Rationalizing Medical Relation Prediction from Corpus-level Statistics

Zhen Wang[1], **Jennifer Lee**[2,3], **Simon Lin**[4], **Huan Sun**[1]
[1]The Ohio State University
[2]Department of Family Medicine, The Ohio State University Wexner Medical Center
[3]Department of Physician Informatics, Nationwide Children's Hospital
[4]Abigail Wexner Research Institute at Nationwide Children's Hospital
{wang.9215, sun.397}@osu.edu
{Jennifer.Lee2, Simon.Lin}@nationwidechildrens.org

Abstract

Nowadays, the interpretability of machine learning models is becoming increasingly important, especially in the medical domain. Aiming to shed some light on how to rationalize medical relation prediction, we present a new interpretable framework inspired by existing theories on how human memory works, e.g., theories of recall and recognition. Given the corpus-level statistics, i.e., a global co-occurrence graph of a clinical text corpus, to predict the relations between two entities, we first *recall* rich contexts associated with the target entities, and then *recognize* relational interactions between these contexts to form model rationales, which will contribute to the final prediction. We conduct experiments on a real-world public clinical dataset and show that our framework can not only achieve competitive predictive performance against a comprehensive list of neural baseline models, but also present rationales to justify its prediction. We further collaborate with medical experts deeply to verify the usefulness of our model rationales for clinical decision making[1].

1 Introduction

Predicting relations between entities from a text corpus is a crucial task in order to extract structured knowledge, which can empower a broad range of downstream tasks, e.g., question answering (Xu et al., 2016), dialogue systems (Lowe et al., 2015), reasoning (Das et al., 2017), etc. There has been a large amount of existing work focusing on predicting relations based on *raw texts* (e.g., sentences, paragraphs) mentioning two entities (Hendrickx et al., 2010; Zeng et al., 2014; Zhou et al., 2016; Mintz et al., 2009; Riedel et al., 2010; Lin et al., 2016; Verga et al., 2018; Yao et al., 2019).

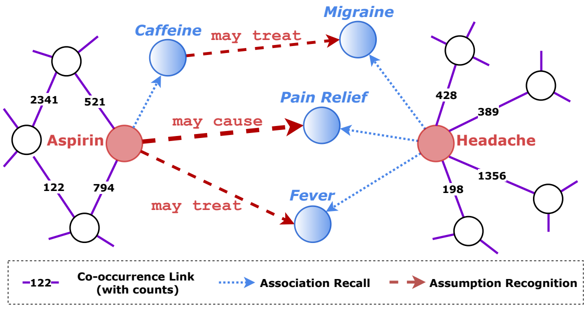


Figure 1: Our intuition for how to rationalize relation prediction based on the corpus-level statistics. To infer the relation between the target entities (red nodes), we recall (blue dashed line) their associated entities (blue nodes) and infer their relational interactions (red dashed line), which will serve as assumptions or model rationales to support the target relation prediction.

In this paper, we study a relatively new setting in which we predict relations between entities based on the *global co-occurrence statistics* aggregated from a text corpus, and focus on medical relations and clinical texts in Electronic Medical Records (EMRs). The corpus-level statistics present a *holistic graph view* of all entities in the corpus, which will greatly facilitate the relation inference, and can better preserve patient privacy than raw or even de-identified textual content and are becoming a popular substitute for the latter in the research community for studying EMR data (Finlayson et al., 2014; Wang et al., 2019).

To predict relations between entities based on a global co-occurrence graph, intuitively, one can first optimize the graph embedding or global word embedding (Pennington et al., 2014; Perozzi et al., 2014; Tang et al., 2015), and then develop a relation classifier (Nickel et al., 2011; Socher et al., 2013; Yang et al., 2015; Wang et al., 2018) based on the embedding vectors of the two entities. However, such kind of neural frameworks often lack the desired *interpretability*, which is especially important for the medical domain. In general, despite

[1] Our code and datasets are available at: https://github.com/zhenwang9102/X-MedRELA

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8078–8092
July 5 - 10, 2020.

their superior predictive performance in many NLP tasks, the opaque decision-making process of neural models has concerned their adoption in high stakes domains like medicine, finance, and judiciary (Rudin, 2019; Murdoch et al., 2019). Building models that provide reasonable explanations and have increased transparency can remarkably enhance user trust (Ribeiro et al., 2016; Miller, 2019). In this paper, we aim to develop such a model for our medical relation prediction task.

To start with, we draw inspiration from the existing theories on cognitive processes about how human memory works, e.g., two types of memory retrieval (recall and recognition) (Gillund and Shiffrin, 1984). Basically, in the *recall* process, humans tend to retrieve contextual associations from long-term memory. For example, given the word "Paris", one may think of "Eiffel Tower" or "France", which are strongly associated with "Paris" (Nobel and Shiffrin, 2001; Kahana et al., 2008; Budiu, 2014). Besides, there is a strong correlation between the association strength and the co-occurrence graph (Spence and Owens, 1990; Lundberg and Lee, 2017). In the *recognition* process, humans typically recognize if they have seen a certain piece of information before. Figure 1 shows an example in the context of relation prediction. Assume a model is to predict whether *Aspirin* may treat *Headache* or not (That "*Aspirin* may treat *Headache*" is a known fact, and we choose this relation triple for illustration purposes). It is desirable if the model could perform the aforementioned two types of memory processes and produce rationales to base its prediction upon: (1) Recall. What entities are associated with *Aspirin*? What entities are associated with *Headache*? (2) Recognition. Do those associated entities hold certain relations, which can be leveraged as clues to predict the target relation? For instance, a model could first retrieve a relevant entity *Pain Relief* for the tail entity *Headache* as they co-occur frequently, and then recognize there is a chance that *Aspirin* can lead to *Pain Relief* (i.e., formulate model rationales or assumptions), based on which it could finally make a correct prediction (*Aspirin*, may treat, *Headache*).

Now we formalize such intuition to rationalize the relation prediction task. Our framework consists of three stages, *global association recall* (CogStage-1), *assumption formation and representation* (CogStage-2), and *prediction decision making* (CogStage-3), shown in Figure 2. CogStage-1 models the process of recalling diverse contextual entities associated with the target head and tail entities respectively, CogStage-2 models the process of recognizing possible interactions between those recalled entities, which serve as model rationales (or, assumptions[2]) and are represented as semantic vectors, and finally CogStage-3 aggregates all assumptions to infer the target relation. We jointly optimize all three stages using a training set of relation triples as well as the co-occurrence graph. Model rationales can be captured through this process *without any gold rationales* available as direct supervision. Overall, our framework rationalizes its relation prediction and is interpretable to users[3] by providing justifications for (i) why a particular prediction is made, (ii) how the assumptions of the prediction are developed, and (iii) how the particular assumptions are relied on.

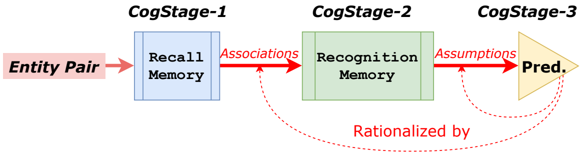


Figure 2: A high-level illustration of our framework.

On a real-life clinical text corpus, we compare our framework with various competitive methods to evaluate the predictive performance and interpretability. We show that our method obtains very competitive performance compared with a comprehensive list of various neural baseline models. Moreover, we follow recent work (Singh et al., 2019; Jin et al., 2020) to quantitatively evaluate model interpretability and demonstrate that rationales produced by our framework can greatly help earn expert trust. To summarize, we study the important problem of rationalizing medical relation prediction based on corpus-level statistics and propose a new framework inspired by cognitive theories, which outperforms competitive baselines in terms of both interpretability and predictive performance.

2 Background

Different from existing work using raw texts for relation extraction, we assume a global co-occurrence graph (i.e., corpus-level statistics) is given, which was pre-constructed based on a text corpus $\mathcal{D}$, and denote it as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where

[2] We use the two terms interchangeably in this paper.

[3] Following Murdoch et al. (2019), desired interpretability is supposed to provide insights to particular audiences, which in our case are medical experts.

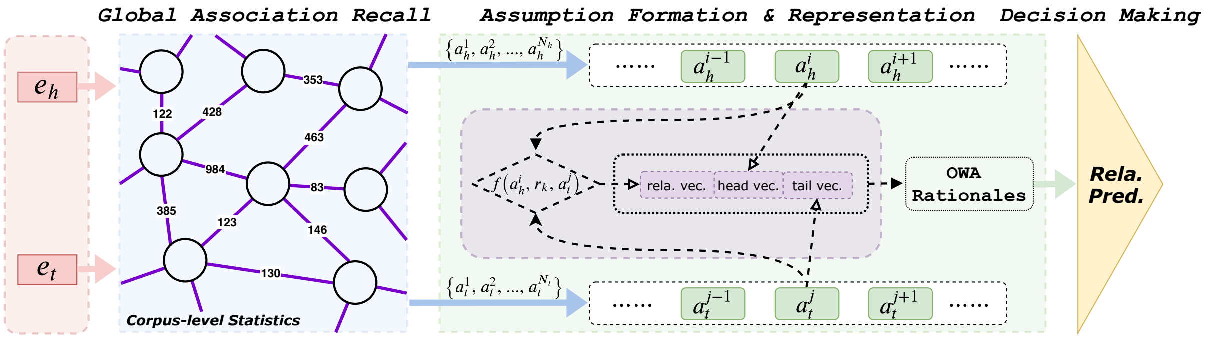


Figure 3: Framework Overview.

each vertex $v \in \mathcal{V}$ represents an entity extracted from the corpus and each edge $e \in \mathcal{E}$ is associated with the global co-occurrence count for the connected nodes. Counts reflect how frequent two entities appear in the same context (e.g., co-occur in the same sentence, document, or a certain time frame). In this paper, we focus on clinical co-occurrence graph in which vertices are medical terms extracted from clinical notes. Nevertheless, as we will see later, our framework is very general and can be applied to other relations with corpus-level statistics.

Our motivation for working under this setting lies in three folds: (1) Such graph data is stripped of raw textual contexts and thus, has a better preserving of patient privacy (Wang et al., 2019), which makes itself easier to be constructed and shared under the HIPPA protected environments (Act, 1996) for medical institutes (Finlayson et al., 2014); (2) Compared with open-domain relation extraction, entities holding a medical relation oftentimes do not co-occur in a local context (e.g., a sentence or paragraph). For instance, we observe that in a widely used clinical co-occurrence graph (Finlayson et al., 2014), which is also employed for our experiments later, of all entity pairs holding the treatment relation according to UMLS (Unified Medical Language System), only about 11.4% have a co-occurrence link (i.e., co-occur in clinical notes within a time frame like 1 day or 7 days); (3) As suggested by cognitive theories (Spence and Owens, 1990), lexical co-occurrence is significantly correlated with association strength in the recall memory process, which further inspires us to utilize such statistics to find associations and form model rationales for relation prediction.

Finally, our relation prediction task is formulated as: Given the global statistics $\mathcal{G}$ and an entity pair, we predict whether they hold a relation r (e.g., MAY_TREAT), and moreover provide a set of model rationales $\mathcal{T}$ composed of relation triples for the prediction. For the example in Figure 1, we aim to build a model that will not only accurately predict the MAY_TREAT relation, but also provide meaningful rationales on how the prediction is made, which are crucial for gaining trust from clinicians.

3 Methodology

Following a high-level framework illustration in Figure 2, we show a more detailed overview in Figure 3 and introduce each component as follows.

3.1 CogStage-1: Global Association Recall

Existing cognitive theories (Kahana et al., 2008) suggest that *recall* is an essential function of human memory to retrieve *associations* for later decision making. On the other hand, the association has been shown to significantly correlate with the lexical co-occurrence from the text corpus (Spence and Owens, 1990; Lund and Burgess, 1996). Inspired by such theories and correlation, we explicitly build up our model based on recalled associations stemming from corpus-level statistics and provide global highly-associated contexts as the source of interpretations.

Given an entity, we build an estimation module to globally infer associations based on the corpus-level statistics. Our module leverages distributional learning to fully explore the graph structure. One can also directly utilize the raw neighborhoods in the co-occurrence graph, but due to the noise introduced in the preprocessing of building the graph, it is a less optimal choice in real practice.

Specifically, for a selected node/entity $e_i \in \mathcal{E}$, our global association recall module estimates a conditional probability $p\,(e_j|e_i)$, representing how likely the entity $e_j \in \mathcal{E}$ is associated with e_i[4]. We formally define such conditional probability as:

$$p\,(e_j|e_i) = \frac{\exp\,(\boldsymbol{v'}^T_{e_j} \cdot \boldsymbol{v}_{e_i})}{\sum_{k=1}^{|\mathcal{V}|} \exp\,(\boldsymbol{v'}^T_{e_k} \cdot \boldsymbol{v}_{e_i})} \qquad (1)$$

[4] We assume all existing entities can be possible associations for the given entity.

where $\boldsymbol{v}_{e_i} \in \mathbb{R}^d$ is the embedding vector of node e_i and $\boldsymbol{v}'_{e_j} \in \mathbb{R}^d$ is the context embedding for e_j.

There are many ways to approximate $p\left(e_j|e_i\right)$ from the global statistics, e.g., using global log-bilinear regression (Pennington et al., 2014). To estimate such probabilities and update entity embeddings efficiently, we optimize the conditional distribution $p\left(e_j|e_i\right)$ to be close to the empirical distribution $\hat{p}\left(e_j|e_i\right)$ defined as:

$$\hat{p}\left(e_j|e_i\right) = \frac{p_{ij}}{\sum_{(i,k)\in\mathcal{E}} p_{ik}} \quad (2)$$

where $\mathcal{E}$ is the set of edges in the co-occurrence graph and p_{ij} is the PPMI value calculated by the co-occurrence counts between node e_i and e_j. We adopt the cross entropy loss for the optimization:

$$\mathcal{L}_n = -\sum_{(e_i,e_j)\in\mathcal{V}} \hat{p}(e_j|e_i)\log\left(p(e_j|e_i)\right) \quad (3)$$

This association recall module will be jointly trained with other objective functions to be introduced in the following sections. After that, given an entity e_i, we can select the top-N_c entities from $p(\cdot|e_i)$ as e_i's associative entities for subsequent assumption formation.

3.2 CogStage-2: Assumption Formation and Representation

As shown in Figure 3, with the associative entities from CogStage-1, we are ready to *formulate* and *represent* assumptions. In this paper, we define model assumptions as *relational interactions between associations*, that is, as shown in Figure 1, the model may identify (*Caffeine*, MAY_TREAT, *Migraine*) as an assumption, which could help predict *Aspirin* may treat *Headache* (*Caffeine* and *Migraine* are associations for *Aspirin* and *Headache* respectively). Such relational rationales are more concrete and much easier for humans to understand than the widely-adopted explanation strategy (Yang et al., 2016; Mullenbach et al., 2018; Vashishth et al., 2019) in NLP that is based on pure attention weights on local contexts.

One straightway way to obtain such rationales is to query existing medical knowledge bases (KBs), e.g., (*Caffeine*, MAY_TREAT, *Migraine*) may exist in SNOMED CT[5] and can serve as a model rationale. We refer to rationales acquired in this way as the *Closed-World Assumption* (CWA) (Reiter, 1981) setting since only KB-stored facts are considered and trusted in a closed world. In contrast to the CWA rationales, considering the sparsity and incompleteness issues of KBs that are even more severe in the medical domain, we also propose the *Open-World Assumptions* (OWA) (Ceylan et al., 2016) setting to discover richer rationales by estimating all potential relations between associative entities based on a seed set of relation triples (which can be regarded as prior knowledge).

In general, the CWA rationales are relatively more accurate as each fact triple has been verified by the KB, but would have a low coverage of other possibly relevant rationales for the target prediction. On the other hand, the OWA rationales are more comprehensive but could be noisy and less accurate, due to the probabilistic estimation procedure and the limited amount of prior knowledge. However, as we will see, by aggregating all OWA rationales into the whole framework with an attention-based mechanism, we can select high-quality and most relevant rationales for prediction. For the rest of the paper, by default we adopt the OWA setting in our framework and describe its details as follows.

Specifically, given a pair of head and tail entity, $e_h, e_t \in \mathcal{V}$, let us denote their association sets as $\mathcal{A}(e_h) = \{a_h^i\}_{i=1}^{N_h}$ and $\mathcal{A}(e_t) = \{a_t^j\}_{j=1}^{N_t}$, where N_h, N_t are the number of associative entities a_h, a_t to use. Each entity has been assigned an embedding vector by the previous association recall module. We first measure the probability of relations holding for the pair. Given $a_h^i \in \mathcal{A}(e_h), a_t^j \in \mathcal{A}(e_t)$ and a relation $r_k \in \mathcal{R}$, we define a scoring function as Bordes et al. (2013) to estimate triple quality:

$$s_k^{ij} = f(a_h^i, r_k, a_t^j) = -||\boldsymbol{v}_{a_h^i} + \boldsymbol{\xi}_k - \boldsymbol{v}_{a_t^j}||_1 \quad (4)$$

where $\boldsymbol{v}_{a_h^i}$ and $\boldsymbol{v}_{a_t^j}$ are embedding vectors, relations are parameterized by a relation matrix $R \in \mathbb{R}^{N_r\times d}$ and $\boldsymbol{\xi}_k$ is its k-th row vector. Such a scoring function encourages larger value for correct triples. Additionally, in order to filter unreliable estimations, we define an $\mathtt{NA}$ relation to represent other trivial relations or no relation as the score, $s_{\mathtt{NA}}^{ij} = f(a_h^i, \mathtt{NA}, a_t^j)$, which can be seen as a dynamic threshold to produce reasonable rationales.

Now we *formulate* OWA rationales by calculating the conditional probability of a relation given a pair of associations as follows (we save the superscript ij for space):

$$p(r_k|a_h^i, a_t^j) = \begin{cases} \dfrac{\exp\left(s_k\right)}{\sum_{s_k\geq s_{\mathtt{NA}}}\exp\left(s_k\right)}, & s_k > s_{\mathtt{NA}} \\ 0, & s_k \leq s_{\mathtt{NA}} \end{cases} \quad (5)$$

[5] https://www.snomed.org/

For each association pair, (a_h^i, a_t^j), we only form an assumption with a relation r_k^* if r_k^* is top ranked according to $p(r_k|a_h^i, a_t^j)$.[6]

To *represent* assumptions, we integrate all relation information per pair into a single vector representation. Concretely, we calculate the assumption representation by treating $p(r_k|a_h^i, a_t^j)$ as weights for all relations as follows:

$$\mathsf{a}_{ij} = \rho(a_h^i, a_t^j; \mathcal{R}) = \sum_{k'=1}^{N_r} p(r_{k'}|a_h^i, a_t^j) \cdot \boldsymbol{\xi}_{k'} \quad (6)$$

Finally, we combine the entity vectors as well as the relation vector to get the final representation of assumptions for association pair (a_h^i, a_t^j), where $c_i \in \mathcal{A}(e_h)$ and $c_j \in \mathcal{A}(e_t)$:

$$\mathsf{e}_{ij} = \tanh([\boldsymbol{v}_{a_h^i}; \boldsymbol{v}_{a_t^j}; \mathsf{a}_{ij}]\boldsymbol{W}_p + \boldsymbol{b}_p) \quad (7)$$

where $[\cdot\,;\cdot]$ represents vector concatenation, $\boldsymbol{W}_p \in \mathbb{R}^{3d \times d_p}$, $\boldsymbol{b}_p \in \mathbb{R}^{d_p}$ are the weight matrix and bias in a fully-connected network.

3.3 CogStage-3: Prediction Decision Making

Analogical to human thinking, our decision making module aggregates all assumption representations and measures their accountability for the final prediction. It learns a distribution over all assumptions and we select the ones with highest probabilities as model rationales. More specifically, we define a scoring function $g(\mathsf{e}_{ij})$ to estimate the accountability based on the assumption representation e_{ij} and normalize $g(\mathsf{e}_{ij})$ as:

$$g(\mathsf{e}_{ij}) = \boldsymbol{v}^T \cdot \tanh(\boldsymbol{W}_a \mathsf{e}_{ij} + \boldsymbol{b}_a) \quad (8)$$

$$p_{ij} = \frac{\exp(g(\mathsf{e}_{ij}))}{\sum_{m=1}^{N_h} \sum_{n=1}^{N_t} \exp(g(\mathsf{e}_{mn}))} \quad (9)$$

where $\boldsymbol{W}_a, \boldsymbol{b}_a$ are the weight matrix and bias for the scoring function. Then we get the weighted rationale representation as:

$$\mathsf{r} = \psi(e_h, e_t) = \sum_{i=1}^{N_h} \sum_{j=1}^{N_t} p_{ij} \mathsf{e}_{ij} \quad (10)$$

With the representation of weighted assumption information for the target pair (e_h, e_t), we calculate the binary prediction probability for relation r as:

$$p(r|e_h, e_t) = \sigma(\boldsymbol{W}_r \mathsf{r} + \boldsymbol{b}_r) \quad (11)$$

where $\sigma(x) = 1/(1 + \exp(-x))$ and $\boldsymbol{W}_r, \boldsymbol{b}_r$ are model parameters.

Rationalizing relation prediction. After fully training the entire model, to recover the most contributing assumptions for predicting the relation between the given target entities (e_h, e_t), we compute the importance scores for all assumptions and select those most important ones as model rationales. In particular, we multiply p_{ij} (the weight for association pair (a_h^i, a_t^j) in Eqn. 9) with $p(r_k|a_h^i, a_t^j)$ (the probability of a relation given the pair (a_h^i, a_t^j) in Eqn. 5) to score the triple (a_h^i, r_k, a_t^j). We rank all such triples for $a_h^i \in \mathcal{A}(e_h), a_t^j \in \mathcal{A}(e_t), r_k \in \mathcal{R}$ and select the top-K triples as model rationales for the final relation prediction.

3.4 Training

We now describe how we train our model efficiently for multiple modules. For relational learning to estimate the conditional probability $p(r_k|a_h^i, a_t^j)$, we utilize training data as the seed set of triples for all relations as correct triples denoted as $(h, r, t) \in \mathcal{P}$. The scoring function in Eqn. 4 is expected to score higher for correct triples than the corrupted ones in which we denote $\mathcal{N}(?, r, t)$ ($\mathcal{N}(t, r, ?)$) as the set of corrupted triples by replacing the head (tail) entity randomly. Instead of using margin-based loss function, we adopt a more efficient training strategy from (Kadlec et al., 2017; Toutanova and Chen, 2015) with a negative log likelihood loss function as:

$$\begin{aligned}\mathcal{L}_r = &- \textstyle\sum_{(h,r,t)\in\mathcal{P}} \log p\,(h|t, r) \\ &- \textstyle\sum_{(h,r,t)\in\mathcal{P}} \log p\,(t|h, r)\end{aligned} \quad (12)$$

where the conditional probability $p(h|t, r)$ is defined as follows ($p(t|h, r)$ is defined similarly):

$$p\,(h|t, r) = \frac{\exp(f\,(h, r, t))}{\sum_{h' \in \mathcal{N}(?,r,t)} \exp(f\,(h', r, t))} \quad (13)$$

For our binary relation prediction task, we define a binary cross entropy loss function with Eqn. 11 as follows:

$$\begin{aligned}\mathcal{L}_p = &- \textstyle\sum_{i=1}^{M}(y_i \cdot \log(p(r|e_h^i, e_t^i)) \\ &+ (1 - y_i) \cdot \log(1 - p(r|e_h^i, e_t^i)))\end{aligned} \quad (14)$$

where M is the number of samples, y_i is the label showing whether e_h, e_t holds a certain relation.

The above three loss functions, i.e., $\mathcal{L}_n$ for global association recall, $\mathcal{L}_r$ for relational learning and $\mathcal{L}_p$ for relation prediction, are all jointly optimized. All three of them share the entity embeddings and $\mathcal{L}_p$ will reuse the relation matrix from $\mathcal{L}_r$ to conduct the rationale generation.

[6] We remove the target relation to predict if it exists in the assumption set.

4 Experiments

In this section, we first introduce our experimental setup, e.g, the corpus-level co-occurrence statistics and datasets used for our experiments, and then compare our model with a list of comprehensive competitive baselines in terms of predictive performance. Moreover, we conduct expert evaluations as well as case studies to demonstrate the usefulness of our model rationales.

4.1 Dataset

We directly adopt a publicly available medical co-occurrence graph for our experiments (Finlayson et al., 2014). The graph was constructed in the following way: Finlayson et al. (2014) first used an efficient annotation tool (LePendu et al., 2012) to extract medical terms from 20 million clinical notes collected by Stanford Hospitals and Clinics, and then computed the co-occurrence counts of two terms based on their appearances in one patient's records within a certain time frame (e.g., 1 day, 7 days). We experiment with their biggest dataset with the largest number of nodes (i.e., the per-bin 1-day graph here[7]) so as to have sufficient training data. The co-occurrence graph contains 52,804 nodes and 16,197,319 edges.

To obtain training labels for relation prediction, we utilize the mapping between medical terms and concepts provided by Finlayson et al. (2014). To be specific, they mapped extracted terms to UMLS concepts with a high mapping accuracy by suppressing the least possible meanings of each term (see Finlayson et al. (2014) for more details). We utilize such mappings to automatically collect relation labels from UMLS. For term e_a and e_b that are respectively mapped to medical concept c_A and c_B, we find the relation between c_A and c_B in UMLS, which will be used as the label for e_a and e_b.

Following Wang and Fan (2014) that studied distant supervision in medical text and identified several crucial relations for clinical decision making, we select 5 important medical relations with no less than 1,000 relation triples in our dataset. Each relation is mapped to UMLS semantic relations, e.g., relation CAUSES corresponds to *cause_of*, *induces*, *causative_agent_of* in UMLS. A full list of mapping is in the appendix. We sample an equal number of negative pairs by randomly pairing head and tail entities with the correct argument types (Wang et al., 2016). We split all samples into train/dev/test sets with a ratio of 70/15/15. Only relation triples in the training set are used to optimize relational parameters. The statistics of the positive samples for relations are summarized in Table 1.

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.

4.2 Predictive Performance Evaluation

Compared Methods. There are a number of advanced neural methods (Tang et al., 2015; Qu et al., 2018; Wang et al., 2018) that have been developed for the link prediction task, i.e., predicting the relation between two nodes in a co-occurrence graph. At the high level, their frameworks comprise of an entity encoder and a relation scoring function. We adapt various existing methods for both the encoder and the scoring functions for comprehensive comparison. Specifically, given the co-occurrence graph, we employ existing distributional representation learning methods to learn entity embeddings. With the entity embeddings as input features, we adapt various models from the knowledge base completion literature as a binary relation classifier. More specifically, for the encoder, we select one word embedding method, Word2vec (Mikolov et al., 2013; Levy and Goldberg, 2014), two graph embedding methods, random-walk based DeepWalk (Perozzi et al., 2014), edge-sampling based LINE (Tang et al., 2015), and one distributional approach REPEL-D (Qu et al., 2018) for weakly-supervised relation extraction that leverages both the co-occurrence graph and training relation triples to learn entity representations. For the scoring functions, we choose DistMult (Yang et al., 2015), RESCAL (Nickel et al., 2011) and NTN (Socher et al., 2013).

Note that one can apply more complex encoders or scoring functions to obtain higher predictive performance; however, in this work, we emphasize more on model interpretability than predictive performance, and unfortunately, all such frameworks are hard to interpret as they provide little or no

[7] https://datadryad.org/stash/dataset/doi:10.5061/dryad.jp917

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.

explanations on how predictions are made.

We also show the predictive performance of our framework under the CWA setting in which the CWA rationales are existing triples in a "closed" knowledge base (i.e., UMLS). We first adopt the pre-trained association recall module to retrieve associative contexts for head and tail entities, then formulate the assumptions using top-ranked triples (that exist in our relation training data), where the rank is based on the product of their retrieval probabilities ($p_{ij} = p(e_i|e_h) \times p(e_j|e_t)$). We keep the rest of our model the same as the OWA setting.

Results. We compare the predictive performance of different models in terms of F1 score under each relation prediction task. As shown in Table 2, our model obtains very competitive performance compared with a comprehensive list of baseline methods. Specifically, on the prediction tasks of MAY_TREAT and CONTRAINDICATES, our model achieves a substantial improvement ($1\sim2$ F1 score) and a very competitive performance on the task of SYMPTOM_OF and MAY_PREVENT. The small amount of training data might partly explain why our model does not perform so well in the CAUSES tasks. Such comparison shows the effectiveness of predicting relations based on associations and their relational interactions. Moreover, compared with those baseline models which encode graph structure into latent vector representation, our model utilizes co-occurrence graph more explicitly by leveraging the associative contexts symbolically to generate human-understandable rationales, which can assist medical experts as we will see shortly. In addition, we observe that our model consistently outperforms the CWA setting: Despite the CWA rationales are true statements on their own, they tend to have a low coverage of possible rationales, and thus, may be not so relevant for the target relation prediction, which leads to a poor predictive performance.

	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.

4.3 Model Rationale Evaluation

To measure the quality of our model rationales (i.e., OWA rationales), as well as to conduct an ablation study of our model, we conduct an expert evaluation for the OWA rationales and also compare them with the CWA rationales. We first collaborate with a physician to explore how much a model's rationales help them better trust the model's prediction following recent work for evaluating model interpretability (Singh et al., 2019; Mullenbach et al., 2018; Atutxa et al., 2019; Jin et al., 2020). Then, we present some case studies to show what kind of rationales our model has learnt. Note that compared with evaluation by human annotators for open-domain tasks (without expertise requirement), evaluation by medical experts is more challenging in general. The physician in our study (an M.D. with 9 years of clinical experience and currently a fellow trained in clinical informatics), who is able to understand the context of terms and the basics of the compared algorithms and can dedicate time, is qualified for our evaluation.

Expert Evaluation. We first explained to the physician about the recall and recognition process in our framework and how model rationales are developed. They endorsed such reasoning process as one possible way to gain their trust in the model. Next, for each target pair for which our model correctly makes the prediction, they were shown the top-5 rationales produced by our framework and were asked whether each rationale helps them better trust the model prediction. For each rationale, they were asked to score it from 0 to 3 in which 0 is *no helpful*, 1 is *a little helpful*, 2 is *helpful* and 3 is *very helpful*. In addition to the individual rationale evaluation, we further compare the overall quality of CWA and OWA rationales, by letting experts rank them based the helpfulness of each set of rationales (the rationale set ranked higher gets 1 ranking score and both get 0 if they have the same rank). We refer readers to the appendix for more details of the evaluation protocol. We randomly select 30 cases in the MAY_TREAT relation and the overall evaluation results are summarized in Table 3. Out of 30, OWA wins in 17 cases and gets higher scores on individual rationales per case on average. There are 8 cases where the two sets of rationales are ranked the same[8] and 5 cases where CWA is better. To get a better idea of how the OWA model obtains more trust, we calculate the average sum score per case, which shows the OWA model gets a higher overall score per case. Considering in some cases only a few rationales are able to get non-zero scores, we also calculate the average max score per case, which shows that our OWA model generally provides one *helpful* rationale (score>2) per case. Overall, as we can see, the OWA rationales are more helpful to gain expert trust.

Case Study. Table 4 shows two concrete examples demonstrating what kind of model rationales our framework bases its predictions on. We highlight the rationales that receive high scores from the physician for being especially useful for trusting the prediction. As we can see, our framework is able to make correct predictions based on reasonable rationales. For instance, to predict that "cephalosporine" may treat "bacterial infection", our model relies on the rationale that "cefuroxime" may treat "infectious diseases". We also note that not all rationales are clinically established facts or even make sense, due to the unsupervised rationale learning and the probabilistic assumption formation process, which leaves space for future work to further improve the quality of rationales. Nevertheless, such model rationales can provide valuable information or new insights for clinicians. For another example, as pointed out by the physician, different medications possibly having the same treatment response, as shown in Case 2, could be clinically useful. That is, if three medications are predicted to possibly treat the same condition and a physician is only aware of two doing so, one might get insights into trying the third one. To summarize, our model is able to provide reasonable rationales and help users understand how model predictions are made in general.

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
Case 2		
azelastine	may_treat	perennial allergic rhinitis
astepro	**may_treat**	**perennial allergic rhinitis**
pseudoephedrine	**may_treat**	**perennial allergic rhinitis**
ciclesonide	**may_treat**	**perennial allergic rhinitis**
overbite	may_treat	perennial allergic rhinitis
diclofenac	may_treat	perennial allergic rhinitis

Table 4: Case studies for rationalizing medical relation prediction. For each case, the first panel is target pair and the second is top-5 rationales (**Bold** ones are useful rationales with high scores from the physician). The left (right) most column is the head (tail) term and their relational associations.

5 Related Work

Relation Extraction (RE) typically focuses on predicting relations between two entities based on their text mentions, and has been well studied in both open domain (Mintz et al., 2009; Zeng et al., 2015; Riedel et al., 2013; Lin et al., 2016; Song et al., 2019; Deng and Sun, 2019) and biomedical domain (Uzuner et al., 2011; Wang and Fan, 2014; Sahu et al., 2016; Lv et al., 2016; He et al., 2019). Among them, most state-of-the-art work develops various powerful neural models by leveraging human annotations, linguistic patterns, distance supervision, etc. More recently, an increasing amount of work has been proposed to improve model's transparency and interpretability. For example, Lee et al. (2019) visualizes self-attention weights learned from BERT (Devlin et al., 2019) to explain relation prediction. However, such text-based interpretable

[8] Of which, 7 cases are indicated equally unhelpful.

models tend to provide explanations within a *local context* (e.g., words in a single sentence mentioning target entities), which may not capture a *holistic view* of all entities and their relations stored in a text corpus. We believe that such a holistic view is important for interpreting relations and can be provided to some degree by the *global statistics* from a text corpus. Moreover, global statistics have been widely used in the clinical domain as they can better preserve patient privacy (Finlayson et al., 2014; Wang et al., 2019).

On the other hand, in recent years, graph embedding techniques (Perozzi et al., 2014; Tang et al., 2015; Grover and Leskovec, 2016; Yue et al., 2019) have been widely applied to learn node representations based on graph structure. Representation learning based on global statistics from a text corpus (i.e., co-occurrence graph) has also been studied (Levy and Goldberg, 2014; Pennington et al., 2014). After employing such methods to learn entity embeddings, a number of relation classifiers (Nickel et al., 2011; Bordes et al., 2013; Socher et al., 2013; Yang et al., 2015; Wang et al., 2018) can be adopted for relation prediction. We compare our method with such frameworks to show its competitive predictive accuracy. However, such frameworks tend to be difficult to interpret as they provide little or no explanations on how decisions are made. In this paper, we focus more on model interpretability than predictive accuracy, and draw inspirations from existing cognitive theories of recall and recognition to develop a new framework, which is our core contribution.

Another line of research related to interpreting relation prediction is path-based knowledge graph (KG) reasoning (Gardner et al., 2014; Neelakantan et al., 2015; Guu et al., 2015; Xiong et al., 2017; Stadelmaier and Padó, 2019). In particular, existing paths mined from millions of relational links in knowledge graphs can be used to provide justifications for relation predictions. For example, to explain *Microsoft* and *USA* may hold the relation *CountryOfHeadquarters*, by traversing a KG, one can extract the path $\textit{Microsoft} \xrightarrow{\text{IsBasedIn}} \textit{Seattle} \xrightarrow{\text{CountryLocatedIn}} \textit{USA}$ as one explanation. However, such path-finding methods typically require large-scale relational links to infer path patterns, and cannot be applied to our co-occurrence graph as the co-occurrence links are unlabeled.

In addition, our work is closely related to the area of rationalizing machine decision by generating justifications/rationales accounting for model's prediction. In some scenarios, human rationales are provided as extra supervision for more explainable models (Zaidan et al., 2007; Bao et al., 2018). However, due to the high cost of manual annotation, model rationales are desired to be learned in an unsupervised manner(Lei et al., 2016; Bouchacourt and Denoyer, 2019; Zhao et al., 2019). For example, Lei et al. (2016) select a subset of words as rationales and Bouchacourt and Denoyer (2019) provide an explanation based on the absence or presence of "concepts", where the selected words and "concepts" are learned unsupervisedly. Different from text-based tasks, in this paper, we propose to rationalize relation prediction based on global co-occurrence statistics and similarly, model rationales in our work are captured without explicit manual annotation either, via a joint training framework.

6 Conclusion

In this paper, we propose an interpretable framework to rationalize medical relation prediction based on corpus-level statistics. Our framework is inspired by existing cognitive theories on human memory recall and recognition, and can be easily understood by users as well as provide reasonable explanations to justify its prediction. Essentially, it leverages corpus-level statistics to recall associative contexts and recognizes their relational connections as model rationales. Compared with a comprehensive list of baseline models, our model obtains competitive predictive performances. Moreover, we demonstrate its interpretability via expert evaluation and case studies.

Acknowledgments

We thank Srinivasan Parthasarathy, Ping Zhang, Samuel Yang and Kaushik Mani for valuable discussions. We also thank the anonymous reviewers for their hard work and constructive feedback. This research was sponsored in part by the Patient-Centered Outcomes Research Institute Funding ME-2017C1-6413, the Army Research Office under cooperative agreements W911NF-17-1-0412, NSF Grant IIS1815674, and Ohio Supercomputer Center (Center, 1987). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S.Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

References

Accountability Act. 1996. Health insurance portability and accountability act of 1996. *Public law*, 104:191.

Aitziber Atutxa, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, and Olatz Perez-de Viñaspre. 2019. Interpretable deep learning to map diagnostic texts to icd-10 codes. *International Journal of Medical Informatics*, 129:49–59.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26*, pages 2787–2795.

Diane Bouchacourt and Ludovic Denoyer. 2019. Educe: Explaining model decisions through unsupervised concepts extraction. *arXiv preprint arXiv:1905.11852*.

Raluca Budiu. 2014. Memory recognition and recall in user interfaces. *Nielsen Norman Group*.

Ohio Supercomputer Center. 1987. Ohio supercomputer center.

Ismail Ilkan Ceylan, Adnan Darwiche, and Guy Van den Broeck. 2016. Open-world probabilistic databases. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Rajarshi Das, Arvind Neelakantan, David Belanger, and Andrew McCallum. 2017. Chains of reasoning over entities, relations, and text using recurrent neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 132–141.

Xiang Deng and Huan Sun. 2019. Leveraging 2-hop distant supervision from table entity pairs for relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 410–420.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Samuel G Finlayson, Paea LePendu, and Nigam H Shah. 2014. Building the graph of medicine from millions of clinical narratives. *Scientific data*, 1:140032.

Matt Gardner, Partha Talukdar, Jayant Krishnamurthy, and Tom Mitchell. 2014. Incorporating vector space similarity in random walk inference over knowledge bases. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 397–406.

Gary Gillund and Richard M Shiffrin. 1984. A retrieval model for both recognition and recall. *Psychological review*, 91(1):1.

Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 855–864.

Kelvin Guu, John Miller, and Percy Liang. 2015. Traversing knowledge graphs in vector space. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 318–327.

Bin He, Yi Guan, and Rui Dai. 2019. Classifying medical relations in clinical text via convolutional neural networks. *Artificial Intelligence in Medicine*, 93:43–49.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2020. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *International Conference on Learning Representations*.

Rudolf Kadlec, Ondrej Bajgar, and Jan Kleindienst. 2017. Knowledge base completion: Baselines strike back. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 69–74.

Michael Kahana, Marc Howard, and Sean Polyn. 2008. Associative retrieval processes in episodic memory. *Psychology*.

D. P. Kingma and J. Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*.

Joohong Lee, Sangwoo Seo, and Yong Suk Choi. 2019. Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing. *Symmetry*, 11(6):785.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Paea LePendu, Srinivasan V Iyer, Cédrick Fairon, and Nigam H Shah. 2012. Annotation analysis for testing drug safety signals using unstructured clinical notes. In *Journal of biomedical semantics*, volume 3, page S5. BioMed Central.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27*, pages 2177–2185.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

Ryan Lowe, Nissan Pow, Iulian Serban, Laurent Charlin, and Joelle Pineau. 2015. Incorporating unstructured textual knowledge sources into neural dialogue systems. In *Neural information processing systems workshop on machine learning for spoken language understanding*.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774.

Xinbo Lv, Yi Guan, Jinfeng Yang, and Jiawei Wu. 2016. Clinical relation extraction with deep learning. *International Journal of Hybrid Information Technology*, 9(7):237–248.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Arvind Neelakantan, Benjamin Roth, and Andrew McCallum. 2015. Compositional vector space models for knowledge base completion. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 156–166.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 809–816.

Peter A Nobel and Richard M Shiffrin. 2001. Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(2):384.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, et al. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 701–710.

Meng Qu, Xiang Ren, Yu Zhang, and Jiawei Han. 2018. Weakly-supervised relation extraction by pattern-enhanced embedding learning. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1257–1266.

Raymond Reiter. 1981. On closed world data bases. In *Readings in artificial intelligence*, pages 119–140. Elsevier.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, ECML PKDD'10, page 148–163.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Sunil Sahu, Ashish Anand, Krishnadev Oruganty, and Mahanandeeshwar Gattu. 2016. Relation extraction from clinical texts using domain invariant convolutional neural network. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 206–215.

Chandan Singh, W. James Murdoch, and Bin Yu. 2019. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems 26*, pages 926–934.

Linfeng Song, Yue Zhang, Daniel Gildea, Mo Yu, Zhiguo Wang, and Jinsong Su. 2019. Leveraging dependency forest for neural medical relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 208–218.

Donald P Spence and Kimberly C Owens. 1990. Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, 19(5):317–330.

Josua Stadelmaier and Sebastian Padó. 2019. Modeling paths for explainable knowledge base completion. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 147–157.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 1067–1077.

Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884.

Chang Wang, Liangliang Cao, and James Fan. 2016. Building joint spaces for relation extraction. In *IJCAI*, pages 2936–2942.

Chang Wang and James Fan. 2014. Medical relation extraction with manifold models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 828–838.

Yanjie Wang, Rainer Gemulla, and Hui Li. 2018. On multi-relational link prediction with bilinear models. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhen Wang, Xiang Yue, Soheil Moosavinasab, Yungui Huang, Simon Lin, and Huan Sun. 2019. Surfcon: Synonym discovery on privacy-aware clinical data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 1578–1586.

Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. DeepPath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 564–573.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on Freebase via relation extraction and textual evidence. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2326–2336.

Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In

Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1480–1489.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Xiang Yue, Zhen Wang, Jingong Huang, Srinivasan Parthasarathy, Soheil Moosavinasab, Yungui Huang, Simon M Lin, Wen Zhang, Ping Zhang, and Huan Sun. 2019. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics*, 36(4):1241–1251.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344.

Jie Zhao, Ziyu Guan, and Huan Sun. 2019. Riker: Mining rich keyword representations for interpretable product question answering. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 1389–1398.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212.

A Appendices

A.1 Implementation Details.

We implemented our model in Pytorch (Paszke et al., 2017) and optimized it by the Adam optimizer (Kingma and Ba, 2015). The dimension of term/node embeddings is set at 128. The number of negative triples for the relational learning is set at 100. The number of association contexts to use for assumption formation, N_c is 32. Early stopping is used when the performance in the dev set does not increase continuously for 10 epochs. We augment the relation triples for optimizing $\mathcal{L}_r$ (Eqn. 12) by adding their reverse relations for better training. We obtain DeepWalk and LINE (2nd) embeddings by OpenNE[9] and word2vec embeddings by doing SVD decomposition over the shifted PPMI co-occurrence matrix (Levy and Goldberg, 2014). Code, dataset and more implementation details are available online[10].

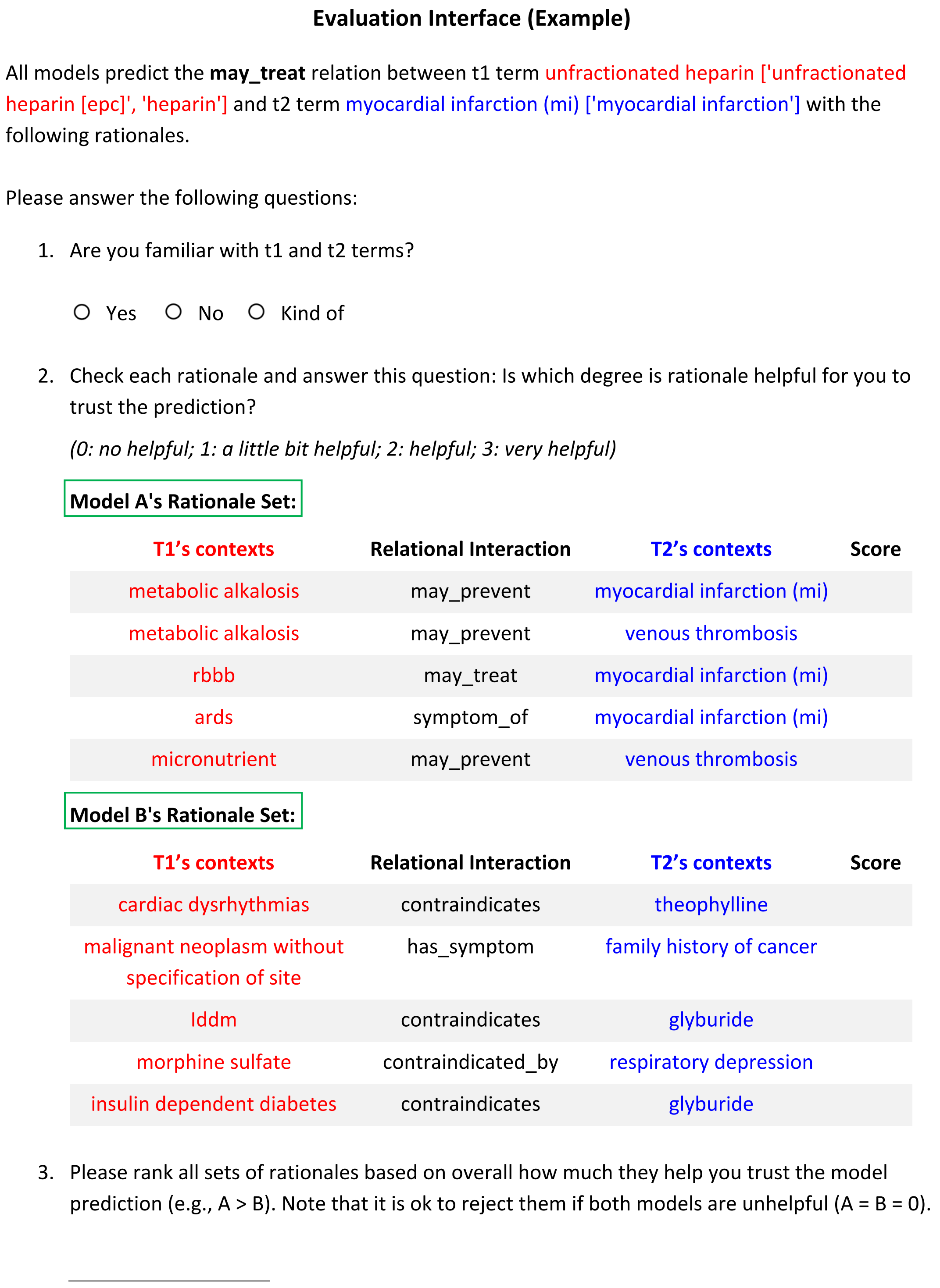
A.2 Training Algorithm

Algorithm 1 CogStage Training Algorithm

```
INPUT: Corpus Statistics $\mathcal{G}$, Gold Triples $\mathcal{P}$, Binary Relation Data $\{(h_k, t_k), y_k\}_{k=1}^{M}$
OUTPUT: Model parameters
repeat
    Sample $\{e_i\}_{i=1}^{b_1}$ with gold contexts from $\mathcal{G}$
    for $i \leftarrow 1 : b_1$ do
        Calculate $p(e_j|e_i)$ and $\hat{p}(e_j|e_i)$
        Optimize $\mathcal{L}_n$ by Eqn. 3
    Sample $\{(h_i, r_i, t_i)\}_{i=1}^{b_2}$ from $\mathcal{P}$
    for $i \leftarrow 1 : b_2$ do
        Generate $N_n$ corrupted triples
        Optimize $\mathcal{L}_r$ by Eqn. 12
    Sample $\{(h_i, t_i), y_i\}_{i=1}^{b_3}$
    for $i \leftarrow 1 : b_3$ do
        Calculate $p(e_j|h_i)$ and $p(e_j|t_i)$
        Get contexts $\{a_h^m\}_{m=1}^{N_c}$ and $\{a_t^n\}_{n=1}^{N_c}$
        Optimize $\mathcal{L}_p$ by Eqn. 14
until Convergence
```

[9] https://github.com/thunlp/OpenNE

[10] https://github.com/zhenwang9102/X-MedRELA



Evaluation Interface (Example)

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.

Please answer the following questions:

1. Are you familiar with t1 and t2 terms?

   ○ Yes ○ No ○ Kind of

2. Check each rationale and answer this question: Is which degree is 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	
Iddm	contraindicates	glyburide	
morphine sulfate	contraindicated_by	respiratory depression	
insulin dependent diabetes	contraindicates	glyburide	

3. Please 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).

   ____

Figure 4: Evaluation interface for expert evaluation.

Relations	UMLS Relations
May_treat	may_treat
May_prevent	may_prevent
Contraindicates	has_contraindicated_drug
Causes	cause_of; induces; causative_agent_of
Symptom of	disease_has_finding; disease_may_have_finding; has_associated_finding; has_manifestation; associated_condition_of; defining_characteristic_of

Table 5: Relations in our dataset and their mapped UMLS semantic relations. (UMLS relation "Treats" does not exist in our dataset and hence is not mapped with the "May_treat" relation.)