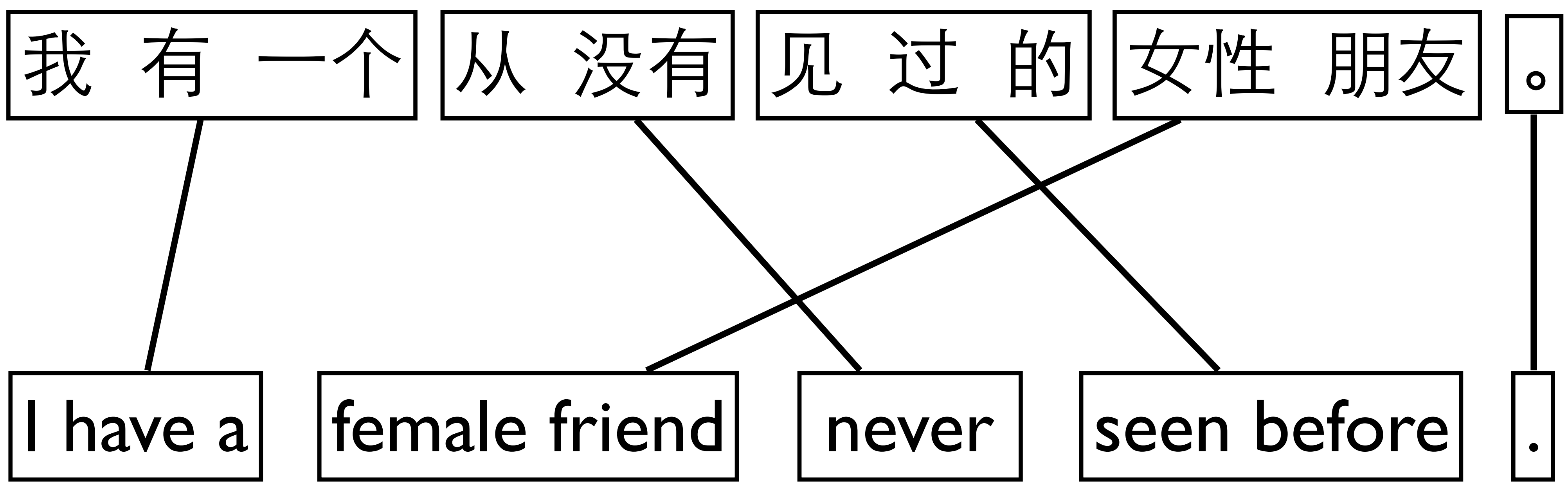
Recursive Autoencoders for ITG-based Translation

Peng Li
Tsinghua University
pengli09@gmail.com

(Joint work with Yang Liu and Maosong Sun)

Overview

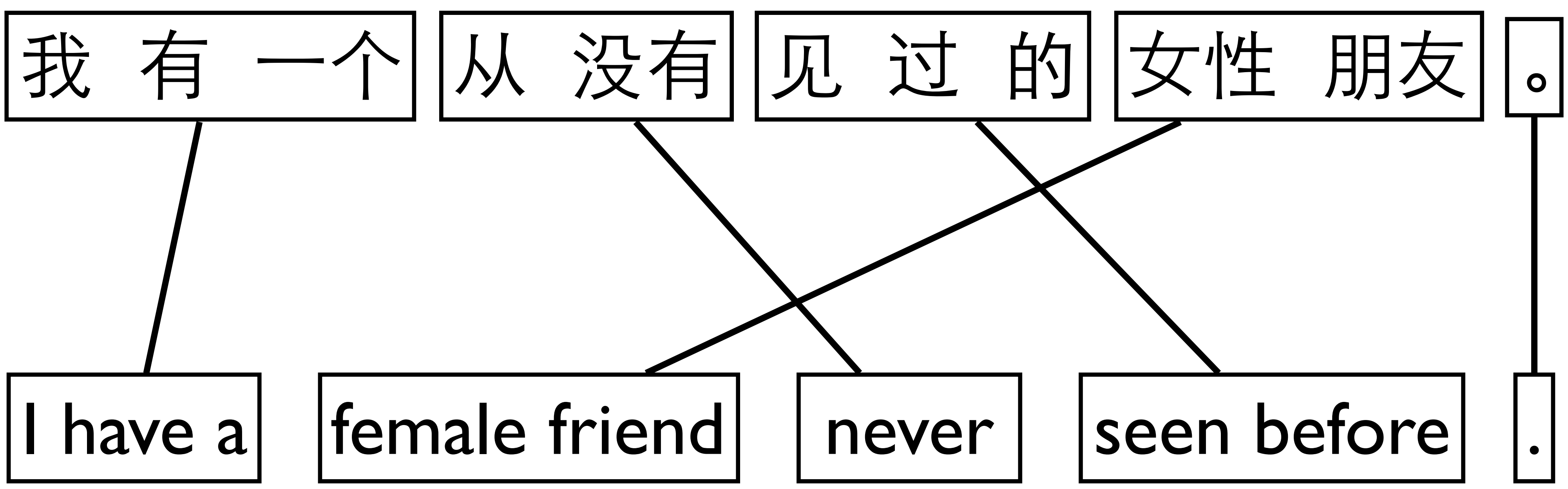
- Phrase reordering model is a critical problem in machine translation (MT), and is NP-complete



(Knight, 1999)

Distortion Models

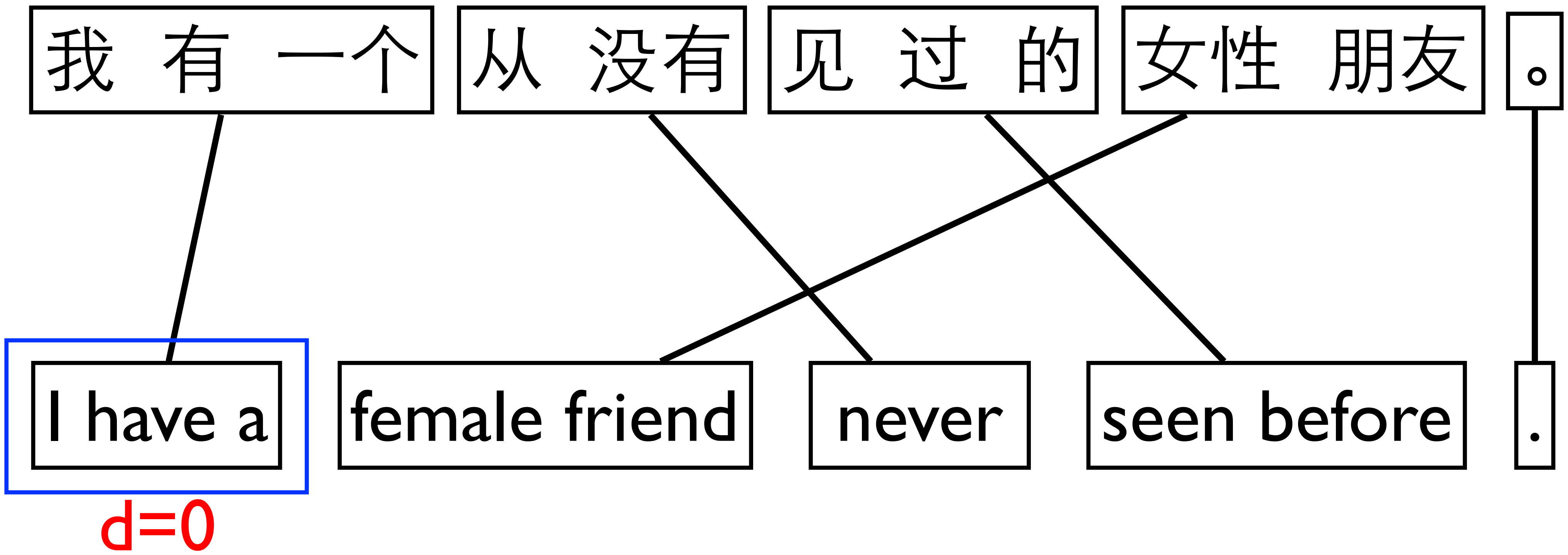
- Distortion models: penalize relative displacement of source phrases



(Koehn et al., 2003; Och and Ney, 2004)

Distortion Models

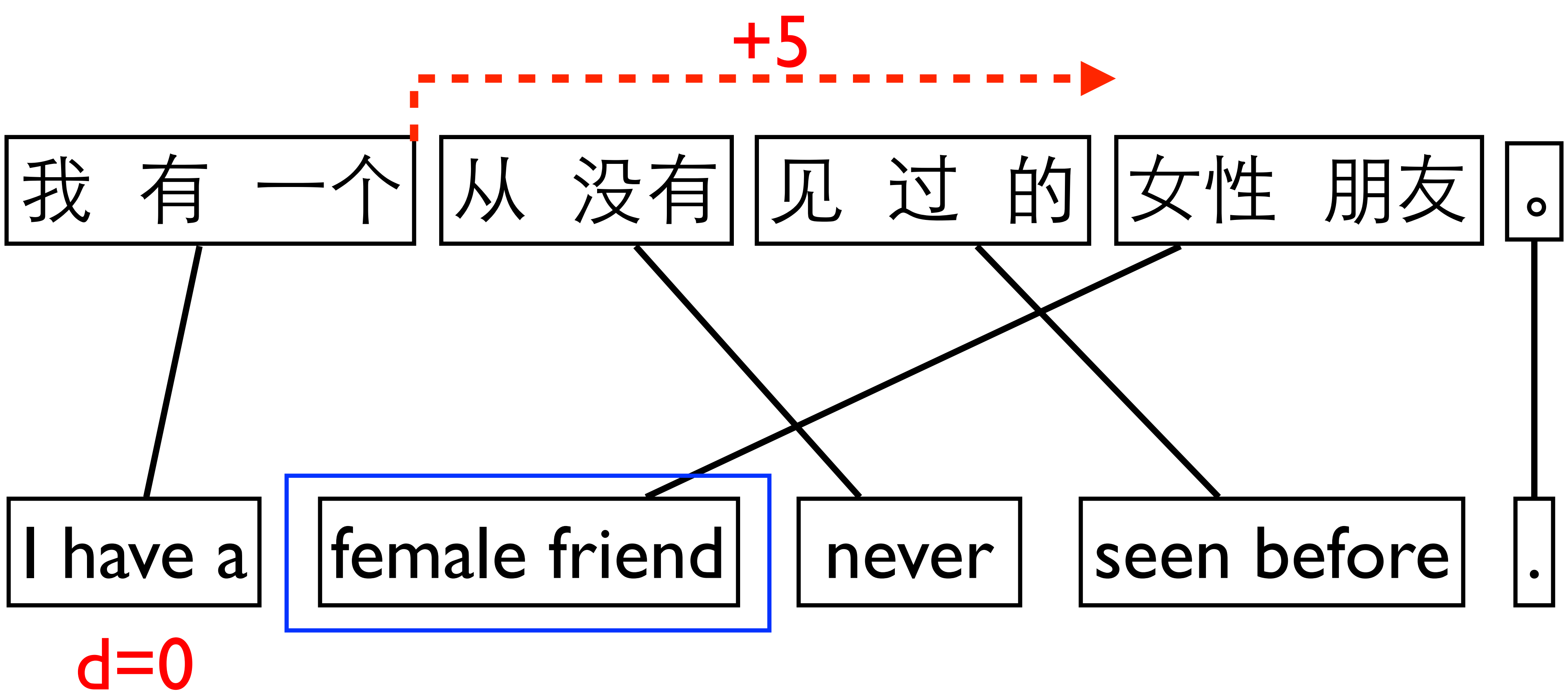
- Distortion models: penalize relative displacement of source phrases



(Koehn et al., 2003; Och and Ney, 2004)

Distortion Models

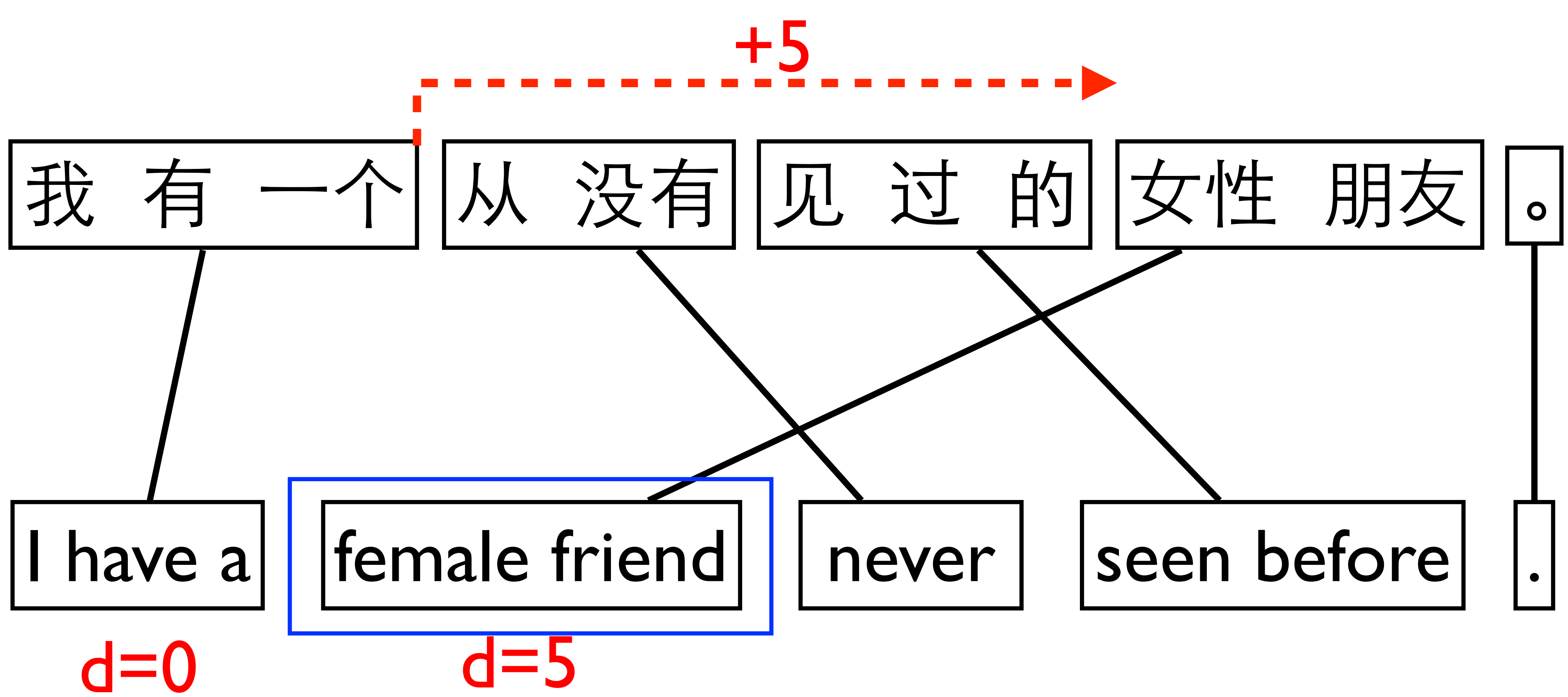
- Distortion models: penalize relative displacement of source phrases



(Koehn et al., 2003; Och and Ney, 2004)

Distortion Models

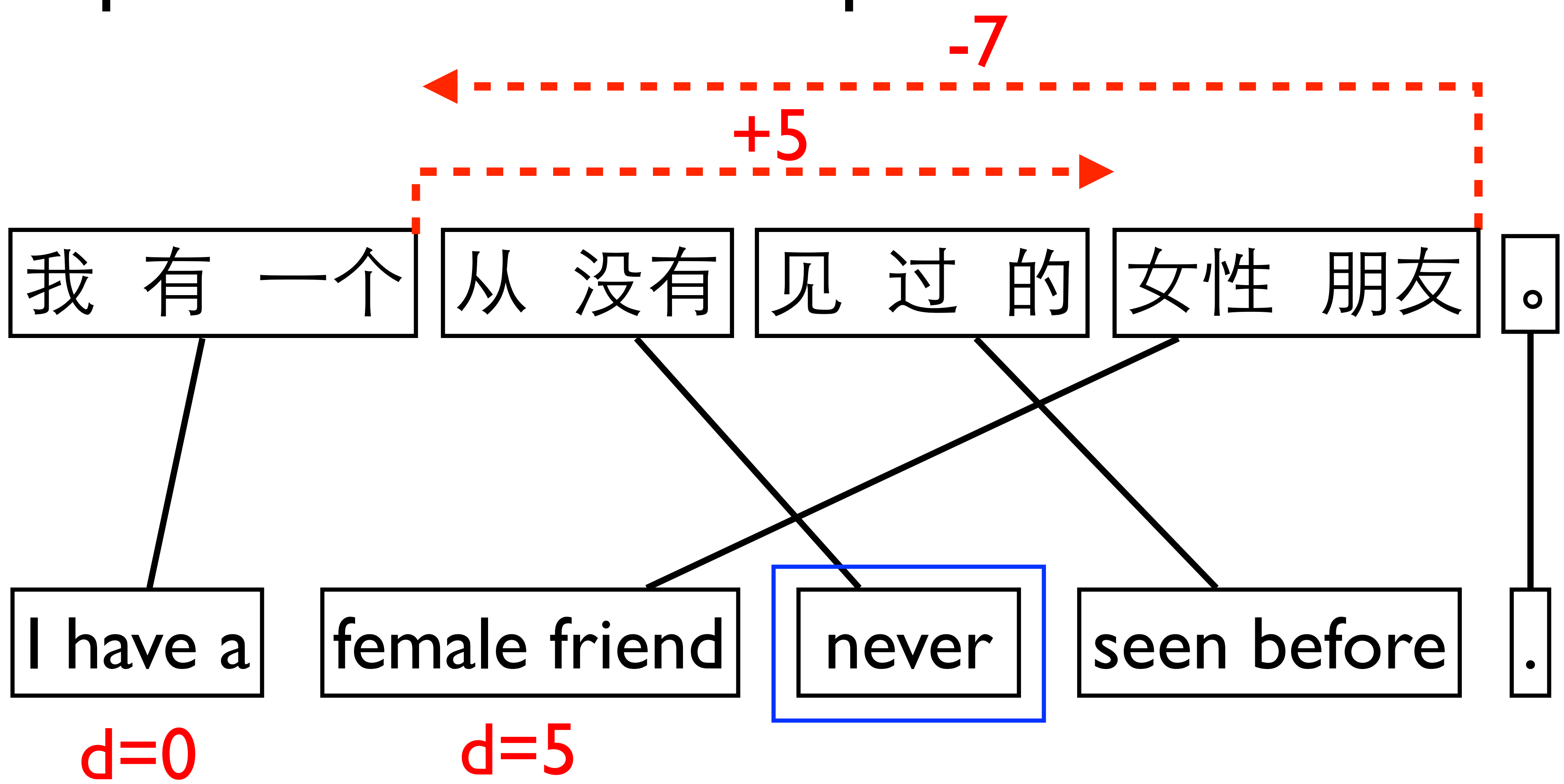
- Distortion models: penalize relative displacement of source phrases



(Koehn et al., 2003; Och and Ney, 2004)

Distortion Models

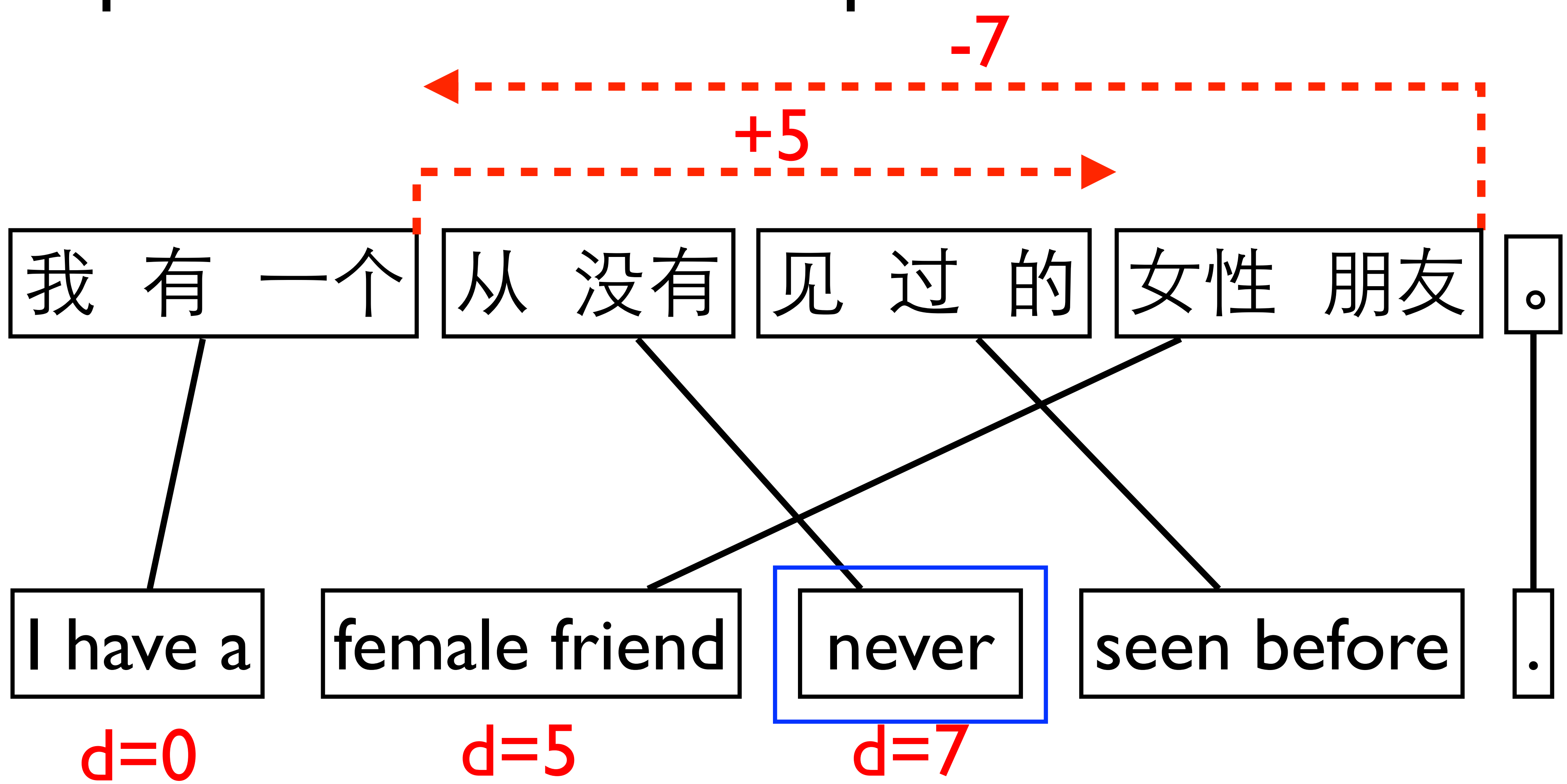
- Distortion models: penalize relative displacement of source phrases



(Koehn et al., 2003; Och and Ney, 2004)

Distortion Models

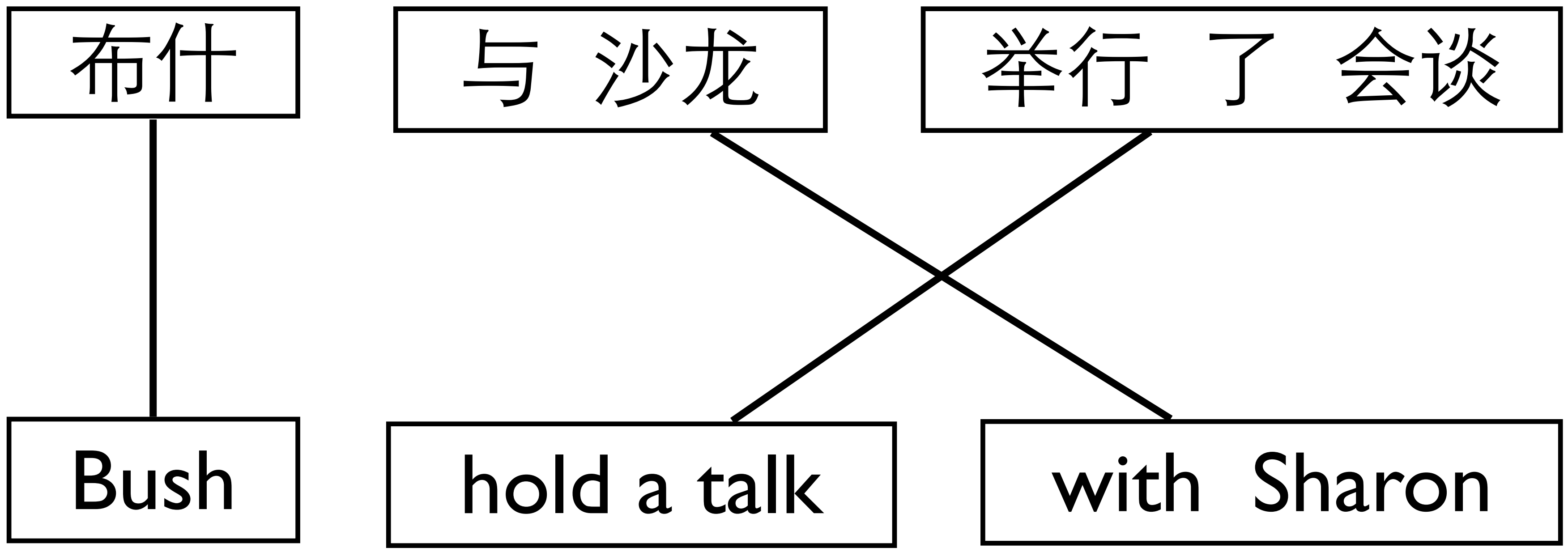
- Distortion models: penalize relative displacement of source phrases



(Koehn et al., 2003; Och and Ney, 2004)

Lexicalized Reordering Models

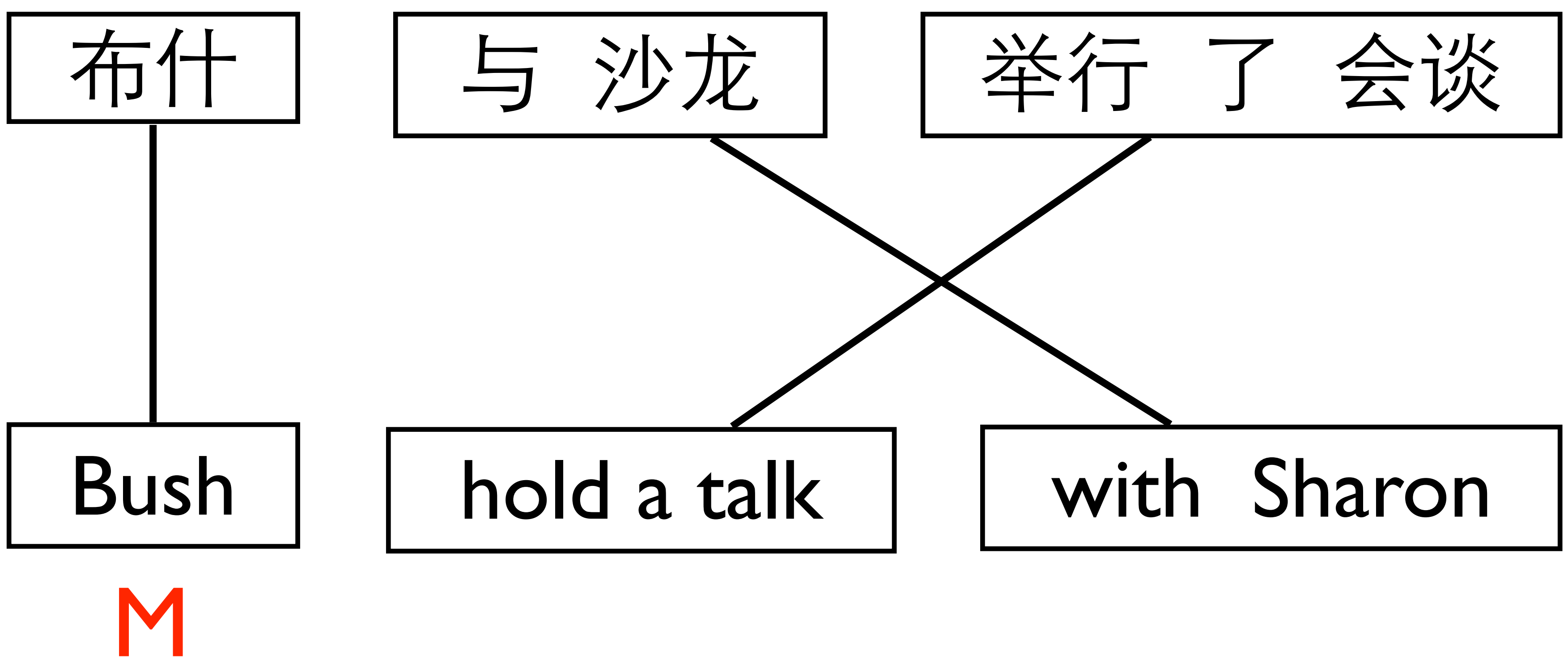
- Lexicalized reordering models: penalize reordering conditioned on both the source and target phrases



(Koehn et al., 2007)

Lexicalized Reordering Models

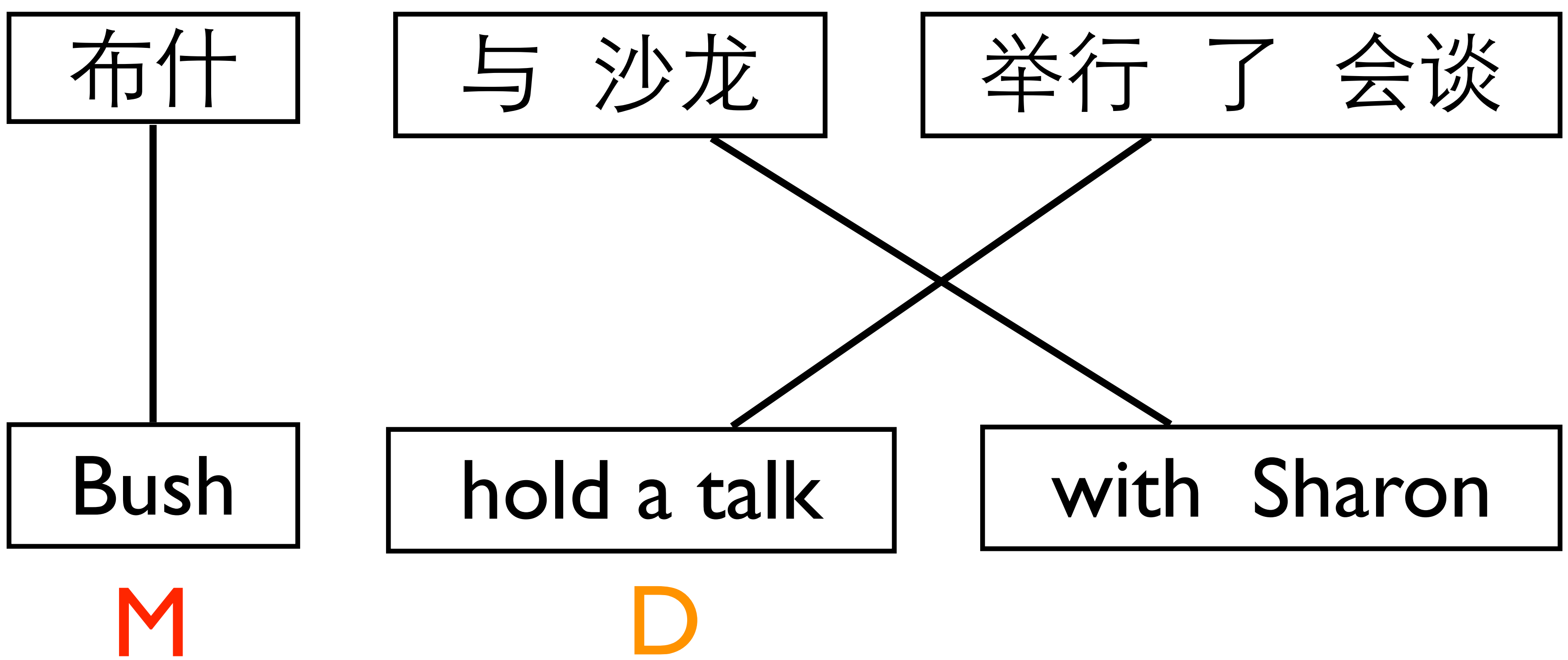
- Lexicalized reordering models: penalize reordering conditioned on both the source and target phrases



(Koehn et al., 2007)

Lexicalized Reordering Models

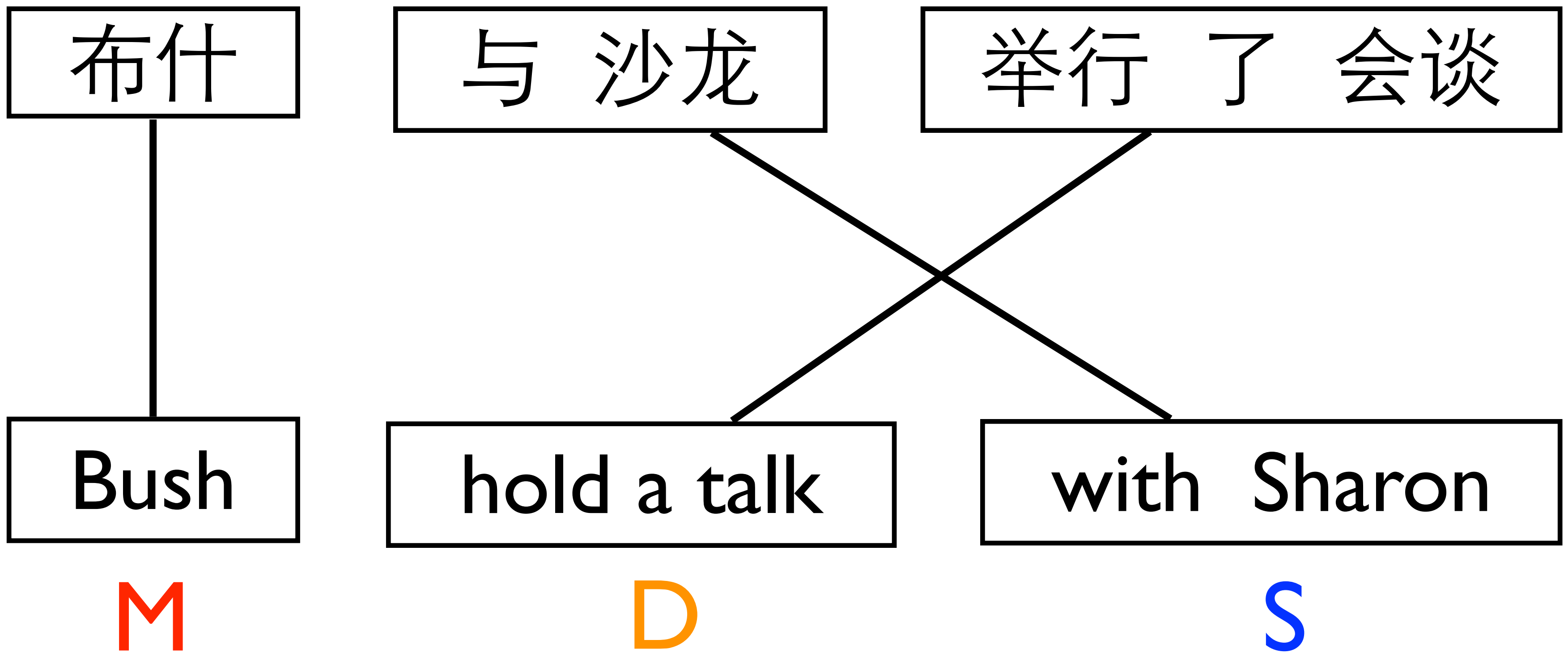
- Lexicalized reordering models: penalize reordering conditioned on both the source and target phrases



(Koehn et al., 2007)

Lexicalized Reordering Models

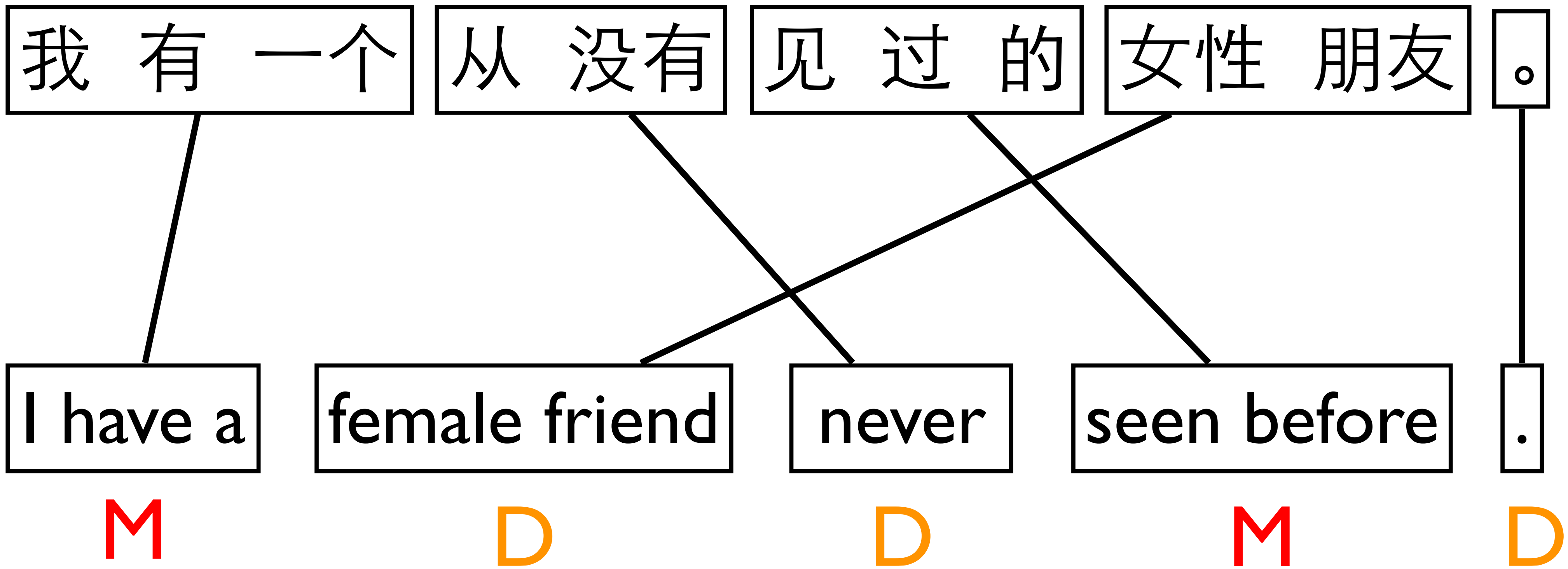
- Lexicalized reordering models: penalize reordering conditioned on both the source and target phrases



(Koehn et al., 2007)

Lexicalized Reordering Models

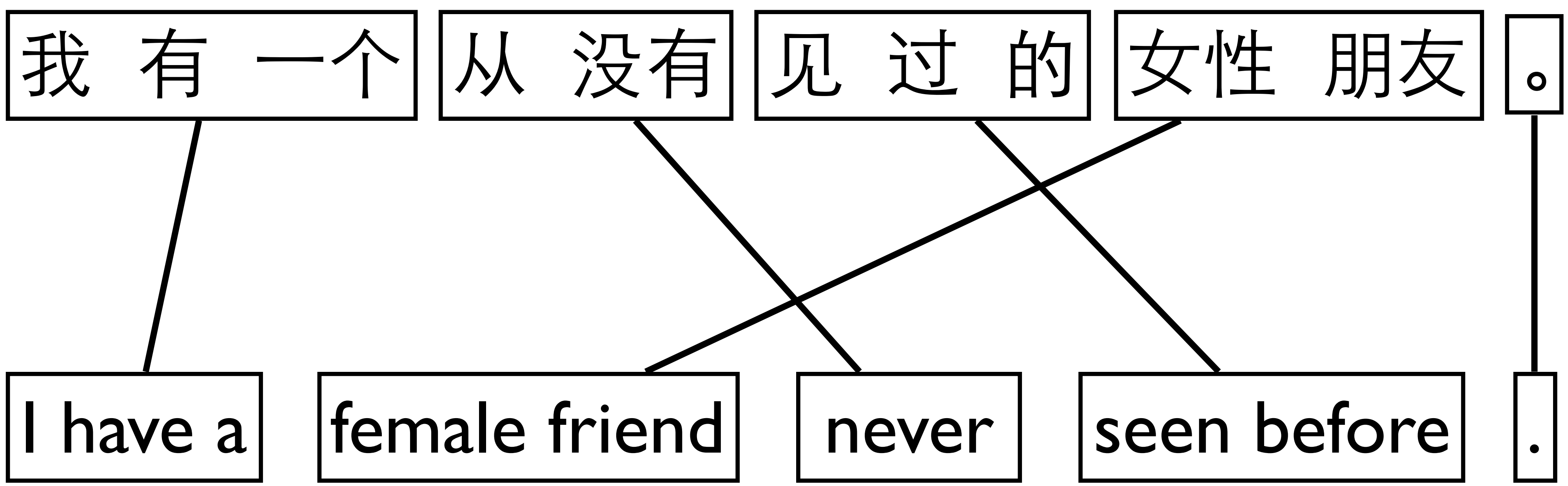
- Lexicalized reordering models: penalize reordering conditioned on both the source and target phrases



(Koehn et al., 2007)

Block Merging

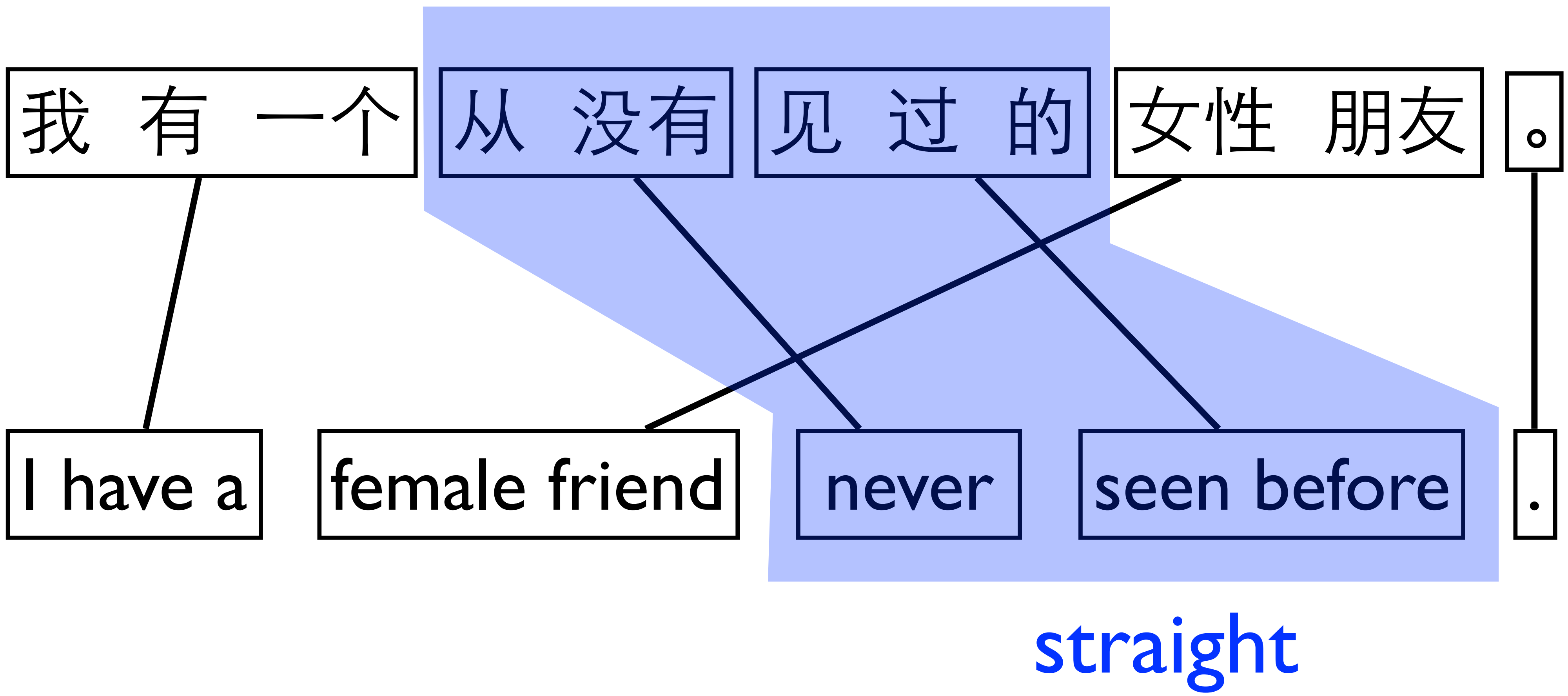
- Reordering as block merging



(Wu, 1997; Xiong et al., 2006)

Block Merging

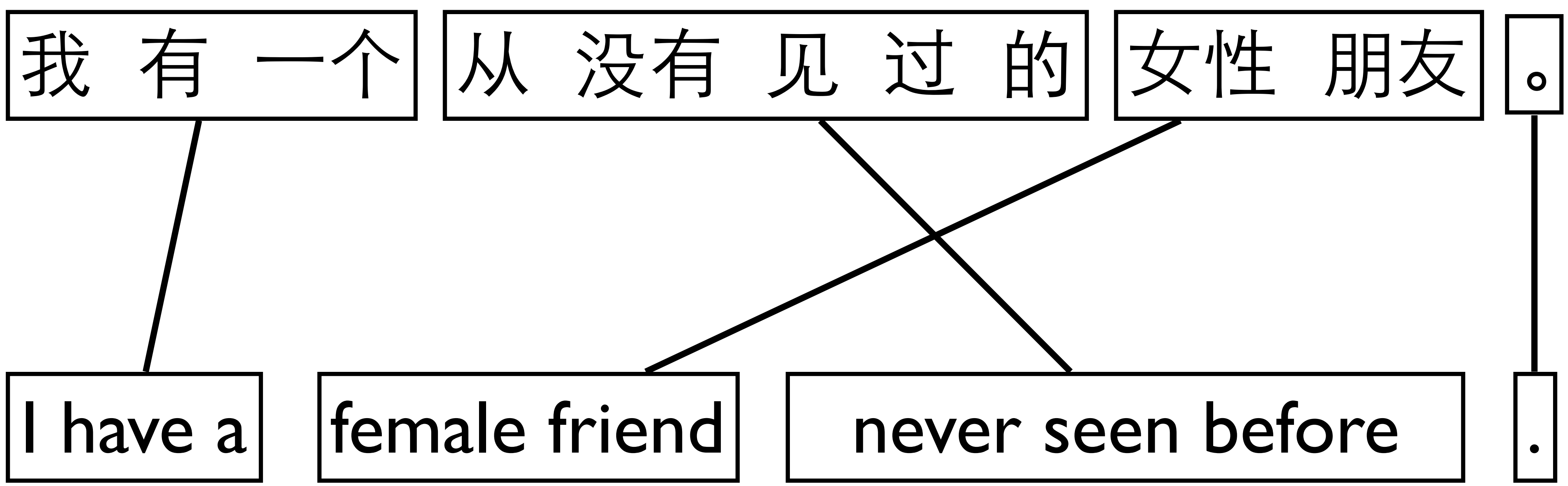
- Reordering as block merging



(Wu, 1997; Xiong et al., 2006)

Block Merging

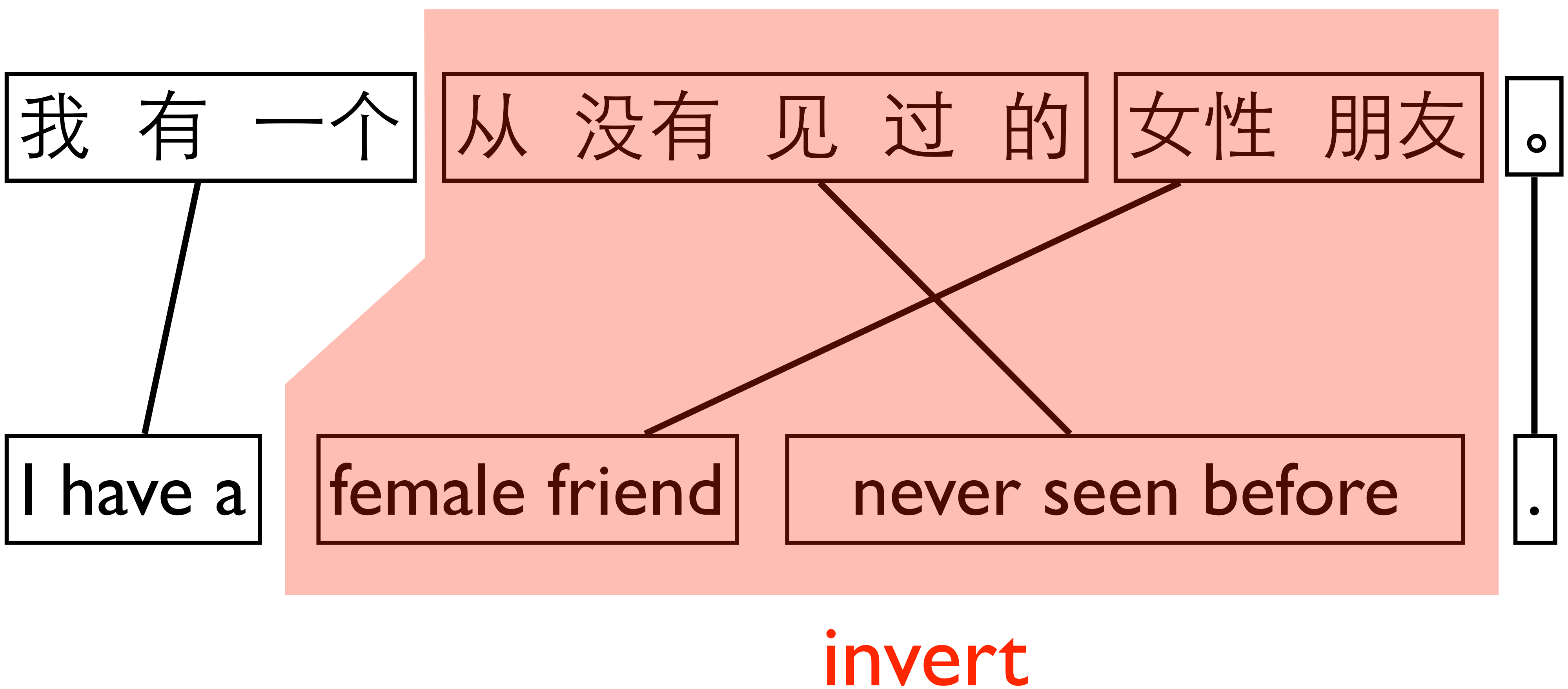
- Reordering as block merging



(Wu, 1997; Xiong et al., 2006)

Block Merging

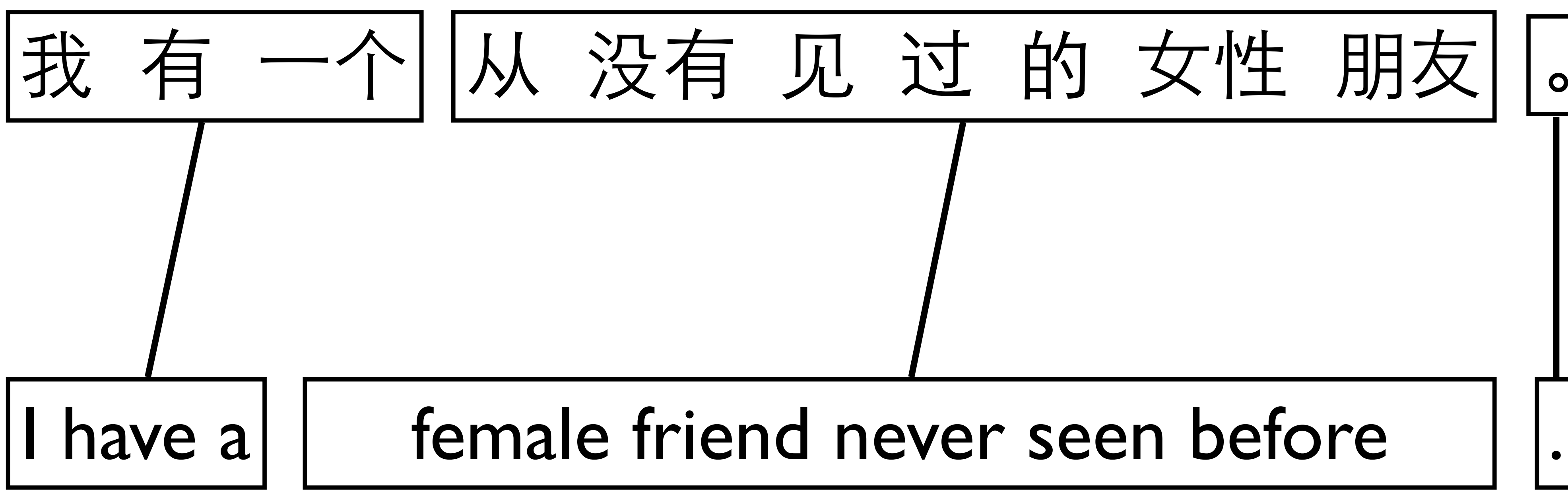
- Reordering as block merging



(Wu, 1997; Xiong et al., 2006)

Block Merging

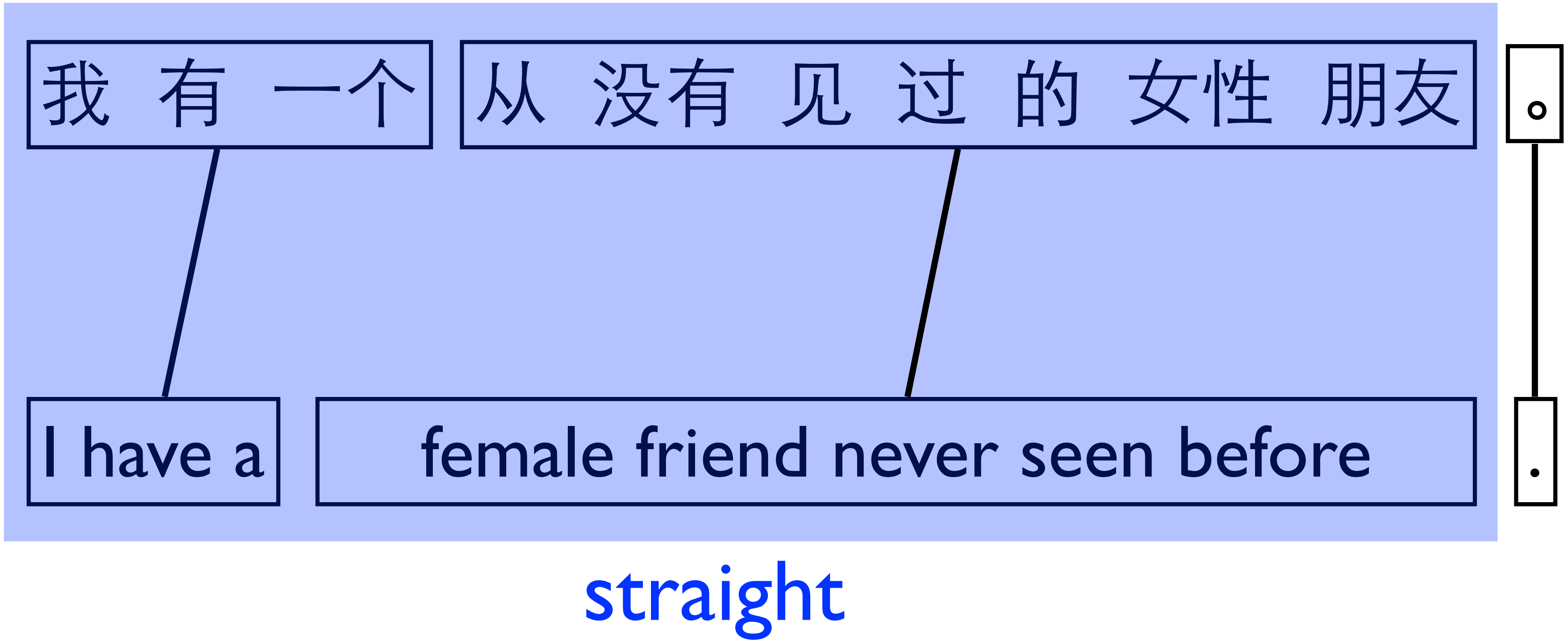
- Reordering as block merging



(Wu, 1997; Xiong et al., 2006)

Block Merging

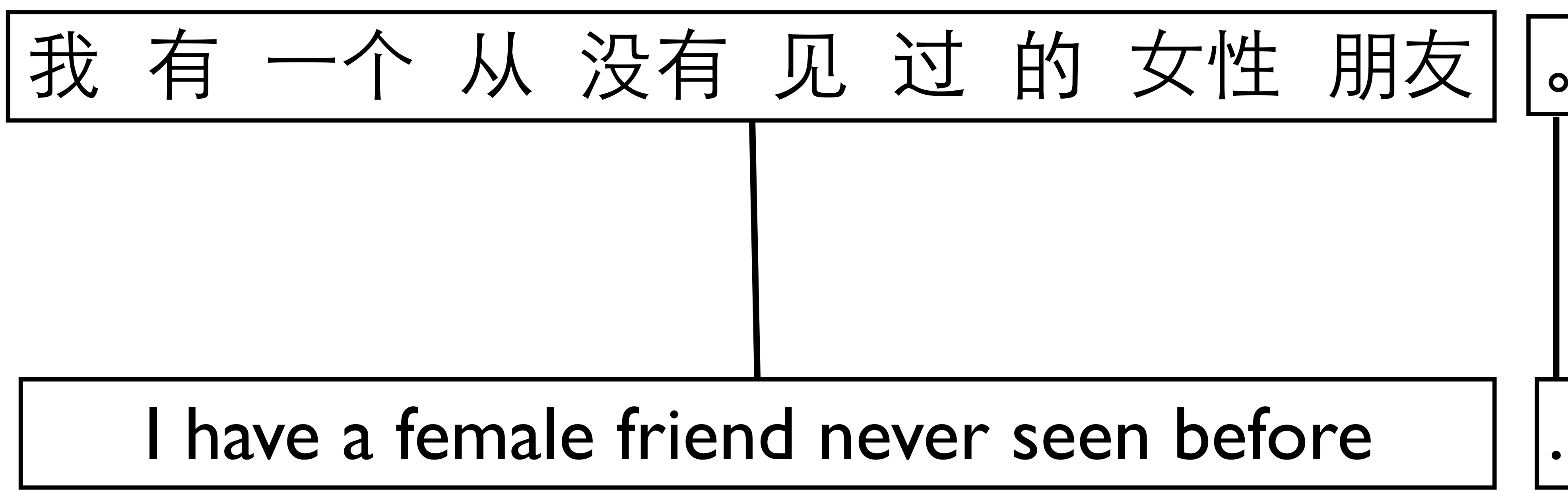
- Reordering as block merging



(Wu, 1997; Xiong et al., 2006)

Block Merging

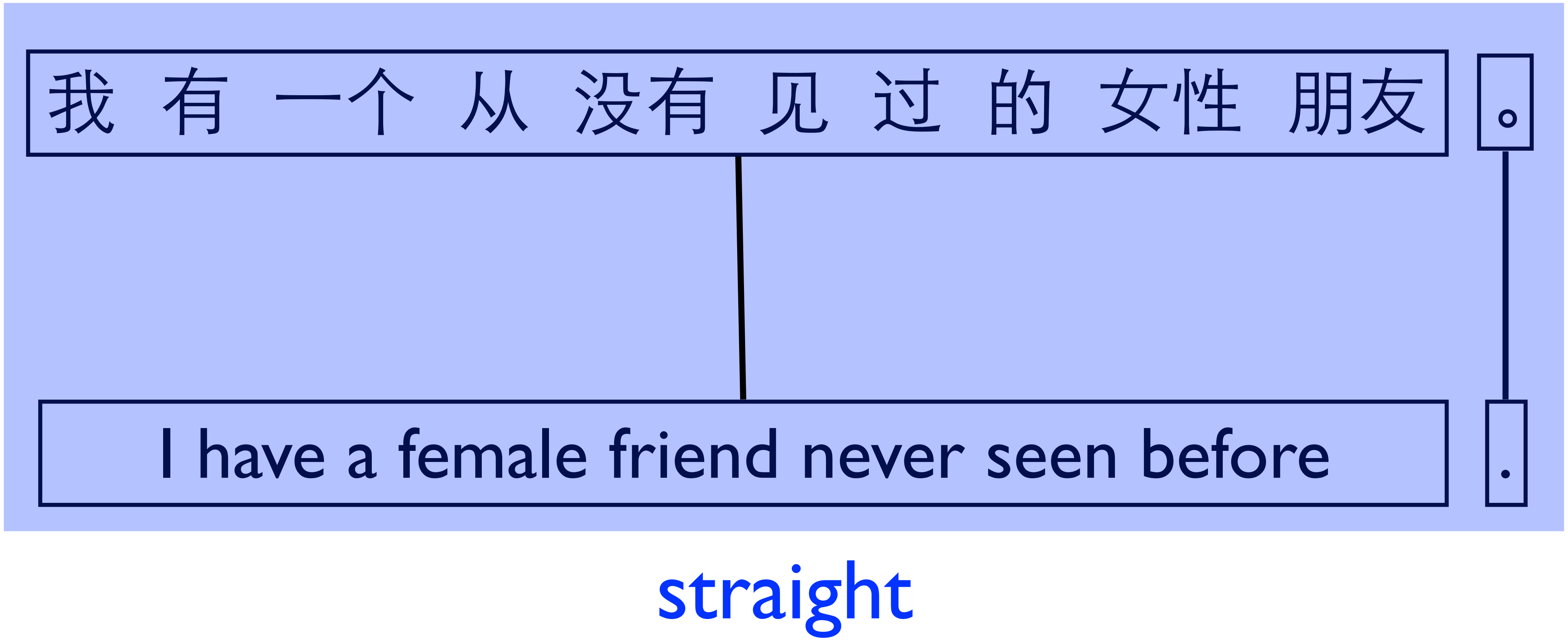
- Reordering as block merging



(Wu, 1997; Xiong et al., 2006)

Block Merging

- Reordering as block merging



(Wu, 1997; Xiong et al., 2006)

Block Merging

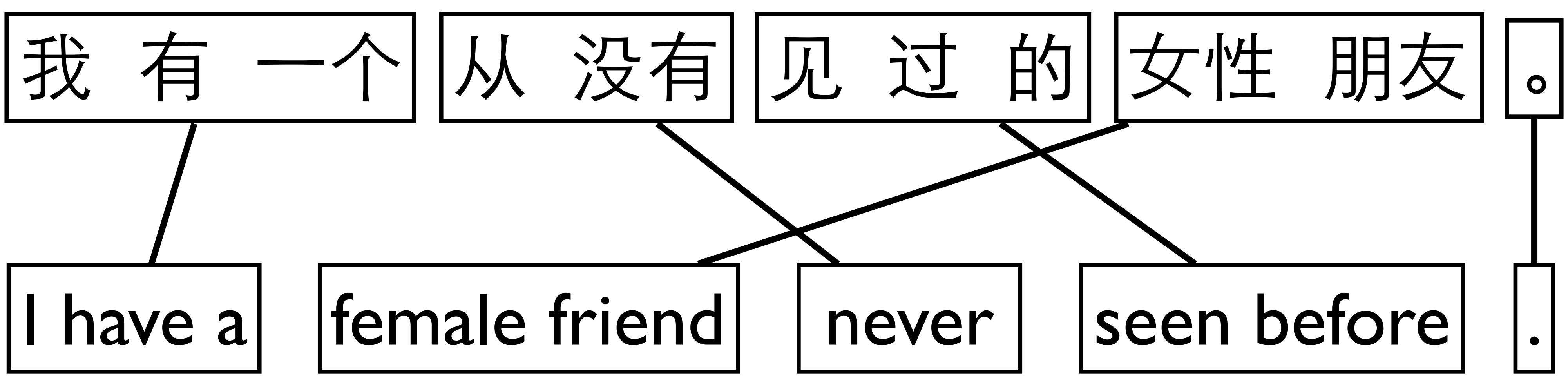
- Reordering as block merging

我 有 一个 从 没有 见 过 的 女性 朋友 。

I have a female friend never seen before .

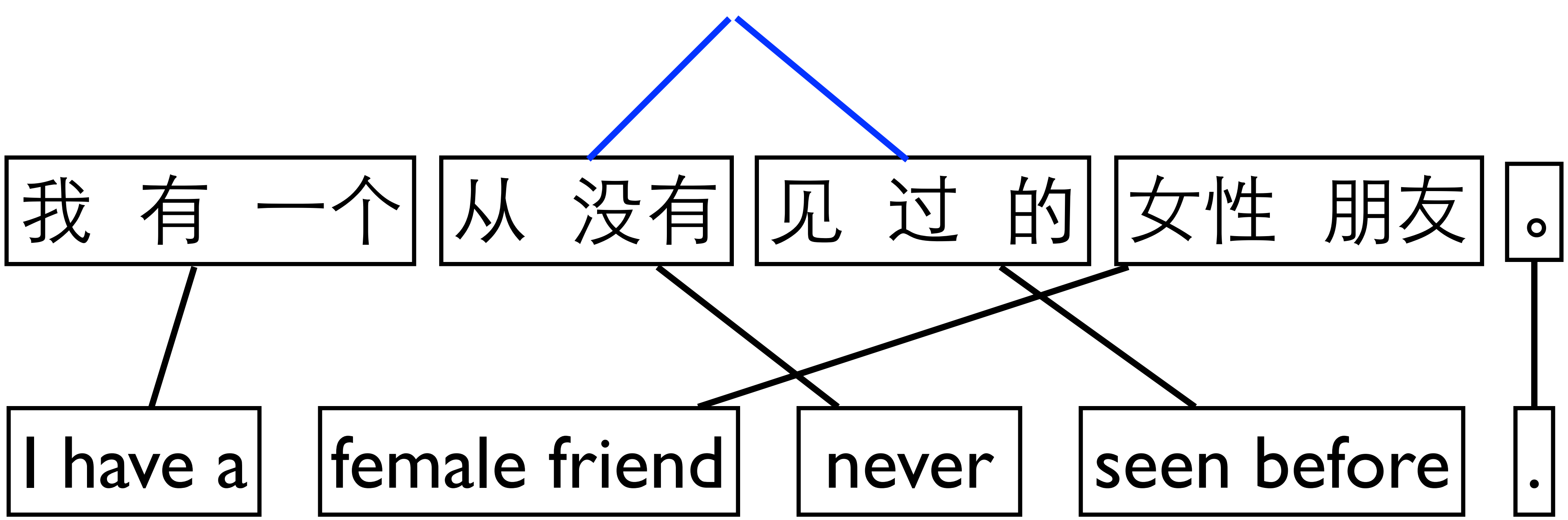
(Wu, 1997; Xiong et al., 2006)

Block Merging



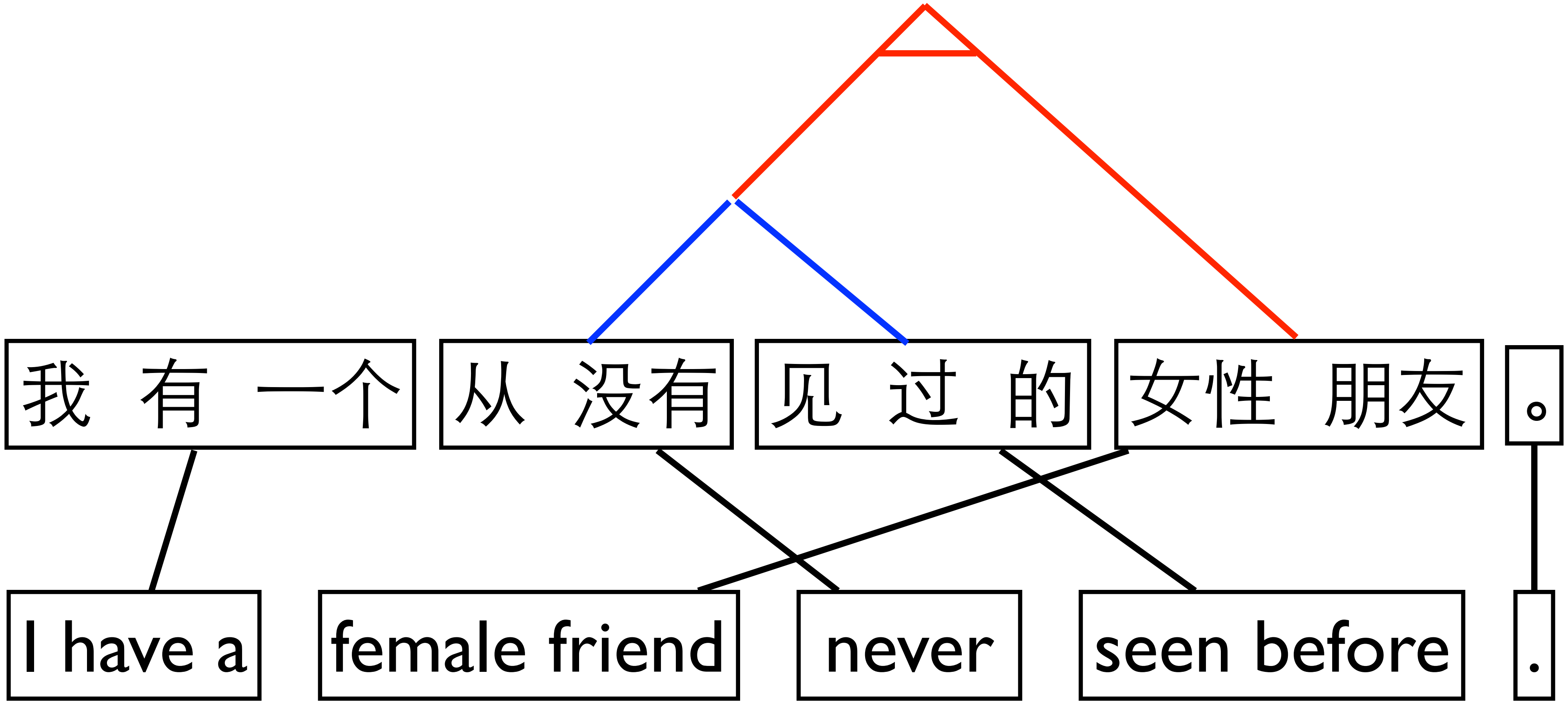
(Wu, 1997; Xiong et al., 2006)

Block Merging



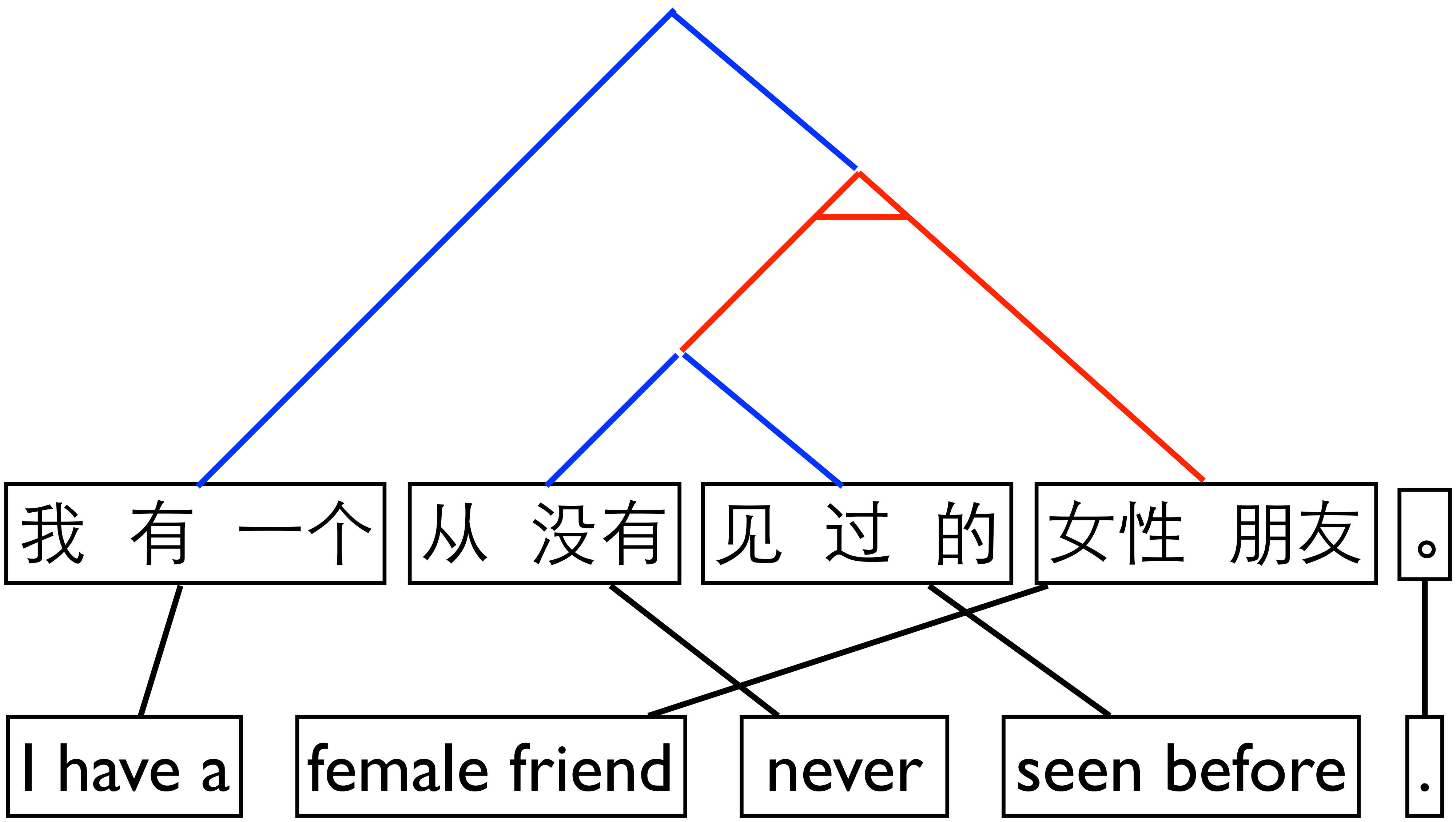
(Wu, 1997; Xiong et al., 2006)

Block Merging



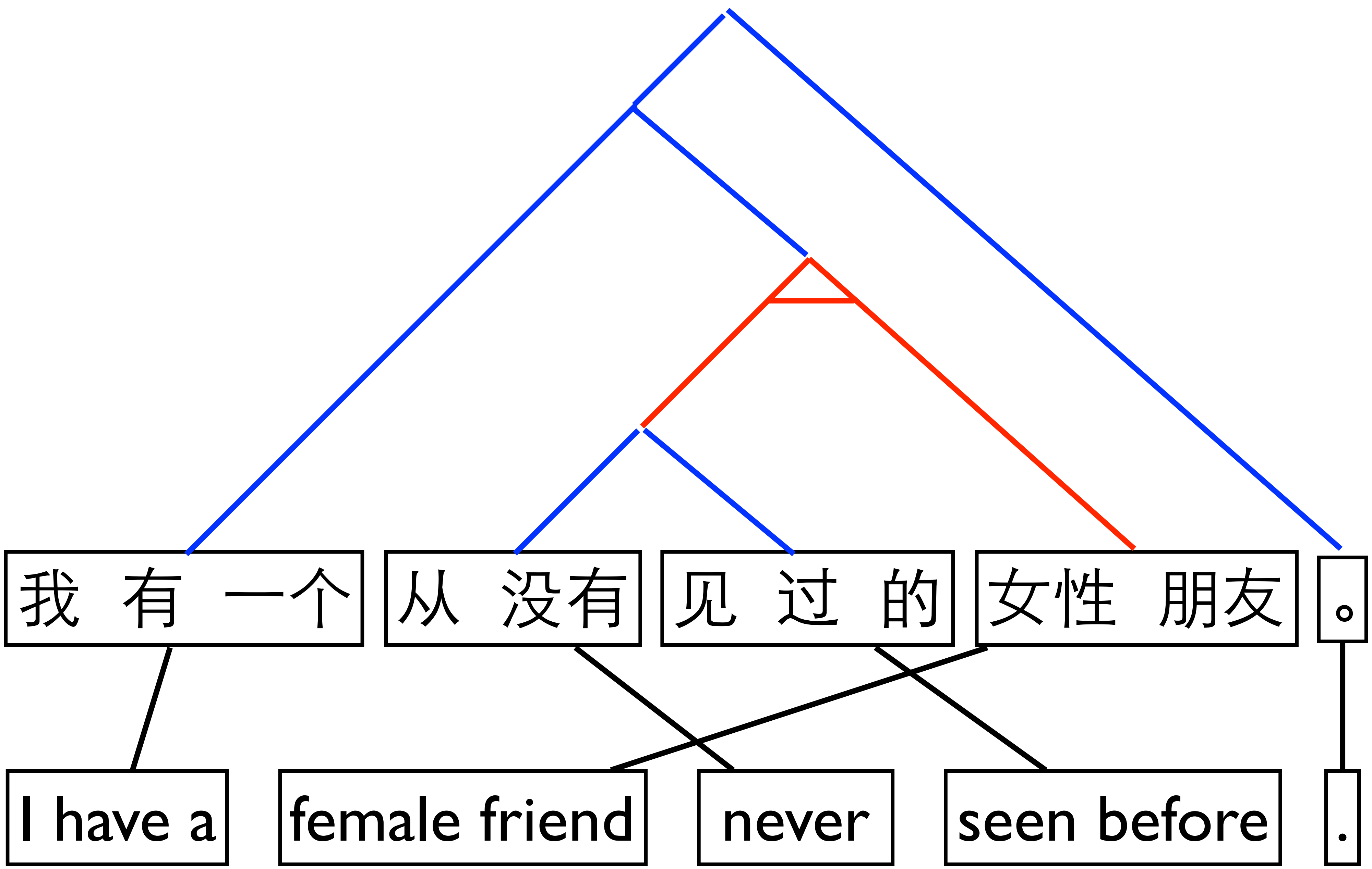
(Wu, 1997; Xiong et al., 2006)

Block Merging



(Wu, 1997; Xiong et al., 2006)

Block Merging



(Wu, 1997; Xiong et al., 2006)

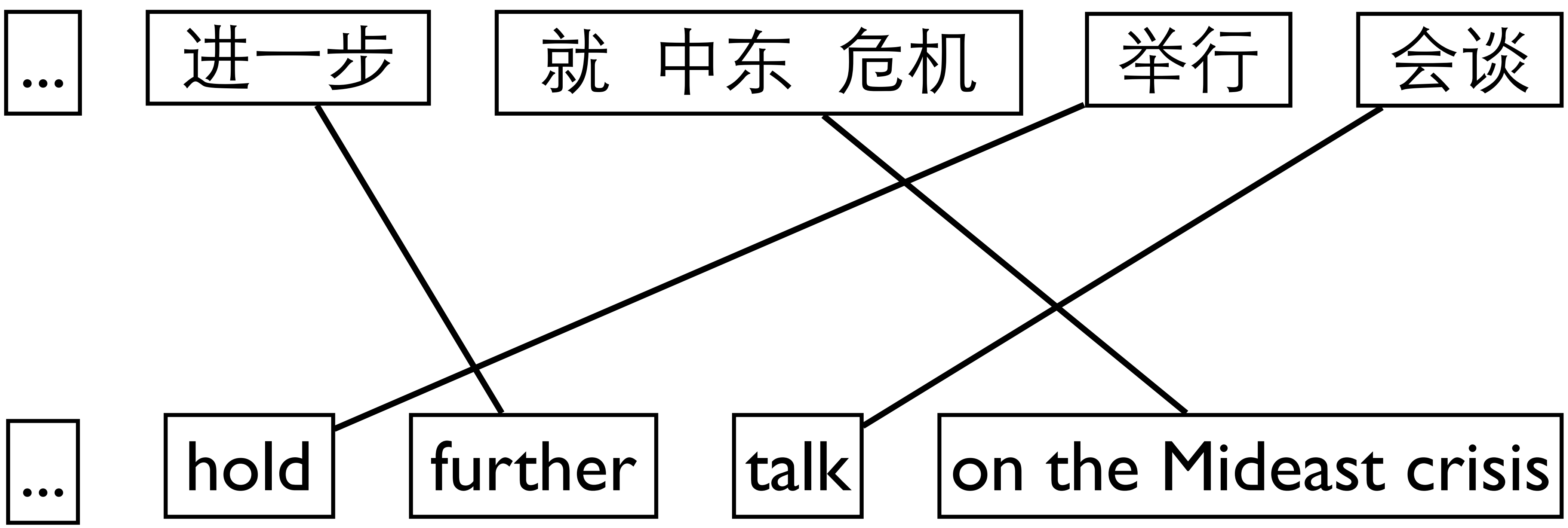
Block Merging

- Can you find a counter example?

(Huang et al., 2009)

Block Merging

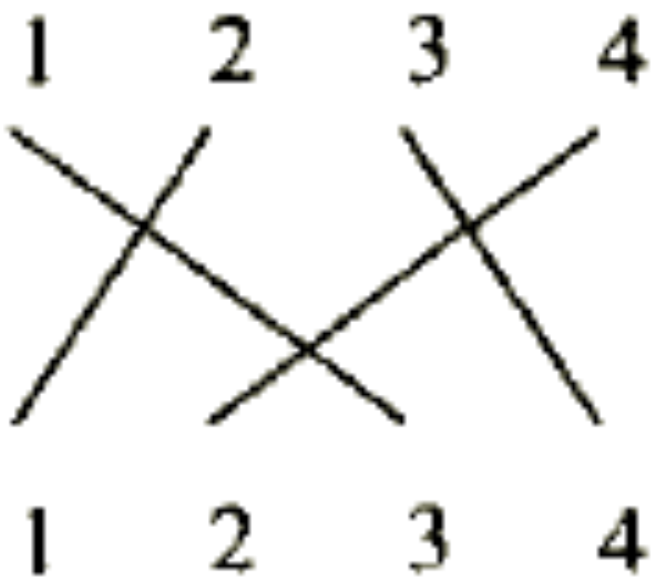
- Can you find a counter example?

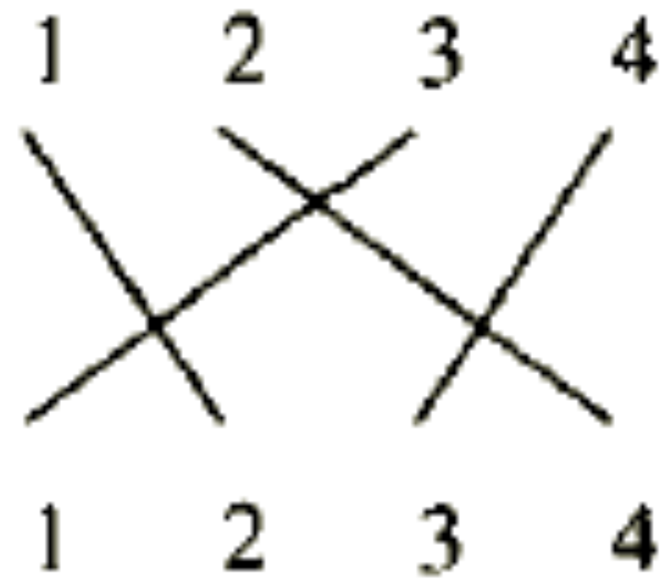


(Huang et al., 2009)

Block Merging

"inside-outside"





(Wu, 1997)

ITG

- Inversion transduction grammar (ITG)

$X \rightarrow [X^1, X^2]$: straight rule

$X \rightarrow \langle X^1, X^2 \rangle$: inverted rule

$X \rightarrow f/e$: lexical rules

(Wu, 1997)

ITG-based Reordering Model

- Type I: Incorporating ITG into left-to-right decoding to constrain the reordering space (e.g., Zens et al., 2004; Feng et al., 2010)
- Type II: Translation as ITG parsing, e.g.
  - Max-Ent ITG reordering model: using maximum entropy (MaxEnt) model to predict which rule to use (Xiong et al., 2006)

MaxEnt ITG Reordering Model

Potentially alleviates the data sparseness problem

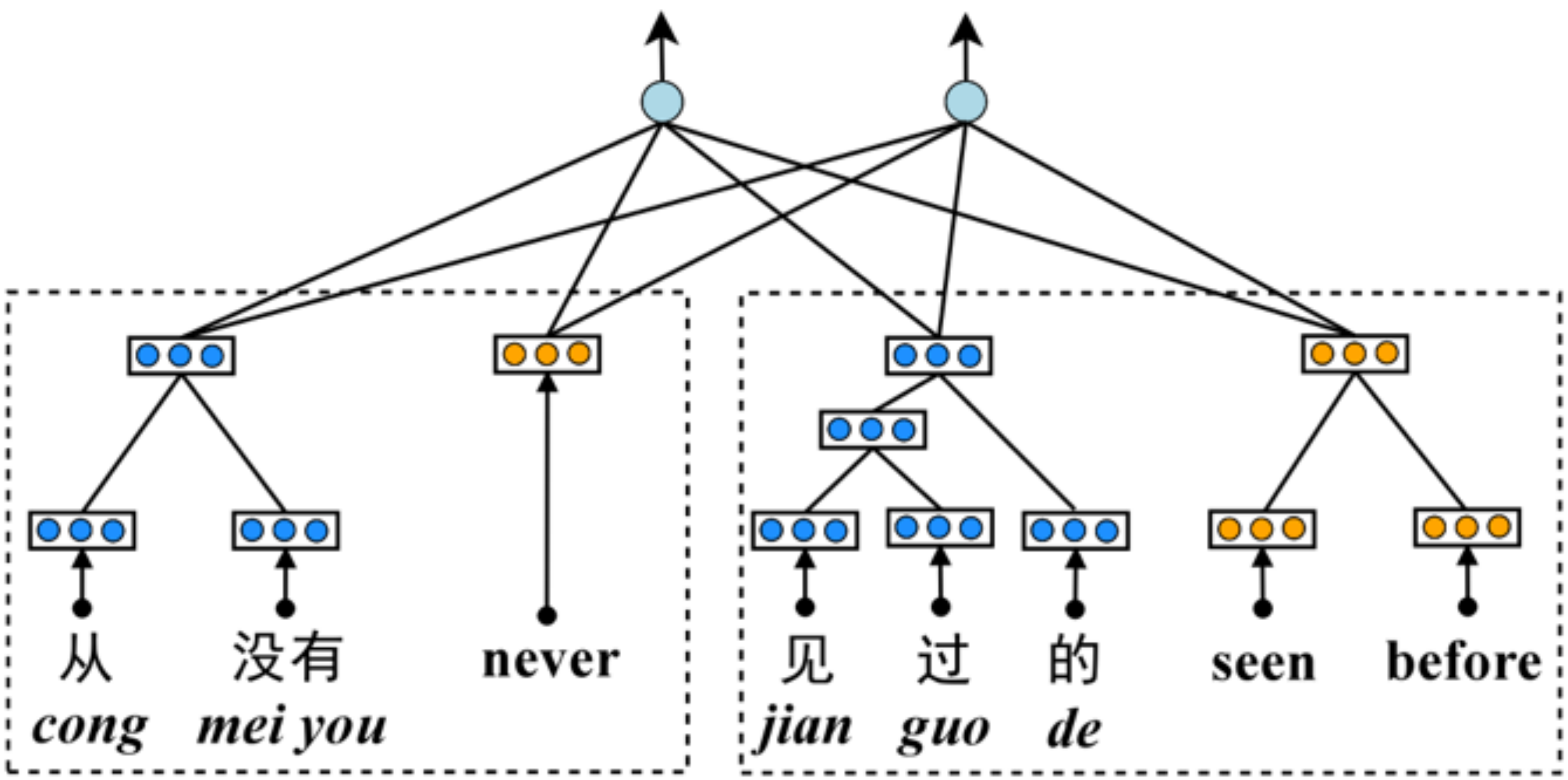
How to extract features from training examples?

- Which words are representative for predicting reordering?
- Xiong et al. (2006) only use boundary words

This Work

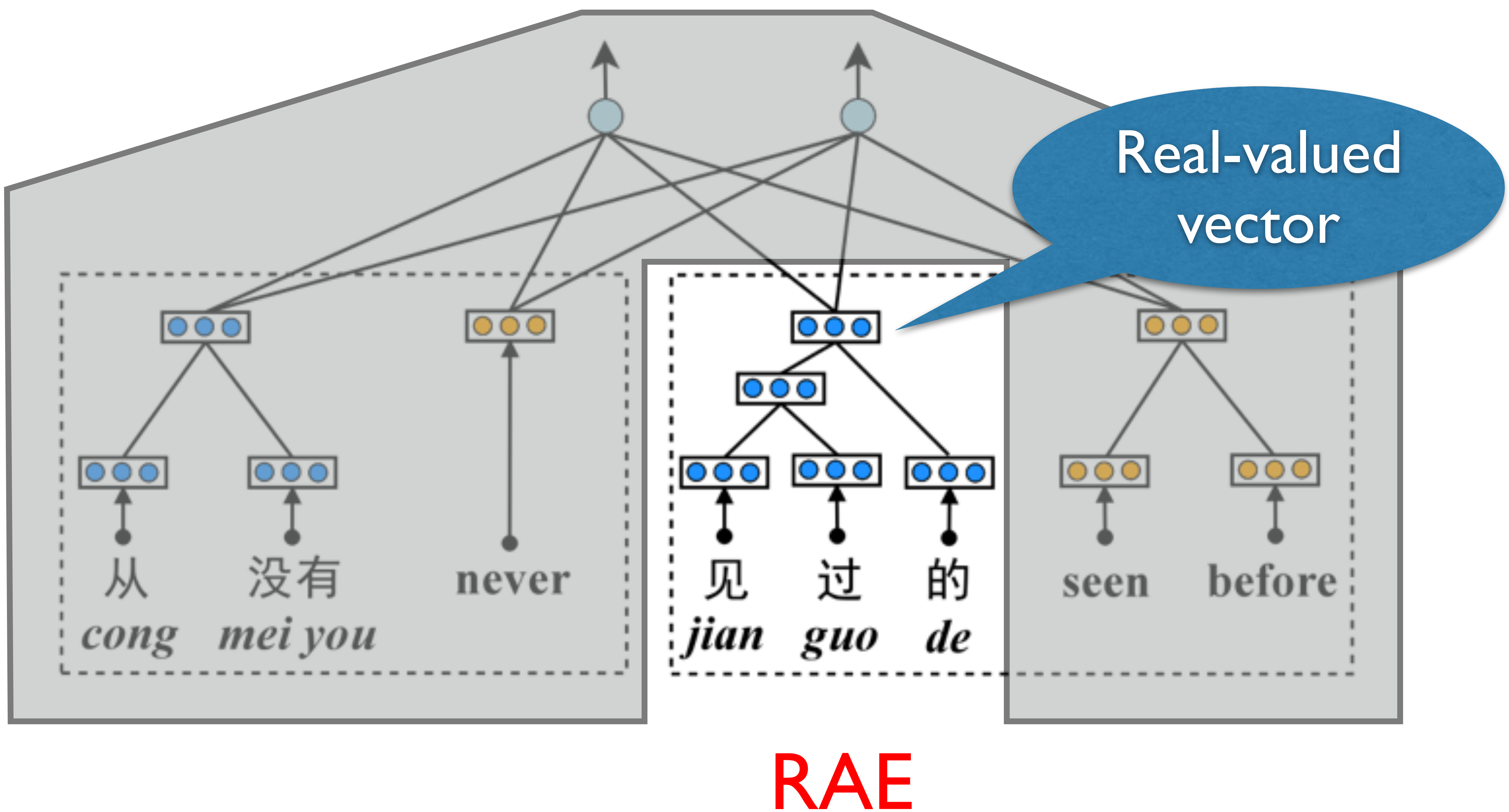
- We propose an ITG reordering classifier based on recursive autoencoders (RAE)
- Our model considers the whole phrases
  - RAEs can produce vector space representations for arbitrary strings
- Our system achieves 1.07 BLEU points improvement on NIST 2008 dataset

Neural ITG Reordering Model

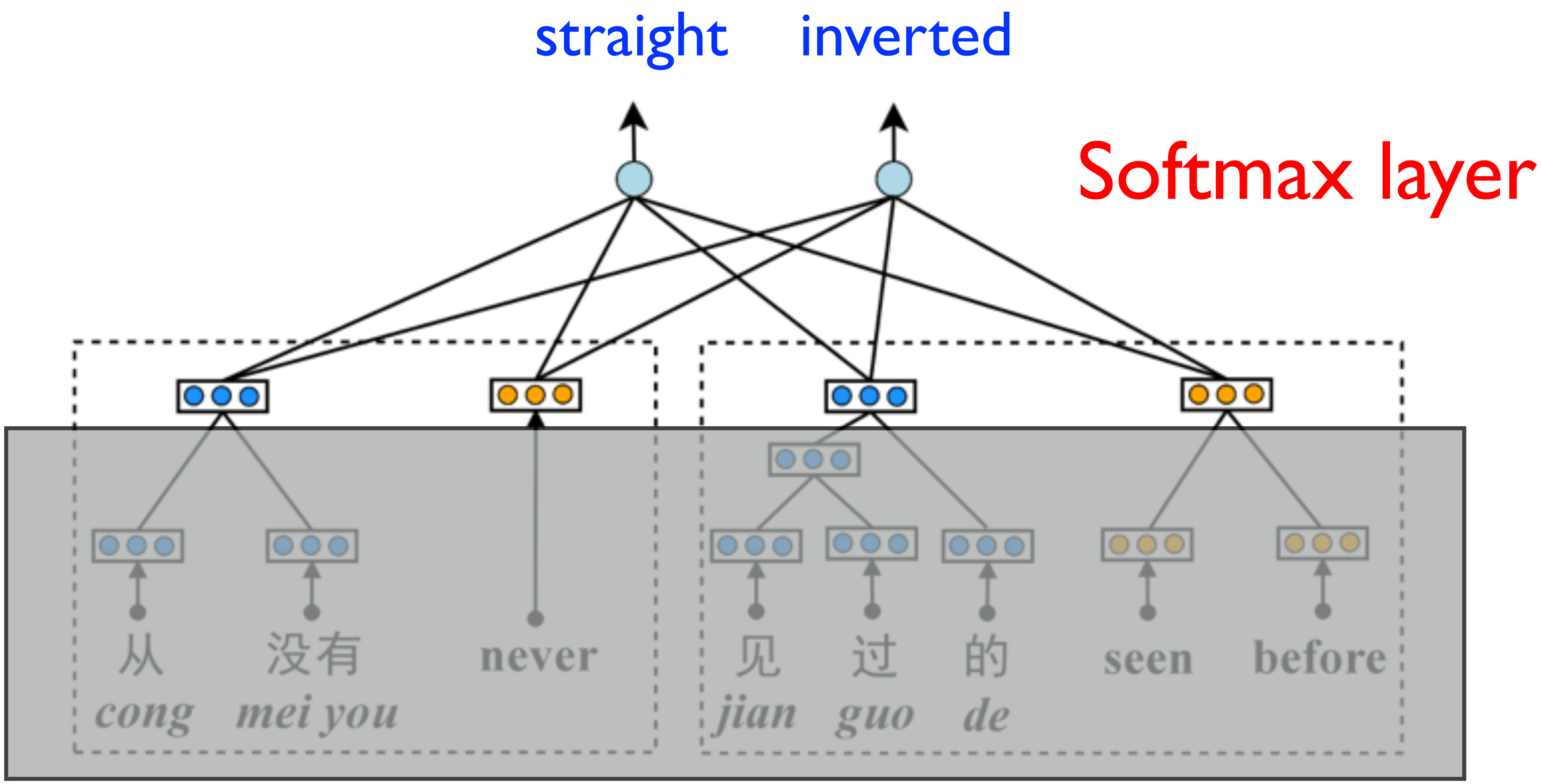


"never seen before" v.s. "seen before never"

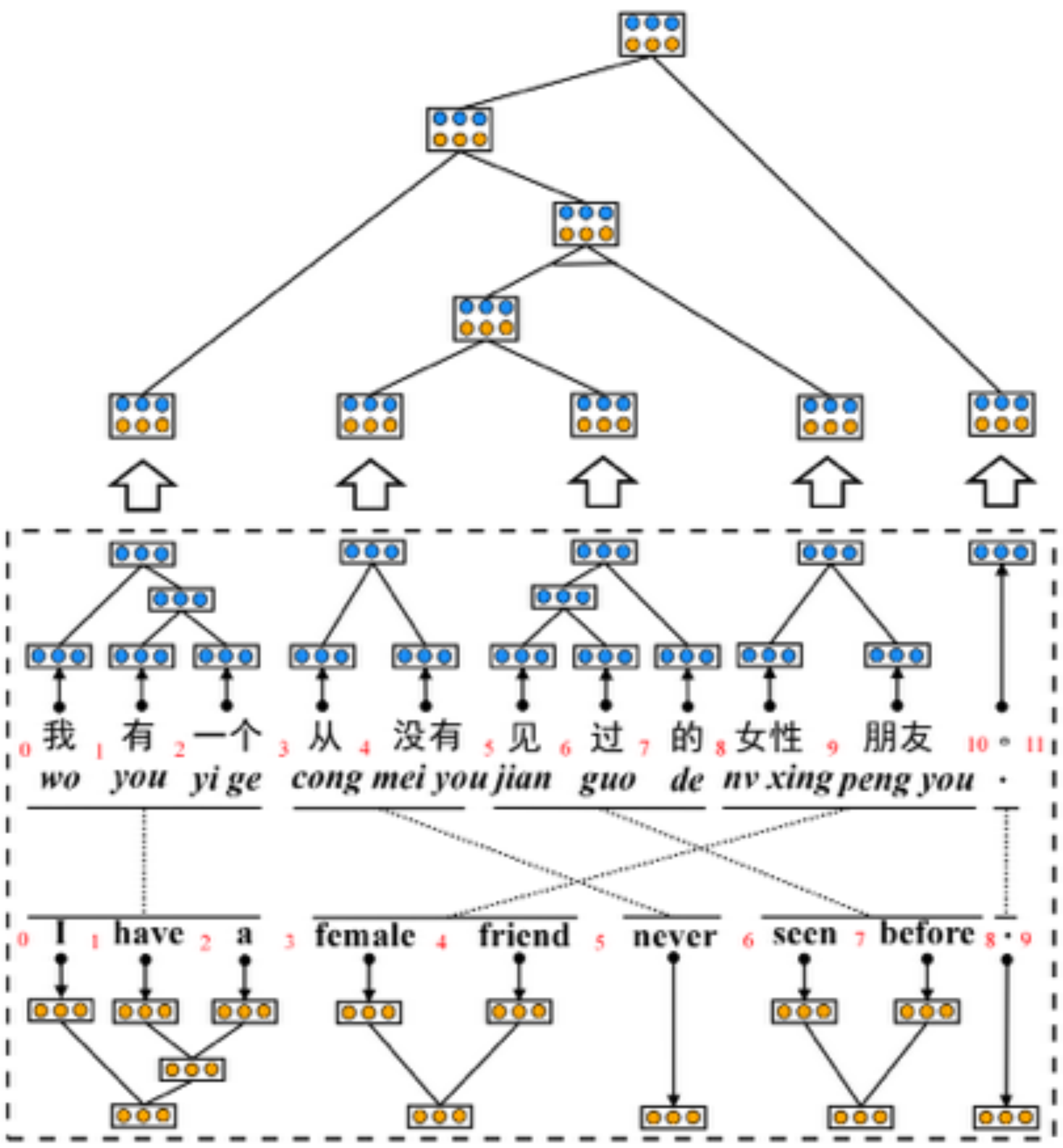
Neural ITG Reordering Model



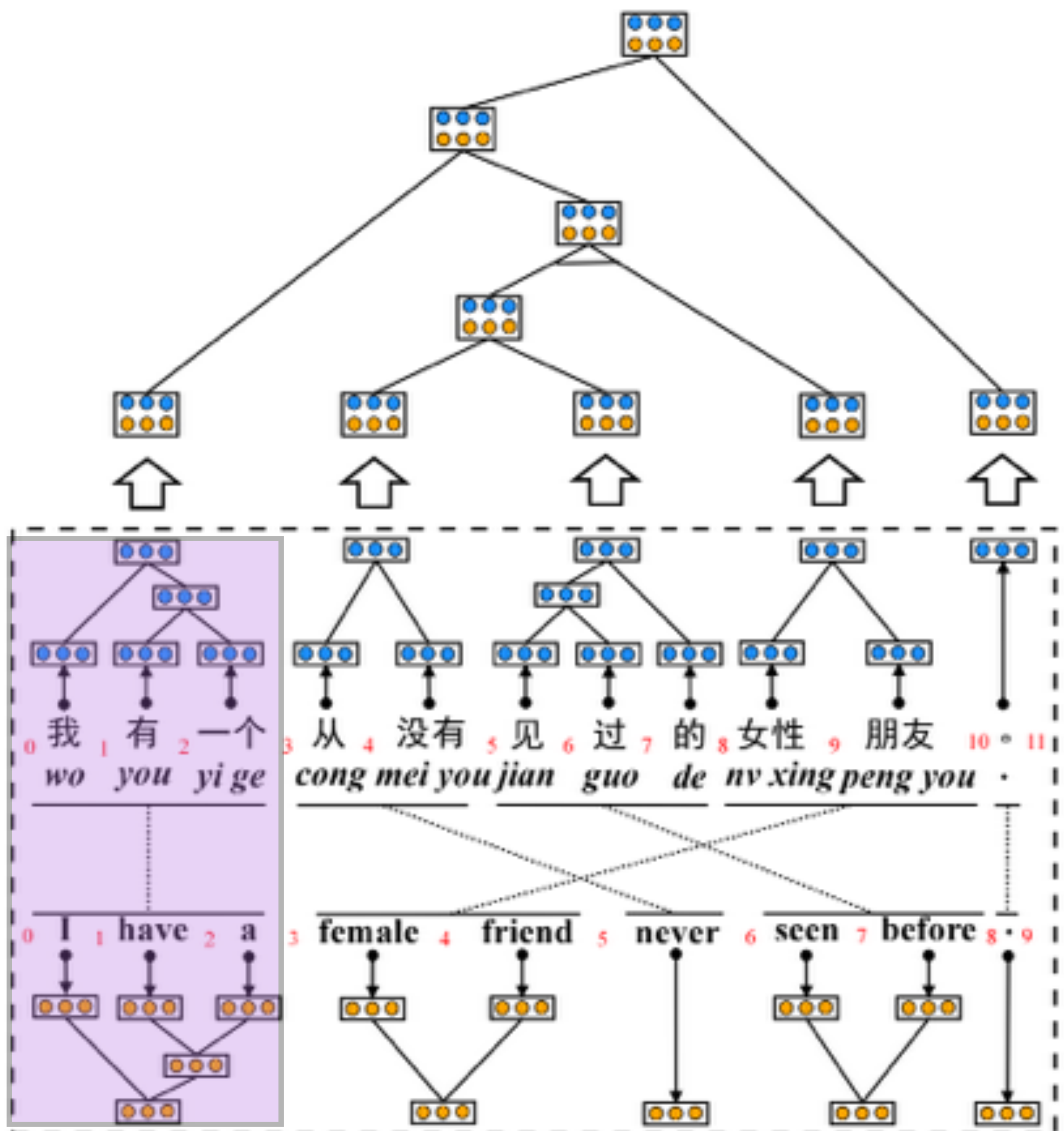
Neural ITG Reordering Model



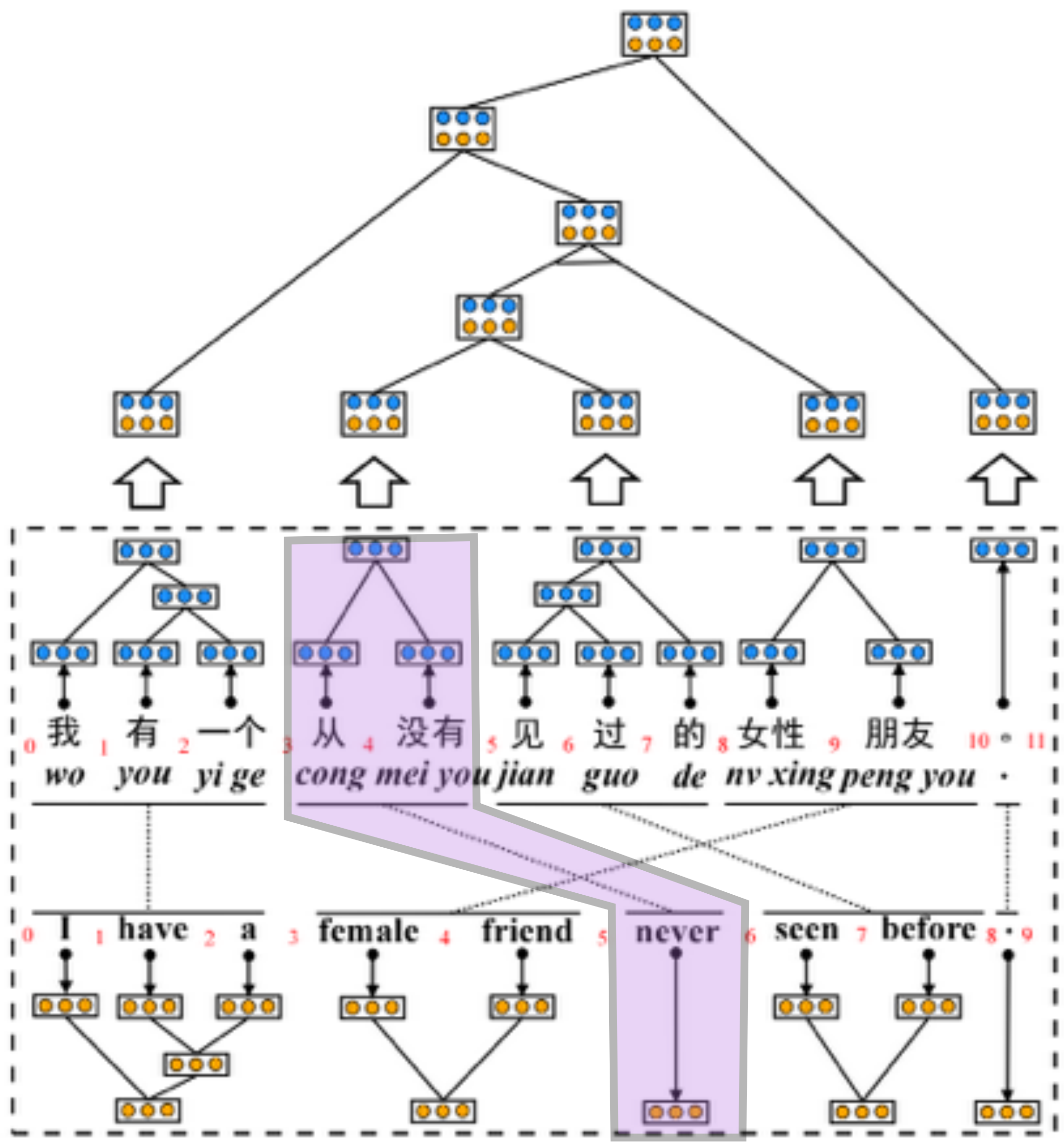
Translation



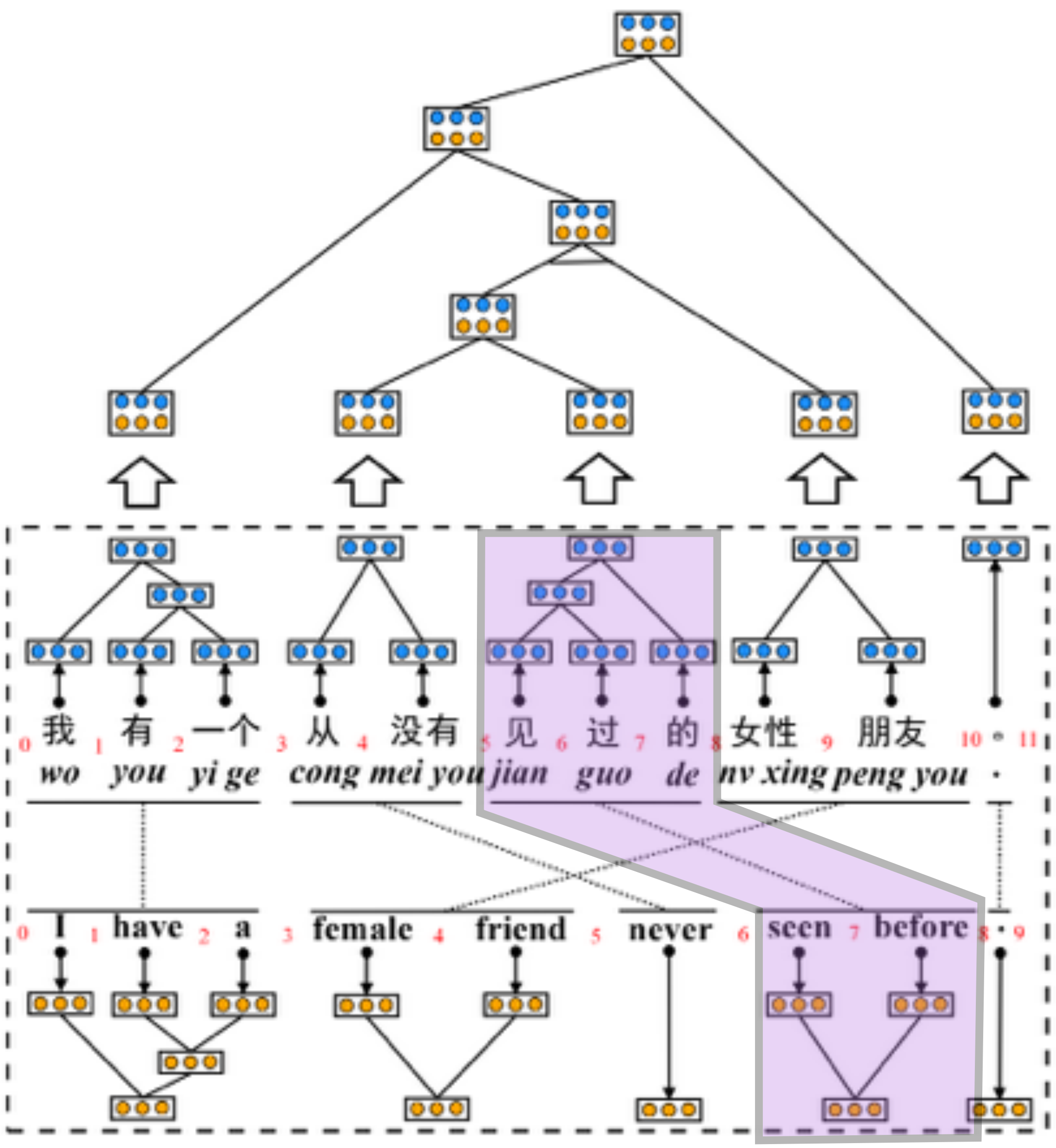
Translation



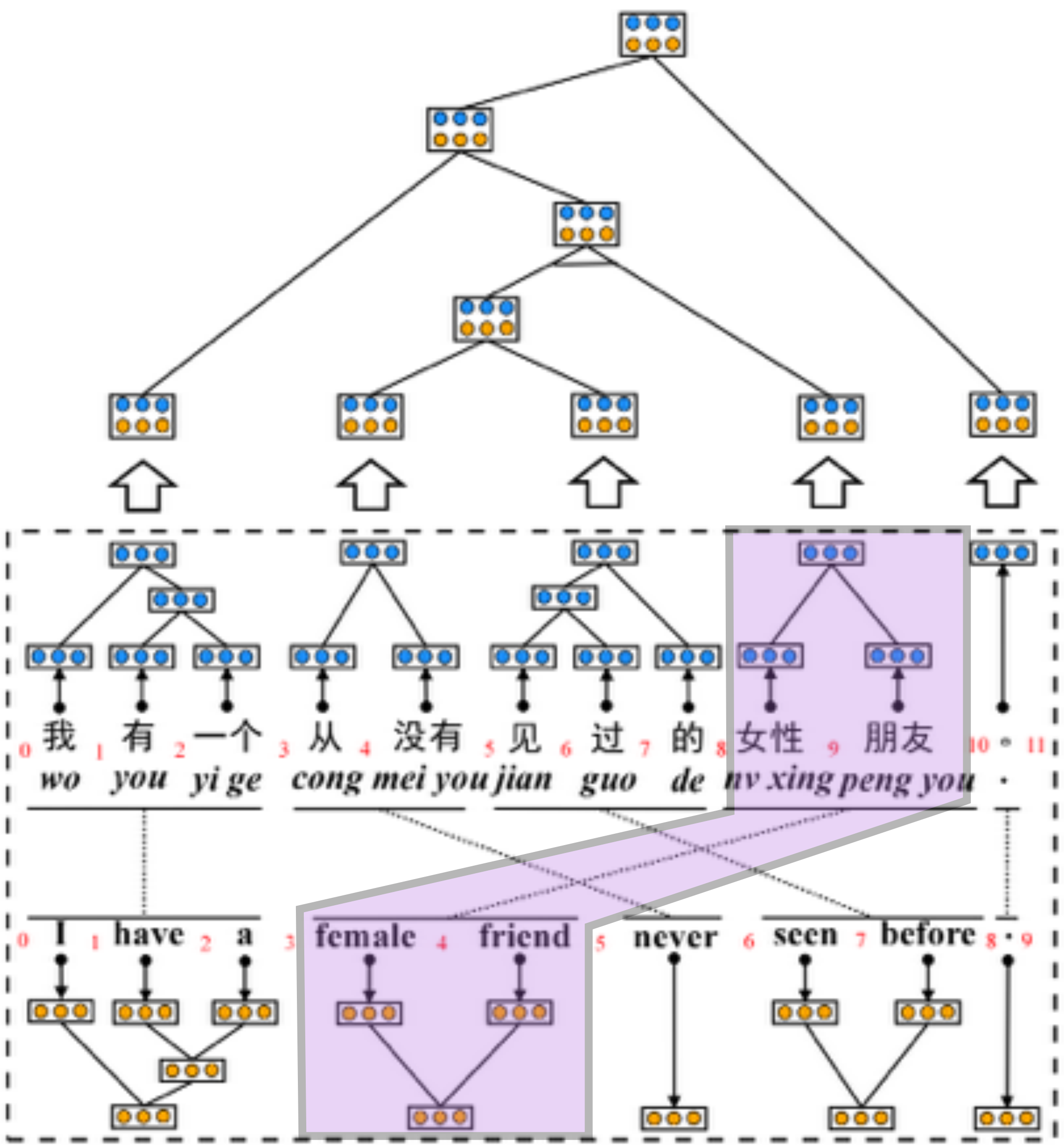
Translation



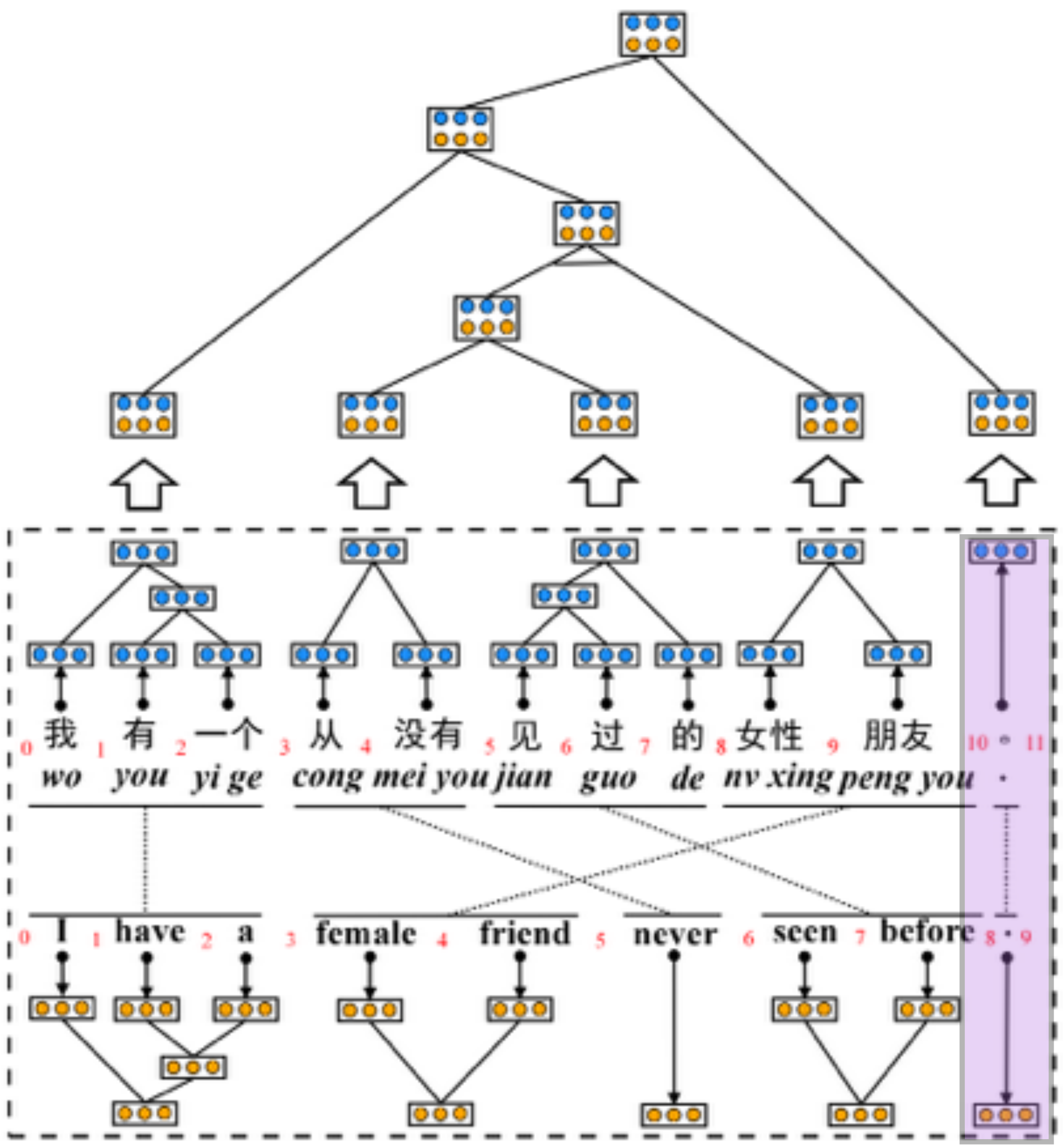
Translation



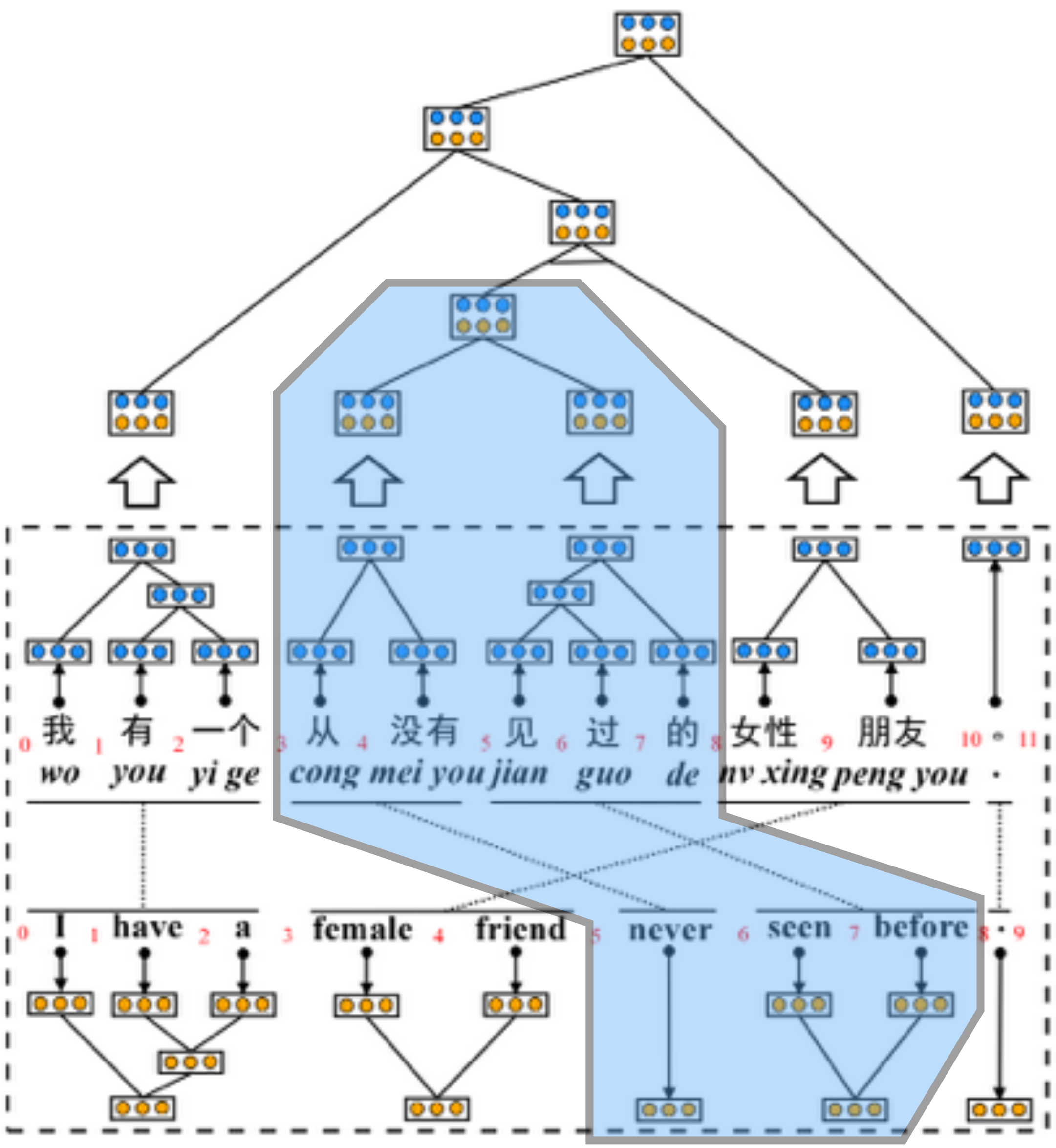
Translation



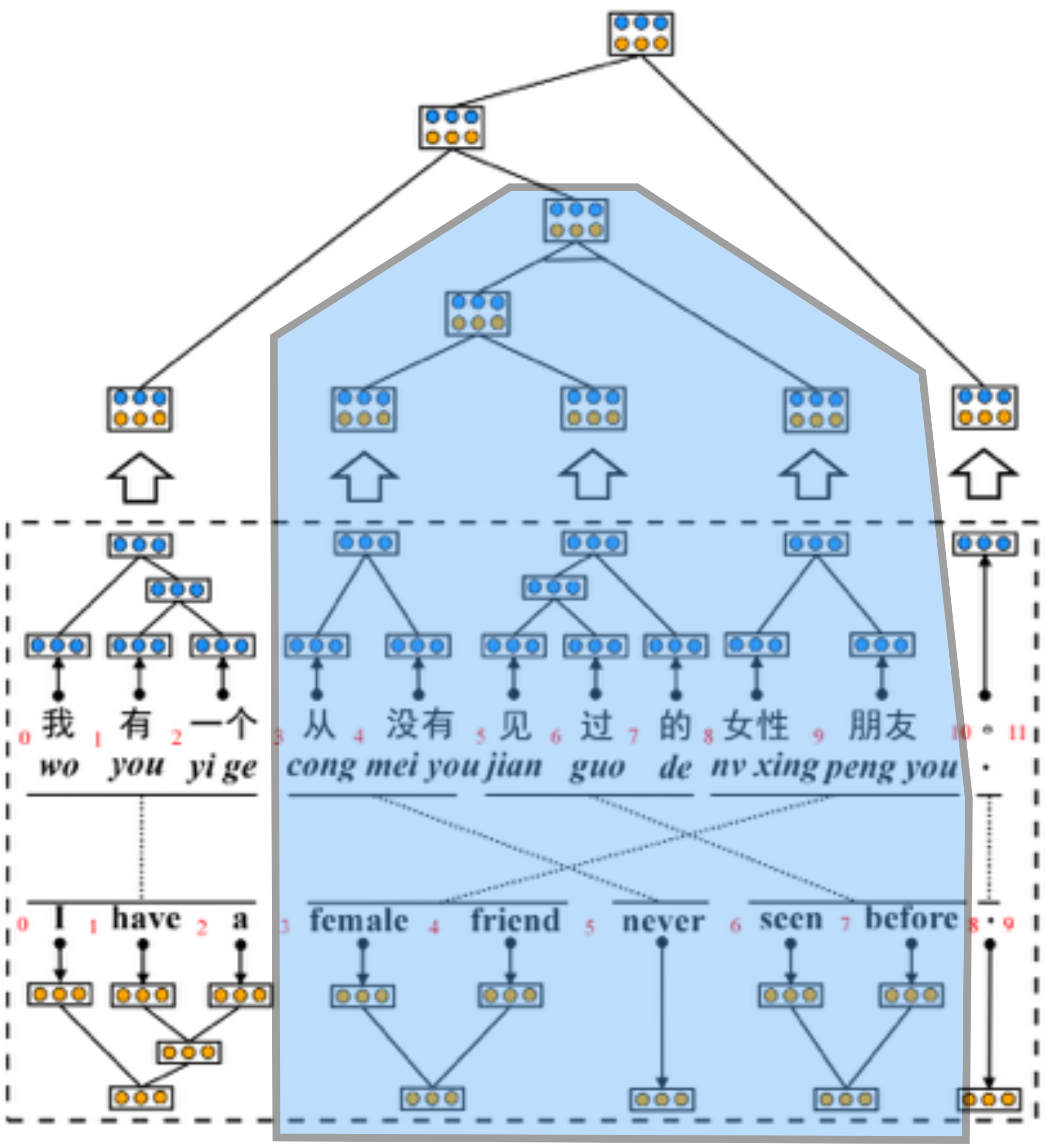
Translation



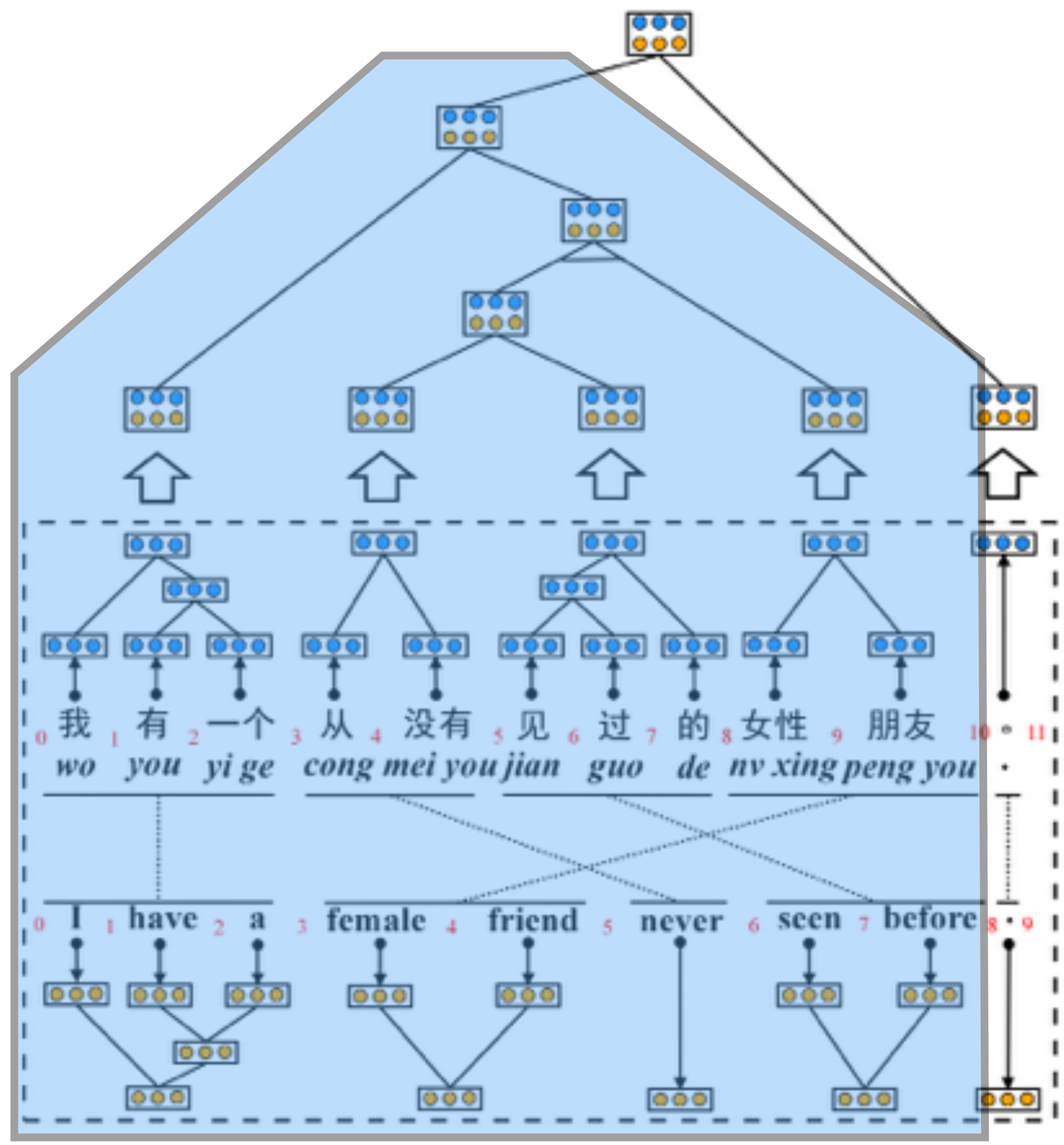
Translation



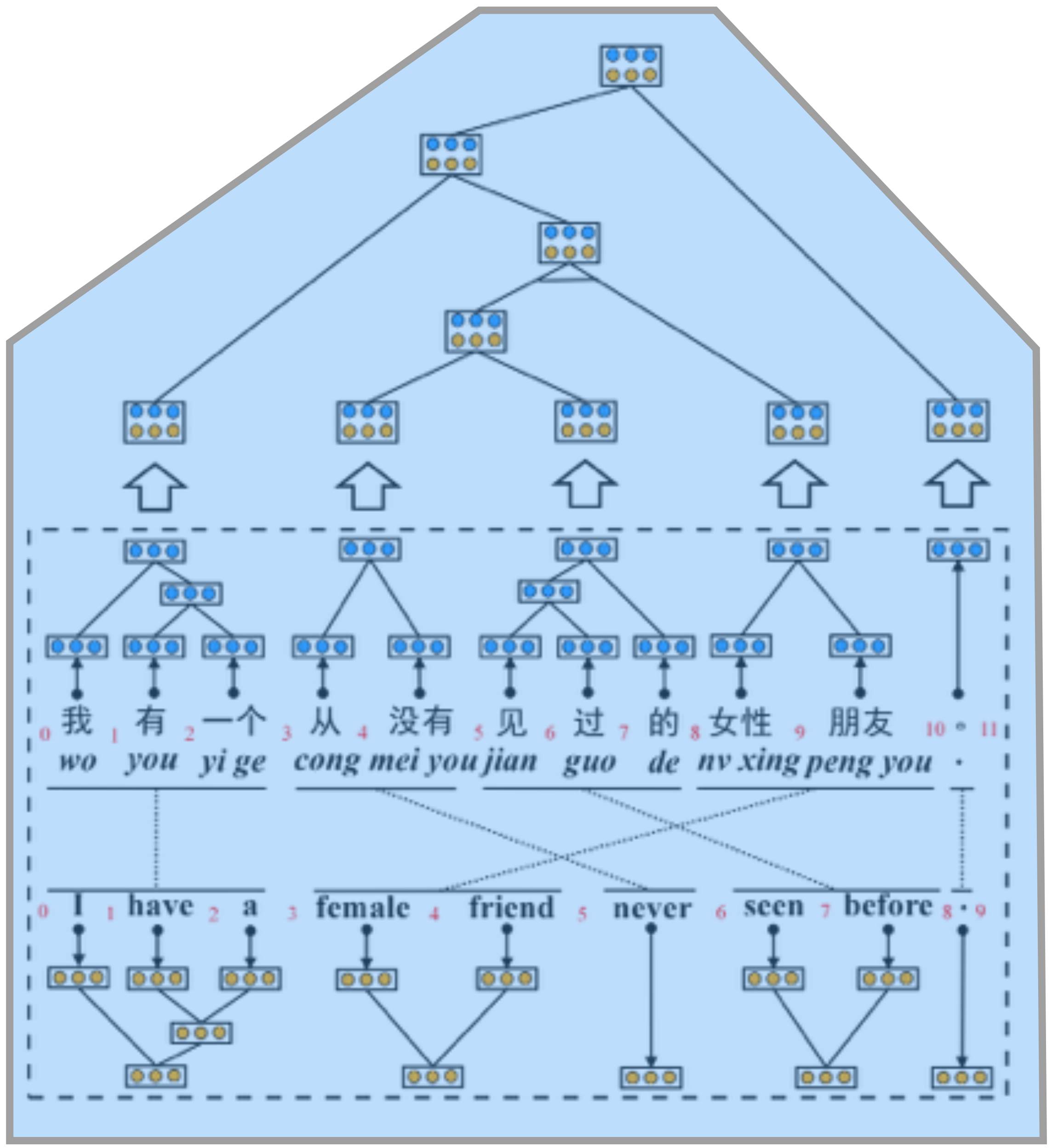
Translation



Translation



Translation

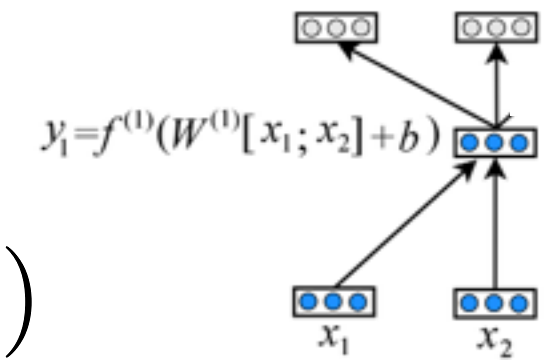


Autoencoders

- Each word is represented as a vector, e.g.
  - “female” ➤ $[0.1\ 0.8\ 0.4]^T$
  - “friend” ➤ $[0.7\ 0.1\ 0.5]^T$
- What is the vector representation of “female friend”?

Autoencoders

- Encoding



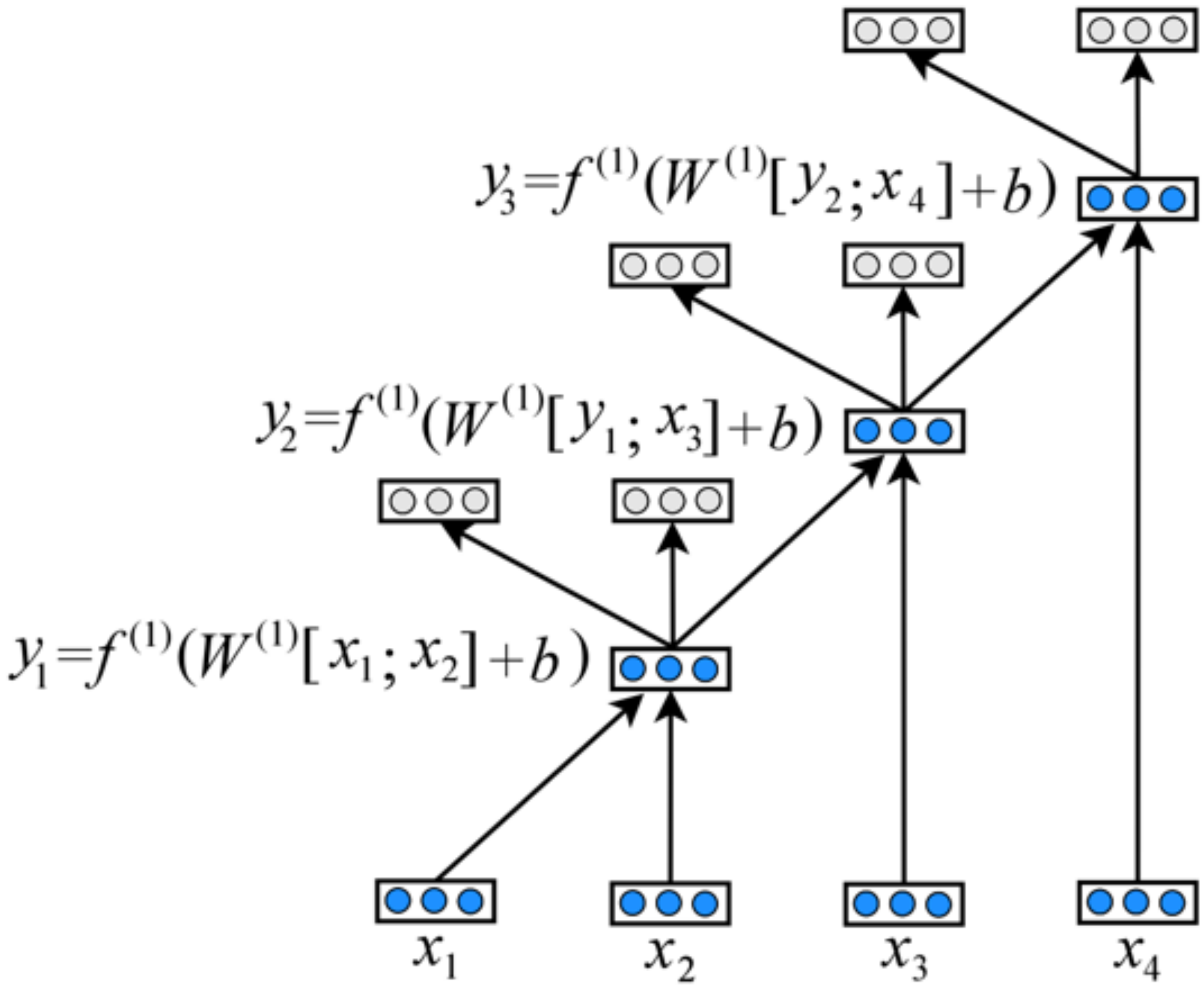
$$p = f^{(1)}(W^{(1)}[c_1; c_2] + b^{(1)})$$

- Decoding

$$[c'_1; c'_2] = f^{(2)}(W^{(2)}p + b^{(2)})$$

- What about multi-word strings?

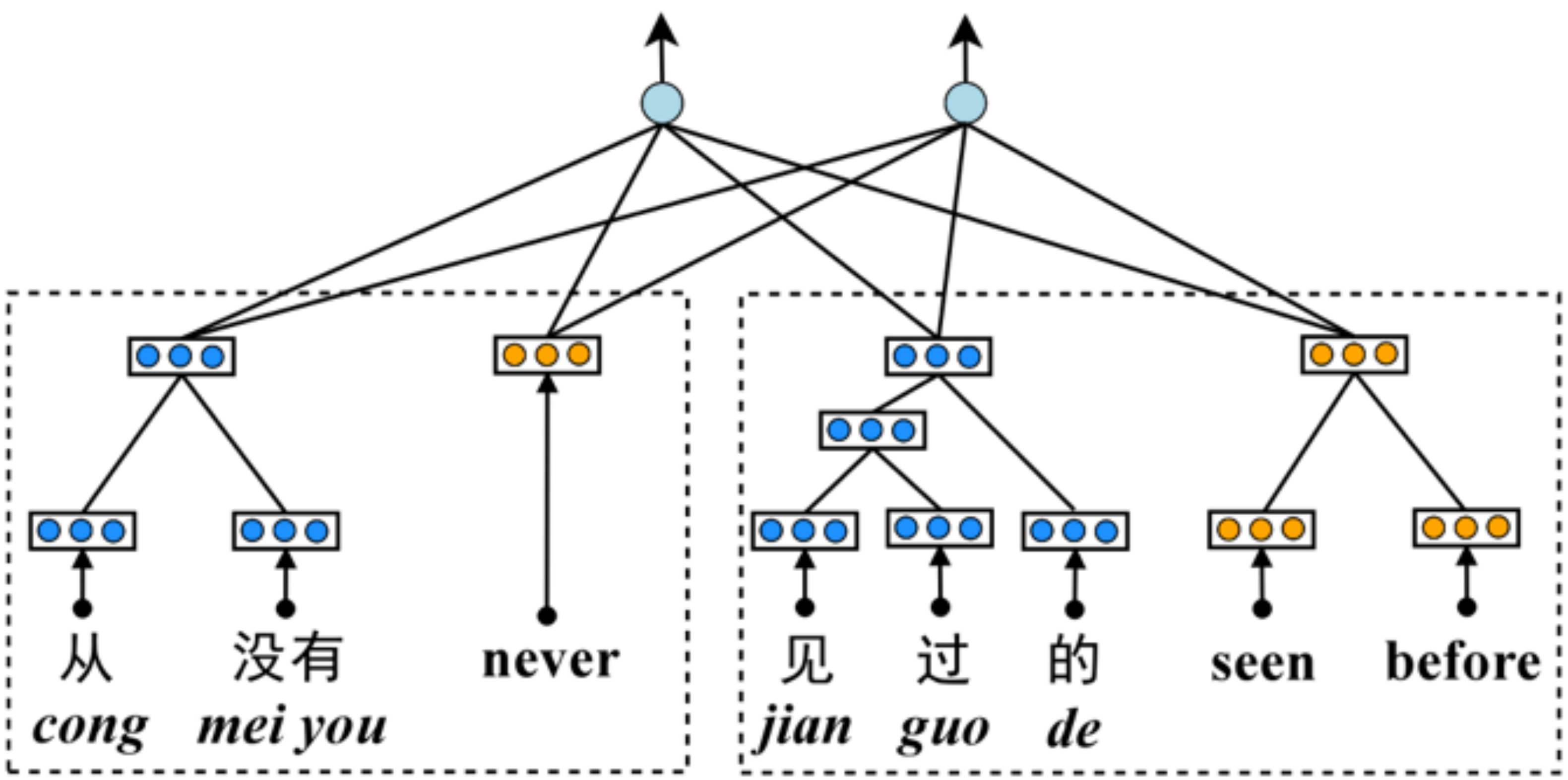
Recursive Autoencoders



(Socher et. al, 2011)

Training

Reordering error: how well the classifier predicts the merging order?



Reconstruction error: how well the learned vector space representations represent the corresponding strings?

Reconstruction Error

- Reconstruction error

$$E_{rec}([c_1; c_2]; \theta) = \frac{1}{2} || [c_1; c_2] - [c'_1; c'_2] ||^2$$

- Source side average reconstruction error

$$E_{rec,s}(S; \theta) = \frac{1}{N_s} \sum_i \sum_{p \in T_R^\theta(t_i, s)} E_{rec}([p.c_1, p.c_2]; \theta)$$

- Total reconstruction error

$$E_{rec}(S; \theta) = E_{rec,s}(S; \theta) + E_{rec,t}(S; \theta)$$

Reordering Error

- Average cross-entropy error

$$E_{reo}(S;\theta) = \frac{1}{|S|}\sum_i \left(-\sum_o d_{t_i}(o) \cdot log(P_\theta(o|t_i)) \right)$$

- Joint training objective

$$J = \alpha E_{rec}(S;\theta) + (1-\alpha)E_{reo}(S;\theta) + R(\theta)$$

$$R(\theta) = \frac{\lambda_L}{2}||\theta_L - \theta_{L_0}||^2 + \frac{\lambda_{rec}}{2}||\theta_{rec}||^2 + \frac{\lambda_{reo}}{2}||\theta_{reo}||^2$$

Optimization

- Hyper-parameters optimization
  - $\alpha, \lambda_L, \lambda_{rec}, \lambda_{reo}$
  - Optimized by random search (Bergstra and Bengio, 2012)
- Training objective optimization: L-BFGS
  - Using backpropagation through structures to compute gradients (Goller and Kuchler, 1996)

Experiments

- Training corpus: 1.23M sentence pairs
- Language model: 4-gram language model trained on the Xinhua portion of the GIGAWORD corpus
- Dev. set: NIST 2006 MT dataset
- Test set: NIST 2008 MT dataset
- Metric: case-insensitive BLEU-4 score

BLEU-4

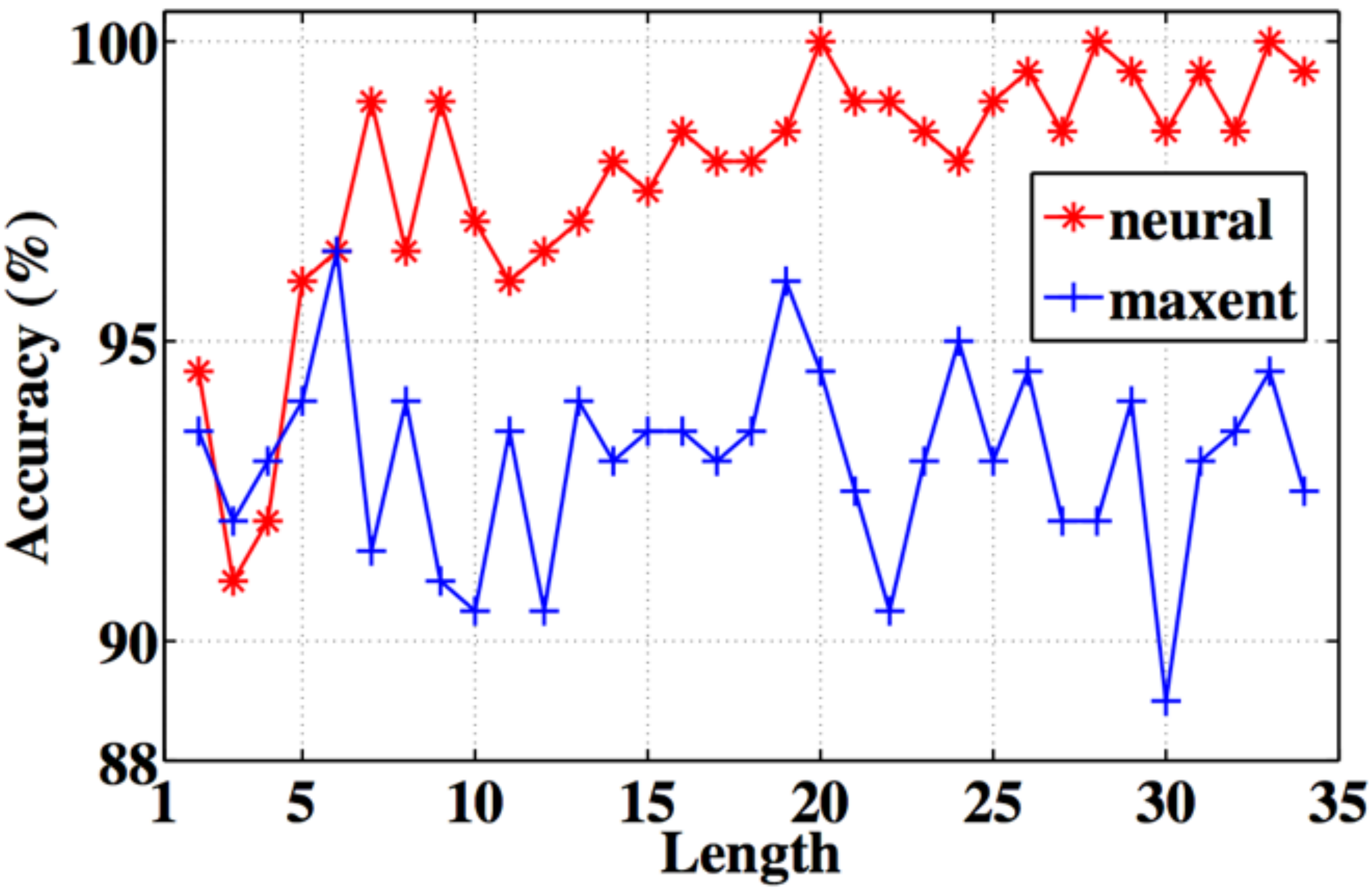
System	NIST06 (dev)	NIST08 (tst)
maxent	30.40	23.75
neural	**31.61***	**24.82***

*: significantly better ($p < 0.01$)

BLEU-4

Sentence Length	>	=	<
[1, 10]	43	121	57
[11, 20]	181	67	164
[21, 30]	170	11	152
[31, 40]	105	3	90
[41, 50]	69	1	53
[51, 119]	40	0	30

Classification Accuracy



Conclusion

- We have presented an ITG reordering classifier based on RAEs
- Feature work
  - Combine linguistically-motivated labels with recursive neural networks
  - Investigate more efficient decoding algorithms
  - Apply our method to other phrase-based and even syntax-based systems

Reference

- Yang Feng, Haitao Mi, Yang Liu, and Qun Liu. 2010. An efficient shift-reduce decoding algorithm for phrased- based machine translation. *In Proceedings of COLING 2010: Posters*, pp. 285–293.
- Christoph Goller and Andreas Kuchler. 1996. Learning task-dependent distributed representations by backpropagation through structure. In *Proceedings of IJCNN 1996*, pp. 347–352.
- Liang Huang, Hao Zhang, Daniel Gildea, Kevin Knight. Binarization of synchronous context-free grammars. *Computational Linguistics*, 35(4), pp. 559–595.
- Kevin Knight. 1999. Decoding complexity in word- replacement translation models. *Computational Linguistics*, 25(4):607–615.

Reference

- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL 2007*, pp. 177–180.
- Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pp. 48–54.
- Franz Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Reference

- Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of EMNLP 2011*, pp. 151–161.
- Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.
- Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of COLING/ACL 2006*, pp. 521–528.
- Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. 2004. Reordering constraints for phrase-based statistical machine translation. In *Proceedings of COLING 2004*, pp. 205–211.

Thanks!

Backup Slides

Training Data Size

# of examples	NIST06 (dev)	NIST08 (tst)
100,000	30.88	23.78
200,000	30.75	23.89
300,000	30.80	24.35
400,000	31.01	24.45
6,004,441	31.61	24.82

Cluster Examples

Cluster 1	Cluster 2	Cluster 3
works for	these people who	of the three
verify on	the reasons why	on the fundamental
tunnels from	the story of how	over the entire
transparency in	the system which	through its own
opinion at	the trend towards	with the best