Mapping Unparalleled Clinical Professional and Consumer Languages with Embedding Alignment

Wei-Hung Weng, MD, MMSc
MIT

2018 KDD
Workshop on Machine Learning for Medicine and Healthcare
Aug 20, 2018

Massachusetts Institute of Technology
CSAIL

Motivation

- Patient-clinician communication → patient-clinician relation
- Huge information gap between professionals and consumers
- Clinical documents
- Discharge summaries
  - `On floor pt found to be hypoxic on O2 4LNC O2 sats 85 %, CXR c/w pulm edema, she was given 40mg IV x 2, nebs, and put out 1.5 L UOP, she was also put on a NRB with improvement in O2 Sats to 95 %`

Problem

- Affect clinical decision making [Nickel 2017]
- Breast cancer / lesion / abnormal cells [Omer 2013]
- PCOS / hormone imbalance [Copp 2017]
- Result in…
  - Overtreatment
  - Overdiagnosis
  - Defensive medicine

Background

Previous studies

- Ontology / Dictionary [Zielstorff 2003, Zeng-Treitler 2007, Alfano 2015]
    - UMLS
    - Consumer health vocabulary (CHV)
- Synonym replacement
- Explanation insertion
- Pattern-based mining with Wikipedia corpus [Vydiswaran 2014]

NLP Techniques

- Unsupervised word vector representation [Mikolov 2013, Bojanowski 2017]
- Embeddings alignment [Conneau 2018, Chung 2018, Lample 2018]

Proposed Approach

- Learning embeddings
  - Word2vec / Fasttext subword
- Procrustes iterative learning [Dinu 2015, Conneau 2018, Xing 2015]
  - Orthogonality constraint on W → Orthogonal Procrustes problem (solve the constrained quadratic problem)

$$W^\star = \underset{W \in \mathbb{R}^{d \times d}}{\text{argmin}} \; \|WX - Y\|^2 \qquad \min_{\bar{W}} \|W - \bar{W}\| \;\; s.t. \;\; \bar{W}^T \bar{W} = I.$$

$$W^\star = \text{argmin}_{W \in \mathbb{R}^{d \times d}} \|WX - Y\|^2 = UV^T, \text{ where } U \Sigma V^T = \text{SVD}(YX^T)$$

- Finding anchor points
  - Identical words
- Linear mapping from anchors between source and target
- Applying mapping to all words

Proposed Approach

- Adversarial training [Goodfellow 2014, Conneau 2018]
    - Discriminator - identifying the origin of embedding

$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{x}\sum_{i=1}^{x}\log\mathbb{P}_{\theta_D}(\text{Pro}=1|Wp_i) - \frac{1}{y}\sum_{j=1}^{y}\log\mathbb{P}_{\theta_D}(\text{Pro}=0|c_j).$$

    - Generator (W) - making aligned source and target embeddings as similar as possible

$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{x}\sum_{i=1}^{x}\log\mathbb{P}_{\theta_D}(\text{Pro}=0|Wp_i) - \frac{1}{y}\sum_{j=1}^{y}\log\mathbb{P}_{\theta_D}(\text{Pro}=1|c_j)$$

- Ignoring all supervision

Dataset

- MIMIC-III [Johnson 2016]
  - 59,654 documents
- Professional language
  - History of present illness / Brief hospital course
  - 443,585 sentences
  - 26,333 unique words
- Consumer language
  - Discharge instruction / Followup instruction
  - 73,349 sentences
  - 6,752 unique words
- Ground truth
  - 100 professional-layman term pairs created by physician

epistaxis	nosebleed
tumor	mass
malignant	cancer
hallucination	confusion
vertigo	dizziness
diaphoresis	sweats
tremor	shake
icterus	jaundice
narcotic	sedative
epileptic	seizure

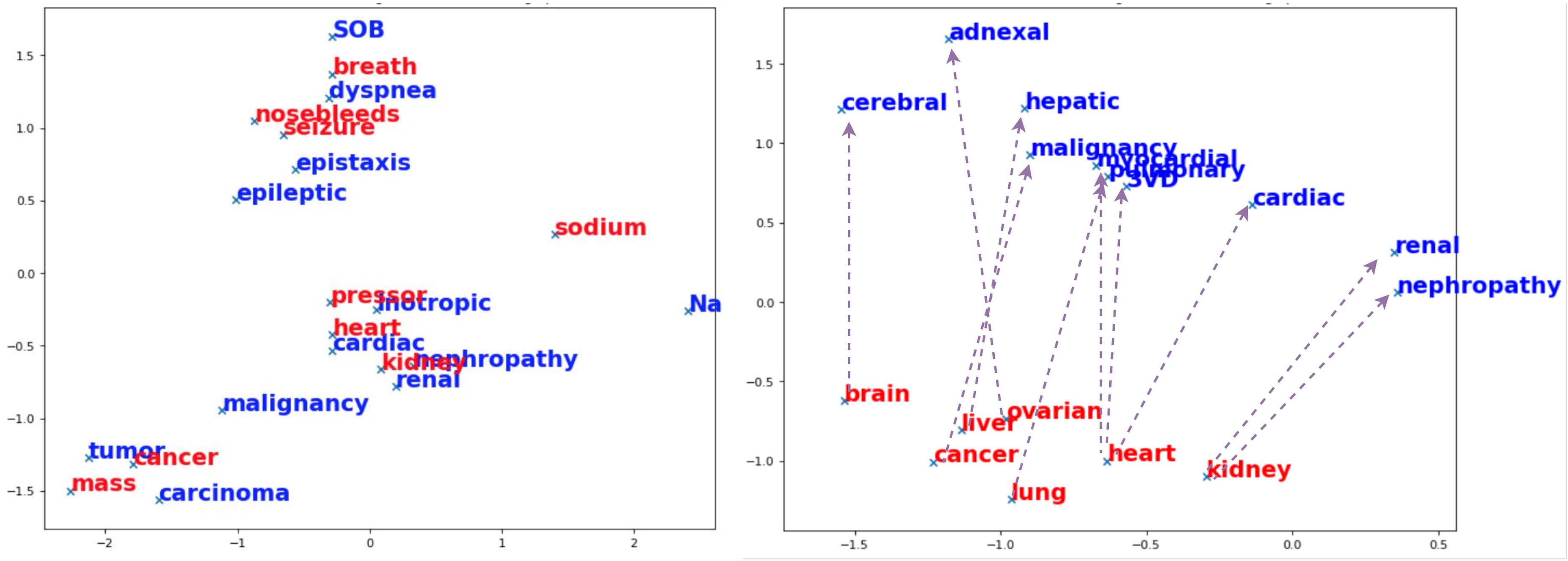
Domain-specific Corpus with Subword Embedding

Source	Target	Embedding	Window	P@1	P@5	P@10
MIMIC-P	MIMIC-C	word	3/3	0.17	0.39	0.48
MIMIC-P	MIMIC-C	word	5/5	0.19	0.42	0.54
MIMIC-P	MIMIC-C	subword	3/3	0.27	**0.57**	**0.78**
MIMIC-P	MIMIC-C	subword	5/5	**0.30**	0.55	0.68
PMC-Pubmed	MIMIC-C	word	30/3	0.26	0.40	0.44
PMC-Pubmed	MIMIC-C	word	30/5	0.18	0.39	0.44
PMC-Pubmed	MIMIC-C	subword	30/3	0.23	0.34	0.44
PMC-Pubmed	MIMIC-C	subword	30/5	0.23	0.41	0.49
PMC-Pubmed	Wikipedia	word	30/10	0.14	0.32	0.41

Clinician Qualitative Evaluation

epistaxis	cardiac	nephropathy	cholangiography	qd	tumor	hepatic	hematemesis
spontaneous	catheterization	**renal**	drug-eluting	EC	tumor	**liver**	**coffee-ground**
coffee-ground	**heart**	**kidney**	**stent**	**once/day**	**cancer**	hepatic	black
light-headed	attack	hepatitis	ureteral	QD	obstructing	Whipple	**bloody**
nosebleed	coronary	HIV	**stented**	Ramipril	resection	gastropathy	tarry
stools	cardiac	aka	circumflex	Zocor	**mass**	Pork	colored
melena	angioplasty	diastolic	bare	meq	metastatic	Mayonnaise	grounds
bleeds	myocardial	pancreatitis	**stents**	Mesalamine	cerebellum	belly	stools
bloody	cardioversion	Epo	sphincterotomy	3.125	occipital	portal	bright
black/tarry	bypass	Diabetes	metal	QHS	hepatocellular	scarring	black/tarry
10days	EP	lupus	**biliary**	162	polypectomy	Y	dark

Embeddings Visualization



Conclusion / Take Home Message

- Professional-Layman language mapping (translation)
- Embeddings alignment
- Supervised Procrustes iterative learning
- Domain-specific corpus with subword embeddings work well
- Future direction
    - Large corpus
    - Concept-level embedding
    - Ground truth
    - Stability of GAN

ckbjimmy@mit.edu

Find Neighbors

- Cross-Domain Similarity Local Scaling (CSLS) to decide mutual nearest neighbors
- Hubness problem
  - points tending to be nearest neighbors of many points in high-dimensional spaces

$$r_{\mathrm{T}}(Wx_s) = \frac{1}{K} \sum_{y_t \in \mathcal{N}_{\mathrm{T}}(Wx_s)} \cos(Wx_s, y_t),$$

$$\mathrm{CSLS}(Wx_s, y_t) = 2\cos(Wx_s, y_t) - r_{\mathrm{T}}(Wx_s) - r_{\mathrm{S}}(y_t).$$