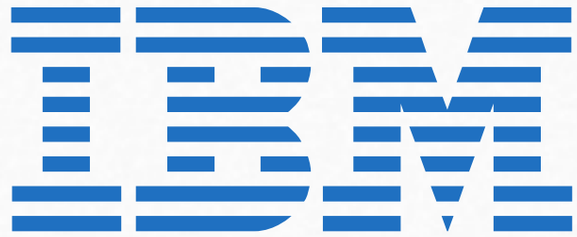
Hierarchical MT Training using Max-Violation Perceptron

Kai Zhao Liang Huang

City University of New York

Haitao Mi Abe Ittycheriah

IBM T.J. Watson Research Center



MT is hard


HERITAGE
WALK


遗产
走路


Watertable
Renaissance Baltimore Harborplace
E Pratt St
Phillips Seafood
USS Constellation
National Aquarium
Chipotle Mexican Grill
Jones Falls
Columbus Park
Miss Shirley's
S President St
Slemmer's Alley
Stiles St
Fawn St
Trinity St
Mince Alley
S Central Ave
S Eden St
Eastern A
Duker Ct
Whole Foods Market
Baltimore Marriott Waterfront
Baltimore Marriott Waterfront
Royal Sonesta Harbor Court Baltimore
Rusty Scupper Restaurant & Bar
Rash Field
Key Hwy
Key Hwy

Discriminative Training for SMT

- discriminative training is not very successful yet in MT
  - although dominant in parsing/tagging
  - can use arbitrary, overlapping, lexicalized features
- most efforts on MT training tune feature weights on the small dev set (~1k sents) not the training set!
  - as a result can only use ~10 dense features (MERT)
  - or ~10k rather impoverished features (MIRA/PRO)
- Liang et al ('06) train on the training set but not successful

MT as Structured Classification

- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron

training example

那	人	咬	了	狗	x

the	man	bit	the	dog	y

(Liang et al '06)

MT as Structured Classification

- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron

training example

那	人	咬	了	狗	x

the	man	bit	the	dog	y

during online learning...

那	人	咬	了	狗	x

(Liang et al '06)

MT as Structured Classification

- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron

training example

那 人 咬 了 狗 x

the man bit the dog y

during online learning...

那 人 咬 了 狗 x

full
search
space

(Liang et al '06)

MT as Structured Classification

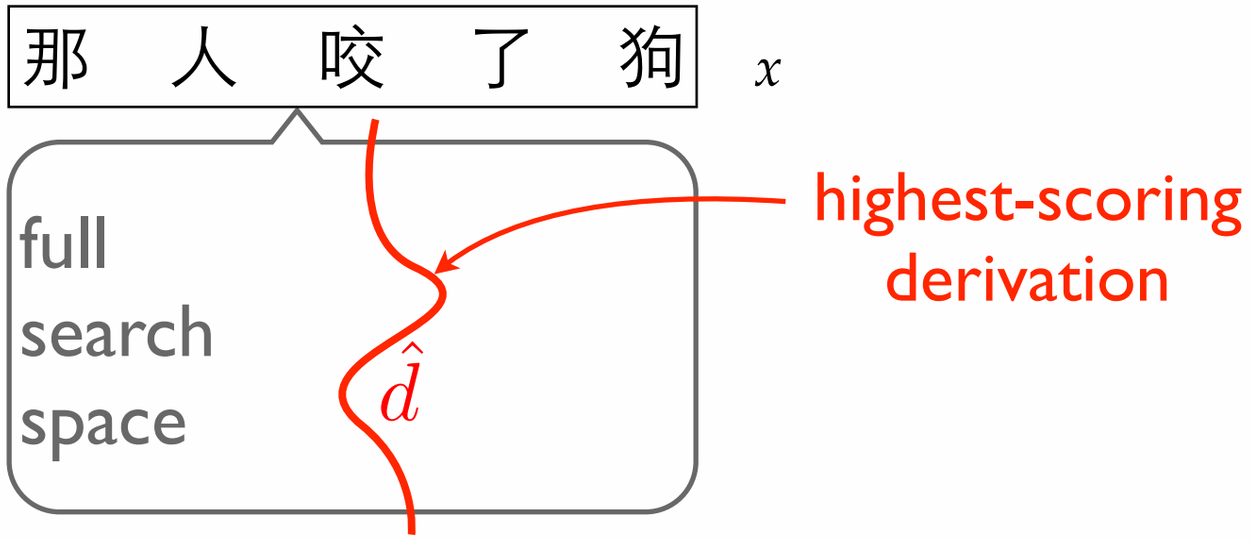
- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron

training example

那 人 咬 了 狗 x

the man bit the dog y

during online learning...



(Liang et al '06)

MT as Structured Classification

- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron

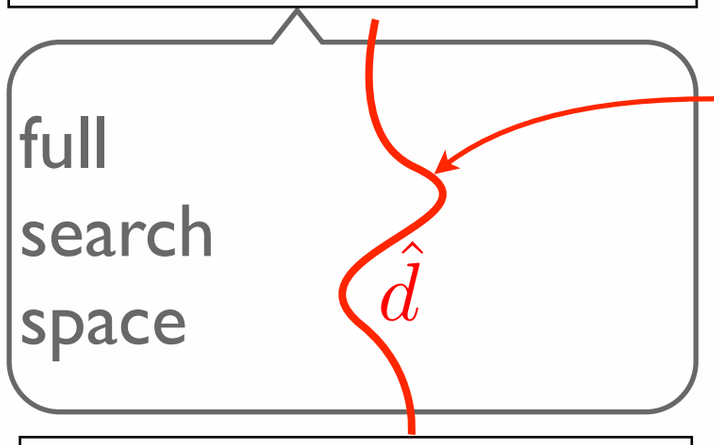
training example

那 人 咬 了 狗 x

the man bit the dog y

during online learning...

那 人 咬 了 狗 x



the dog bit the man z

(Liang et al '06)

MT as Structured Classification

- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron

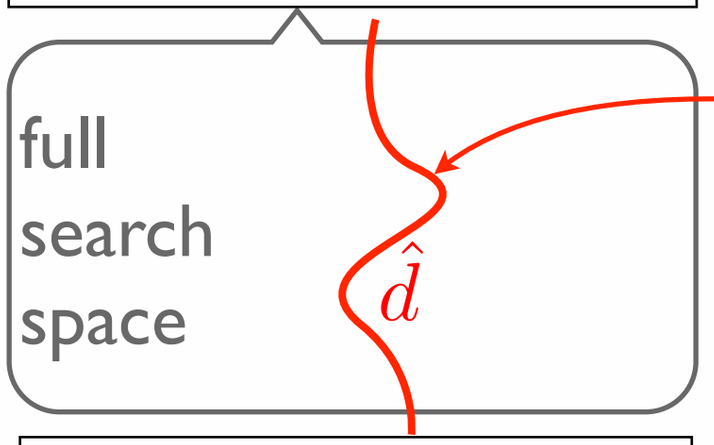
training example

那 人 咬 了 狗 x

the man bit the dog y $\neq$

during online learning...

那 人 咬 了 狗 x

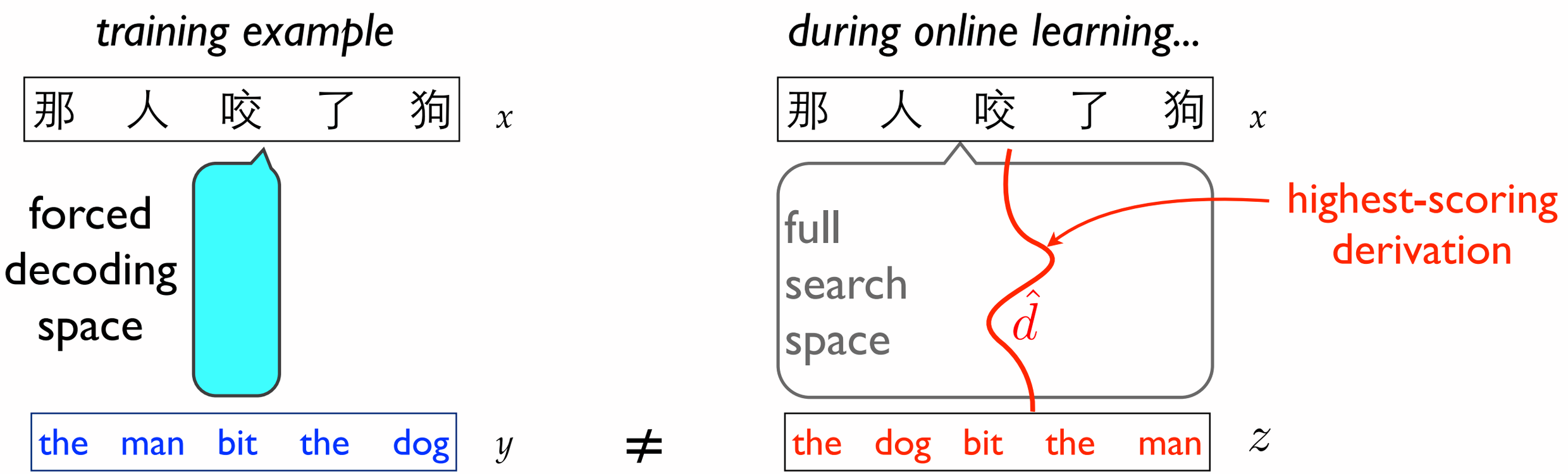


the dog bit the man z

(Liang et al '06)

MT as Structured Classification

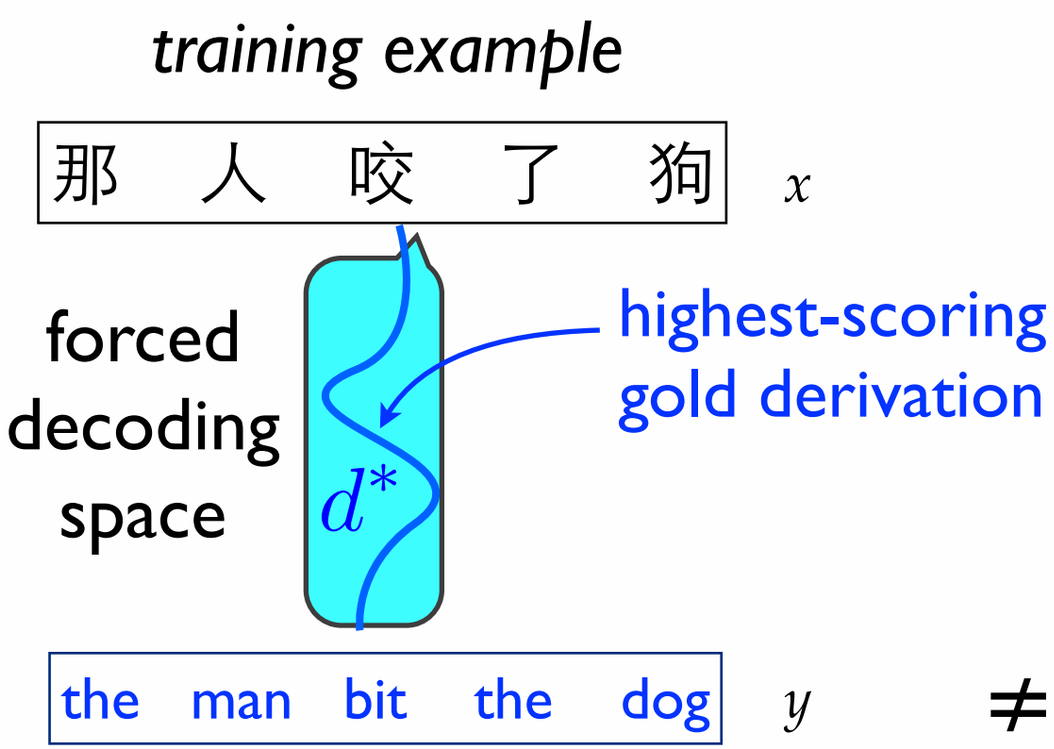
- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron

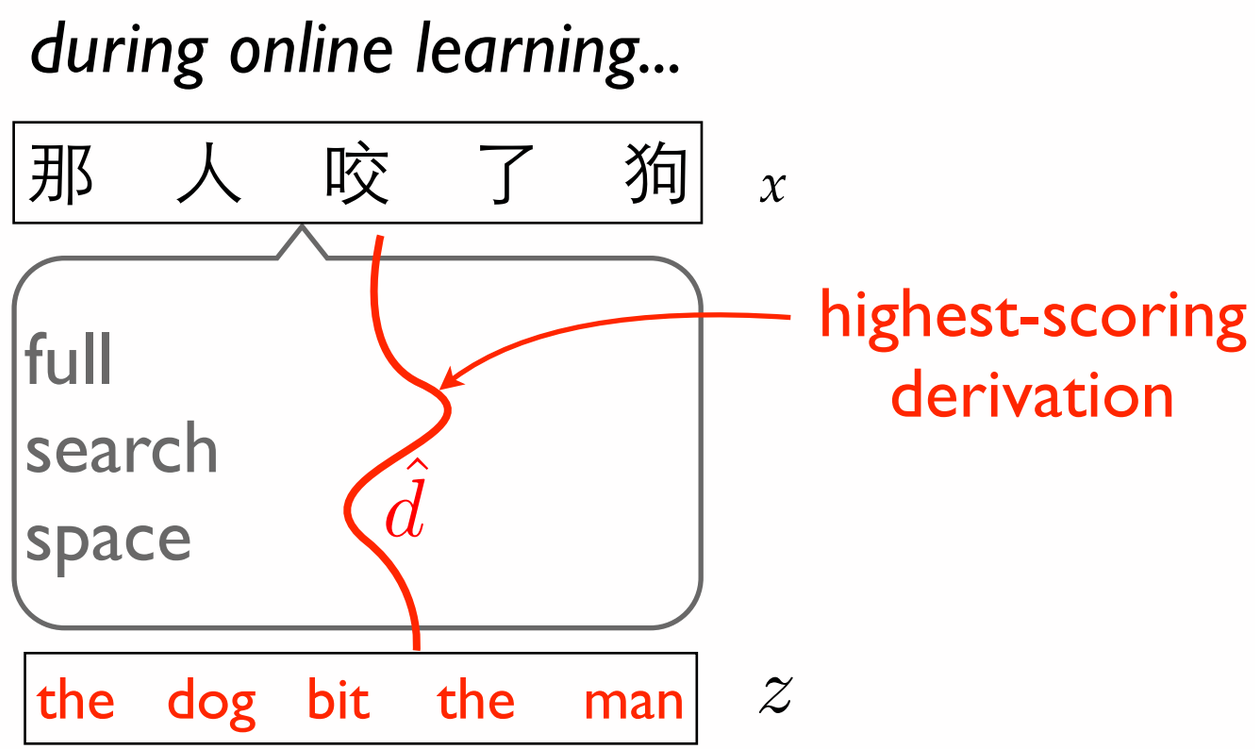


(Liang et al '06)

MT as Structured Classification

- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron

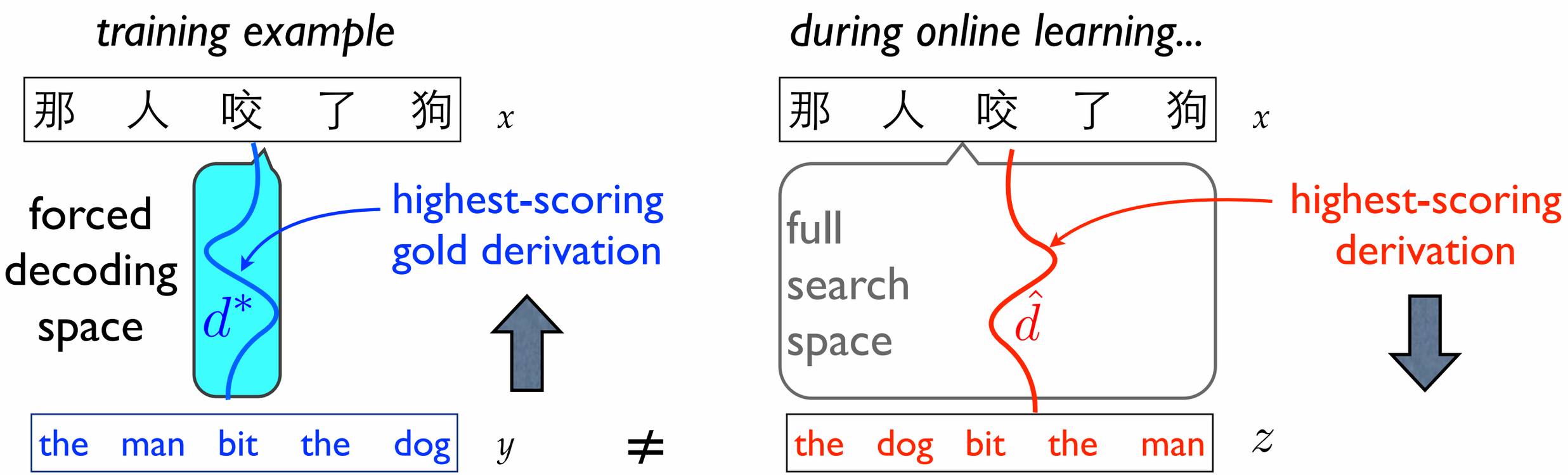




(Liang et al '06)

MT as Structured Classification

- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron



$$\mathbf{w} \leftarrow \mathbf{w} + \Phi(x, d^*) - \Phi(x, \hat{d})$$

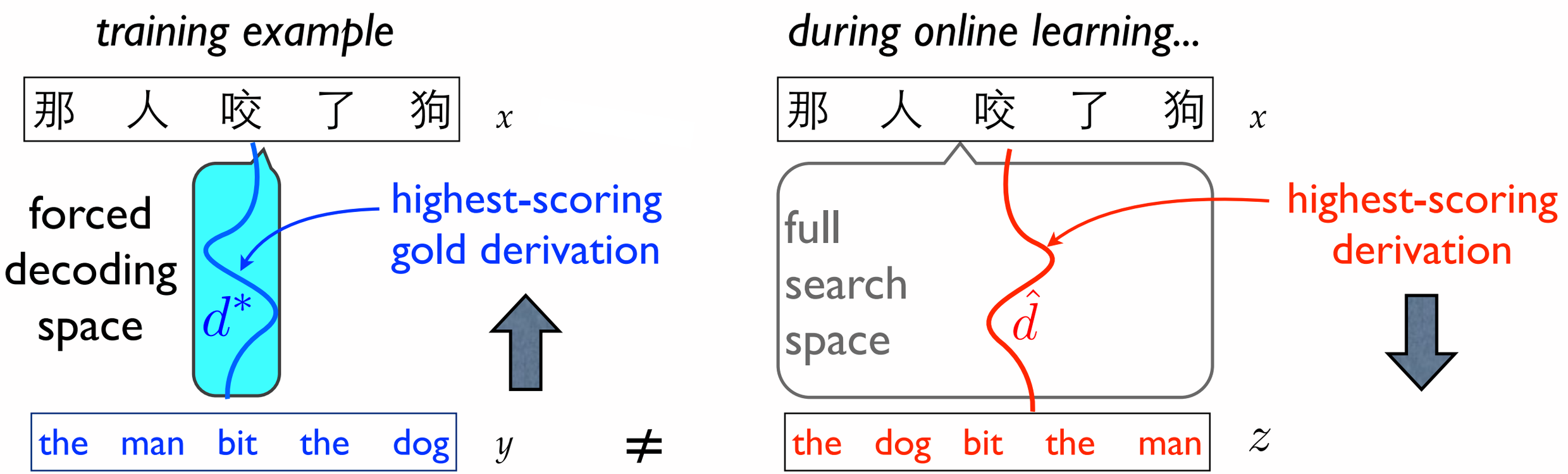
reward correct

penalize wrong

(Liang et al '06)

MT as Structured Classification

- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron



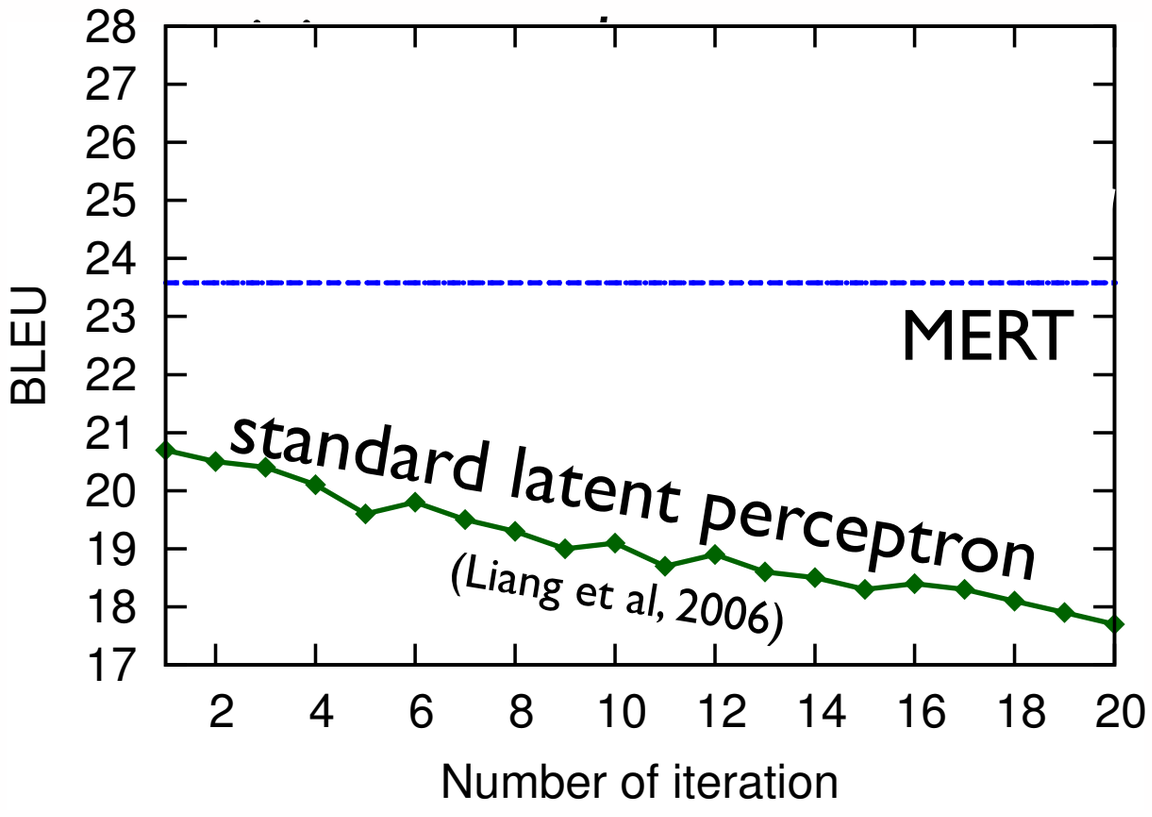
$$\mathbf{w} \leftarrow \mathbf{w} + \Phi(x, d^*) - \Phi(x, \hat{d})$$

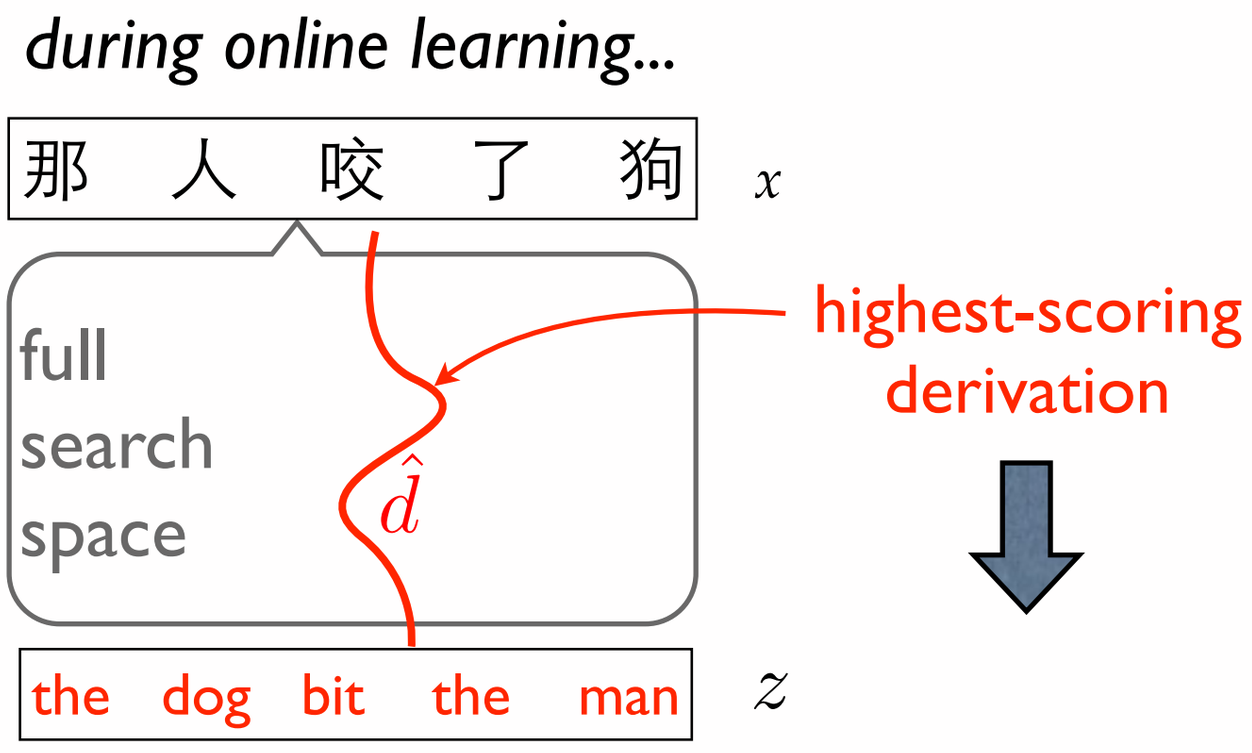
reward correct

penalize wrong

MT as Structured Classification

- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron





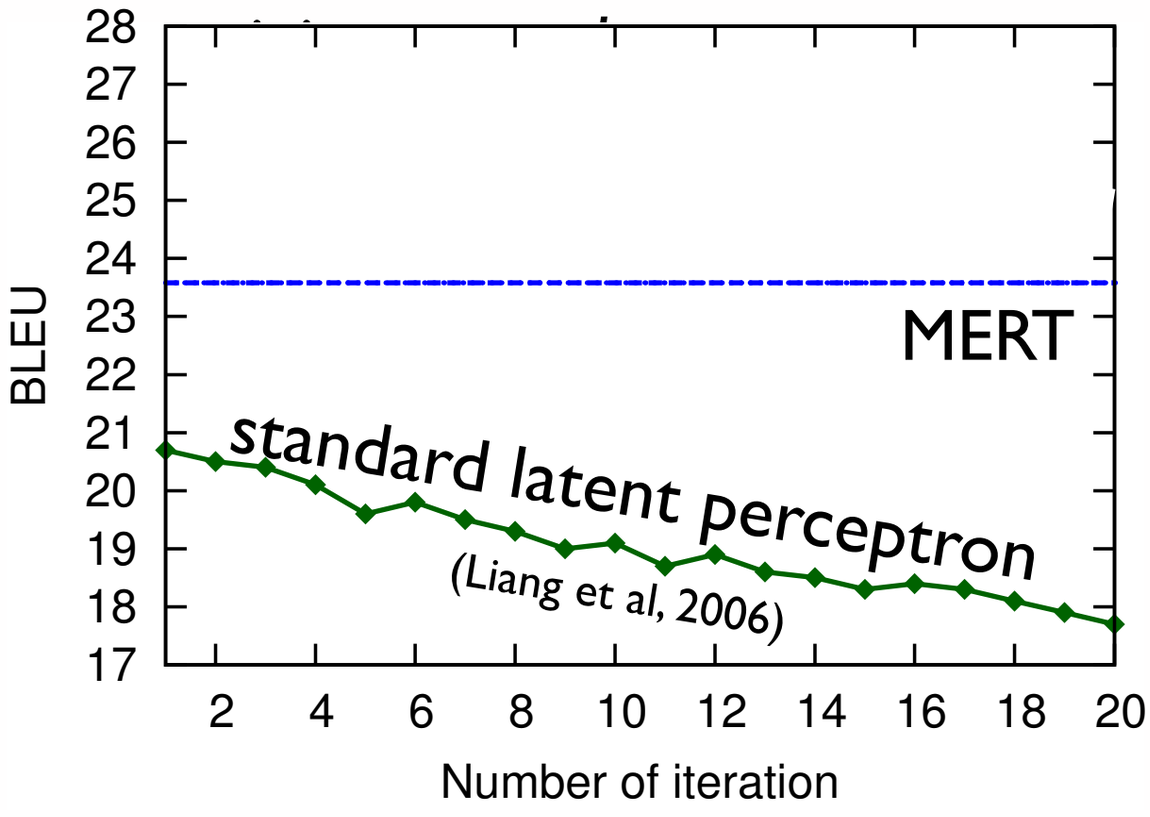
$$\mathbf{w} \leftarrow \mathbf{w} + \Phi(x, d^*) - \Phi(x, \hat{d})$$

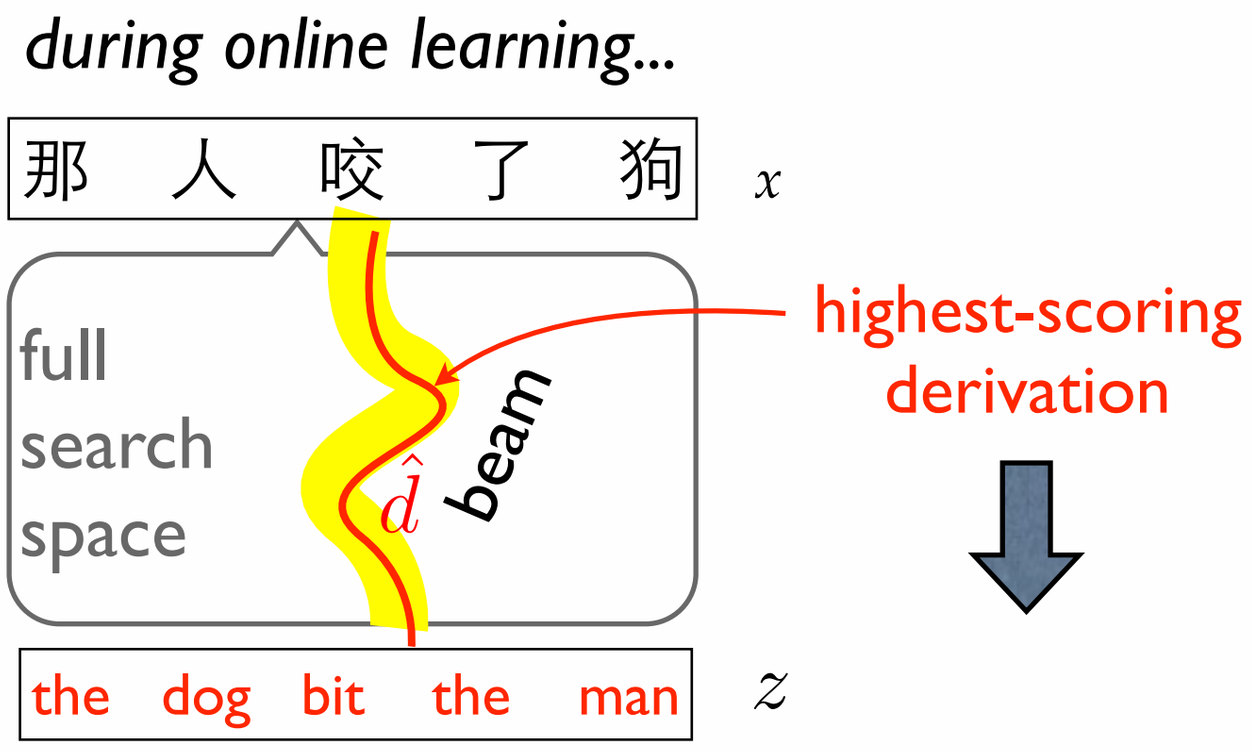
reward correct

penalize wrong

MT as Structured Classification

- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron





$$\mathbf{w} \leftarrow \mathbf{w} + \Phi(x, d^*) - \Phi(x, \hat{d})$$

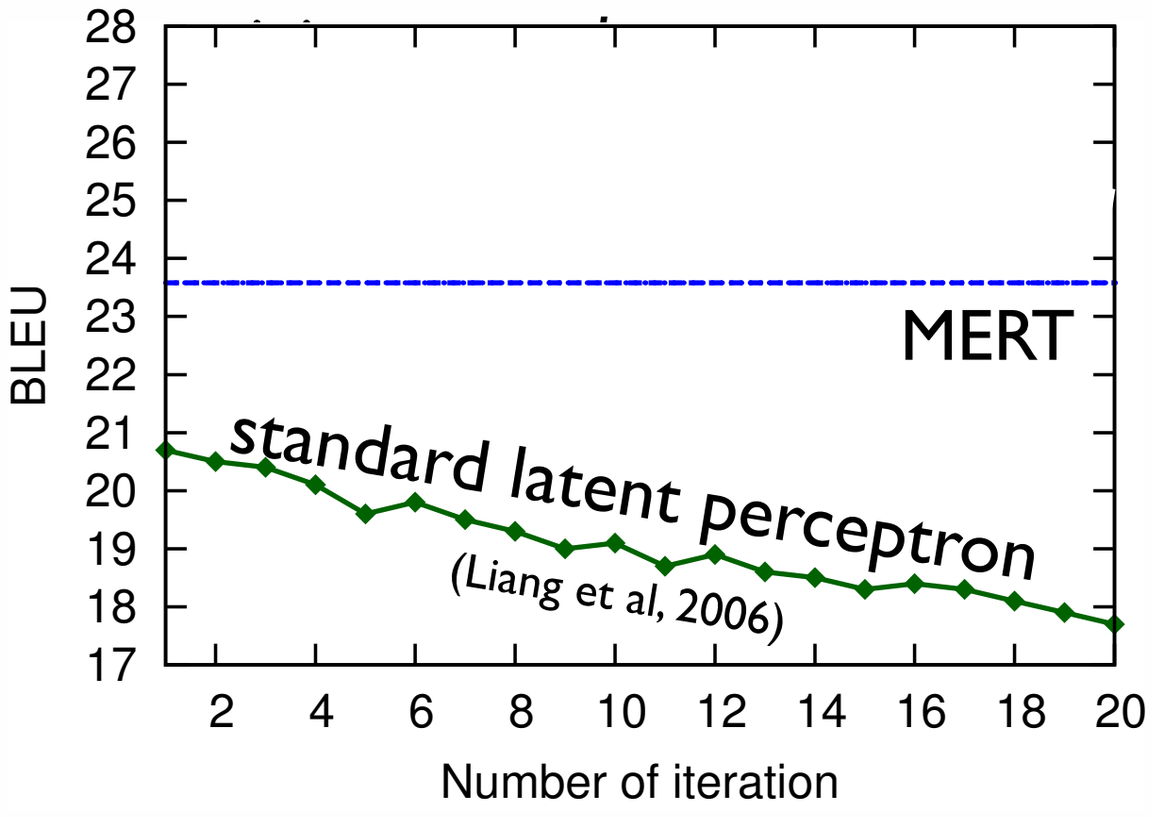
reward correct

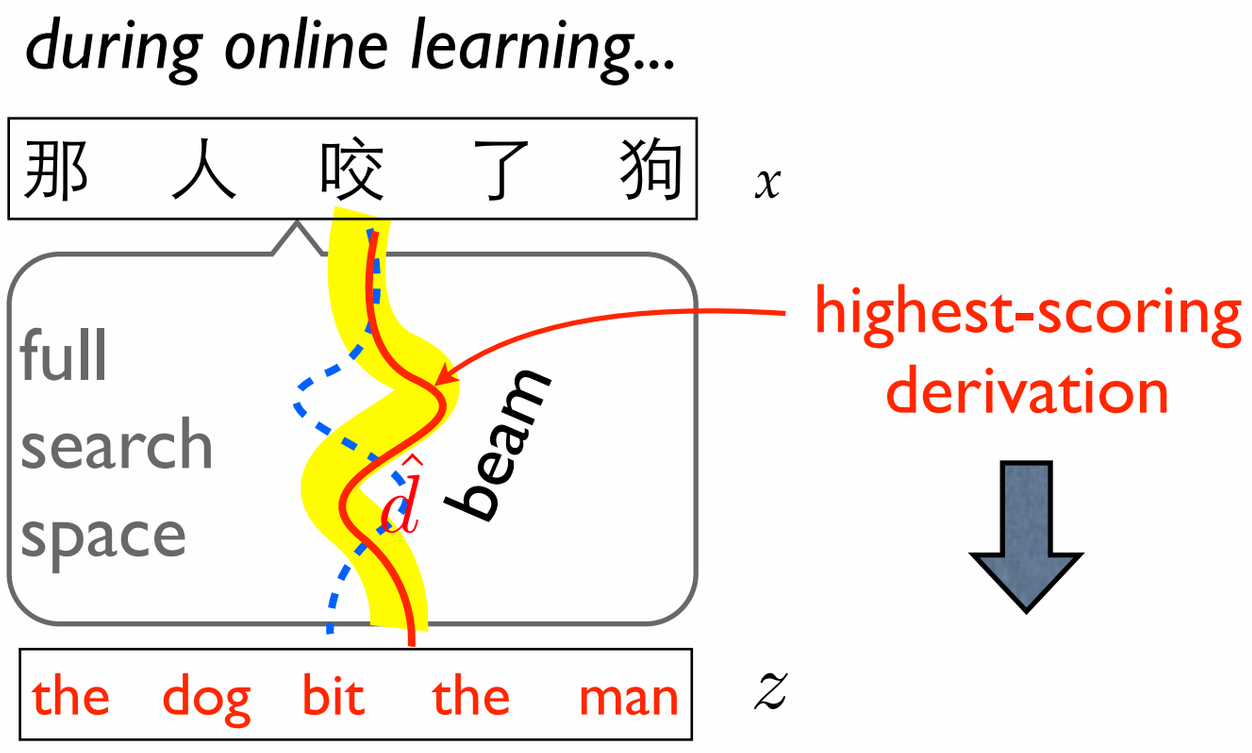
penalize wrong

problem: search errors
(Yu et al 2013)

MT as Structured Classification

- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron





$$\mathbf{w} \leftarrow \mathbf{w} + \Phi(x, d^*) - \Phi(x, \hat{d})$$

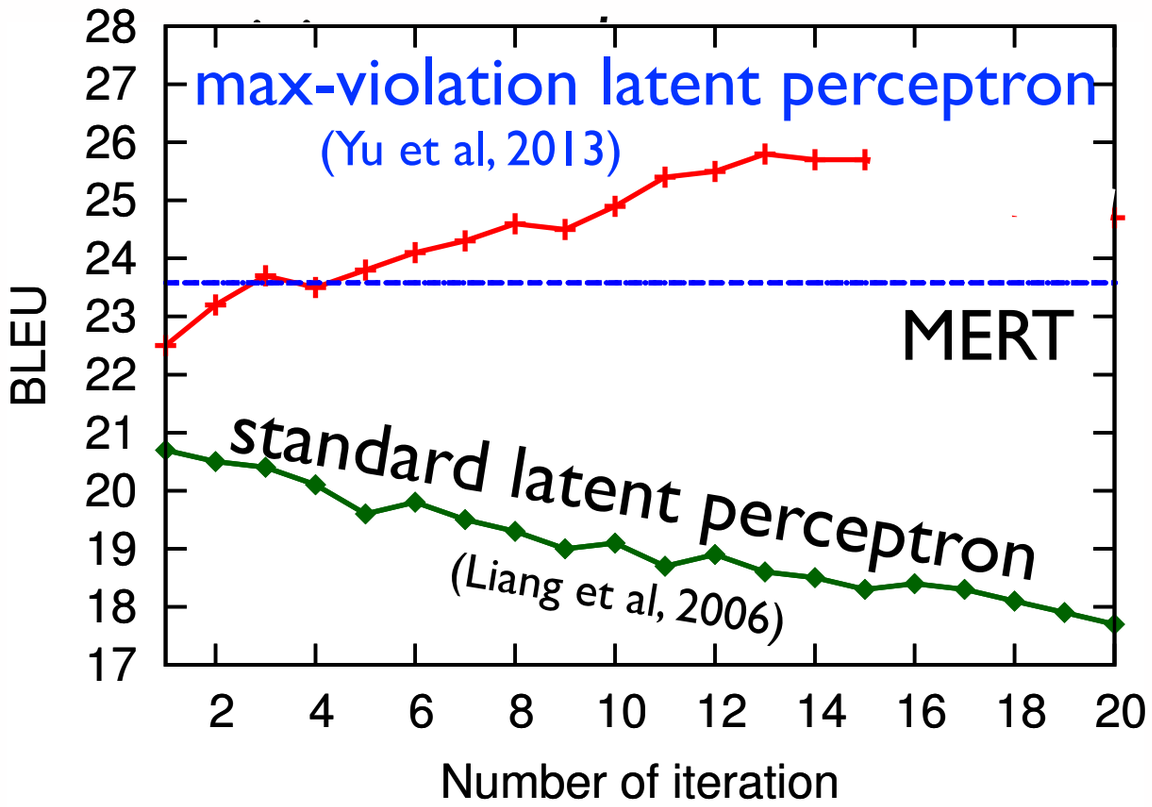
reward correct

penalize wrong

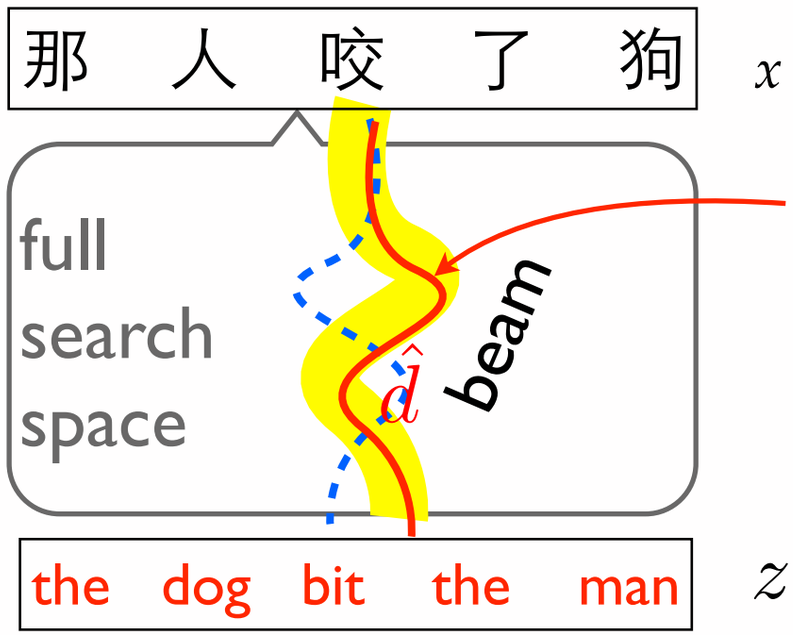
problem: search errors
(Yu et al 2013)

MT as Structured Classification

- with latent variables (hidden derivations)
- structured perceptron => latent-variable structured perceptron



during online learning...



$$\mathbf{w} \leftarrow \mathbf{w} + \Phi(x, d^*) - \Phi(x, \hat{d})$$

reward correct

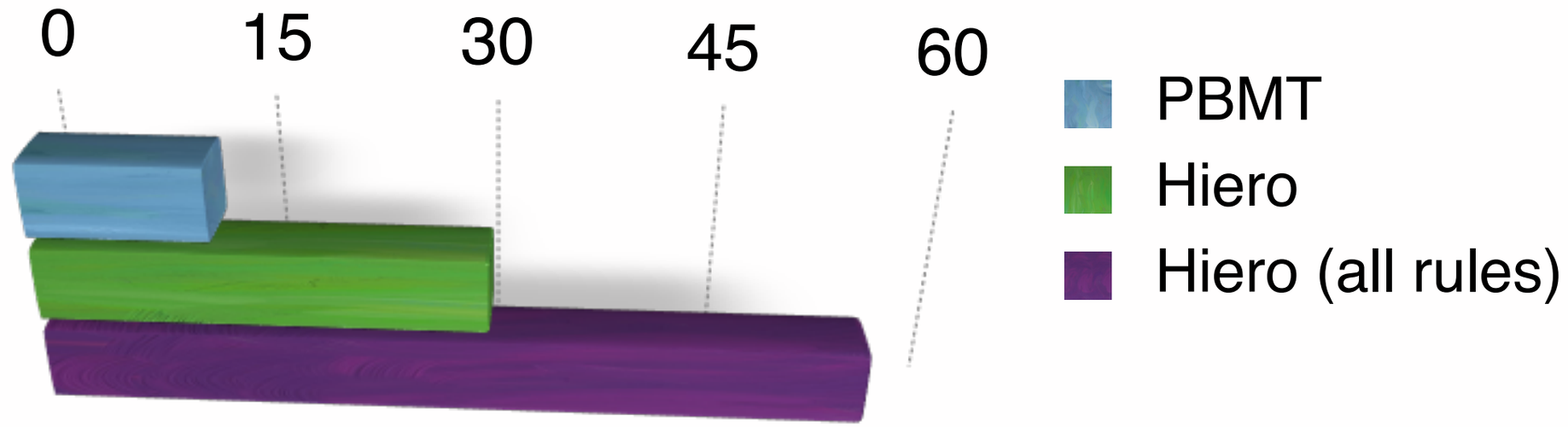
penalize wrong

problem: search errors
(Yu et al 2013)

From PBMT to Hiero

From PBMT to Hiero

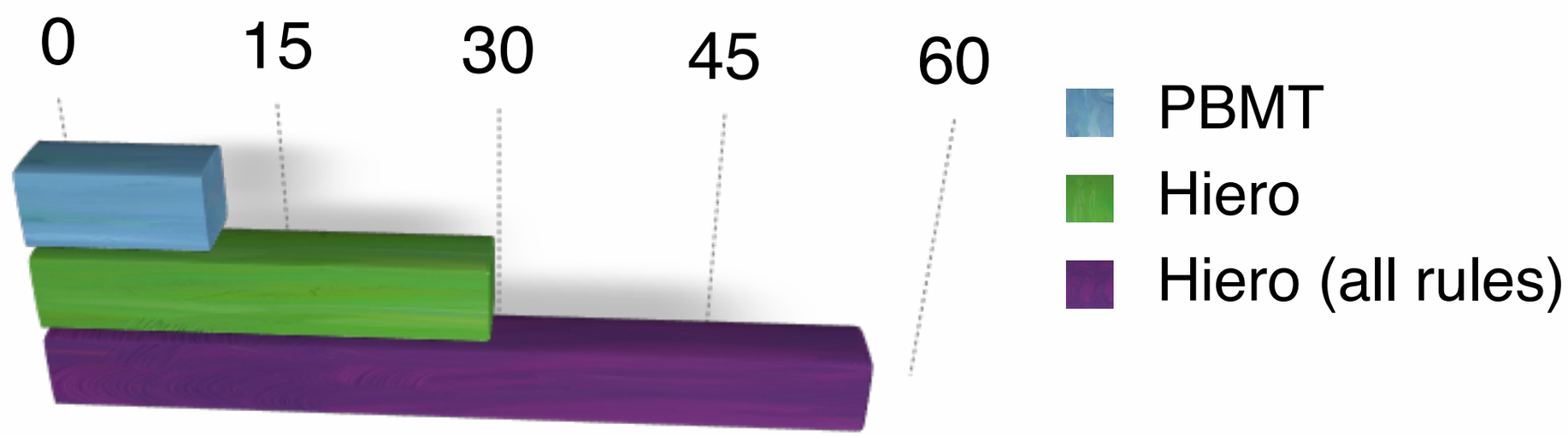
- Phrase-based translation suffers from distortion limit
  - can only use a **small** portion of bitext (low forced decoding reachability)
  - translation quality is often slightly worse than hierarchical models



reachability % (in # of words) on FBIS

From PBMT to Hiero

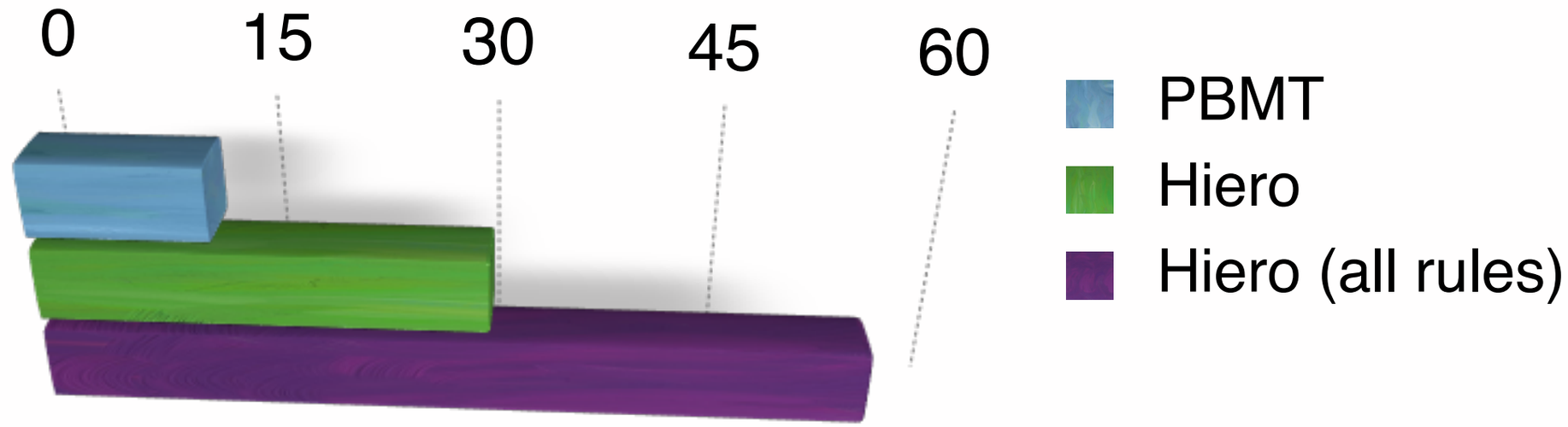
- Phrase-based translation suffers from distortion limit
  - can only use a **small** portion of bitext (low forced decoding reachability)
  - translation quality is often slightly worse than hierarchical models
- Hiero handles reordering better
  - potentially more sentences to train
  - learn more sparse features



reachability % (in # of words) on FBIS

From PBMT to Hiero

- Phrase-based translation suffers from distortion limit
  - can only use a **small** portion of bitext (low forced decoding reachability)
  - translation quality is often slightly worse than hierarchical models
- Hiero handles reordering better
  - potentially more sentences to train
  - learn more sparse features
- Challenge: generalize max-violation latent perceptron to hypergraphs?

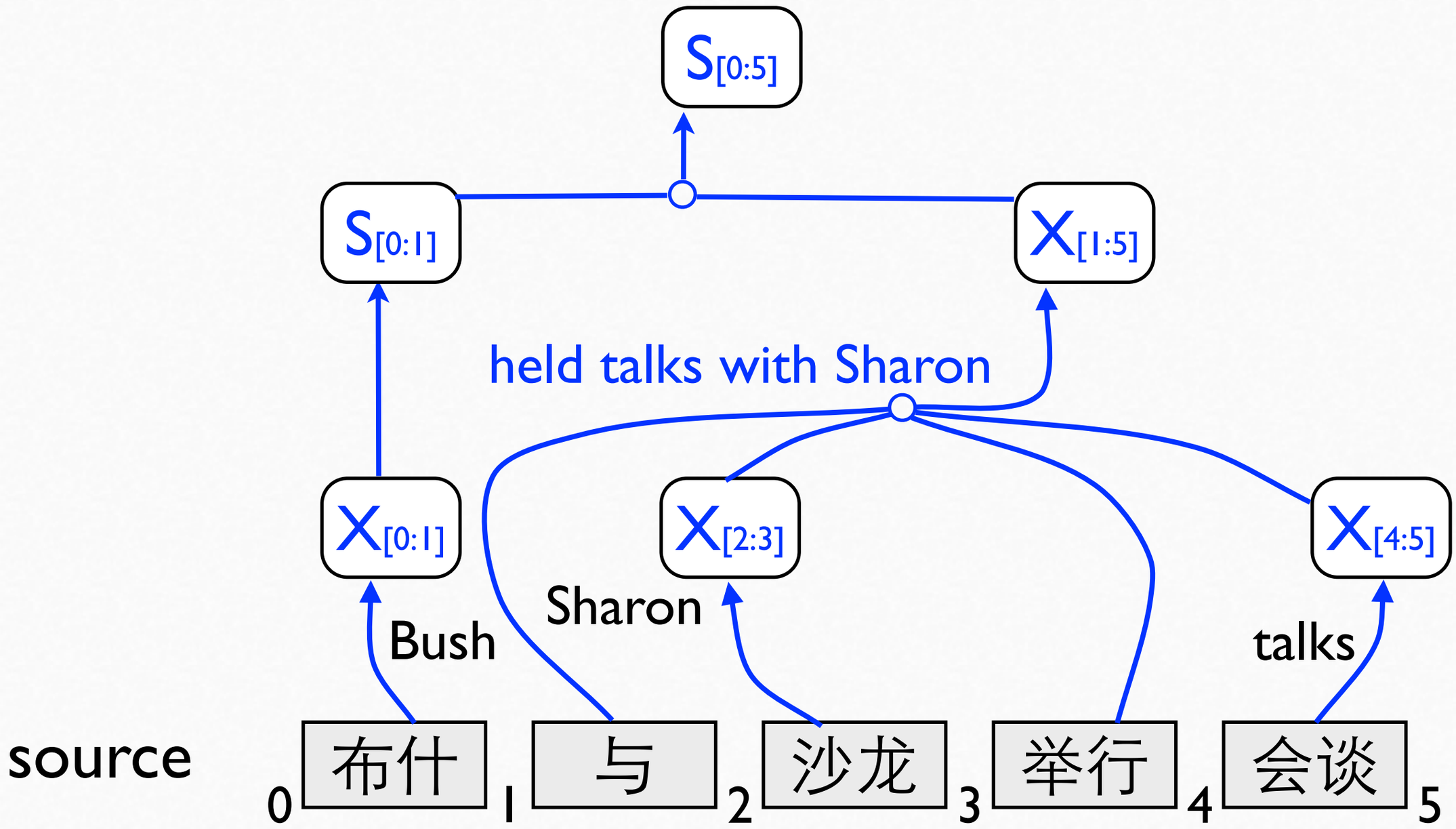


reachability % (in # of words) on FBIS

A Hiero Example

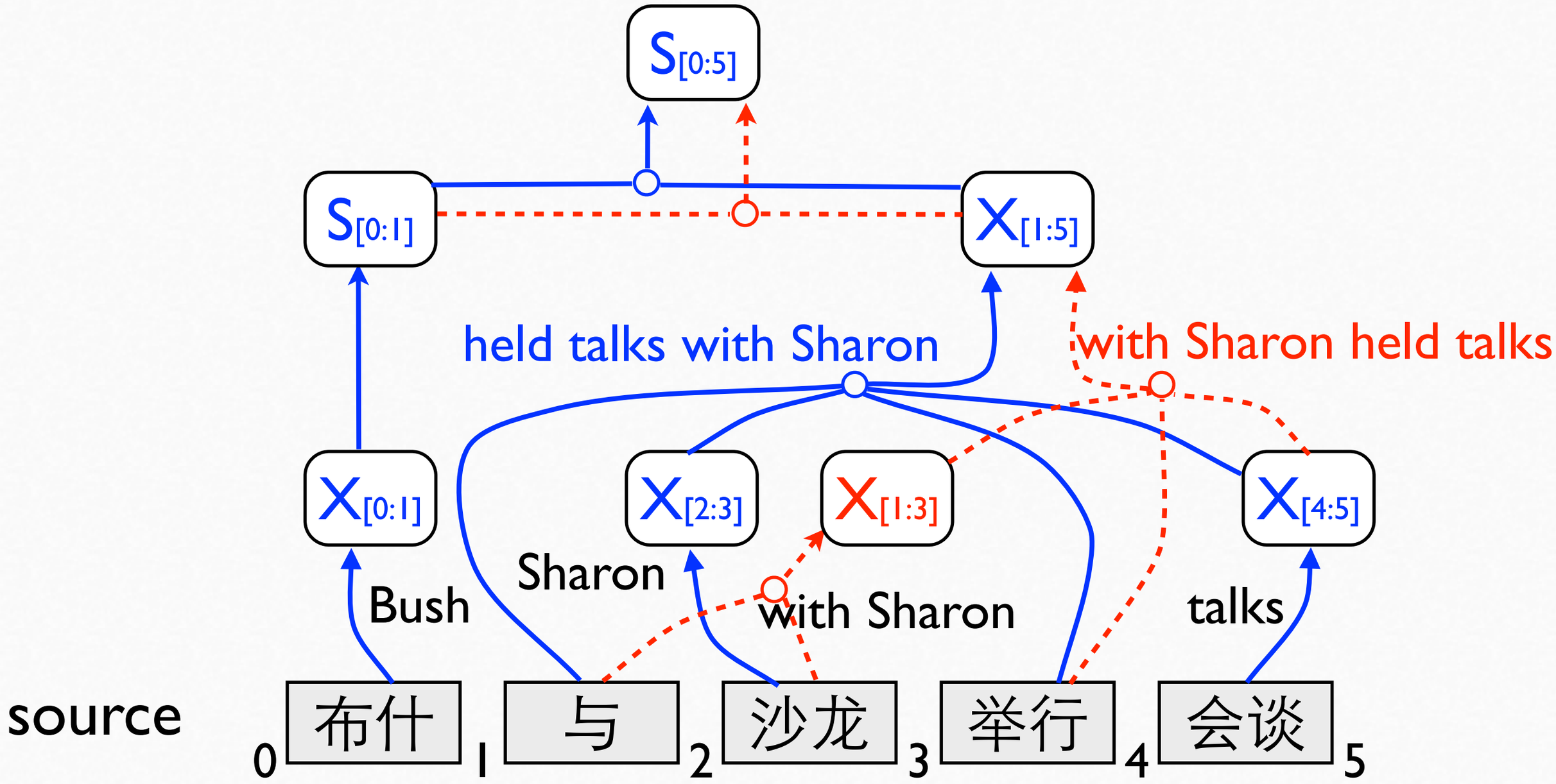
source

A Hiero Example



Reference Bush held talks with Sharon

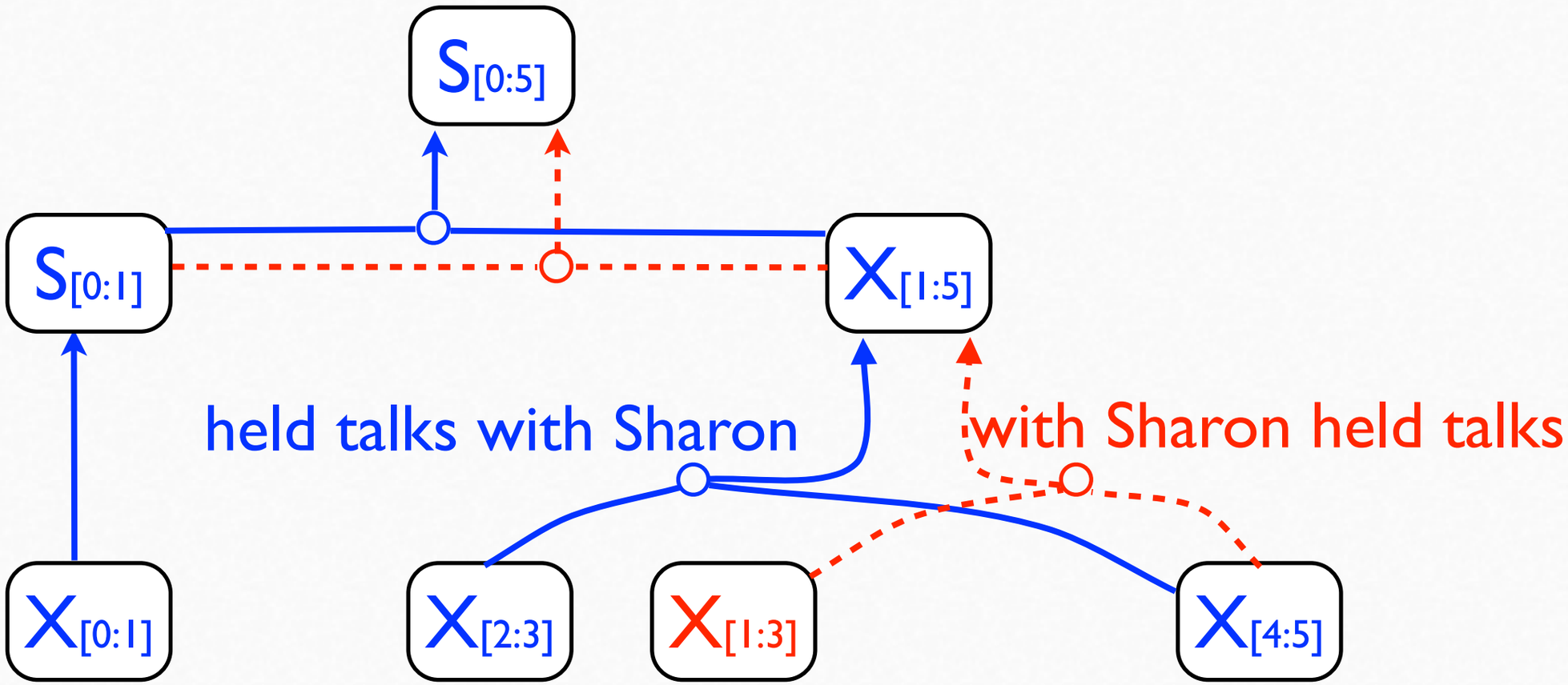
A Hiero Example



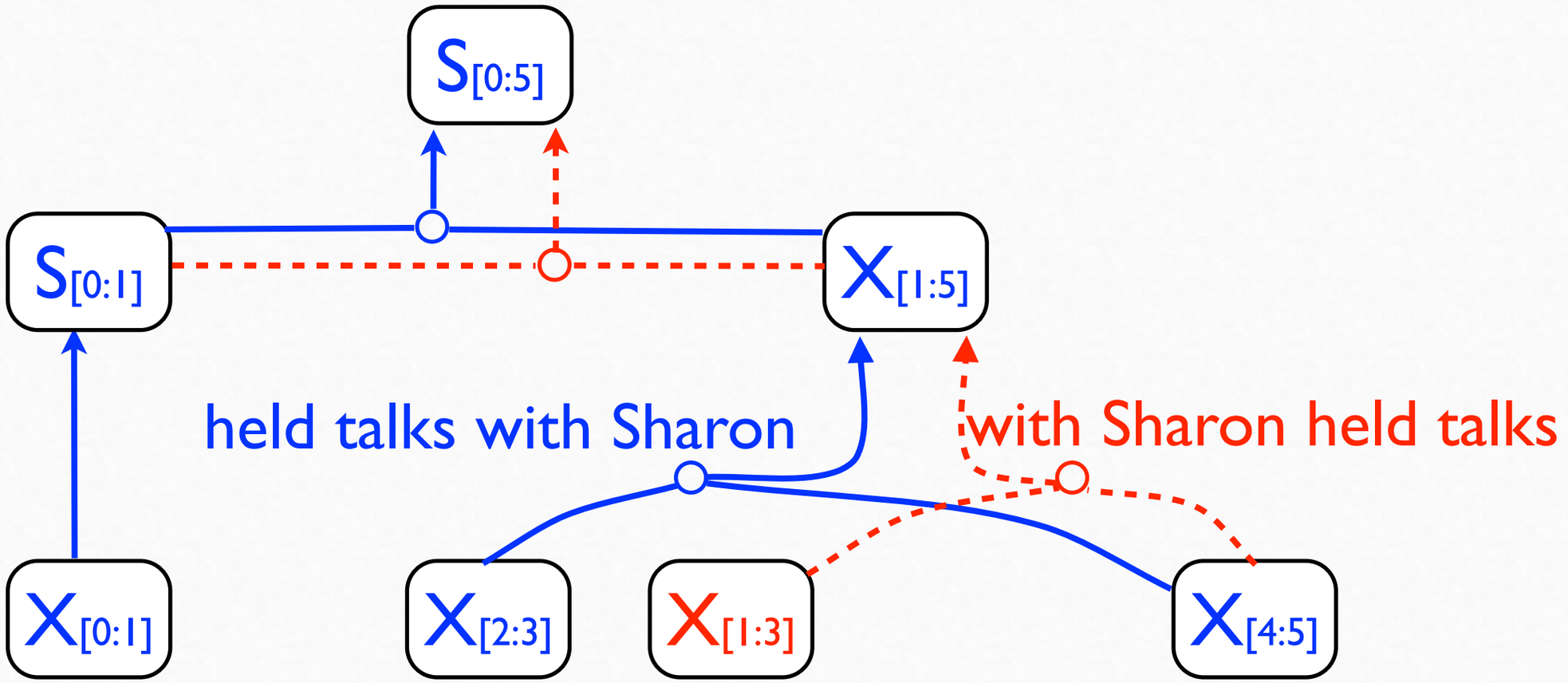
Reference Bush held talks with Sharon

Viterbi Bush with Sharon held talks

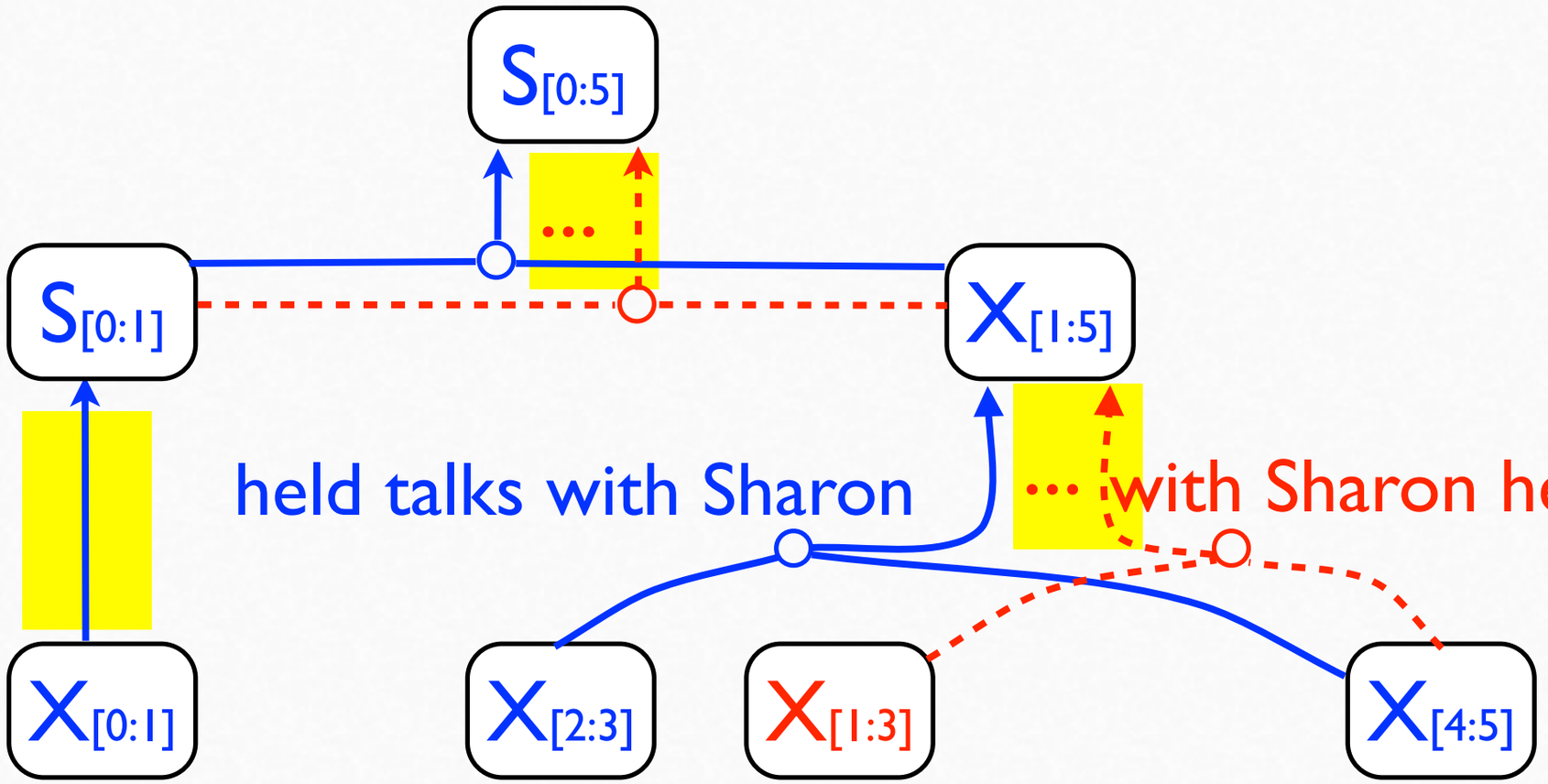
A Hiero Example



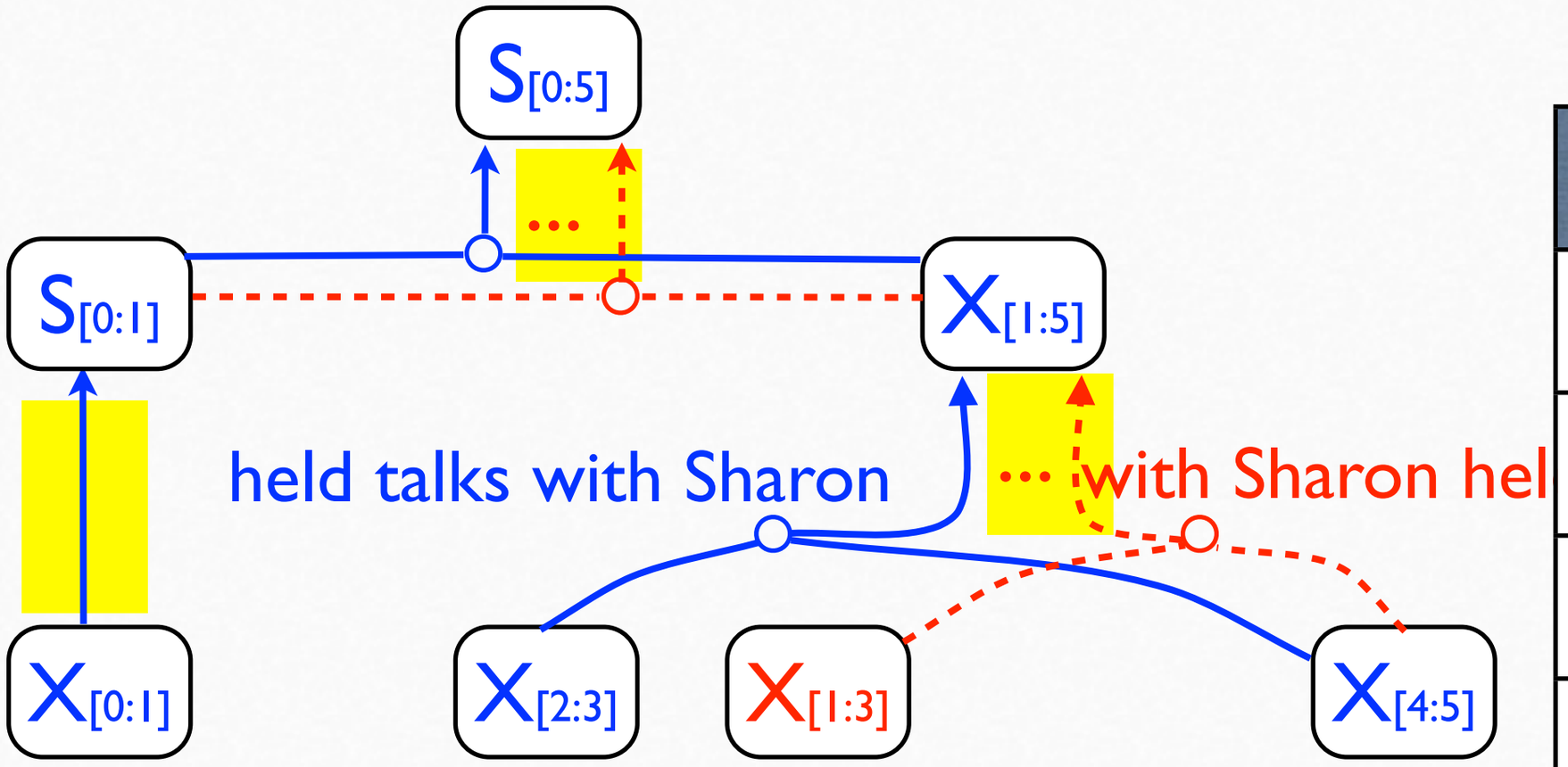
Search Error: Reference Got Pruned



Search Error: Reference Got Pruned



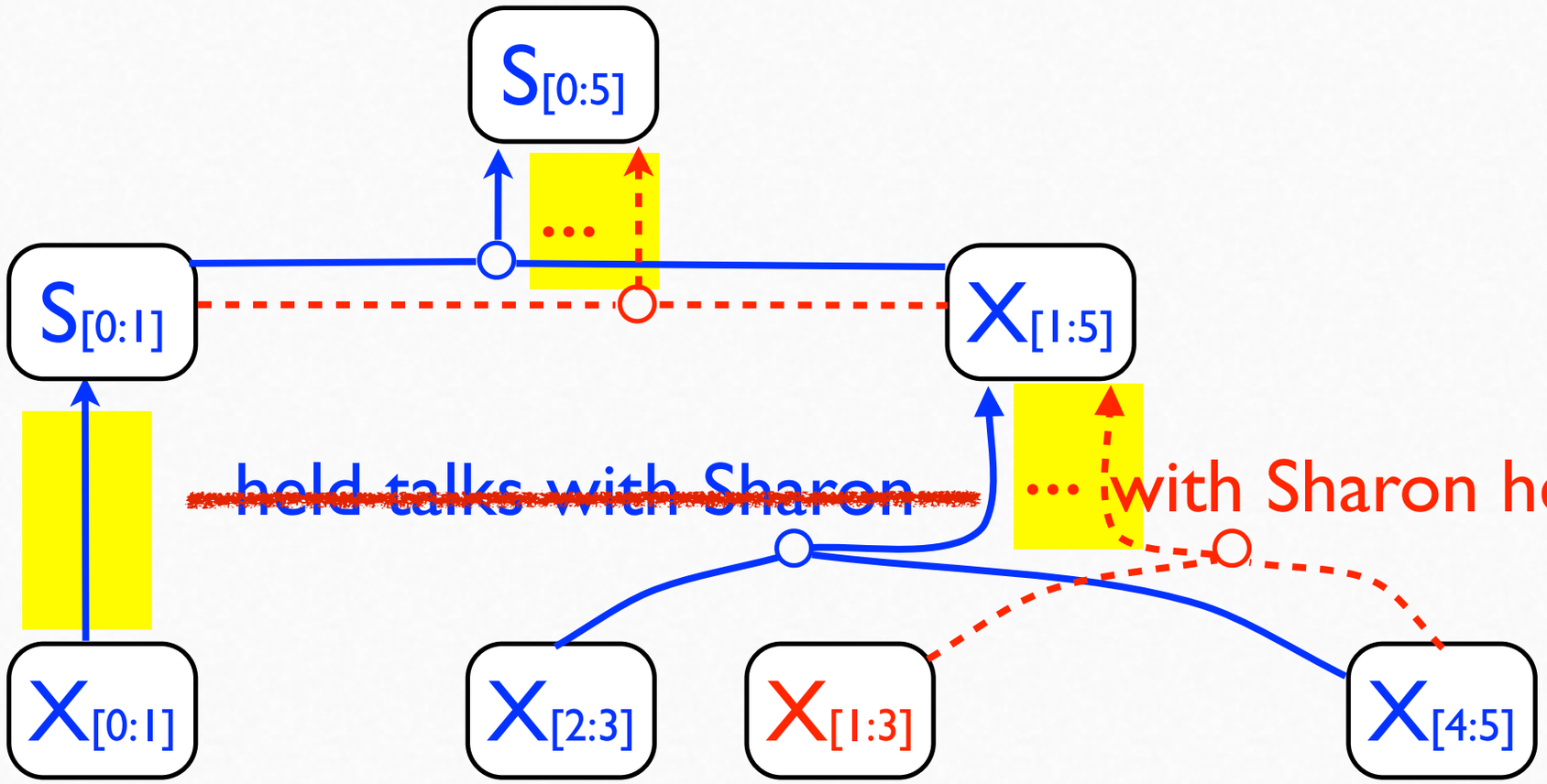
Search Error: Reference Got Pruned



beam at $X_{[1:5]}$

candidate	score
X_1 held X_2	0.88
X_1 holds X_2	0.86
...	...
held X_2 with X_1	0.25
...	...

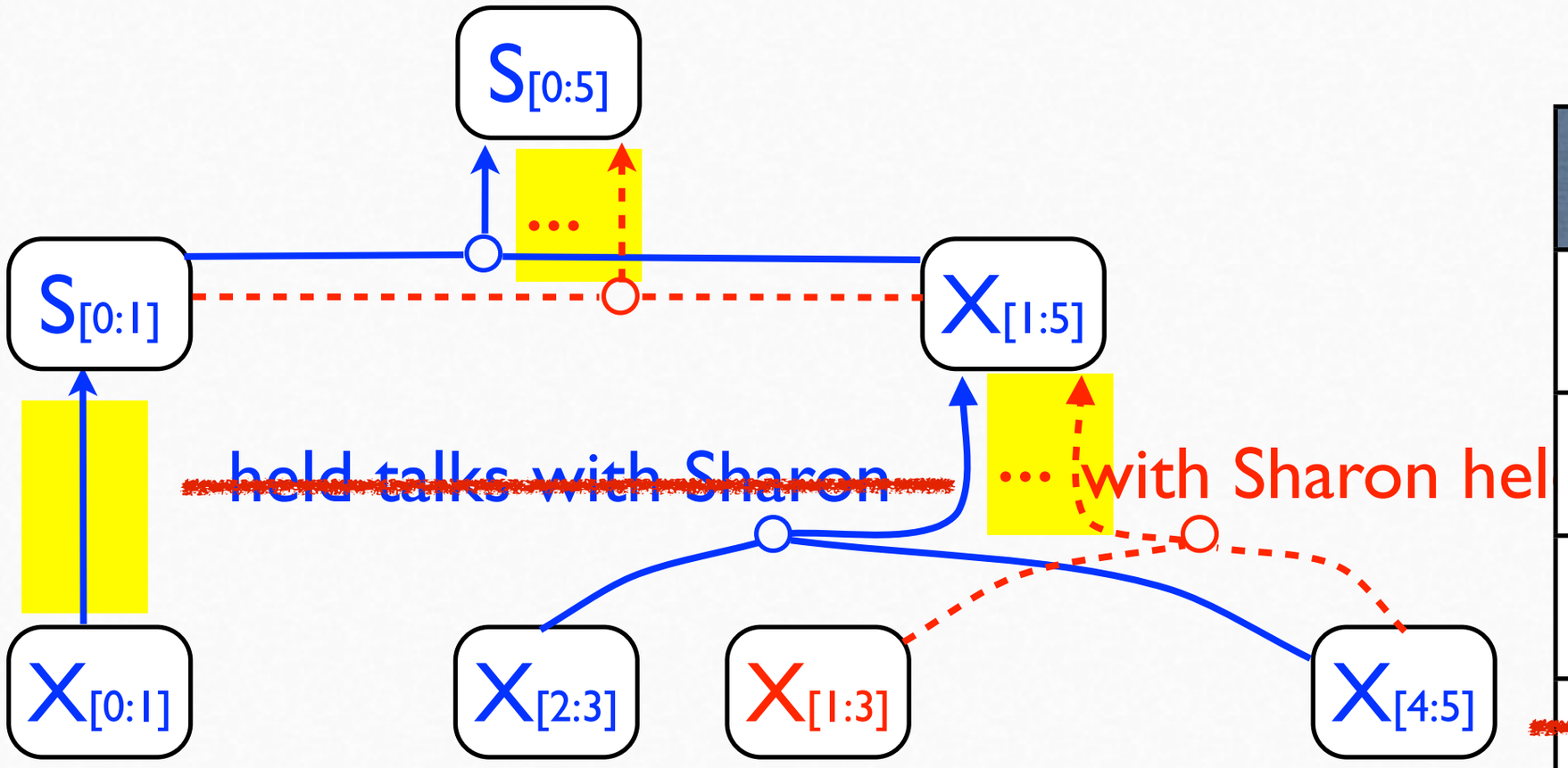
Search Error: Reference Got Pruned



beam at $X_{[1:5]}$

candidate	score
X_1 held X_2	0.88
X_1 holds X_2	0.86
...	...
held X_2 with X_1	0.25
...	...

Search Error: Reference Got Pruned

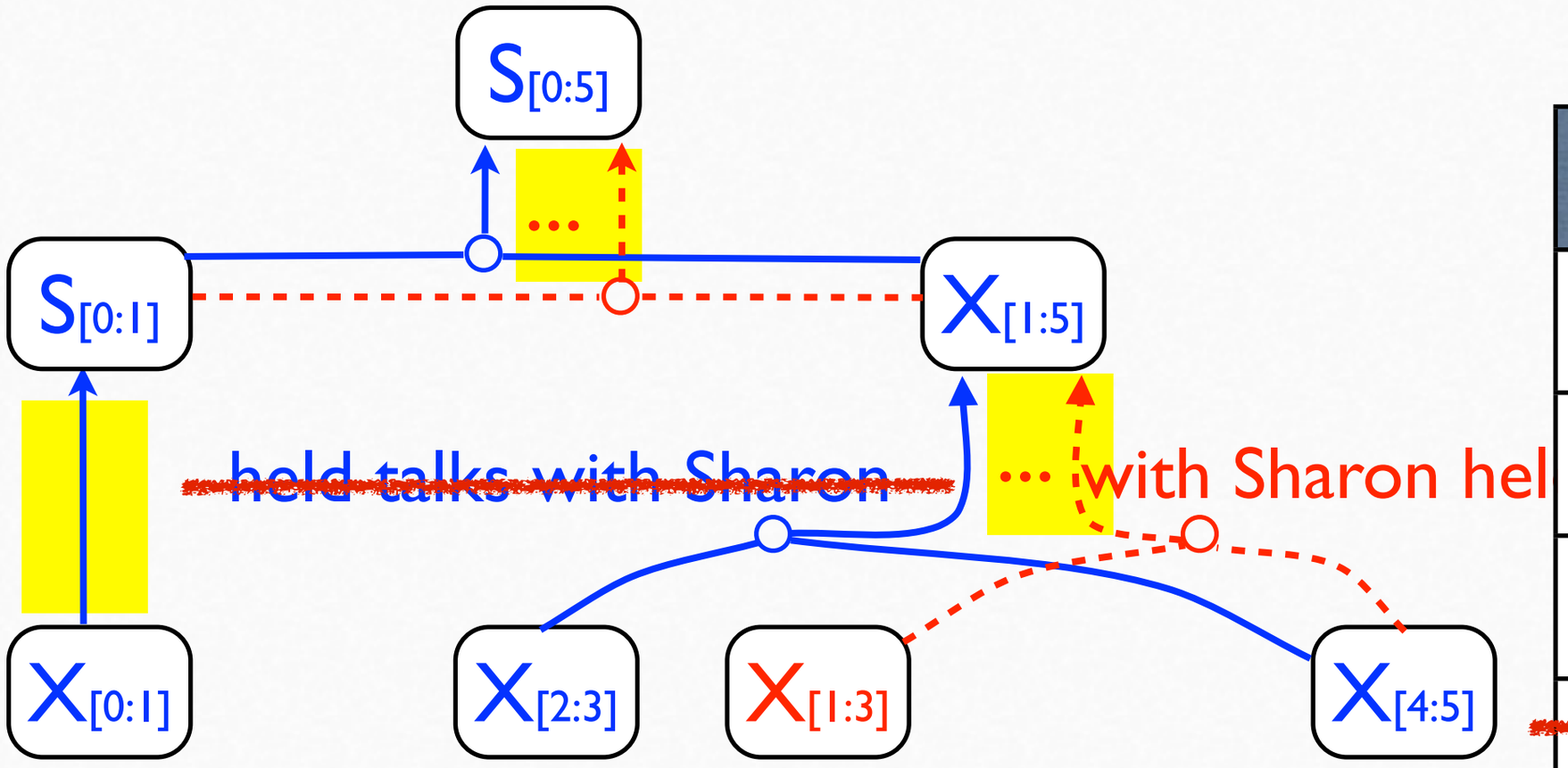


beam at $X_{[1:5]}$

candidate	score
X_1 held X_2	0.88
X_1 holds X_2	0.86
...	...
held X_2 with X_1	0.25
...	...

- Reference derivation rank too low in the beam and gets pruned
- We call it a **Violation** iff. score(viterbi) - score(reference) > 0

Search Error: Reference Got Pruned



beam at $X_{[1:5]}$

candidate	score
X_1 held X_2	0.88
X_1 holds X_2	0.86
...	...
held X_2 with X_1	0.25
...	...

- Reference derivation rank too low in the beam and gets pruned
- We call it a **Violation** iff. score(viterbi) - score(reference) > 0

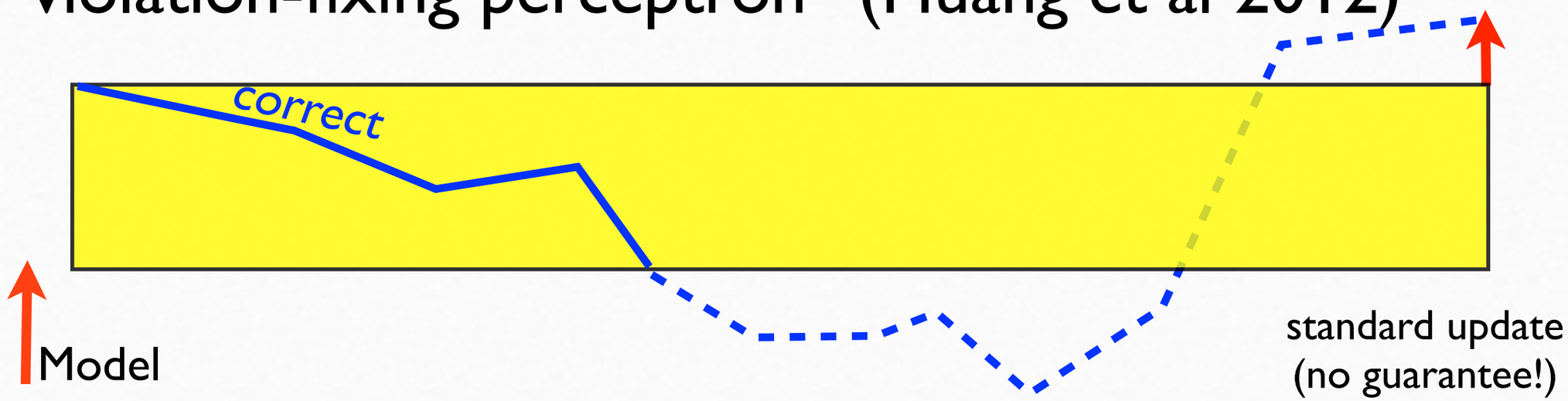
= amount of violation

Fixing Search Error: Max-Violation

- standard perceptron does not guarantee violation
  - w/ pruning, the correct derivation might score higher at the ends
  - called "invalid" update b/c it doesn't fix the search error
- max-violation: update at where the violation is maximum
  - "worst-mistake" in the search
  - learns more and faster
- "violation-fixing perceptron" (Huang et al 2012)

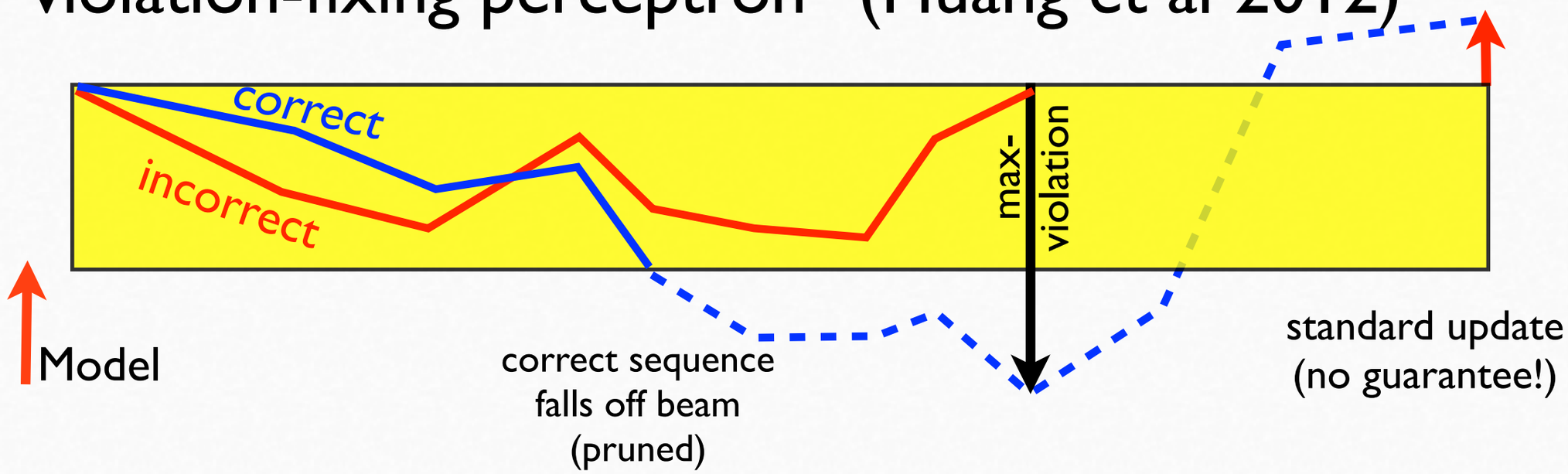
Fixing Search Error: Max-Violation

- standard perceptron does not guarantee violation
  - w/ pruning, the correct derivation might score higher at the ends
  - called "invalid" update b/c it doesn't fix the search error
- max-violation: update at where the violation is maximum
  - "worst-mistake" in the search
  - learns more and faster
- "violation-fixing perceptron" (Huang et al 2012)

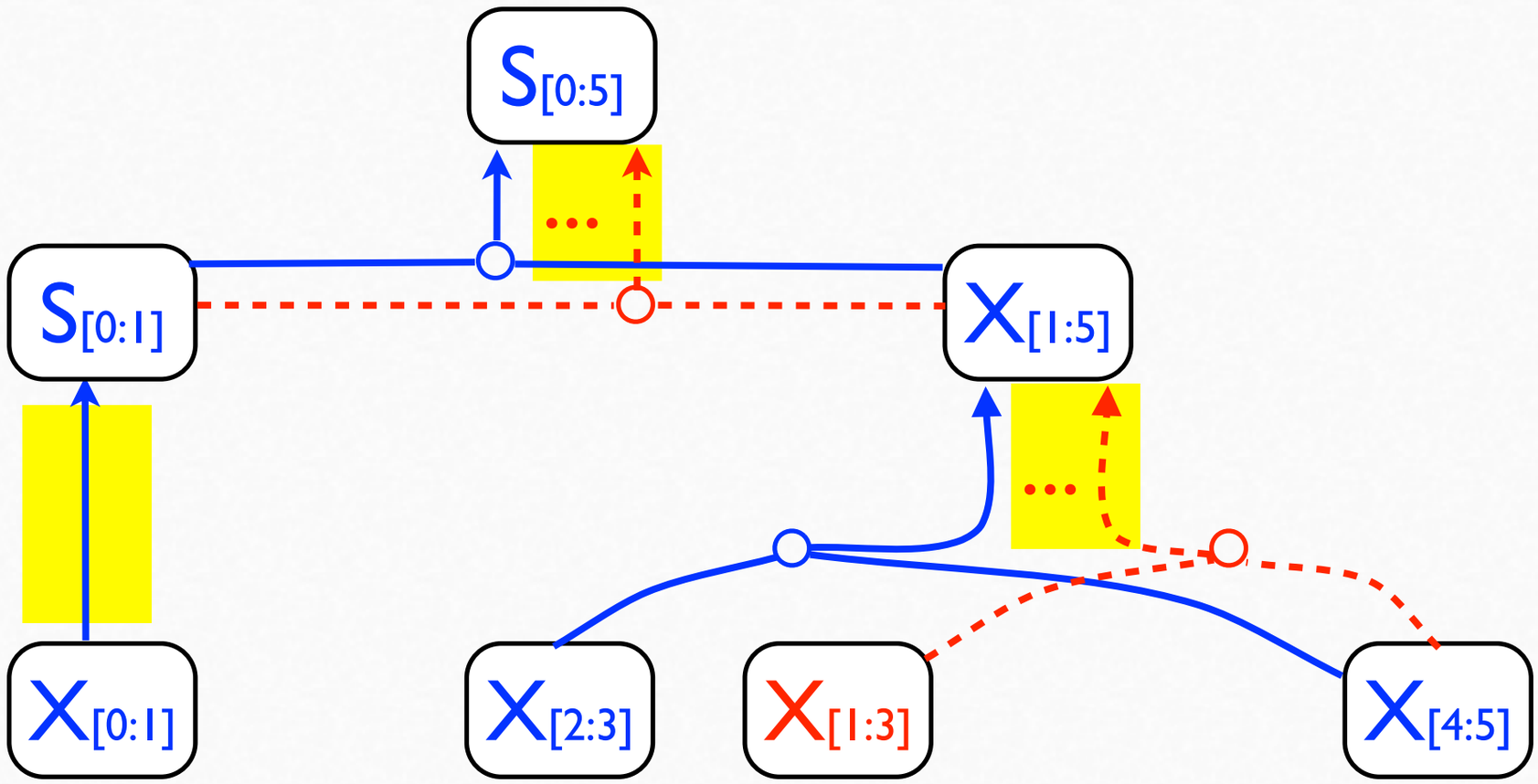


Fixing Search Error: Max-Violation

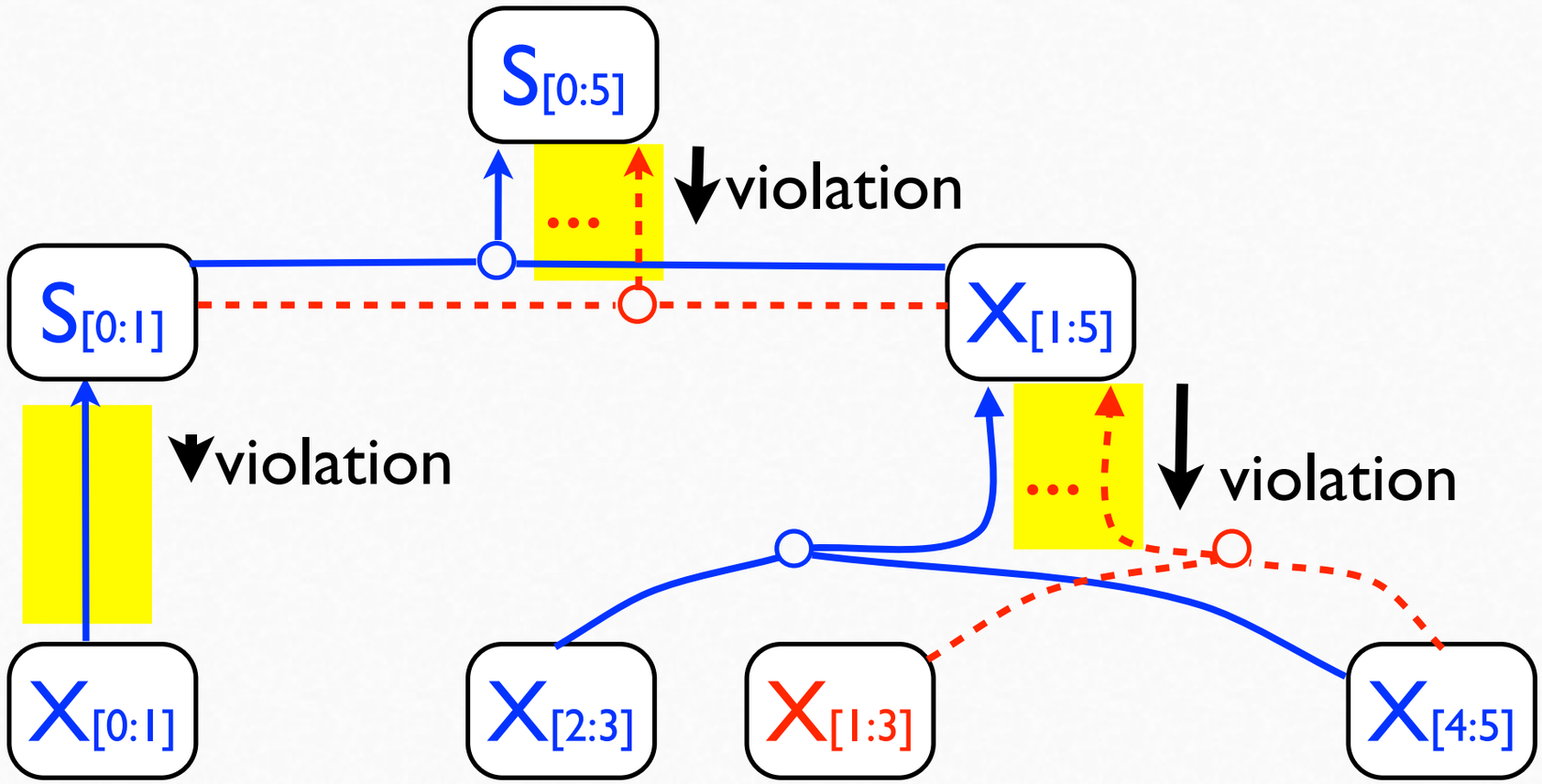
- standard perceptron does not guarantee violation
  - w/ pruning, the correct derivation might score higher at the ends
  - called "invalid" update b/c it doesn't fix the search error
- max-violation: update at where the violation is maximum
  - "worst-mistake" in the search
  - learns more and faster
- "violation-fixing perceptron" (Huang et al 2012)



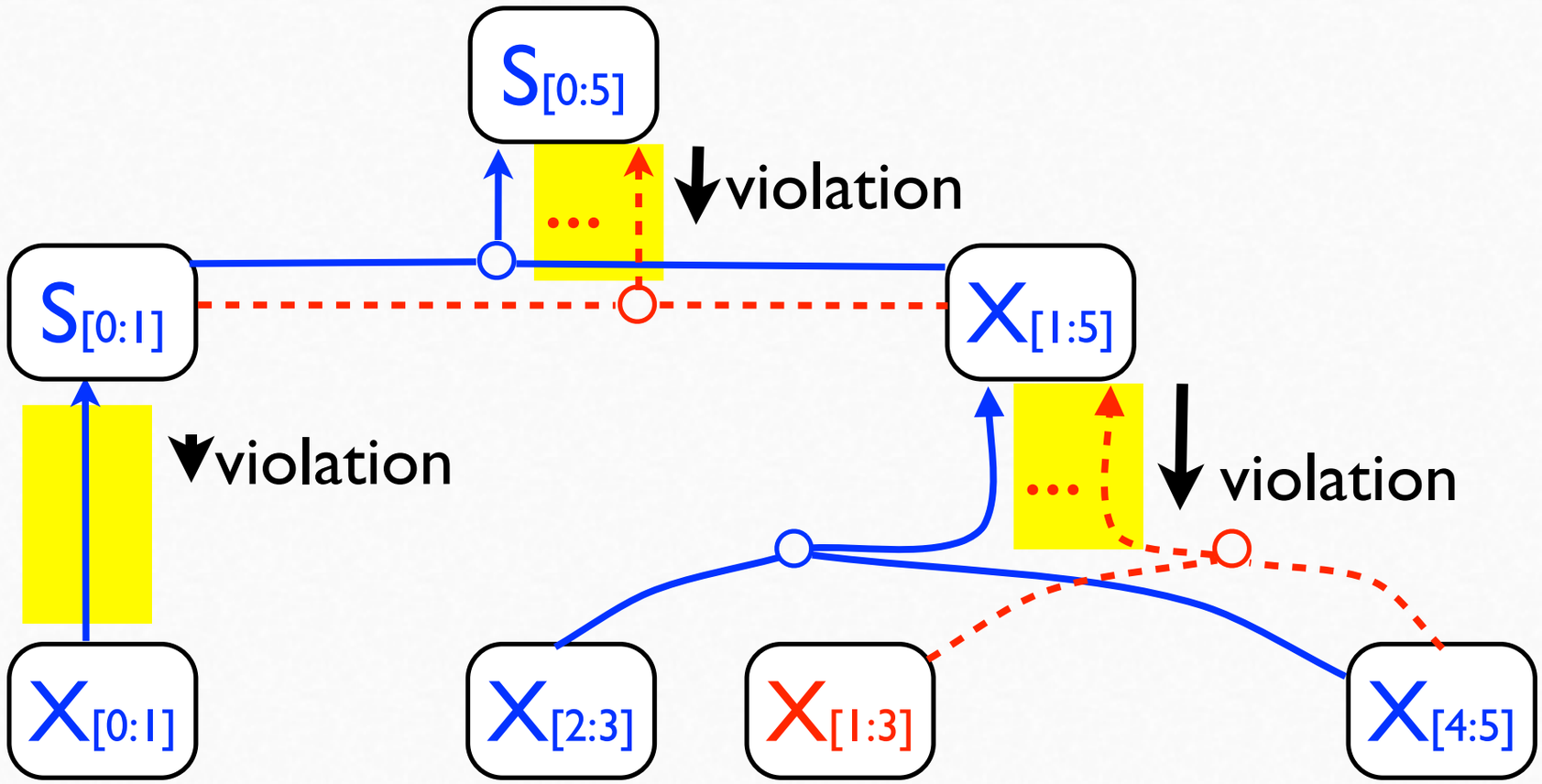
Max-Violation over Hypergraphs



Max-Violation over Hypergraphs

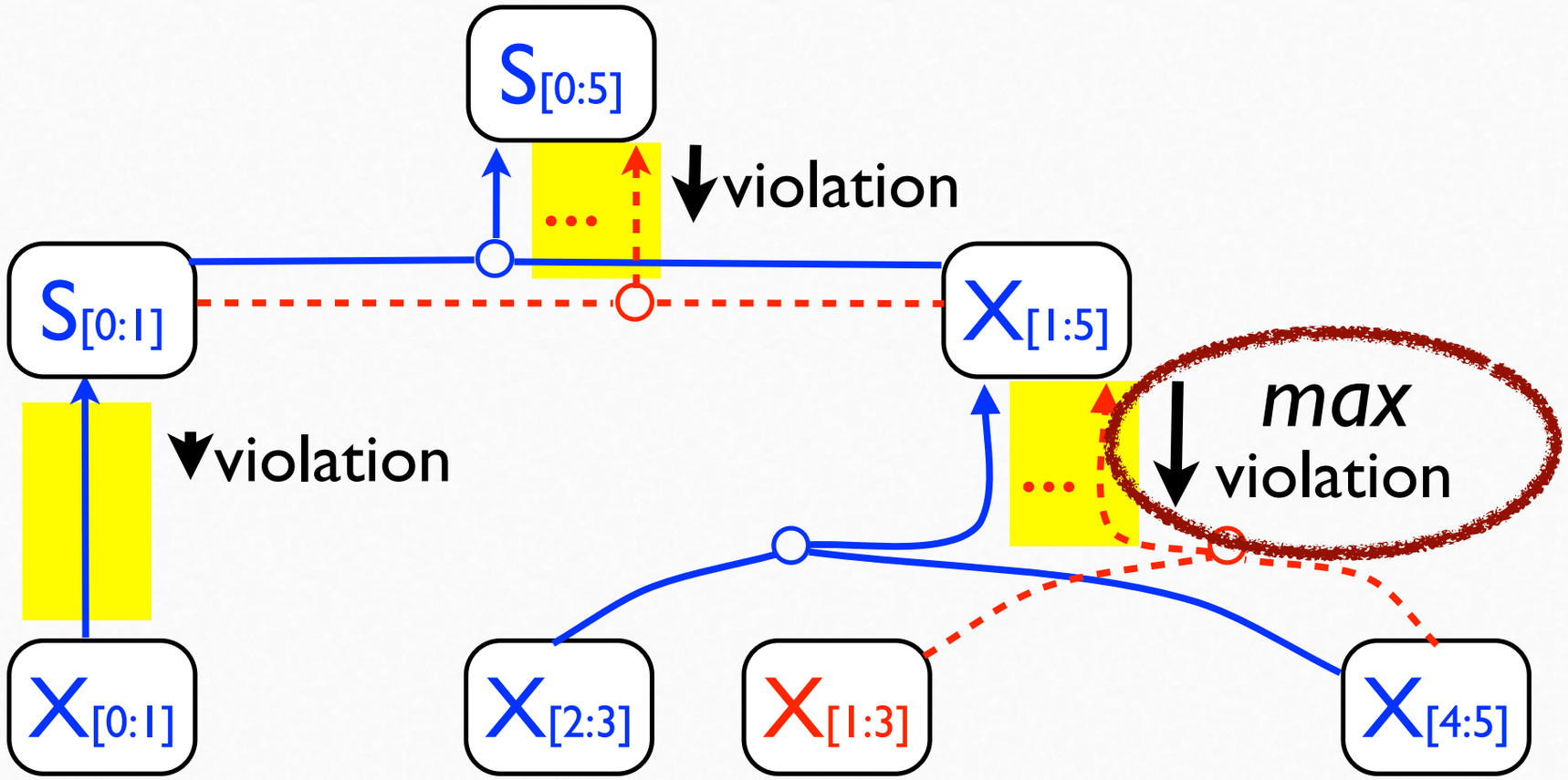


Max-Violation over Hypergraphs



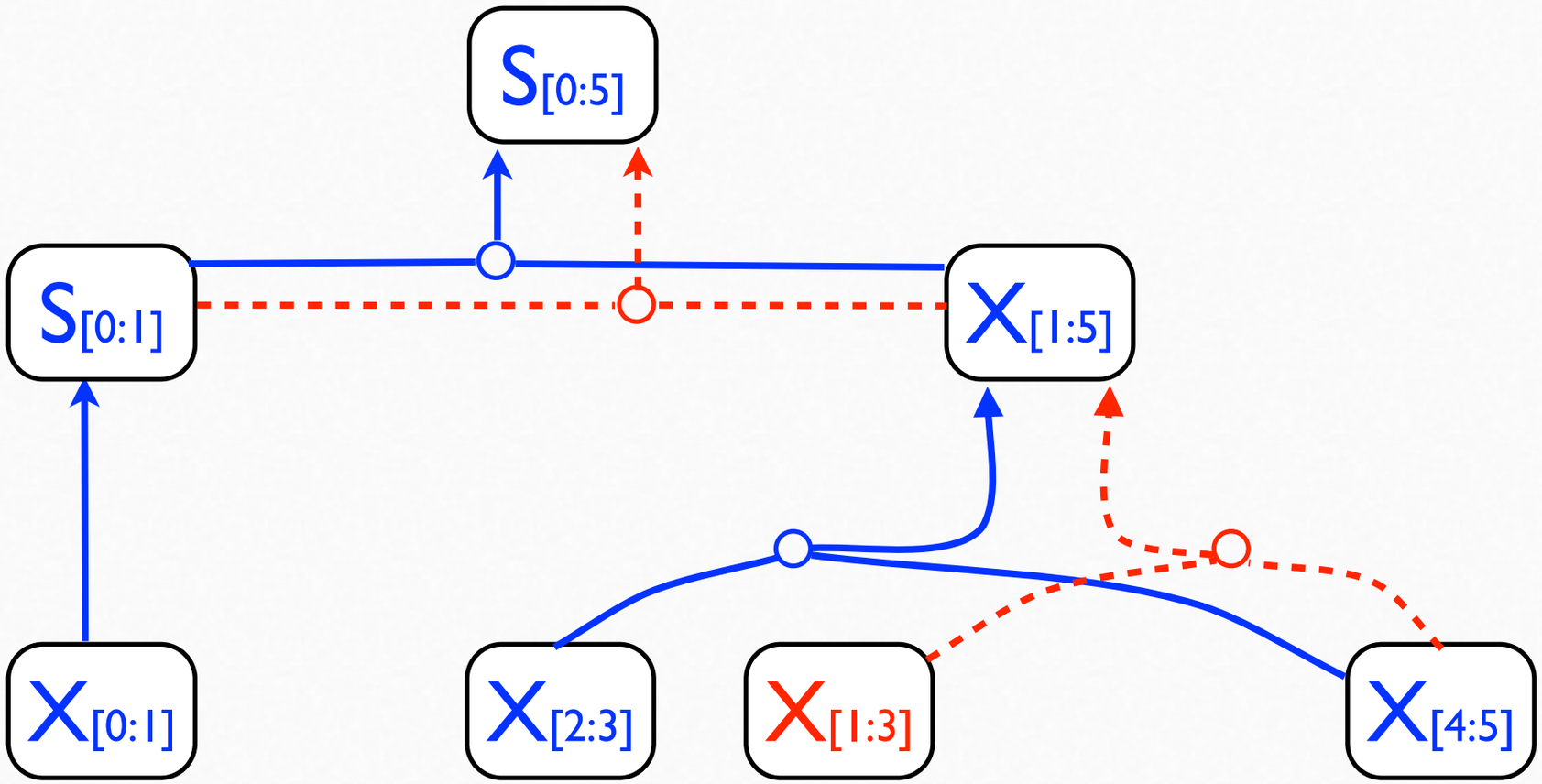
- Update at where the violation b/w the Viterbi and the reference is **maximum** (Zhang et al. '13)

Max-Violation over Hypergraphs

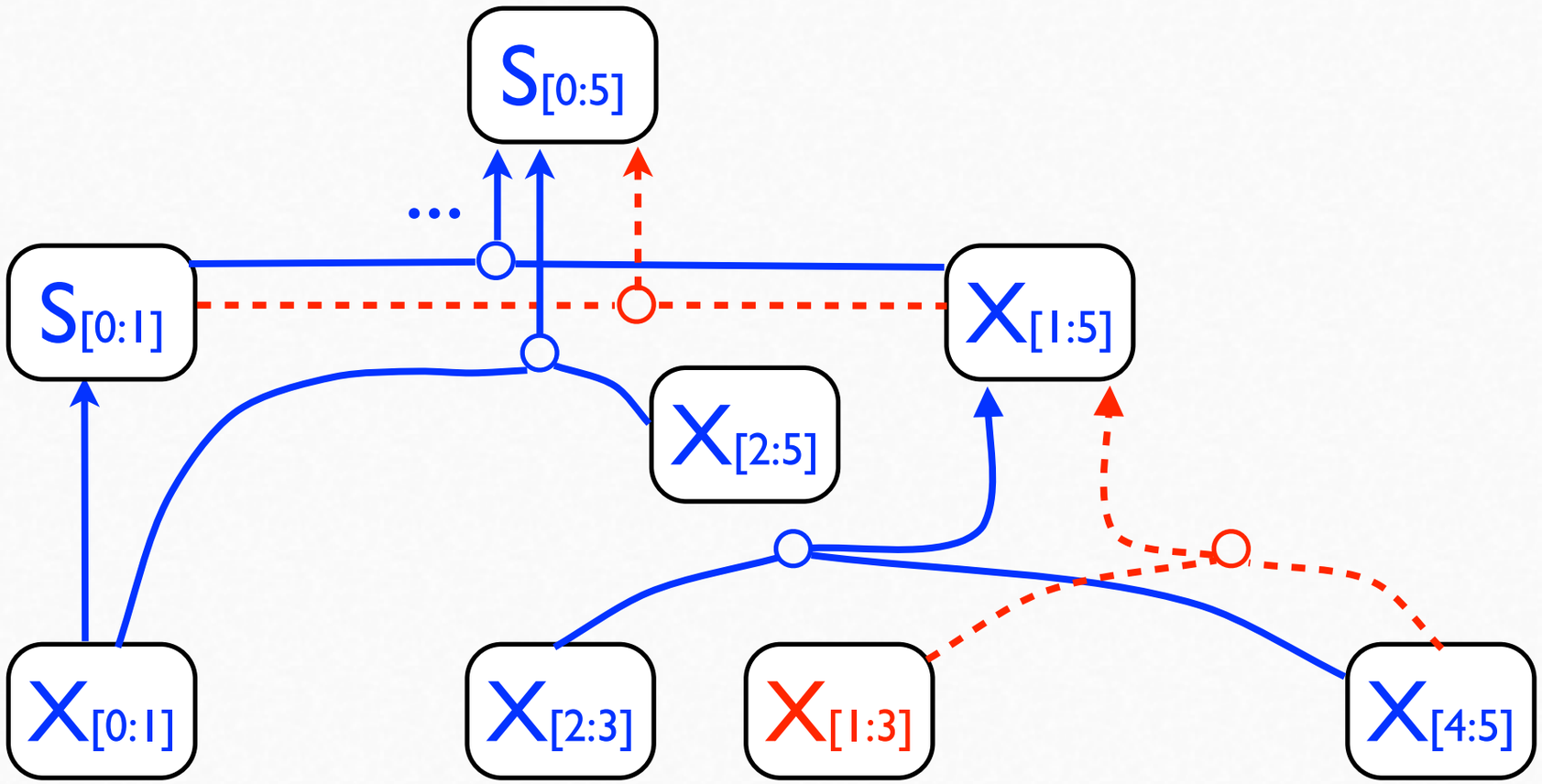


- Update at where the violation b/w the Viterbi and the reference is **maximum** (Zhang et al. '13)

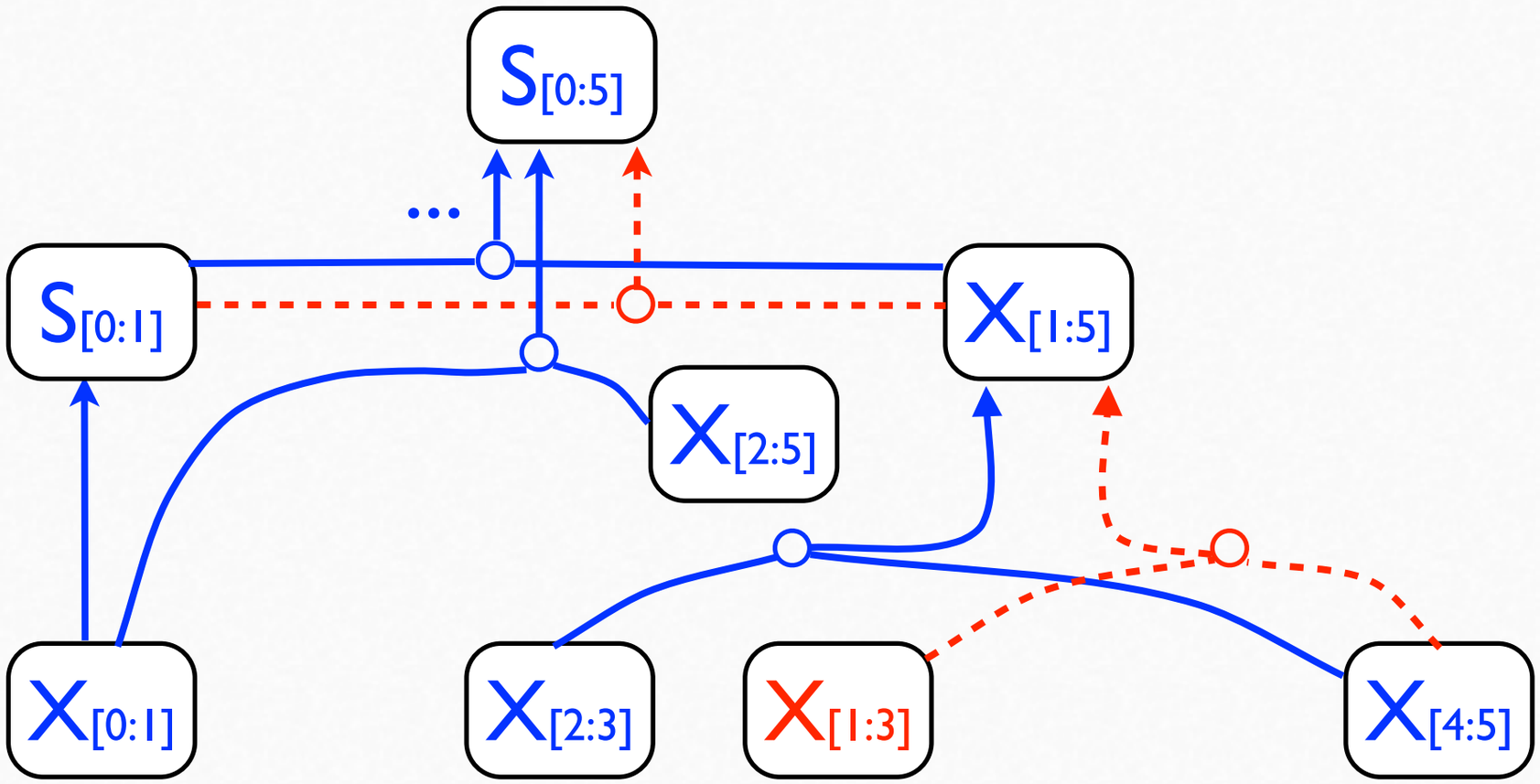
Max-Violation over Hypergraphs w/ Latent Var.



Max-Violation over Hypergraphs w/ Latent Var.

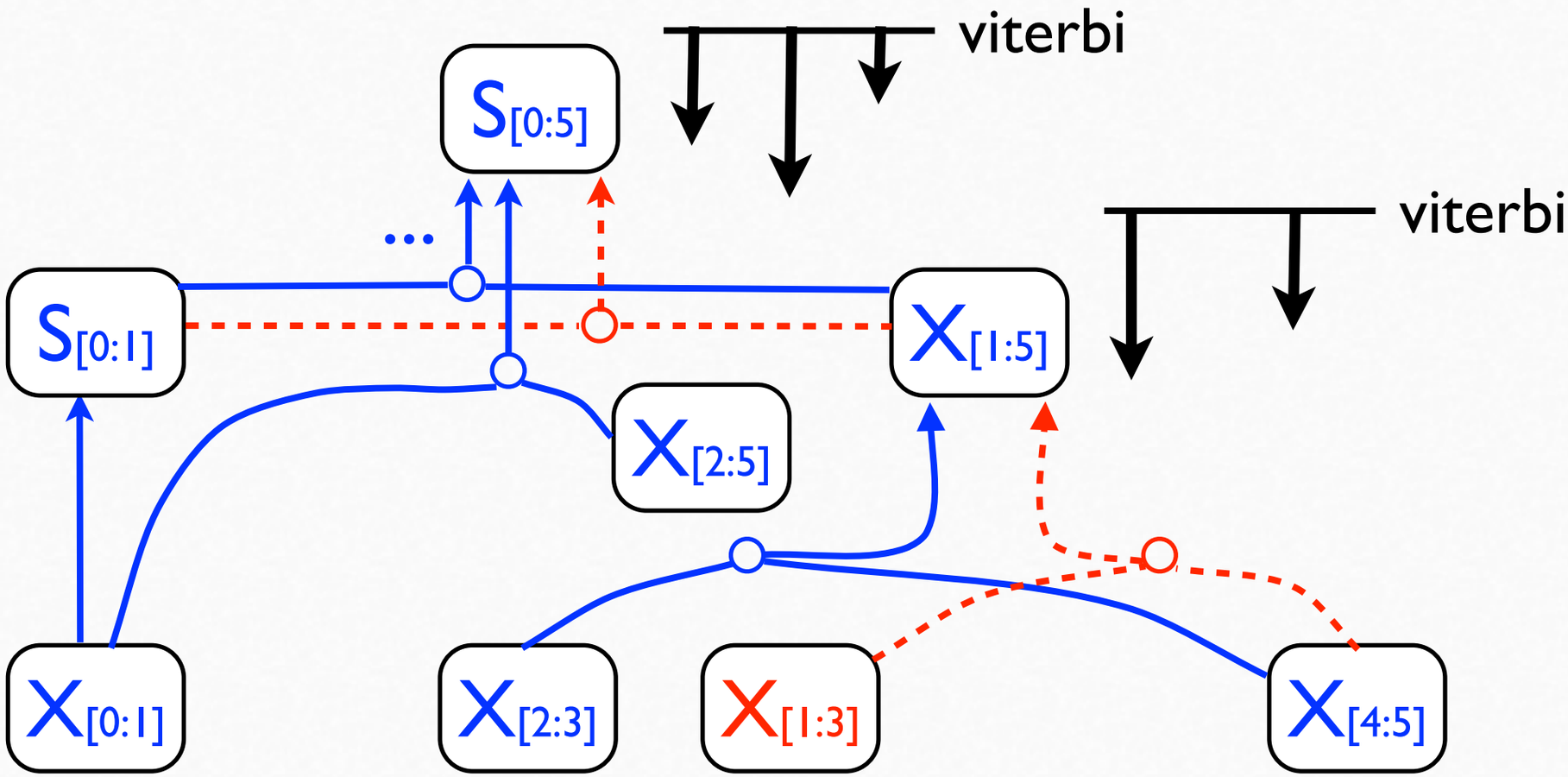


Max-Violation over Hypergraphs w/ Latent Var.



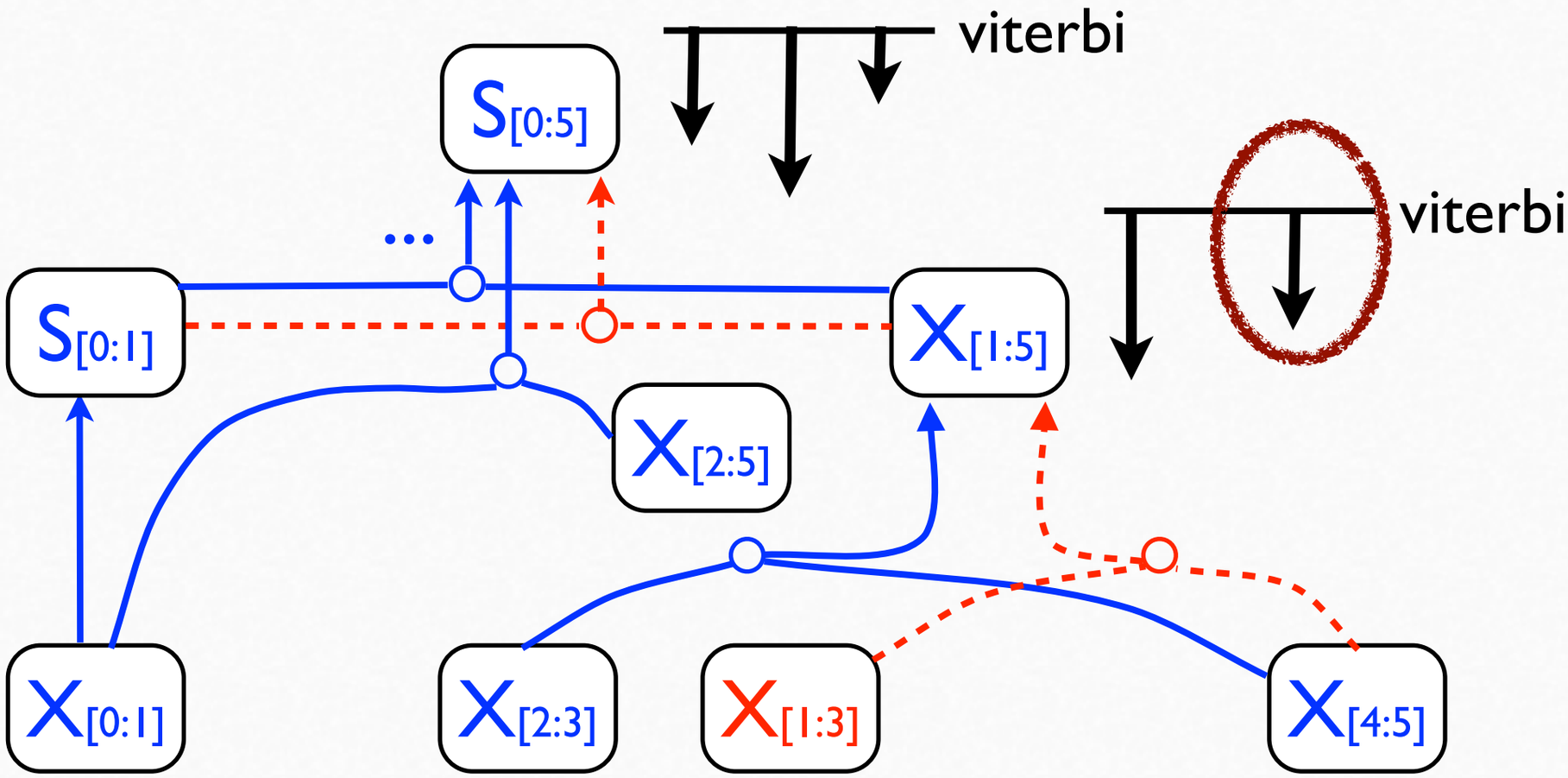
- multiple reference derivations at one node
  - at one node, pick the best reference
  - globally choose max-violated best reference

Max-Violation over Hypergraphs w/ Latent Var.



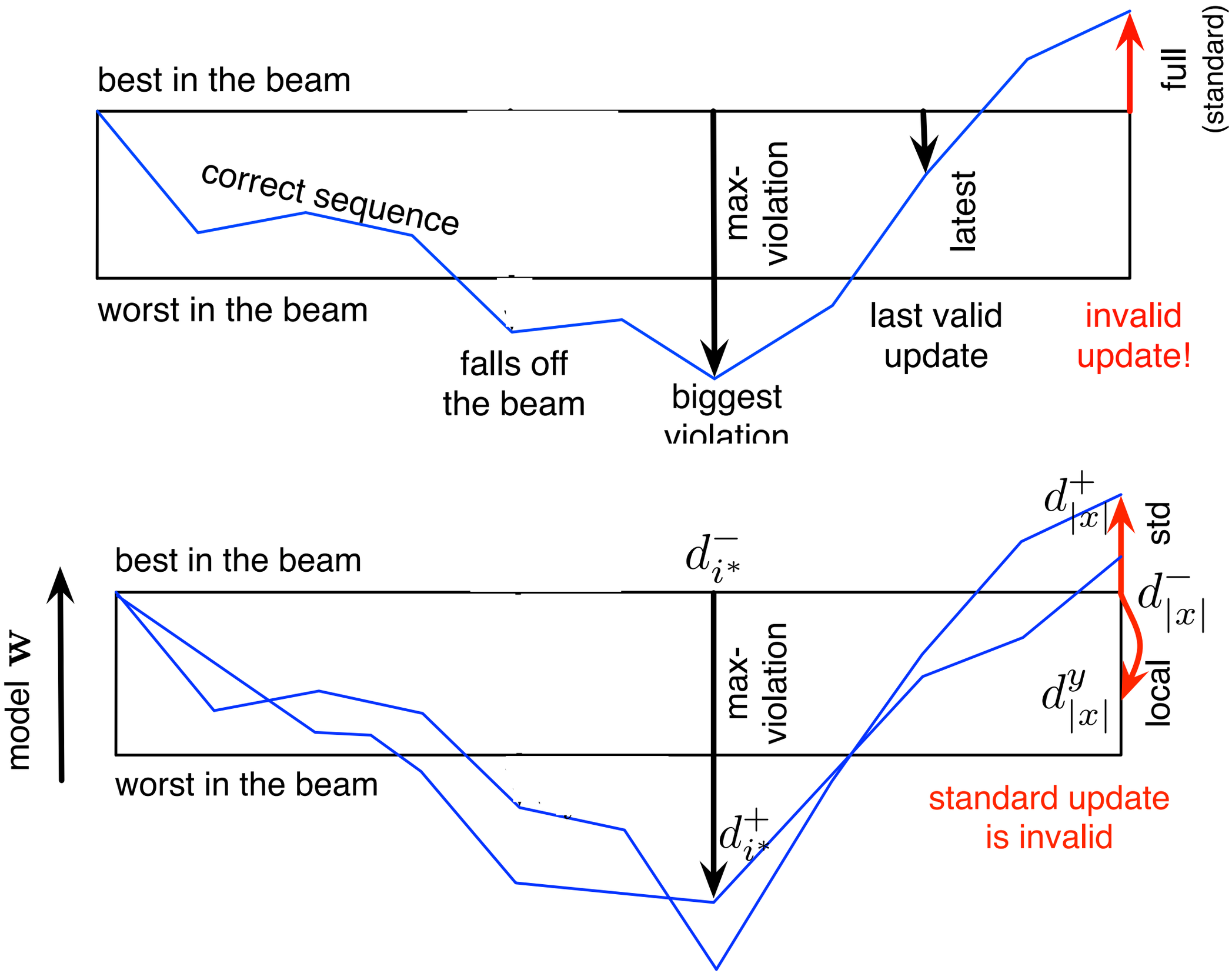
- multiple reference derivations at one node
  - at one node, pick the best reference
  - globally choose max-violated best reference

Max-Violation over Hypergraphs w/ Latent Var.



- multiple reference derivations at one node
  - at one node, pick the best reference
  - globally choose max-violated best reference

Latent-Variable Max-Violation Perceptron



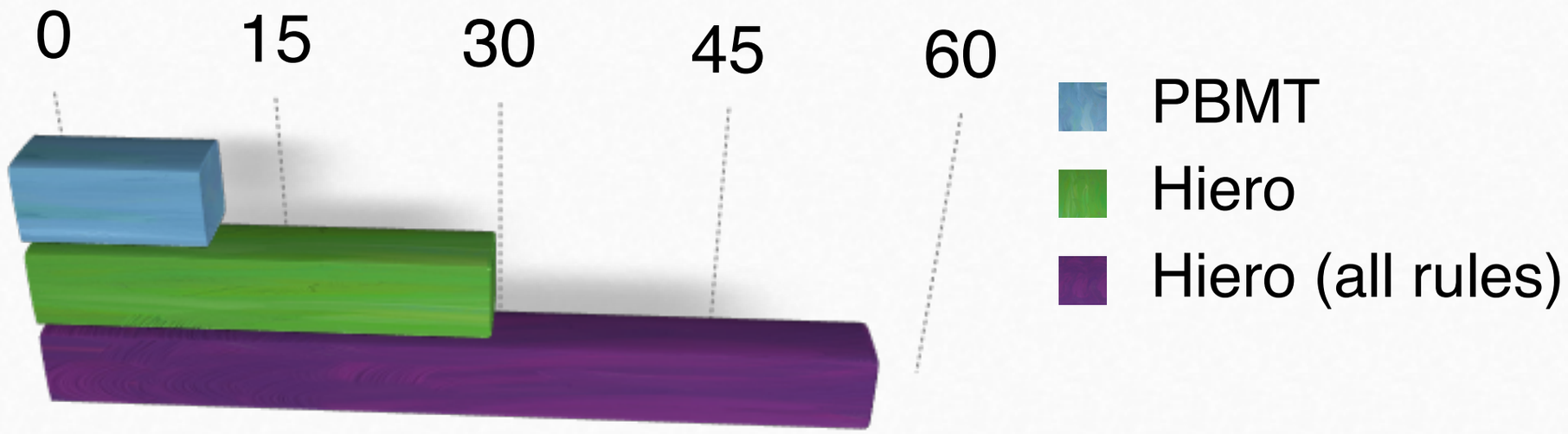
Experiments Setting

train	dev	test
IWSLT09 30k short sentences	IWSLT04 16 references	IWSLT05 16 references
FBIS 240k sentences	NIST06 4 references	NIST08 4 references

- 18 dense features from cdec
- budgeted sparse features based on Word-Edge features
  - atomic features: C/E boundary words; C boundary characters
  - complex features
    - combination of atomic features within limited budget
- we remove all 1-count rules in rule extraction
- trigram LM trained from target side

Experiments

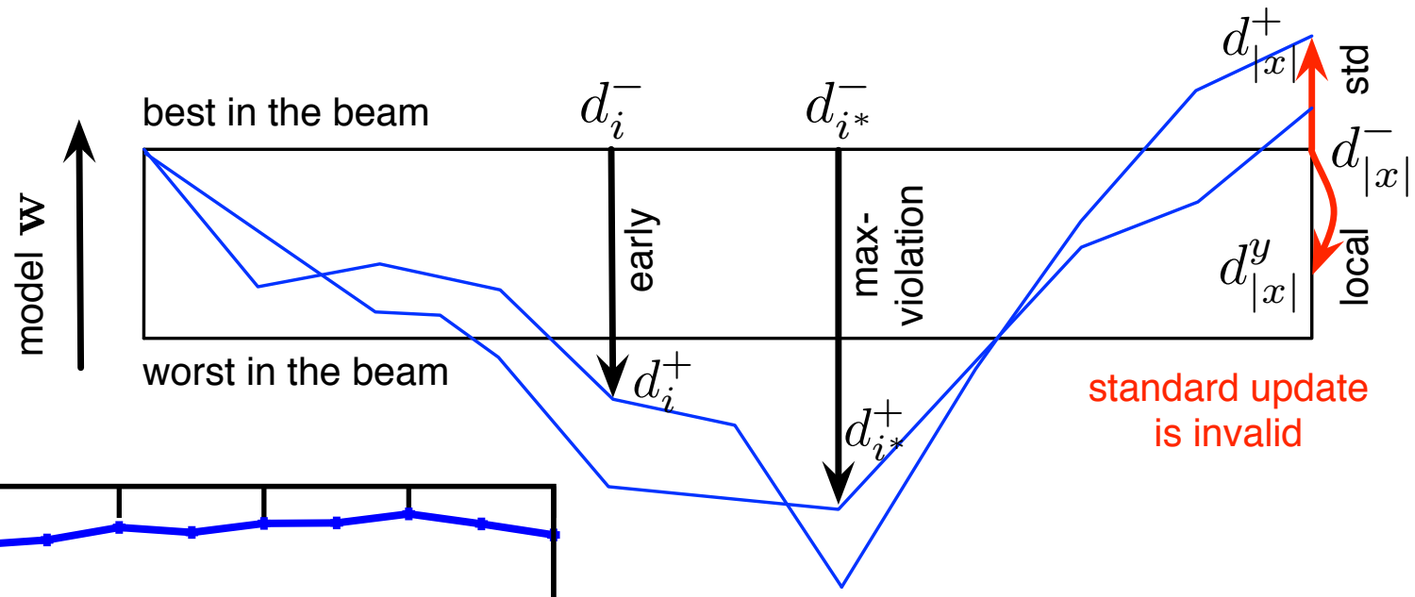
- Reachability on FBIS

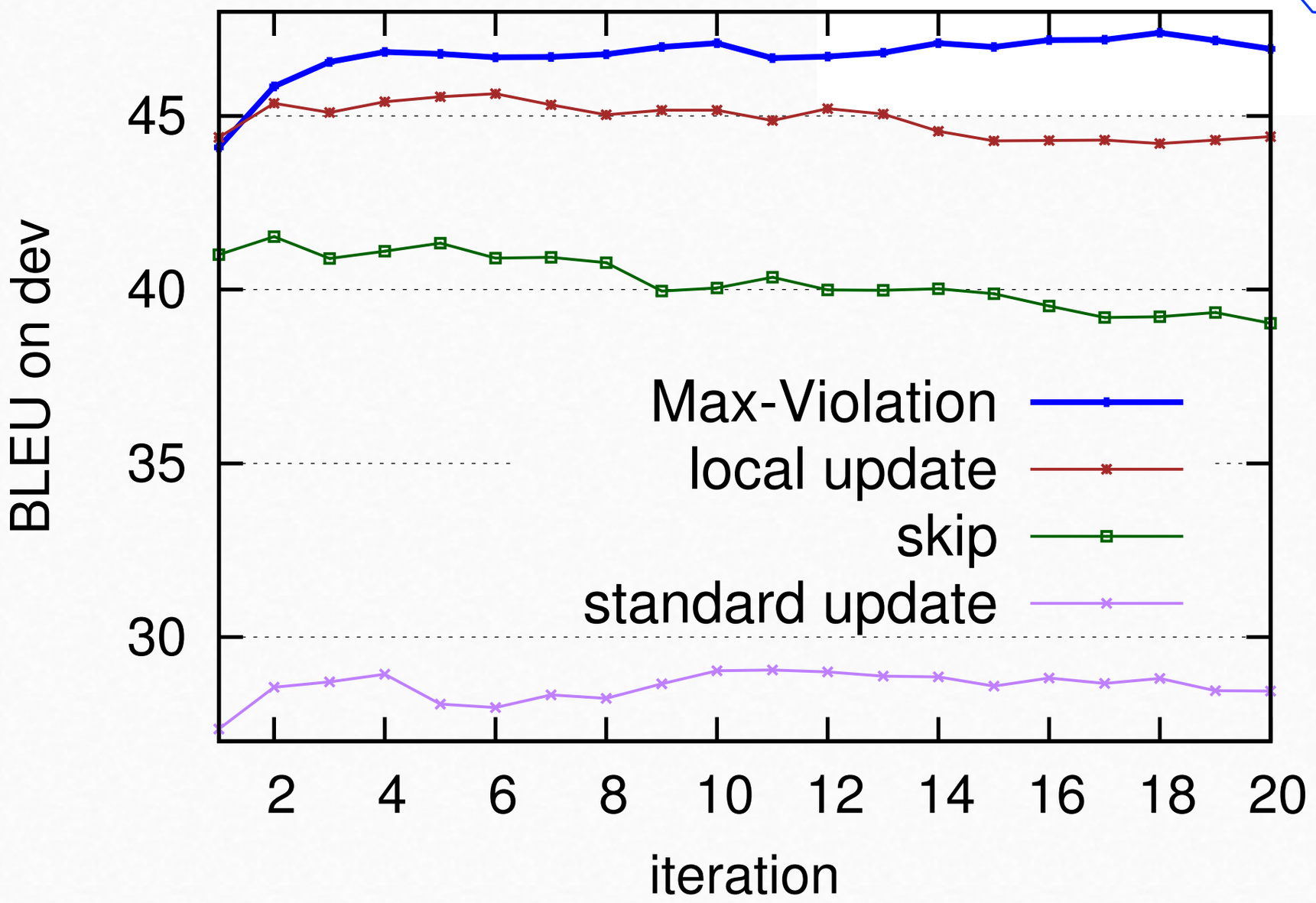


% reachability (in # of words)

Experiments

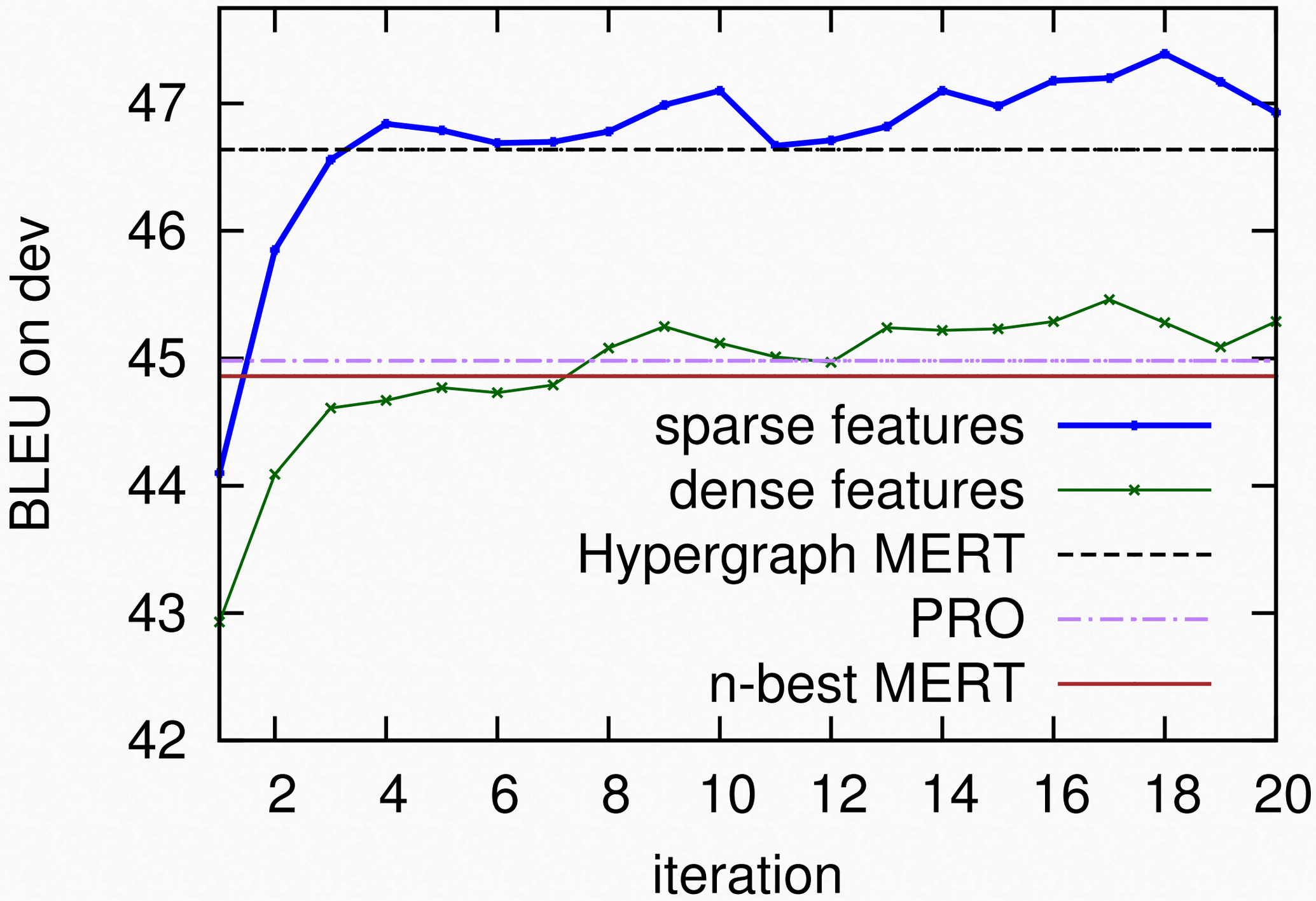
- Comparison of various update methods (IWSLT)





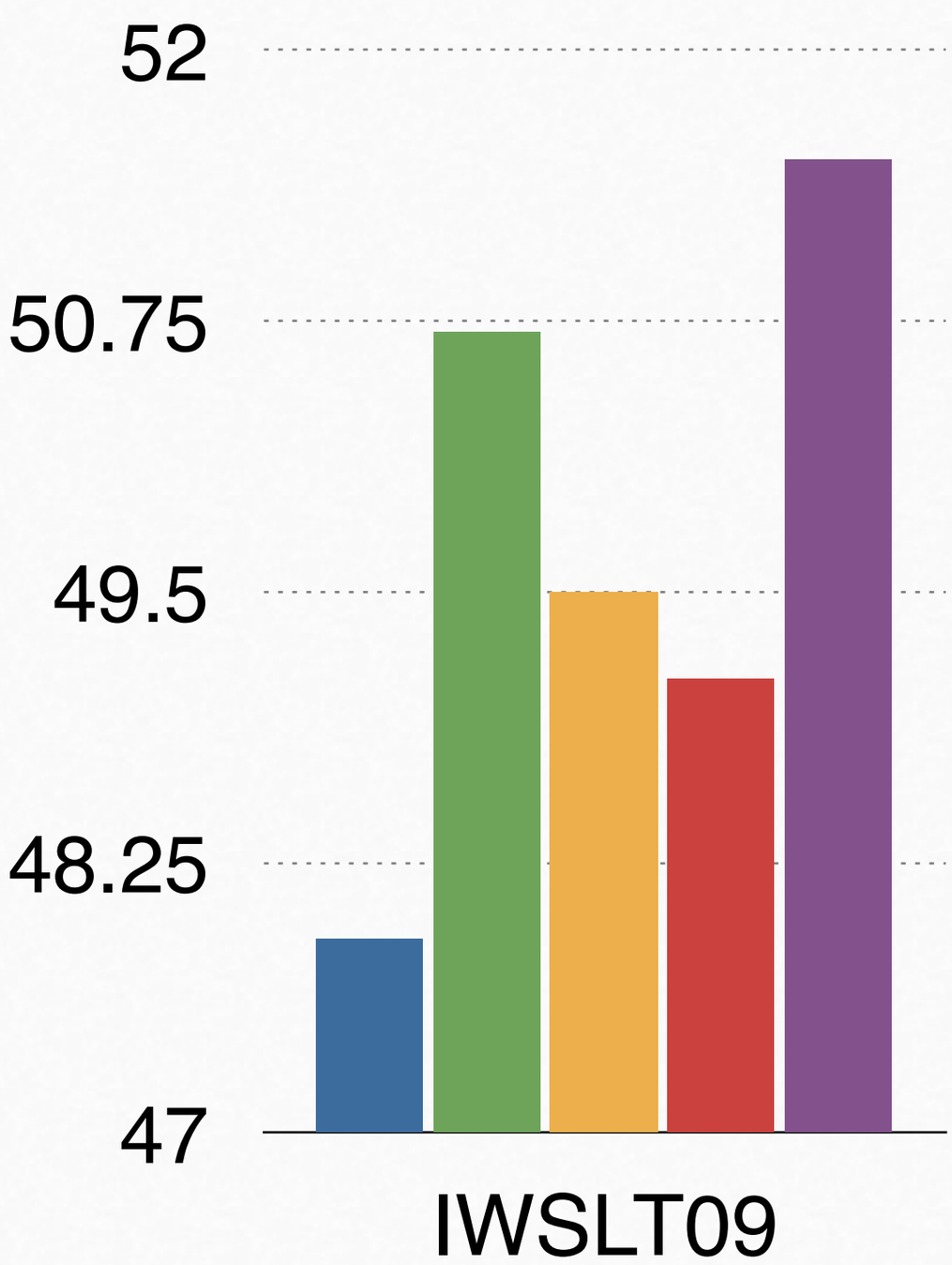
Experiments

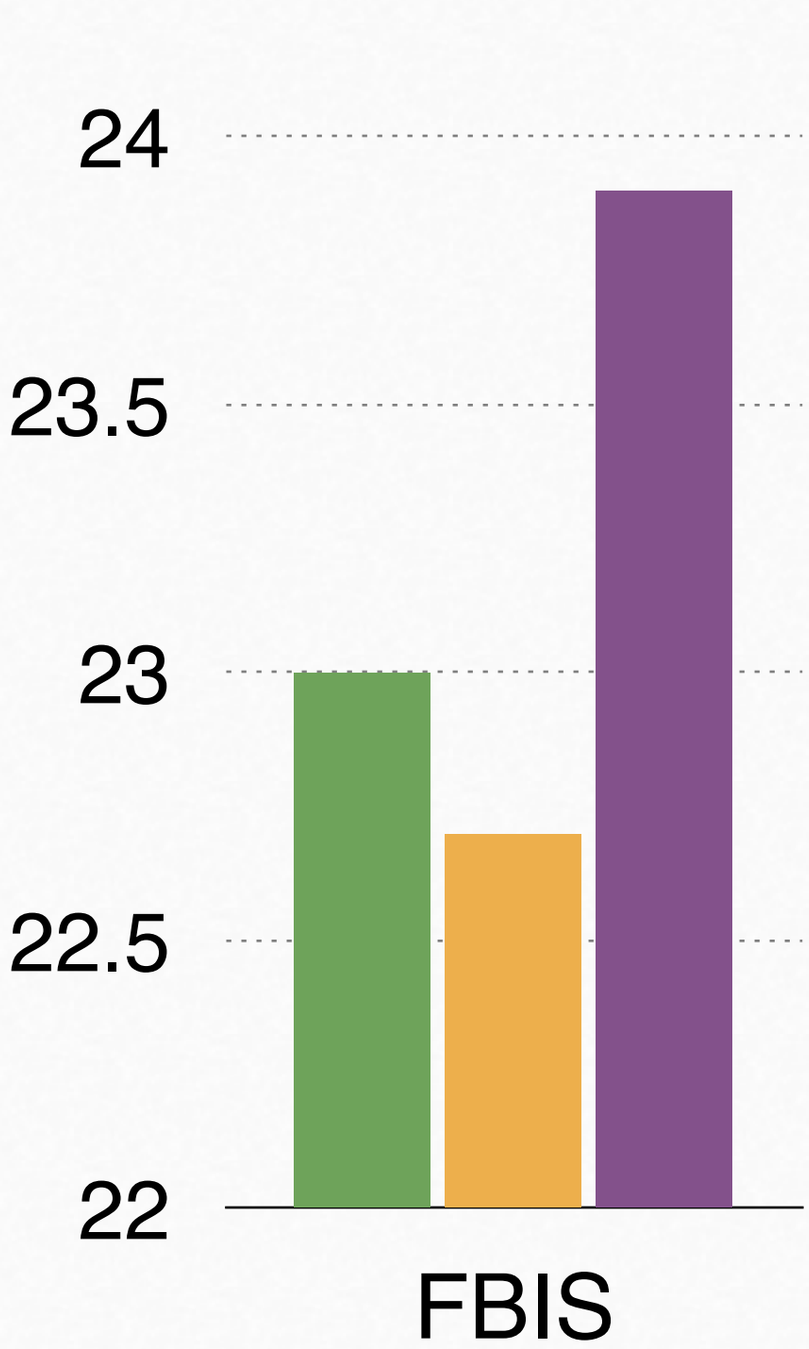
- Contribution of Sparse Features (on IWSLT09)

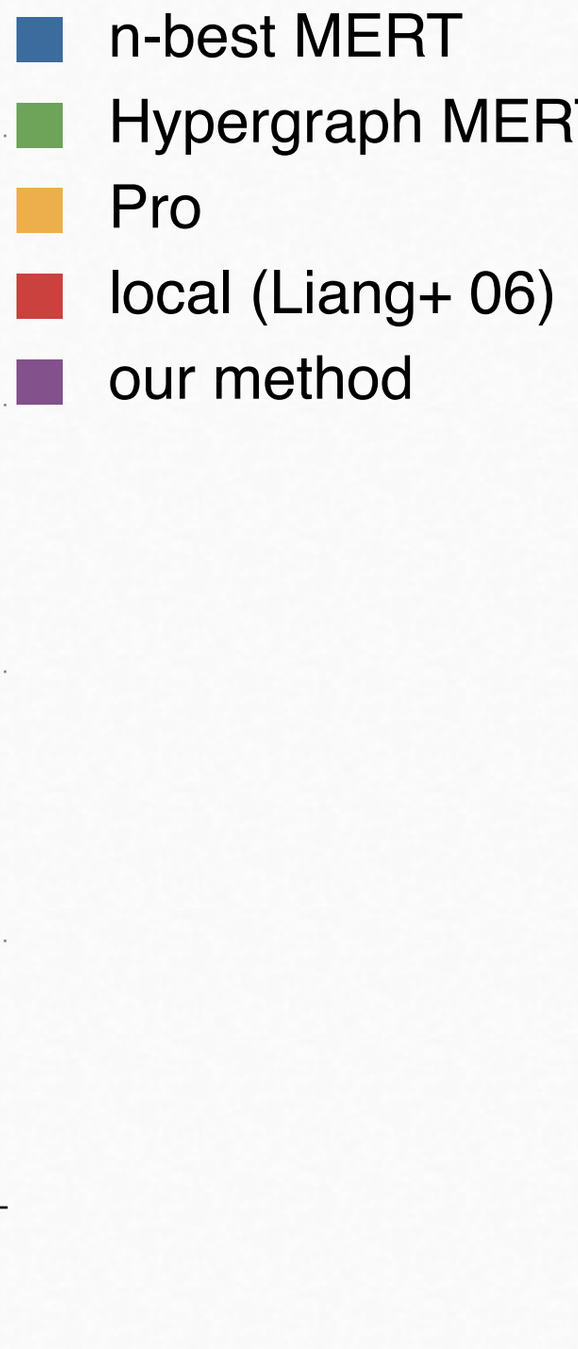


Experiments

- BLEU scores on IWSLT09 (16 refs) and FBIS (4 refs)

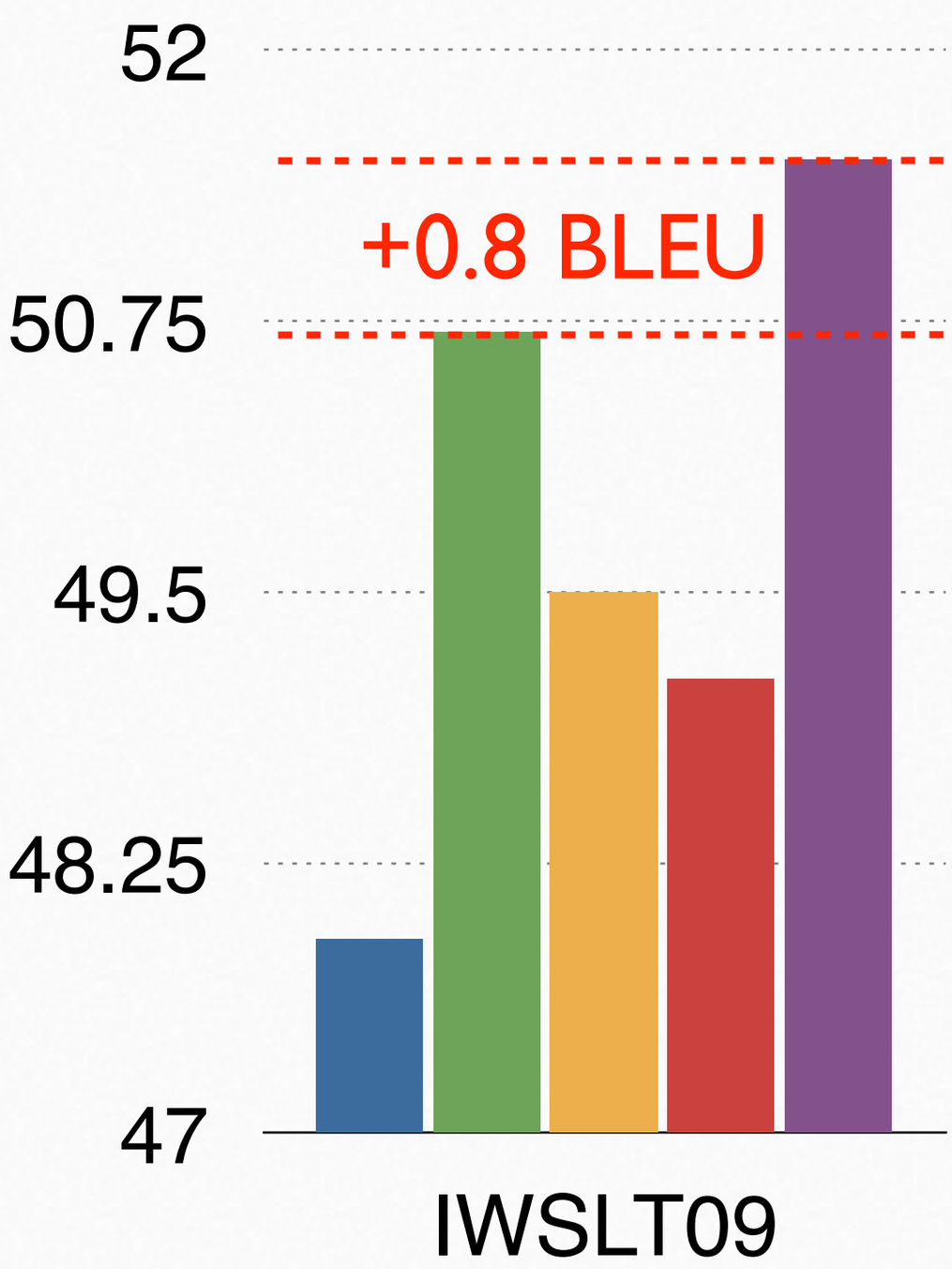


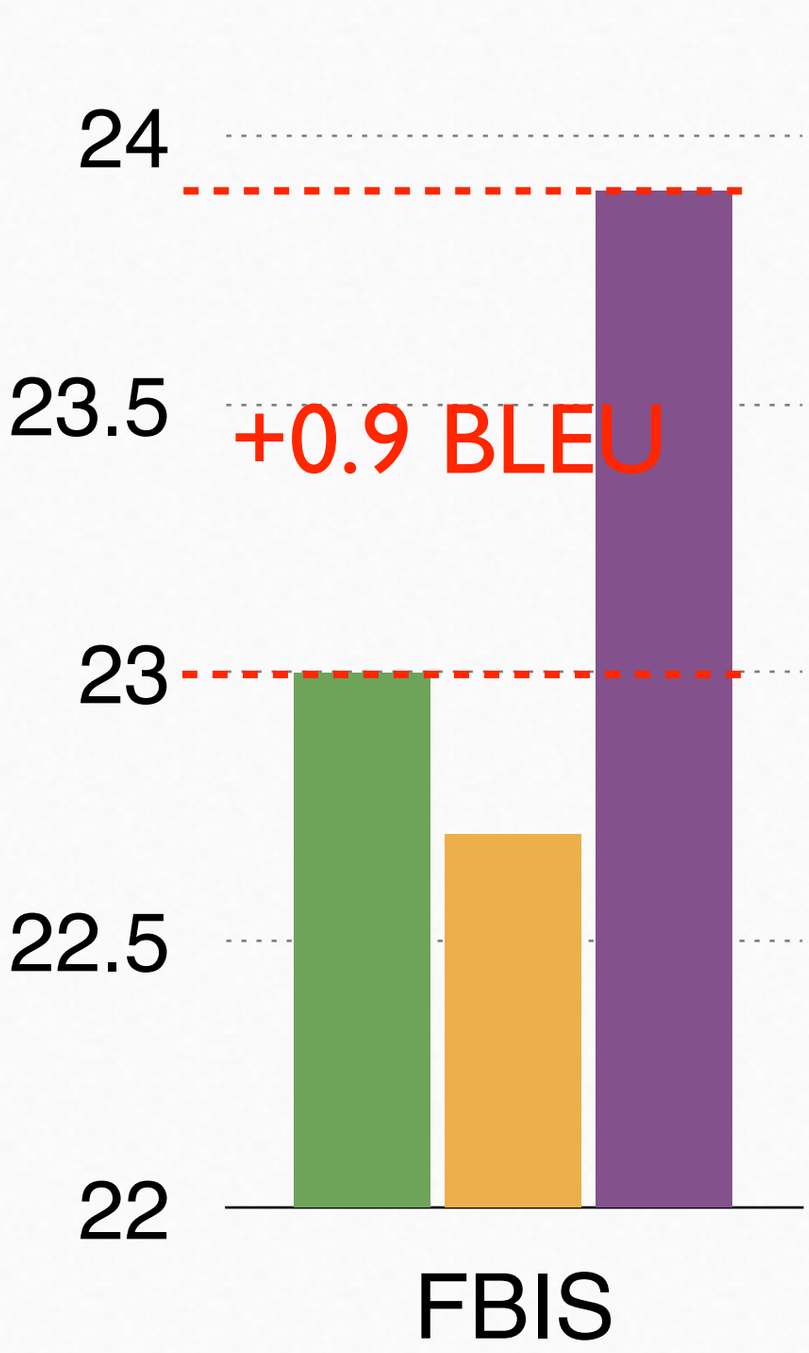


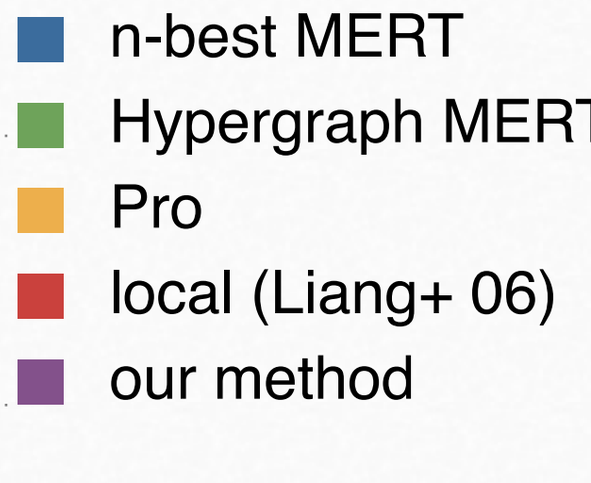


Experiments

- BLEU scores on IWSLT09 (16 refs) and FBIS (4 refs)







Conclusion

- A **latent-variable violation-fixing perceptron** framework for general structured prediction problems with **inexact search** over **hypergraphs**
- Compared with PBMT, it can use **more** training sentences.
- Compared with MERT/PRO, it is **simpler** in theory and practice, and achieves **better** translations