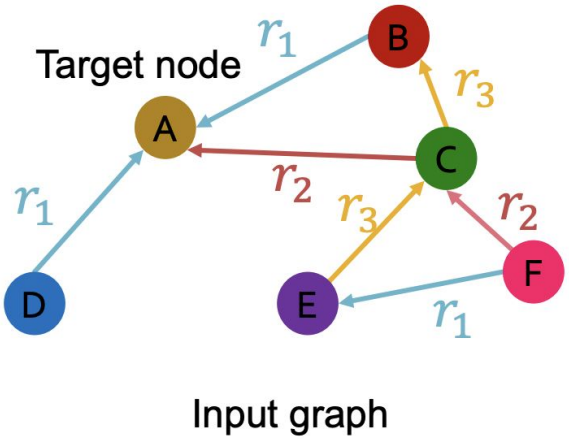
Глубинное обучение в анализе графовых данных

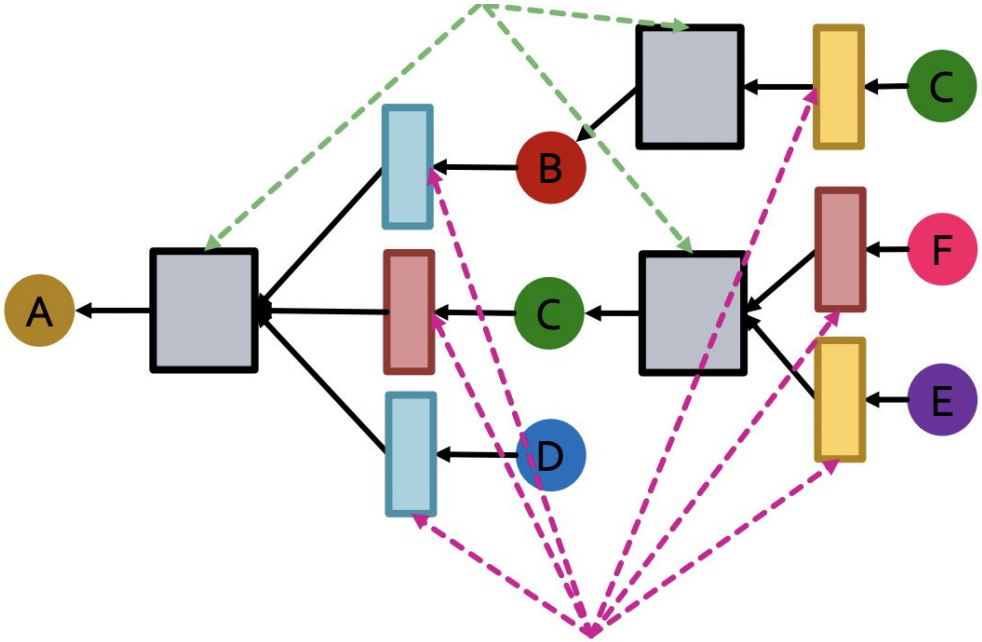
9. Использование эмбеддингов в графах знаний

Высшая Школа Экономики, 2022-2023

в предыдущих сериях…

Вычислительный граф в графе знаний





TransE алгоритм обучения

Algorithm 1 Learning TransE

input Training set $S = \{(h, \ell, t)\}$, entities and rel. sets E and L, margin γ, embeddings dim. k.

1: **initialize** $\boldsymbol{\ell} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each $\ell \in L$
2: $\boldsymbol{\ell} \leftarrow \boldsymbol{\ell} / \|\boldsymbol{\ell}\|$ for each $\ell \in L$
3: $\mathbf{e} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity $e \in E$

Entities and relations are initialized uniformly, and normalized

4: **loop**
5: $\mathbf{e} \leftarrow \mathbf{e} / \|\mathbf{e}\|$ for each entity $e \in E$
6: $S_{batch} \leftarrow \text{sample}(S, b)$ // sample a minibatch of size b
7: $T_{batch} \leftarrow \emptyset$ // initialize the set of pairs of triplets
8: **for** $(h, \ell, t) \in S_{batch}$ **do**
9: $(h', \ell, t') \leftarrow \text{sample}(S'_{(h,\ell,t)})$ // sample a corrupted triplet

Negative sampling with triplet that does not appear in the KG

10: $T_{batch} \leftarrow T_{batch} \cup \{((h, \ell, t), (h', \ell, t'))\}$
11: **end for**
12: Update embeddings w.r.t.

$$\sum_{((h,\ell,t),(h',\ell,t')) \in T_{batch}} \nabla \left[\gamma + d(\boldsymbol{h} + \boldsymbol{\ell}, \boldsymbol{t}) - d(\boldsymbol{h'} + \boldsymbol{\ell}, \boldsymbol{t'})\right]_+$$

positive sample

negative sample

d represents distance (negative of score)

13: **end loop**

Contrastive loss: favors lower distance (or higher score) for valid triplets, high distance (or lower score) for corrupted ones

Шаблоны отношений

- (анти)Симметрия : $r(h,t) \Rightarrow r(t,h)$ $(r(h,t) \Rightarrow \neg r(t,h))$ $\forall h,t$
- Инверсия : $r_2(h,t) \Rightarrow r_1(t,h)$
- Транзитивность : $r_1(x,y) \wedge r_2(y,z) \Rightarrow r_3(x,z)$ $\forall x,y,z$
- 1 to N : $r(h,t_1), r(h,t_2), \ldots, r(h,t_n)$ are all True

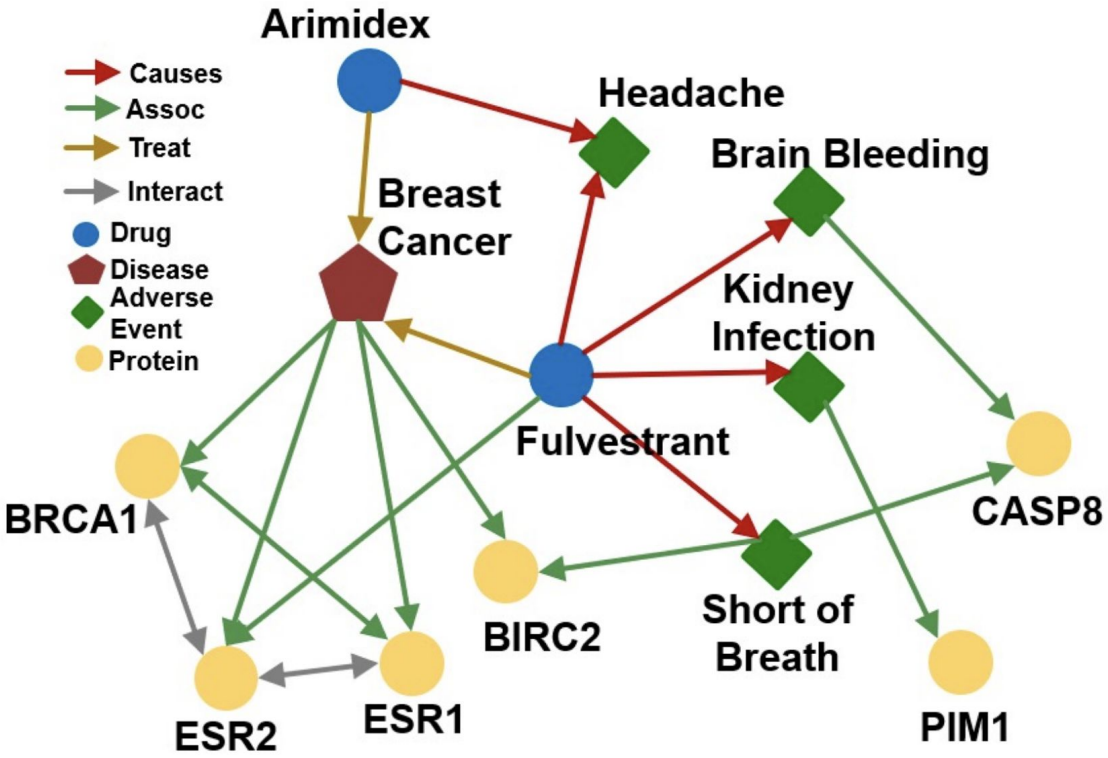
“Запросы” в графах знаний

Мультизапросы

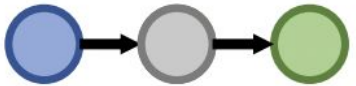
Можно ли с помощью графов знаний работать не только с тройками, но с длинными цепочками отношений?

Да можно.

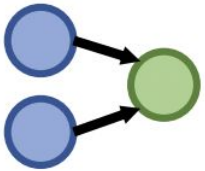
Типы запросов



One-hop Queries



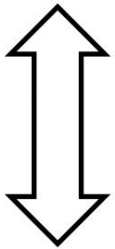
Path Queries



Conjunctive Queries

One-hop

- **KG completion:** Is link (h, r, t) in the KG?



- **One-hop query:** Is t an answer to query (h, r)?

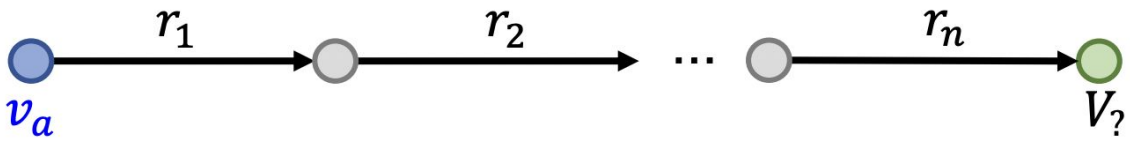
Path

- An n-hop path query q can be represented by

$$q = (v_a, (r_1, \ldots, r_n))$$

  - v_a is an "anchor" entity,
  - Let answers to q in graph G be denoted by $[\![q]\!]_G$.

Query Plan of q:

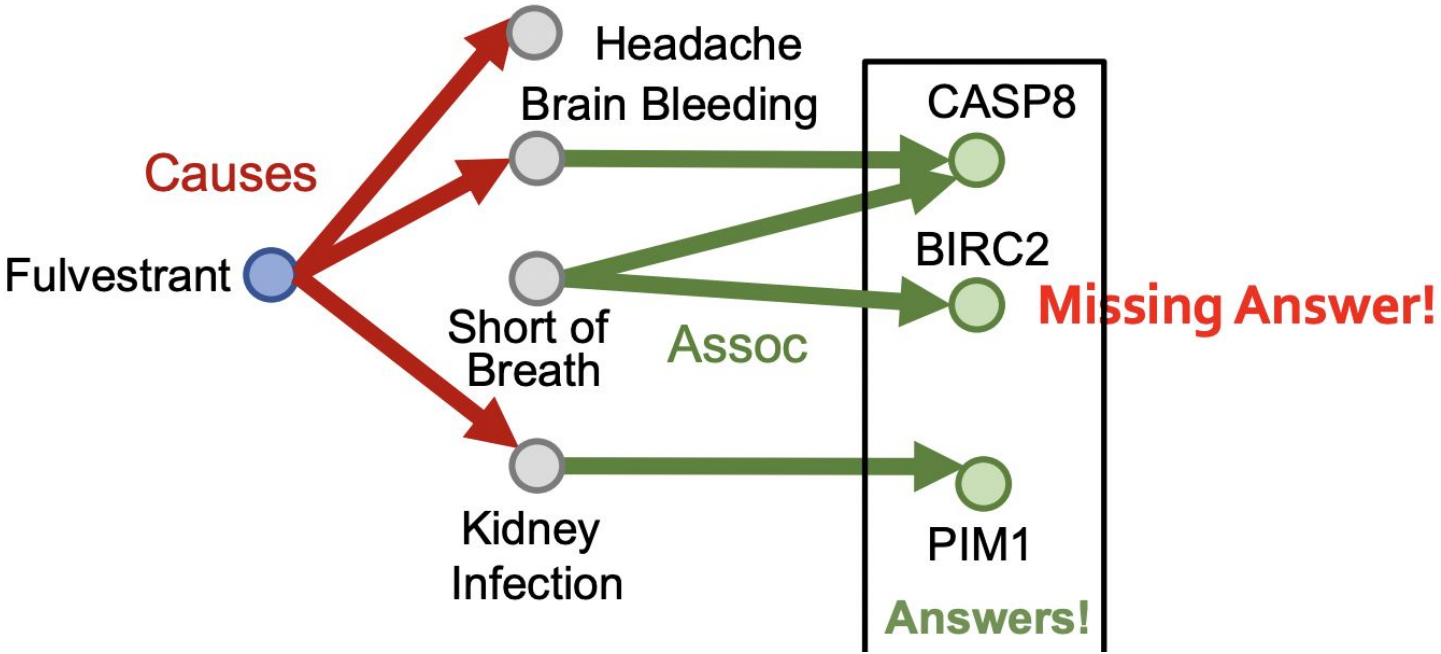


Как отвечать

Интуитивно - обход

Проблема - много пропущенных данных

Пример проблемы пропущенных

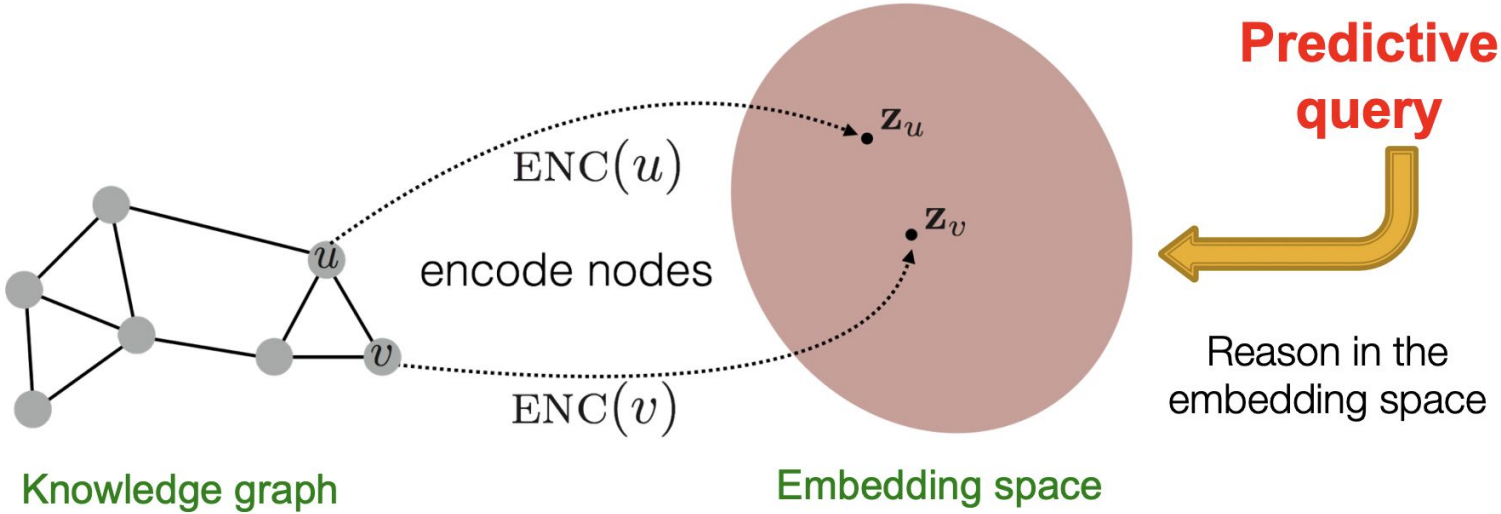


Использование восстановления

Использовать подход восстановления графа знаний не очень помогает - граф становится плотным, сложность обхода становится экспоненциальной

Что делать

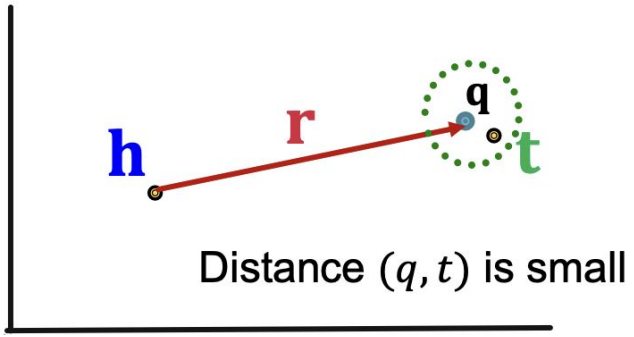
Вспомним об эмбеддингах

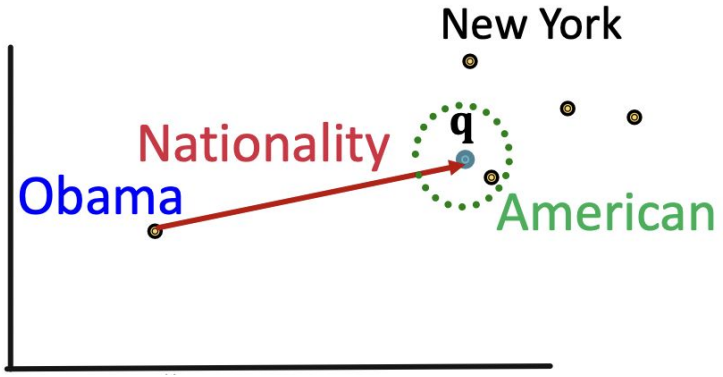


Будем использовать TransE

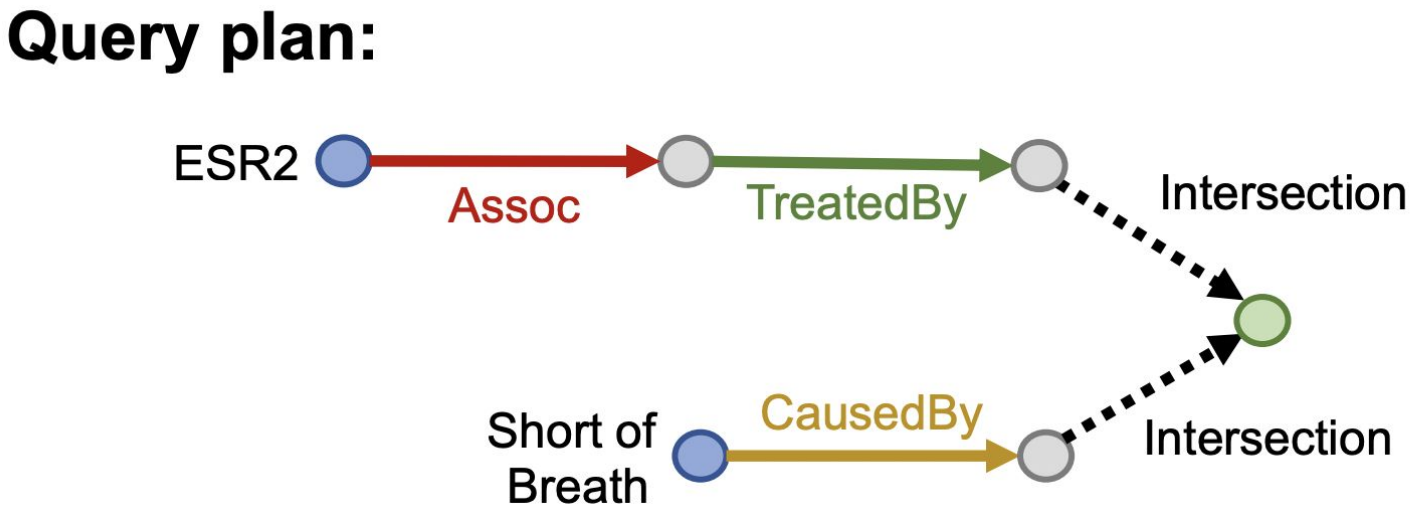
- Another way to interpret this is that:
  - **Query embedding**: $\mathbf{q} = \mathbf{h} + \mathbf{r}$
  - Goal: **query embedding** $\mathbf{q}$ is **close** to the **answer embedding** $\mathbf{t}$

$$f_q(t) = -\|\mathbf{q} - \mathbf{t}\|$$

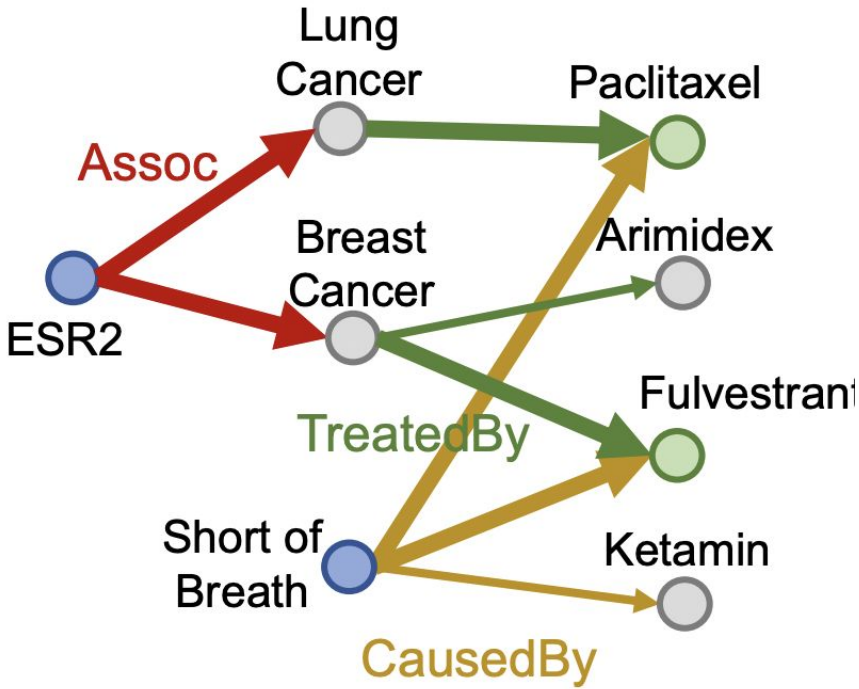




Conjuctive

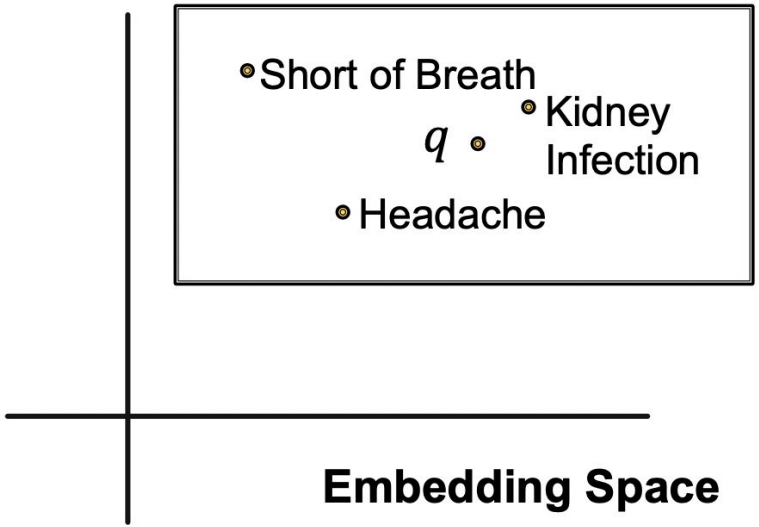

Query plan:
ESR2
Assoc
TreatedBy
Intersection
Short of Breath
CausedBy
Intersection

Пример запроса



Box embeddings

$$\mathbf{q} = (Center(q), Offset(q))$$



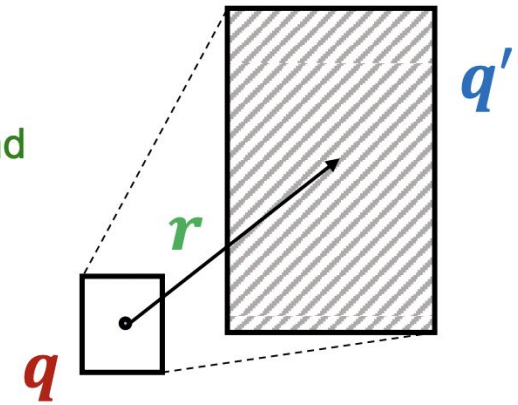
For example, we can embed the adverse events of Fulvestrant with **a box that enclose all the answer entities**.

Оператор проекции

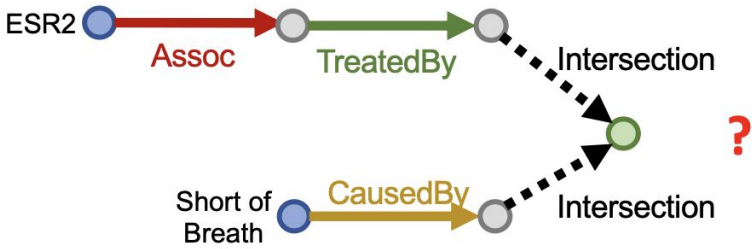
- $\mathcal{P} : \text{Box} \times \text{Relation} \rightarrow \text{Box}$

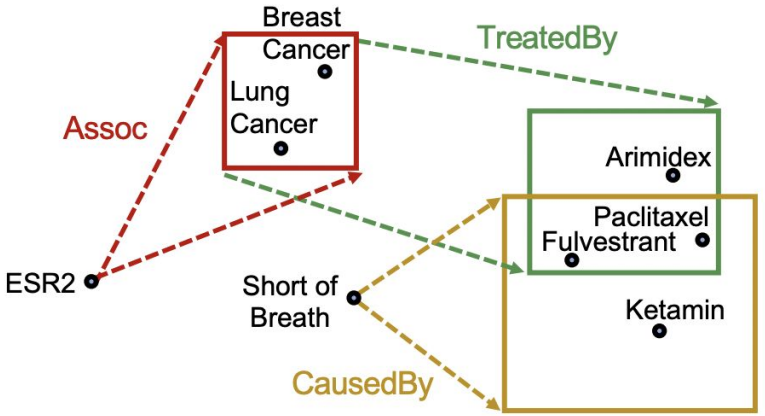
$$Cen(q') = Cen(q) + Cen(r)$$
$$Off(q') = Off(q) + Off(r)$$

"×" (cross) means the projection operator is a **relation** from any box and **relation** to a new box



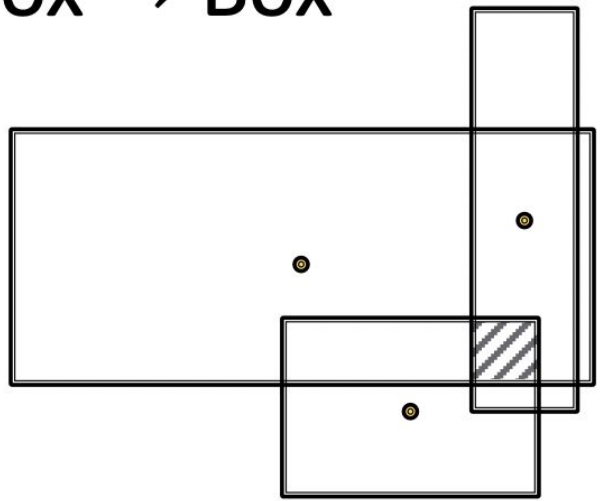
Пример работы





Оператор пересечения

- $\mathcal{I} : \text{Box} \times \cdots \times \text{Box} \rightarrow \text{Box}$



Оператор пересечения

- $\mathcal{I} : \text{Box} \times \cdots \times \text{Box} \rightarrow \text{Box}$

Hadamard product
(element-wise product)

$$Cen(q_{inter}) = \sum_i \boldsymbol{w}_i \odot Cen(q_i)$$

$$\boldsymbol{w}_i = \frac{\exp(f_{cen}(Cen(q_i)))}{\sum_j \exp(f_{cen}(Cen(q_j)))}$$

$$Cen(q_i) \in \mathbb{R}^d$$
$$\boldsymbol{w}_i \in \mathbb{R}^d$$

Оператор пересечения

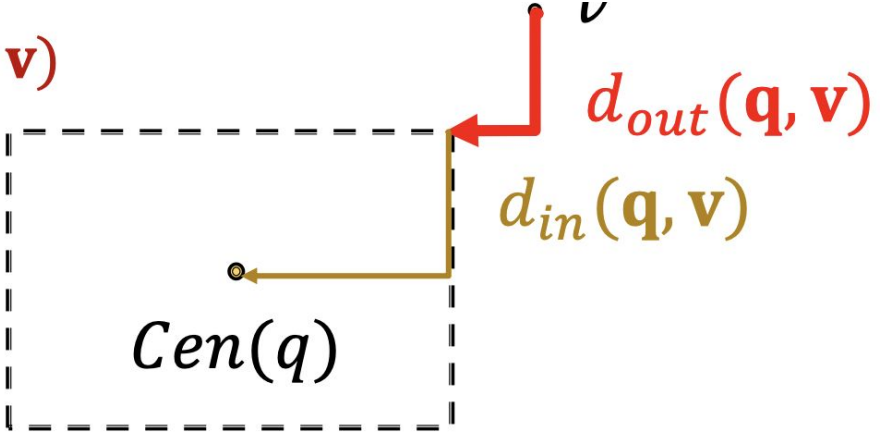
- $\mathcal{I}$: Box $\times \cdots \times$ Box $\rightarrow$ Box

$$Off(q_{inter}) = \min\big(Off(q_1), \ldots, Off(q_n)\big) \odot \sigma(f_{off}(Off(q_1), \ldots, Off(q_n)))$$

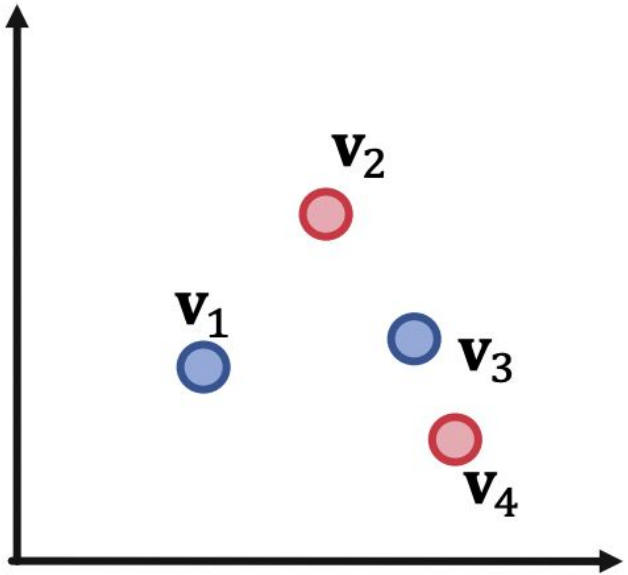
guarantees shrinking

$$d_{box}(\mathbf{q}, \mathbf{v}) = d_{out}(\mathbf{q}, \mathbf{v}) + \alpha \cdot d_{in}(\mathbf{q}, \mathbf{v})$$

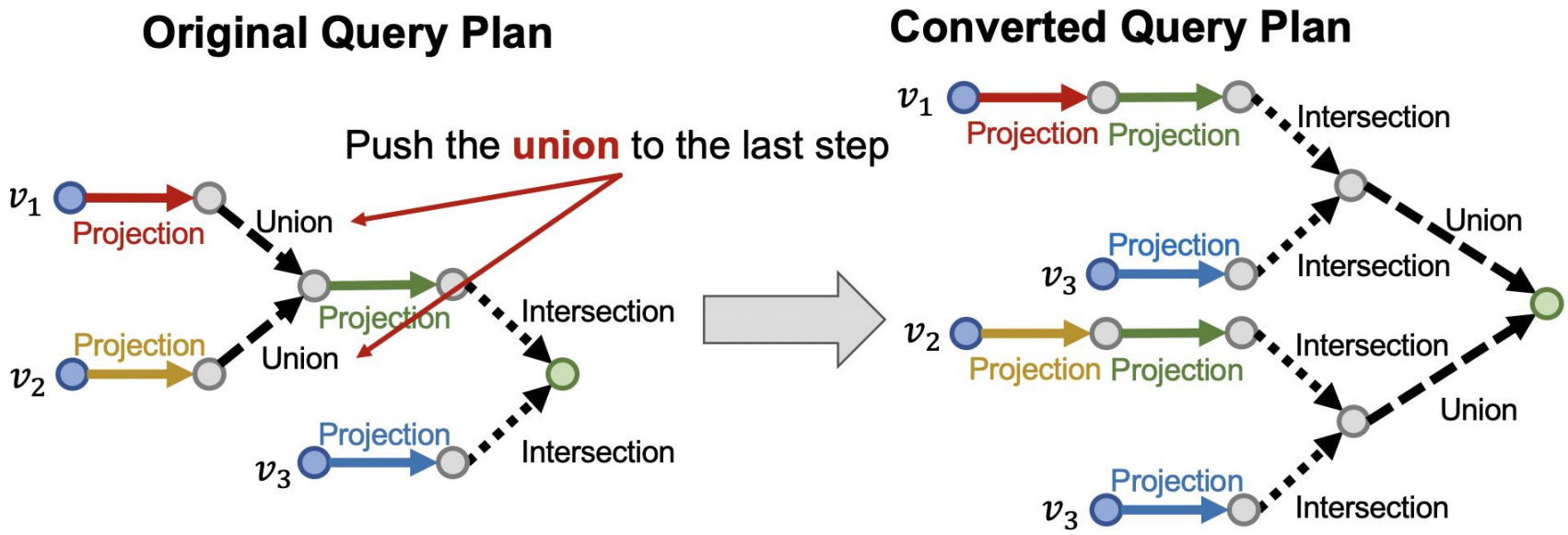
- $f_q(v) = -d_{box}(\mathbf{q}, \mathbf{v})$



Объединение



Объединение



- **Distance** between entity embedding and a DNF $q = q_1 \vee q_2 \vee \cdots \vee q_m$ is defined as:

$$d_{box}(\mathbf{q}, \mathbf{v}) = min(d_{box}(\mathbf{q}_1, \mathbf{v}), \ldots, d_{box}(\mathbf{q}_m, \mathbf{v}))$$

  - **The process of embedding any AND-OR query q**
    1. Transform q to **equivalent DNF** $q_1 \vee \cdots \vee q_m$
    2. **Embed** q_1 to q_m
    3. Calculate the (box) distance $d_{box}(\mathbf{q}_i, \mathbf{v})$
    4. Take the **minimum** of all distance
    5. **The final score** $f_q(v) = -d_{box}(\mathbf{q}, \mathbf{v})$

Обучение

- **Training**:
  1. Randomly sample a query q from the training graph G_{train}, answer $v \in [\![q]\!]_{G_{train}}$, and a negative sample $v' \notin [\![q]\!]_{G_{train}}$.
     - Negative sample: Entity of same type as v but not answer.
  2. Embed the query $\mathbf{q}$.
  3. Calculate the score $f_q(v)$ and $f_q(v')$.
  4. Optimize the loss ℓ to maximize $f_q(v)$ while minimize $f_q(v')$:

$$\ell = -\log \sigma\left(f_q(v)\right) - \log(1 - \sigma(f_q(v')))$$