# New Embedded Representations and Evaluation Protocols for Inferring Transitive Relations

Sandeep Subramanian Soumen Chakrabarti

## Point embeddings for words and entities

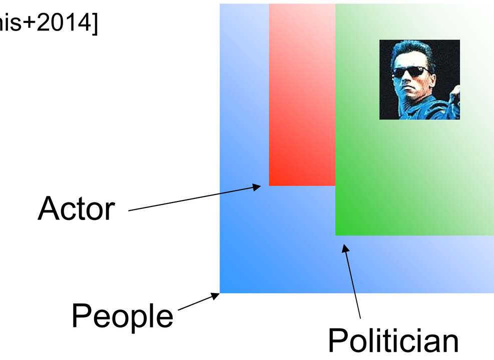
- Each word *w* has focus embedding $\boldsymbol{u}_w$ and context embedding $\boldsymbol{v}_w$
- Word *c* (not) in context of word *f* ⇒ want large (small) $\boldsymbol{u}_f \cdot \boldsymbol{v}_c$ … GloVe, word2vec
- Entities treated like words over multi-token spans
  - Young Albert took violin lessons from …
  - Albert linked to entity ID `m.0jcx`
  - Embeddings of `m.0jcx` (but not other Alberts) and *relativity* become similar

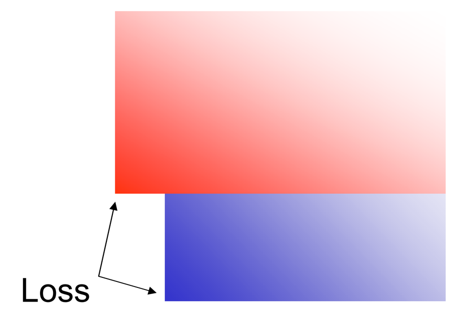
## How to represent types?

- Points, like entities
  - Entity instance-of type and type subtype-of type
  - Antisymmetry? Transitivity?
- Subspace of $\mathbb{R}^D$ [Jameel+2016]
  - Type *t* represented by *D*+1 points $\boldsymbol{t}^0, \ldots, \boldsymbol{t}^D$
  - Minimize nuclear norm of $[\boldsymbol{t}^1 - \boldsymbol{t}^0, \ldots, \boldsymbol{t}^D - \boldsymbol{t}_0]$
  - Also $\min_{\{\boldsymbol{e}, \vec{\boldsymbol{t}}\}} \sum_t \sum_{e \in t} \min_{(\lambda_0, \ldots, \lambda_D) \in \Delta^{D+1}} \left\| \boldsymbol{e} - \sum_d \lambda_d \boldsymbol{t}^d \right\|_2^2$
  - Degeneracies; also does not handle transitivity

## How to represent types?

- Each type is a Gaussian density [Vilnis+2014]
  - Asymmetry via KL (closed form)
  - No antisymmetry or transitivity
  - Uncertainty or extended region?
- Dominance [Vendrov+2015] (Order Embeddings or OE)
  - $x \subseteq y$ represented by $\boldsymbol{u}_x \geq \boldsymbol{u}_y$ (elementwise)




## Order embedding (OE) loss

- "*x* less preferred than *y*": $x \prec y$
- OE defines $\ell(x,y) = \|\max\{\boldsymbol{0}, \boldsymbol{u}_y - \boldsymbol{u}_x\}\|_2^2$
- If $x \prec y$ ($x \not\prec y$) then we want small (large) loss
- Overall loss is sum of

$$\mathcal{L}_+ = \sum_{x \prec y} \ell_+(x,y) = \sum_{x \prec y} \ell(x,y)$$

$$\mathcal{L}_- = \sum_{x \not\prec y} \ell_-(x,y) = \sum_{x \not\prec y} \max\{0, \alpha - \ell(x,y)\},$$

Does not recognize that violation in one dimension enough

## Improved OE loss

If $x \prec y$, then, for *all* dimensions, we want $u_{x,d} \geq u_{y,d}$

$$\ell_+(x,y) = \max_{d \in [1,D]} [u_{y,d} - u_{x,d}]_+,$$

$$\ell_-(x,y) = \min_{d \in [1,D]} [u_{x,d} - u_{y,d}]_+,$$

If $x \not\prec y$, then, for *some* dimension, we want $u_{x,d} \leq u_{y,d}$

Normalize per-instance loss using sigmoid:

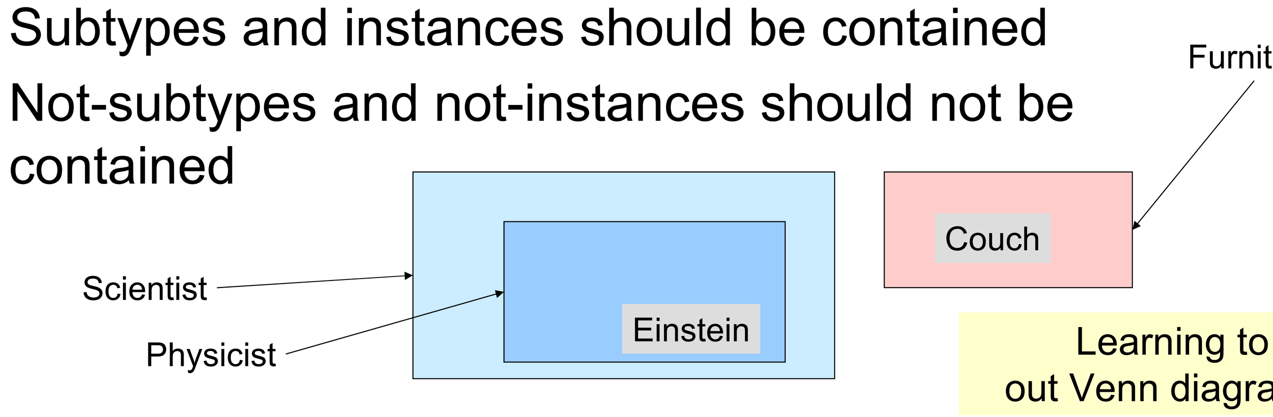
$$\ell_+(x,y) = \sigma\left(\psi \max_{d \in [1,D]} [u_{y,d} + \underset{\text{margin}}{\Delta} - u_{x,d}]_+\right) - \frac{1}{2}$$

stiffness

Far from convex, but …

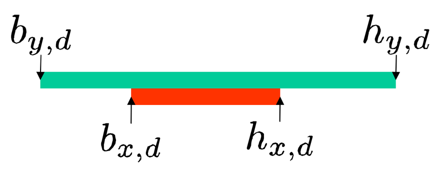

## Rectangle embeddings

- OE models each type as infinite dominating cone
  - All types have same measure; no negative correlation
    - X is-a furniture ⇒¬(X is-a scientist)
- Fix: type is a range in each dimension
  - Subtypes and instances should be contained
  - Not-subtypes and not-instances should not be contained




Learning to lay out Venn diagrams

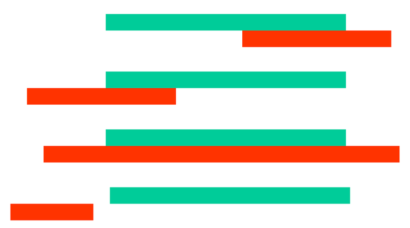
## Rectangle losses




If $x \prec y$, then, for *all* dimensions, interval of *x* must be contained in interval of *y*

$$\delta_+(x,y,d) = \max_{d \in [D]} \max\{[b_{y,d} - b_{x,d}]_+, [h_{x,d} - h_{y,d}]_+\}$$

$$\delta_-(x,y,d) = \min_{d \in [D]} \max\{[b_{x,d} - b_{y,d}]_+, [h_{y,d} - h_{x,d}]_+\}$$



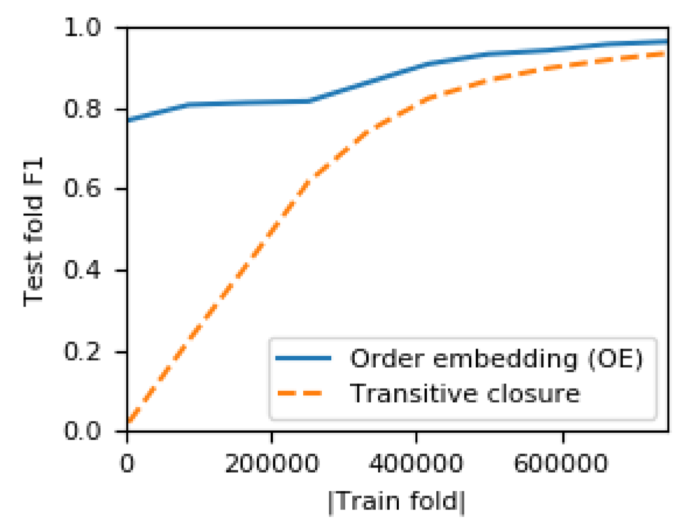
If $x \not\prec y$, then, for *some* dimension, interval of *x* must stray outside the interval of *y*

## "Leaky" evaluation protocol

- *T* = available partial order, clo(*T*) transitive closure
  - WordNet has |*T*|=82115, |clo(*T*)|=838073
- Sample positive eval fold $E_+ \subset \mathrm{clo}(T)$ …4000
- Sample positive dev fold $D_+ \subset \mathrm{clo}(T) \setminus E_+$ …4000
- Rest learn fold $L_+ = \mathrm{clo}(T) \setminus (E_+ \cup D_+)$ …830k(!)
- Plus negative samples $L_-$, $D_-$, $E_-$, typically 1:1
  - Local closed world assumption
- With 830k of 838k known, what's left to discover?

## OE with leaky protocol

- TC = transitive closure on $L_+$ gave 0/1 acc = 88.6%
  - Gaussian embedding worse, 86.6%; OE at 90.6%
- OE was actually undersold, gains more when training fold shrunk
- Still rewarded for computing transitive closure




## Sanitized evaluation protocol

- Give no credit for computing transitive closure
- Sample $L_+ \subset \mathrm{clo}(T)$
- Perturb each $(x,y) \in L_+$ to negative sample $(x', y')$
- Discard if $(x', y') \in \mathrm{clo}(T)$
- Sample pos dev $D_+ \subset (\mathrm{clo}(T) \setminus \mathrm{clo}(L_+))$
  - Discard $(x,y)$ from $D_+$ if *x* or *y* is o.o.v. from $L_+ \cup L_-$
- Sample pos eval $E_+ \subset \mathrm{clo}(T) \setminus (\mathrm{clo}(L_+) \cup \mathrm{clo}(D_+))$
- Perturb and discard as before

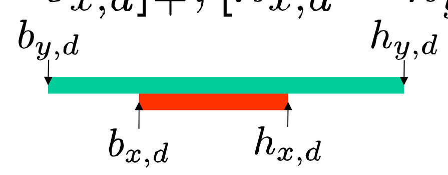
## Results with leaky and sanitized protocols

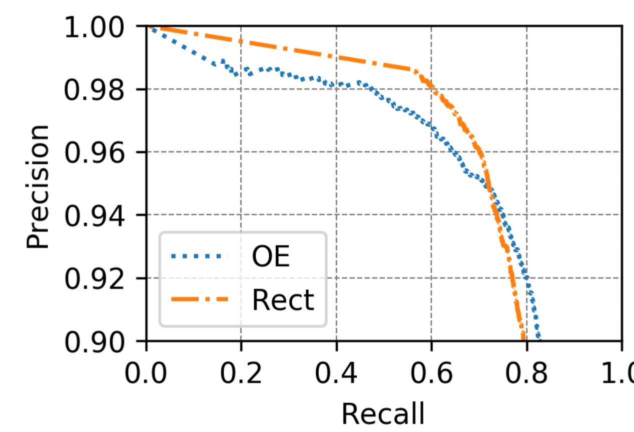
| | Leaky protocol | | | Sanitized protocol | | |
|---|---|---|---|---|---|---|
| | OE | σOE | Rect | OE | σOE | Rect |
| Acc | 0.922 | 0.921 | 0.926 | 0.574 | 0.742 | 0.767 |
| AP | 1 | 1 | 1 | 0.977 | 0.969 | 0.986 |
| P | 0.994 | 0.915 | 0.973 | 0.987 | 0.925 | 0.983 |
| R | 0.850 | 0.929 | 0.877 | 0.151 | 0.527 | 0.544 |
| F1 | 0.916 | 0.922 | 0.923 | 0.262 | 0.671 | 0.722 |

- Sanitized protocol destroys recall and F1 of OE
- Rectangle is better than σOE is (much) better than OE
- σOE and rectangle better even with leaky protocol

## Precision vs. recall for ranking

- Sort (*x*,*y*) by increasing containment violation $\max_{d \in [D]} \max\{[b_{y,d} - b_{x,d}]_+, [h_{x,d} - h_{y,d}]_+\}$
- $x \prec y$ = "relevant"
- $x \not\prec y$ irrelevant
- AUC, MAP, NDCG etc.
- Rectangle has higher precision at lower recall than OE




gitlab.com/soumen.chakrabarti/rectangle