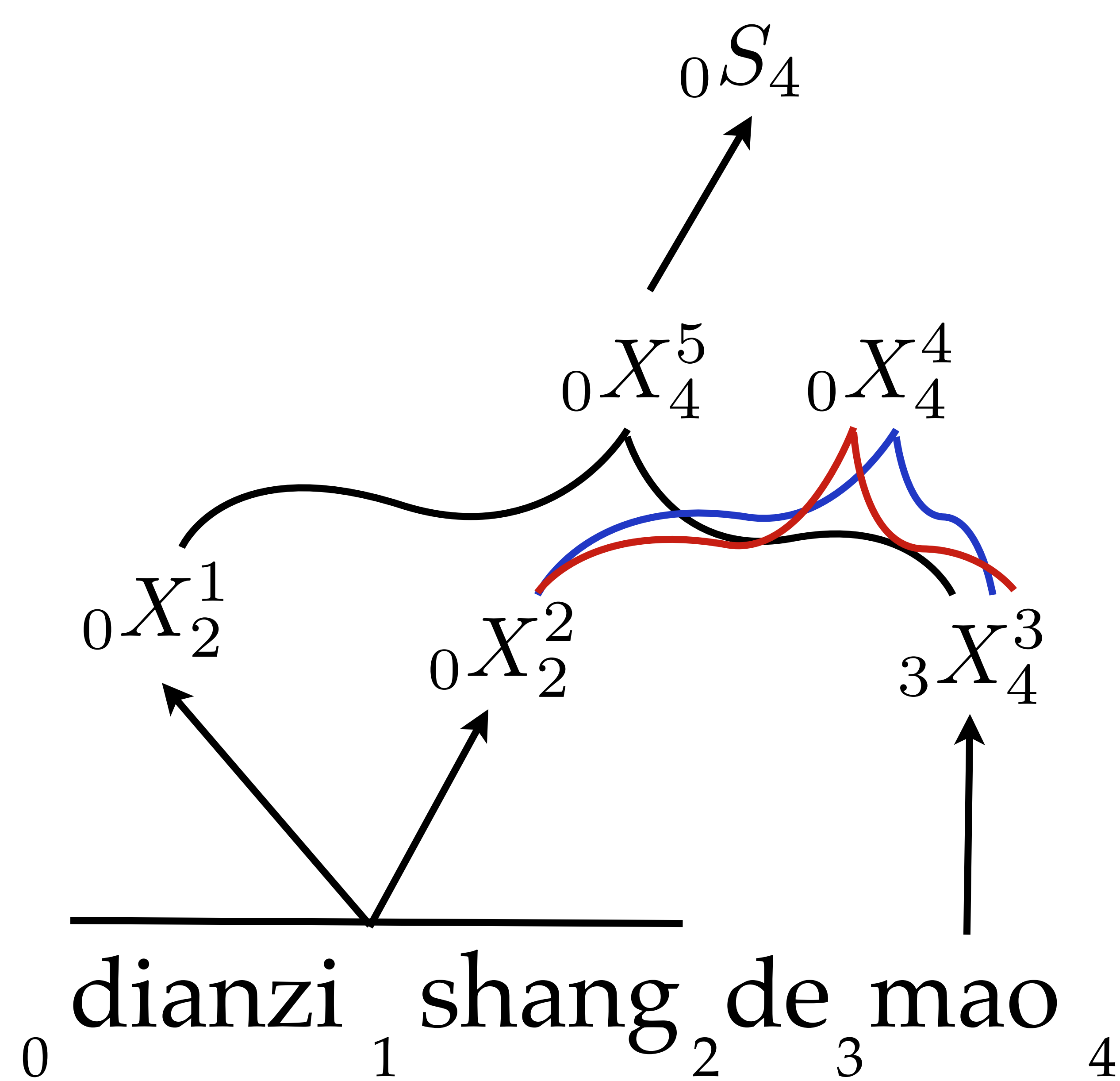
$S \rightarrow X^4$

$S \rightarrow X^5$



Parsing

$$X^1 \rightarrow dianzi\ shang$$
$$X^2 \rightarrow dianzi\ shang$$
$$X^3 \rightarrow mao$$
$$X^4 \rightarrow X^1\ de\ X^3$$
$$X^4 \rightarrow X^2\ de\ X^3$$
$$X^5 \rightarrow X^1\ de\ X^3$$
$$S \rightarrow X^4$$
$$S \rightarrow X^5$$



Parsing

$X^1 \rightarrow \textit{dianzi shang}$

$X^2 \rightarrow \textit{dianzi shang}$

$X^3 \rightarrow \textit{mao}$

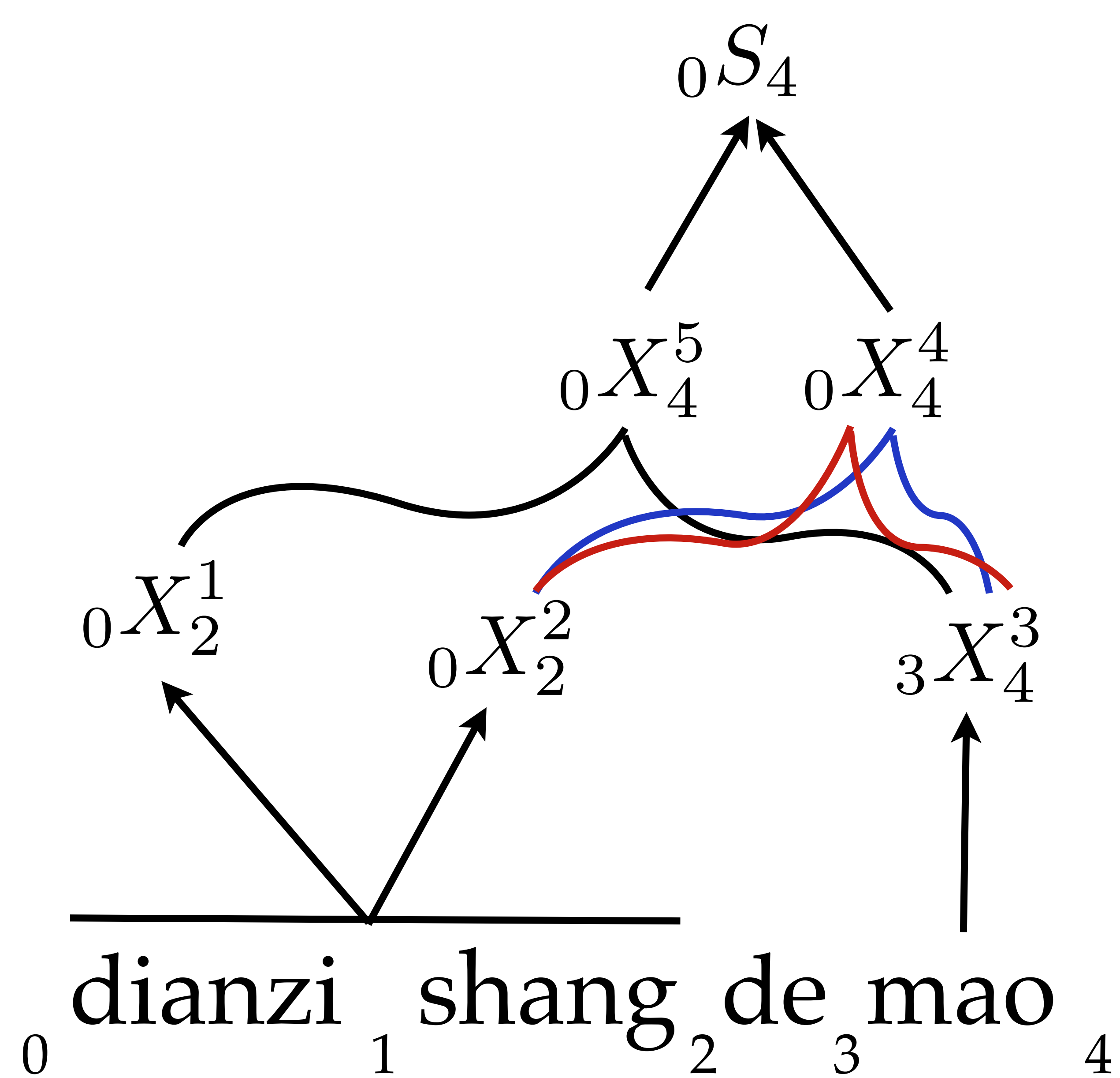
$X^4 \rightarrow X^1 \textit{ de } X^3$

$X^4 \rightarrow X^2 \textit{ de } X^3$

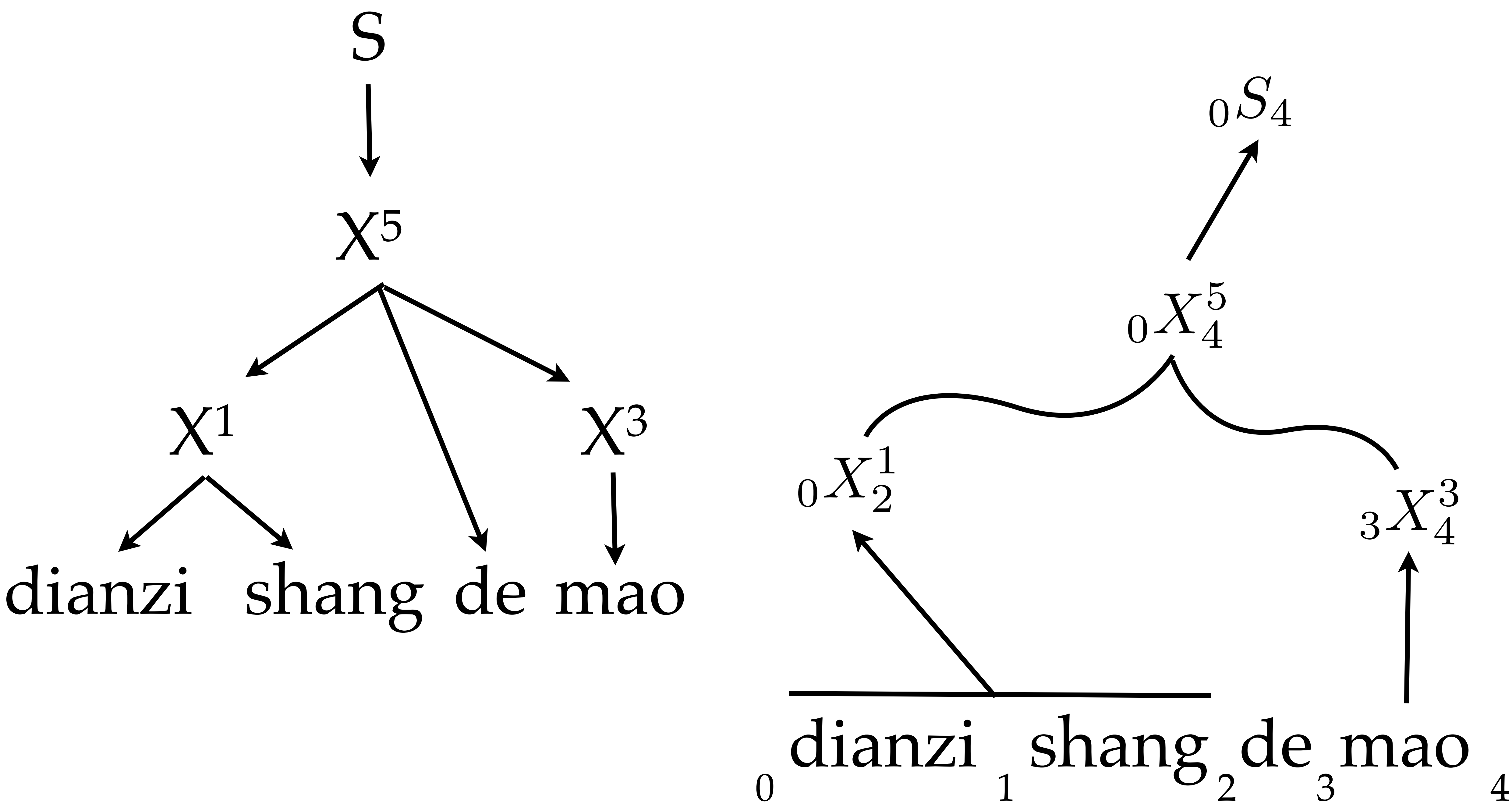
$X^5 \rightarrow X^1 \textit{ de } X^3$

$S \rightarrow X^4$

$S \rightarrow X^5$

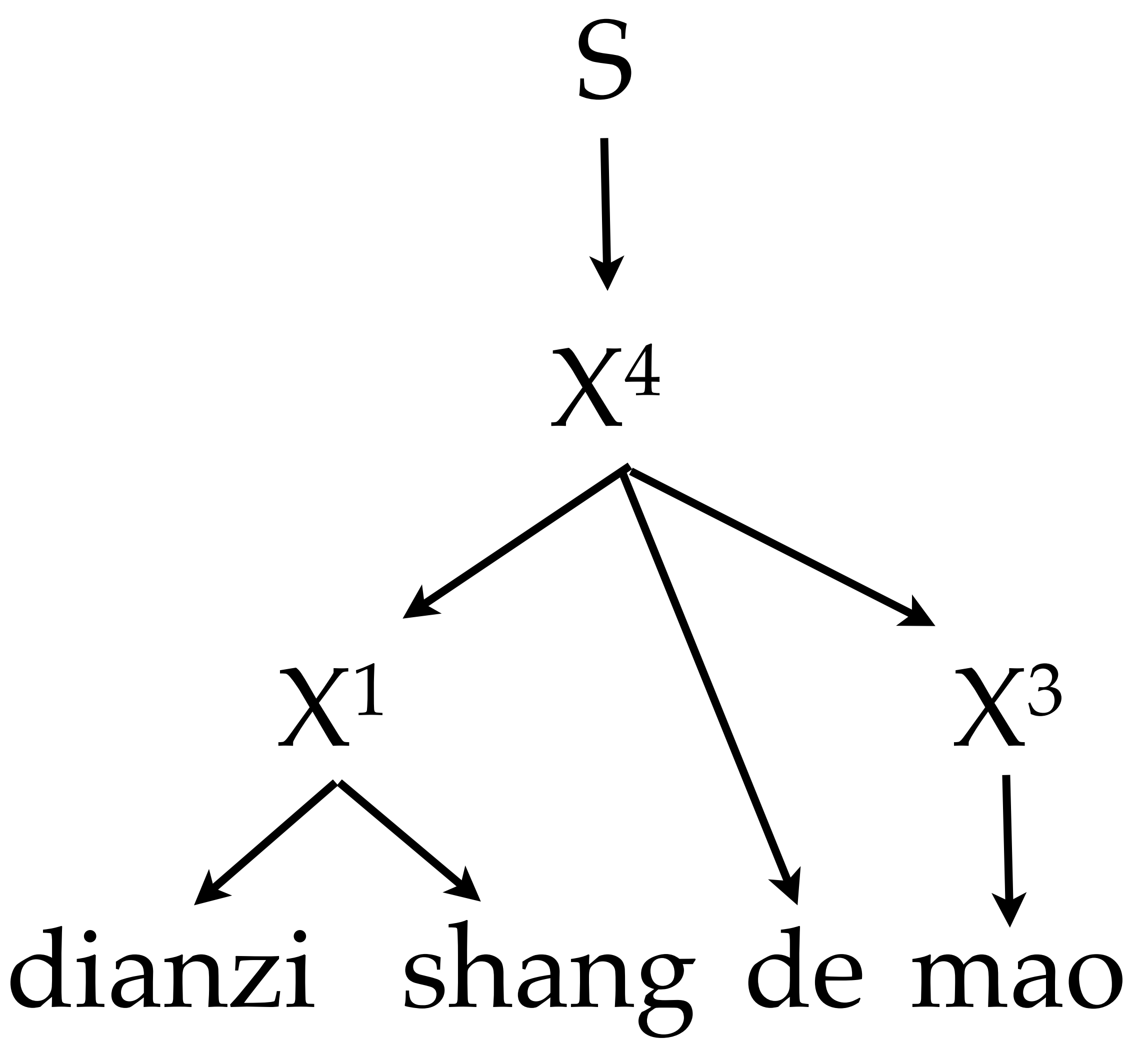


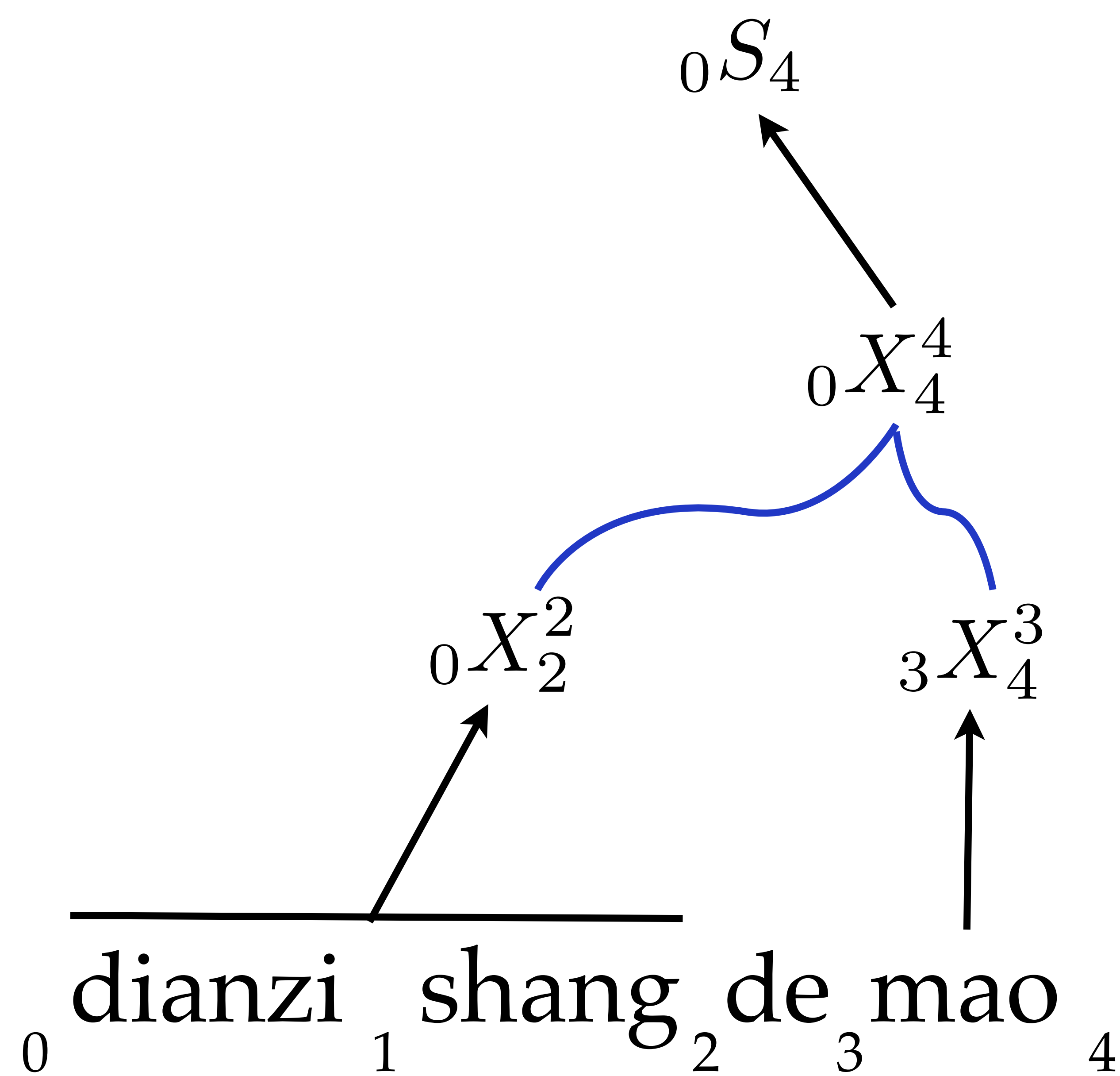
Parsing



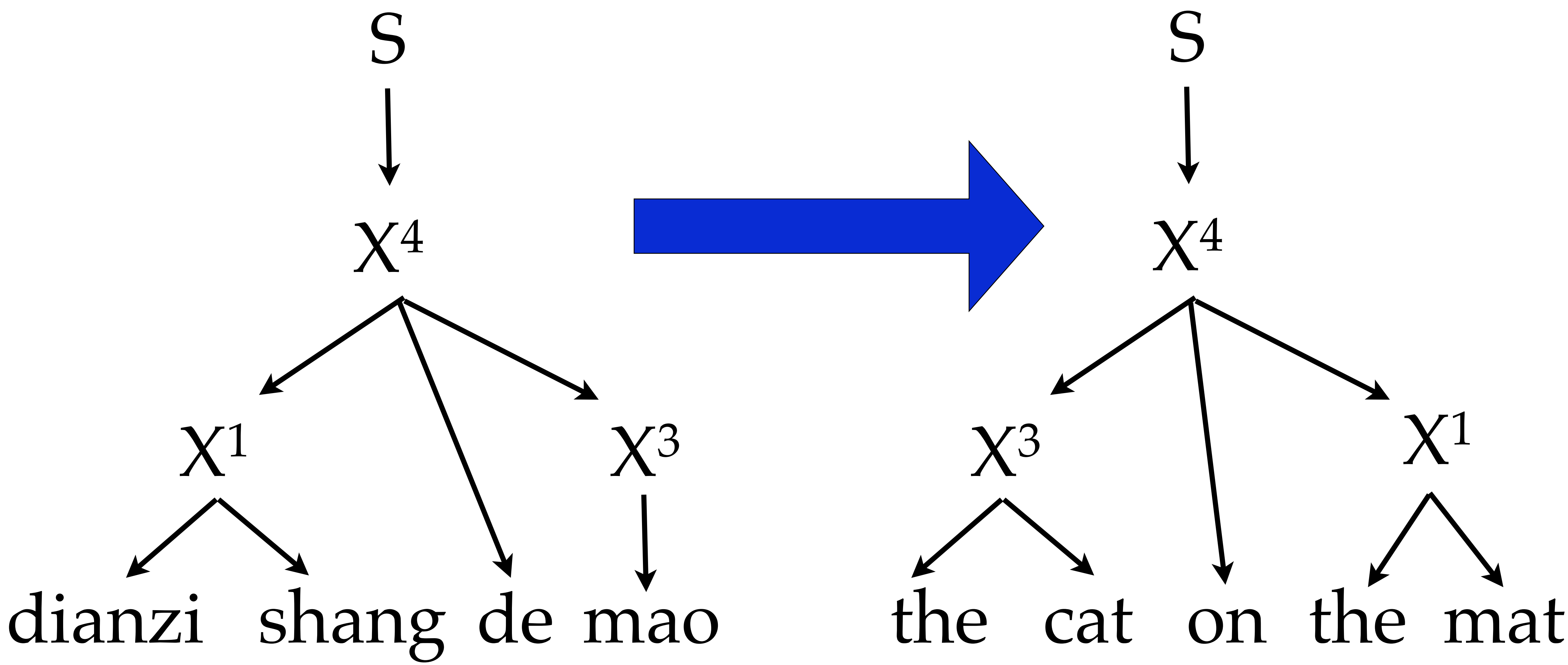
S
X5
X1
X3
dianzi shang de mao
$_0S_4$
$_0X_4^5$
$_0X_2^1$
$_3X_4^3$
0 dianzi 1 shang 2 de 3 mao 4

Parsing


S
X4
X1
X3
dianzi
shang
de
mao


$_0S_4$
$_0X_4^4$
$_0X_2^2$
$_3X_4^3$
$_0$ dianzi $_1$ shang $_2$ de $_3$ mao $_4$

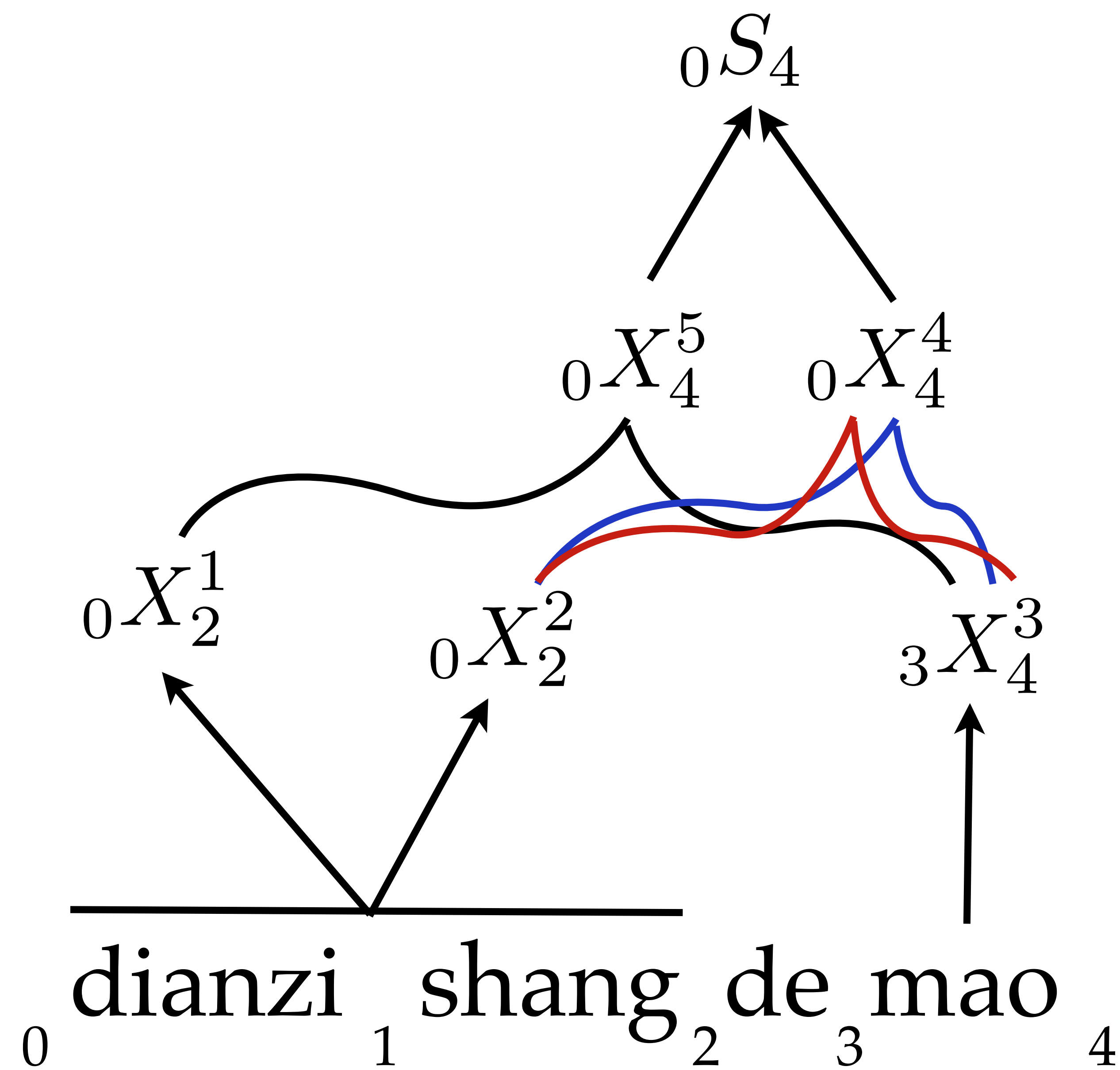
Parsing



Analysis

Nodes: O(Gn²)

Edges: O(G³n³)



Not so fast...

- Speed and memory footprint matter for both evaluation and tuning.
- What if *G* is really big?
- What happens when we add an *n*-gram language model?

Not so fast...

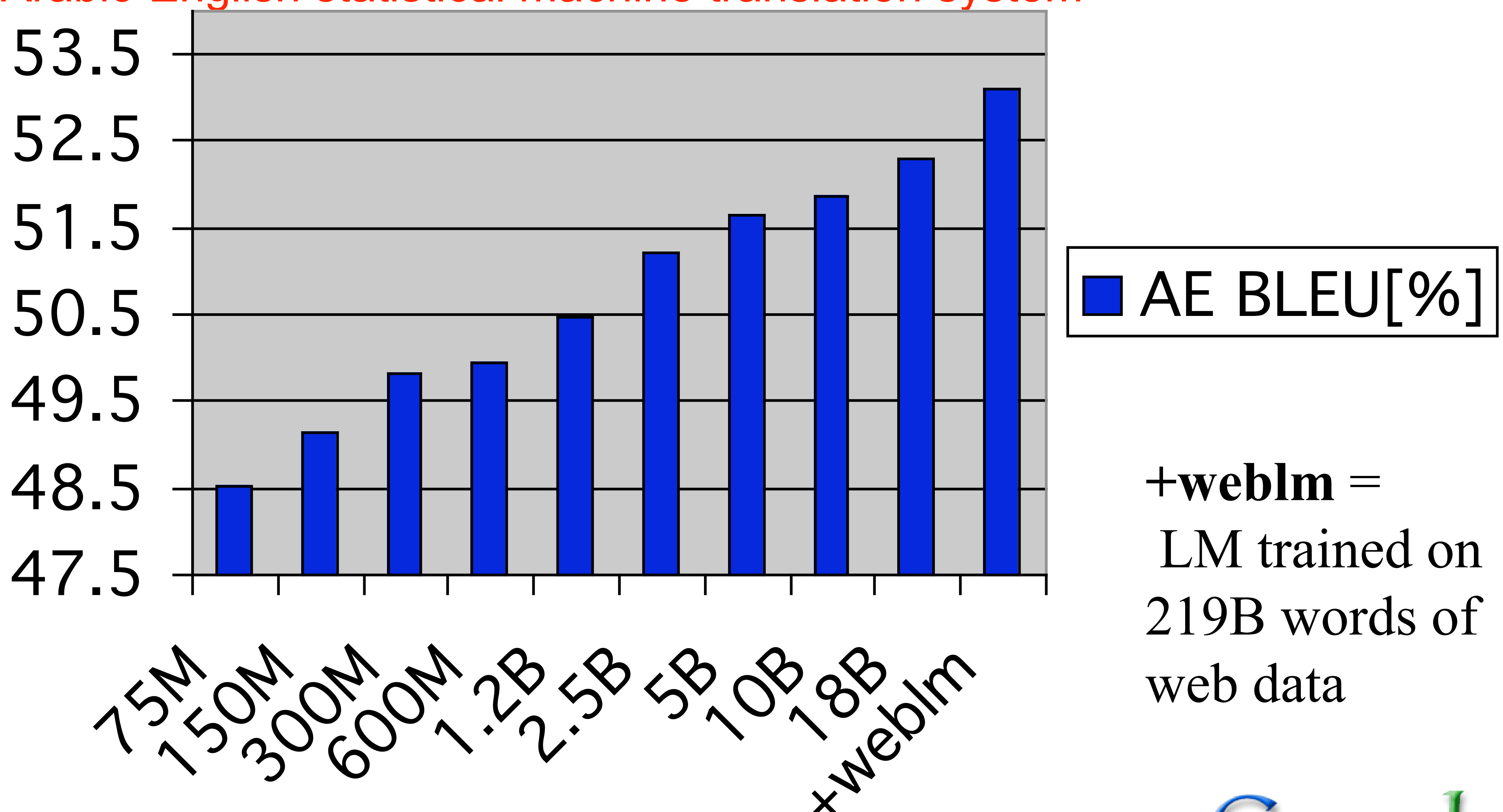
- Speed and memory footprint matter for both evaluation and tuning.
- What if *G* is really big?
- What happens when we add an *n*-gram language model?

$$argmax_{English} p(Urdu|English)p(English)$$

Language Models are Important

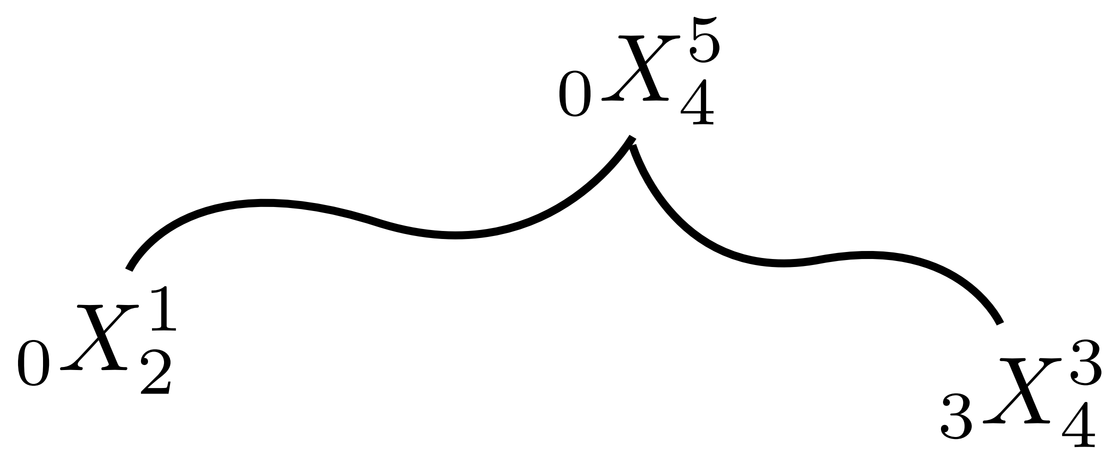
Language Models are Important

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system



+weblm = LM trained on 219B words of web data

+LM Dynamic Programming



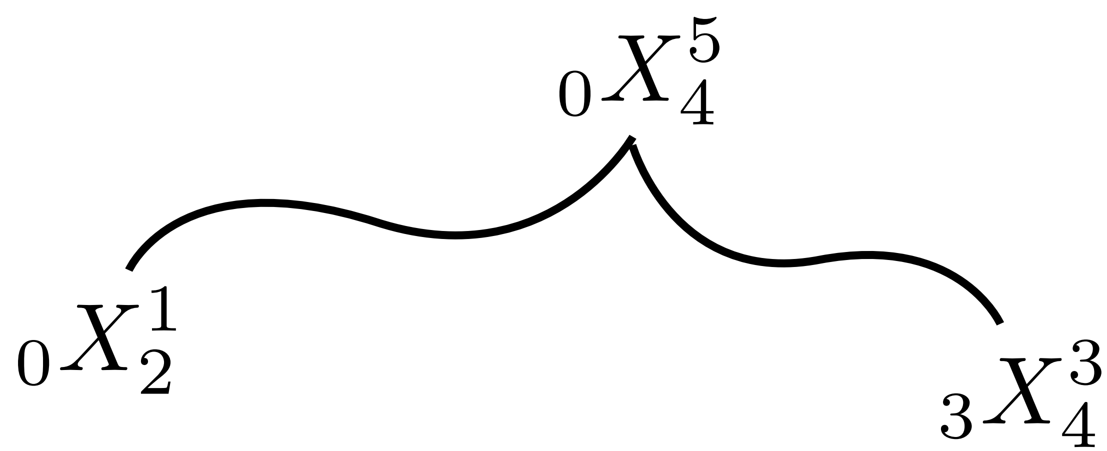
+LM Dynamic Programming

$${}_0X_4^5$$

$${}_0X_2^1 \qquad\qquad\qquad {}_3X_4^3$$

the ... mat
a ... mat
mat ... mat

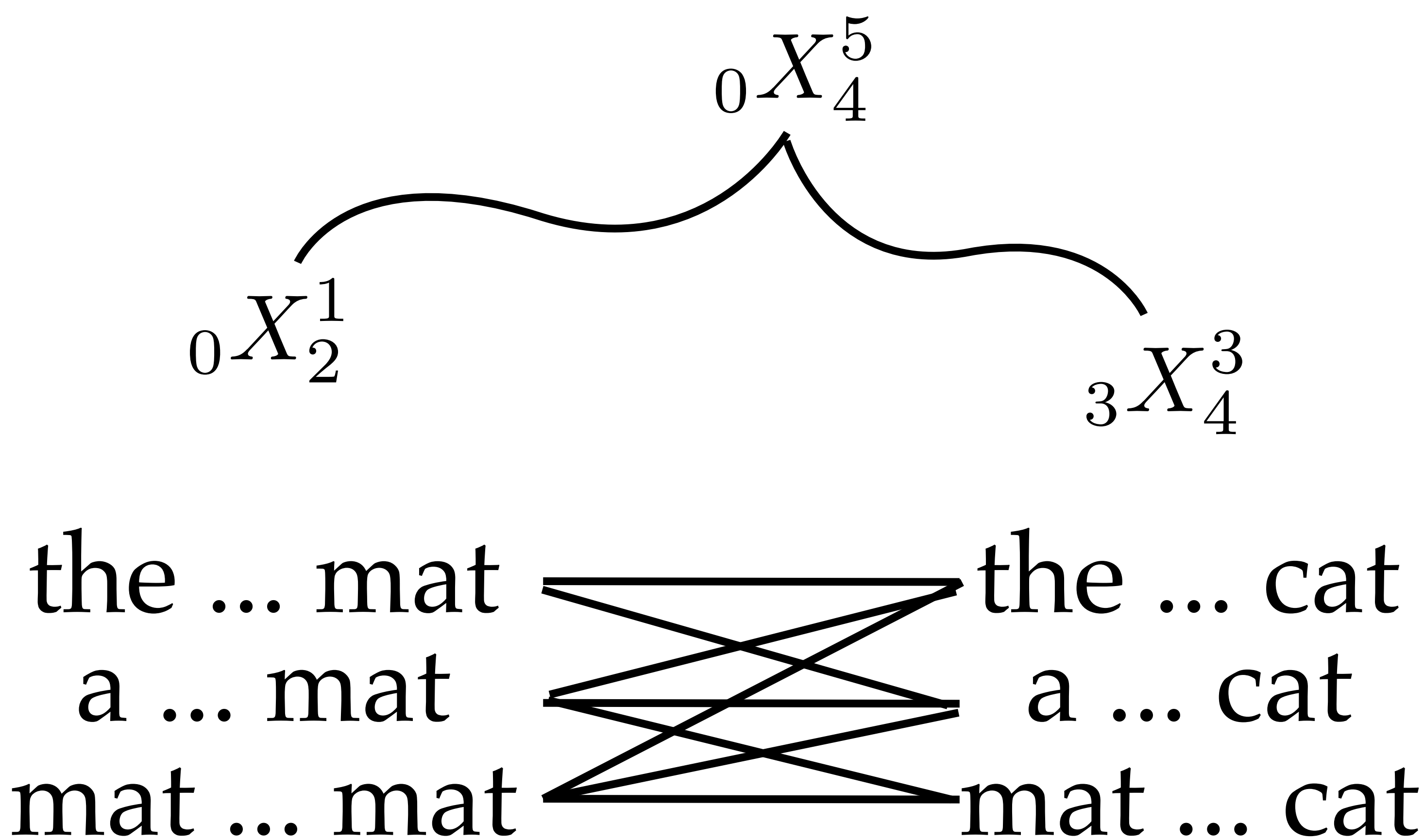
+LM Dynamic Programming



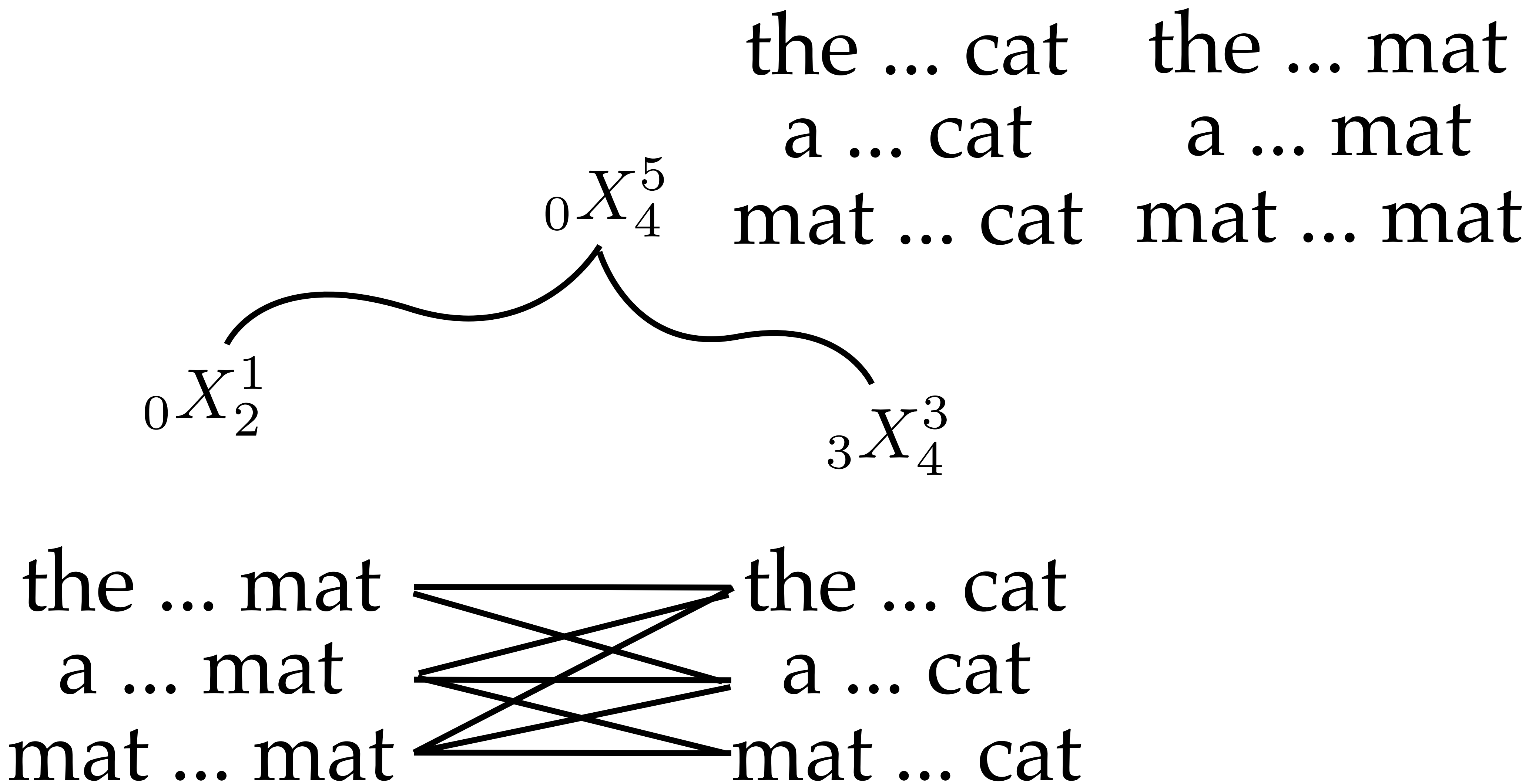
the ... mat
a ... mat
mat ... mat

the ... cat
a ... cat
mat ... cat

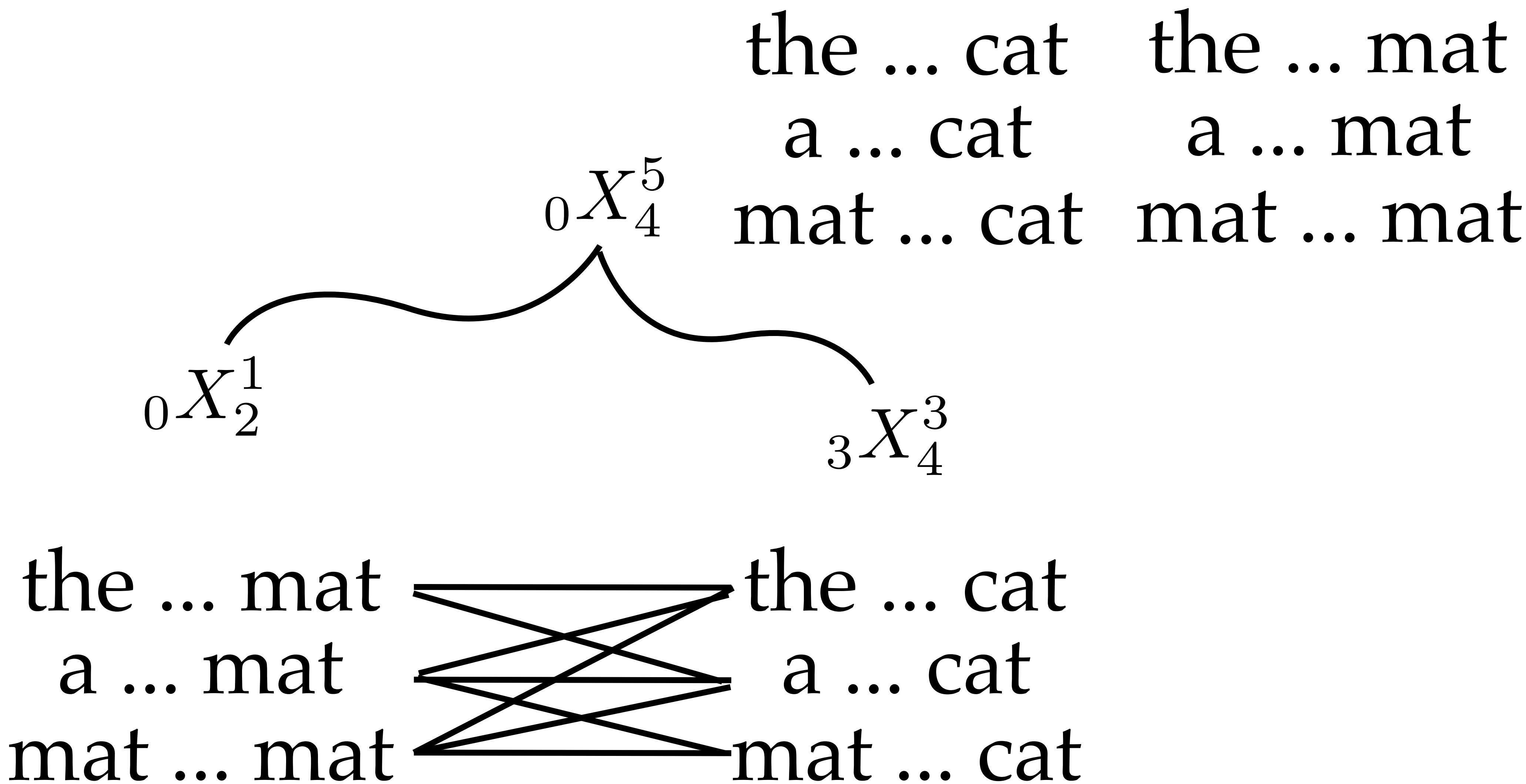
+LM Dynamic Programming



+LM Dynamic Programming



+LM Dynamic Programming



Item (node) form: $X_{i,j,q,r}$

Cube Pruning Summary

- Parse Source
  - Result: -LM Hypergraph
- Incorporate n-grams bottom up, pruning +LM items along the way
  - Result: +LM Hypergraph

Experimental Sandbox

- Urdu 25 category grammar
- cdec decoder (Dyer et al., ACL 2010)
  - http://www.cdec-decoder.org
  - http://code.google.com/p/ws10smt
  - http://github.com/alopez/cdec

cdec

- Why it's awesome:
  - Supports multiple models: linear chain CRF, SCFG, phrase-based
  - Generic hypergraph algorithms
  - Implements baseline: cube pruning

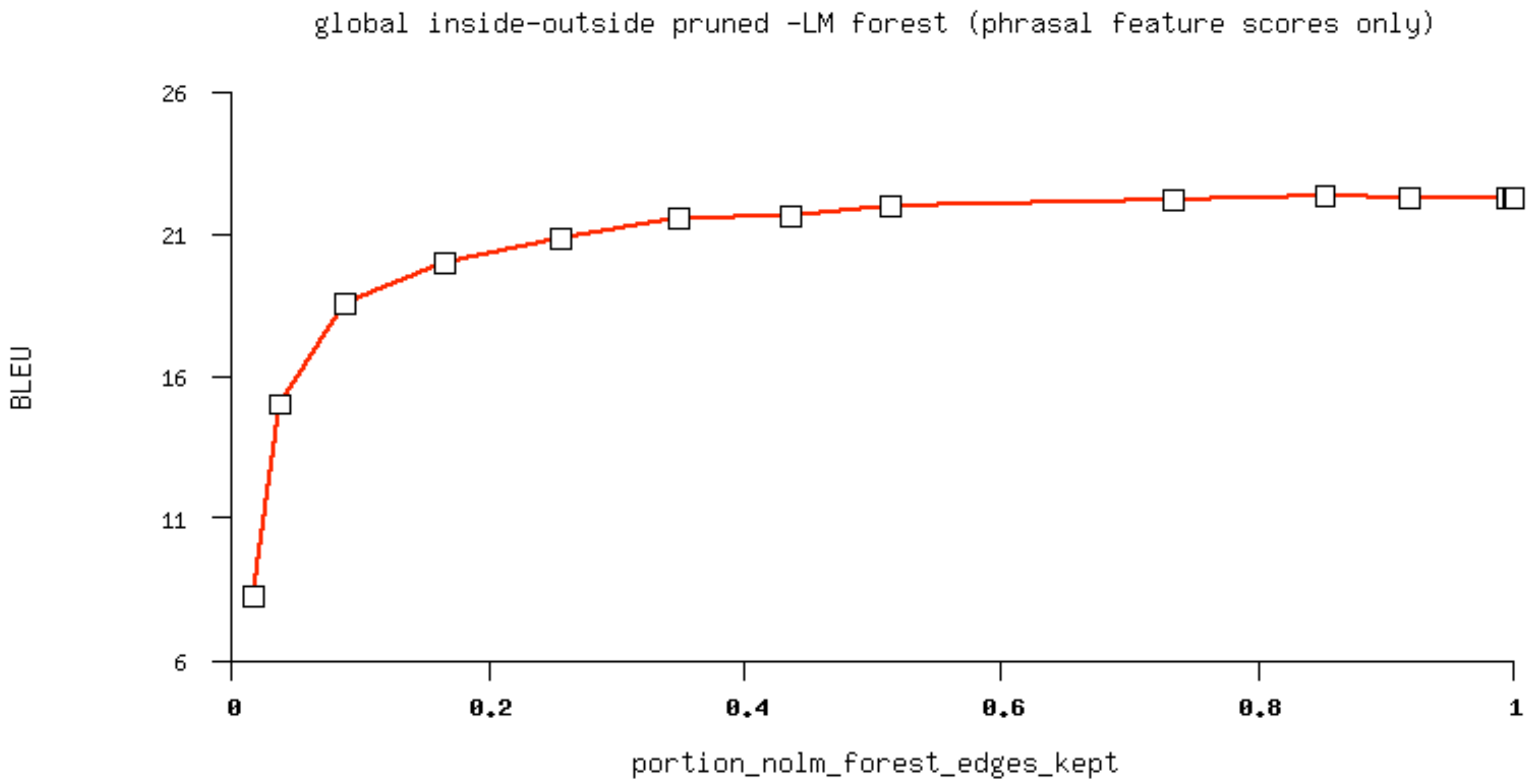
The Size of Source Forests

- 1 Category (baseline)
  - Edges per sentence: 188,954
  - Decode time per sentence: 3.0 seconds
- 25 Categories
  - Edges per sentence: 1,242,410
  - Decode time per sentence: 52 seconds

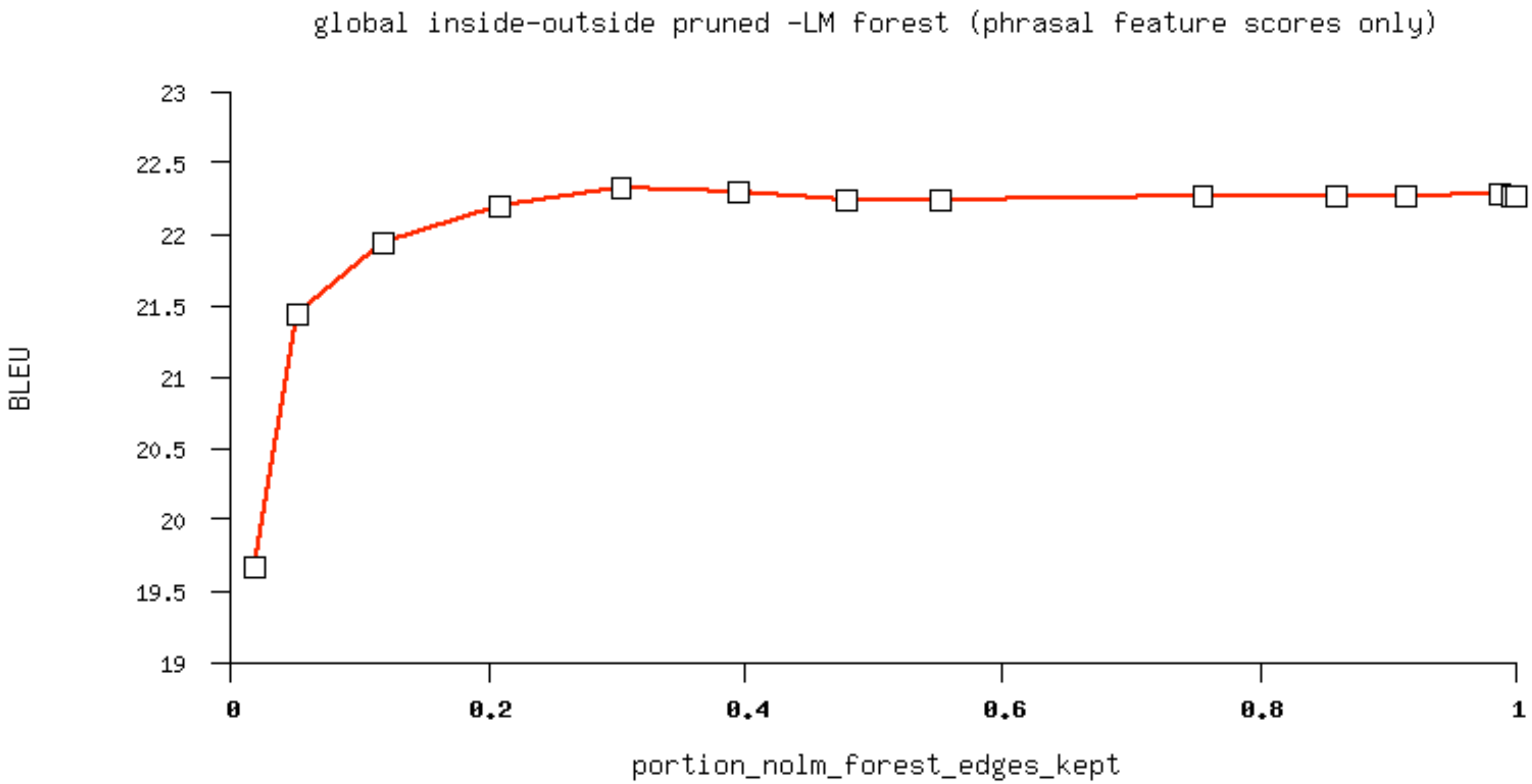
Oracle: Does -LM Pruning Help?

- Generate *unpruned* -LM graph
- Prune using inside-outside
- Cube pruning to obtain +LM graph

Oracle: Source Forest Pruning



Oracle: *Tuned* Source Forest Pruning



Coarse-to-fine Parsing

$$X^1 \rightarrow dianzi\ shang/the\ mat$$
$$X^2 \rightarrow dianzi\ shang/mat$$
$$X^3 \rightarrow mao/the\ cat$$
$$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$$
$$X^4 \rightarrow X^2\ de\ X^3/X^3\ of\ X^2$$
$$X^5 \rightarrow X^1\ de\ X^3/X^1\ 's\ X^3$$
$$S \rightarrow X^4/X^4$$
$$S \rightarrow X^5/X^5$$

Coarse-to-fine Parsing

$X^1 \rightarrow dianzi\ shang/the\ mat$

$X^2 \rightarrow dianzi\ shang/mat$

$X^3 \rightarrow mao/the\ cat$

$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$

$X^4 \rightarrow X^2\ de\ X^3/X^3\ of\ X^2$

$X^5 \rightarrow X^1\ de\ X^3/X^1\ 's\ X^3$

$S \rightarrow X^4/X^4$

$S \rightarrow X^5/X^5$

$X \rightarrow dianzi\ shang/the\ ma$

$X \rightarrow dianzi\ shang/mat$

$X \rightarrow mao/the\ cat$

$X \rightarrow X\ de\ X/X\ on\ X$

$X \rightarrow X\ de\ X/X\ of\ X$

$X \rightarrow X\ de\ X/X\ 's\ X$

$S \rightarrow X/X$

Coarse-to-fine Parsing

$X^1 \rightarrow dianzi\ shang/the\ mat$
$X^2 \rightarrow dianzi\ shang/mat$
$X^3 \rightarrow mao/the\ cat$
$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$
$X^4 \rightarrow X^2\ de\ X^3/X^3\ of\ X^2$
$X^5 \rightarrow X^1\ de\ X^3/X^1\ 's\ X^3$
$S \rightarrow X^4/X^4$
$S \rightarrow X^5/X^5$

$X \rightarrow dianzi\ shang/the\ ma$
$X \rightarrow dianzi\ shang/mat$
$X \rightarrow mao/the\ cat$
$X \rightarrow X\ de\ X/X\ on\ X$
$X \rightarrow X\ de\ X/X\ of\ X$
$X \rightarrow X\ de\ X/X\ 's\ X$
$S \rightarrow X/X$

Good news: Grammar constant shrinks

Coarse-to-fine Parsing

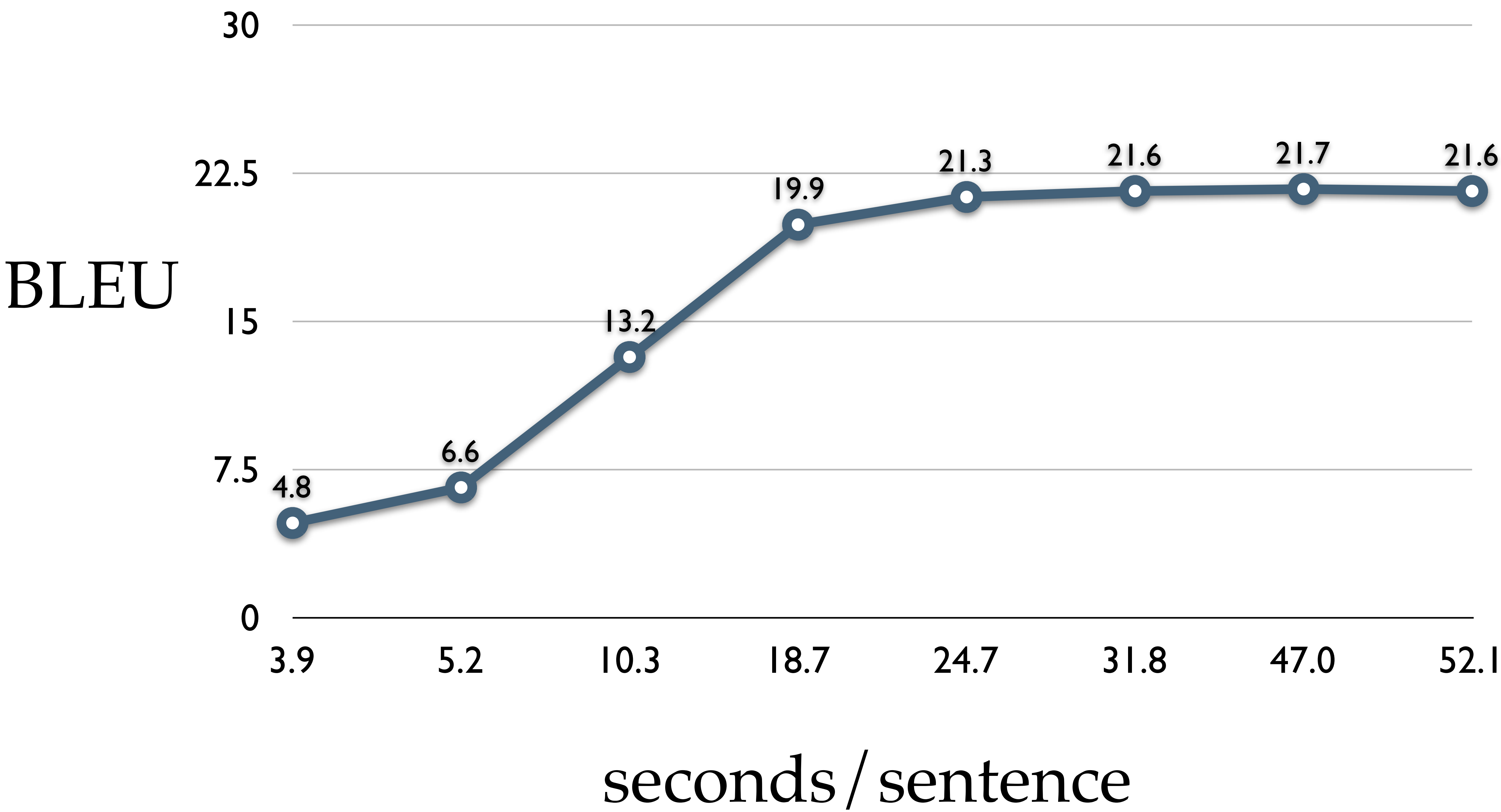
$$X^1 \rightarrow dianzi\ shang/the\ mat$$
$$X^2 \rightarrow dianzi\ shang/mat$$
$$X^3 \rightarrow mao/the\ cat$$
$$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$$
$$X^4 \rightarrow X^2\ de\ X^3/X^3\ of\ X^2$$
$$X^5 \rightarrow X^1\ de\ X^3/X^1\ 's\ X^3$$
$$S \rightarrow X^4/X^4$$
$$S \rightarrow X^5/X^5$$

$$X \rightarrow dianzi\ shang/the\ ma$$
$$X \rightarrow dianzi\ shang/mat$$
$$X \rightarrow mao/the\ cat$$
$$X \rightarrow X\ de\ X/X\ on\ X$$
$$X \rightarrow X\ de\ X/X\ of\ X$$
$$X \rightarrow X\ de\ X/X\ 's\ X$$
$$S \rightarrow X/X$$

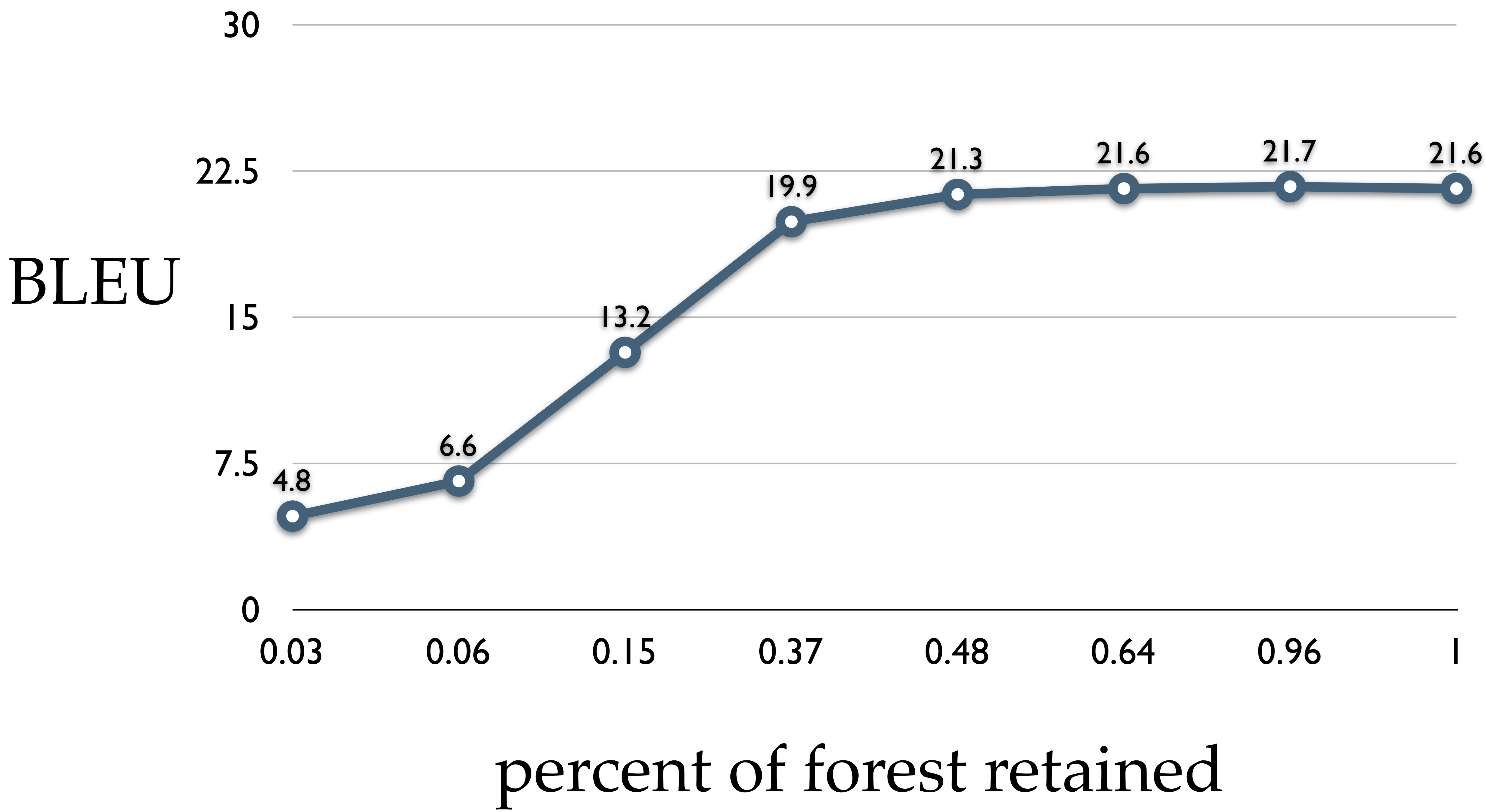
Good news: Grammar constant shrinks

Bad news: Can sometimes prune away all parses

Coarse-to-fine Parsing



Coarse-to-fine Parsing



Grammar Pruning

$X^1 \rightarrow dianzi\ shang/the\ mat$

$X^2 \rightarrow dianzi\ shang/mat$

$X^3 \rightarrow mao/the\ cat$

$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$

$X^4 \rightarrow X^2\ de\ X^3/X^3\ of\ X^2$

$X^5 \rightarrow X^1\ de\ X^3/X^1\ 's\ X^3$

$S \rightarrow X^4/X^4$

$S \rightarrow X^5/X^5$

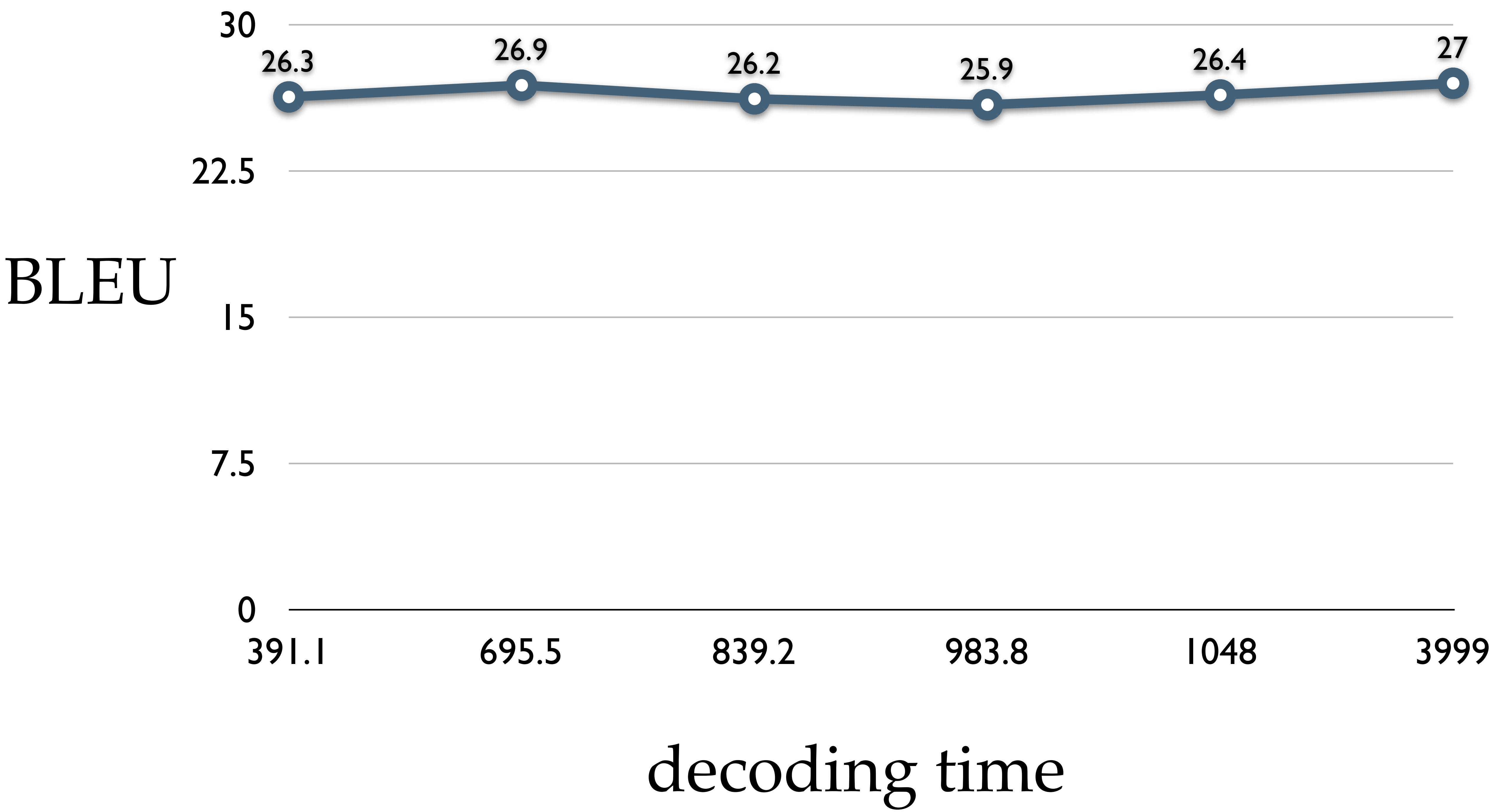
Grammar Pruning

$X^1 \rightarrow dianzi\ shang/the\ mat$
$X^2 \rightarrow dianzi\ shang/mat$
$X^3 \rightarrow mao/the\ cat$
$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$
$X^4 \rightarrow X^2\ de\ X^3/X^3\ of\ X^2$
$X^5 \rightarrow X^1\ de\ X^3/X^1\ 's\ X^3$
$S \rightarrow X^4/X^4$
$S \rightarrow X^5/X^5$

$X^1 \rightarrow dianzi\ shang/the\ mat$
$X^3 \rightarrow mao/the\ cat$
$X^4 \rightarrow X^1\ de\ X^3/X^3\ on\ X^1$
$S \rightarrow X^4/X^4$

Grammar pruning

(note: different data conditions)



Current Results Summary

Current Results Summary

- Source parse forest pruning works.

Current Results Summary

- Source parse forest pruning works.
- Can reduce overall decoding time by 40% with *unpruned* grammars.

Current Results Summary

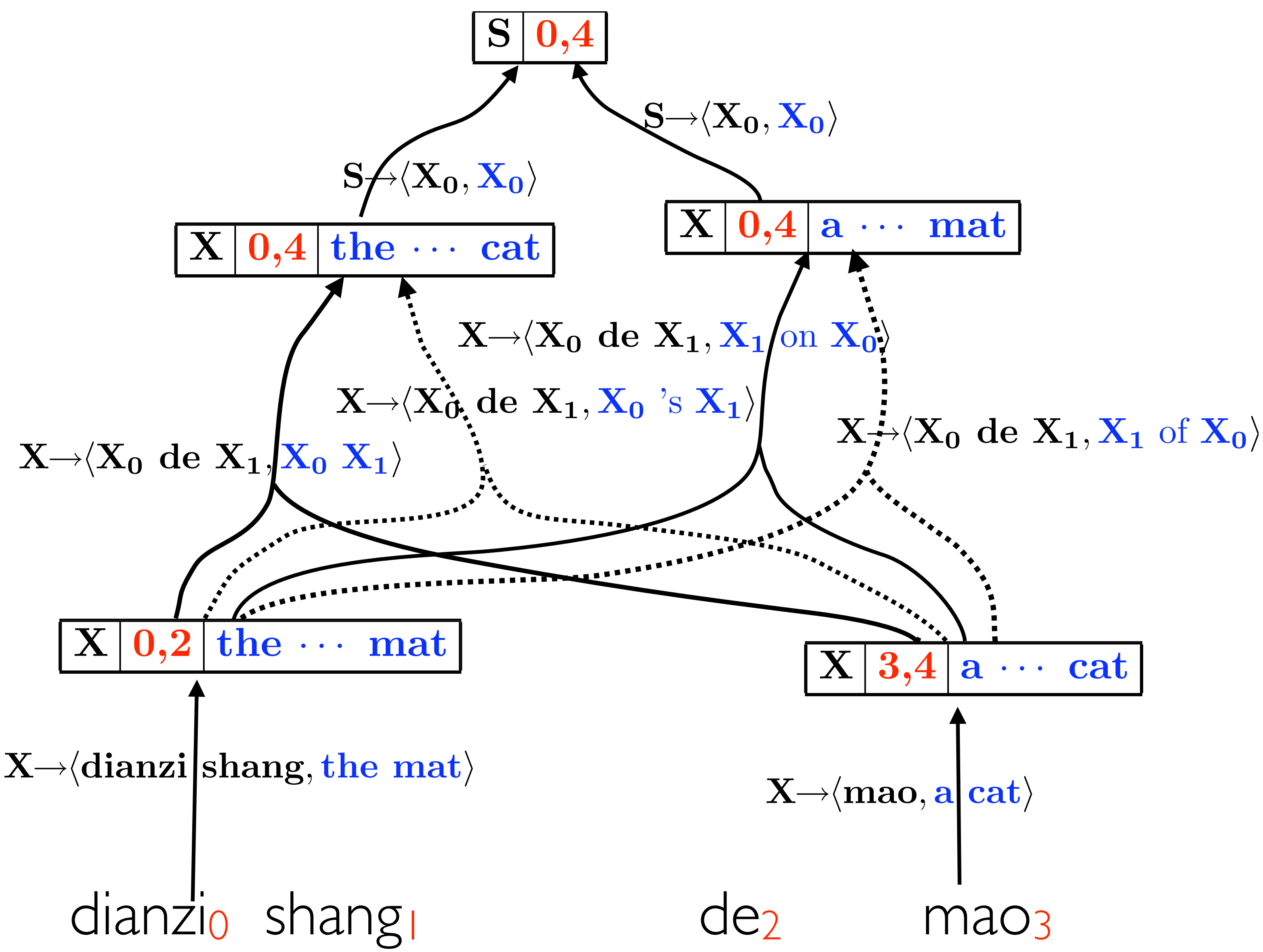
- Source parse forest pruning works.
- Can reduce overall decoding time by 40% with *unpruned* grammars.
- Can reduce decoding time by an order of magnitude with pruned grammars (*maybe* at some cost in BLEU).

Current Results Summary

- Source parse forest pruning works.
- Can reduce overall decoding time by 40% with *unpruned* grammars.
- Can reduce decoding time by an order of magnitude with pruned grammars (*maybe* at some cost in BLEU).
- Ongoing work on more interesting algorithms...

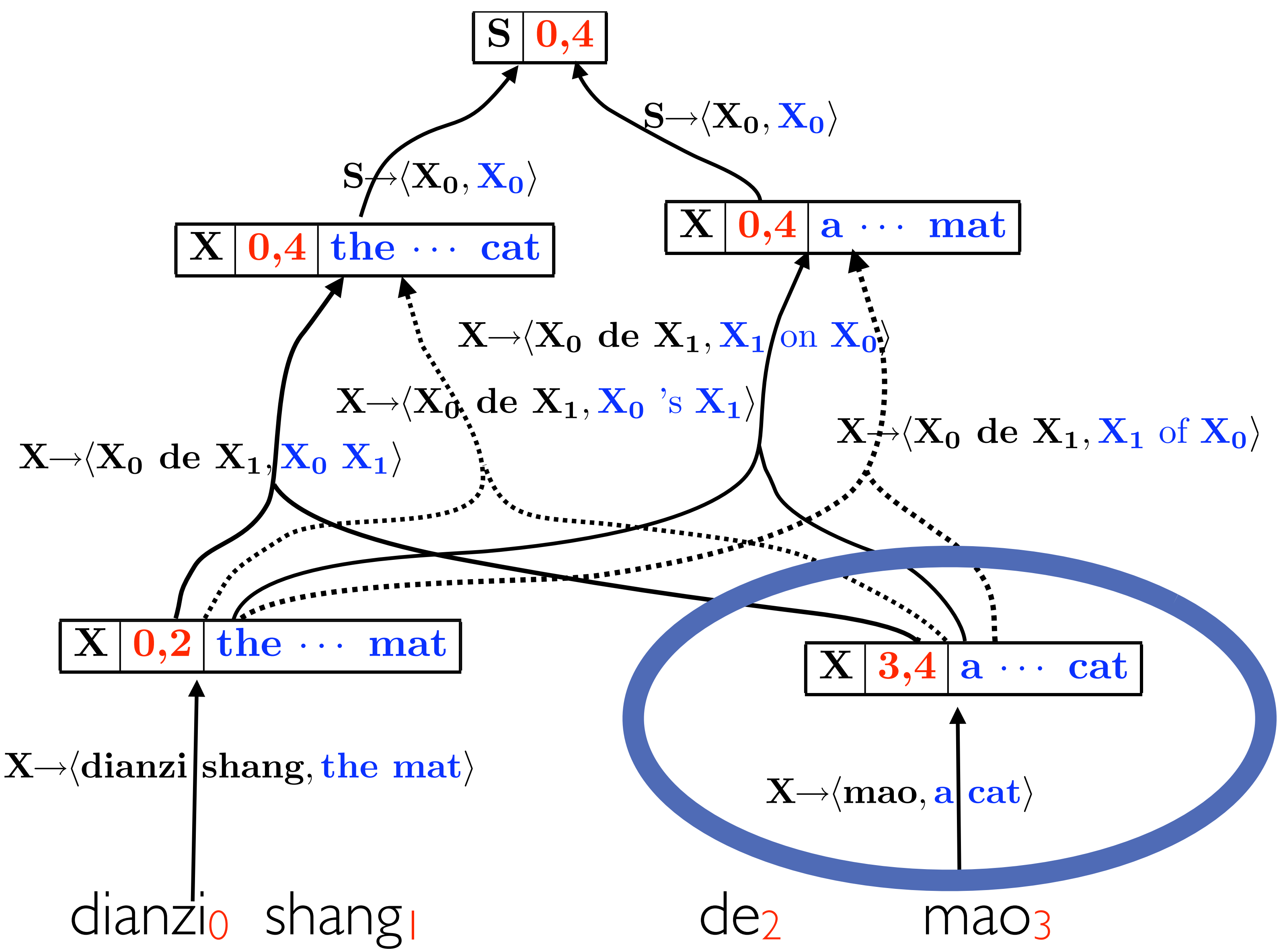
Parse Forests as Grammars

Input: <*dianzi shiang de mao* , *a cat on the mat*>



Parse Forests as Grammars

Input: <*dianzi shiang de mao* , *a cat on the mat*>

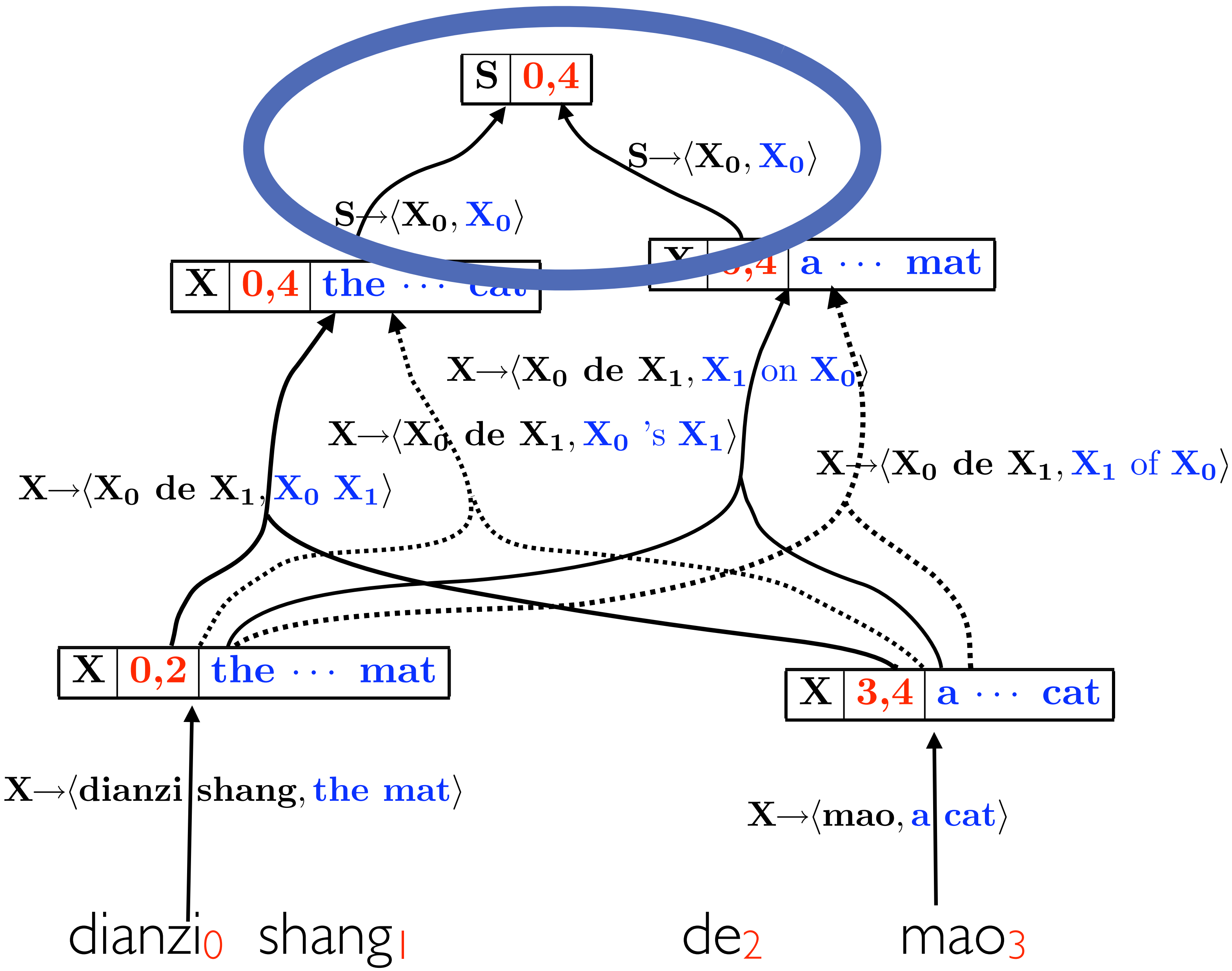


Isomorphic CFG

[X34] → *a cat*

Parse Forests as Grammars

Input: <*dianzi shiang de mao* , *a cat on the mat*>

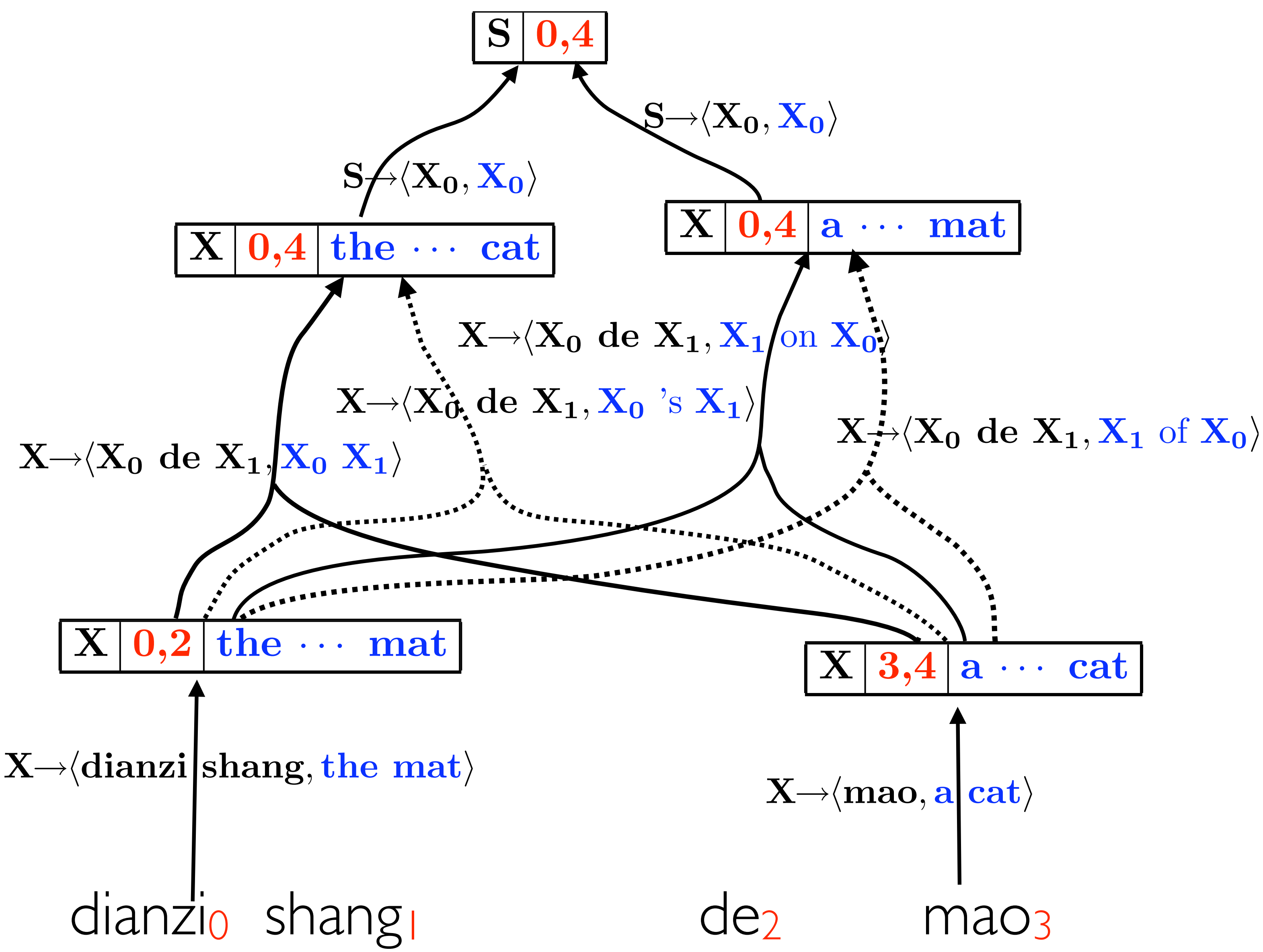


Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

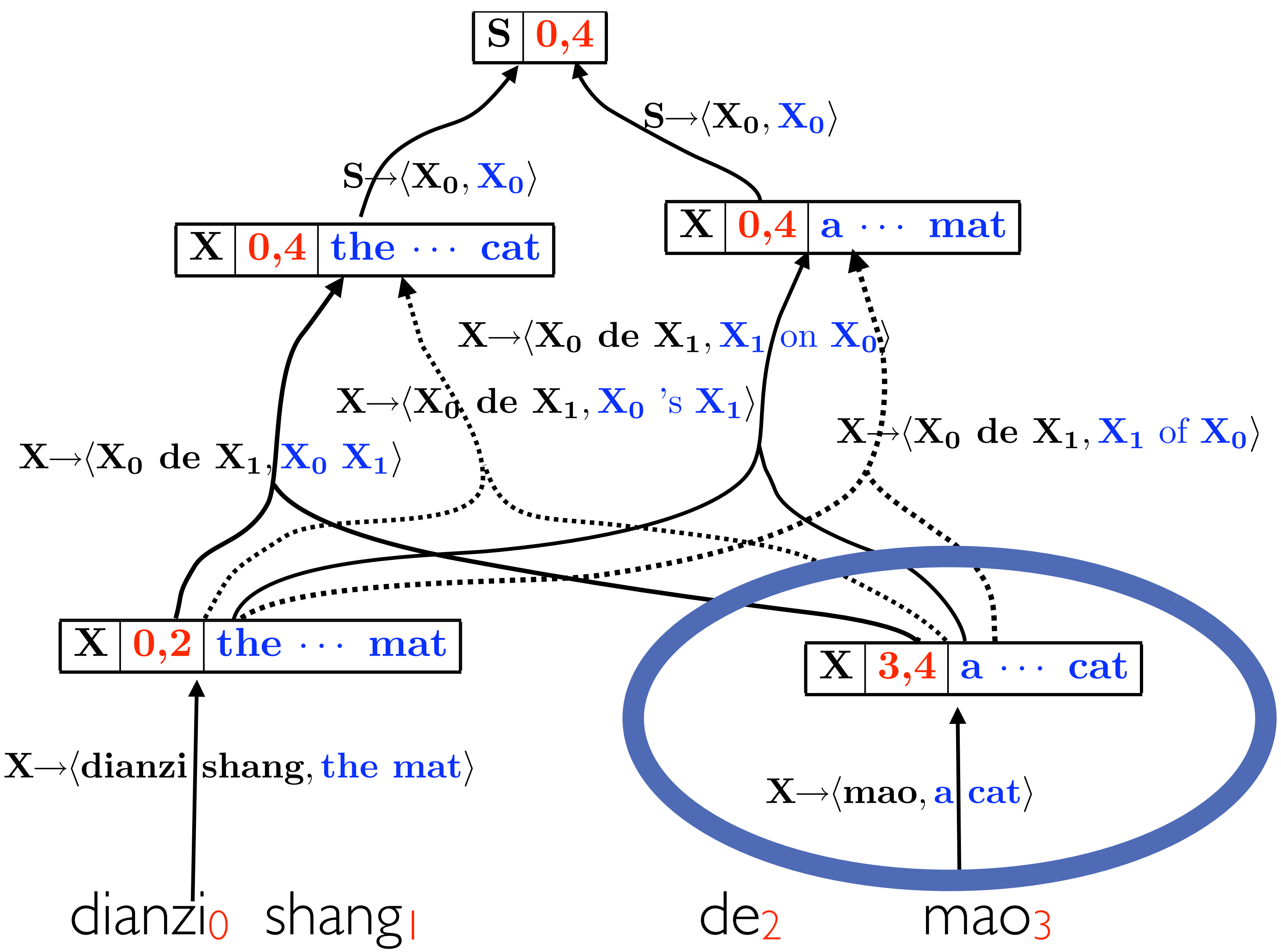
Parse Forests as Grammars

Input: <*dianzi shiang de mao* , *a cat on the mat*>



Parse Forests as Grammars

Input: <*dianzi shiang de mao* , *a cat on the mat*>



Isomorphic CFG

[X34] → *a cat*

Translation is Intersection

- Translation by parsing is intersection
- Intersect source with grammar
- Yields a parse forest → target grammar
- Generate with target grammar
- Intersect with target language model (regular language)

Generation from Grammars

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

$$S_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X^{a}_{0,4}$$

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

$$X^a_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X_{3,4} \text{ on } X_{0,2}$$

$$S_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X^a_{0,4}$$

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

$$X^a_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X_{3,4} \text{ on } X_{0,2}$$

$$S_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X^a_{0,4}$$

$$X_{3,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet \text{a cat}$$

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

$$X^a_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X_{3,4} \text{ on } X_{0,2}$$

$$X_{3,4,\langle s\rangle,\text{a}} \longrightarrow \text{a} \bullet \text{cat}$$

$$S_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X^a_{0,4}$$

$$X_{3,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet\text{a cat}$$

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

$$X_{3,4,\langle s\rangle,\mathrm{cat}} \longrightarrow \text{a cat} \bullet$$

$$X^{a}_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X_{3,4} \text{ on } X_{0,2}$$

$$X_{3,4,\langle s\rangle,\mathrm{a}} \longrightarrow \text{a} \bullet \text{cat}$$

$$S_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X^{a}_{0,4}$$

$$X_{3,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet \text{a cat}$$

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

$$X_{3,4,\langle s \rangle,\text{cat}} \longrightarrow \text{a cat} \bullet$$

$$X^a_{0,4,\langle s \rangle,\langle s \rangle} \longrightarrow \bullet X_{3,4} \text{ on } X_{0,2}$$

$$X_{3,4,\langle s \rangle,\text{a}} \longrightarrow \text{a} \bullet \text{cat}$$

$$S_{0,4,\langle s \rangle,\langle s \rangle} \longrightarrow \bullet X^a_{0,4}$$

$$X_{3,4,\langle s \rangle,\langle s \rangle} \longrightarrow \bullet \text{a cat}$$

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

$$X_{3,4,\langle s\rangle,\text{cat}} \longrightarrow \text{a cat} \bullet$$

$$X^{a}_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X_{3,4} \text{ on } X_{0,2}$$

$$X_{3,4,\langle s\rangle,\text{a}} \longrightarrow \text{a} \bullet \text{cat}$$

$$S_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X^{a}_{0,4}$$

$$X_{3,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet \text{a cat}$$

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

$$X^{a}_{0,4,\langle s\rangle,\text{cat}} \longrightarrow X_{3,4} \bullet \ \text{on} \ X_{0,2}$$

$$X_{3,4,\langle s\rangle,\text{cat}} \longrightarrow \text{a cat} \bullet$$

$$X^{a}_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X_{3,4} \ \text{on} \ X_{0,2}$$

$$X_{3,4,\langle s\rangle,\text{a}} \longrightarrow \text{a} \bullet \text{cat}$$

$$S_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X^{a}_{0,4}$$

$$X_{3,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet \text{a cat}$$

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

$$X^{a}_{0,4,\langle s\rangle,\text{cat}} \longrightarrow X_{3,4} \bullet \ \text{on} \ X_{0,2}$$

$$X_{3,4,\langle s\rangle,\text{cat}} \longrightarrow \text{a cat} \bullet$$

$$X^{a}_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X_{3,4} \ \text{on} \ X_{0,2}$$

$$X_{3,4,\langle s\rangle,\text{a}} \longrightarrow \text{a} \bullet \text{cat}$$

$$S_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X^{a}_{0,4}$$

$$X_{3,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet \text{a cat}$$

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

$$X_{0,4,\langle s \rangle,\text{on}} \longrightarrow X_{3,4} \text{ on} \bullet X_{0,2}$$

$$X^{a}_{0,4,\langle s \rangle,\text{cat}} \longrightarrow X_{3,4} \bullet \text{ on } X_{0,2}$$

$$X_{3,4,\langle s \rangle,\text{cat}} \longrightarrow \text{a cat} \bullet$$

$$X^{a}_{0,4,\langle s \rangle,\langle s \rangle} \longrightarrow \bullet X_{3,4} \text{ on } X_{0,2}$$

$$X_{3,4,\langle s \rangle,\text{a}} \longrightarrow \text{a} \bullet \text{cat}$$

$$S_{0,4,\langle s \rangle,\langle s \rangle} \longrightarrow \bullet X^{a}_{0,4}$$

$$X_{3,4,\langle s \rangle,\langle s \rangle} \longrightarrow \bullet \text{a cat}$$

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

$$X_{0,4,\langle s\rangle,\mathrm{on}} \longrightarrow X_{3,4}\ \mathrm{on} \bullet X_{0,2}$$

$$X_{0,2,\mathrm{on},\mathrm{mat}} \longrightarrow \mathrm{the\ mat} \bullet$$

$$X^{a}_{0,4,\langle s\rangle,\boxed{\mathrm{cat}}} \longrightarrow X_{3,4} \bullet\ \mathrm{on}\ X_{0,2}$$

$$\boxed{X_{3,4}},\boxed{\langle s\rangle},\boxed{\mathrm{cat}} \longrightarrow \mathrm{a\ cat} \bullet$$

$$X^{a}_{0,4,\langle s\rangle,\boxed{\langle s\rangle}} \longrightarrow \bullet \boxed{X_{3,4}}\ \mathrm{on}\ X_{0,2}$$

$$X_{3,4,\langle s\rangle,\mathrm{a}} \longrightarrow \mathrm{a} \bullet \mathrm{cat}$$

$$S_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X^{a}_{0,4}$$

$$X_{3,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet \mathrm{a\ cat}$$

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*

[X02] → *the mat*

[X04a] → [X34] *on* [X02]

[X04a] → [X34] *of* [X02]

[X04b] → [X02] *'s* [X34]

[X04b] → [X02] [X34]

[S] → [X04a]

[S] → [X04b]

$$X_{0,4,\langle s\rangle,\mathrm{mat}} \longrightarrow X_{3,4} \text{ on } X_{0,2} \bullet$$

$$X_{0,4,\langle s\rangle,\mathrm{on}} \longrightarrow X_{3,4} \text{ on} \bullet X_{0,2}$$

$$X_{0,2,\mathrm{on},\mathrm{mat}} \longrightarrow \text{the mat} \bullet$$

$$X^{a}_{0,4,\langle s\rangle,\boxed{\mathrm{cat}}} \longrightarrow X_{3,4} \bullet \text{ on } X_{0,2}$$

$$\boxed{X_{3,4}}{}_{,\boxed{\langle s\rangle},\boxed{\mathrm{cat}}} \longrightarrow \text{a cat} \bullet$$

$$X^{a}_{0,4,\langle s\rangle,\boxed{\langle s\rangle}} \longrightarrow \bullet \boxed{X_{3,4}} \text{ on } X_{0,2}$$

$$X_{3,4,\langle s\rangle,\mathrm{a}} \longrightarrow \text{a} \bullet \text{cat}$$

$$S_{0,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet X^{a}_{0,4}$$

$$X_{3,4,\langle s\rangle,\langle s\rangle} \longrightarrow \bullet \text{a cat}$$

Generation from Grammars

$$X_{0,4,\langle s\rangle,\mathrm{mat}} \rightarrow X_{3,4} \text{ on } X_{0,2}$$

$$X_{0,2,\mathrm{on},\mathrm{mat}} \rightarrow \text{the mat}$$

$$X_{3,4,\langle s\rangle,\mathrm{cat}} \rightarrow \text{a cat}$$

Generation from Grammars

Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*
[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *'s* [X34]
[X04b] → [X02] [X34]
[S] → [X04a]
[S] → [X04b]

$$X_{0,4,\langle s \rangle,\mathrm{mat}} \longrightarrow X_{3,4} \text{ on } X_{0,2}$$

$$X_{0,2,\mathrm{on},\mathrm{mat}} \longrightarrow \text{the mat}$$

$$X_{3,4,\langle s \rangle,\mathrm{cat}} \longrightarrow \text{a cat}$$

Theoretical Outcomes

- Works for arbitrary grammars, not just binary grammars
- Asymptotically faster than cube pruning ("hook trick", Liang et al. 2006).
- Produces lots of admissible heuristics (A*)
- No cube pruning: everything is monotonic.

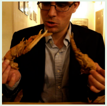
Conclusions

- Faster algorithms are needed to make induced grammars practical.
- Workshop made significant progress towards this goal.
- More improvements are underway.

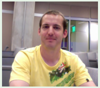
Outline

Vlad Eidelman

Ziyuan Wang



Adam Lopez



Jon Graehl

ThuyLinh Nguyen

- 3:20pm Parametric models: posterior regularisation. Desai
- 3:35pm Training models with rich features spaces. Vlad
- 3:50pm Decoding with complex grammars. Adam
- 4:20pm Closing remarks. Phil
- 4:25pm Finish.

Statistical machine translation: state-of-the-art

Urdu → English

اس حملہ کے بعد بڑی تعداد میں مقامی باشندوں نے علاقوں کو خالی کردیا ہے .

In this attack a large number of local residents has should vacate areas.

- Current state-of-the-art translation models struggle with language pairs which exhibit large differences in structure.

Statistical machine translation: our unsupervised grammars

Urdu → English

اس حملہ کے بعد بڑی تعداد میں مقامی باشندوں نے علاقوں کو خالی کردیا ہے .

After this attack, a large number of local residents have to vacate the areas.

- In this workshop we've made some small steps towards better translations for difficult language pairs.

Statistical machine translation: our unsupervised grammars

Urdu → English

اس حملہ کے بعد بڑی تعداد میں مقامی باشندوں نے علاقوں کو خالی کردیا ہے ۔

↓

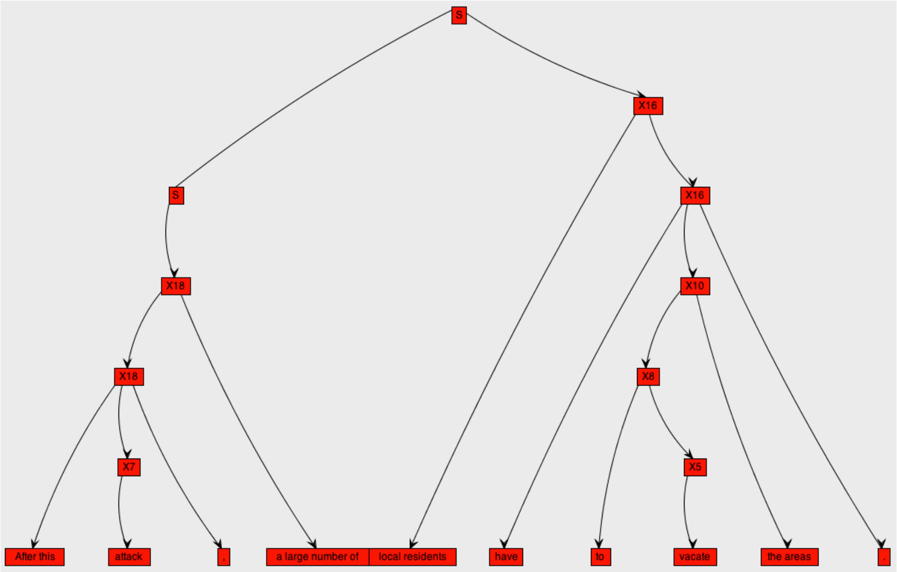
After this attack, a large number of local residents have to vacate the areas.

- In this workshop we've made some small steps towards better translations for difficult language pairs.

Google Translate:

**After the attack a number of local residents has blank areas.*

Induced Translation Structure



What we've achieved:

- The first unsupervised labelled SCFG induction algorithms:
  - by clustering translation phrases which occur in the same context we can learn which phrases are substituteable,
  - we have implemented parametric and non-parametric Bayesian clustering algorithms and shown positive results on real translation tasks.
- Improved SCFG decoders that efficiently decode grammars with many labels:
  - we have created faster search algorithms tuned for syntactic grammars.
- Discriminative training regimes to leverage features extracted from these grammars:
  - we've implemented two large scale discriminative algorithms for training our models.

What we've achieved:

- The first unsupervised labelled SCFG induction algorithms:
  - by clustering translation phrases which occur in the same context we can learn which phrases are substituteable,
  - we have implemented parametric and non-parametric Bayesian clustering algorithms and shown positive results on real translation tasks.

- Improved SCFG decoders that efficiently decode grammars with many labels:
  - we have created faster search algorithms tuned for syntactic grammars.

- Discriminative training regimes to leverage features extracted from these grammars:
  - we've implemented two large scale discriminative algorithms for training our models.

Thank you.