ACL 2020

THE OHIO STATE UNIVERSITY

NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.

Rationalizing Medical Relation Prediction from Corpus-level Statistics

Zhen Wang

The Ohio State University

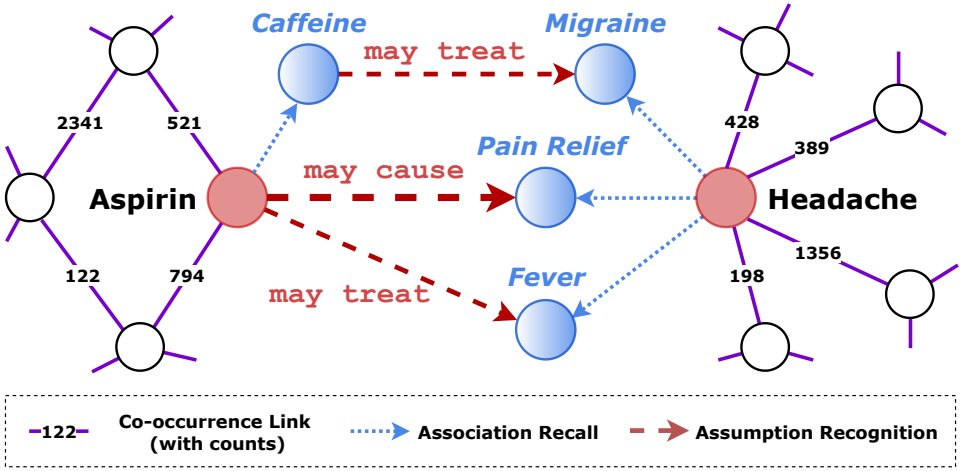
In collaboration with Jennifer Lee (NCH), Simon Lin (NCH), Huan Sun (OSU)

One-Minute Summary

What is this paper about?

Task: Predicting relations between two given terms from a text corpus

Goal: Make accurate prediction & Provide justifications for it (Rationalization)



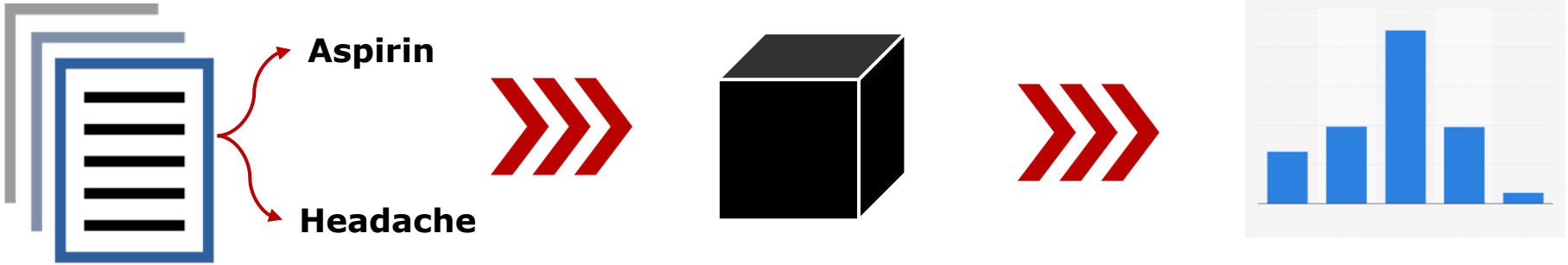
1. Draw inspirations from human memory **recall** and **recognition**

2. **Recall** global associations (blue) from corpus-level statistics and **recognize** meaningful relational assumptions (red) between them

3. Justify the prediction (*Aspirin may treat headache*) by highlighting important associations and assumptions

Black-Box Relation Classifier

Medical Records



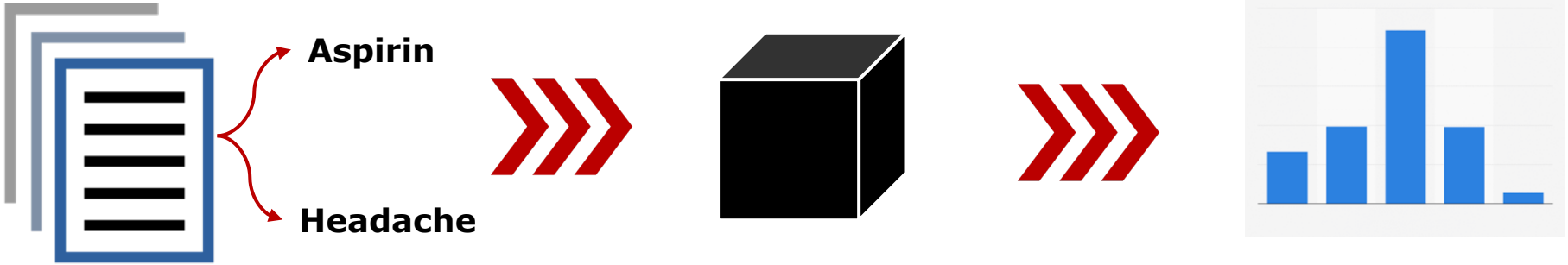
Input Pair

Black-Box Classifier

Class Distribution

Black-Box Relation Classifier

Medical Records



Input Pair

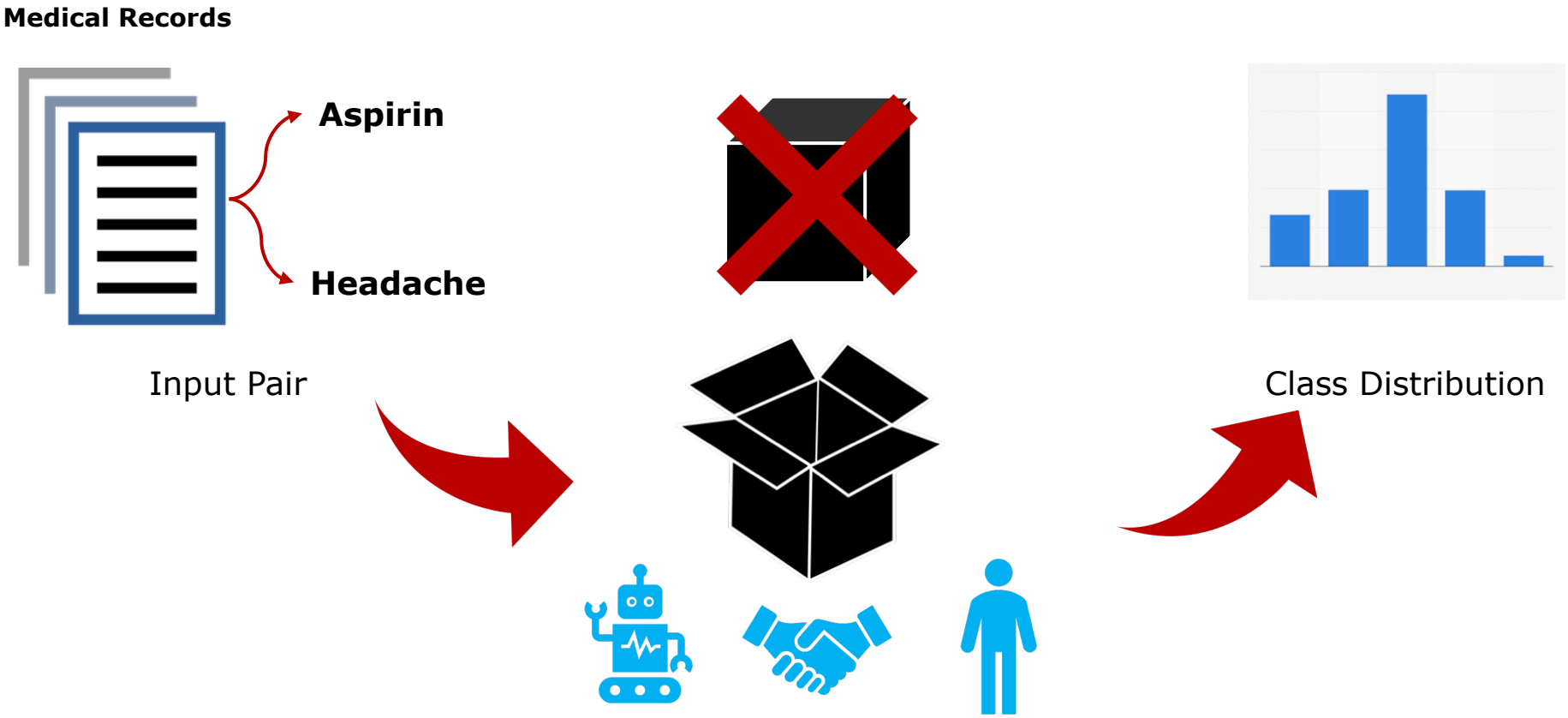
Black-Box Classifier

Class Distribution

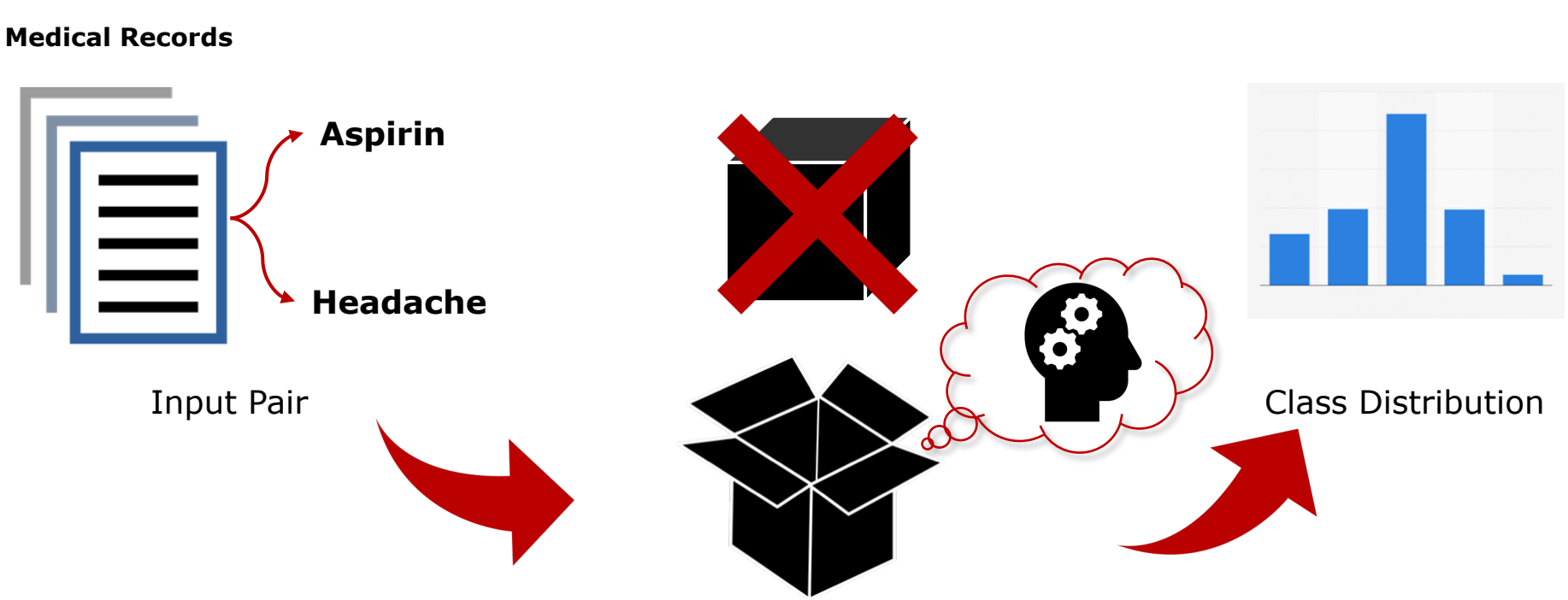
High Risk!

Medicine **Finance** **Judiciary**

Can We Open the Black-Box?



Can We Open the Black-Box?



To solve the problem, we draw inspirations from human memory theories – Recall and Recognition!

Memory Recall and Recognition

Recall

Retrieve association information from long-term memory

Recognition

Identify previously learned information

Memory Recall and Recognition

Recall

Retrieve association information from long-term memory

Recall my friend's name

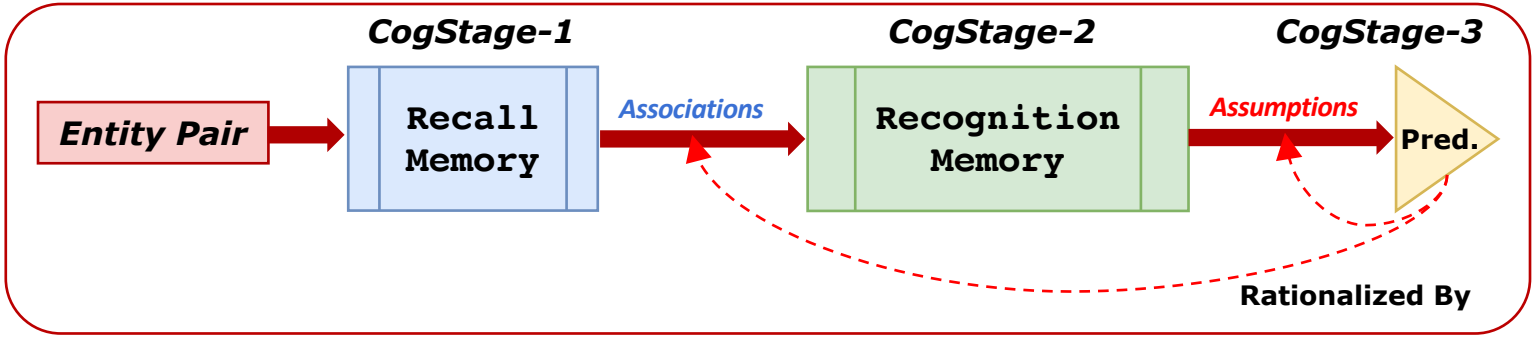
Recognition

Identify previously learned information

Recognize my friend's face

Example from http://webdesign-review.blogspot.com/2016/04/recognition-vs-recall-in-mobile-web.html

Rationalizing Medical Relation Prediction



Rationalizing Medical Relation Prediction

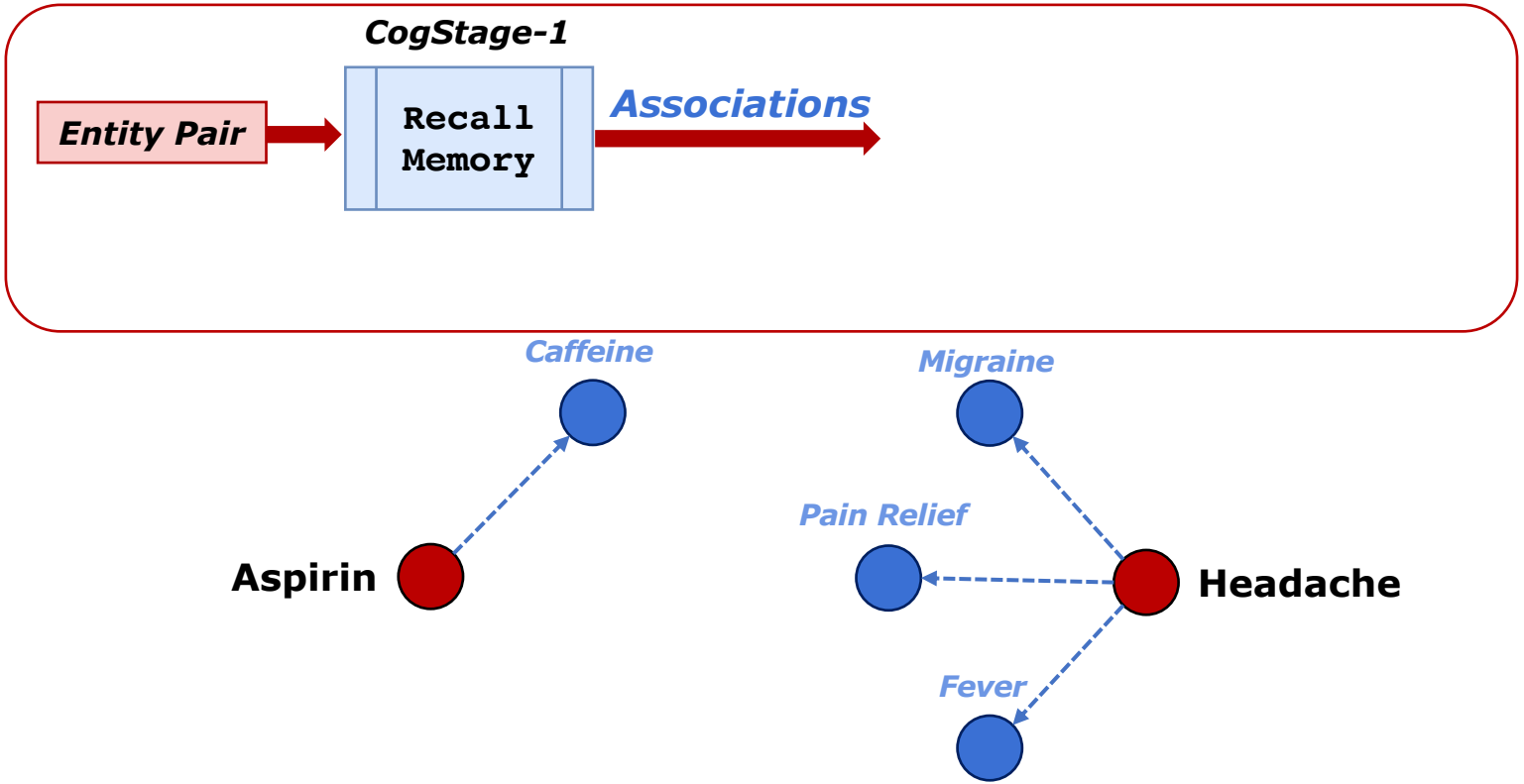
Entity Pair

Rationalizing Medical Relation Prediction

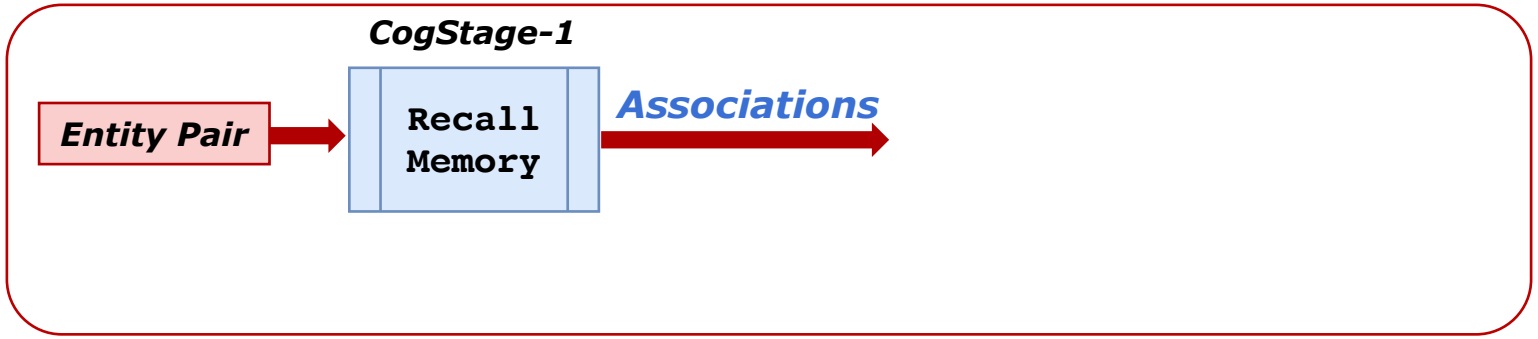
Aspirin

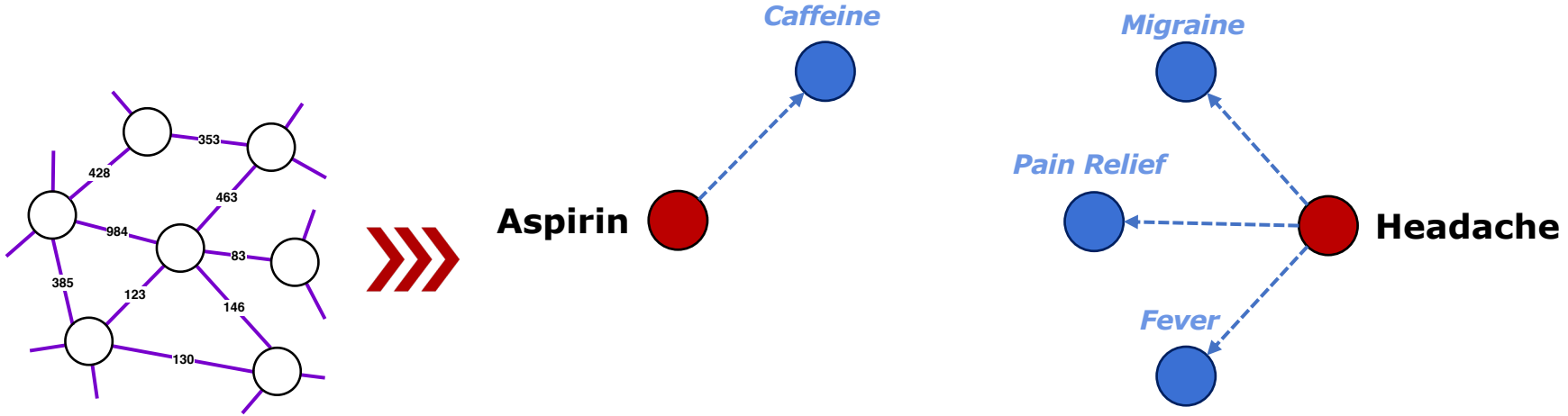
Headache

Rationalizing Medical Relation Prediction



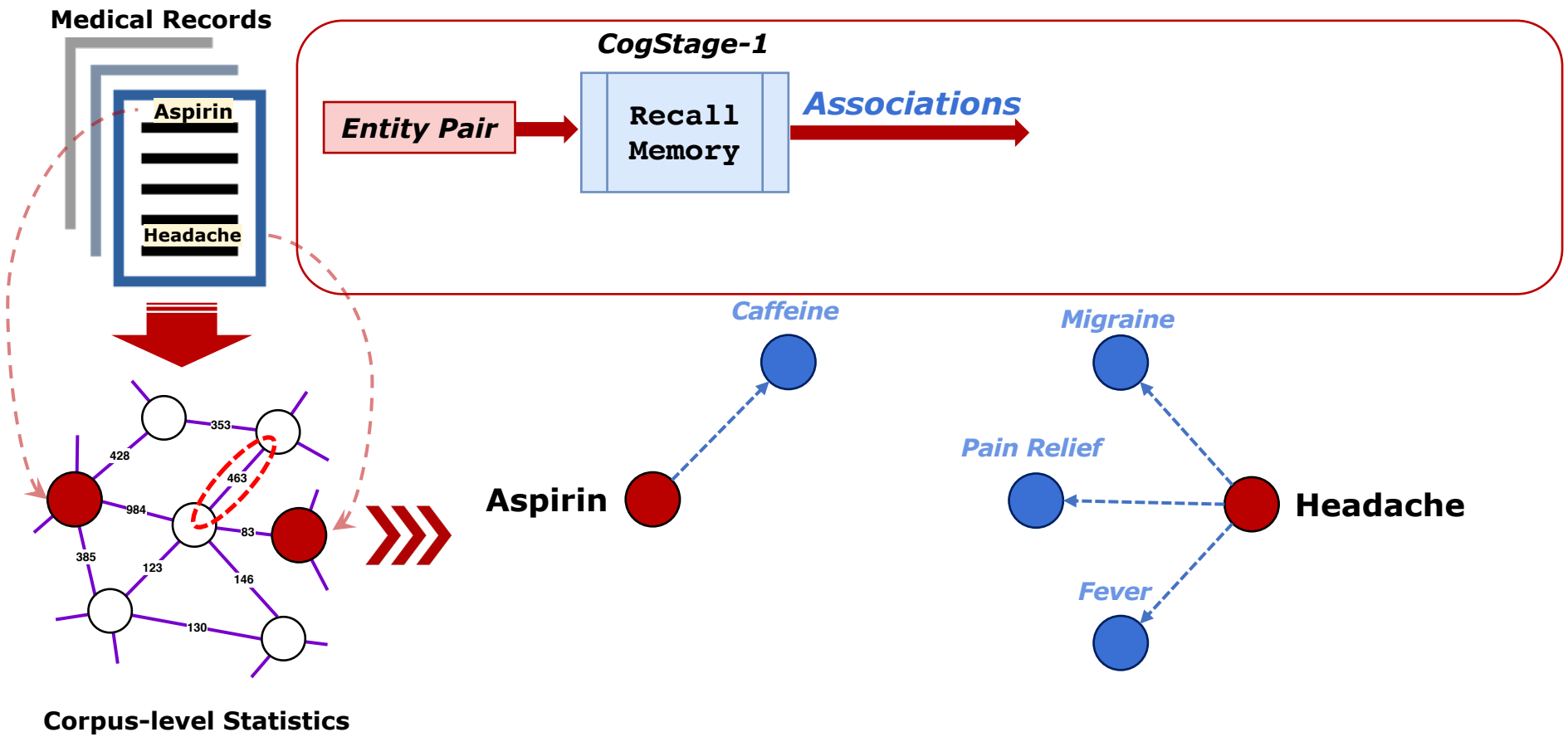
Rationalizing Medical Relation Prediction





Corpus-level Statistics

Rationalizing Medical Relation Prediction



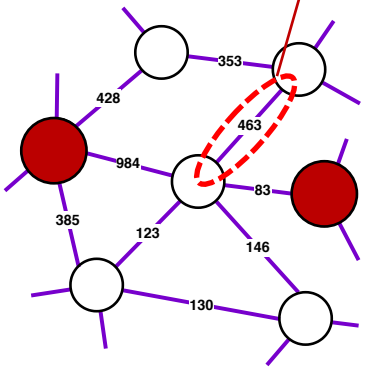
Rationalizing Medical Relation Prediction

Medical Records

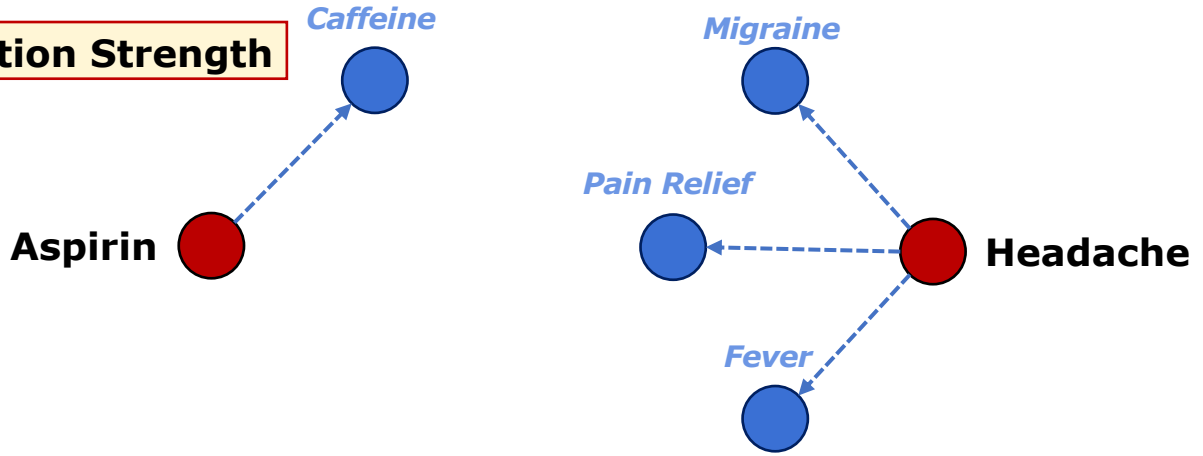
CogStage-1

Entity Pair → Recall Memory → *Associations*

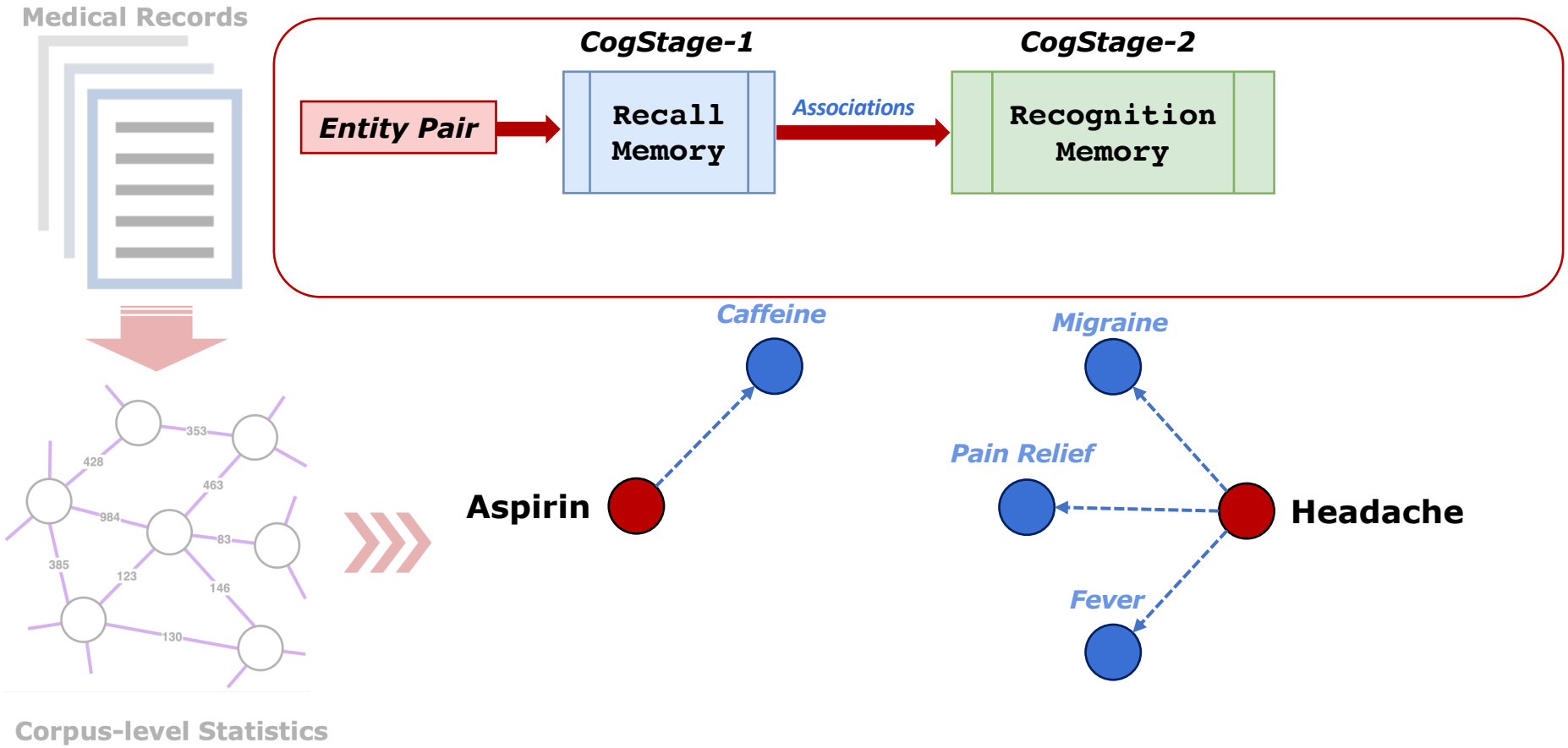
Association Strength



Corpus-level Statistics

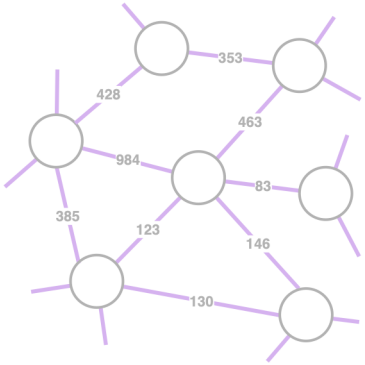


Rationalizing Medical Relation Prediction

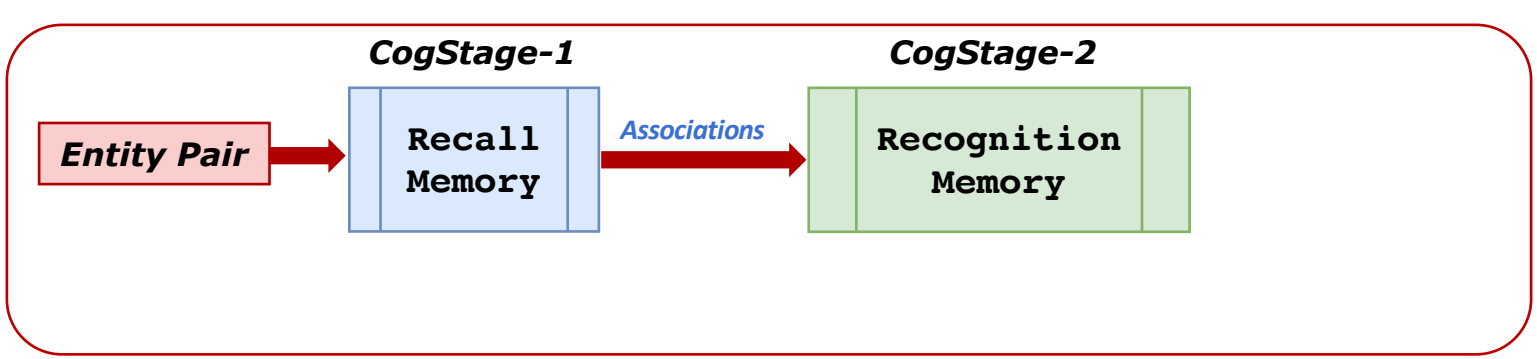


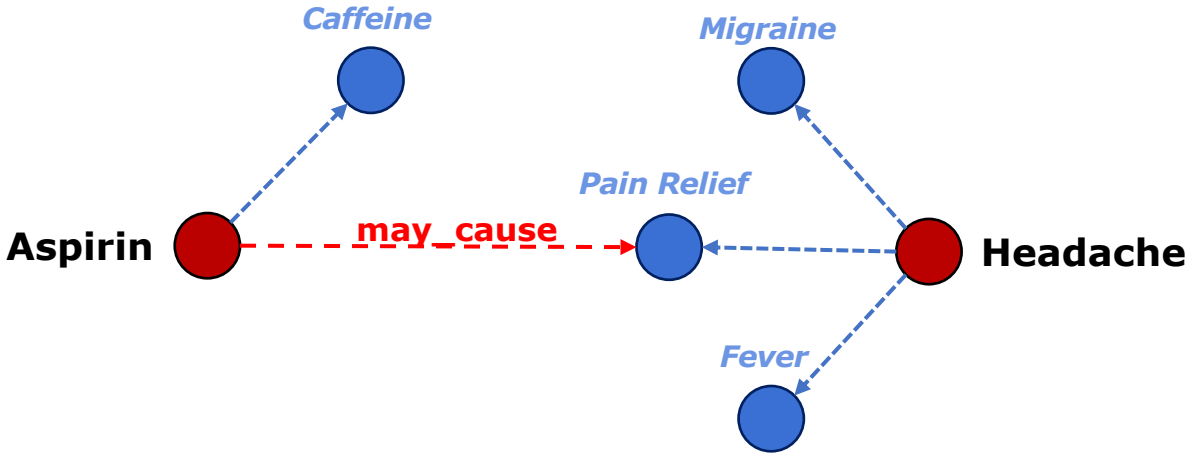
Rationalizing Medical Relation Prediction

Medical Records



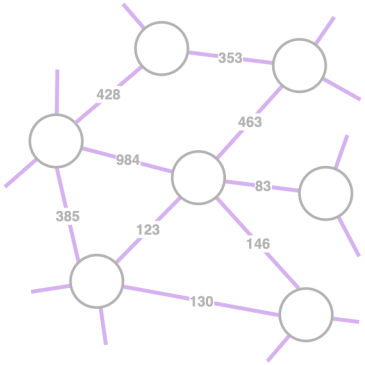
Corpus-level Statistics



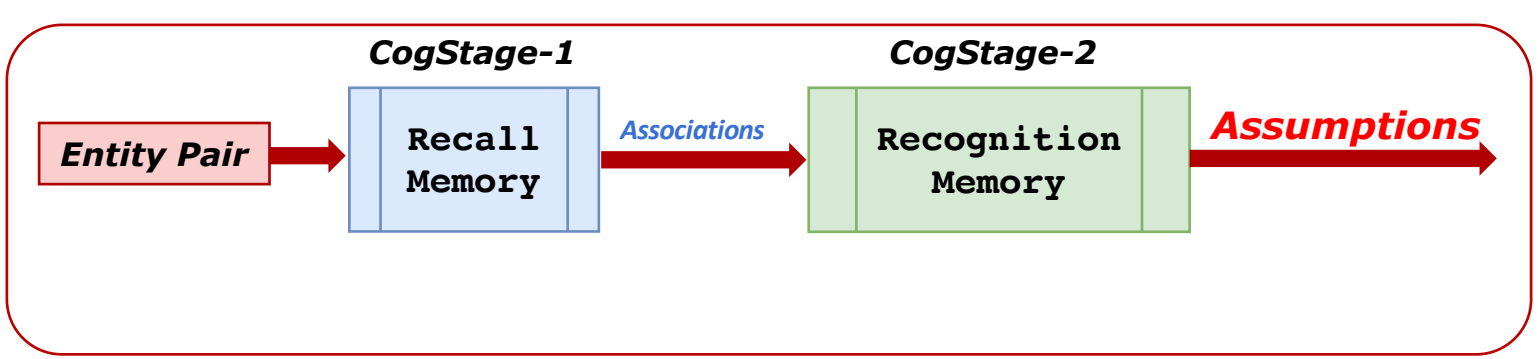


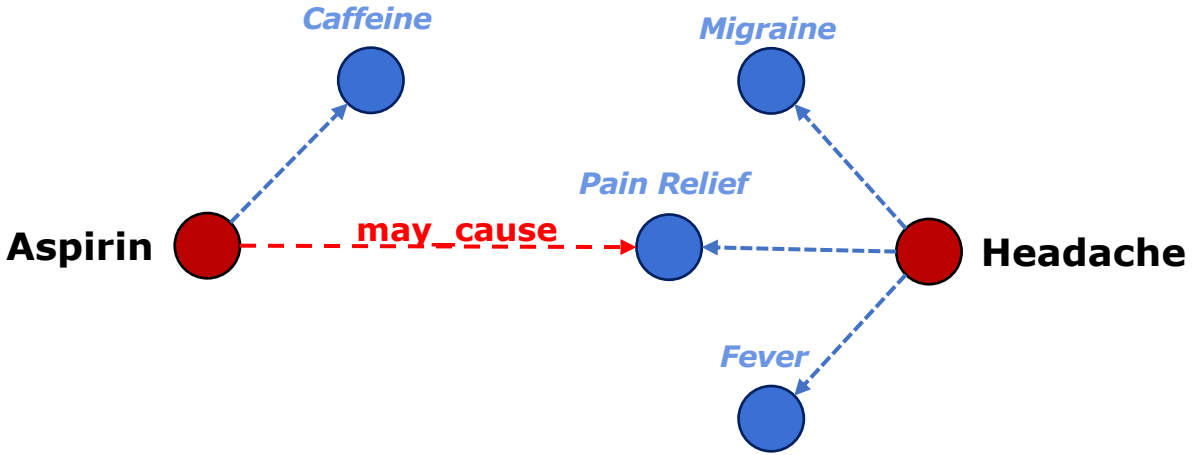
Rationalizing Medical Relation Prediction

Medical Records

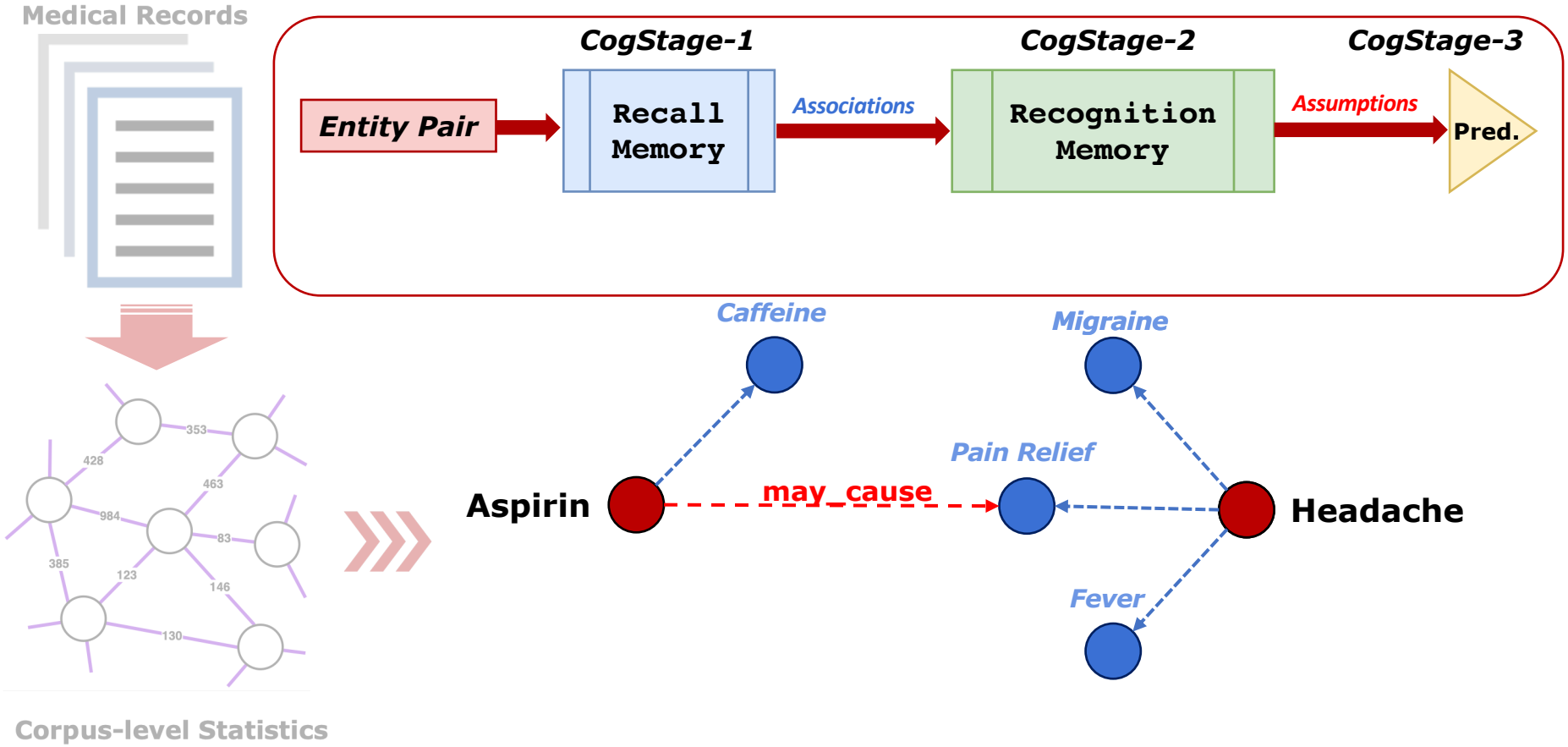


Corpus-level Statistics



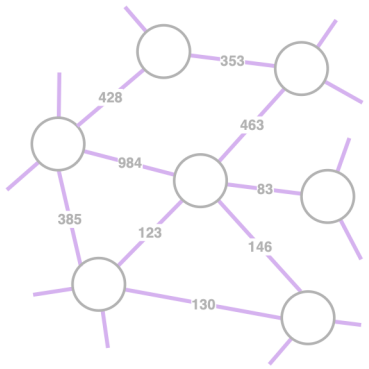


Rationalizing Medical Relation Prediction

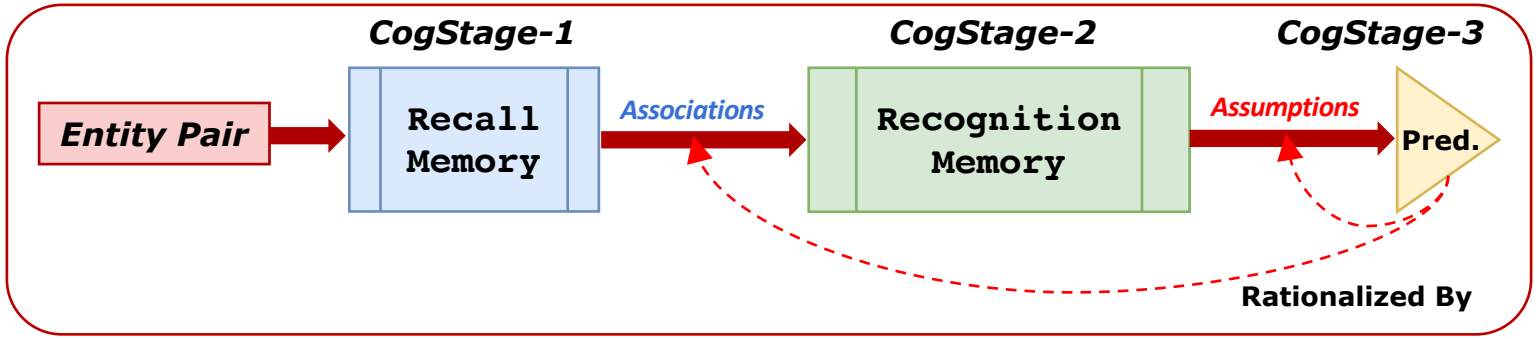


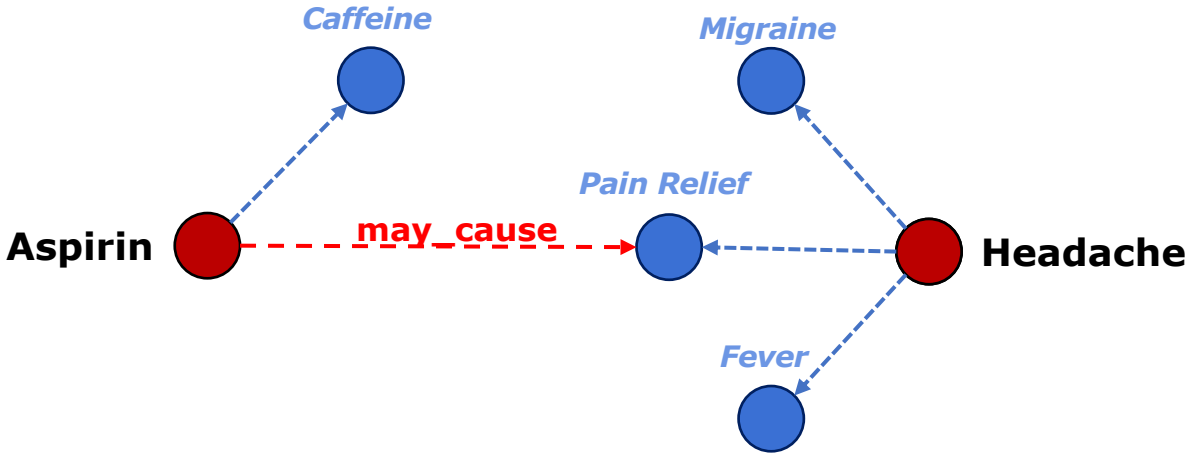
Rationalizing Medical Relation Prediction

Medical Records

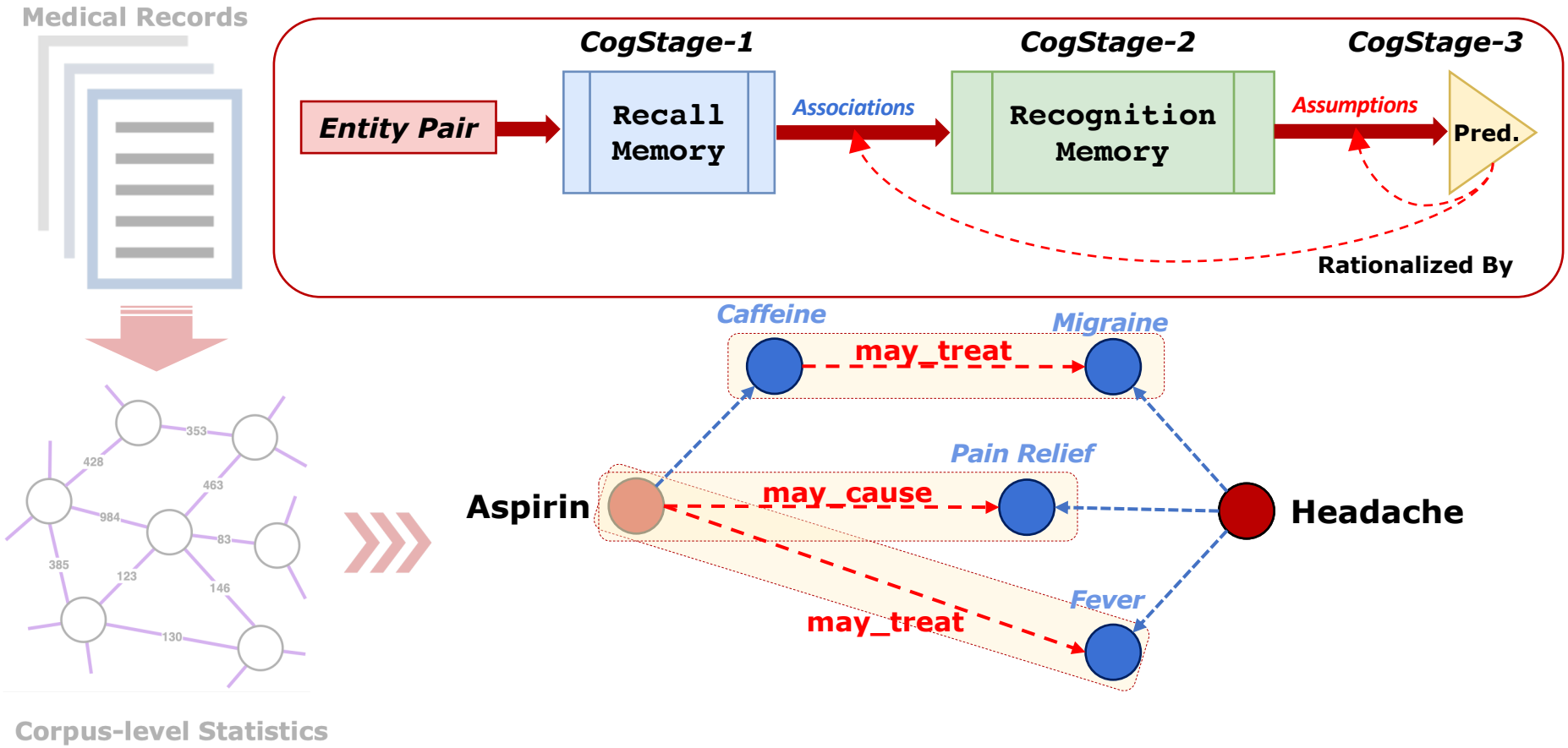


Corpus-level Statistics

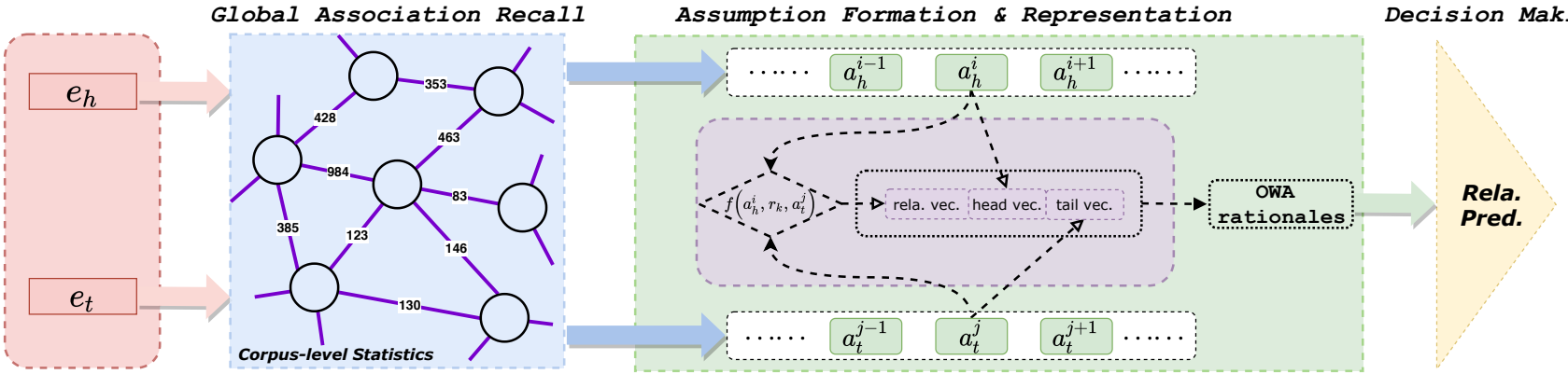




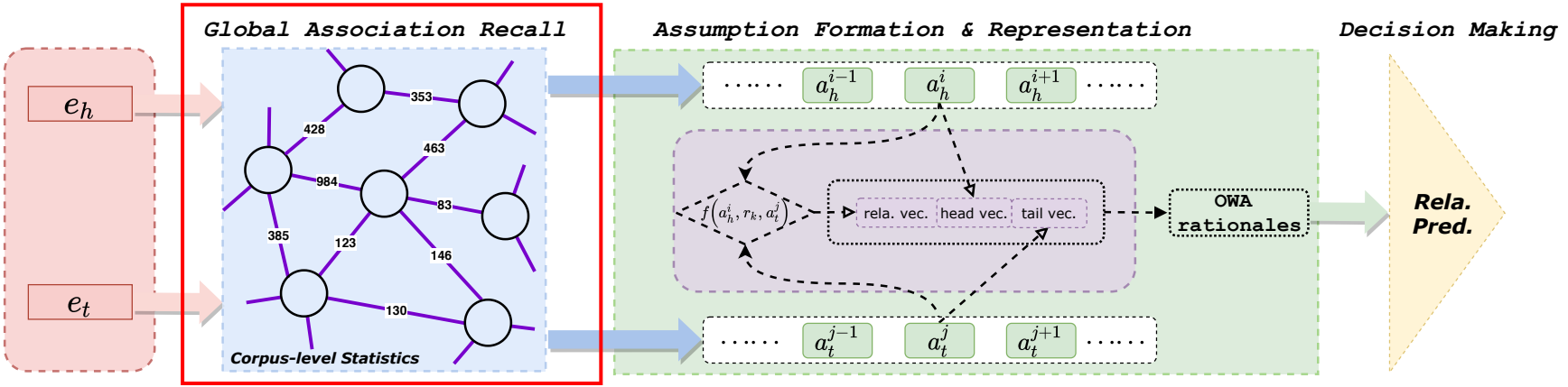
Rationalizing Medical Relation Prediction



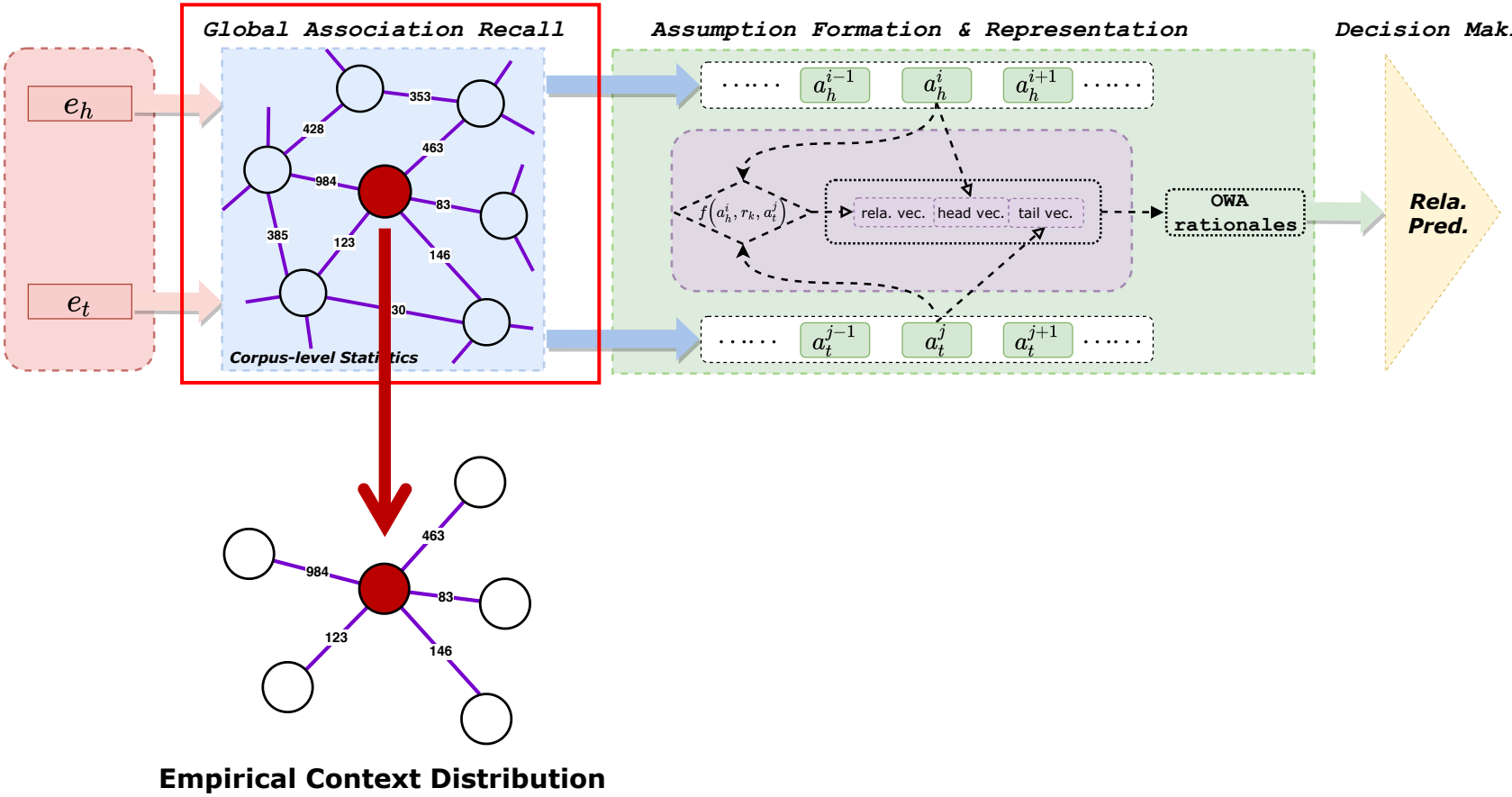
Framework Overview



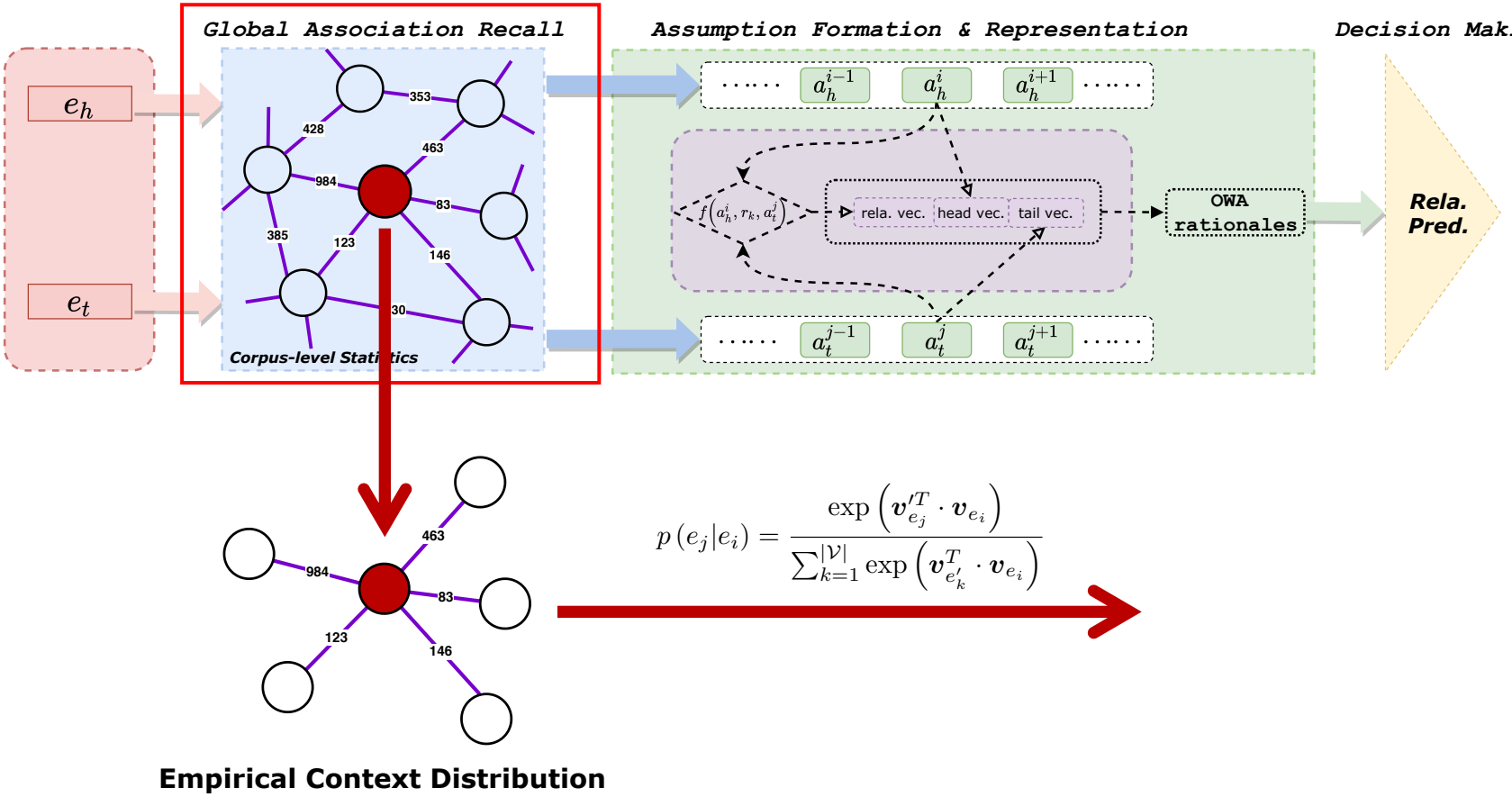
CogStage-1: Global Association Recall



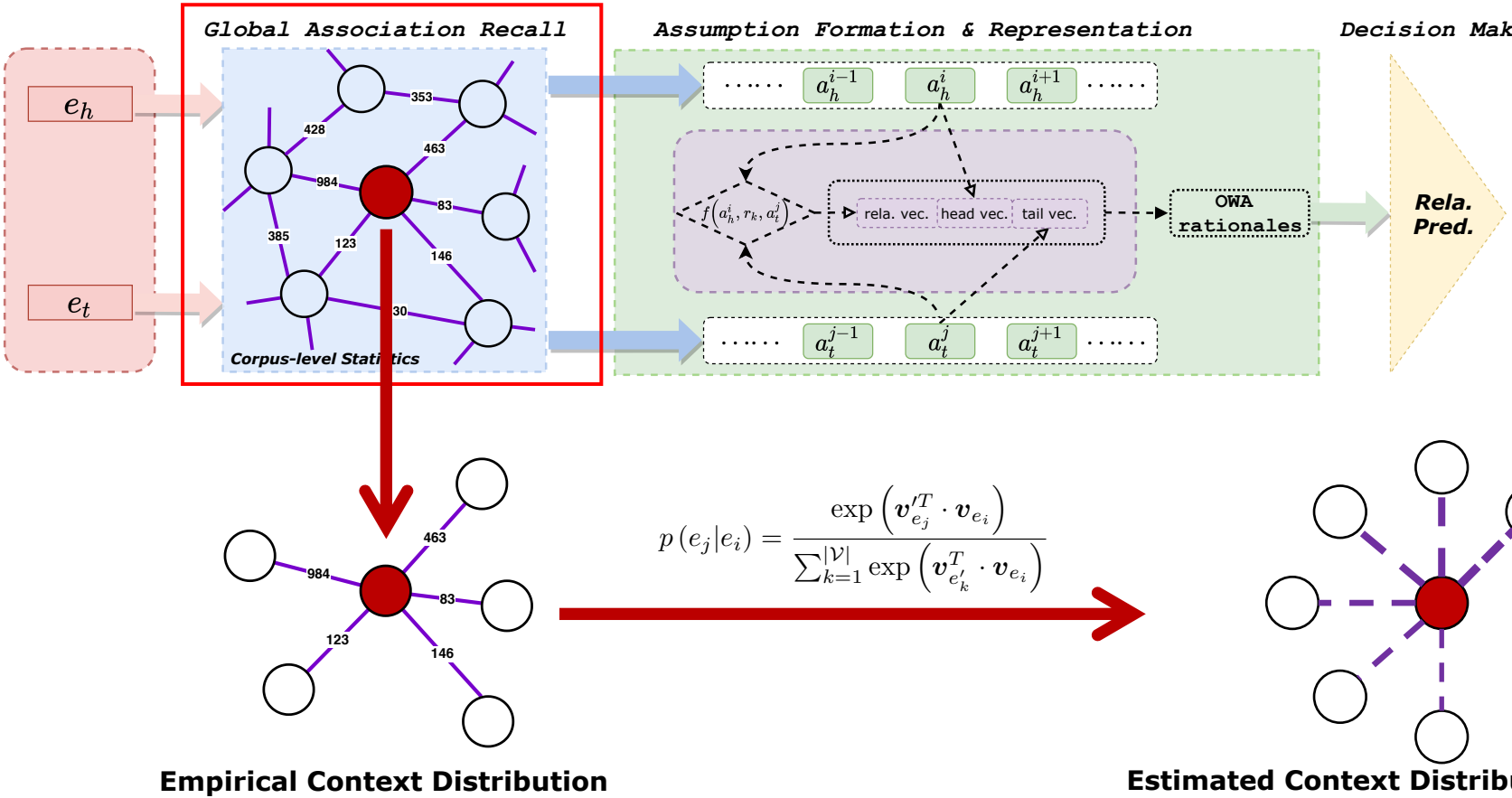
CogStage-1: Global Association Recall



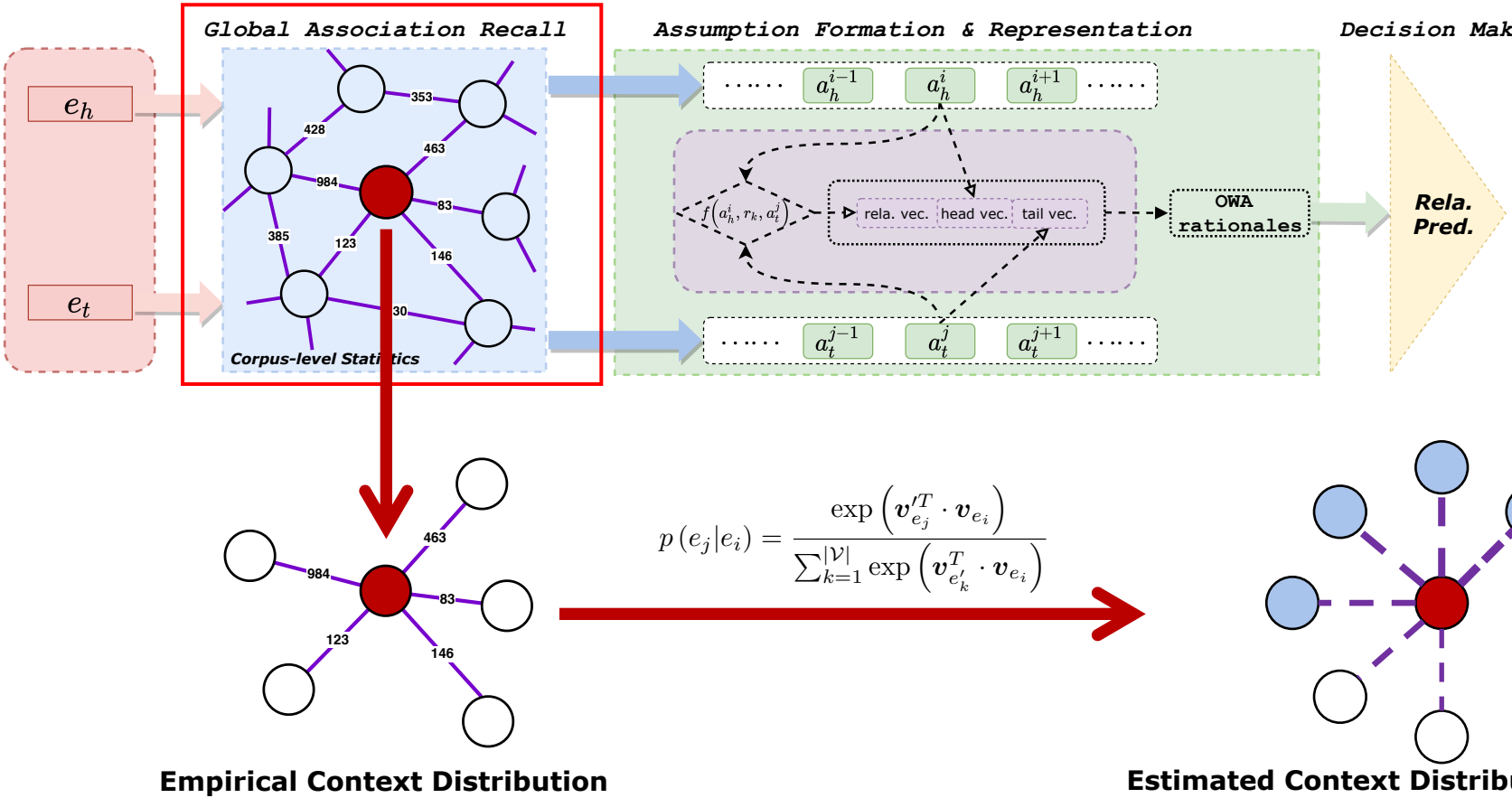
CogStage-1: Global Association Recall



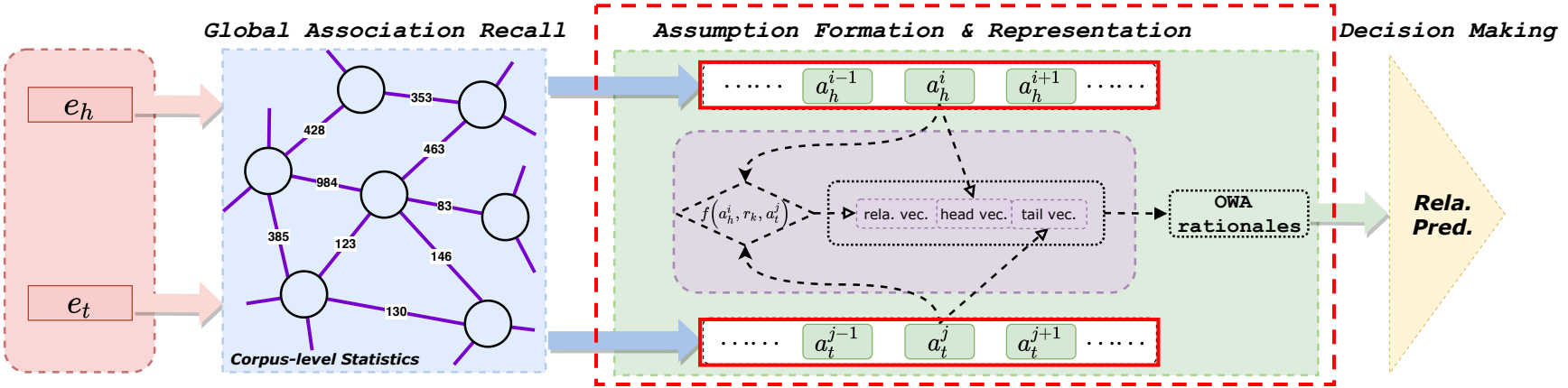
CogStage-1: Global Association Recall



CogStage-1: Global Association Recall

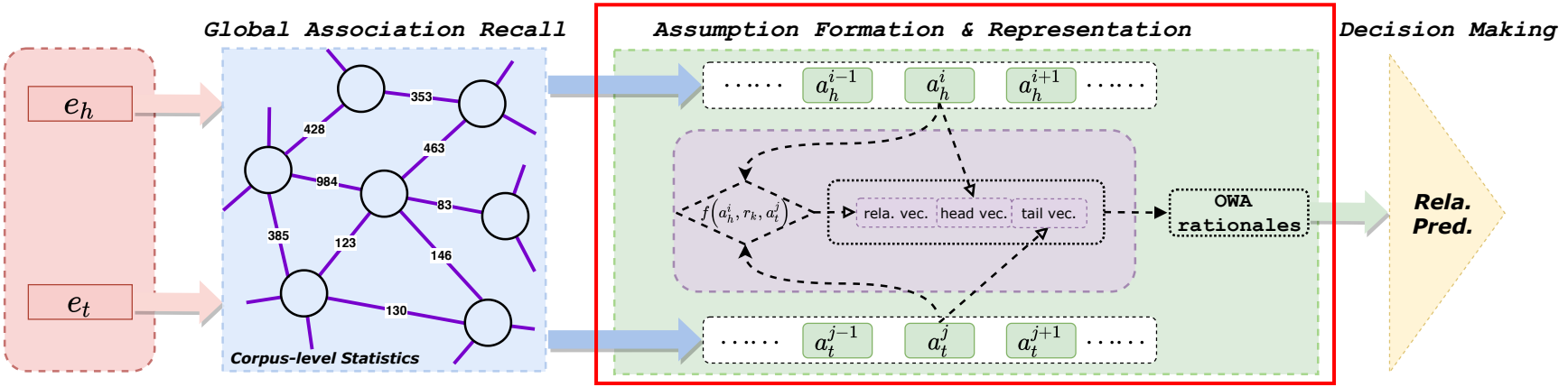


CogStage-2: Assumption Formation and Representation



Assumption: Recognize whether context a_h^i and a_t^j hold a relationship r_k?

CogStage-2: Assumption Formation and Representation

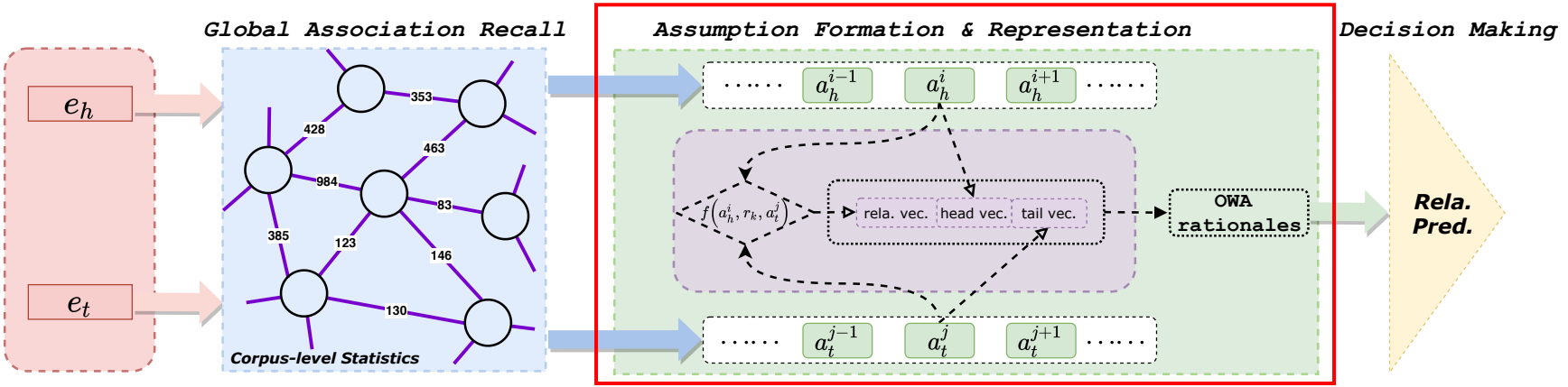


Assumption: Recognize whether context a_h^i and a_t^j hold a relationship r_k?

Closed World Assumption (CWA):

Open World Assumption (OWA):

CogStage-2: Assumption Formation and Representation



Assumption: Recognize whether context a_h^i and a_t^j hold a relationship r_k?

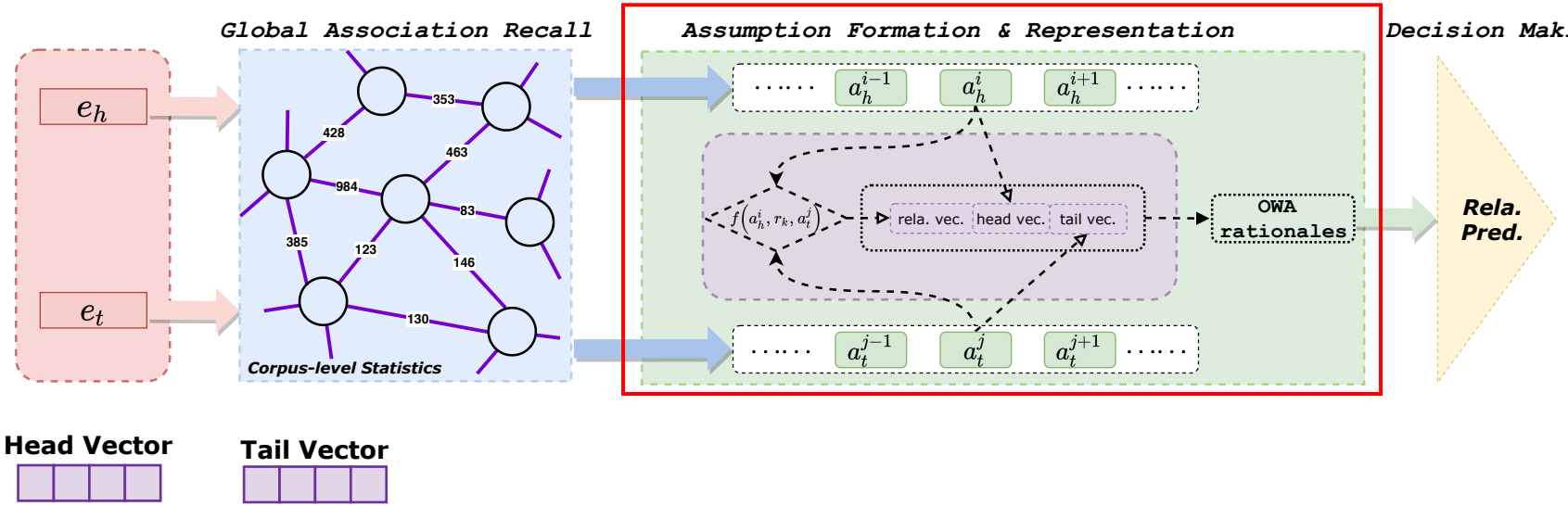
Closed World Assumption (CWA):

Only consider the facts that exist in KBs

Open World Assumption (OWA):

Consider all possible combinations and select the best ones.

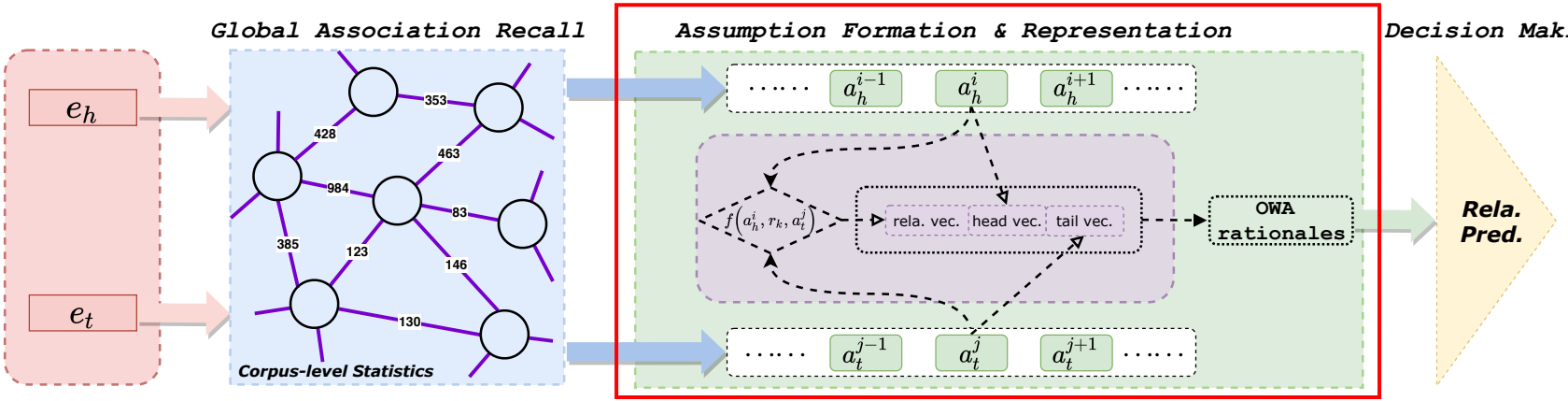
CogStage-2: Assumption Formation and Representation

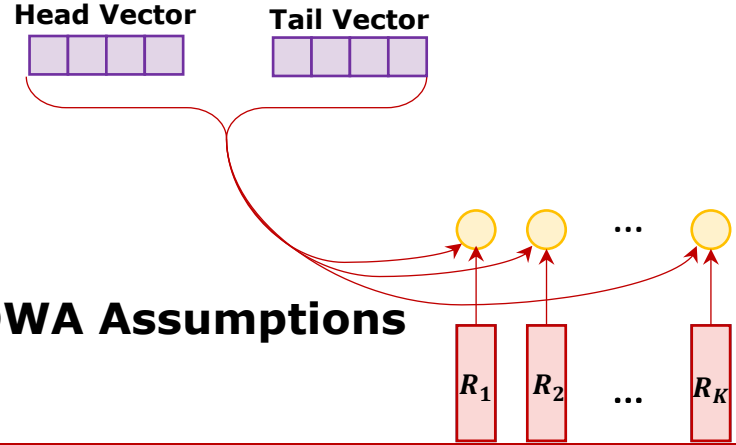


OWA Assumptions

R_1 R_2 ... R_K **Relation Embeddings**

CogStage-2: Assumption Formation and Representation



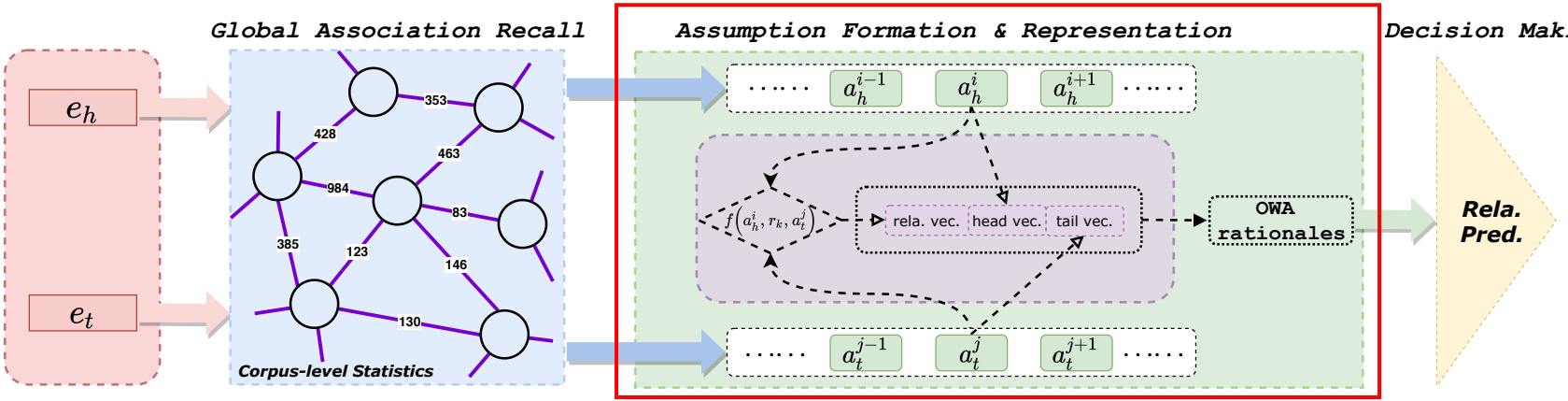


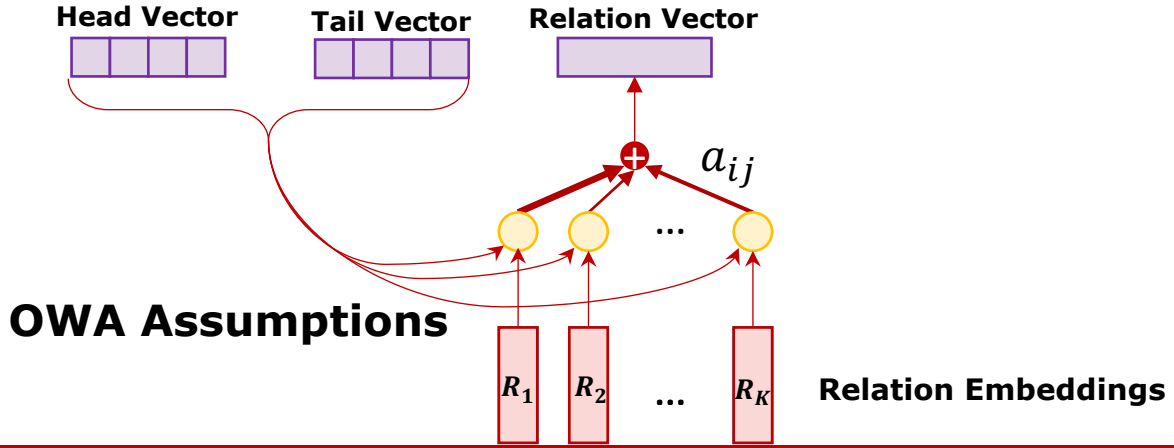
OWA Assumptions

$$s_k^{ij} = f\left(a_h^i, r_k, a_t^j\right) = -\left\|\boldsymbol{v}_{a_h^i} + \boldsymbol{\xi}_k - \boldsymbol{v}_{a_t^j}\right\|_1$$

Relation Embeddings

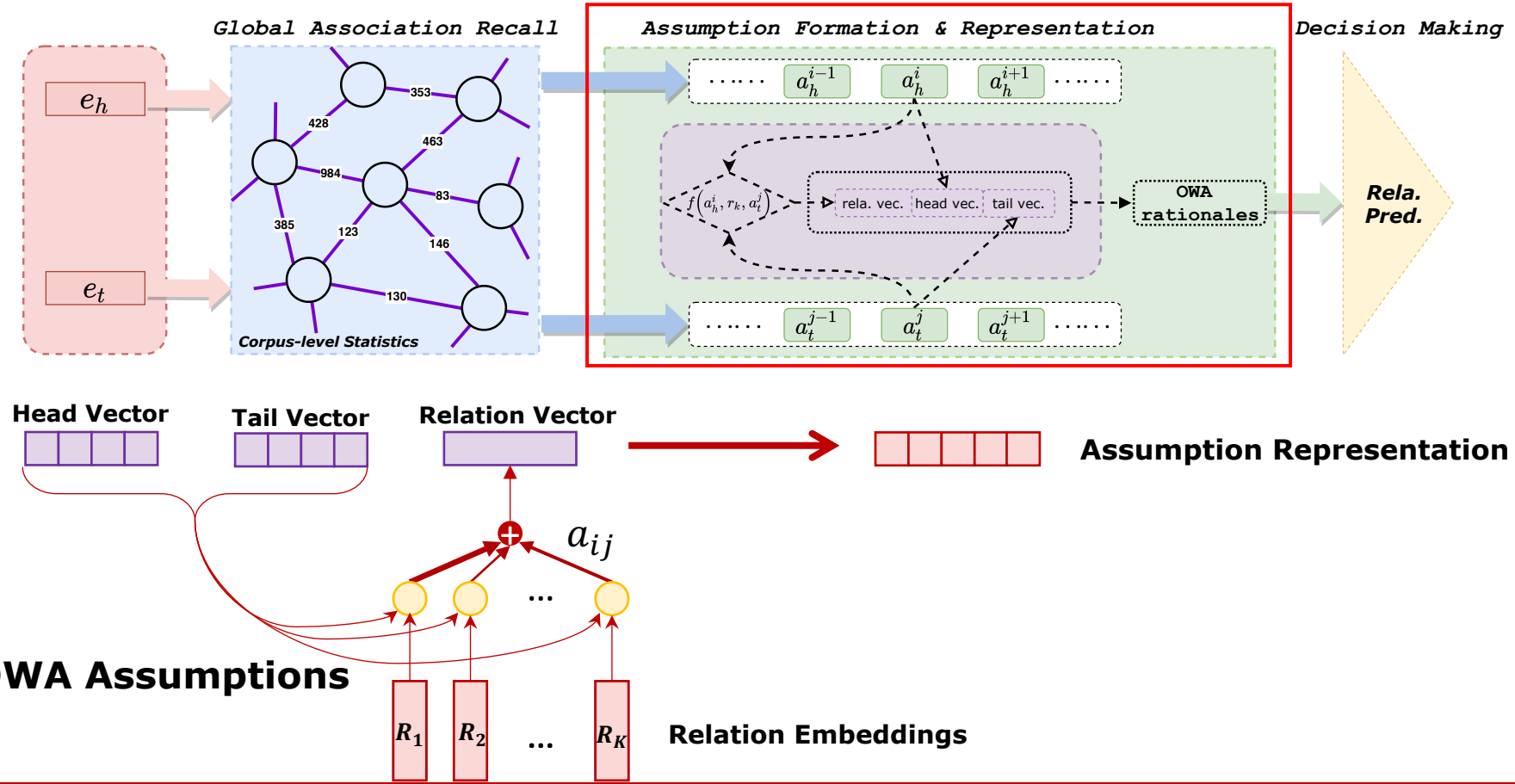
CogStage-2: Assumption Formation and Representation



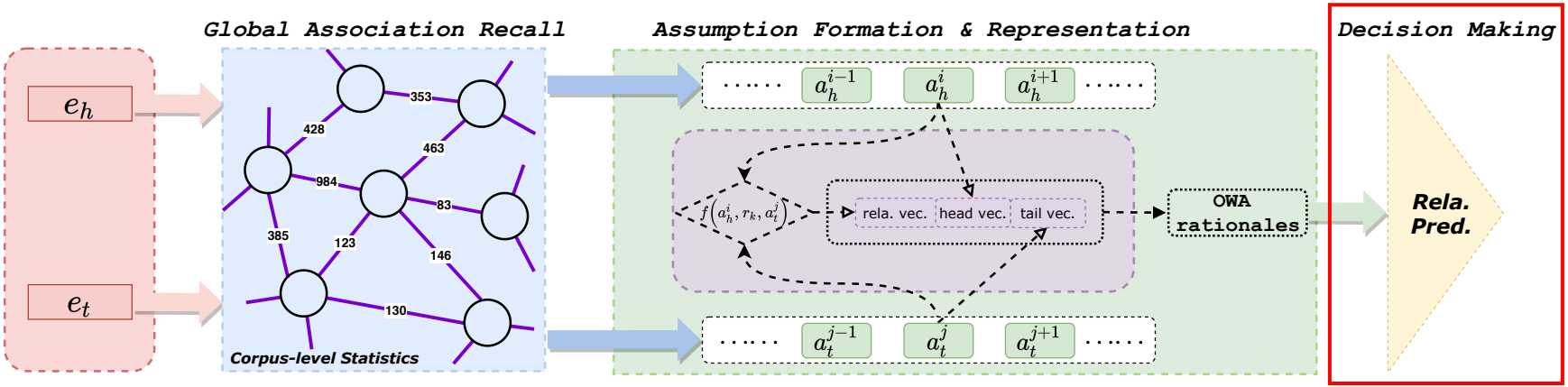


OWA Assumptions

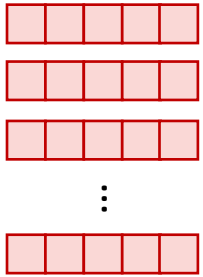
CogStage-2: Assumption Formation and Representation



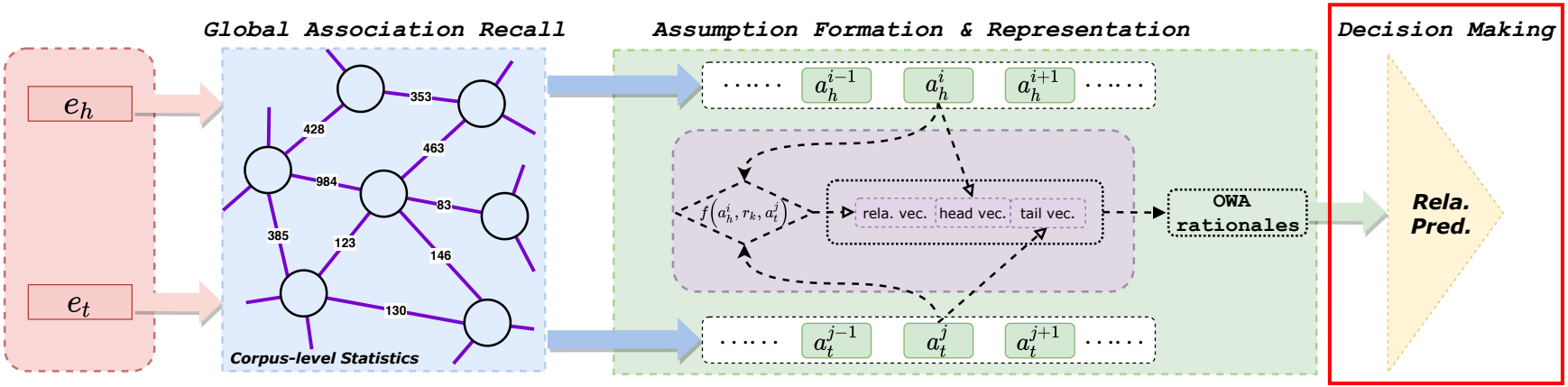
CogStage-3:Prediction Decision Making



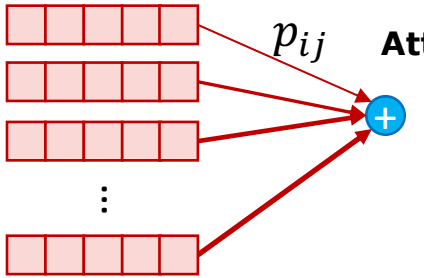
Representations for all assumptions



CogStage-3:Prediction Decision Making

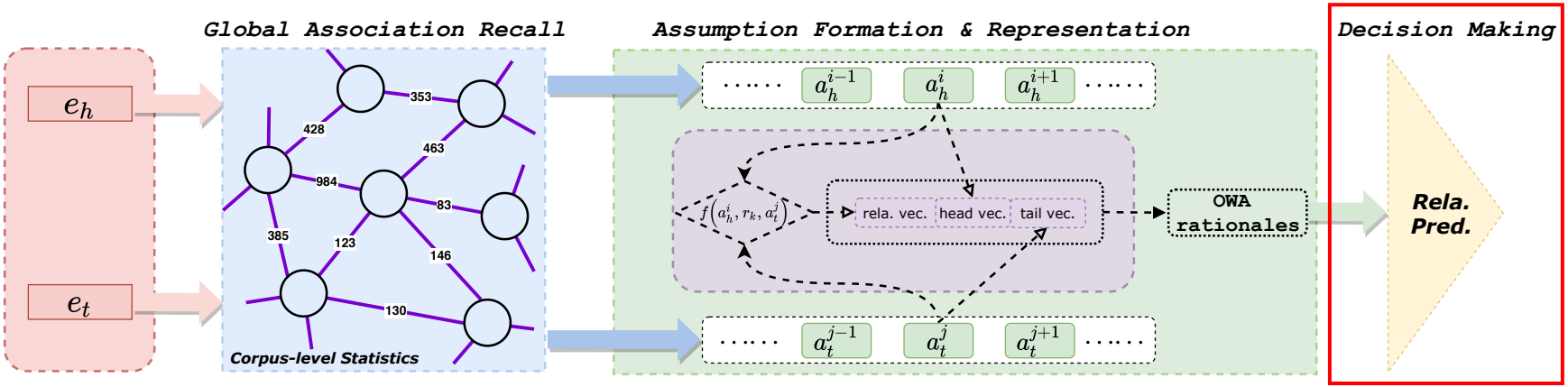


Representations for all assumptions

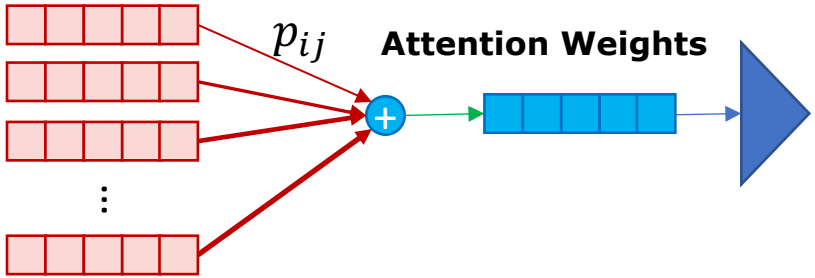


Attention Weights

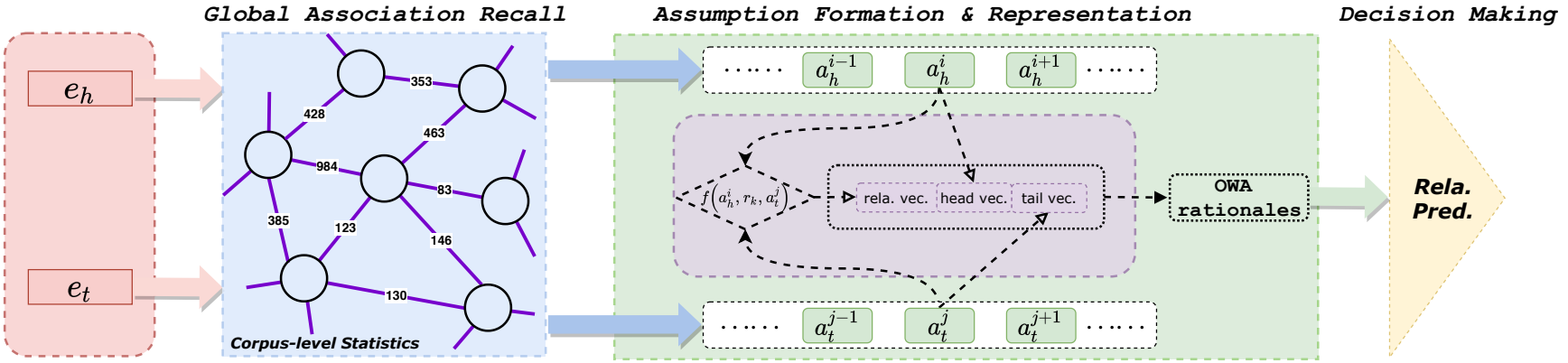
CogStage-3:Prediction Decision Making



Representations for all assumptions



Framework Training



$$\mathcal{L}_n = -\sum_{(e_i,e_j)\in\mathcal{V}} \hat{p}\left(e_j|e_i\right)\log\left(p\left(e_j|e_i\right)\right)$$

$$\mathcal{L}_r = -\sum_{(h,r,t)\in\mathcal{P}} \log p(h|t,r) - \sum_{(h,r,t)\in\mathcal{P}} \log p(t|h,r)$$

$$\mathcal{L}_p = -\sum_{i=1}^{M} \left(y_i \cdot \log\left(p\left(r|e_h^i, e_t^i\right)\right) + (1-y_i)\cdot\log\left(1-p\left(r|e_h^i, e_t^i\right)\right)\right)$$

Experimental Setup

- Datasets
  - Medical Term-Term Co-occurrence Graph
  - 20 Million Clinical Notes from Stanford Hospital and Clinics since 1995
  - 52,804 Nodes, 16,197,319 Edges

Experimental Setup

- Datasets
  - Medical Term-Term Co-occurrence Graph
  - 20 Million Clinical Notes from Stanford Hospital and Clinics since 1995
  - 52,804 Nodes, 16,197,319 Edges
  - Five Popular Medical Relations

Med Relations	Train	Dev	Test
Symptom of	14,326	3,001	3,087
May treat	12,924	2,664	2,735
Contraindicates	10,593	2,237	2,197
May prevent	2,113	440	460
Causes	1,389	305	354
Total	41.3k	8.6k	8.8k

Table 1: Dataset Statistics.

Experimental Setup

- Datasets
  - Medical Term-Term Co-occurrence Graph
  - 20 Million Clinical Notes from Stanford Hospital and Clinics since 1995
  - 52,804 Nodes, 16,197,319 Edges
  - Five Popular Medical Relations
- Baseline Methods
  - Entity Encoder + Relation Scoring

Word2vec	**DistMult**
DeepWalk	**RESCAL**
LINE	**NTN**
REPEL-D	

Med Relations	Train	Dev	Test
Symptom of	14,326	3,001	3,087
May treat	12,924	2,664	2,735
Contraindicates	10,593	2,237	2,197
May prevent	2,113	440	460
Causes	1,389	305	354
Total	41.3k	8.6k	8.8k

Table 1: Dataset Statistics.

Experiment Results: Predictive Performance

Methods	MAY_TREAT	CONTRAIN.	SYMPTOM_OF	MAY_PREVENT	CAUSES	Avg.
Word2vec + DistMult	0.767 (±0.008)	0.777 (±0.013)	0.815 (±0.005)	0.649 (±0.018)	0.671 (±0.015)	0.736
Word2vec + RESCAL	0.743 (±0.010)	0.767 (±0.003)	0.808 (±0.009)	0.658 (±0.023)	0.659 (±0.039)	0.727
Word2vec + NTN	0.693 (±0.013)	0.758 (±0.005)	0.808 (±0.004)	0.605 (±0.022)	0.631 (±0.017)	0.699
DeepWalk + DistMult	0.740 (±0.003)	0.776 (±0.004)	0.805 (±0.003)	0.608 (±0.014)	0.650 (±0.018)	0.716
DeepWalk + RESCAL	0.671 (±0.010)	0.778 (±0.003)	0.800 (±0.003)	0.600 (±0.023)	**0.708 (±0.011)**	0.711
DeepWalk + NTN	0.696 (±0.006)	0.778 (±0.005)	0.787 (±0.005)	0.614 (±0.016)	0.674 (±0.024)	0.710
LINE + DistMult	0.767 (±0.003)	0.783 (±0.002)	0.795 (±0.003)	0.621 (±0.015)	0.641 (±0.024)	0.721
LINE + RESCAL	0.725 (±0.003)	0.771 (±0.002)	0.801 (±0.001)	0.613 (±0.013)	0.694 (±0.015)	0.721
LINE + NTN	0.733 (±0.002)	0.773 (±0.003)	0.800 (±0.001)	0.601 (±0.015)	0.706 (±0.013)	0.723
REPEL-D + DistMult	0.784 (±0.002)	0.797 (±0.002)	0.809 (±0.003)	0.681 (±0.010)	0.694 (±0.022)	0.751
REPEL-D + RESCAL	0.726 (±0.003)	0.780 (±0.002)	0.776 (±0.002)	**0.685 (±0.010)**	**0.708 (±0.003)**	0.737
REPEL-D + NTN	0.736 (±0.004)	0.780 (±0.002)	0.773 (±0.001)	0.667 (±0.015)	0.694 (±0.024)	0.731
Ours (w/ CWA)	0.709 (±0.005)	0.751 (±0.009)	0.744 (±0.007)	0.667 (±0.008)	0.661 (±0.032)	0.706
Ours	**0.805 (±0.017)**	**0.811 (±0.006)**	**0.816 (±0.004)**	0.676 (±0.020)	0.684 (±0.017)	**0.758**

Table 2: Comparison of model predictive performance. We run all methods for five times and report the averaged F1 scores with standard deviations.

Experiment Results: Predictive Performance

Methods	MAY_TREAT	CONTRAIN.	SYMPTOM_OF	MAY_PREVENT	CAUSES	Avg.
Word2vec + DistMult	0.767 (±0.008)	0.777 (±0.013)	0.815 (±0.005)	0.649 (±0.018)	0.671 (±0.015)	0.736
Word2vec + RESCAL	0.743 (±0.010)	0.767 (±0.003)	0.808 (±0.009)	0.658 (±0.023)	0.659 (±0.039)	0.727
Word2vec + NTN	0.693 (±0.013)	0.758 (±0.005)	0.808 (±0.004)	0.605 (±0.022)	0.631 (±0.017)	0.699
DeepWalk + DistMult	0.740 (±0.003)	0.776 (±0.004)	0.805 (±0.003)	0.608 (±0.014)	0.650 (±0.018)	0.716
DeepWalk + RESCAL	0.671 (±0.010)	0.7	0.8	0.6	**0.708 (±0.011)**	0.711
DeepWalk + NTN	0.696 (±0.006)	0.				0.710
LINE + DistMult	0.767 (±0.003)	0.				0.721
LINE + RESCAL	0.725 (±0.003)	0.771 (±0.002)	0.801 (±0.001)	0.613 (±0.013)	0.694 (±0.015)	0.721
LINE + NTN	0.733 (±0.002)	0.773 (±0.003)	0.800 (±0.001)	0.601 (±0.015)	0.706 (±0.013)	0.723
REPEL-D + DistMult	0.784 (±0.002)	0.797 (±0.002)	0.809 (±0.003)	0.681 (±0.010)	0.694 (±0.022)	0.751
REPEL-D + RESCAL	0.726 (±0.003)	0.780 (±0.002)	0.776 (±0.002)	**0.685 (±0.010)**	**0.708 (±0.003)**	0.737
REPEL-D + NTN	0.736 (±0.004)	0.780 (±0.002)	0.773 (±0.001)	0.667 (±0.015)	0.694 (±0.024)	0.731
Ours (w/ CWA)	0.709 (±0.005)	0.751 (±0.009)	0.744 (±0.007)	0.667 (±0.008)	0.661 (±0.032)	0.706
Ours	**0.805 (±0.017)**	**0.811 (±0.006)**	**0.816 (±0.004)**	0.676 (±0.020)	0.684 (±0.017)	**0.758**

Competitive predictive performance compared with a comprehensive list of baselines

Table 2: Comparison of model predictive performance. We run all methods for five times and report the averaged F1 scores with standard deviations.

Experiment Results: Predictive Performance

Methods	MAY_TREAT	CONTRAIN.	SYMPTOM_OF	MAY_PREVENT	CAUSES	Avg.
Word2vec + DistMult	0.767 (±0.008)	0.777 (±0.013)	0.815 (±0.005)	0.649 (±0.018)	0.671 (±0.015)	0.736
Word2vec + RESCAL	0.743 (±0.010)	0.767 (±0.003)	0.808 (±0.009)	0.658 (±0.023)	0.659 (±0.039)	0.727
Word2vec + NTN	0.693 (±0.013)	0.758 (±0.005)	0.808 (±0.004)	0.605 (±0.022)	0.631 (±0.017)	0.699
DeepWalk + DistMult	0.740 (±0.003)	0.776 (±0.004)	0.805 (±0.003)	0.608 (±0.014)	0.650 (±0.018)	0.716
DeepWalk + RESCAL	0.671 (±0.010)	0.7	0.8	0.6	**0.708 (±0.011)**	0.711
DeepWalk + NTN	0.696 (±0.006)	0.				0.710
LINE + DistMult	0.767 (±0.003)	0.				0.721
LINE + RESCAL	0.725 (±0.003)	0.771 (±0.002)	0.801 (±0.001)	0.613 (±0.013)	0.694 (±0.015)	0.721
LINE + NTN	0.733 (±0.002)	0.773 (±0.003)	0.800 (±0.001)	0.601 (±0.015)	0.706 (±0.013)	0.723
REPEL-D + DistMult	0.784 (±0.002)	0.797 (±0.002)	0.809 (±0.003)	0.681 (±0.010)	0.694 (±0.022)	0.751
REPEL-D + RESCAL	0.726 (±0.003)	0.780 (±0.002)	0.776 (±0.002)	**0.685 (±0.010)**	**0.708 (±0.003)**	0.737
REPEL-D + NTN	0.736 (±0.004)	0.780 (±0.002)	0.773 (±0.001)	0.667 (±0.015)	0.694 (±0.024)	0.731
Ours (w/ CWA)	0.709 (±0.005)	0.751 (±0.009)	0.744 (±0.007)	0.667 (±0.008)	0.661 (±0.032)	0.706
Ours	**0.805 (±0.017)**	**0.811 (±0.006)**	**0.816 (±0.004)**	0.676 (±0.020)	0.684 (±0.017)	**0.758**

Best predictive performance when the training data is large

Table 2: Comparison of model predictive performance. We run all methods for five times and report the averaged F1 scores with standard deviations.

Experiment Results: Human Evaluation

Human Evaluation Interface

All models predict the **may_treat** relation between t1 term unfractionated heparin ['unfractionated heparin [epc]', 'heparin'] and t2 term myocardial infarction (mi) ['myocardial infarction'] with the following rationales.

1. Check each rationale and answer this question: To which degree is this rationale helpful for you to trust the prediction? *(0: no helpful; 1: a little bit helpful; 2: helpful; 3: very helpful)*

Model A's Rationale Set:

T1's contexts	Relational Interaction	T2's contexts	Score
metabolic alkalosis	may_prevent	myocardial infarction (mi)	
metabolic alkalosis	may_prevent	venous thrombosis	
rbbb	may_treat	myocardial infarction (mi)	
ards	symptom_of	myocardial infarction (mi)	
micronutrient	may_prevent	venous thrombosis	

Model B's Rationale Set:

T1's contexts	Relational Interaction	T2's contexts	Score
cardiac dysrhythmias	contraindicates	theophylline	
malignant neoplasm without specification of site	has_symptom	family history of cancer	
lddm	contraindicates	glyburide	
morphine sulfate	contraindicated_by	respiratory depression	
insulin dependent diabetes	contraindicates	glyburide	

2. Please also rank all sets of rationales based on overall how much they help you trust the model prediction (e.g., A > B). Note that it is ok to reject them if both models are unhelpful (A = B = 0).

Experiment Results: Human Evaluation

Human Evaluation Interface

All models predict the **may_treat** relation between t1 term unfractionated heparin ['unfractionated heparin [epc]', 'heparin'] and t2 term myocardial infarction (mi) ['myocardial infarction'] with the following rationales.

1. Check each rationale and answer this question: To which degree is this rationale helpful for you to trust the prediction? *(0: no helpful; 1: a little bit helpful; 2: helpful; 3: very helpful)*

Model A's Rationale Set:

T1's contexts	Relational Interaction	T2's contexts	Score
metabolic alkalosis	may_prevent	myocardial infarction (mi)	
metabolic alkalosis	may_prevent	venous thrombosis	
rbbb	may_treat	myocardial infarction (mi)	
ards	symptom_of	myocardial infarction (mi)	
micronutrient	may_prevent	venous thrombosis	

Model B's Rationale Set:

T1's contexts	Relational Interaction	T2's contexts	Score
cardiac dysrhythmias	contraindicates	theophylline	
malignant neoplasm without specification of site	has_symptom	family history of cancer	
lddm	contraindicates	glyburide	
morphine sulfate	contraindicated_by	respiratory depression	
insulin dependent diabetes	contraindicates	glyburide	

2. Please also rank all sets of rationales based on overall how much they help you trust the model prediction (e.g., A > B). Note that it is ok to reject them if both models are unhelpful (A = B = 0).

_____

Experiment Results: Human Evaluation

Human Evaluation Interface

All models predict the **may_treat** relation between t1 term unfractionated heparin ['unfractionated heparin [epc]', 'heparin'] and t2 term myocardial infarction (mi) ['myocardial infarction'] with the following rationales.

1. Check each rationale and answer this question: To which degree is this rationale helpful for you to trust the prediction? *(0: no helpful; 1: a little bit helpful; 2: helpful; 3: very helpful)*

Model A's Rationale Set:

T1's contexts	Relational Interaction	T2's contexts	Score
metabolic alkalosis	may_prevent	myocardial infarction (mi)	
metabolic alkalosis	may_prevent	venous thrombosis	
rbbb	may_treat	myocardial infarction (mi)	
ards	symptom_of	myocardial infarction (mi)	
micronutrient	may_prevent	venous thrombosis	

Model B's Rationale Set:

T1's contexts	Relational Interaction	T2's contexts	Score
cardiac dysrhythmias	contraindicates	theophylline	
malignant neoplasm without specification of site	has_symptom	family history of cancer	
lddm	contraindicates	glyburide	
morphine sulfate	contraindicated_by	respiratory depression	
insulin dependent diabetes	contraindicates	glyburide	

2. Please also rank all sets of rationales based on overall how much they help you trust the model prediction (e.g., A > B). Note that it is ok to reject them if both models are unhelpful (A = B = 0).

Human Evaluation Score

	OWA Rationales	CWA Rationales
Ranking Score	17	5
Avg. Sum Score/Case	6.14	2.24
Avg. Max Score/Case	2.04	0.77

Table 3: Human evaluation on the quality of rationales.

Experiment Results: Human Evaluation

Human Evaluation Interface

All models predict the **may_treat** relation between t1 term unfractionated heparin ['unfractionated heparin [epc]', 'heparin'] and t2 term myocardial infarction (mi) ['myocardial infarction'] with the following rationales.

1. Check each rationale and answer this question: To which degree is this rationale helpful for you to trust the prediction? *(0: no helpful; 1: a little bit helpful; 2: helpful; 3: very helpful)*

Model A's Rationale Set:

T1's contexts	Relational Interaction	T2's contexts	Score
metabolic alkalosis	may_prevent	myocardial infarction (mi)	
metabolic alkalosis	may_prevent	venous thrombosis	
rbbb	may_treat	myocardial infarction (mi)	
ards	symptom_of	myocardial infarction (mi)	
micronutrient	may_prevent	venous thrombosis	

Model B's Rationale Set:

T1's contexts	Relational Interaction	T2's contexts	Score
cardiac dysrhythmias	contraindicates	theophylline	
malignant neoplasm without specification of site	has_symptom	family history of cancer	
lddm	contraindicates	glyburide	
morphine sulfate	contraindicated_by	respiratory depression	
insulin dependent diabetes	contraindicates	glyburide	

2. Please also rank all sets of rationales based on overall how much they help you trust the model prediction (e.g., A > B). Note that it is ok to reject them if both models are unhelpful (A = B = 0).

Human Evaluation Score

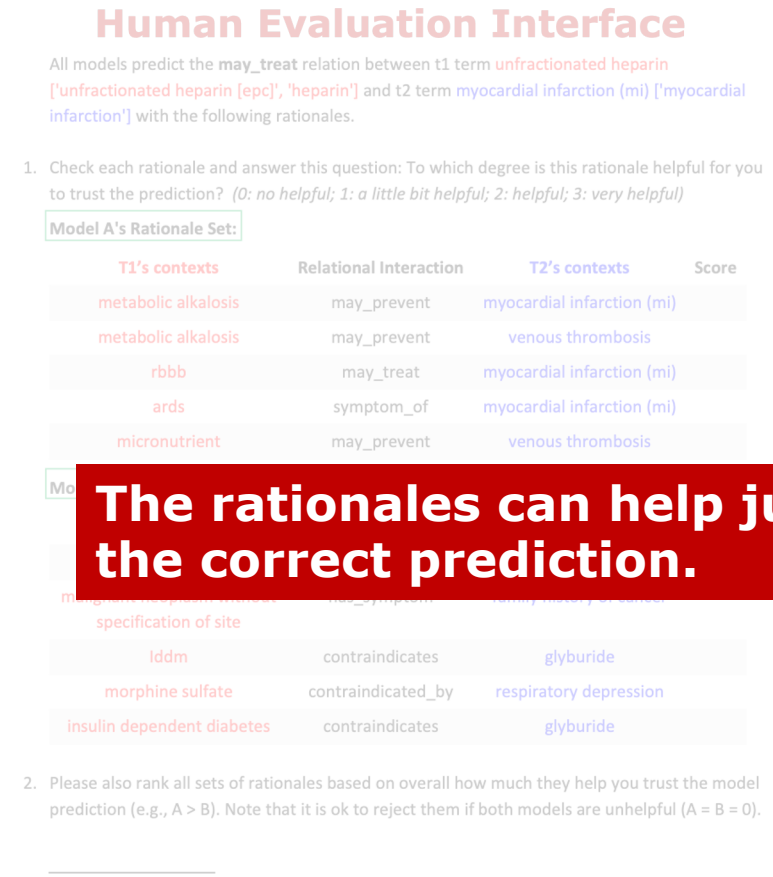
	OWA Rationales	CWA Rationales
Ranking Score	17	5
Avg. Sum Score/Case	6.14	2.24
Avg. Max Score/Case	2.04	0.77

Table 3: Human evaluation on the quality of rationales.

Case Study

Case 1		
cephalosporins	may_treat	bacterial infection
cefuroxime	may_treat	viral syndrome
cefuroxime	may_treat	low grade fever
cefuroxime	**may_treat**	**infectious diseases**
cefuroxime	may_prevent	low grade fever
sulbactam	may_treat	low grade fever

Experiment Results: Human Evaluation



The rationales can help justify the correct prediction.

	OWA Rationales	CWA Rationales
Ranking Score	17	5
Avg. Sum Score/Case	6.14	2.24
Avg. Max Score/Case	2.04	0.77

Table 3: Human evaluation on the quality of rationales.

Case Study

Case 1		
cephalosporins	may_treat	bacterial infection
cefuroxime	may_treat	viral syndrome
cefuroxime	may_treat	low grade fever
cefuroxime	**may_treat**	**infectious diseases**
cefuroxime	may_prevent	low grade fever
sulbactam	may_treat	low grade fever

See more details in the paper.

Conclusions

- We propose an interpretable framework to rationalize medical relation prediction based on corpus-level statistics

Conclusions

- We propose an interpretable framework to rationalize medical relation prediction based on corpus-level statistics

- Inspired by existing cognitive theories, the reasoning process can be easily understood by users and provides reasonable explanations to justify its prediction

Conclusions

- We propose an interpretable framework to rationalize medical relation prediction based on corpus-level statistics
- Inspired by existing cognitive theories, the reasoning process can be easily understood by users and provides reasonable explanations to justify its prediction
- We demonstrate the model effectiveness by its predictive performance and human evaluation.

Thank You!

THE OHIO STATE UNIVERSITY

Zhen Wang

The Ohio State University

wang.9125@osu.edu

Code available at:
https://github.com/zhenwang9102/X-MedRELA

Supported By:

pcori®

NSF

SCIENCE TO SHAPE THE FUTURE OF THE ARMY
ARMY RESEARCH OFFICE
1951