Mapping Unparalleled Clinical Professional and Consumer Languages with Embedding Alignment

Wei-Hung Weng, Peter Szolovits Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

Massachusetts Institute of Technology | CSAIL

Abstract

Mapping and translating professional but arcane clinical jargons to consumer language is essential to improve the patient-clinician communication. We utilized the embeddings alignment method for the word mapping between unparalleled clinical professional and consumer language embeddings.

To map semantically similar words in two different word embeddings, we independently trained word embeddings on the professional and layman clinical corpora. Then, we aligned the embeddings by the Procrustes and adversarial algorithms with refinement. We evaluated the quality of the alignment through the similar words retrieval both by the model precision and humnan judging.

The Procrustes algorithm can be performant for the alignment, whereas adversarial training with refinement may find some relations between professional and consumer language embeddings.

Background

Motivation

- Patient-clinician communication
- Clinical documents
- Huge information gap between professionals and consumers
- `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 %`

Affecting Clinical Decision Making

- Breast cancer/lesion/abnormal cells [Omer 2013]
- PCOS/hormone imbalance [Copp 2017]
- Result in...
  - Overtreatment / overdiagnosis / defensive medicine

Current Solutions

- Ontology/Dictionary-based [Zielstorff 2003, Zeng-Treitler 2007, Alfano 2015]
  - UMLS /Consumer health vocabulary (CHV)
  - Synonym replacement
  - Explanation insertion
  - Problem - expert efforts to manually build the dictionary, which is hard to be generalized and scalable
- Pattern-based mining with Wikipedia corpus [Vydiswaran 2014]
  - Problem - not specific for professional and consumer languages

Proposed Approach

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

Methods

Learning Word Embeddings

- Word-level (word2vec) [Mikolov 2013]
- Subword-level (fasttext) [Bojanowski 2017]
- Different window sizes (k=3, 5)

Alignment - Procrustes Algorithm

- Constructing the synthetic mapping dictionary to learn a linear mapping matrix between the two embedding spaces

$$W^\star = \underset{W \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \|WX - Y\|^2$$

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

Alignment - Adversarial Training

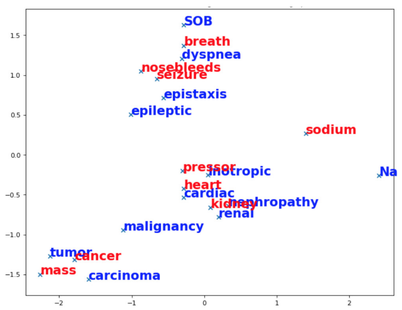
- Making the aligned embeddings indistinguishable

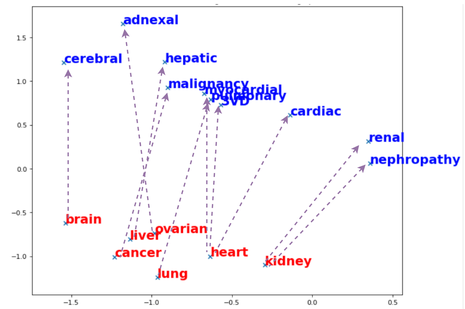
$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{x}\sum_{i=1}^{x} \log \mathbb{P}_{\theta_D}(\text{Pro} = 1|Wp_i) - \frac{1}{y}\sum_{j=1}^{y} \log \mathbb{P}_{\theta_D}(\text{Pro} = 0|c_j)$$

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

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

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





Reference

- Conneau et al. Word translation without parallel data. ICLR, 2018.
- Lample et al. Unsupervised machine translation using monolingual corpora only. ICLR 2018.
- Chung et al. Unsupervised cross-modal alignment of speech and text embedding spaces. arXiv, 2018.
- Sogaard et al. On the Limitations of Unsupervised Bilingual Dictionary Induction. arXiv, 2018.
- Omer et al. Impact of ductal carcinoma in situ terminology on patient treatment preferences. JAMA IM 2013.
- Zielstorff et al. Controlled vocabularies for consumer health. Journal of Biomedical Informatics, 2003.

Experiment

Data Source

- 59,654 free-text discharge summaries in MIMIC-III version 1.4.
- Professional corpus (443,585 sentences / 26,333 vocabularies)
  - History of present illness
  - Brief hospital course
- Consumer corpus (73,349 sentences / 6,752 vocabularies)
  - Discharge instruction
  - Followup instruction
- Source and target corpora are not parallel
- Overlapping English terms as anchors
- Ground truth
  - List of 100 professional-consumer term pairs created by the clinician

Quantitative Evaluation

- Mapping word retrieval task, compute precision
- Query the nearest k words (k=1, 5, 10) from the consumer language embedding using each professional term in the aligned professional language embedding.

Qualitative Evaluation

- Subword-level embedding > word-level embedding
  - Utilizing the character-level n-grams information
  - Capturing morphological patterns and therefore may enhance the information about word semantics
- MIMIC > PMC-Pubmed
  - Suitable to represent the clinical professional and consumer language, comparing with the general PMC-Pubmed and Wikipedia corpora.
- Limitation
  - Training sample size
  - Unsimilar shapes of distribution of embeddings

Conclusion

We demonstrate the capability of embeddings alignment for mapping unparalleled clinical professional and consumer languages in word-level, without the knowledge and supervision from biomedical ontologies and dictionaries, and just use the minimal supervision using the identical strings across corpora. We found that the Procrustes algorithm with anchors approach with the subword-level word embeddings trained on clinical narrative texts, rather than larger general corpus, outperformed the other combinations.

The aligned embeddings learned from the adversarial training approach reveal the relation between professional and consumer anatomy-related terms. Further investigations include exploring larger datasets and extending word-level to concept-level embeddings.