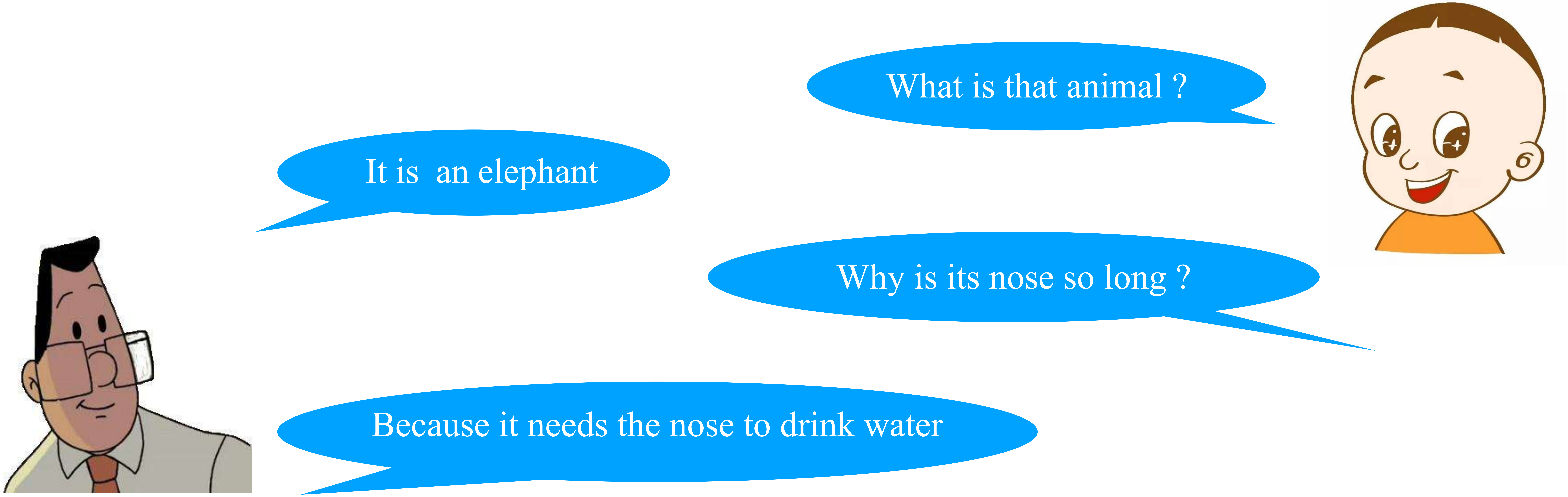
Conversational Word Embedding for Retrieval-based Dialog System

WENTAO MA, YIMING CUI, TING LIU, DONG WANG, SHIJIN WANG, GUOPING HU

ACL 2020

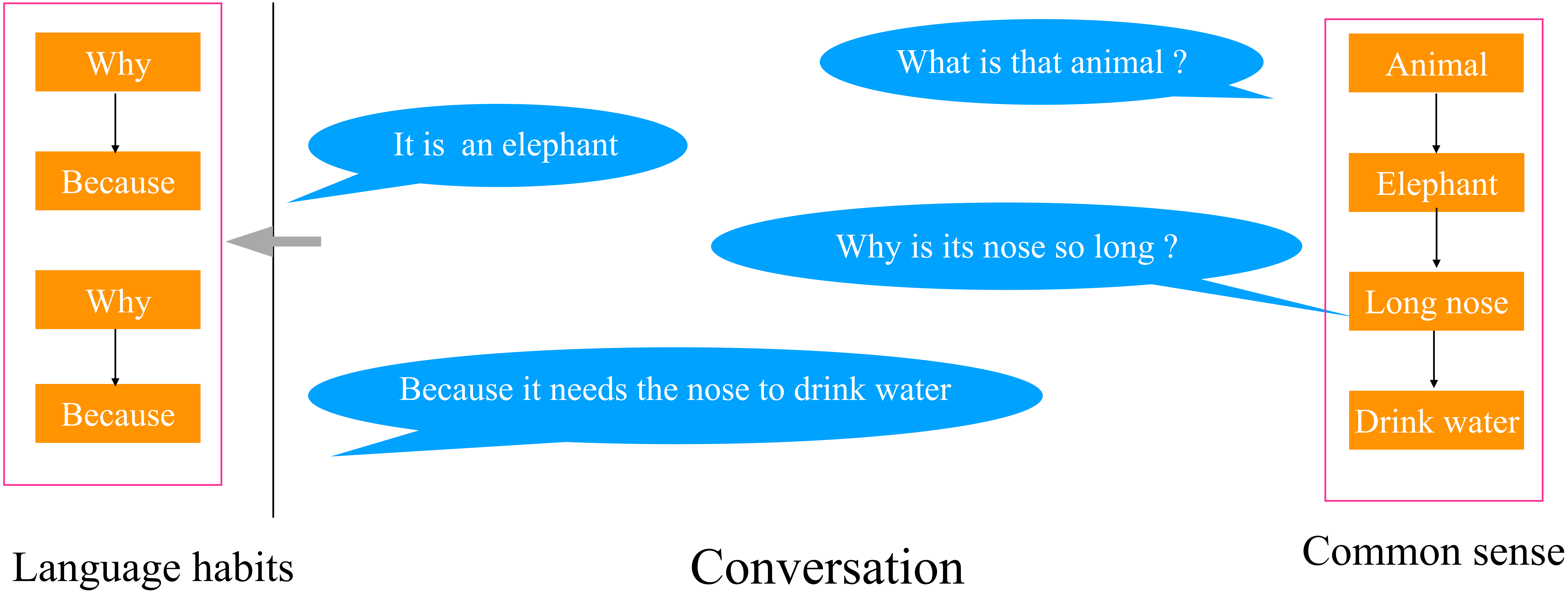
State Key Laboratory of Cognitive Intelligence, iFLYTEK Research,
Research Center for SCIR, Harbin Institute of Technology
iFLYTEK AI Research (Hebei)

Motivation



A Conversation between a father and his son in a zoo

Motivation

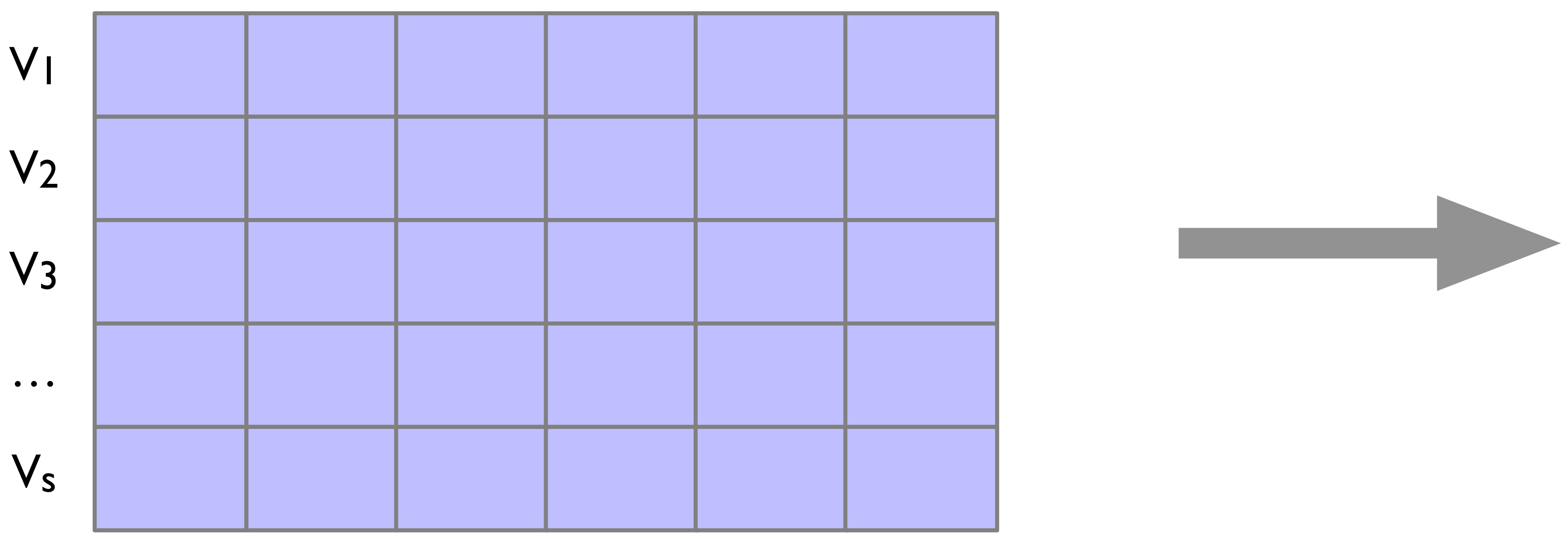

Why
Because
Why
Because
What is that animal ?
It is an elephant
Why is its nose so long ?
Because it needs the nose to drink water
Animal
Elephant
Long nose
Drink water
Language habits
Conversation
Common sense

Motivation

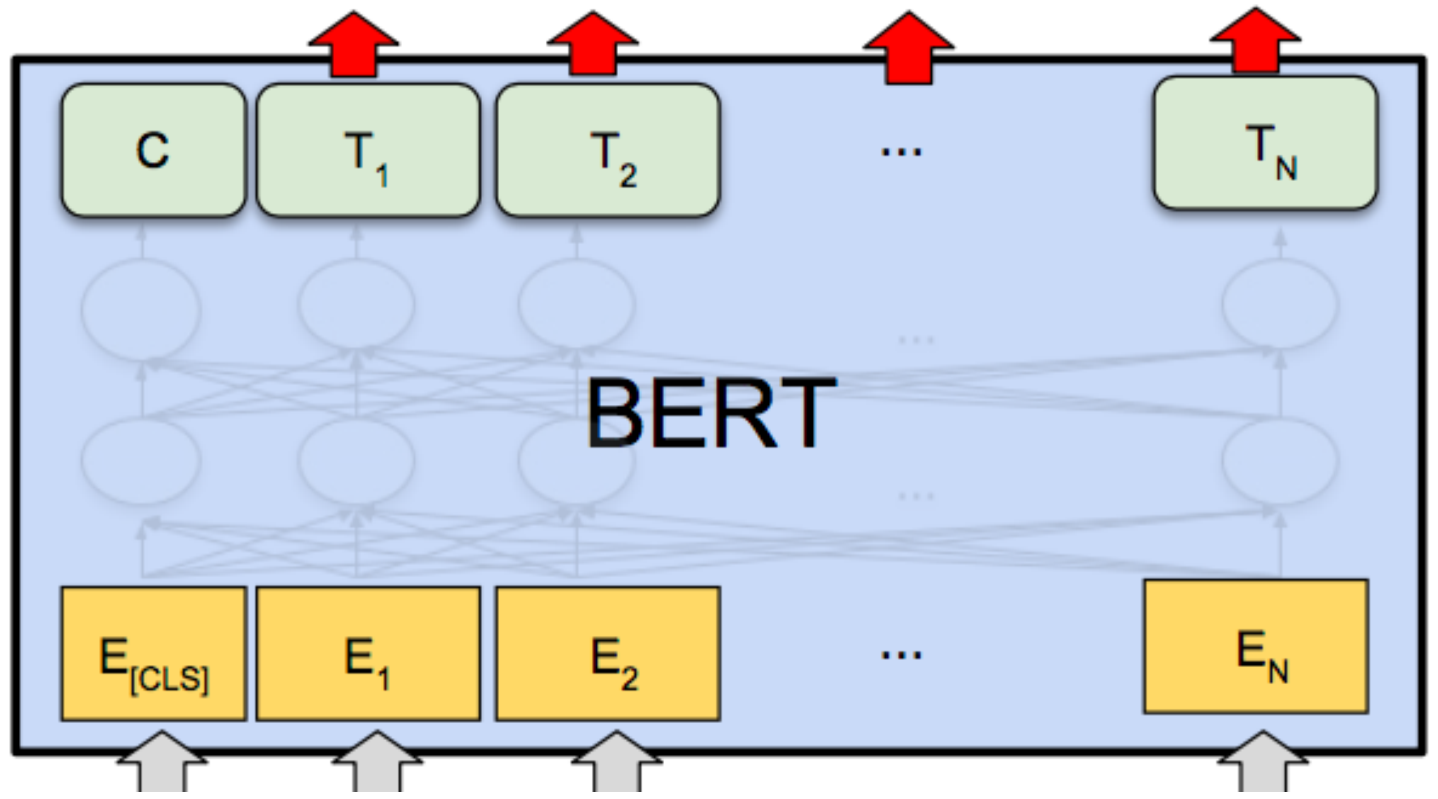
- Human conversations contain many types of information, e.g., common sense, language habits and knowledge.
  - cross-sentence: exist in conversation pair instead of single sentence
  - asymmetric: some language habits are directional, such as
    - ‘why’ $\longrightarrow$ ‘because’,
    - ‘congratulation’ $\longrightarrow$ ‘thanks’

Related works

- Word representation methods
  - Static word embedding: Word2vec, GloVe, fastText…
  - Contextual word embedding: ELMo, BERT, XLNet…



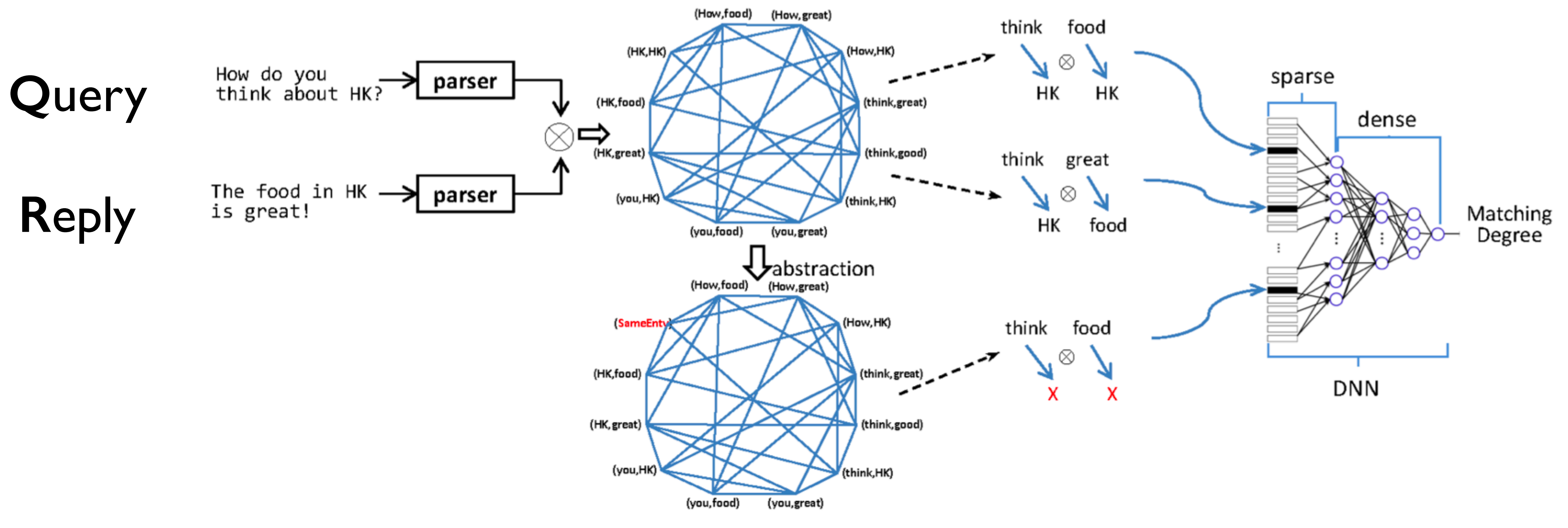
Embedding Matrix



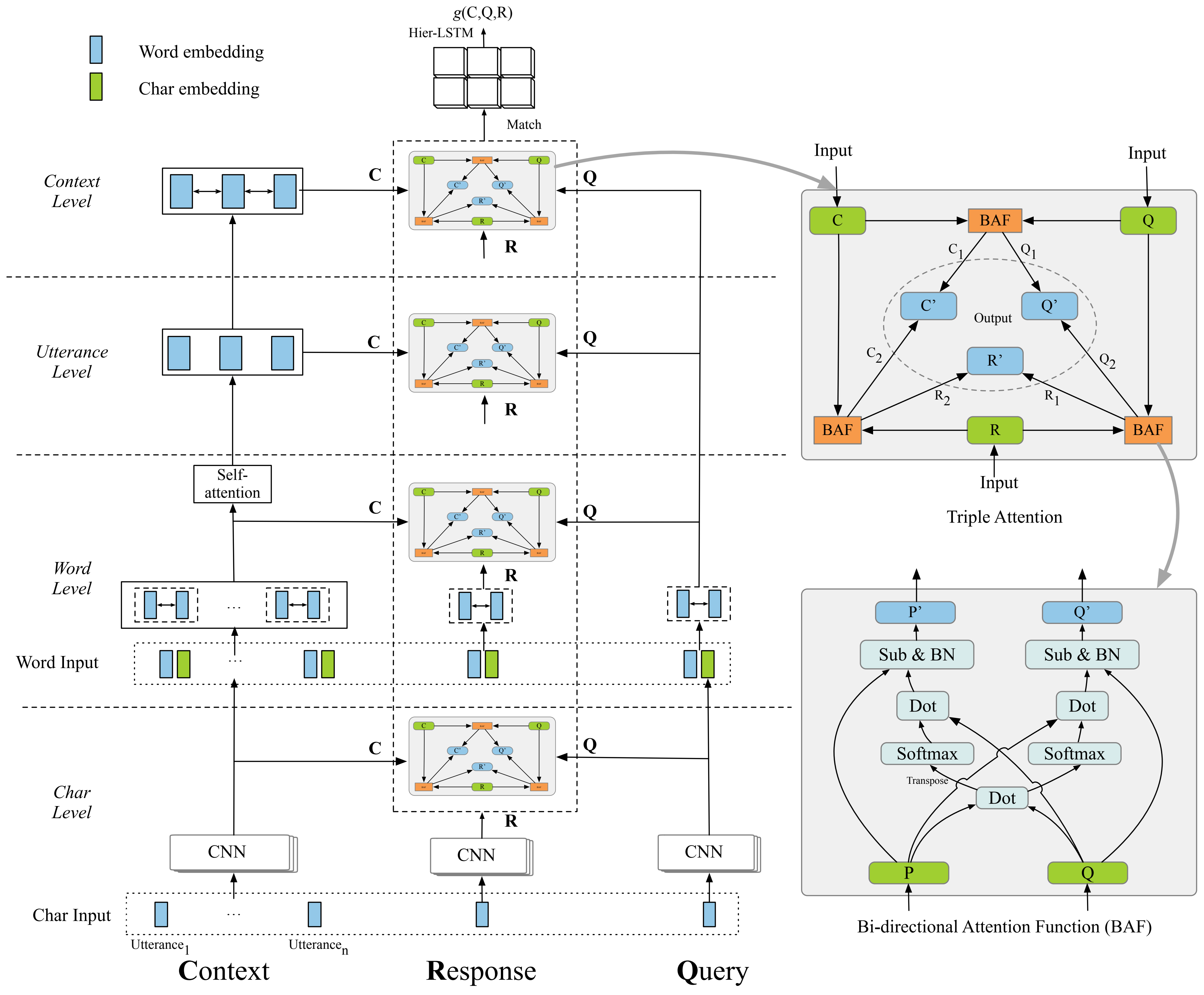
Pretraining Model

Related works

- Retrieval-based Dialog System
  - Single-turn Response Selection
  - Multi-turn Response Selection



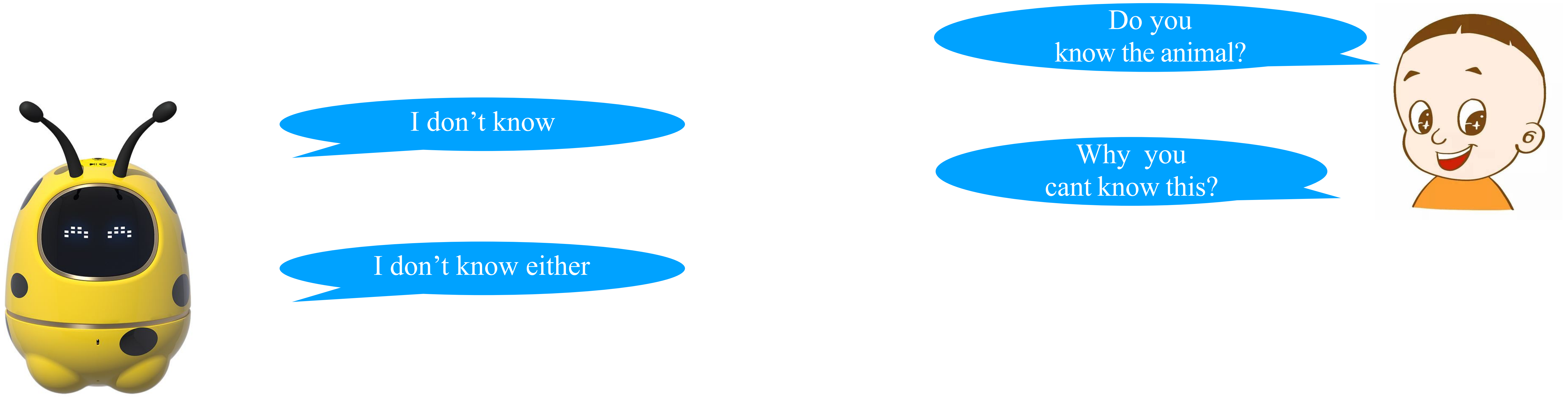
Single-turn g(Q,R)



Multi-turn g(C,Q,R)

Motivation

- Previous word embedding methods for conversation
  - Single sentence: the semantic correlation beyond a single sentence is missing
  - Single vector space: map the post and reply into the same vector space, which leads the reply with repeated words is easy to be selected



Contribution

PR-Embedding: learn conversational word embedding from conversation pairs in two different vector spaces.

Notation

Vocabulary

$$V^p := \{v_1^p, v_2^p, ..., v_s^p\}$$

$$V^r := \{v_1^r, v_2^r, ..., v_s^r\}$$

Embedding Matrix

$$\mathbf{E_p} = \begin{array}{c} v_1^p \\ v_2^p \\ v_3^p \\ \dots \\ v_s^p \end{array} \begin{bmatrix} \; & \; & \; & \; \\ \; & \; & \; & \; \\ \; & \; & \; & \; \\ \; & \; & \; & \; \\ \; & \; & \; & \; \end{bmatrix}$$

$$\mathbf{E_r} = \begin{array}{c} v_1^r \\ v_2^r \\ v_3^r \\ \dots \\ v_s^r \end{array} \begin{bmatrix} \; & \; & \; & \; \\ \; & \; & \; & \; \\ \; & \; & \; & \; \\ \; & \; & \; & \; \\ \; & \; & \; & \; \end{bmatrix}$$

Sequence

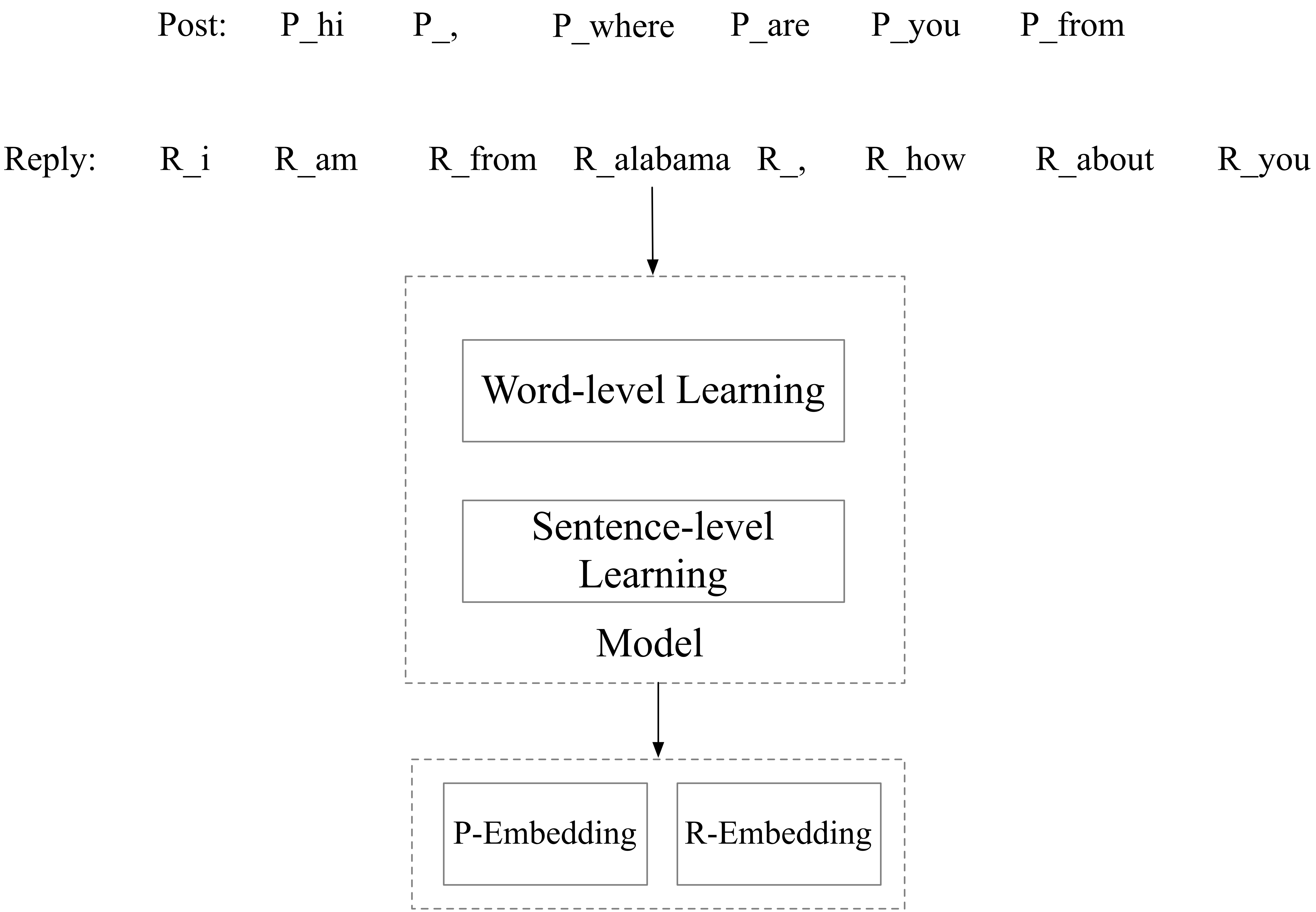
$$P = (p_1, ..., p_m)$$

$$R = (r_1, ..., r_n)$$

Post

Reply

Model


Post: P_hi P_, P_where P_are P_you P_from
Reply: R_i R_am R_from R_alabama R_, R_how R_about R_you
Word-level Learning
Sentence-level Learning
Model
P-Embedding
R-Embedding

Word-level Learning

Post: P_hi P_, P_where P_are P_you P_from

Reply: R_i R_am R_from R_alabama R_, R_how R_about R_you

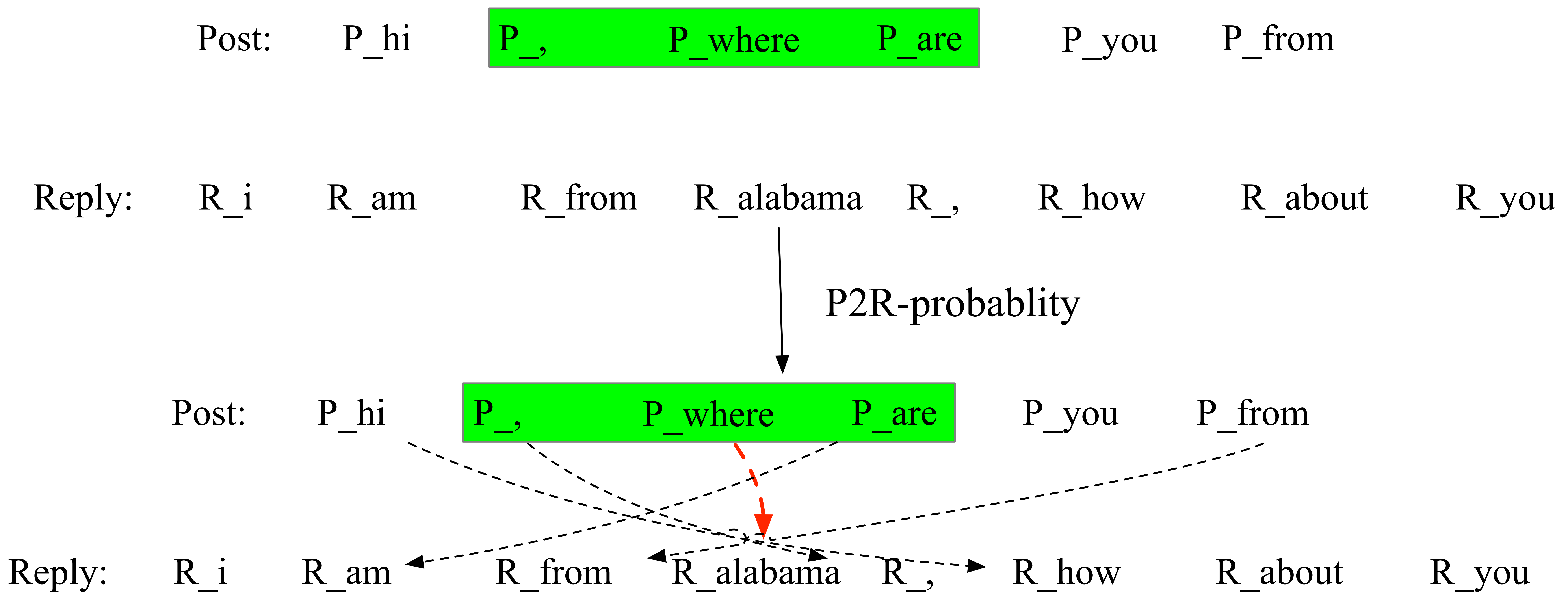
Model

Post: P_hi P_, P_where P_are P_you P_from

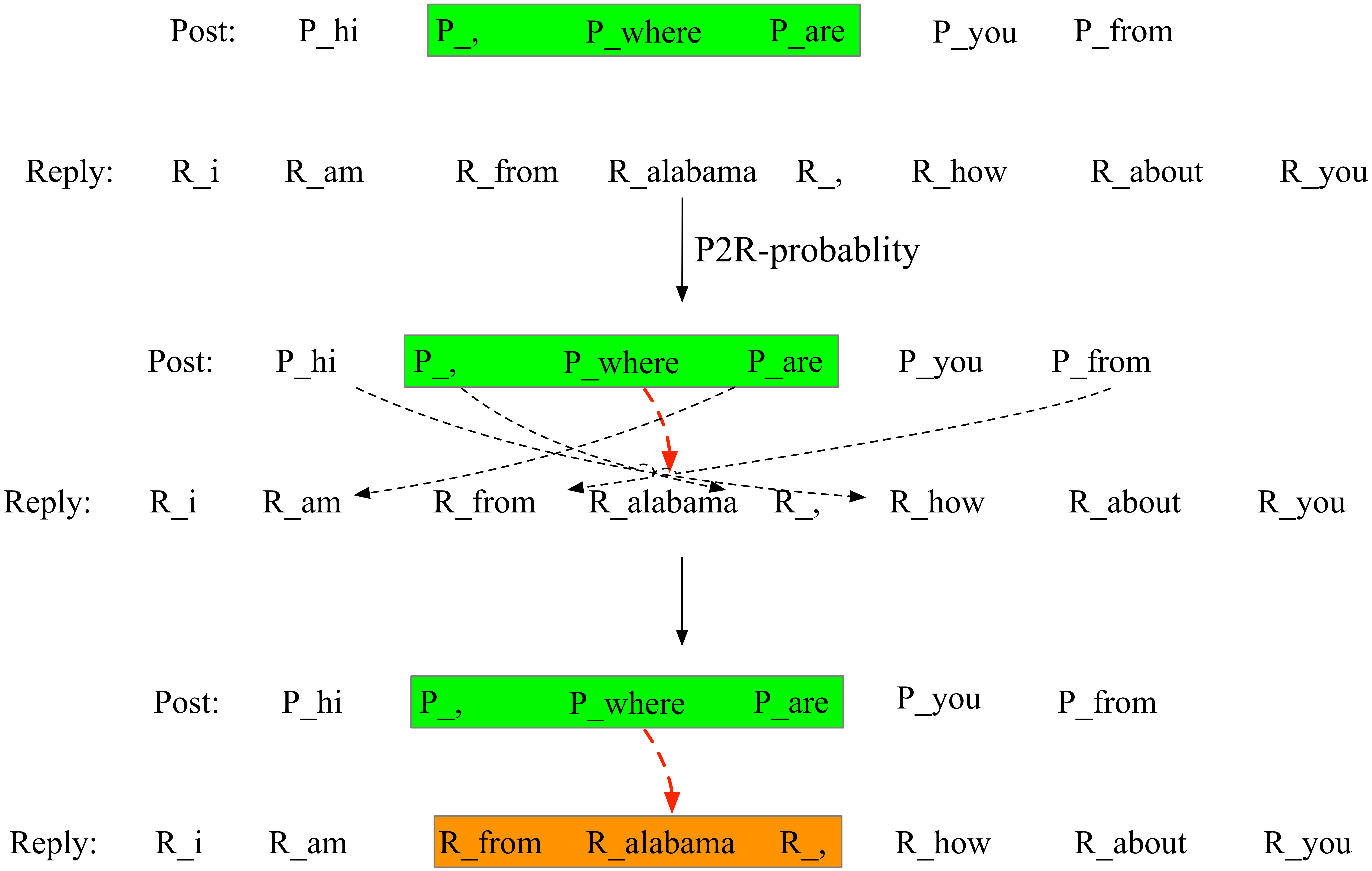
Reply: R_i R_am R_from R_alabama R_, R_how R_about R_you

How to generate the cross-sentence co-occurrence window ?

Model


Post: P_hi P_, P_where P_are P_you P_from
Reply: R_i R_am R_from R_alabama R_, R_how R_about R_you
P2R-probablity
Post: P_hi P_, P_where P_are P_you P_from
Reply: R_i R_am R_from R_alabama R_, R_how R_about R_you

Model



Post: P_hi P_, P_where P_are P_you P_from
Reply: R_i R_am R_from R_alabama R_, R_how R_about R_you
P2R-probablity
Post: P_hi P_, P_where P_are P_you P_from
Reply: R_i R_am R_from R_alabama R_, R_how R_about R_you
Post: P_hi P_, P_where P_are P_you P_from
Reply: R_i R_am R_from R_alabama R_, R_how R_about R_you

Model

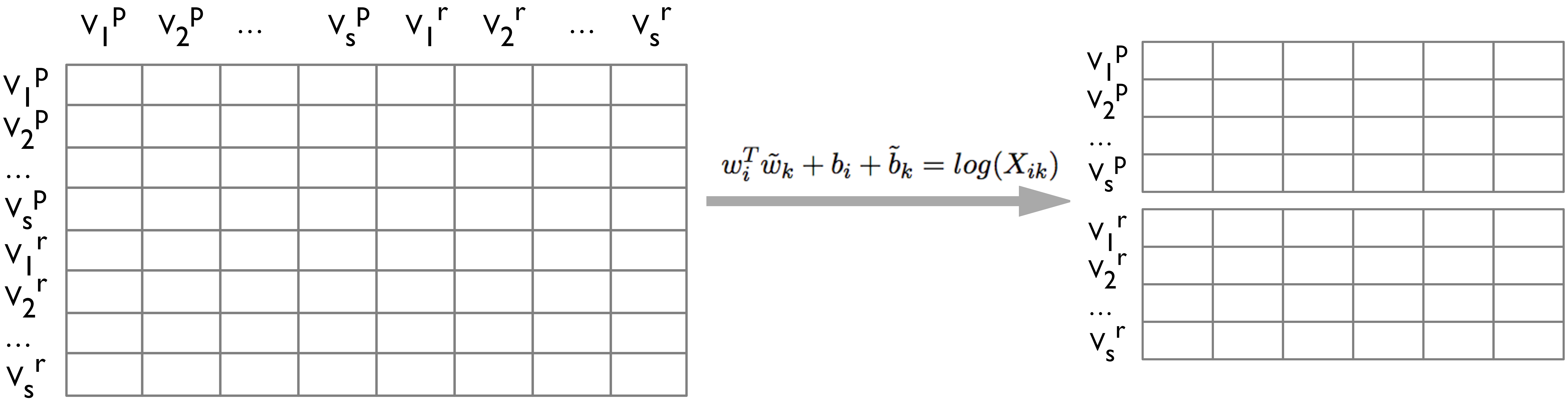
Word-level Learning

	v_1^p	v_2^p	…	v_s^p	v_1^r	v_2^r	…	v_s^r
v_1^p								
v_2^p								
…								
v_s^p								
v_1^r								
v_2^r								
…								
v_s^r								

Word-level Co-occurrence

Model

Word-level Learning



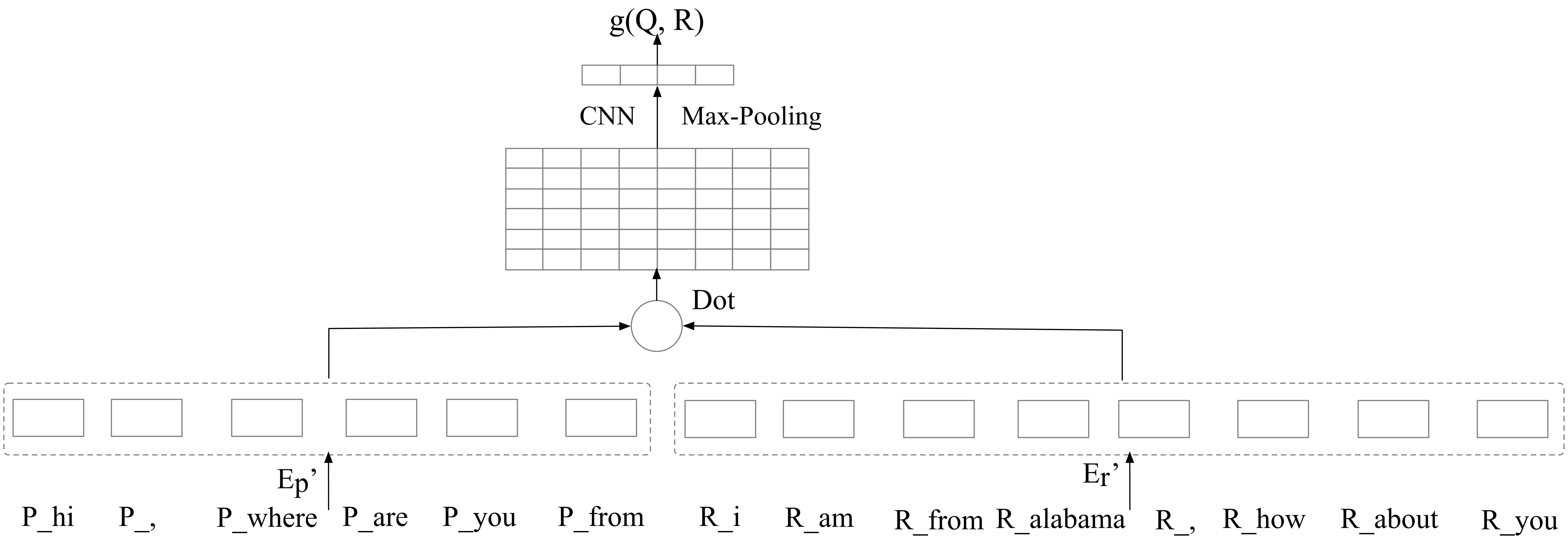
$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = log(X_{ik})$$

Word-level Co-occurrence

Embedding Matrix Ep', Er'

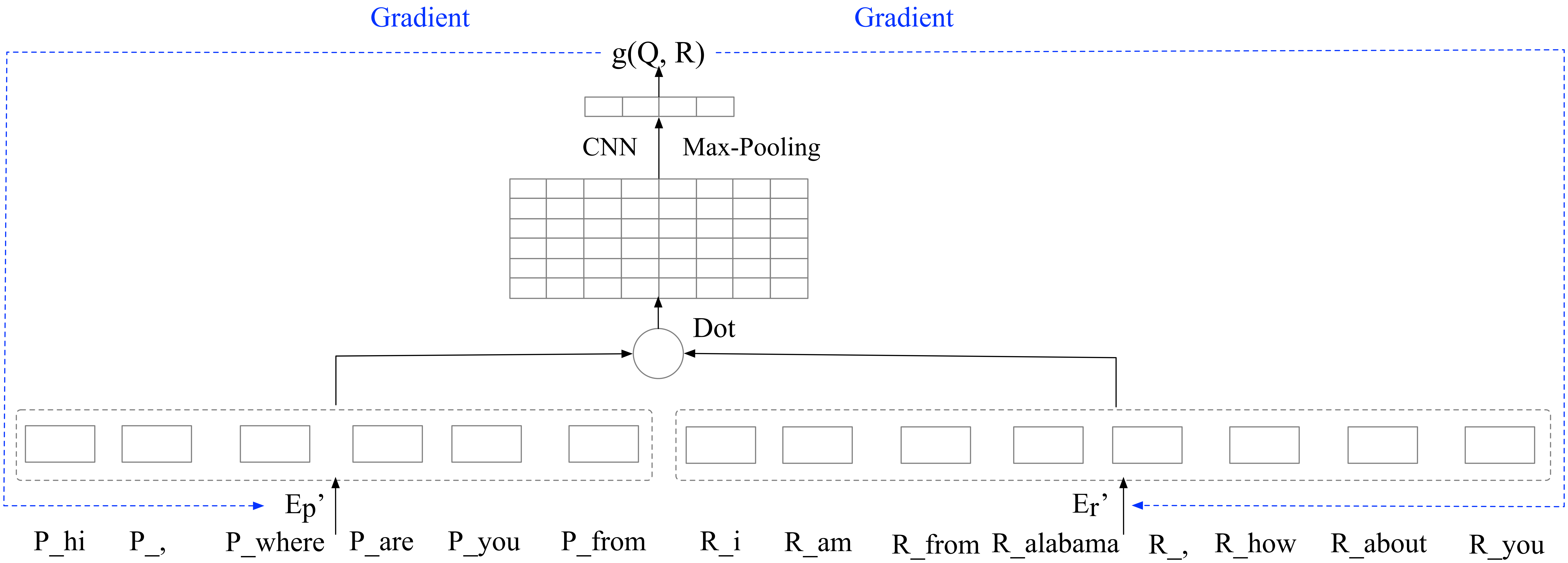
Model

Sentence-level Learning



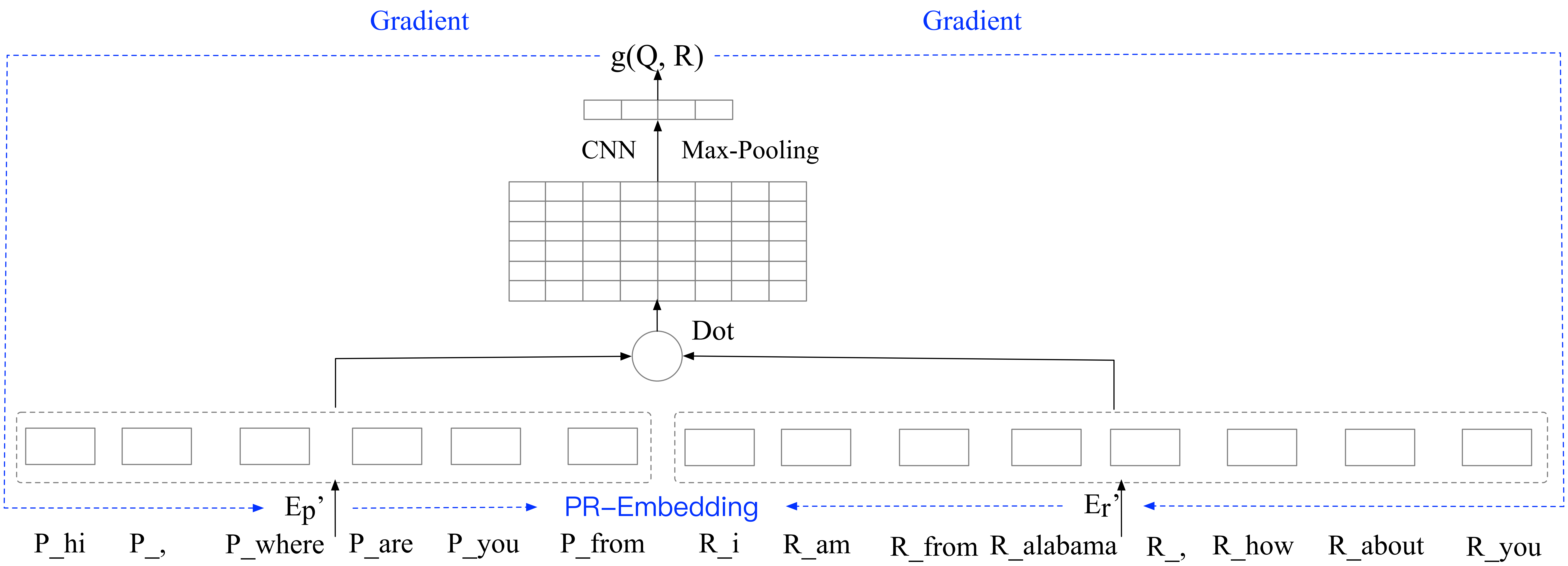
Model

Sentence-level Learning



Model

Sentence-level Learning



Experiment

- Datasets
  - PersonaChat dataset (Zhang et al., 2018)
    - English, multi-turn conversation dataset with profile
    - Train/Dev/Test: 133.5k/15.7k/15.1k utterance
    - Evaluation Metrics: hit@k
  - In-house conversation dataset
    - Chinese, single-turn conversation dataset
    - Test: 935 posts and 12,767 candidate replies (label with ‘good, middle, bad’)

      Train: 1.07 million pairs after cleaning, from Baidu Zhidao
    - Evaluation Metrics: NDCG, P@1

Experiment

- Result on PersonaChat
  - Single-turn task: compare the embeddings based on BOW (bag-of-words, the average of all word embeddings), only use the current query for prediction
  - Multi-turn task: compare the embeddings based on a neural network KVMemnn, use all the context for prediction

	hits@1	hits@5	hits@10
GloVe_{train}	12.6	39.6	63.7
GloVe_{emb}	18.0	44.6	66.9
BERT_{emb}	15.4	41.0	62.9
Fasttext_{emb}	17.8	44.9	67.2
PR-Embedding	**22.4**	**60.0**	**81.1**
IR baseline†	21.4	-	-
Starpace†	31.8	-	-
Profile Memory†	31.8	-	-
KVMemnn	32.3	62.0	79.2
+PR-Embedding	35.9	66.1	82.6
KVMemnn (GloVe)	36.8	68.1	83.6
+PR-Embedding	**39.9**	**72.4**	**87.0**

Experiment

- Result on In-house dataset
  - Single-turn task, compare with GloVe and the public embedding of DSG.
  - P@1(s): only use the candidate reply labeled with 'good' as true
- Ablation study
  - w/o PR: change the two vector spaces with the single one, just as the previous method
  - w/o SLL: remove the sentence-level learning

	NDCG	NDCG@5	P@1	P@1(s)
GloVe_{train}	69.97	48.87	51.23	33.48
DSG_{emb}	70.82	50.45	52.19	35.61
BERT_{emb}	70.06	48.45	51.66	35.08
PR-Emb	**74.79**	**58.16**	**62.03**	**45.99**
w/o PR	70.68	50.60	50.48	35.19
w/o SLL	71.65	52.03	53.48	40.86

Analysis

- Nearest tokens
  - Four nearest tokens for the three selected words in the whole vector space
  - For PR-Embedding, we select the words from the post vocabulary and give the nearest words both in post and reply space

Why			Thanks			Congratulations		
GloVe	P-Emb	R-Emb	GloVe	P-Emb	R-Emb	GloVe	P-Emb	R-Emb
why	why	because	thanks	thanks	welcome	congratulations	congratulations	thank
know	understand	matter	thank	asking	problem	congrats	ah	thanks
guess	oh	idea	fine	thank	today	goodness	fantastic	appreciate
so	probably	reason	asking	good	bill	yum	bet	problem

Summary

- We proposed a conversational word embedding method PR-Embedding, which is learned from conversational pairs in two different spaces;
- We introduce the word alignment model from SMT to generate the cross-sentence window, and train the embedding in word and sentence level;
- The experimental results shows PR-Embedding can help the models select better reply by catching the information among the pairs.

wtma@iflytek.com

https://github.com/wtma/PR-Embedding

Thank you !