Pointwise HSIC: A Linear-Time Kernelized Co-occurrence Norm for Sparse Linguistic Expressions [EMNLP2018]

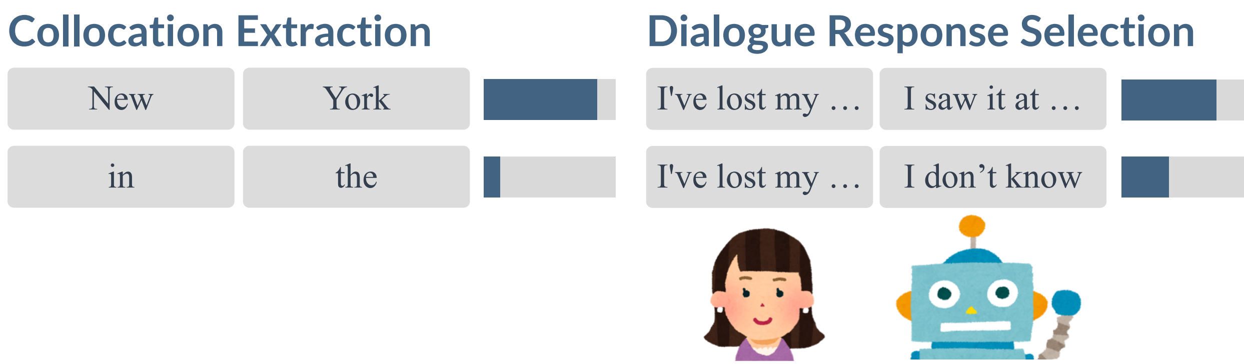
Sho Yokoi (Tohoku U./AIP)*, Sosuke Kobayashi (PFN), Kenji Fukumizu (ISM), Jun Suzuki *, Kentaro. Inui *

1809.00800 https://bit.ly/2SfpZmv github.com/cl-tohoku/phsic pip install phsic-cli

Computing Co-occurrence Strength in NLP

observed: paired data $\mathcal{D} = \{(x_i, y_i)\} \sim \mathbf{P}_{XY}$

task: compute "co-occurrence/association/relation strength" of $(x, y) \in \mathcal{X} \times \mathcal{Y}$



Pointwise HSIC: A Linear-Time Kernelized Co-occurrence Norm for Sparse Linguistic Expressions [EMNLP2018]

Sho Yokoi (Tohoku U./AIP)*, Sosuke Kobayashi (PFN), Kenji Fukumizu (ISM), Jun Suzuki *, Kentaro. Inui *

1809.00800 https://bit.ly/2SfpZmv github.com/cl-tohoku/phsic pip install phsic-cli

Computing Co-occurrence Strength in NLP

observed: paired data $\mathcal{D} = \{(x_i, y_i)\} \sim \mathbf{P}_{XY}$

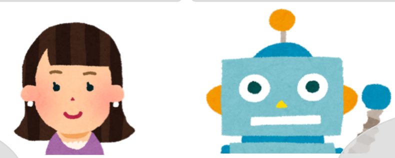
task: compute "co-occurrence/association/relation strength" of $(x, y) \in \mathcal{X} \times \mathcal{Y}$

Collocation Extraction

New	York
in	the

Dialogue Response Selection

I've lost my …	I saw it at …
I've lost my …	I don't know



De facto Measure: Pointwise Mutual Information

$$\mathrm{PMI}(x, y) = \log \frac{\mathbf{P}_{XY}(x, y)}{\mathbf{P}_X(x)\mathbf{P}_Y(y)}$$

co-occurrence **actual**

co-occurrence **by chance**

$$\widehat{\mathrm{PMI}}(x, y) = \log \frac{n \cdot \#(x, y)}{\#(x, \cdot)\#(\cdot, y)}$$

just **counting**

- ✔ easy to learn
- ✘ inappropriate to sparse data

$$\widehat{\mathrm{PMI}}(x, y) = \log \frac{\widehat{\mathbf{P}}_{\mathrm{RNN}}(y|x)}{\widehat{\mathbf{P}}_{\mathrm{RNN}}(y)}$$

using **RNNs**

- ✘ tough to learn
- ✔ applicable to sparse data

Proposed Measure for Computing Co-occurrence Strength: Pointwise HSIC

Dependence of X and Y	**Co-occurrence** of x and y

Mutual Information

$$\mathrm{MI}(X,Y) = \mathrm{KL}[\mathbf{P}_{XY} \| \mathbf{P}_X \mathbf{P}_Y]$$

$$= \mathop{\mathbf{E}}_{(x,y)}\left[\boxed{\log \frac{\mathbf{P}_{XY}(x,y)}{\mathbf{P}_X(x)\mathbf{P}_Y(y)}}\right]$$

contribute

Pointwise Mutual Information

$$\mathrm{PMI}(x,y;X,Y)$$

$$= \boxed{\log \frac{\mathbf{P}_{XY}(x,y)}{\mathbf{P}_X(x)\mathbf{P}_Y(y)}}$$

Kernels

$k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

HSIC [Gretton+'05]

$$\mathrm{HSIC}(X,Y;\boxed{k},\boxed{\ell}) = \mathrm{MMD}^2_{k,\ell}[\mathbf{P}_{XY}, \mathbf{P}_X \mathbf{P}_Y]$$

$$= \mathop{\mathbf{E}}_{(x,y)}\left[\boxed{(\phi(x) - m_X)^\top C_{XY} (\psi(y) - m_Y)}\right]$$

$$= \mathop{\mathbf{E}}_{(x,y)}\left[\boxed{\mathop{\mathbf{E}}_{(x',y')}[\tilde{k}(x,x')\tilde{\ell}(y,y')]}\right]$$

contribute

Pointwise HSIC

$$\mathrm{PHSIC}(x,y;X,Y,\boxed{k},\boxed{\ell})$$

$$= \boxed{(\phi(x) - m_X)^\top C_{XY} (\psi(y) - m_Y)}$$

$$= \boxed{\mathop{\mathbf{E}}_{(x',y')}[\tilde{k}(x,x')\tilde{\ell}(y,y')]}$$

- **Requires very short learning time**
  1000 times faster than RNN-based PMI
- **Applicable to sparse objects**
  PHSIC allows various and **available similarity metrics** to be plugged in as kernels

Pointwise HSIC: A Linear-Time Kernelized Co-occurrence Norm for Sparse Linguistic Expressions (EMNLP2018)

Sho Yokoi (Tohoku U./AIP)*, **Sosuke Kobayashi** (PFN), **Kenji Fukumizu** (ISM), **Jun Suzuki** *, **Kentaro. Inui** *

1809.00800 https://bit.ly/2SfpZmv github.com/cl-tohoku/phsic pip install phsic-cli

input: paired data $\mathcal{D} = \{(x_i, y_i)\} \sim \mathbf{P}_{XY}$ **goal:** compute "co-occurrence strength" of $(x, y) \in \mathcal{X} \times \mathcal{Y}$

Collocation Extraction

New	York
in	the

$$\widehat{\mathrm{PMI}}(x, y) = \log \frac{n \cdot \#(x, y)}{\#(x, \cdot)\#(\cdot, y)}$$

- easy to learn
- inappropriate to sparse data

Dialogue Response Selection

I've lost my ...	I saw it at ...
I've lost my ...	I'm so sleepy

$$\widehat{\mathrm{PMI}}(x, y) = \log \frac{\widehat{\mathbf{P}}_{\mathrm{RNN}}(y|x)}{\widehat{\mathbf{P}}_{\mathrm{RNN}}(y)}$$

- tough to learn
- applicable to sparse data

Proposed Measure (PHSIC)

$$\widehat{\mathrm{PHSIC}}(x, y; k, \ell) = \frac{1}{n} \sum_i \widehat{\widetilde{k}}(x, x_i) \widehat{\widetilde{\ell}}(y, y_i)$$

- easy to learn
- applicable to sparse data