A Neural Reordering Model for Phrase-based Translation

Peng Li
Tsinghua University
pengli09@gmail.com

joint work with Yang Liu, Maosong Sun, Tatsuya Izuha, Dakun Zhang

Phrase-based Translation

布什　　与　　沙龙　　举行　　了　　会谈

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

Phrase-based Translation

布什	与　沙龙	举行　了　会谈

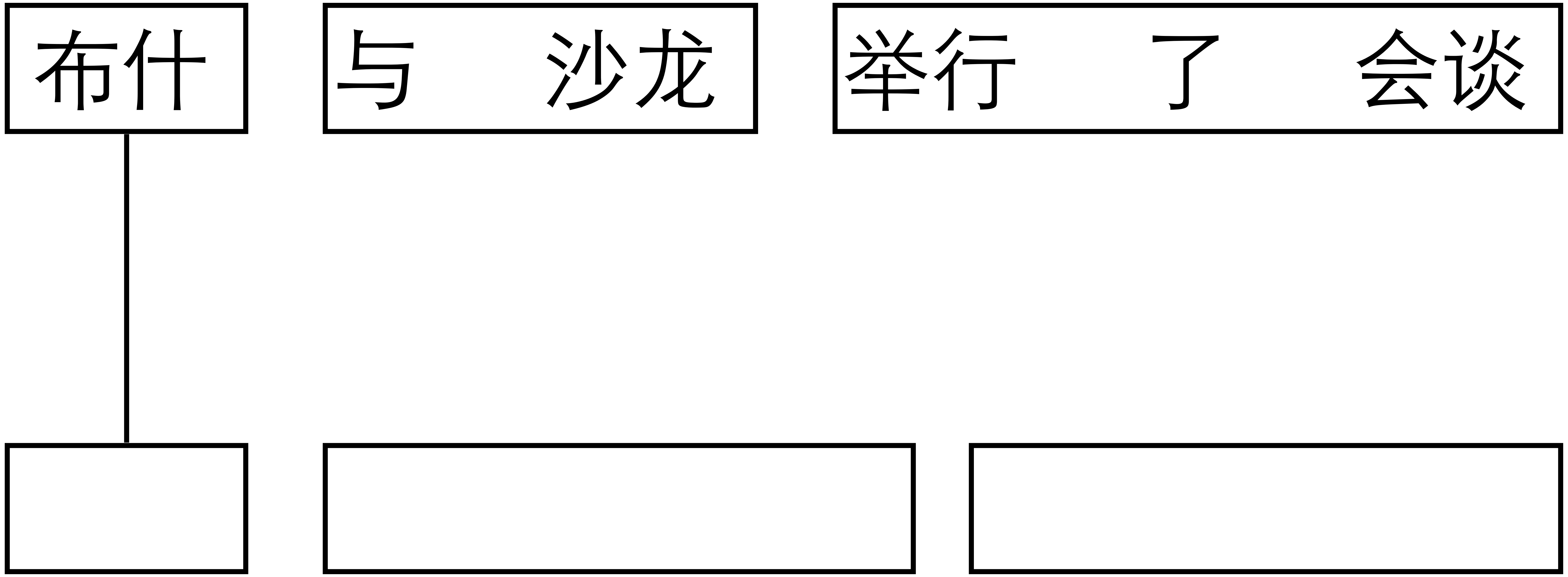
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Translation

布什　与　沙龙　举行　了　会谈

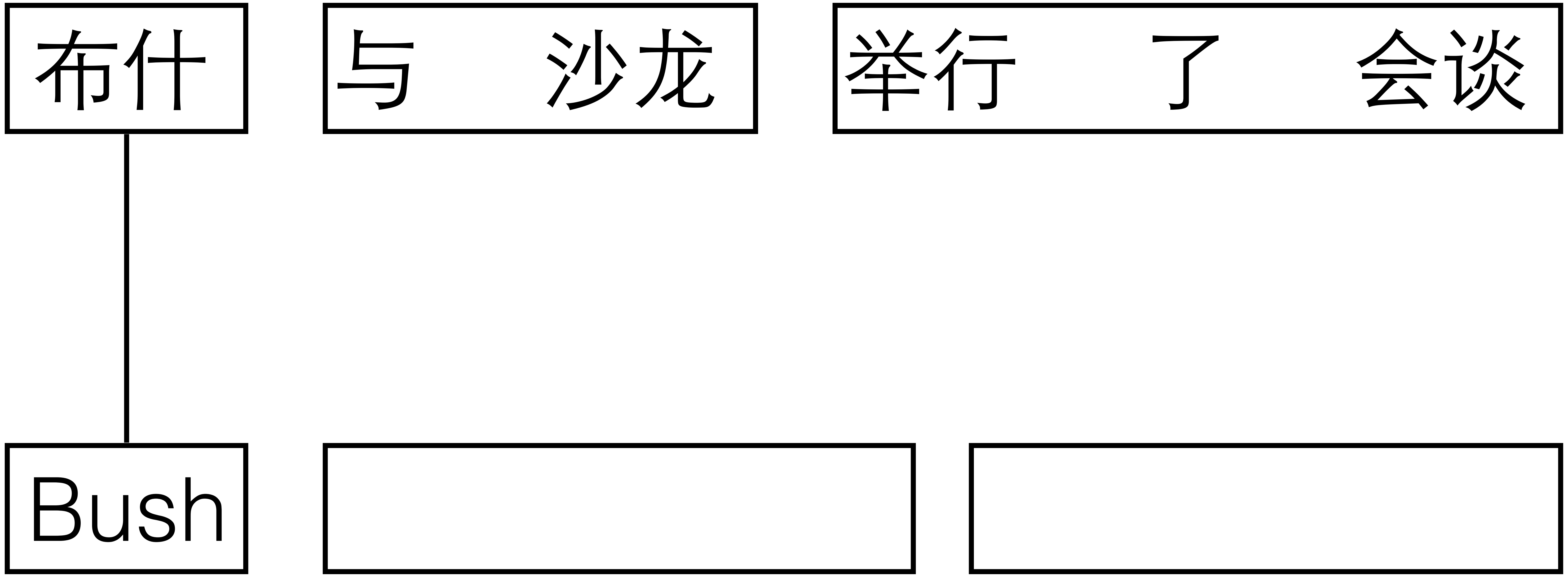
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Translation



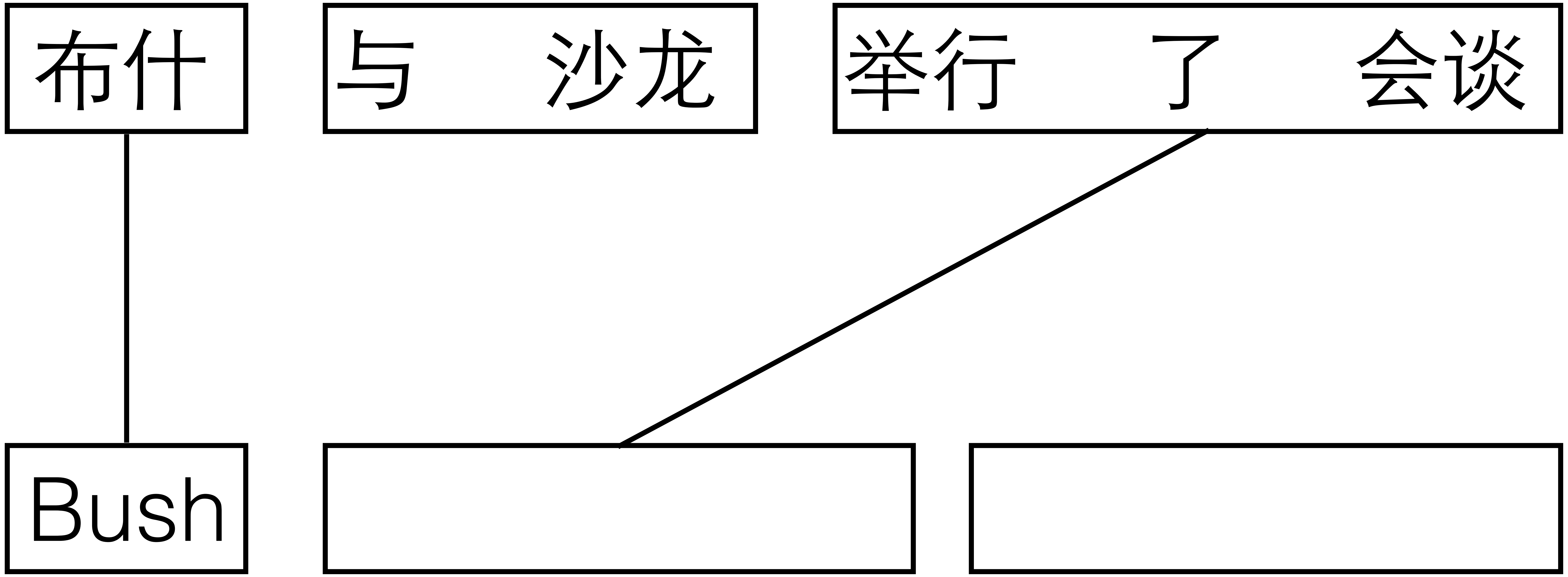
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Translation



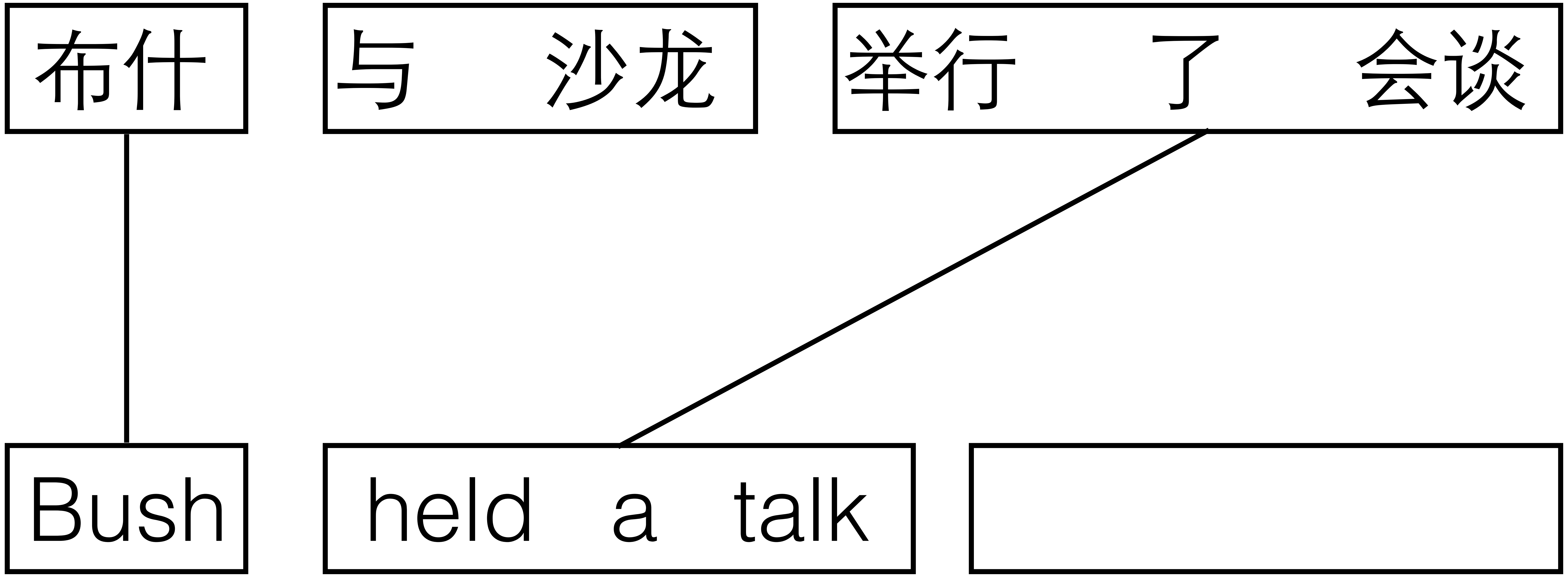
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Translation



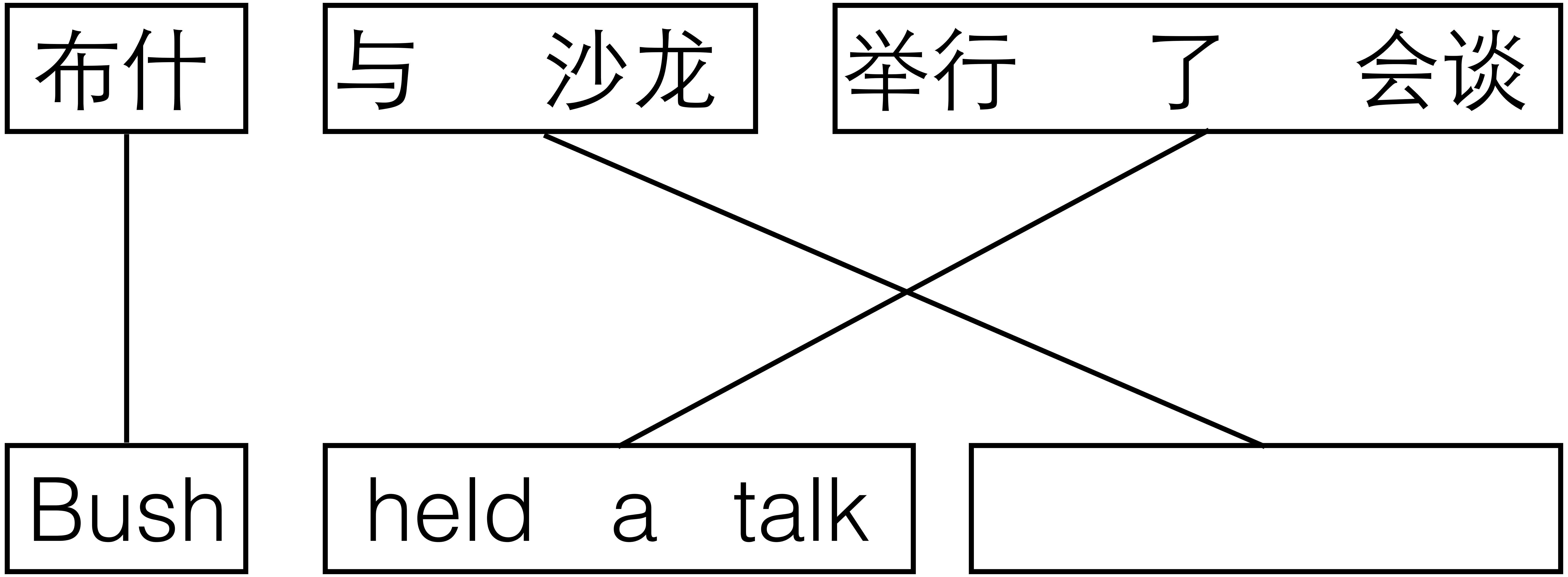
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Translation



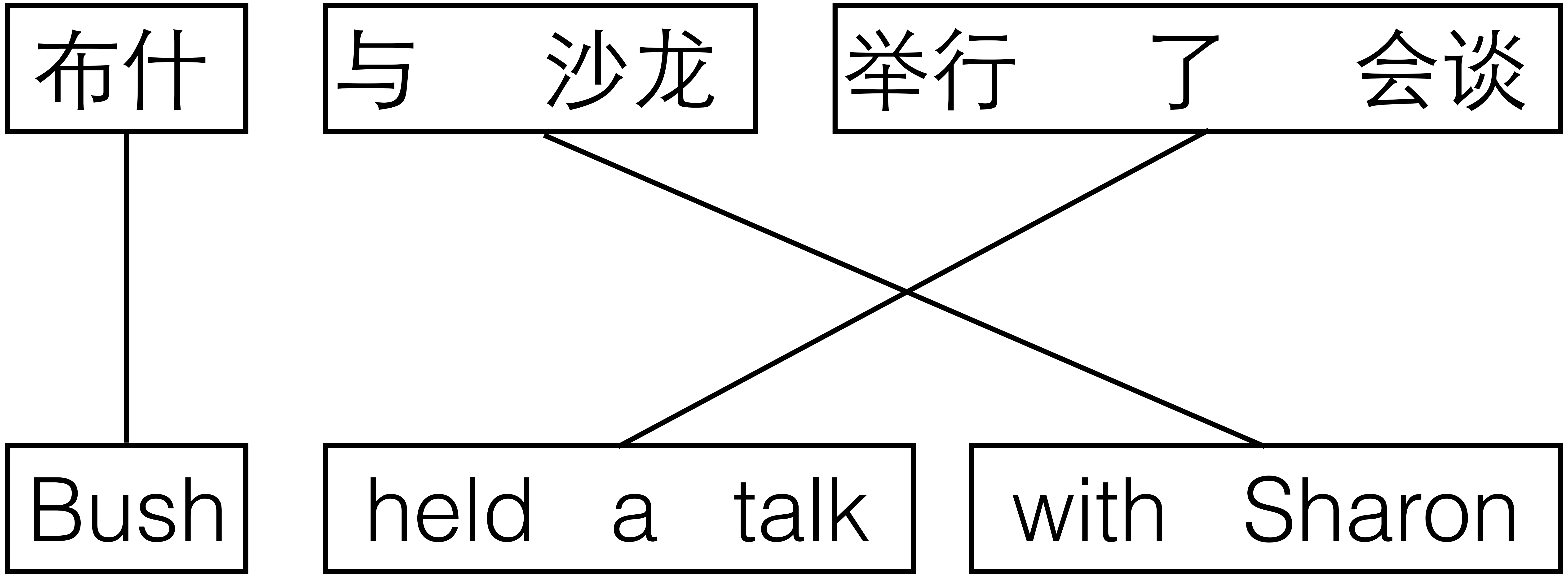
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Translation



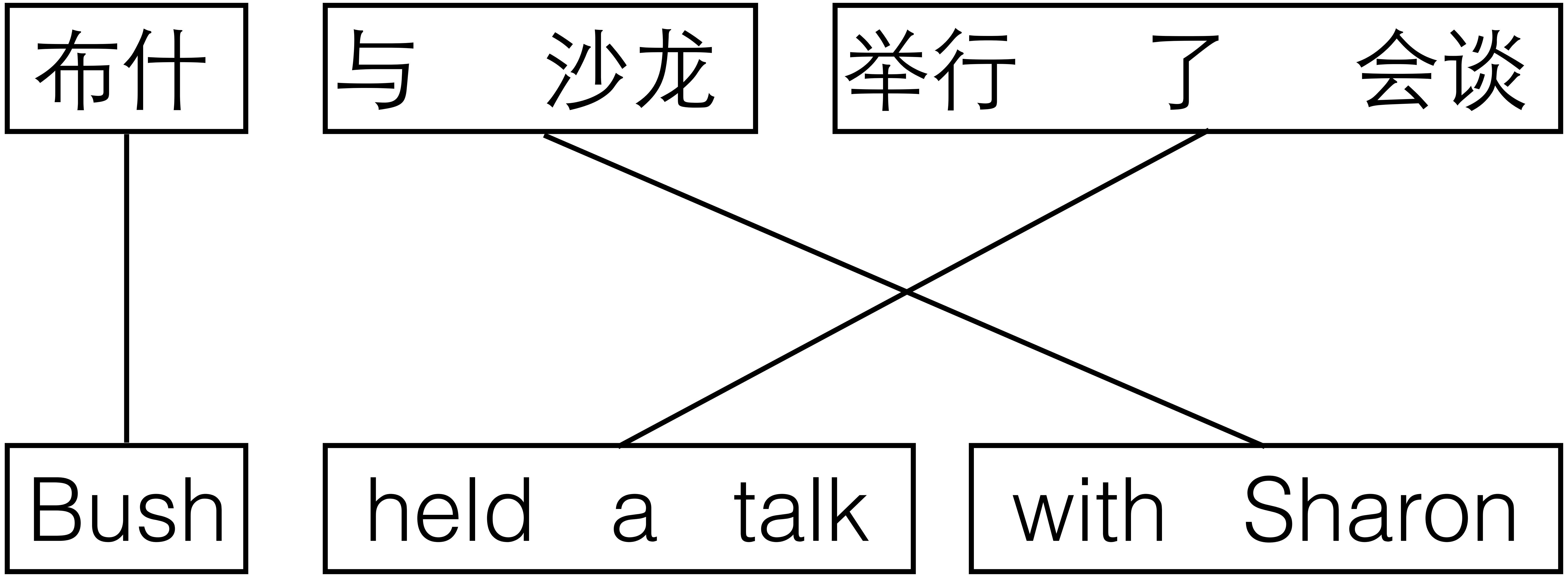
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Translation



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

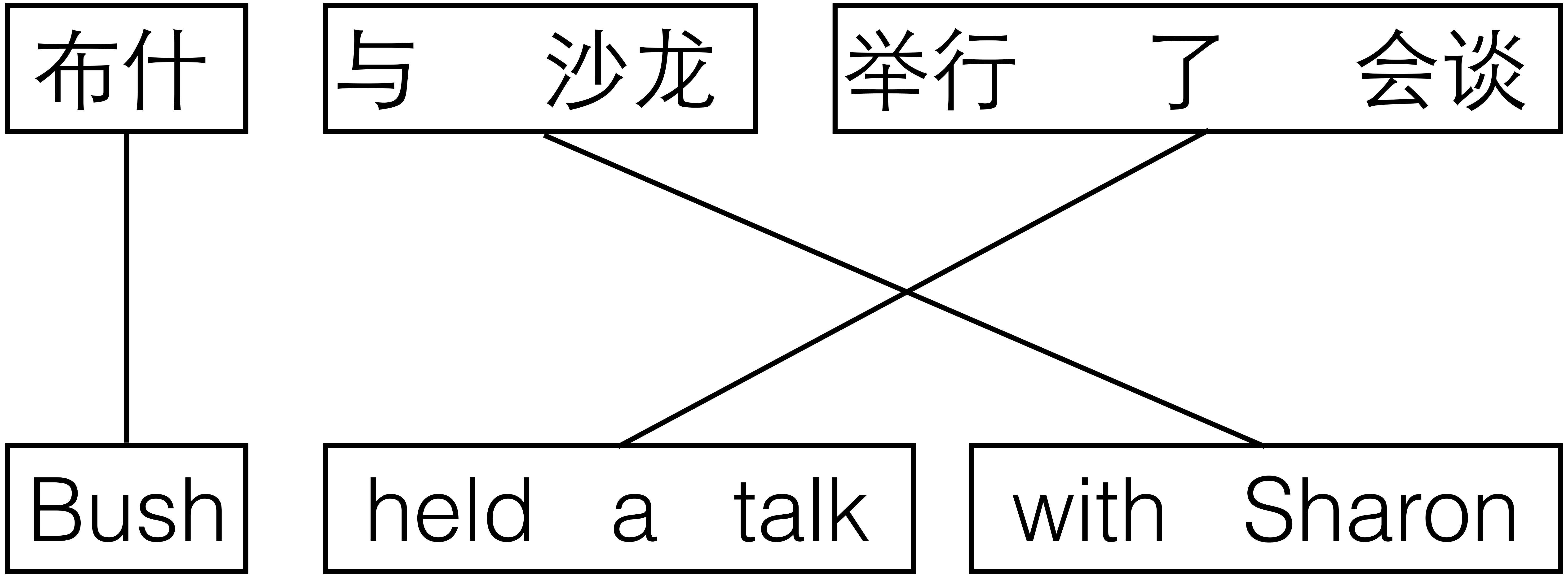
Phrase-based Translation



segmentation

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

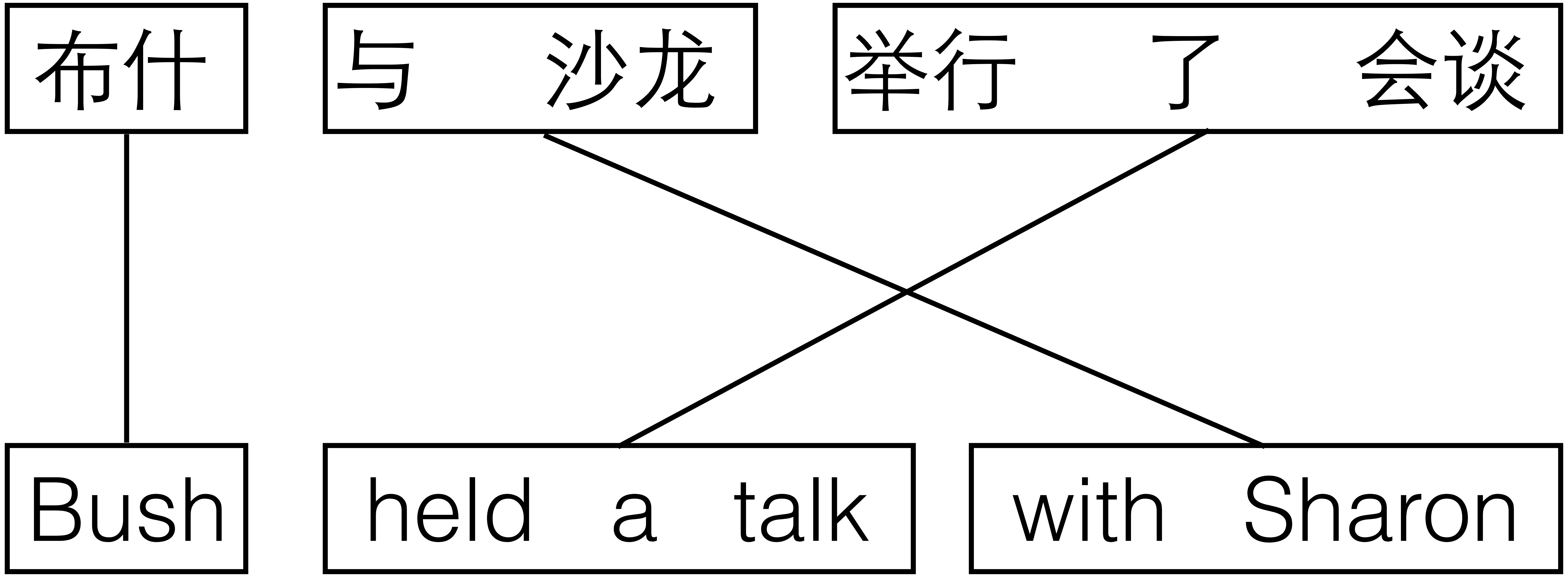
Phrase-based Translation



segmentation reordering

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

Phrase-based Translation



segmentation reordering translation

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

Reordering is Hard

high-stakes of issues his

and

counterpart Chinese talks Barack

in

President Us number

days Obama

a

Jinping open Xi

California of two on

Reordering is Hard

high-stakes
of
issues
his
and
talks
Barack
counterpart
Chinese
in
President
Us
number
days
Obama
a
Jinping
open
Xi
California
of
two
on

Q: Can you figure out a sentence using these words?

Reordering is Hard

Chinese President Xi Jinping and his Us counterpart

Barack Obama open two days of talks in California

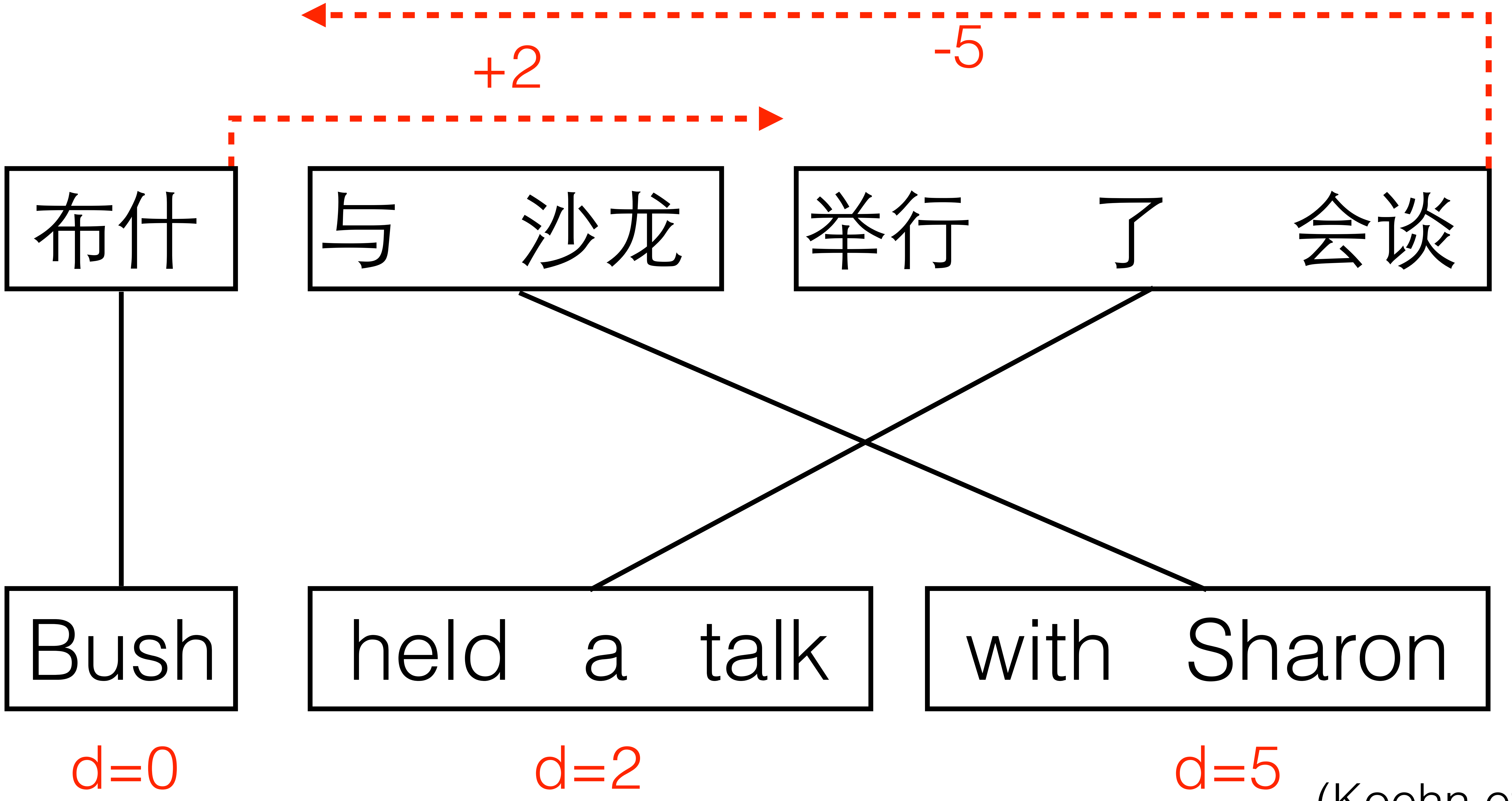
on a number of high-stakes issues

Q: Can you figure out a sentence using these words?

Reordering is Hard

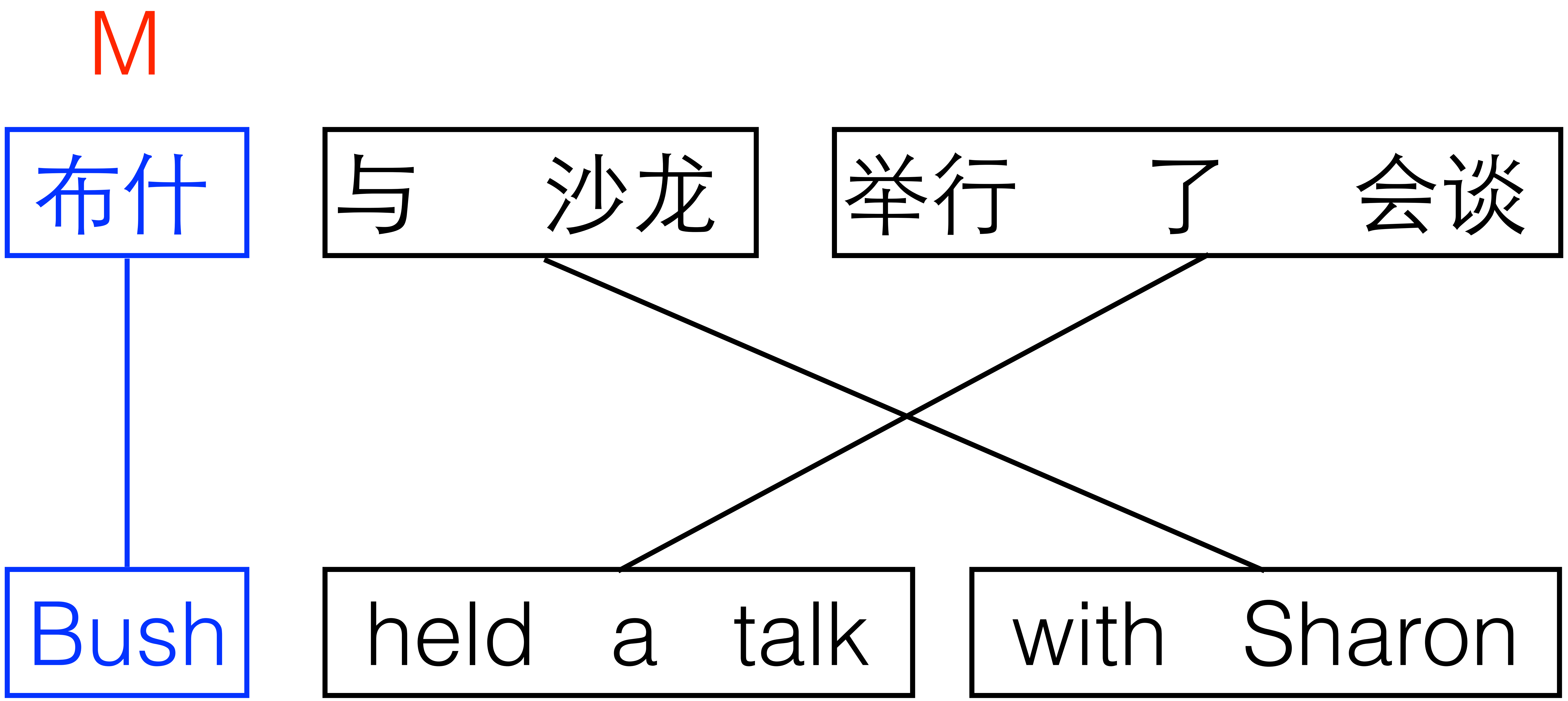
- An NP-complete problem (Knight, 1999; Zaslavskiy et al., 2009)
- Reordering modeling has attracted intensive attention, e.g.
    - Distance-based model (Koehn et al., 2003)
    - Word-based lexicalized model (Koehn et al., 2007)
    - Phrase-based lexicalized model (Tillman, 2004)
    - Hierarchical phrase-based lexicalized model (Galley and Manning, 2008)

Distance-based Model



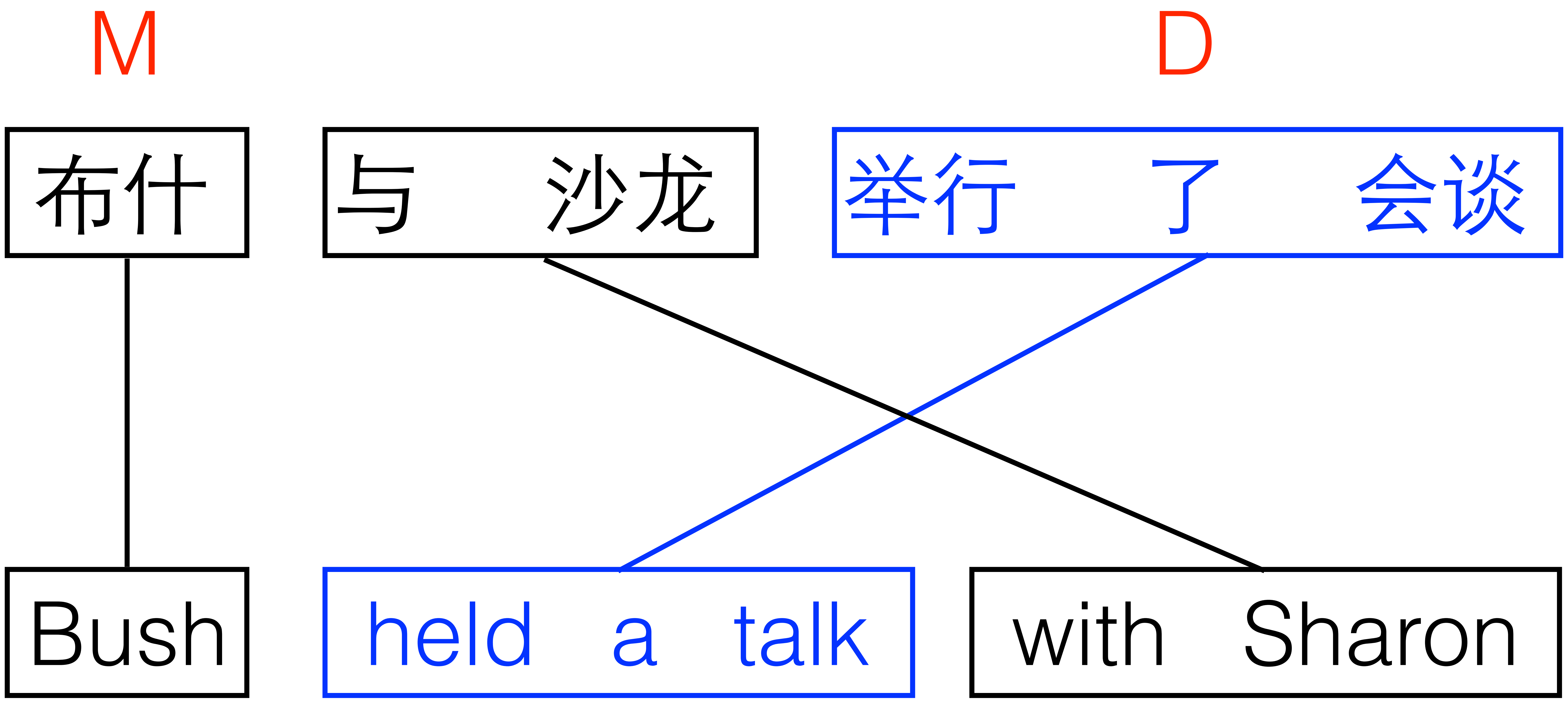
(Koehn et al., 2003)

Lexicalized Models



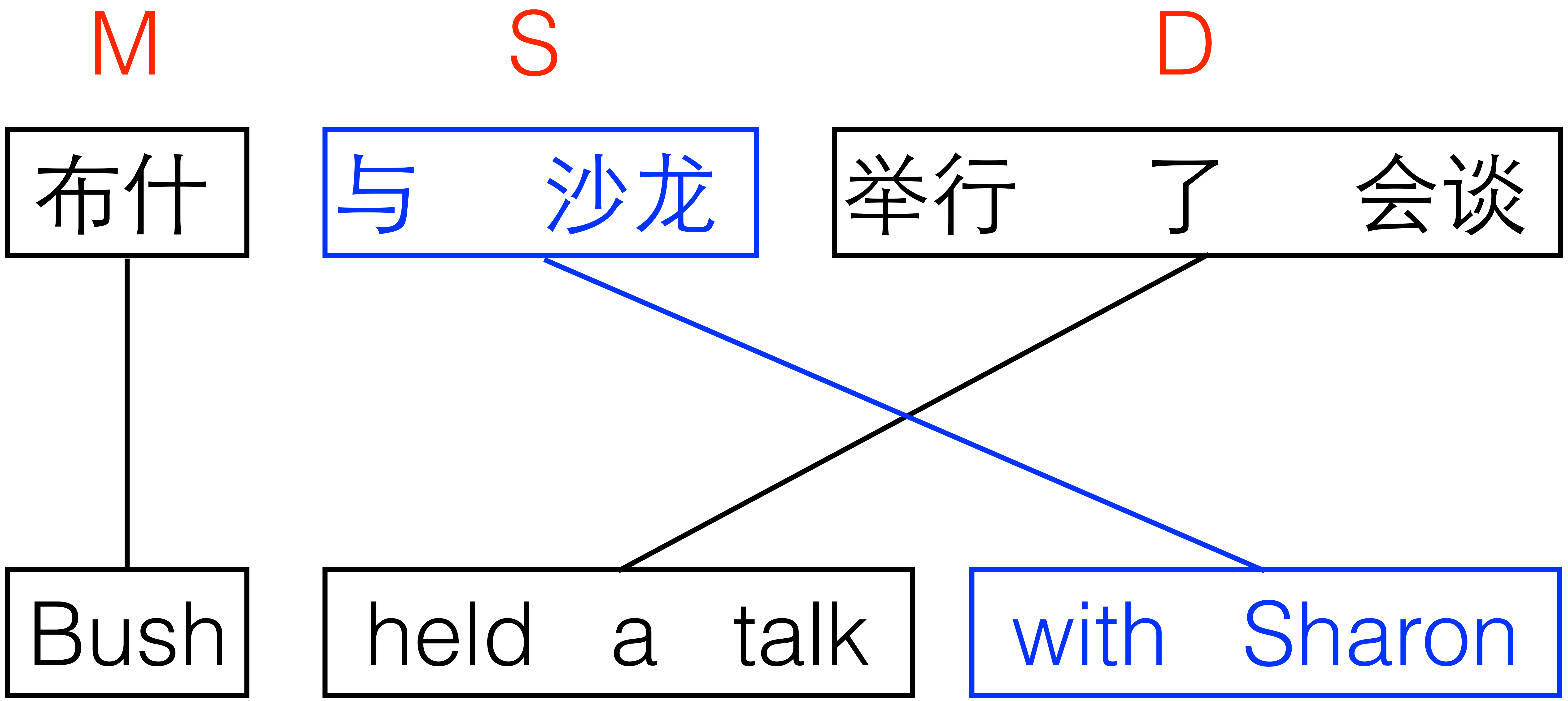
(Koehn et al., 2007; Tillman, 2004; Galley and Manning, 2008)

Lexicalized Models



(Koehn et al., 2007; Tillman, 2004; Galley and Manning, 2008)

Lexicalized Models



(Koehn et al., 2007; Tillman, 2004; Galley and Manning, 2008)

Challenge #1: Sparsity

Source Phrase	Target Phrase	M	S	D
布什	Bush	0.7	0.2	0.1
举行 了 会谈	held a talk	0.1	0.1	0.8
与 沙龙	with Sharon	0.7	0.1	0.2
举行 了	held a	0.6	0.1	0.3
会谈	talk	0.4	0.3	0.3

Challenge #1: Sparsity

- Probability distributions are estimated by MLE

Challenge #1: Sparsity

- Probability distributions are estimated by MLE

$$P\left(\color{blue}{D} \,\middle|\, \begin{array}{c} \boxed{\text{举行 了 会谈}} \\ | \\ \boxed{\text{held a talk}} \end{array} \right) = \rule{8cm}{0.4pt}$$

Challenge #1: Sparsity

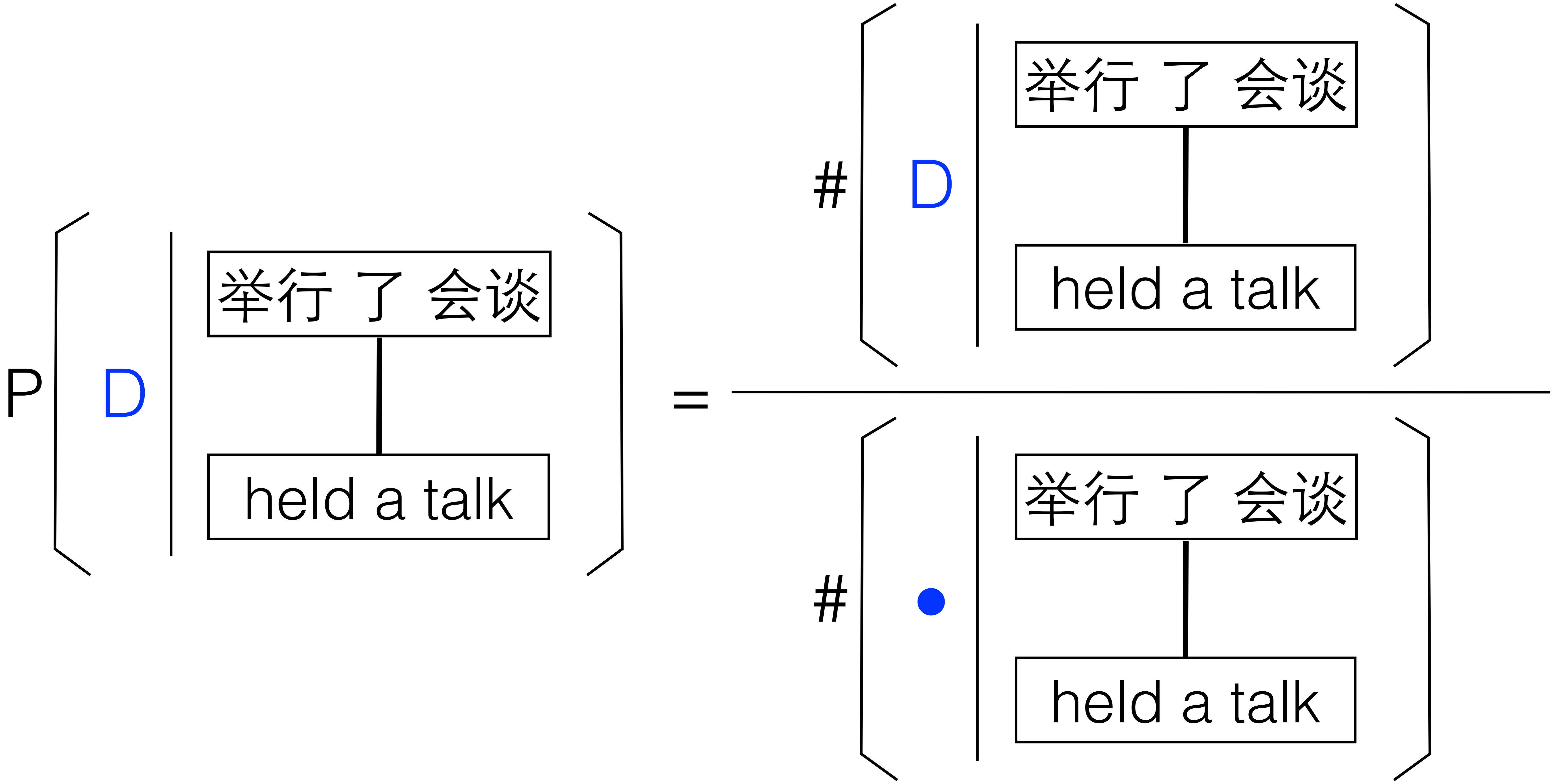
- Probability distributions are estimated by MLE

D 举行 了 会谈 held a talk

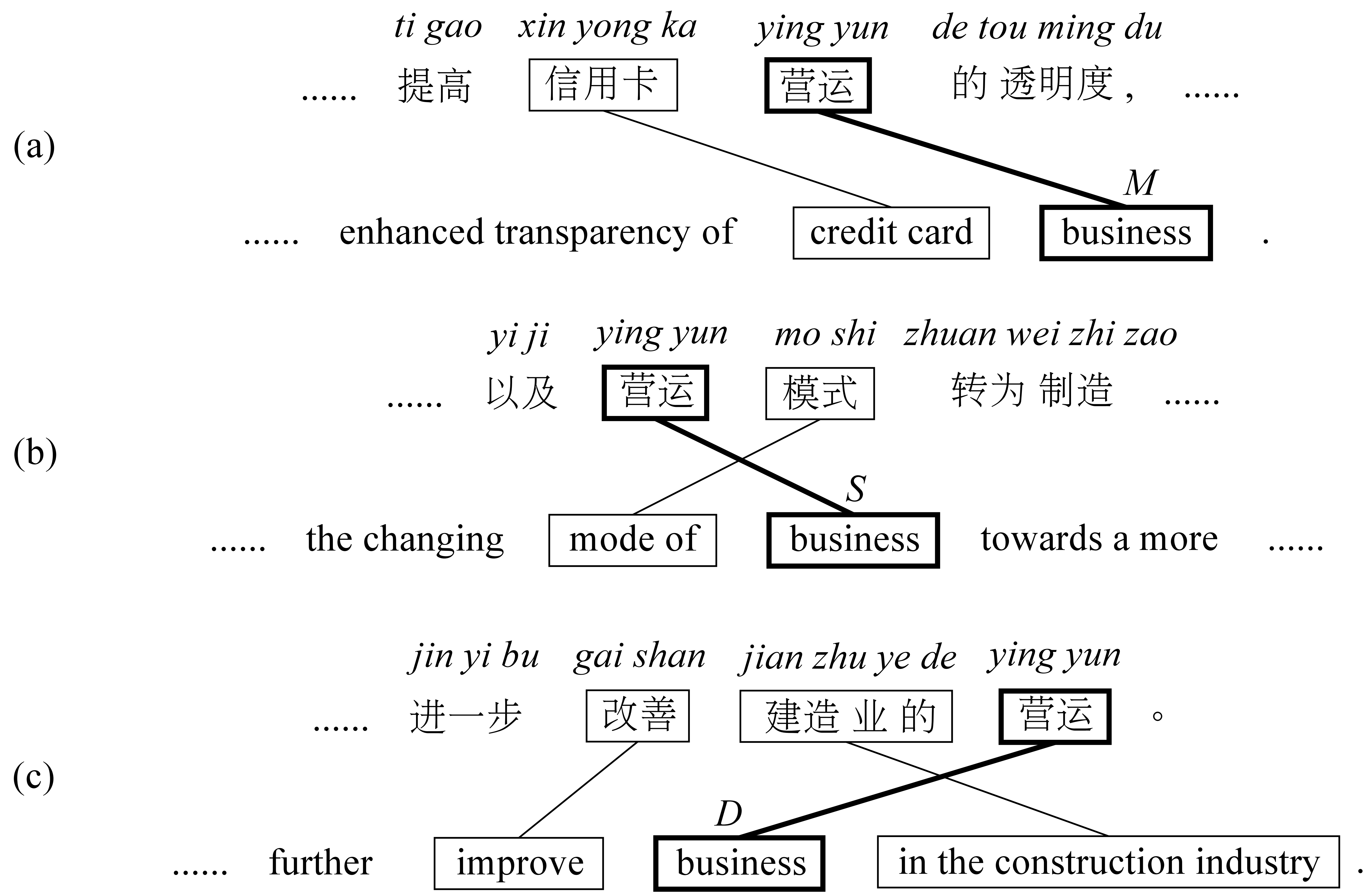
P D 举行 了 会谈 held a talk =

Challenge #1: Sparsity

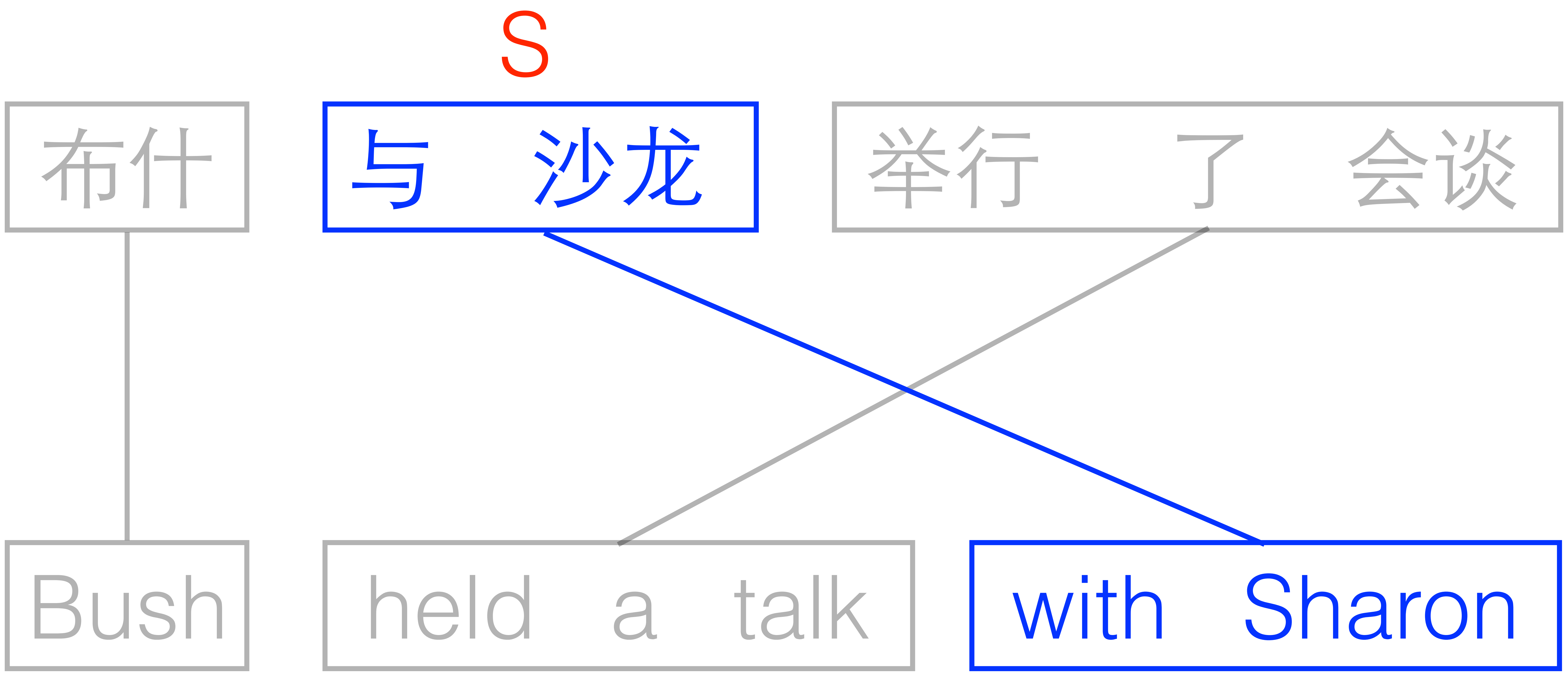
- Probability distributions are estimated by MLE



Challenge #2: Ambiguity



Challenge #3: Context Insensitivity



Challenge #3: Context Insensitivity

S

How to resolve the three challenges?

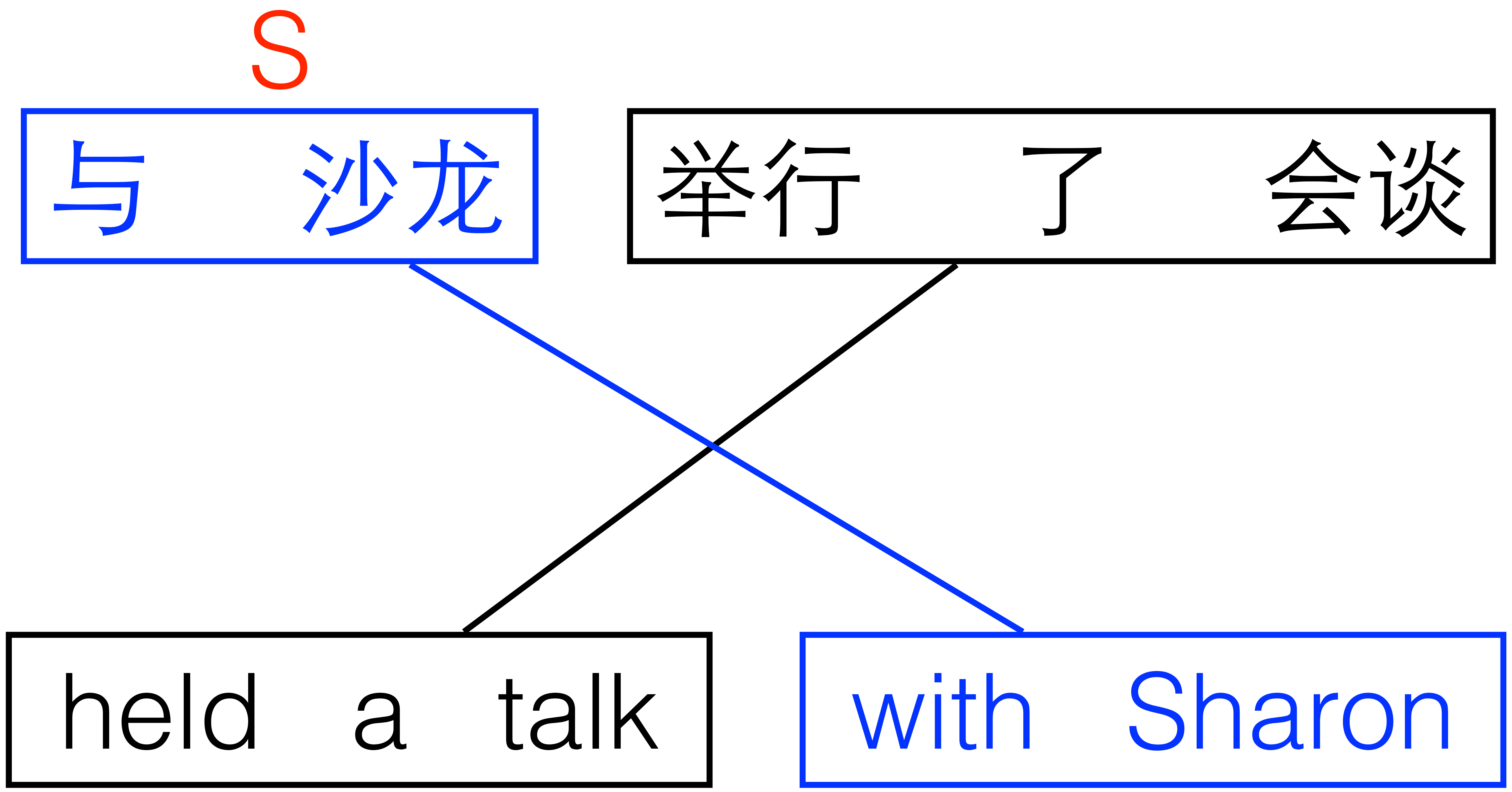
Bush | held a talk | with Sharon

Including More Contexts

Sparsity

Ambiguity

Context Insensitivity

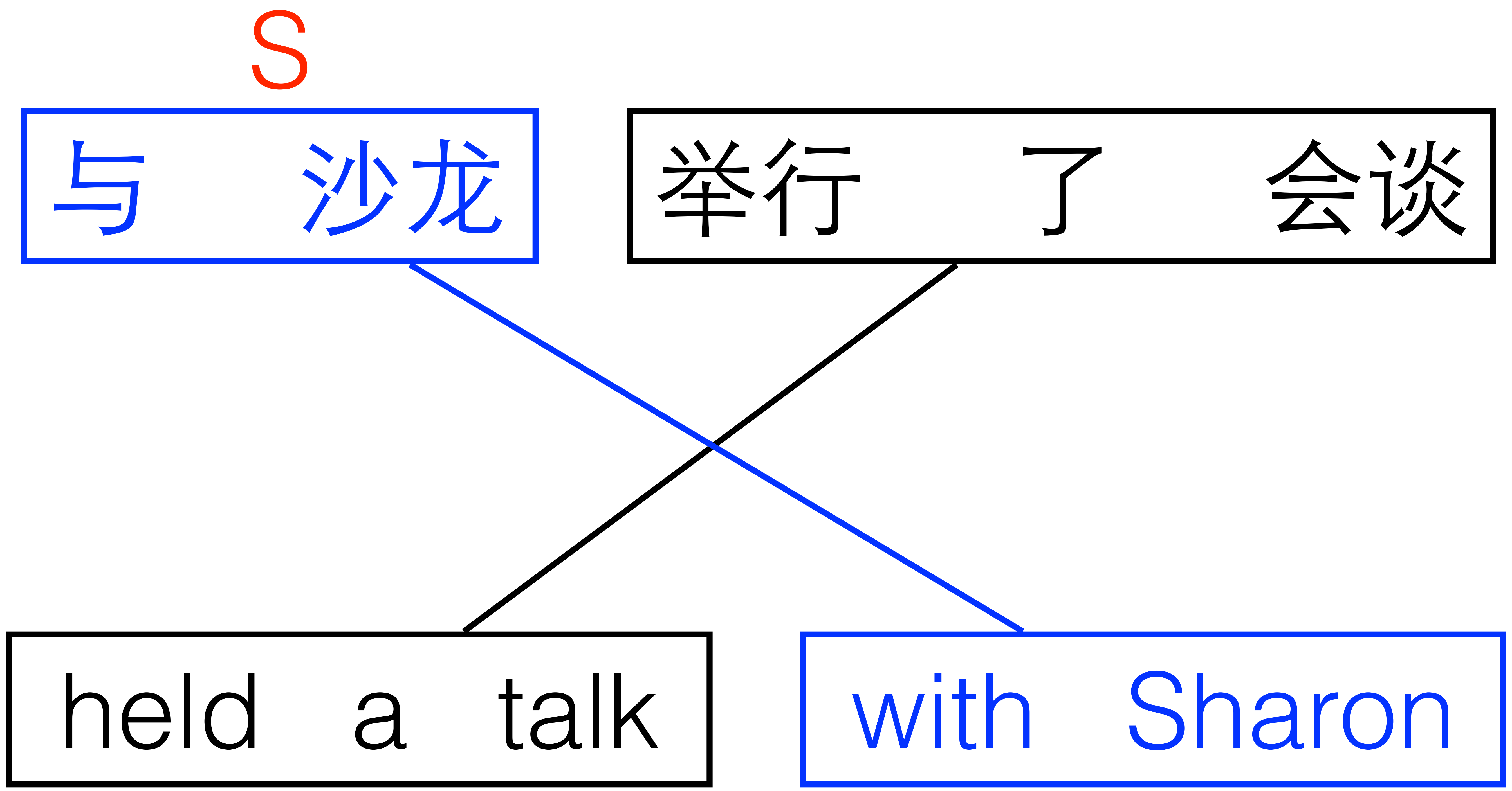


Including More Contexts

Sparsity

✔ Ambiguity

✔ Context Insensitivity

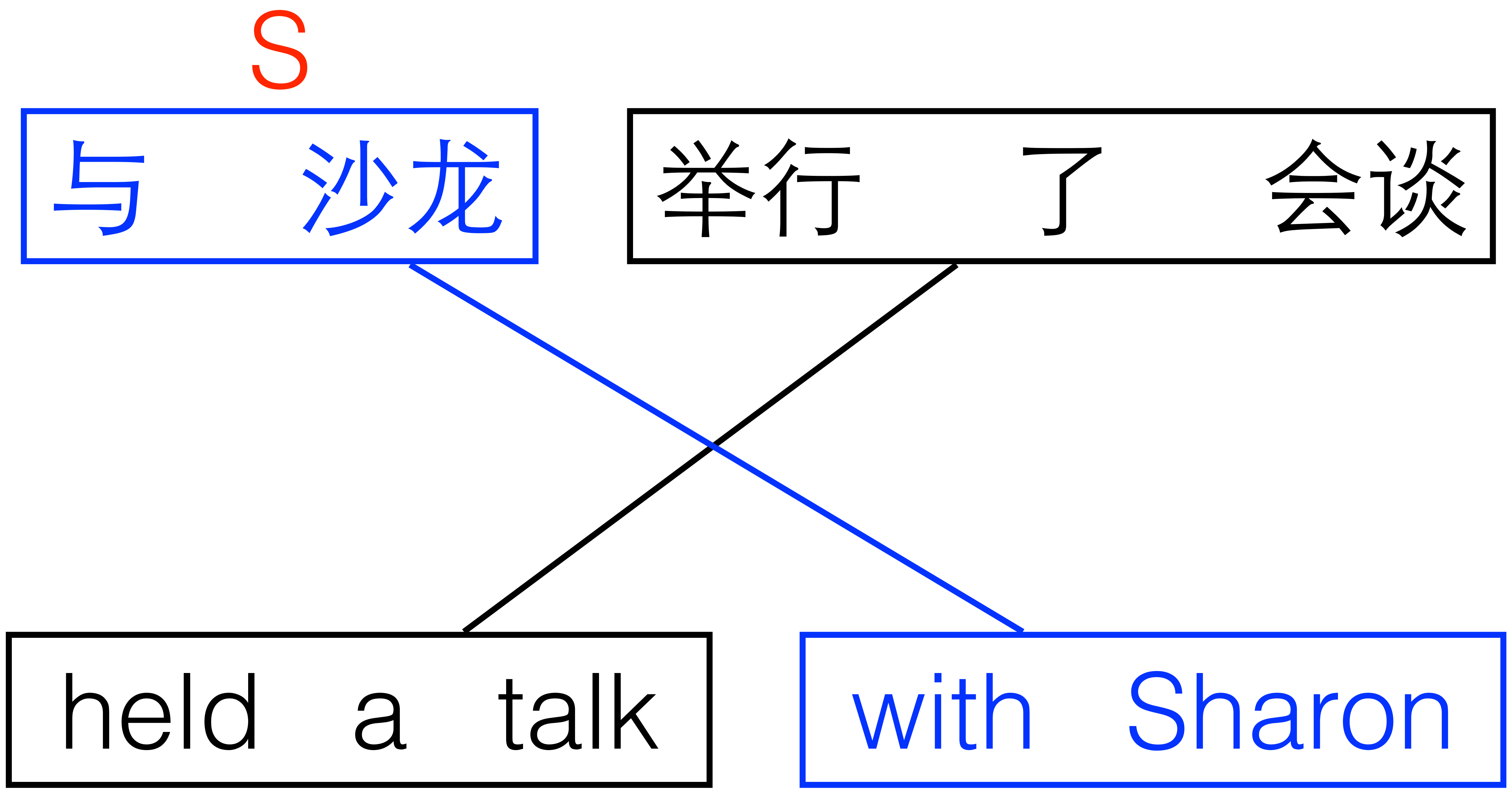


Including More Contexts

? Sparsity

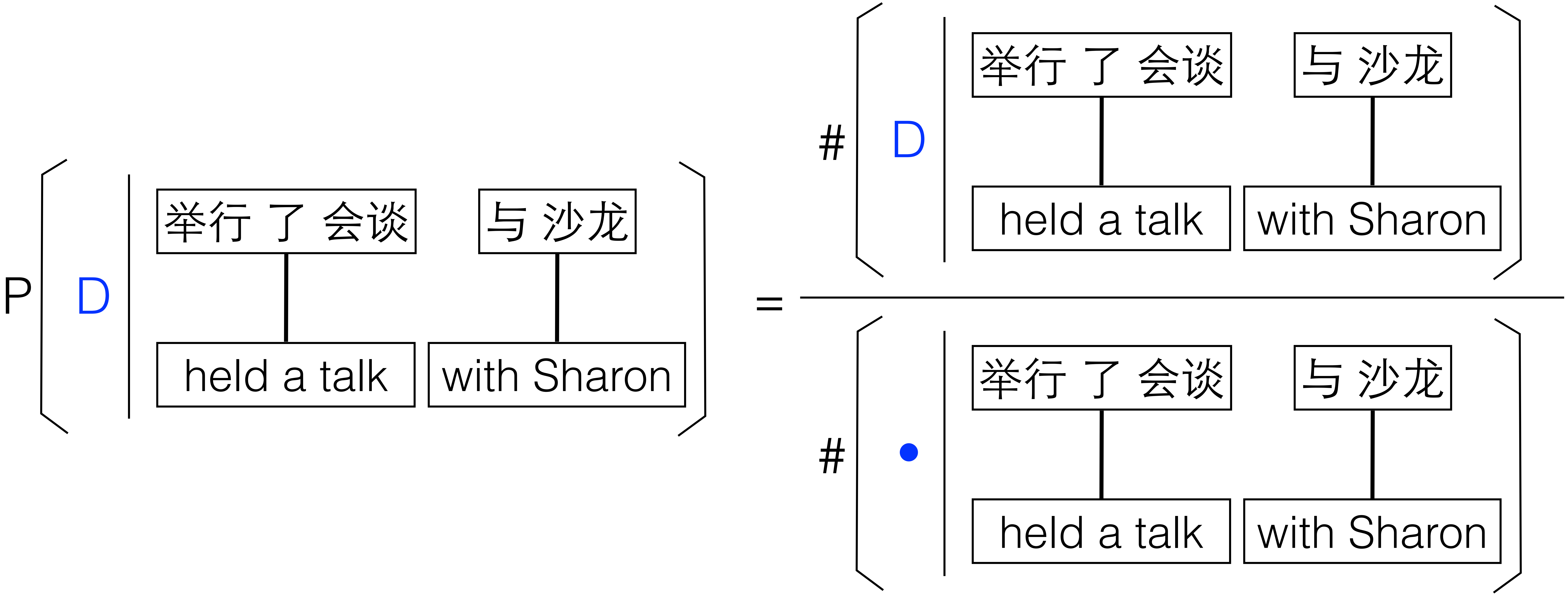
✔ Ambiguity

✔ Context Insensitivity



Sparsity

- Including more contexts leads to severer sparsity



Sparsity

- Including more contexts leads to severer sparsity

Reordering as Classification

Neural Reordering Model

Neural Reordering Model

- A neural classifier for predicting reordering orientations

Neural Reordering Model

- A neural classifier for predicting reordering orientations
- Conditioned on both the current and previous phrase pairs
    - Improves context sensitivity
    - Reduces reordering ambiguity

Neural Reordering Model

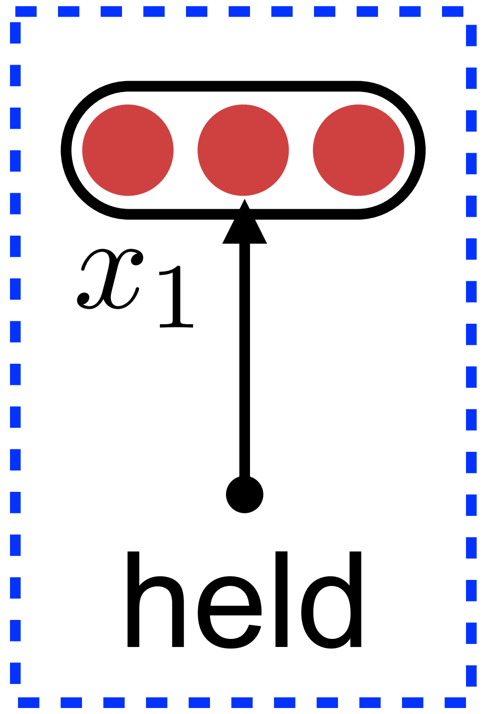
- A neural classifier for predicting reordering orientations
- Conditioned on both the current and previous phrase pairs
    - Improves context sensitivity
    - Reduces reordering ambiguity
- A single classifier for all phrase pairs
    - Uses vector space representations
    - Alleviates the data sparsity problem

Recursive Autoencoder (RAE)

held a talk

(Pollack; 1990; Socher et. al, 2011)

Recursive Autoencoder (RAE)

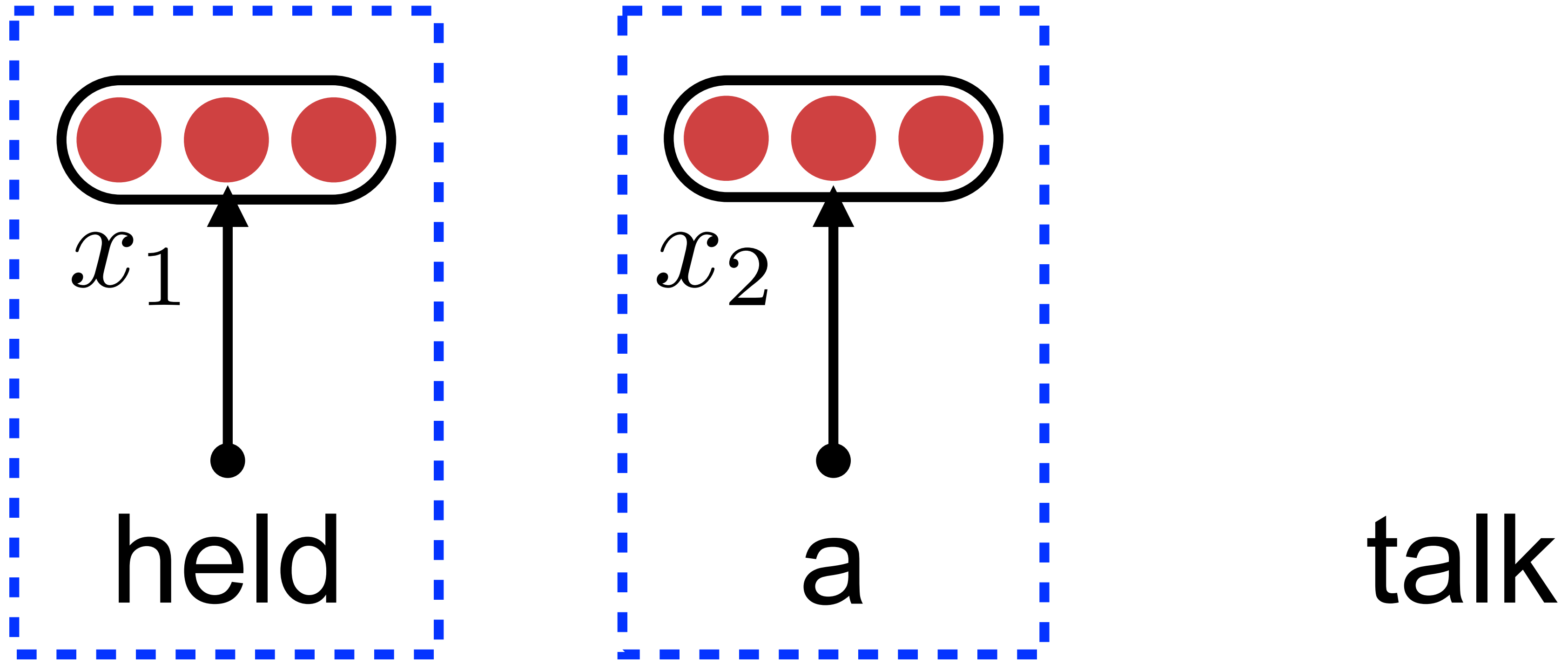


a

talk

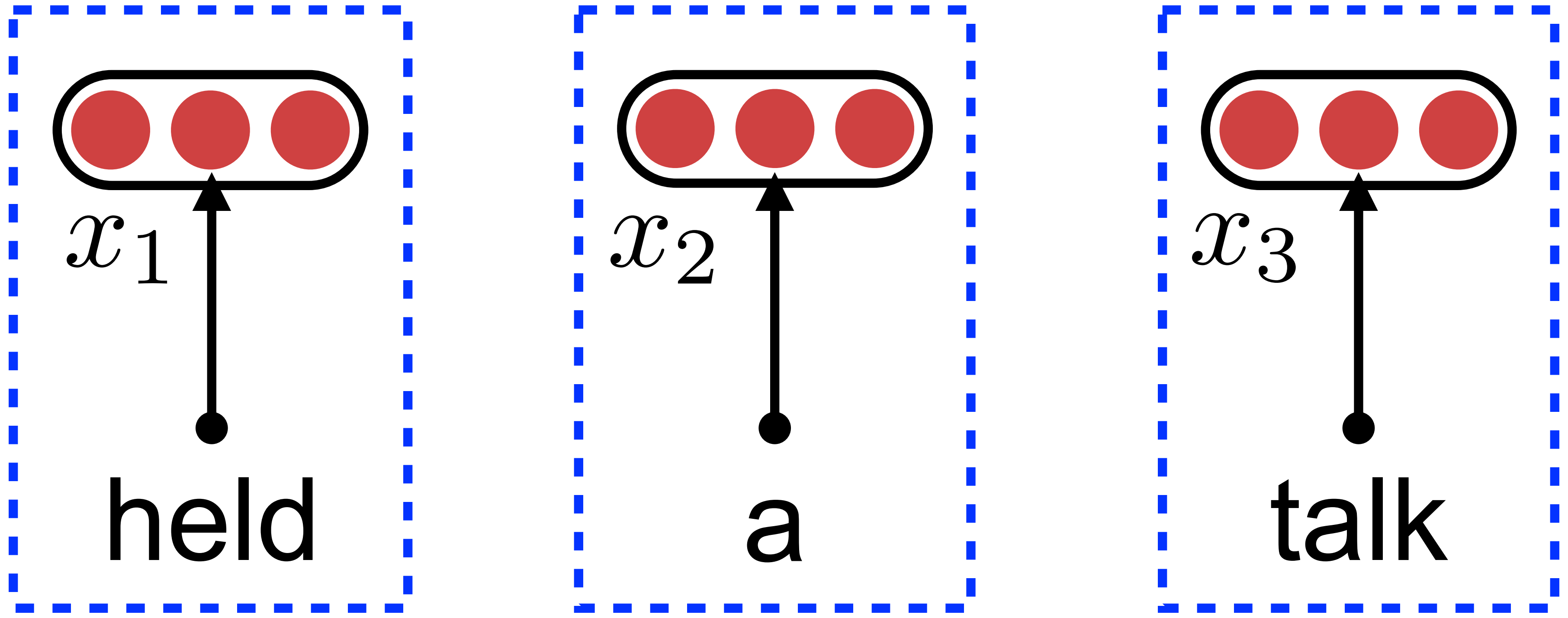
(Pollack; 1990; Socher et. al, 2011)

Recursive Autoencoder (RAE)



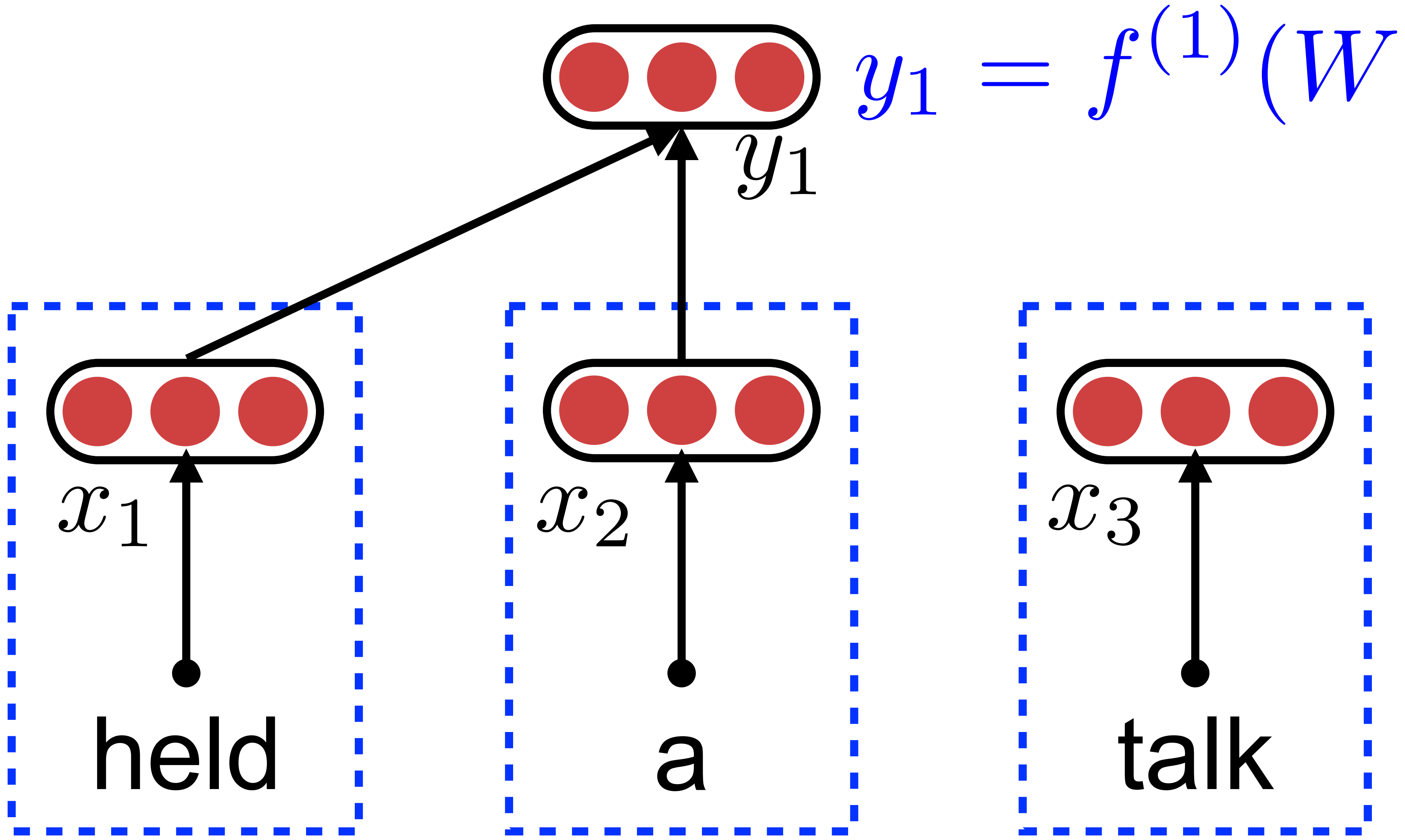
(Pollack; 1990; Socher et. al, 2011)

Recursive Autoencoder (RAE)



(Pollack; 1990; Socher et. al, 2011)

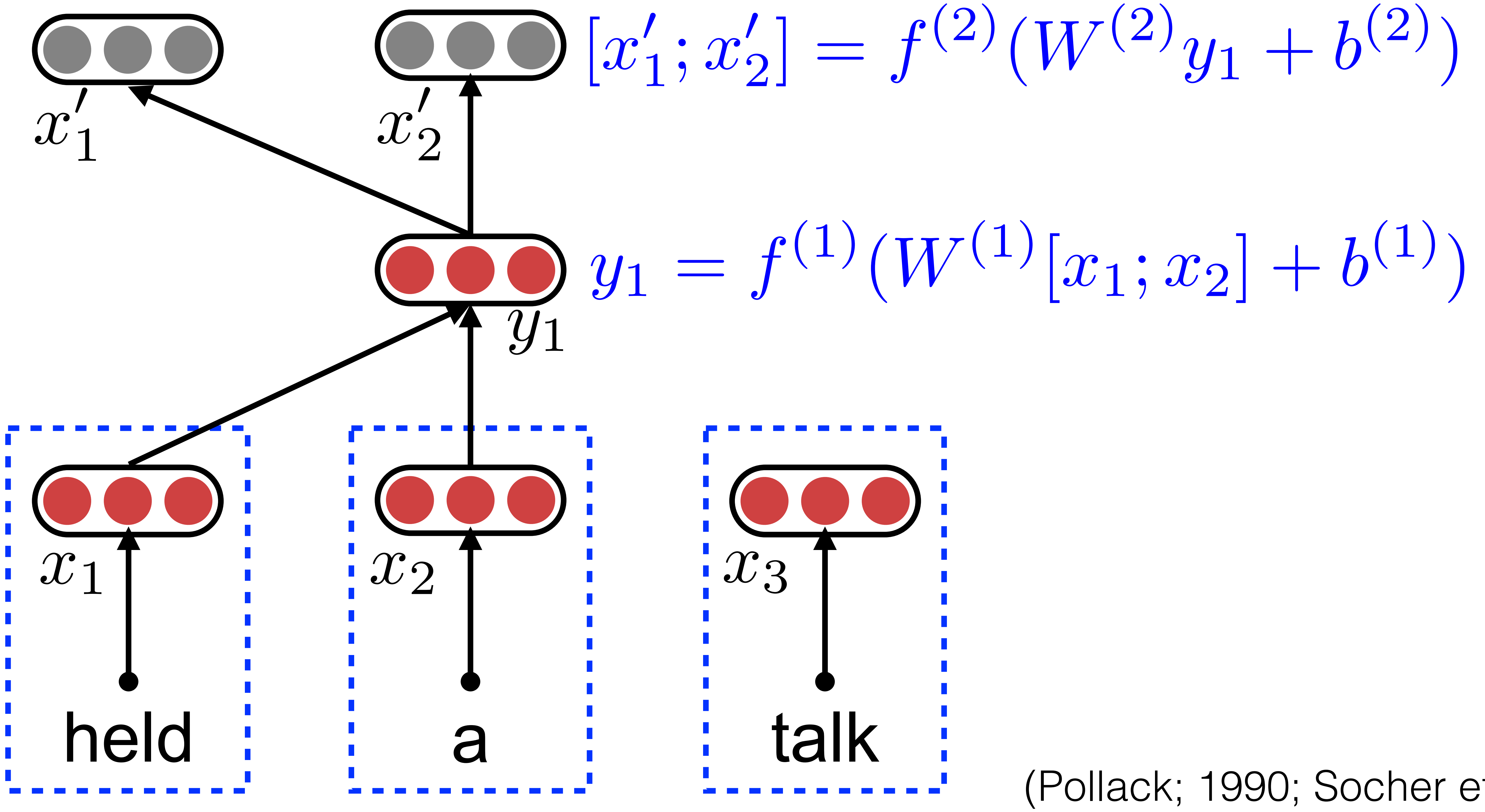
Recursive Autoencoder (RAE)



$$y_1 = f^{(1)}(W^{(1)}[x_1; x_2] + b^{(1)})$$

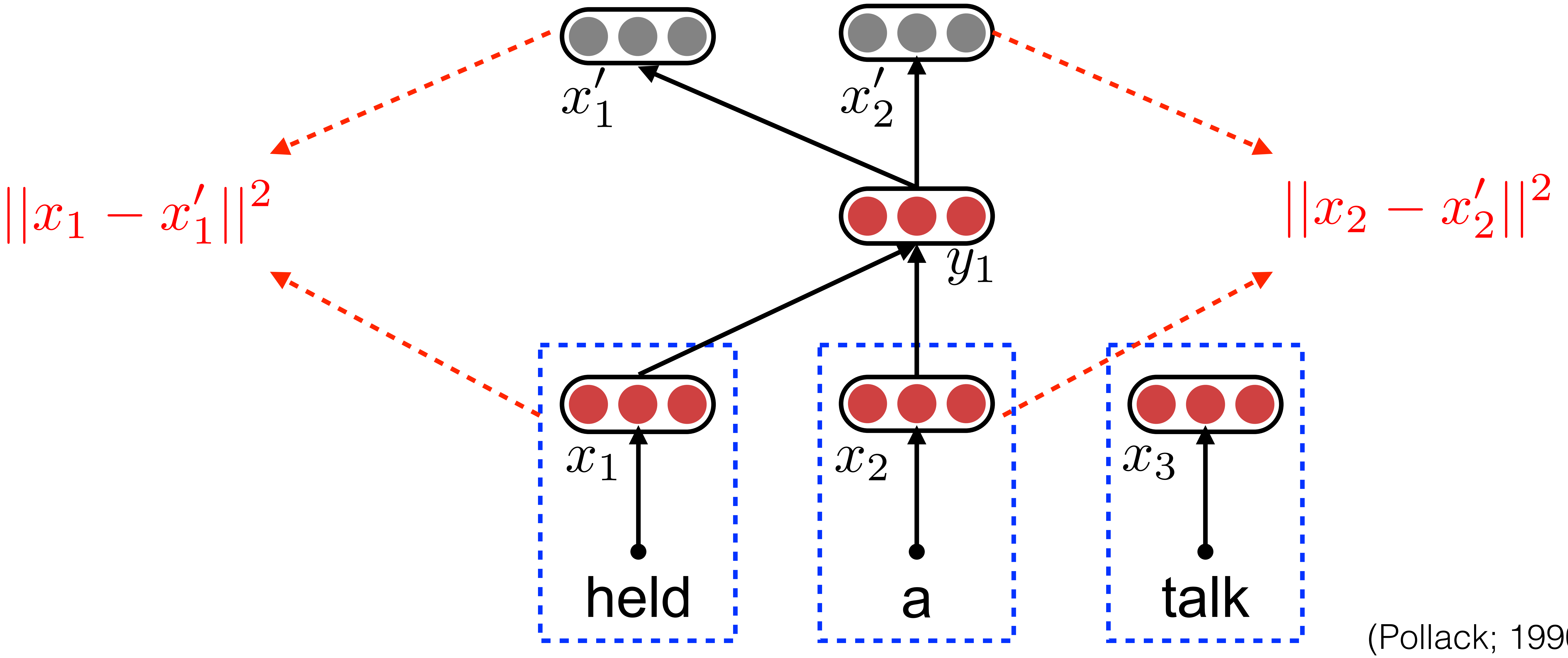
(Pollack; 1990; Socher et. al, 2011)

Recursive Autoencoder (RAE)



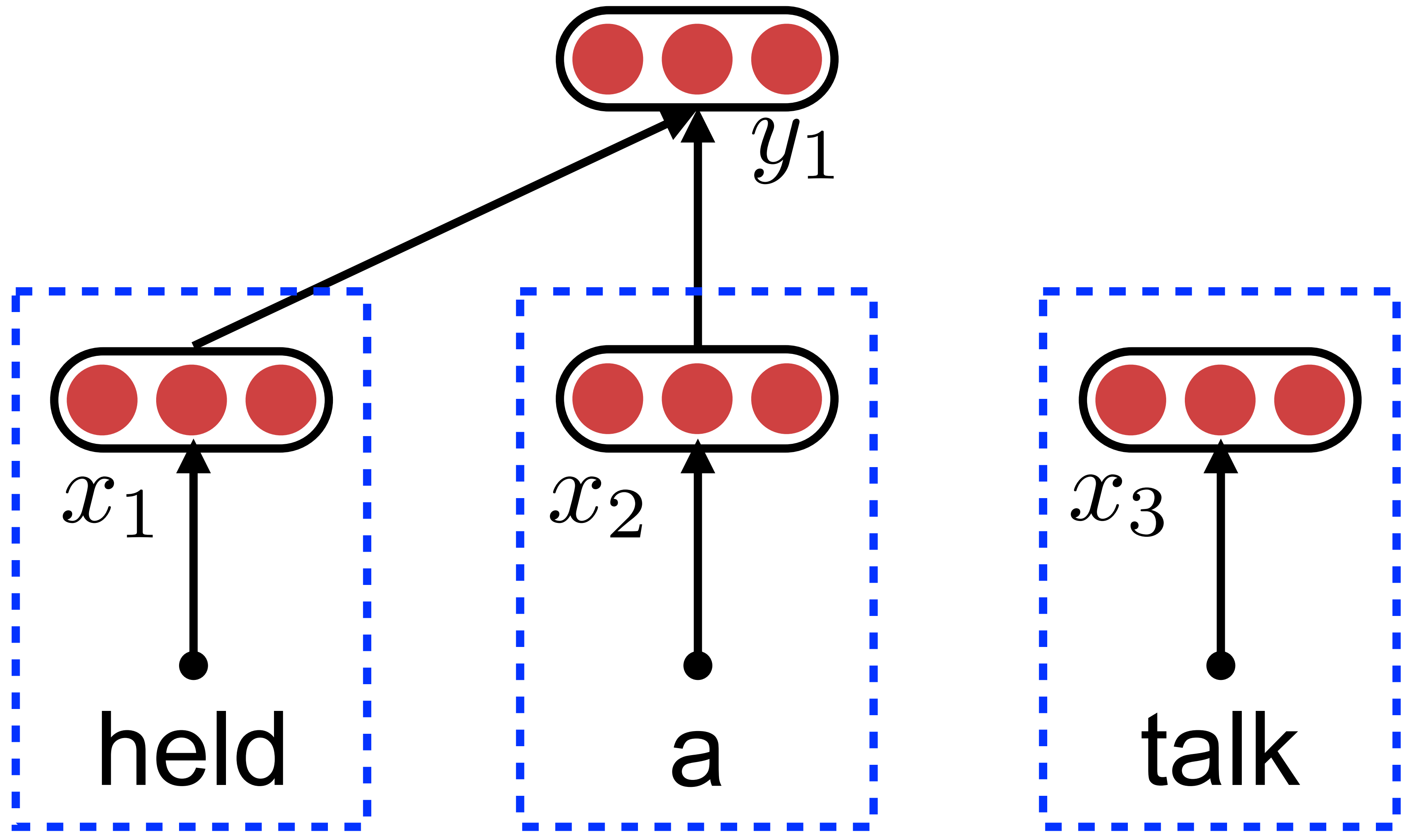
(Pollack; 1990; Socher et. al, 2011)

Recursive Autoencoder (RAE)



(Pollack; 1990; Socher et. al, 2011)

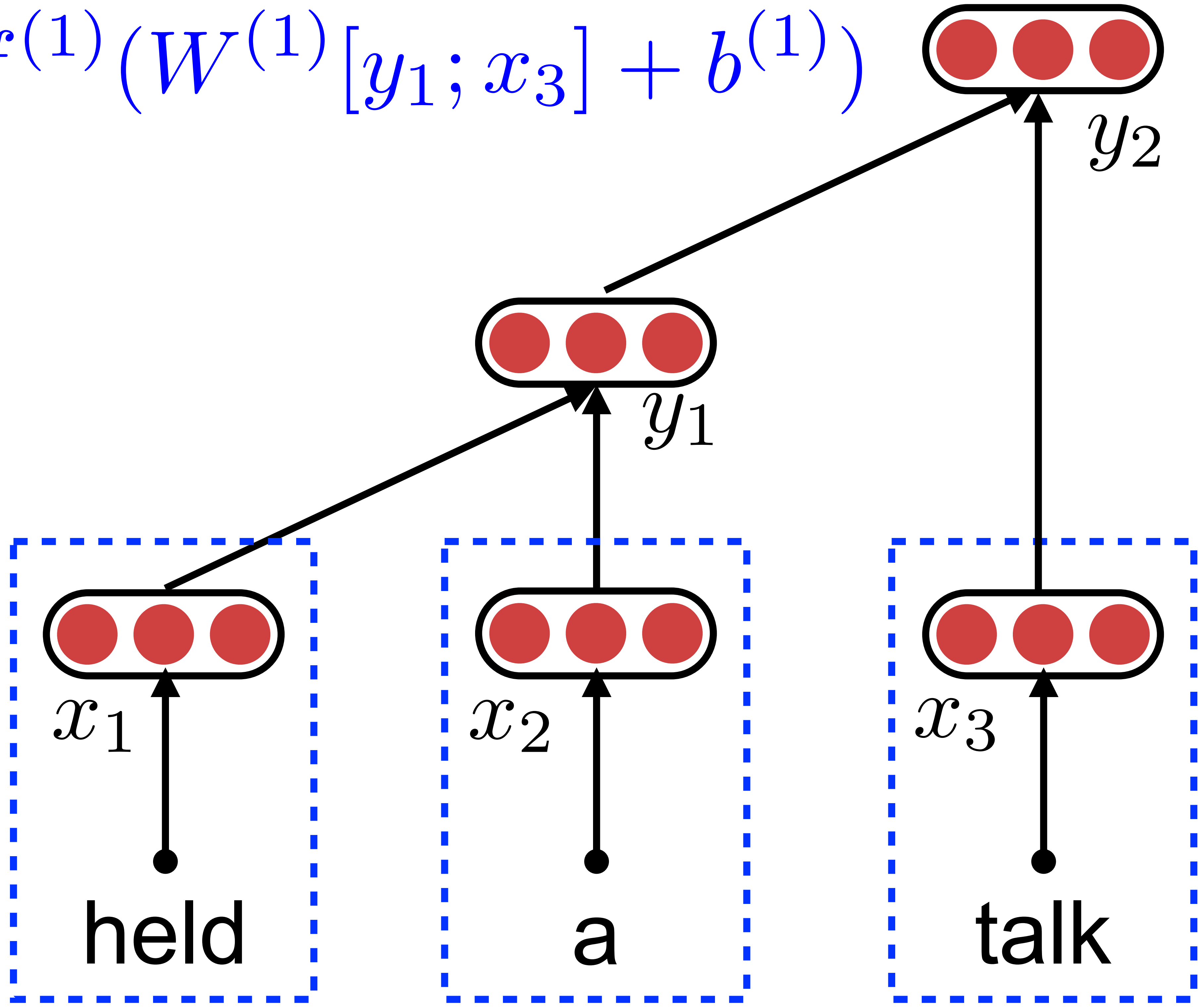
Recursive Autoencoder (RAE)



(Pollack; 1990; Socher et. al, 2011)

Recursive Autoencoder (RAE)

$$y_2 = f^{(1)}(W^{(1)}[y_1; x_3] + b^{(1)})$$

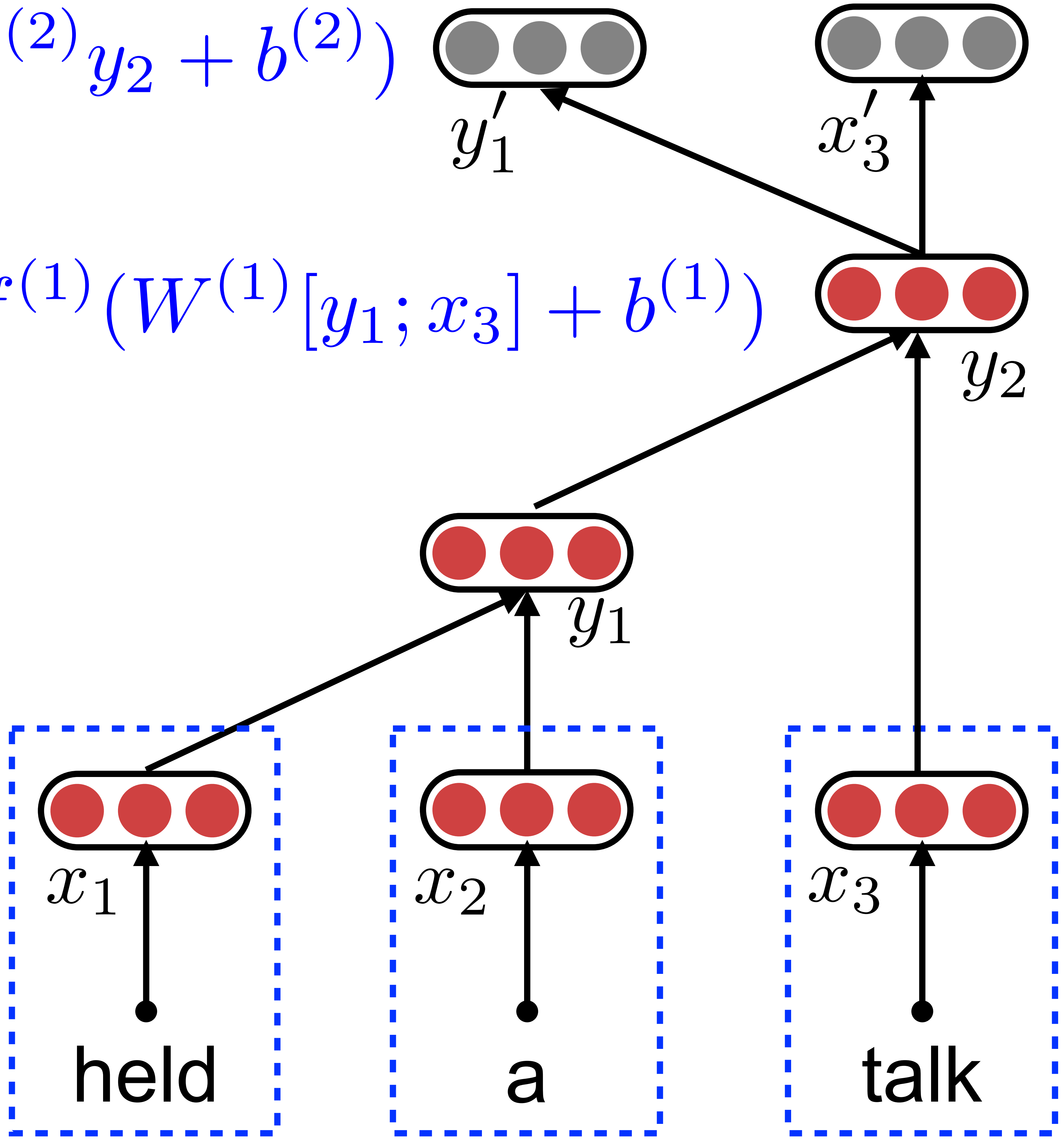


(Pollack; 1990; Socher et. al, 2011)

Recursive Autoencoder (RAE)

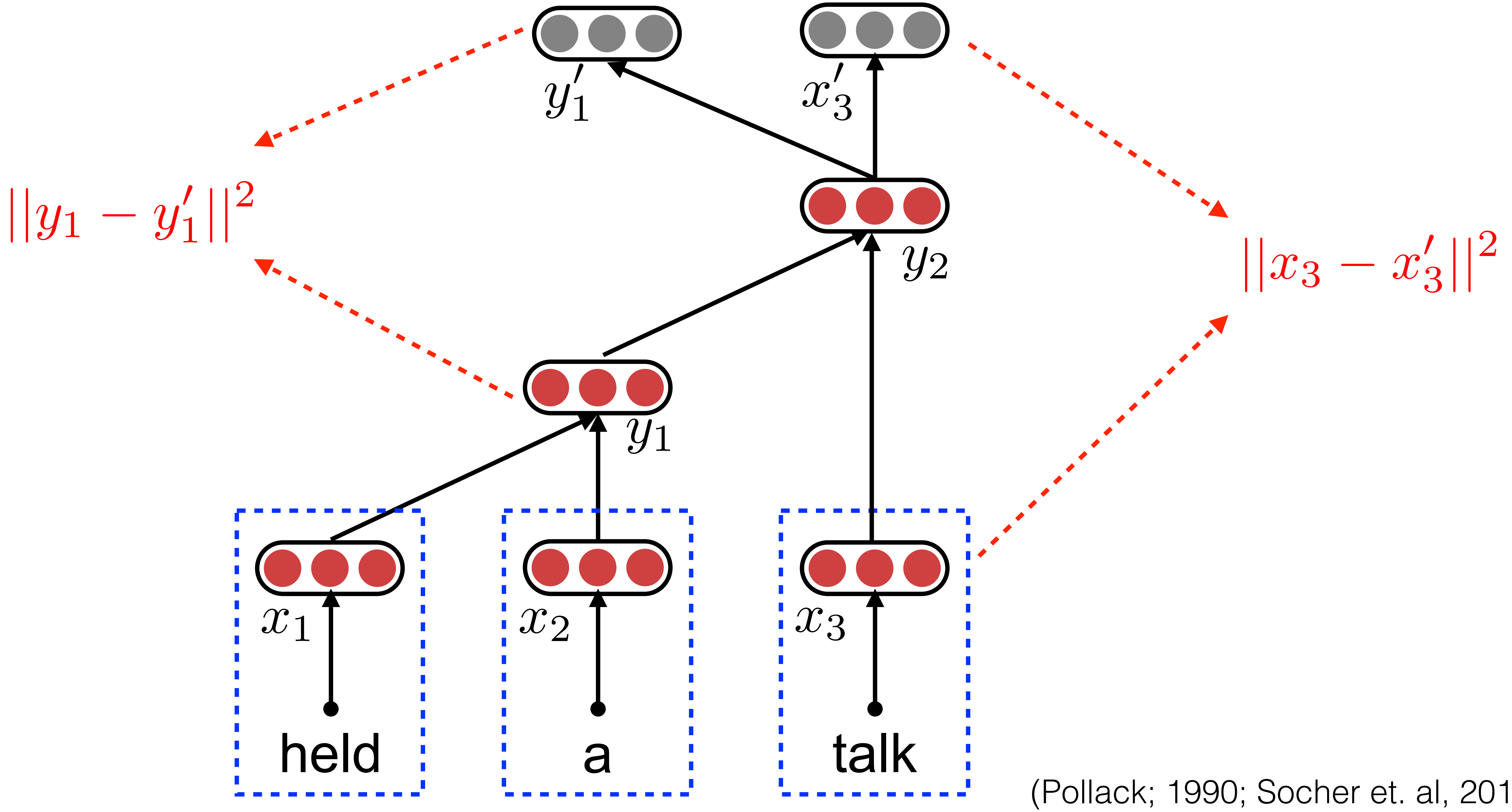
$$[y_1'; x_3'] = f^{(2)}(W^{(2)} y_2 + b^{(2)})$$

$$y_2 = f^{(1)}(W^{(1)}[y_1; x_3] + b^{(1)})$$



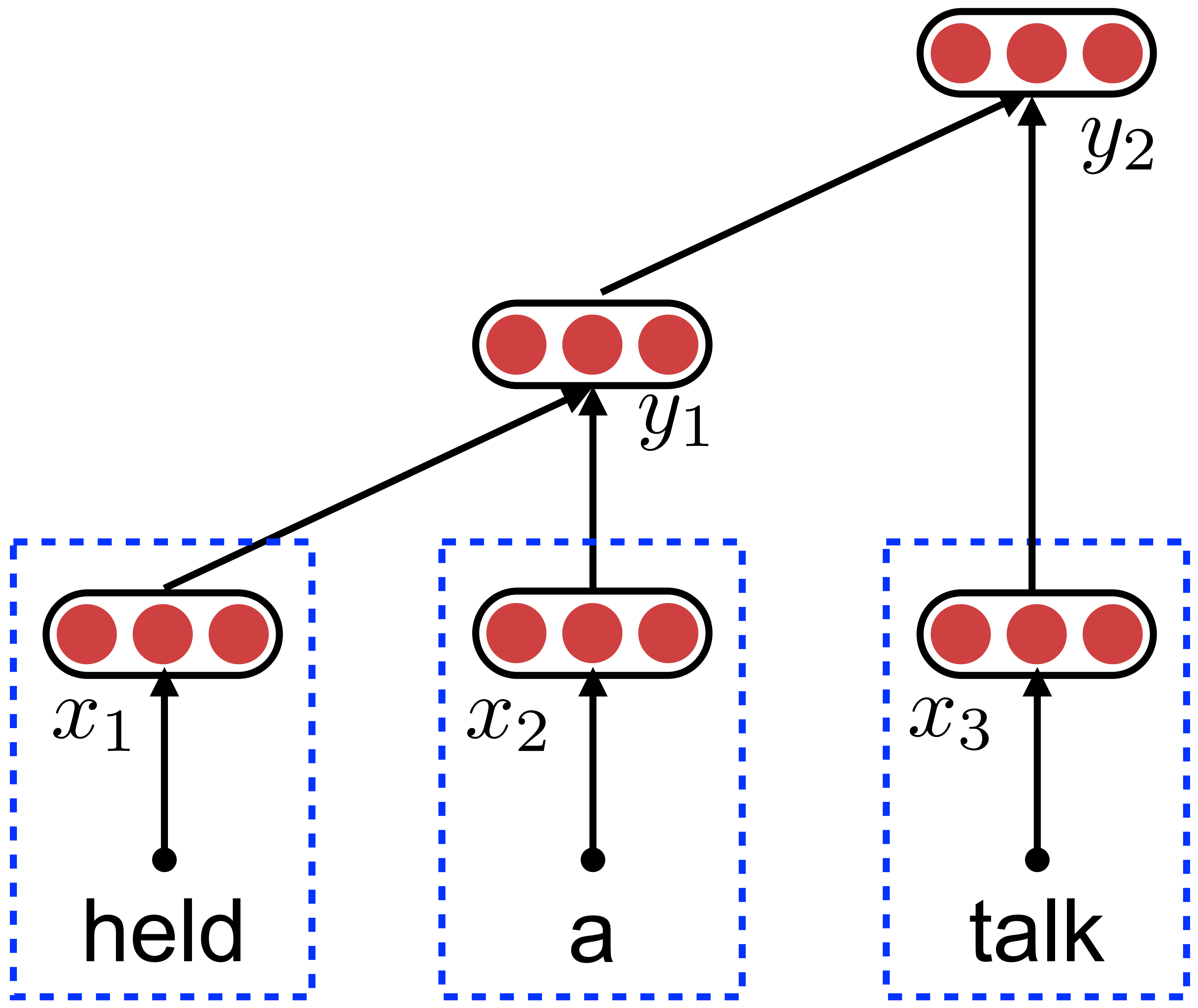
(Pollack; 1990; Socher et. al, 2011)

Recursive Autoencoder (RAE)



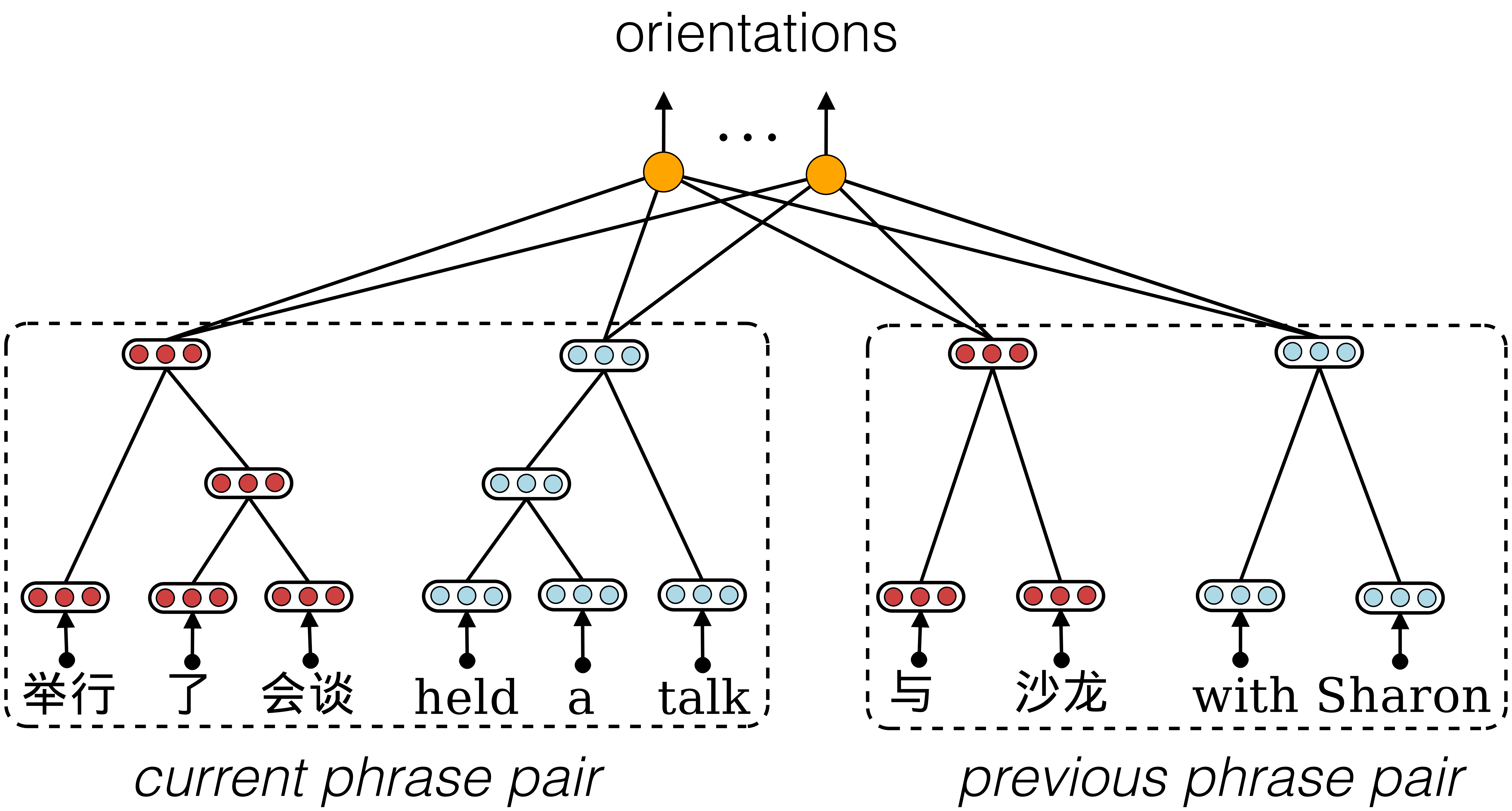
(Pollack; 1990; Socher et. al, 2011)

Recursive Autoencoder (RAE)

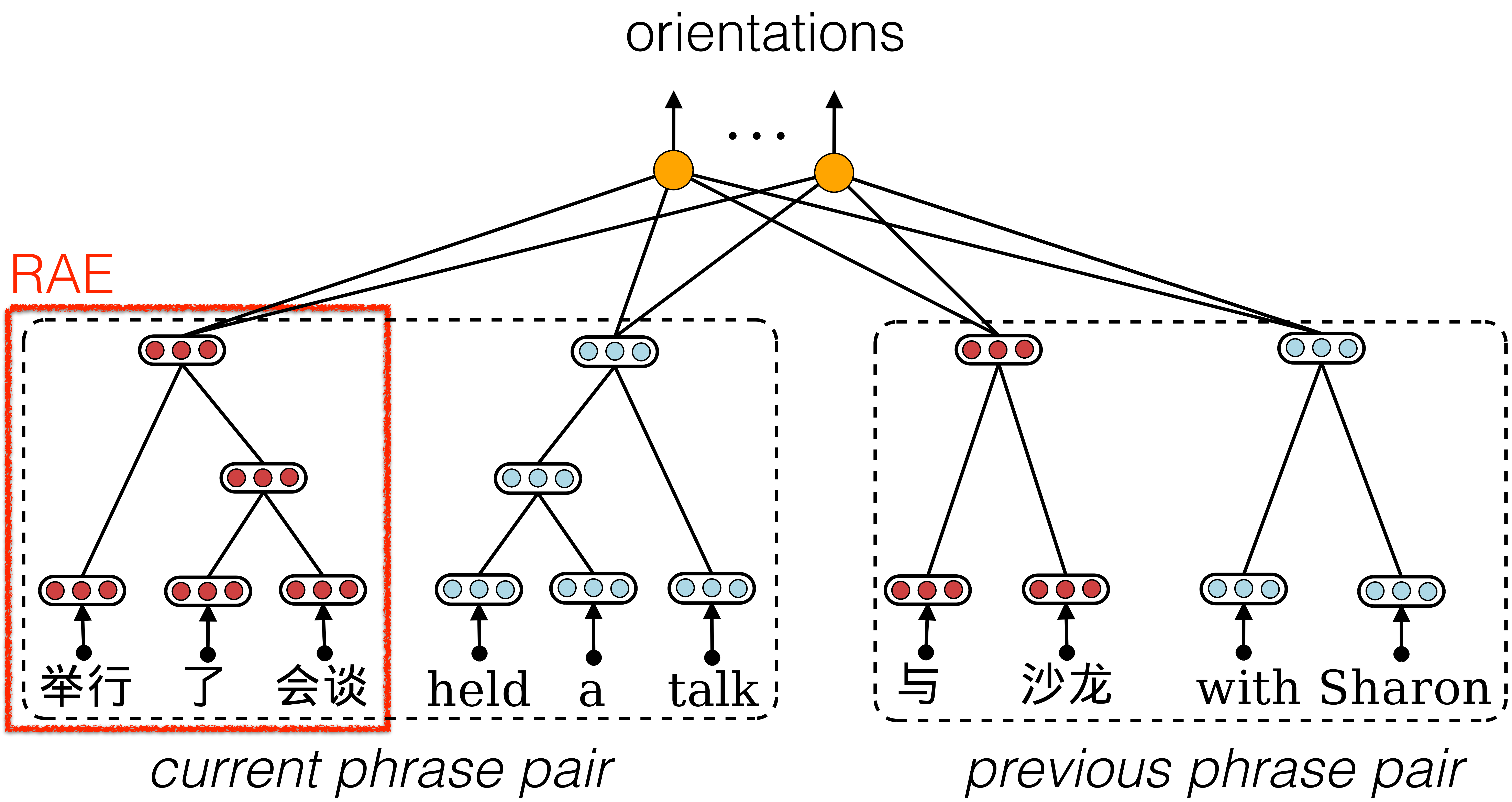


(Pollack; 1990; Socher et. al, 2011)

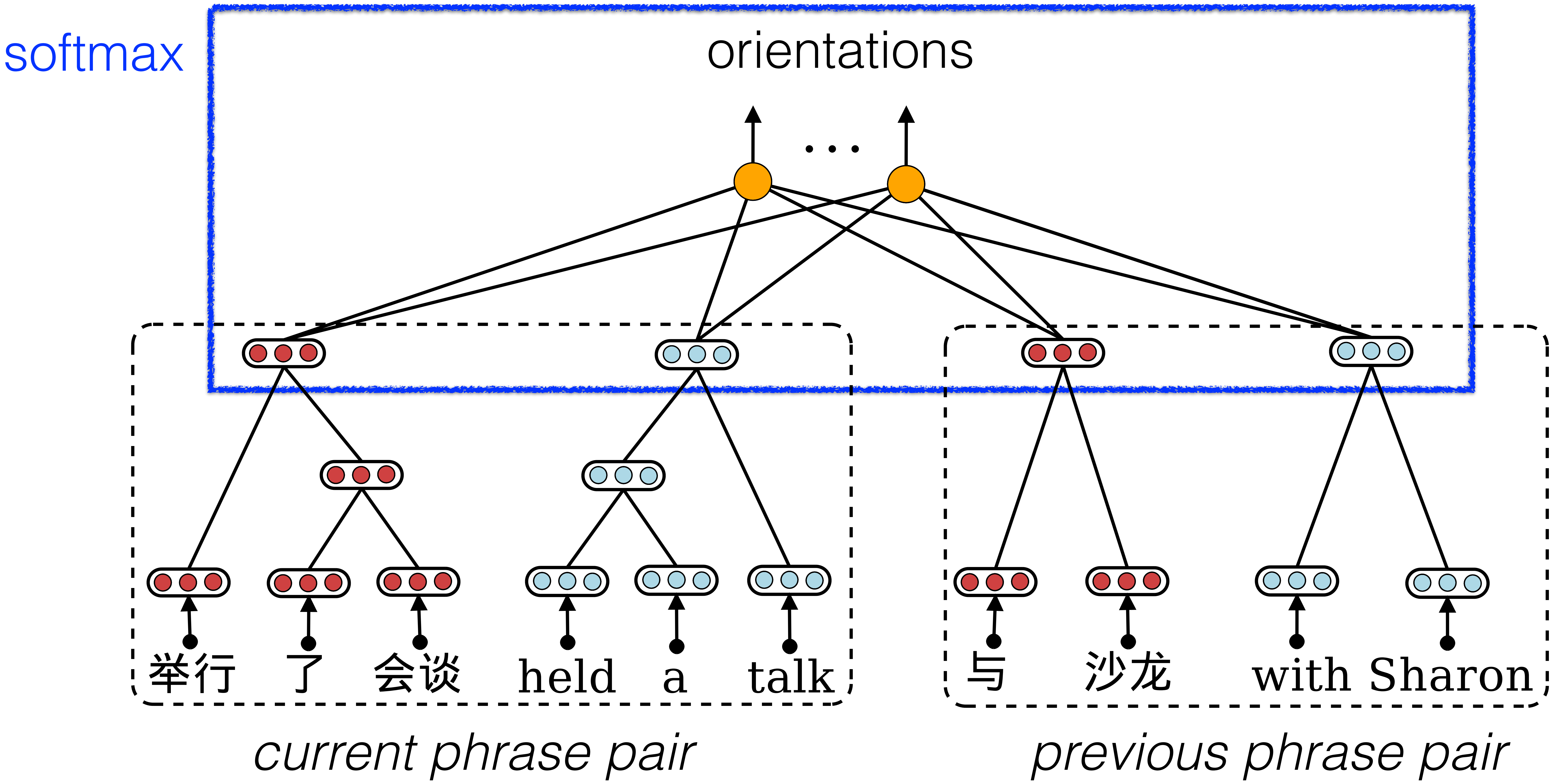
Neural Classifier



Neural Classifier

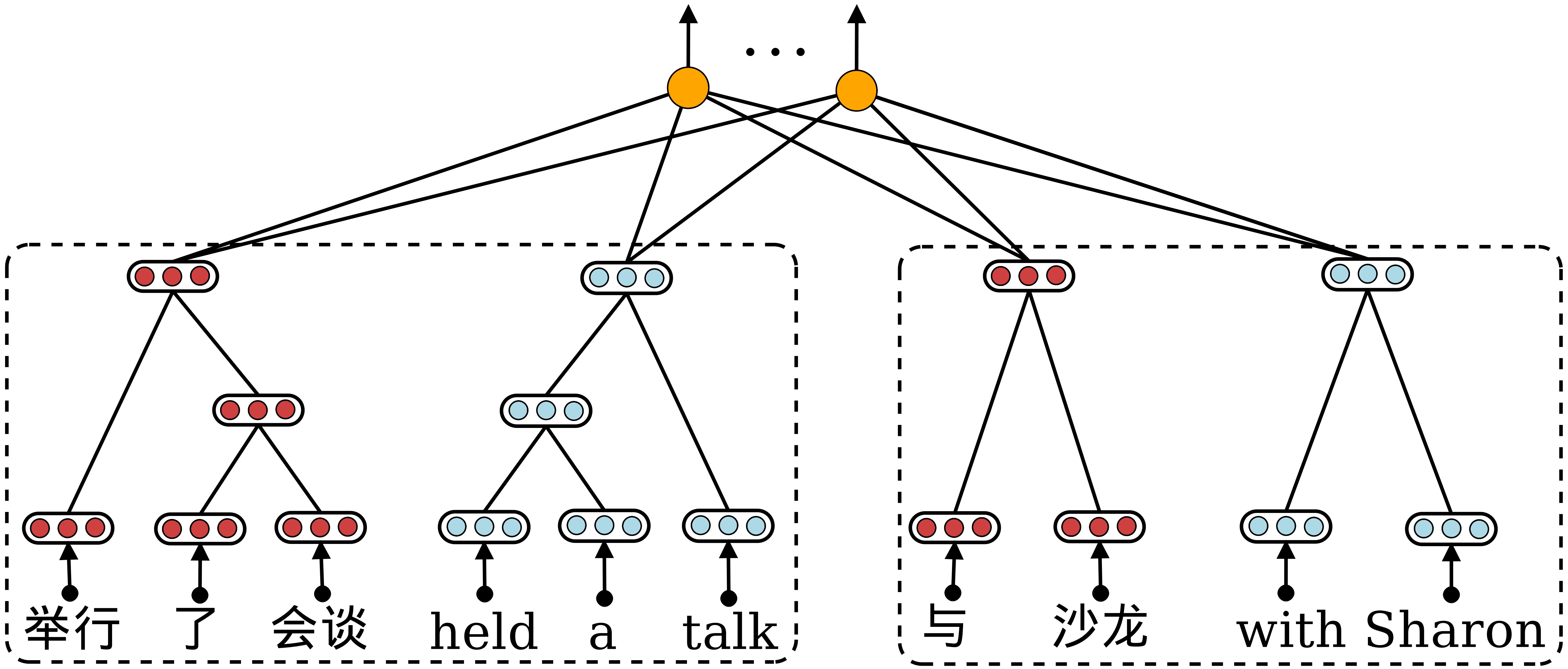


Neural Classifier



Training

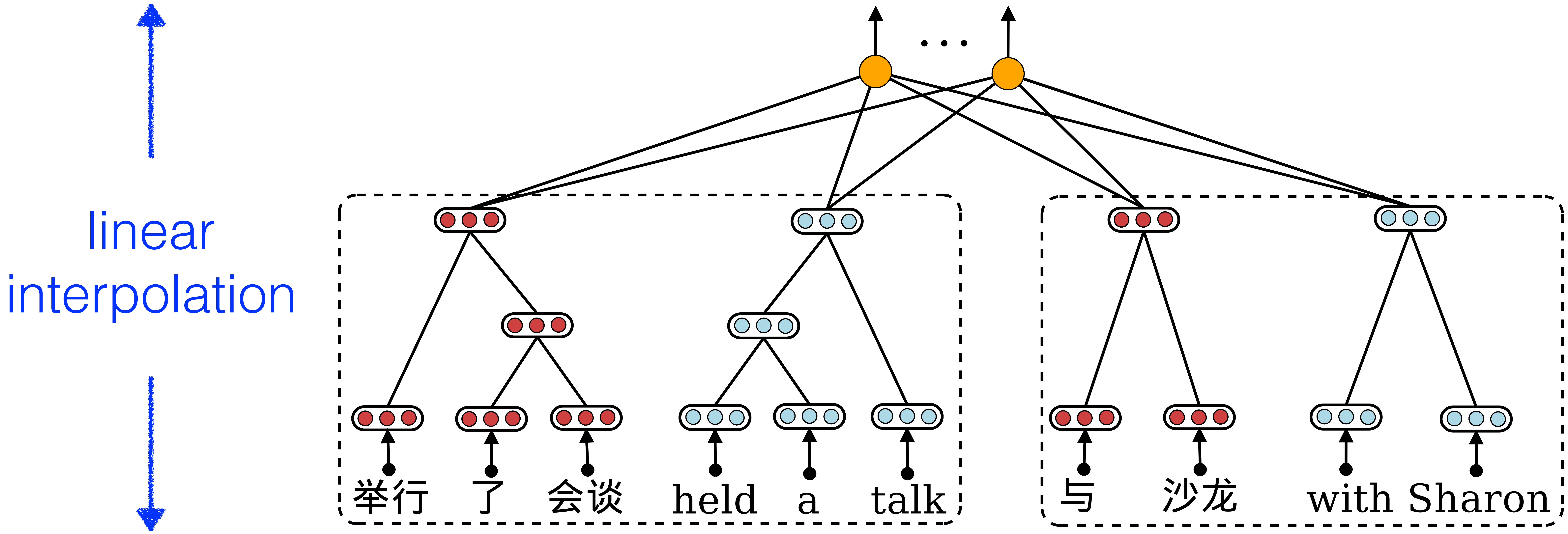
Reordering error on predicting orientations



Reconstruction error on recovering training examples

Training

Reordering error on predicting orientations



Reconstruction error on recovering training examples

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 the gradients (Goller and Kuchler, 1996)

Experiments

- Chinese-English translation
- Training: 1.2M sentence pairs
- LM: 4-gram, 397.6M words
- Dev. set: NIST 06
- Test set: NIST 02-05, 08
- Case-insensitive BLEU

Experiments

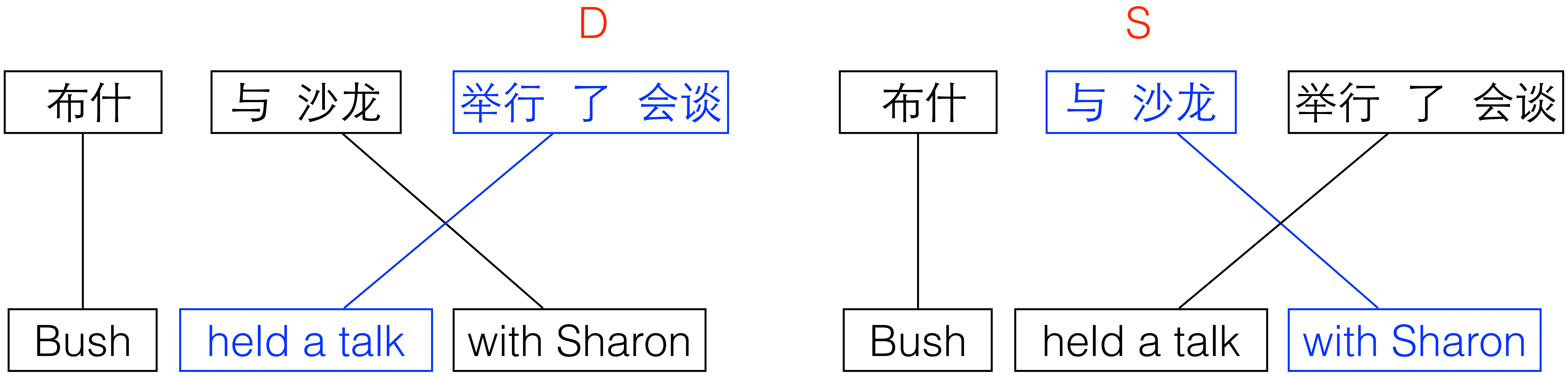
- Chinese-English translation
- Training: 1.2M sentence pairs
- LM: 4-gram, 397.6M words
- Dev. set: NIST 06
- Test set: NIST 02-05, 08
- Case-insensitive BLEU

- Baselines
  - Distance-based model
  - Lexicalized model

$$\underset{\text{model type}}{\begin{pmatrix} \text{word-based} \\ \text{phrase-based} \\ \text{hier. phrase-based} \end{pmatrix}} \times \underset{\text{orientations}}{\begin{pmatrix} \text{M/S/D} \\ \text{left/right} \end{pmatrix}}$$

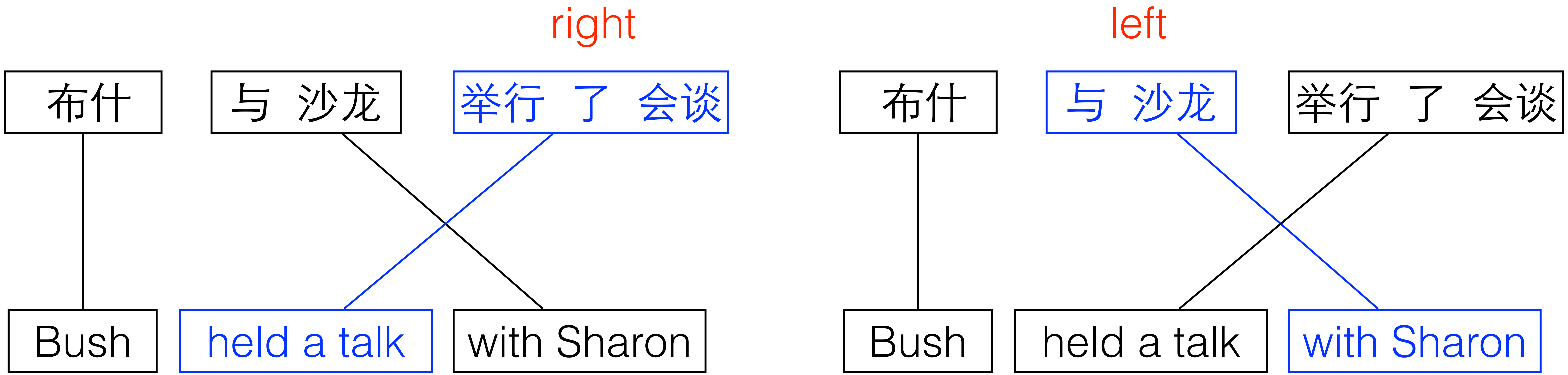
M/S/D Orientations

- Care about relative position and adjacency



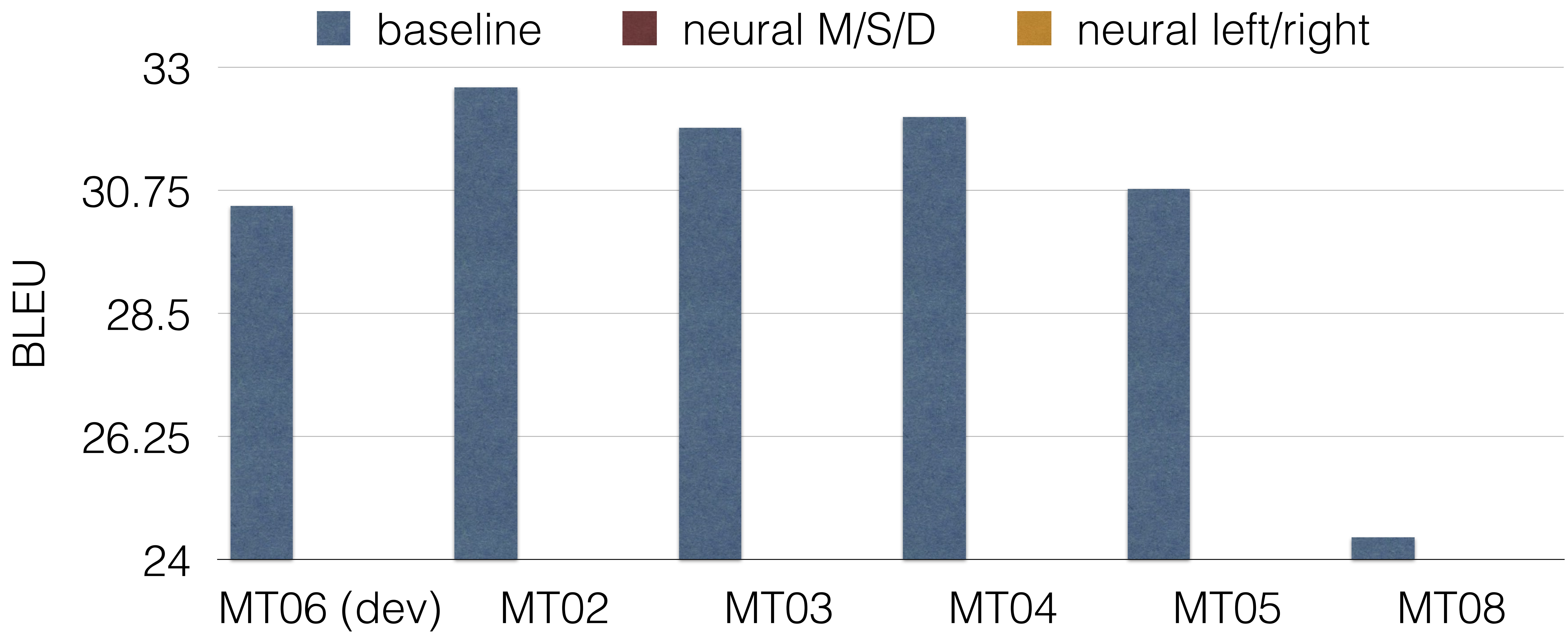
Left/Right Orientations

- Only care about relative position

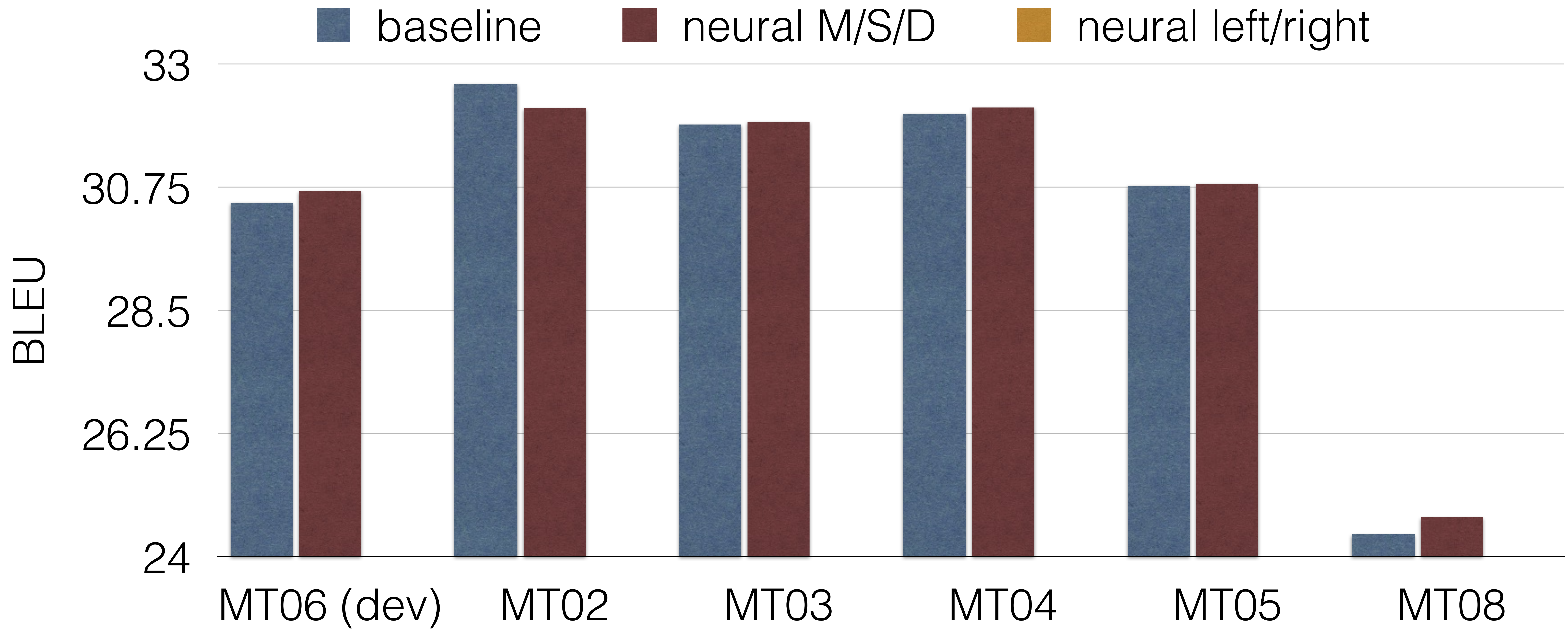


Translation

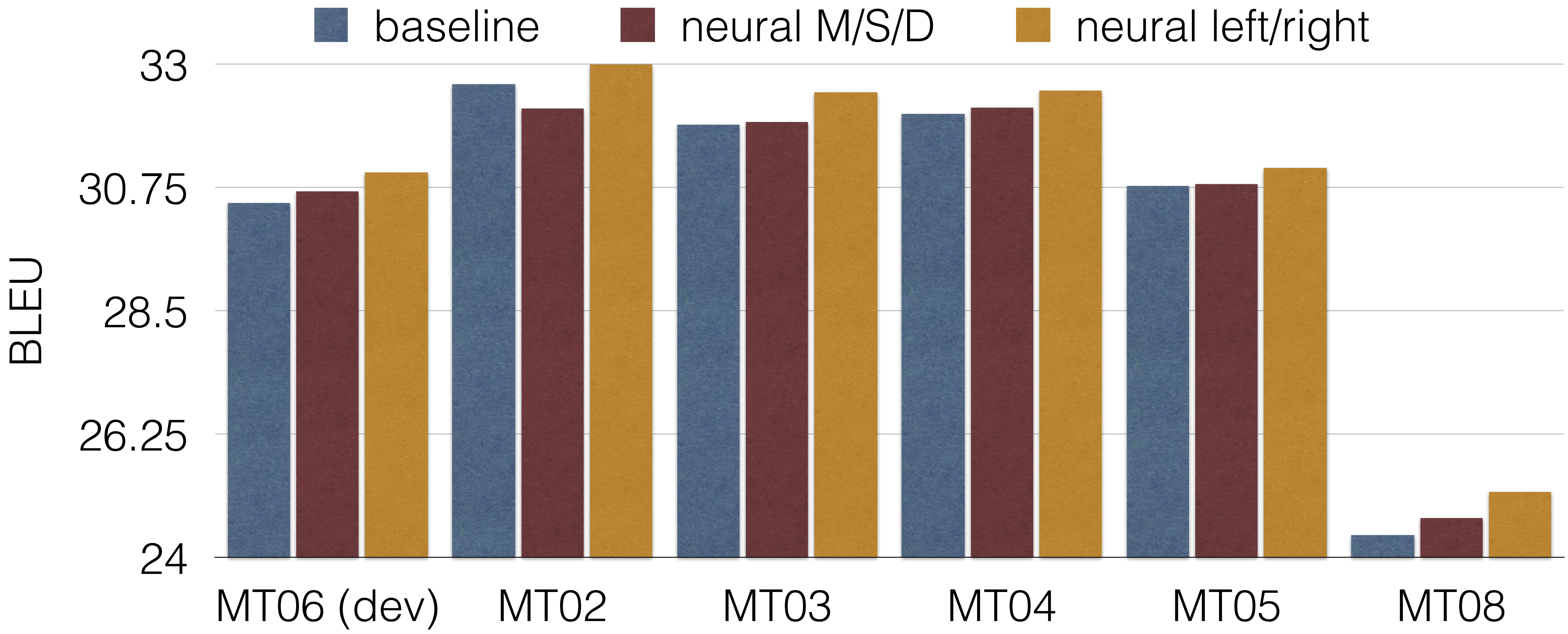
Translation



Translation

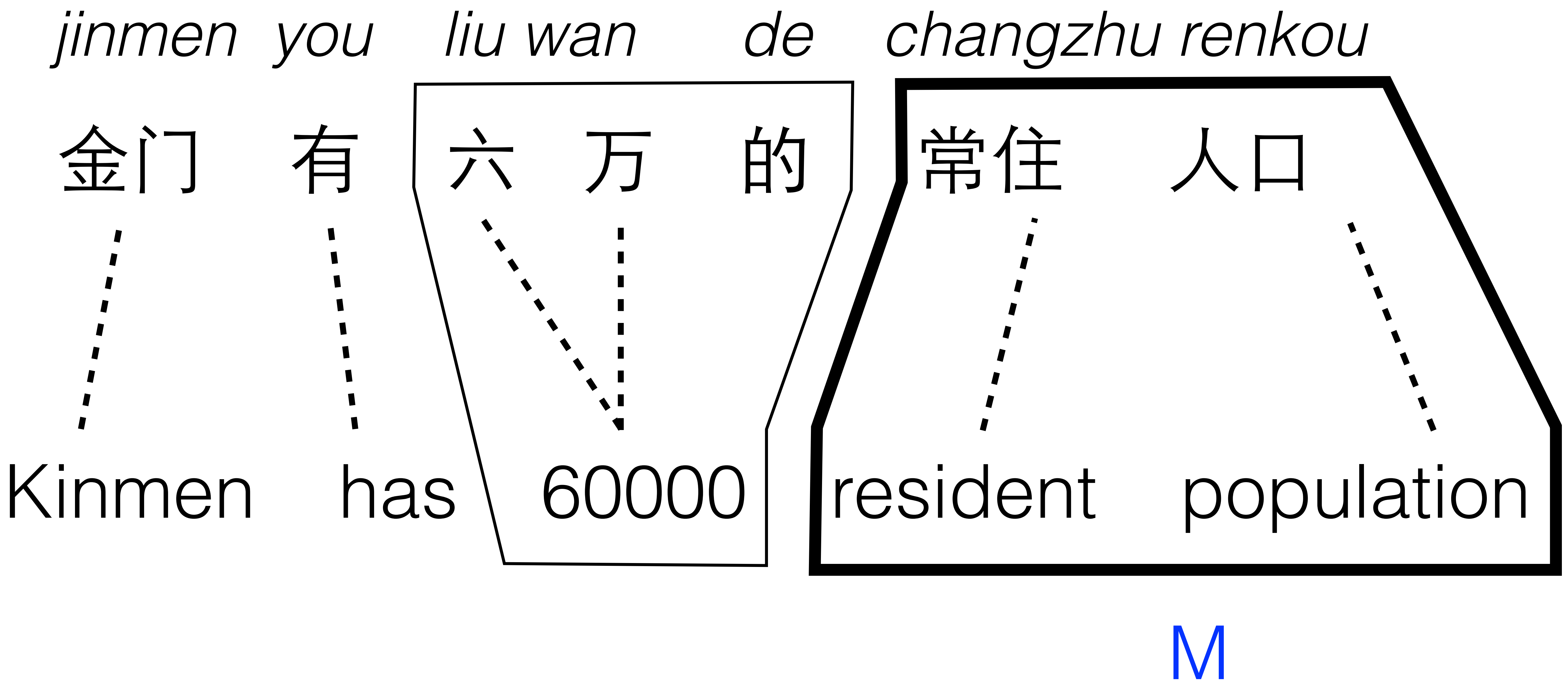


Translation



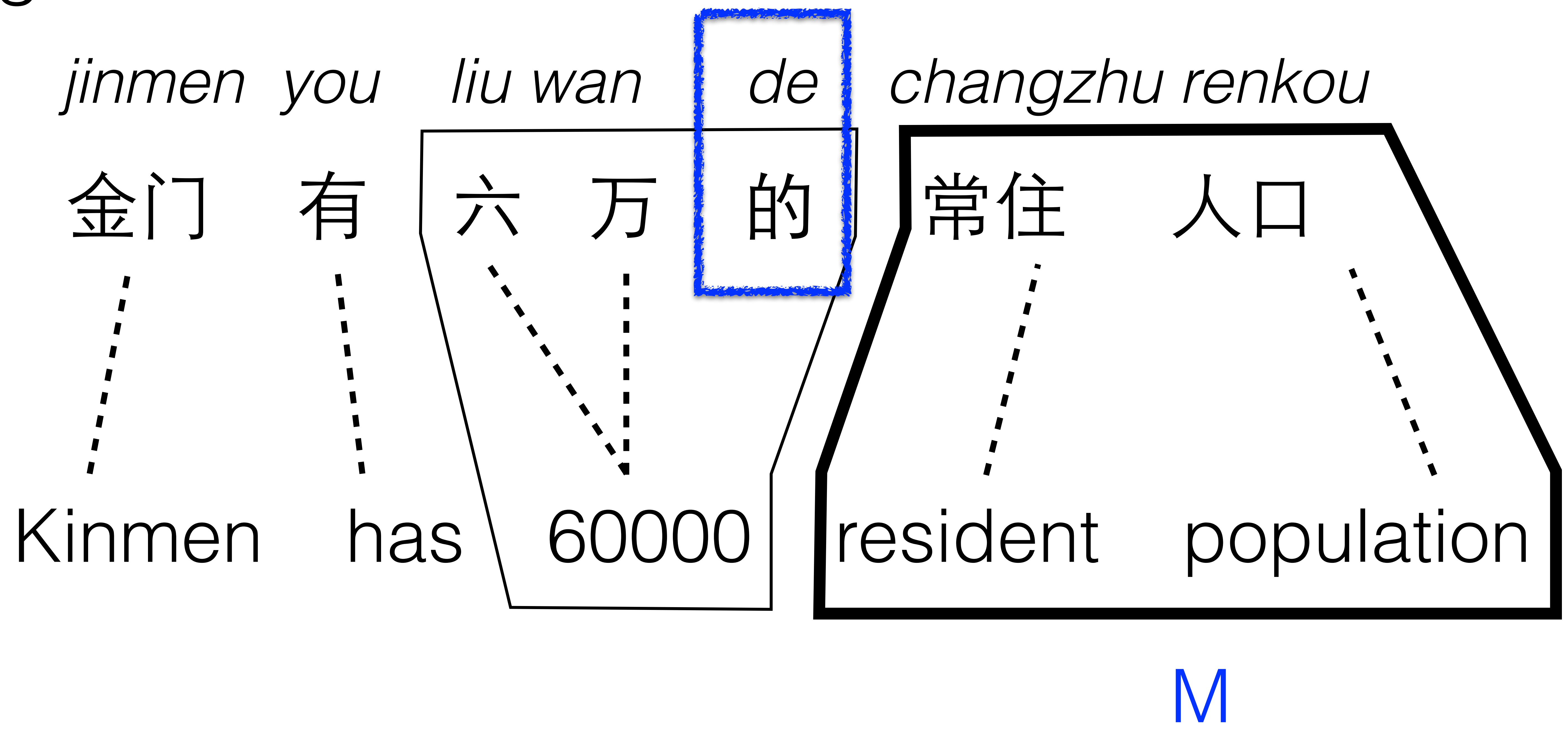
Non-Separability

- The unaligned Chinese word “de” makes a big difference in determining M/S/D orientations



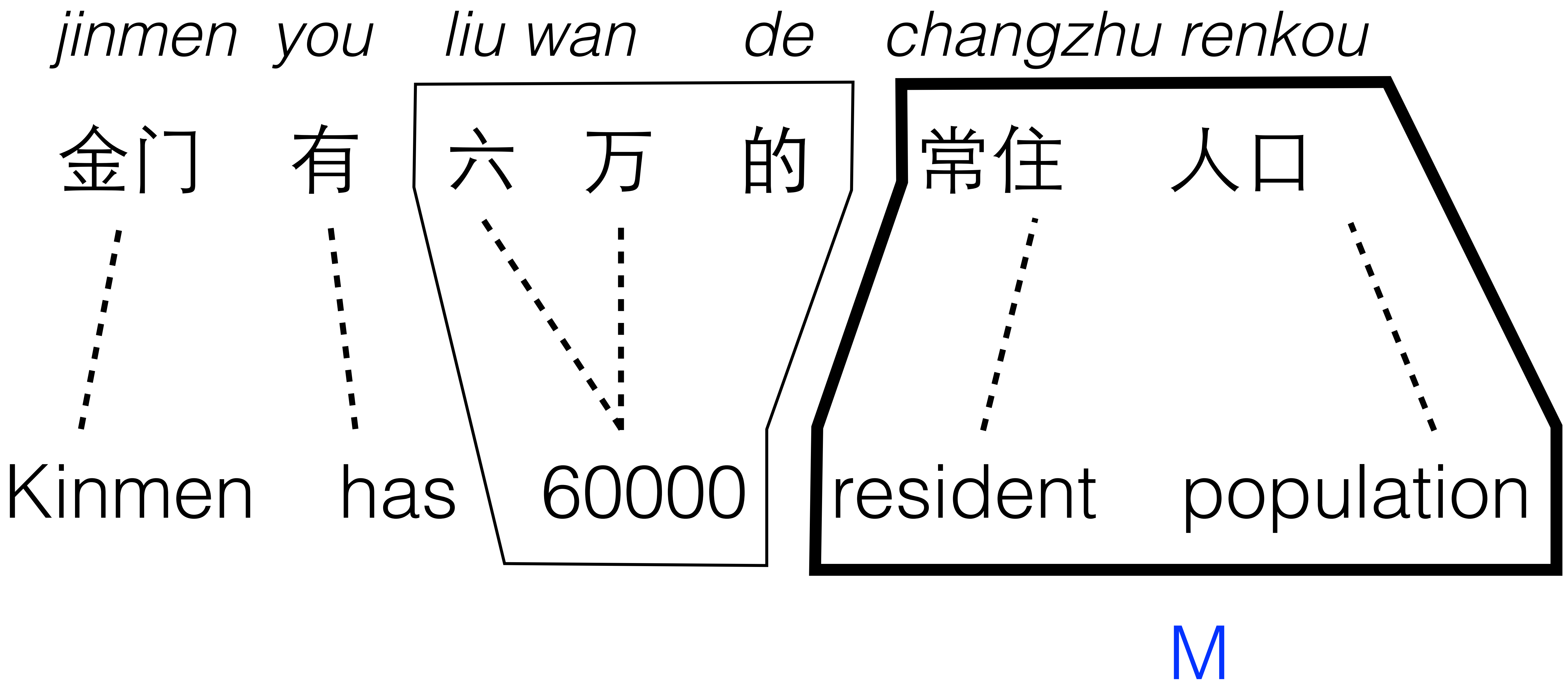
Non-Separability

- The unaligned Chinese word “de” makes a big difference in determining M/S/D orientations



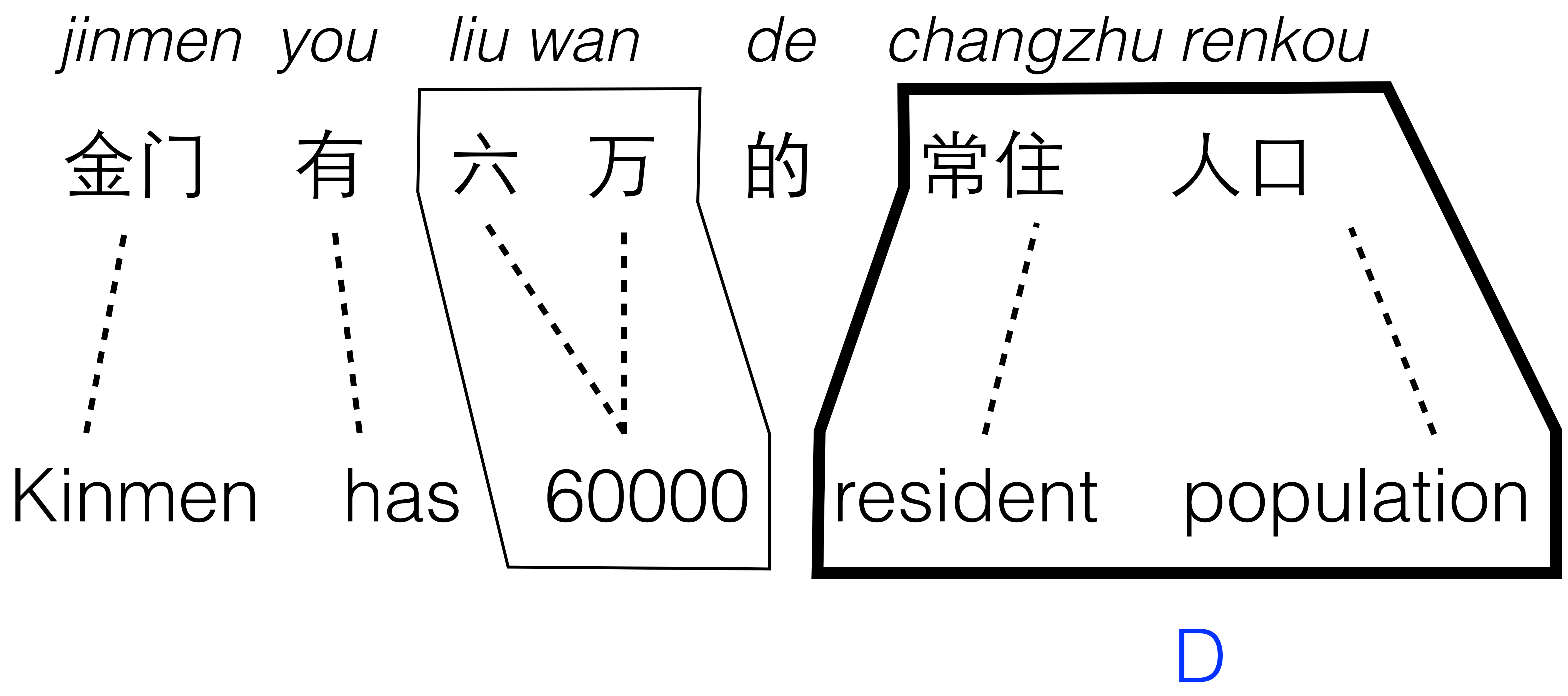
Non-Separability

- The unaligned Chinese word "de" makes a big difference in determining M/S/D orientations

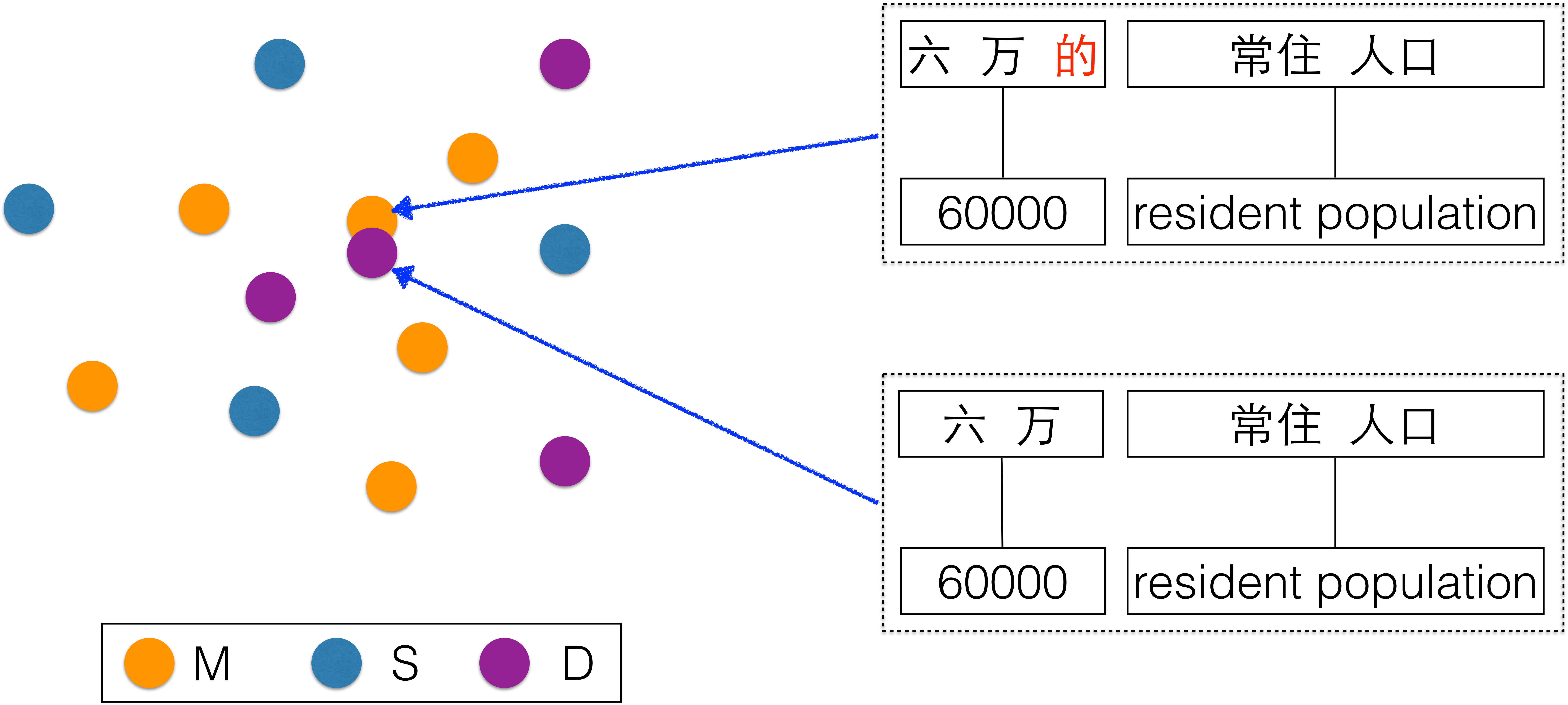


Non-Separability

- The unaligned Chinese word “de” makes a big difference in determining M/S/D orientations

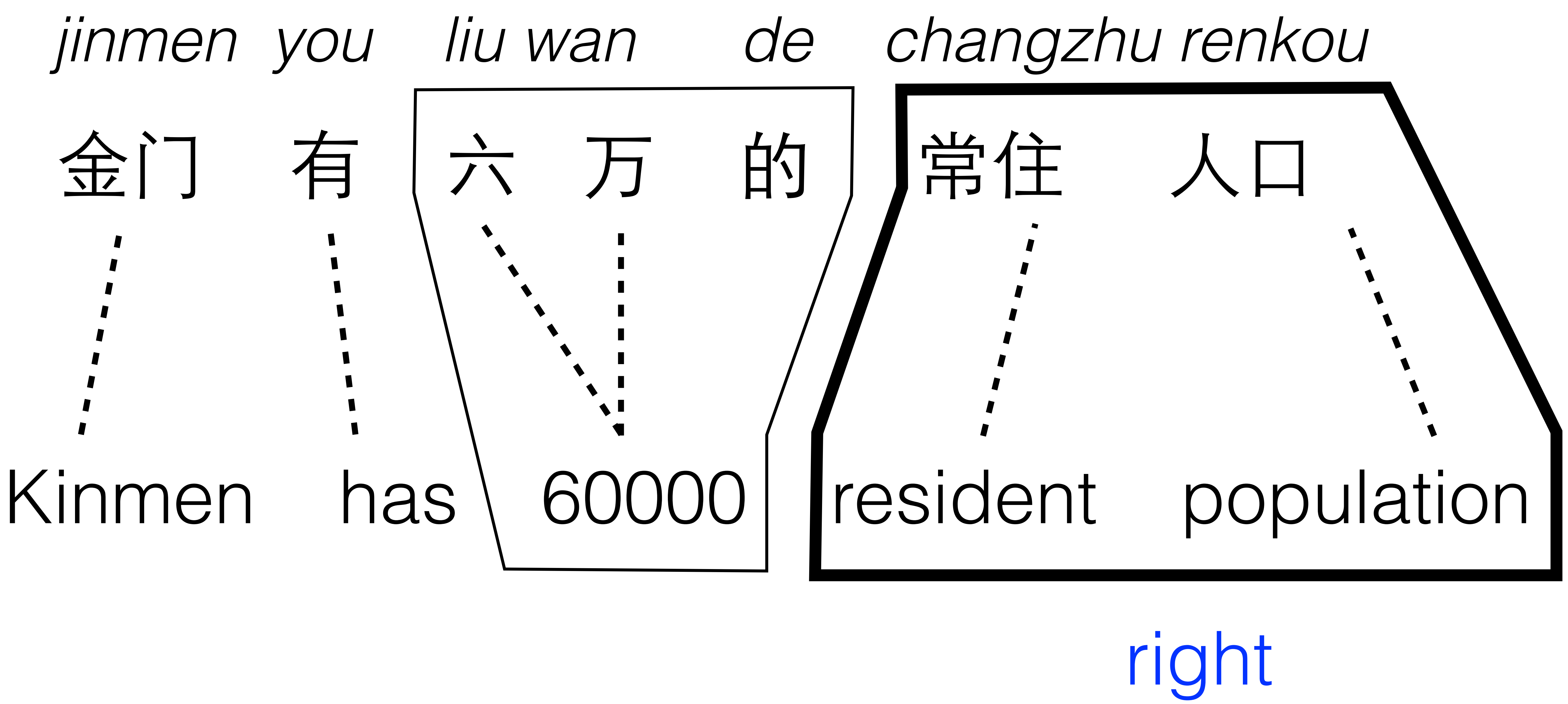


Non-Separability



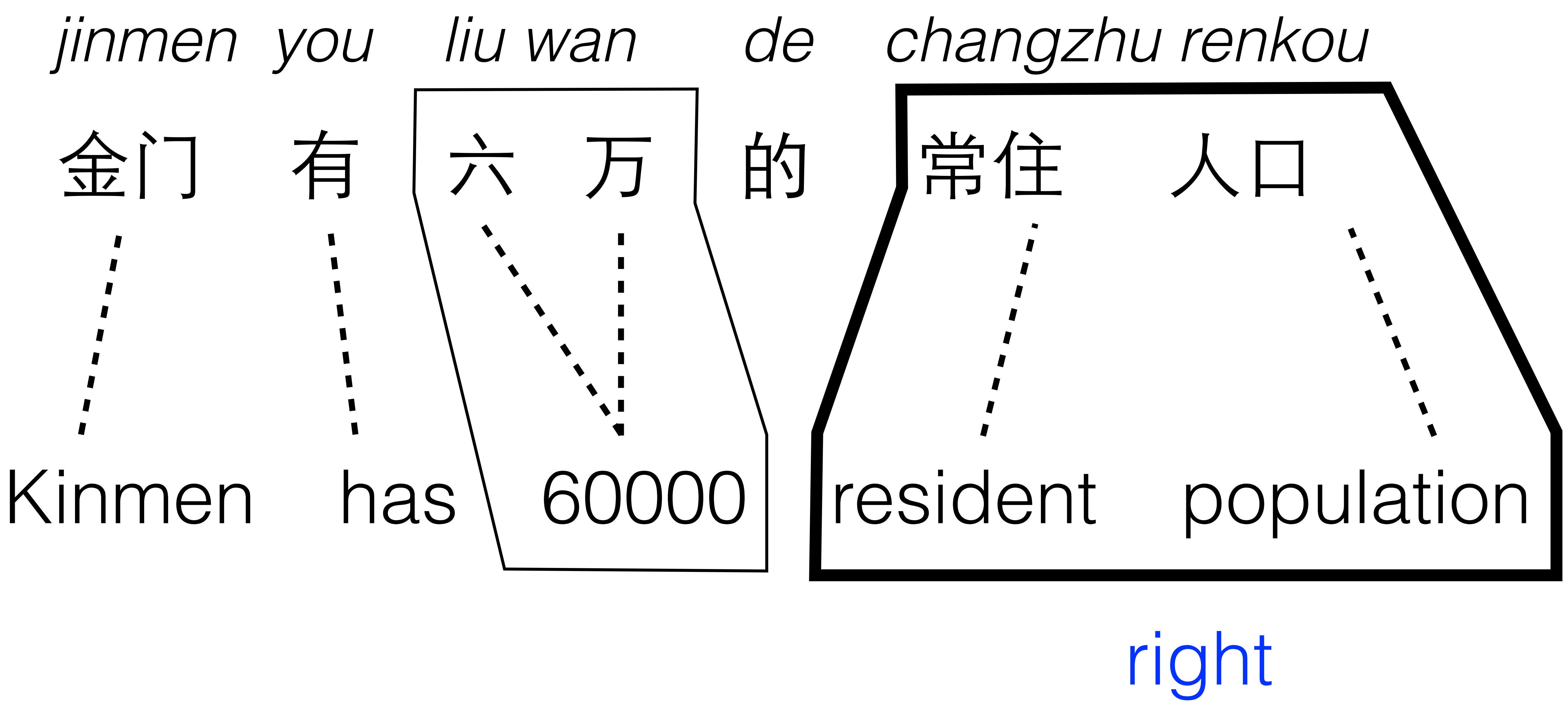
Non-Separability

- Left/right orientations are not so sensitive to unaligned words

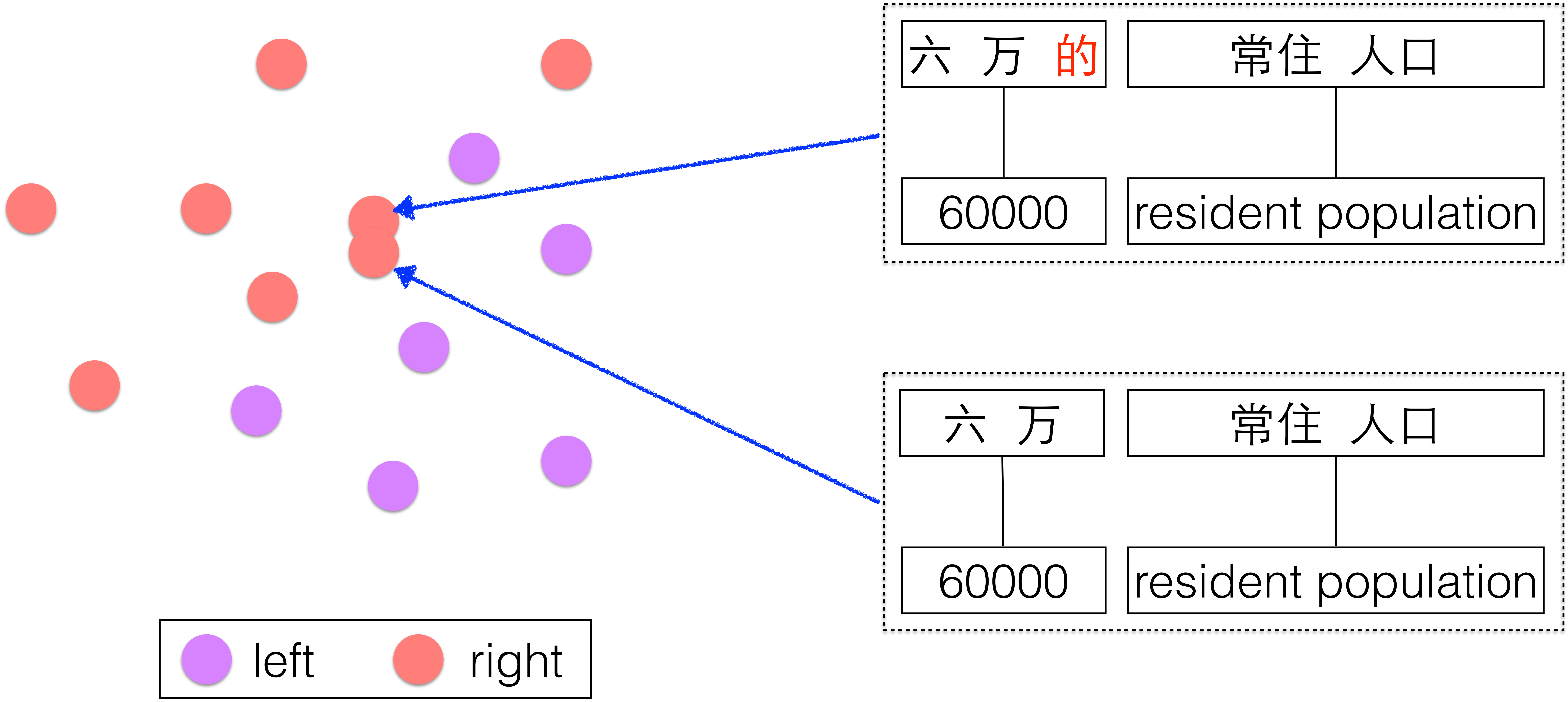


Non-Separability

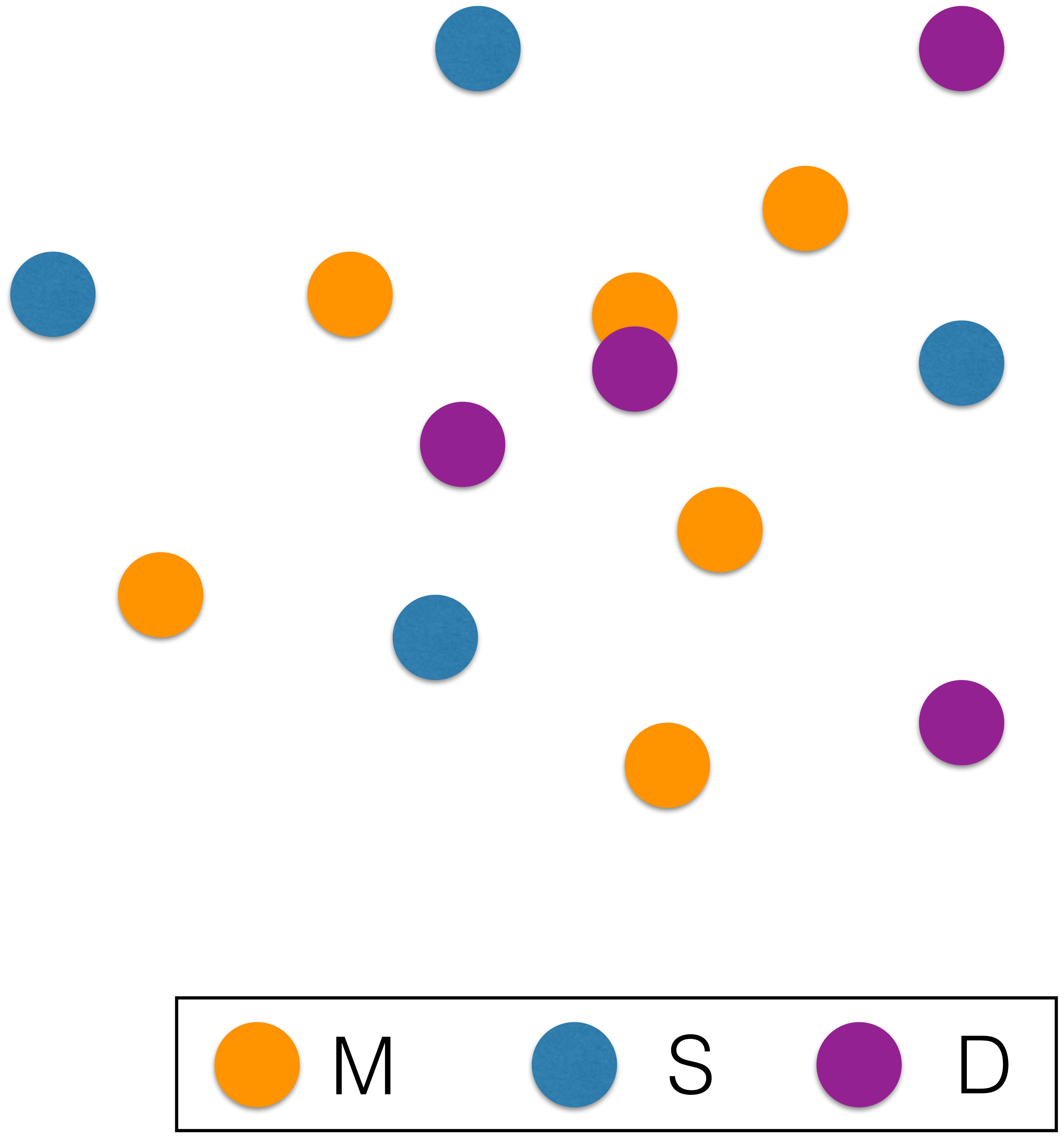
- Left/right orientations are not so sensitive to unaligned words

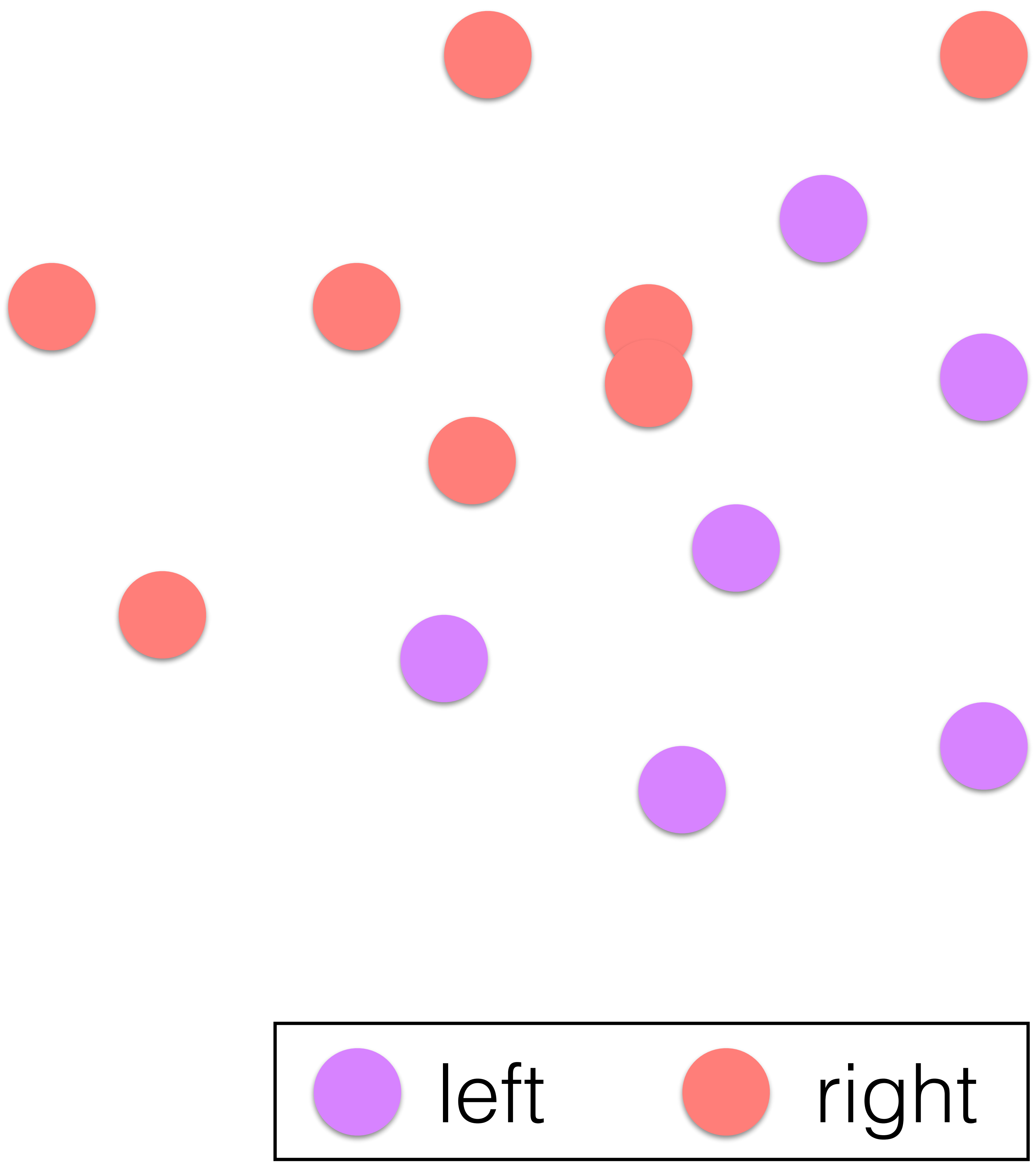


Non-Separability

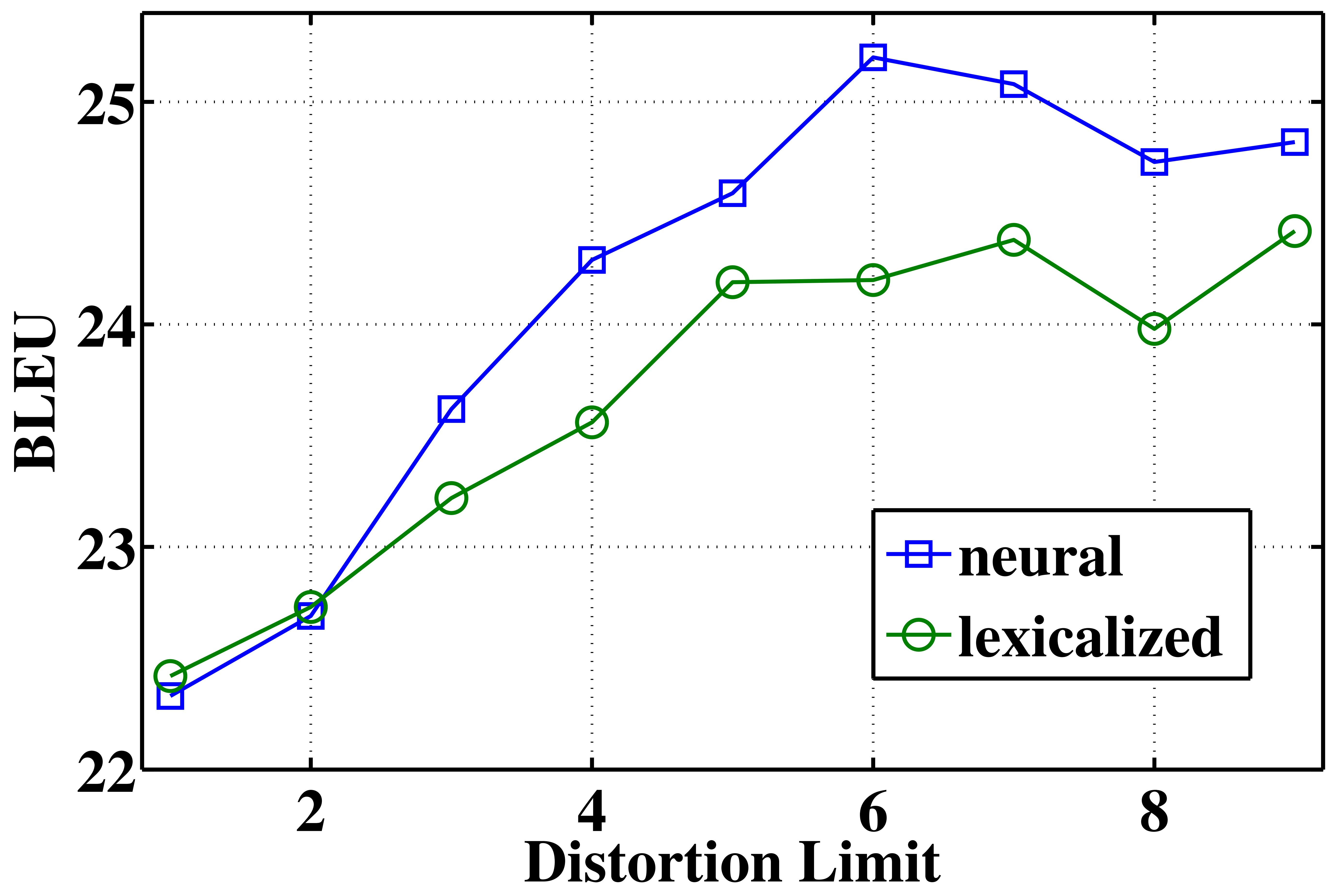


Non-Separability

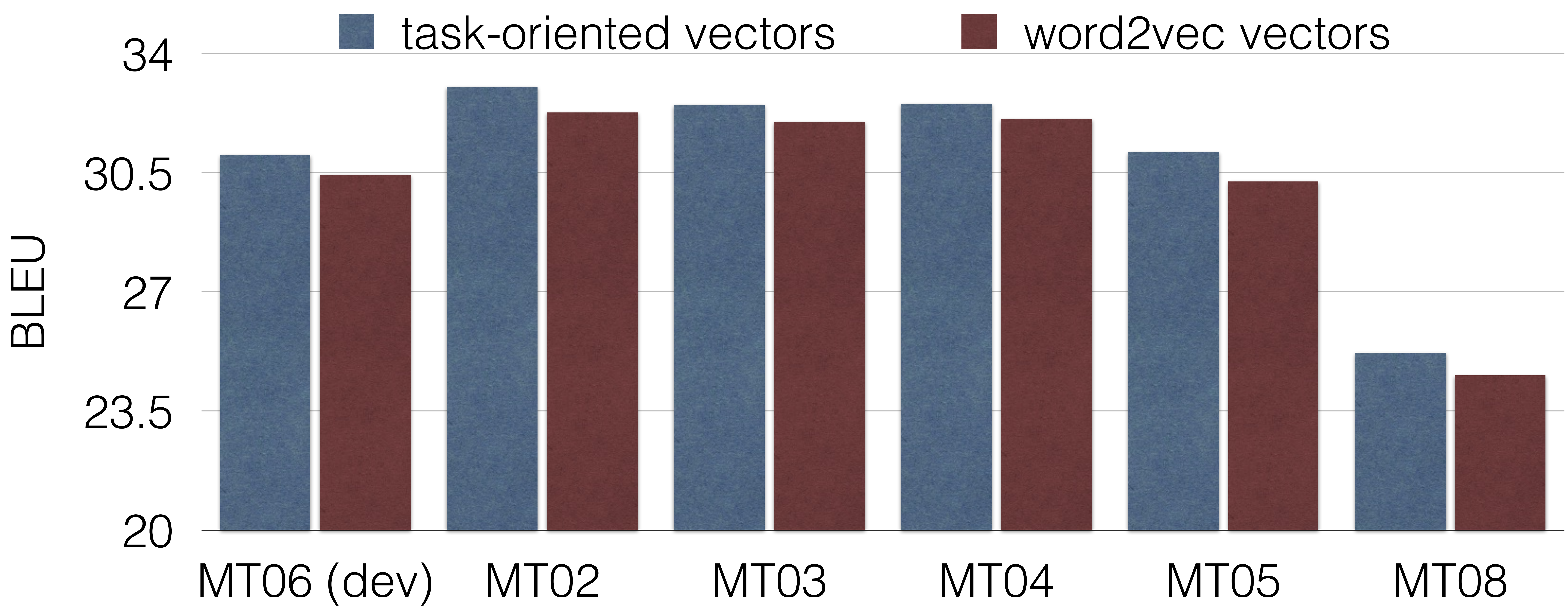




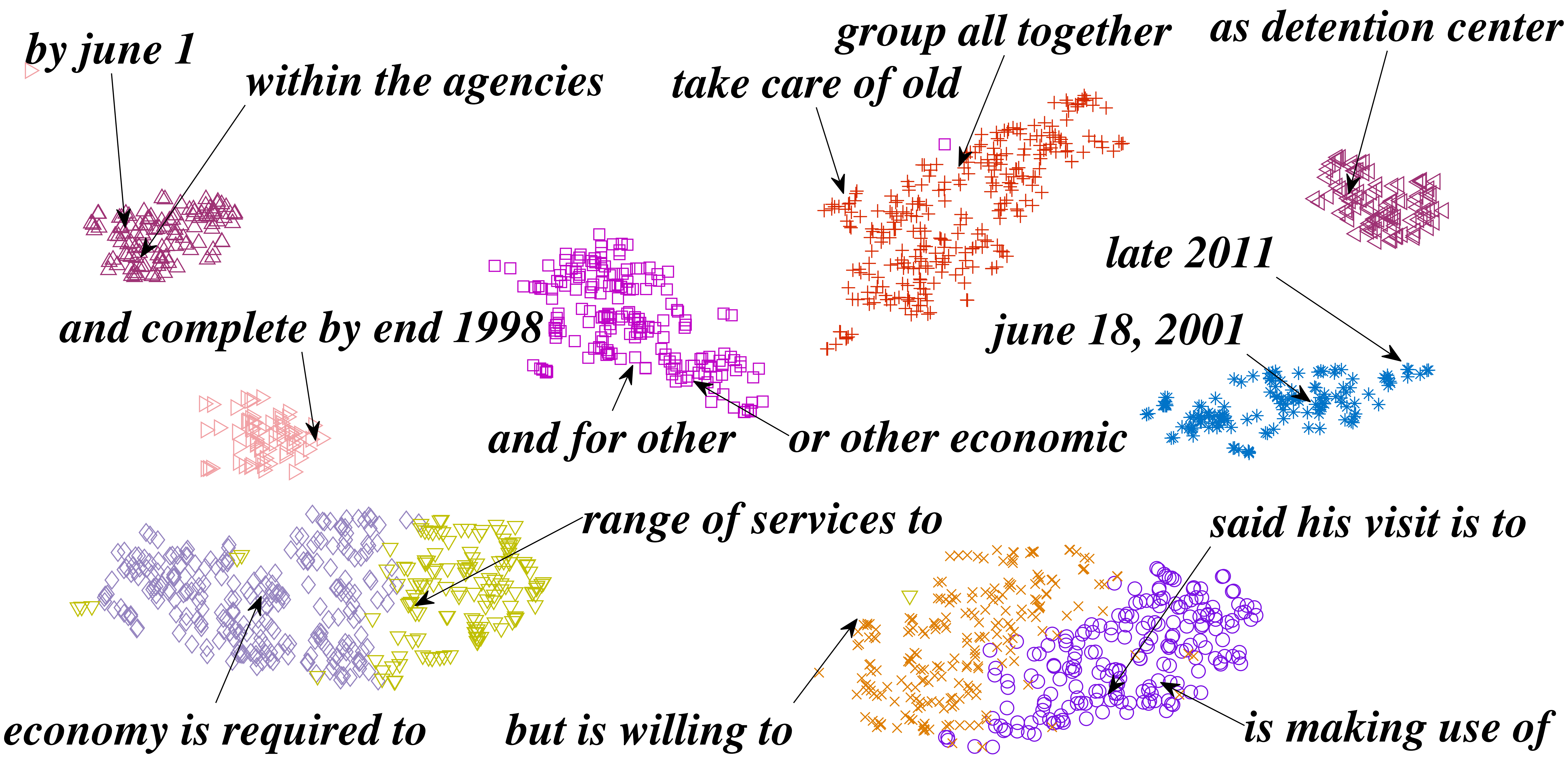
Distortion Limit



Word Vectors



Vector Space Representations



Conclusion

- We propose a neural reordering model for phrase-based translation
- It improves the context sensitivity, reduces ambiguity and alleviates the data sparsity problem

Conclusion

- We propose a neural reordering model for phrase-based translation
- It improves the context sensitivity, reduces ambiguity and alleviates the data sparsity problem
- Future work
  - Train MT system and neural classifier jointly
  - Develop more efficient models to leverage larger contexts
  - Extend our work to syntax-based and n-gram based models

Thanks!