Learning Low-Dimensional Representations of Medical Concepts

Youngduck Choi, Yi-I Chiu, David Sontag

Courant Institute of Mathematical Sciences, New York University

February 23, 2016

Disclosure

The authors have no relationships with commercial interests.

Low-Dimensional Representations of Medical Concepts

- The central objects of study are low dimensional representations of medical concepts, also referred to as medical concept embeddings.

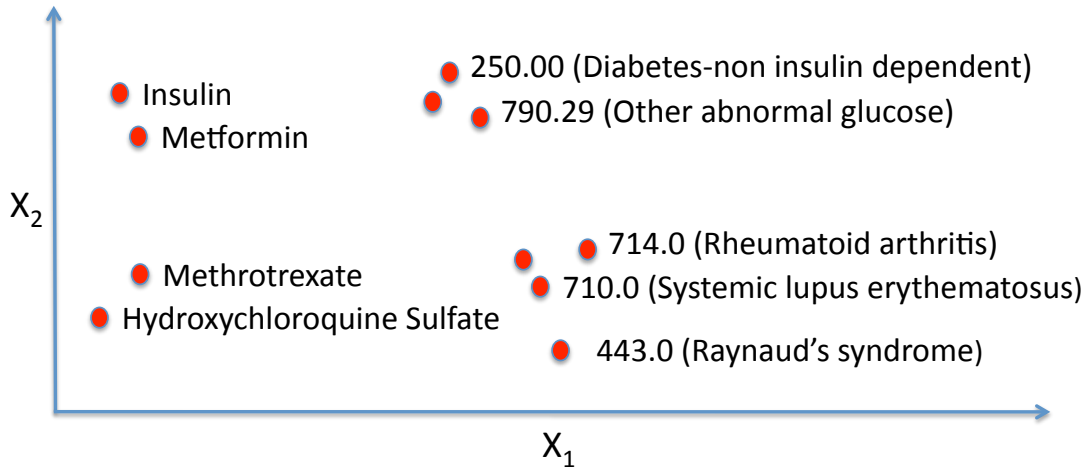

Figure 1: Illustration a low-dimensional representation (in this case, 2 dimensions) of medical concepts. Similar concepts are close to each other in Euclidean space.

Motivation: Why Medical Concept Embeddings?

- Learning distributed low-dimensional representations, word embeddings, has proven particularly useful in various language processing tasks, ranging from simple language modeling to information extraction.

- Many Deep Learning models in natural processing, such as recurrent neural networks, rely on the concept of low dimensional representations.

- The dimensionality of medical concepts, identified with standard medical ontologies are roughly around $100,000$ (UMLS, ICD9, NDC, etc.), which suffers from the same problem of curse of dimensionality.

Previous Work

- Algorithmically, we leverage and modify the work of Mikolov et al., Word2Vec, and Levy et al, SVD method with Shifted Positive Pointwise Mutual Information.

- There are previous works, specifically for learning medical concept embeddings, using the techniques above (DeVine et al., Minaro-Gimenez et al). We will discuss them further in the later section.

Outline

The talk will include two main parts:

- Description of the introduced embeddings: the distributions of interest and necessary algorithmic adjustments.
- Quantitative analysis of the embeddings: an identification of abstract properties and their associated metrics.

Three Types of Medical Concept Embeddings

- Medical Concept Embeddings from Medical Journals(Previous work; MCEMJ)
- Medical Concept Embeddings from Medical Claims (New; MCEMC)
- Medical Concept Embeddings from Clinical Narratives (New; MCECN)

Background

- In natural language processing, the key information, that allows one to learn a sensible set of embedding of words, is the distribution of nearby words, observed at the document level.
- With the distribution of nearby words, one can choose to train a neural network that optimizes the prediction of nearby words, or use techniques for producing low-dimensional factorization of matrices.
- Previous works of medical concept embeddings naturally extend the idea by using the same idea, but restricting the algorithm to medical texts and only medically relevant terms.

We will instead focus on the distribution of temporally nearby medical concepts, within available context of a given patient.

Medical Concept Embeddings from Medical Claims (MCEMC)

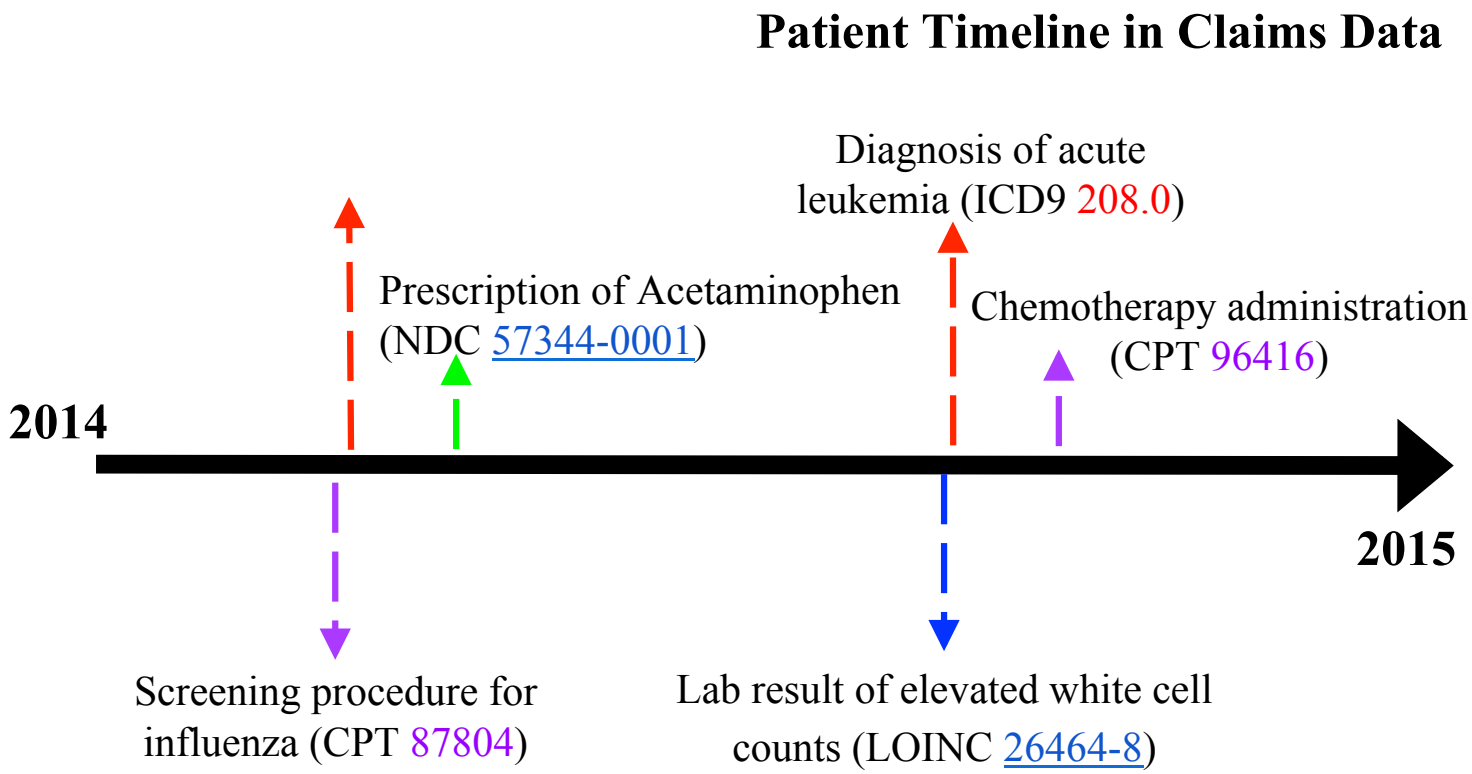

Figure 2: Illustration of the data used to learn embeddings of medical concepts, for a single patient.

Medical Concept Embeddings from Clinical Narratives (MCECN)

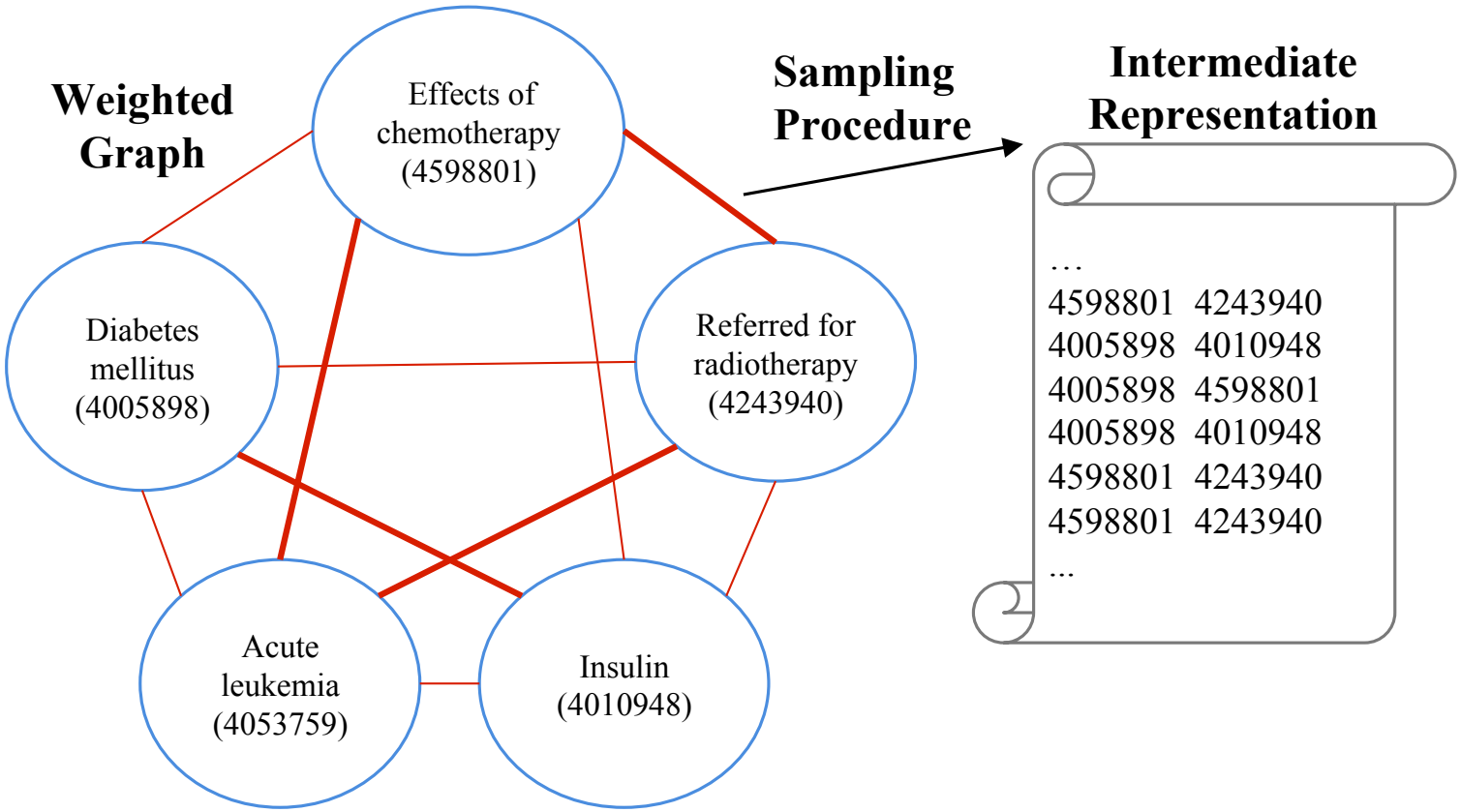

Figure 3: Shown on the left is the input, a weighted graph of medical concept pairs with their temporal frequencies from Shah et al.

Preliminary Studies: the Neighborhood Structures Look Different.

- It is standard to look at the neighborhood structures of word embeddings in natural language processing community.
- Even from our preliminary studies, we qualitatively observed that there were significant differences in the respective neighborhood structures among the embeddings.
- For instance, in one set of embeddings, diagnosis codes for various kinds of cancer formed a cluster, whereas in another set, a cluster around lung cancer would include drugs and treatments related to lung cancer.

Limitations of Qualitative Analysis

- Although it is standard to do such analysis by observing the selected subset of embeddings, it is not entirely obvious, if the claim holds true for the entire set of embeddings ($100,000$ medical concepts)

- What do we exactly mean by medically related?

Can we provide a concrete, quantitative methodology, with which we could precisely characterize each set of embeddings?

Call for Methodology: Two Abstract Properties

To address the issue, we introduce the notion of two abstract properties to formalize the intuition, that we gathered from the preliminary studies:

- The Conceptual Similarity Property
- The Medical Relatedness Property

Defining Surrogate Measures: Medical Ontologies

- We leverage existing medical resources, such as AHRQ's clinical classification software(CCS), ULMS, and the National Drug File - Reference Terminology (NDF-RT) to define surrogate measures that reasonably compute the degrees to which a specific embedding space exhibit the abstract properties.

- The details of the experimental setup can be found directly on the paper, which also contains the link to the code repository.

Definition of the Conceptual Similarity Property

Concretely, we define the Medical Conceptual Similarity Measure (MCSM) of a set of concepts V with respect to a conceptual type set T induced by the UMLS (e.g., neoplastic process), parameterized by a size of the neighborhood k, as:

$$\text{MCSM}_{\text{UMLS}}(V, T, k) \quad = \quad \frac{1}{V(T)} \sum_{v \in V(T)} \sum_{i=1}^{k} \frac{1_T(v(i))}{\log_2(i+1)},$$

where $V(T) \subset V$ is the set of concepts of type T, $v(i)$ denotes the ith closest neighbor of the chosen medical concept v, and 1_T is an indicator function which is 1 if concept $v(i)$ is of type T, and 0 otherwise.

Definition of the Medical Relatedness Property

Concretely, for NDF-RT we define the Medical Relatedness Measure (MRM) of a set of concepts V with respect to a medical relation R, parameterized by a size of the neighborhood k and choice of a seed pair s, as:

$$\text{MRM}_{\text{NDF-RT}}(V, R, k, s) = \frac{1}{|V^*|} \sum_{v \in V^*} 1_R\Big(\bigcup_{i=1}^{k} (v - s)(i) \Big),$$

where $V^* \subset V$ denotes the set of concepts for which NDF-RT specifies at least one pharmacological substance with the given relation, and 1_R is the indicator function which returns 1 if *any* of the medical concepts in the top-k neighborhood of the selected medical concept is an element with the given relation R, and 0 otherwise.

Ontologies are parameters as well.

Computation in Action I: *MCSM*

Table 1: Display of a sub-computation for $\mathrm{MCSM}_{\mathrm{UMLS}}(\mathrm{MCECN}, \text{Neoplastic Process}, 8)$.

Neighbors of CUI 4003436 (Carcinoma, non-small-cell lung) ['Neoplastic Process']
4069419 (small cell carcinoma of lung, C0149925, ['Neoplastic Process']) : 0.956
4394316 (carcinoma of lung, C0684249, ['Neoplastic Process']) : 0.934
4125384 (malignant neoplasm of lung, C0242379, ['Neoplastic Process']) : 0.929
4070138 (adenocarcinoma of lung (disorder), C0152013, ['Neoplastic Process']) : 0.925
4555365 (tarceva, C1135136, ['Organic Chemical', 'Pharmacologic Substance']) : 0.918
4069342 (lung mass, C0149726, ['Finding']) : 0.914
4542086 (alimta, C1101816, ['Organic Chemical', 'Pharmacologic Substance']) : 0.903
4148168 (non-small cell lung cancer metastatic, C0278987, ['Neoplastic Process']) : 0.900

Result I: *MCSM*

Table 2: Medical conceptual similarity property comparison of MCEMJ and MCECN-SGD through $\text{MCSM}_{\text{UMLS}}$.

	$\text{MCSM}_{\text{UMLS}}$(MCEMJ[**?**], −, 40)	$\text{MCSM}_{\text{UMLS}}$(MCECN-SGD, −, 40)
Pharmacologic Substance	**6.74 ± 3.21**	2.95 ± 2.15
Disease or Syndrome	**5.41 ± 2.48**	4.28 ± 1.60
Neoplastic Process	**6.74 ± 3.47**	4.54 ± 0.11
Clinical Drug	**1.01 ± 0.12**	0.12 ± 0.18
Finding	**2.85 ± 1.90**	2.15 ± 1.35
Injury or Poisoning	2.67 ± 2.40	**2.92 ± 2.80**

Computation in Action II: *MRM*

Table 3: Display of a sub-computation for $\mathrm{MRM}_{\text{NDF-RT}}(\text{MCECN}, \text{May-Treat}, 8, -)$.

Neighbors of CUI 4003436 (Carcinoma, non-small-cell lung) ['Neoplastic Process']
4069419 (small cell carcinoma of lung, C0149925, ['Neoplastic Process']) : 0.956
4394316 (carcinoma of lung, C0684249, ['Neoplastic Process']) : 0.934
4125384 (malignant neoplasm of lung, C0242379, ['Neoplastic Process']) : 0.929
4070138 (adenocarcinoma of lung (disorder), C0152013, ['Neoplastic Process']) : 0.925
4555365 (tarceva, C1135136, ['Organic Chemical', 'Pharmacologic Substance']) : 0.918
4069342 (lung mass, C0149726, ['Finding']) : 0.914
4542086 (alimta, C1101816, ['Organic Chemical', 'Pharmacologic Substance']) : 0.903
4148168 (non-small cell lung cancer metastatic, C0278987, ['Neoplastic Process']) : 0.900

Results II: MRM_{NDF-RT}

Table 4: The Medical Relatedness Property comparison of various embeddings through $\text{MRM}_{\text{NDF-RT}}$. The results are of the form (neighbors/avg-seed/max-seed).

	$\text{MRM}_{\text{NDF-RT}}(-, \text{May Treat}, 40)$	$\text{MRM}_{\text{NDF-RT}}(-, \text{May Prevent}, 40)$
$\text{MCEMJ}_{\text{r=5,d=200}}$ [?]	12.59% / 31.56% / 53.92%	18.12% / **35.20%** / 55.88%
$\text{MCEMC}_{\text{month,ns20}}$	10.93% / 28.67% / 57.01%	5.88 % / 29.45% / **57.35%**
$\text{MCEMC}_{\text{month,hs}}$	19.24% / **37.68%** / **60.57 %**	8.82 % / 30.20% / **57.35%**
$\text{MCECN-SGD}_{\text{1Bil,7d,ns20}}$	36.81% / 33.94% / 57.48%	27.94% / 30.42% / 45.59%
$\text{MCECN-SGD}_{\text{10Bil,7d,ns20}}$	38.72% / 34.90% / 57.95%	32.95 % / 31.99% / 48.53%
$\text{MCECN-SVD}_{\text{7d,ns10}}$	**52.26%** / 35.70 % / 53.21%	**39.71%** / 32.32% / 50.00%

Results III: MRM_{CCS}

Table 5: The Medical Relatedness Property comparison of various embeddings through $\mathrm{MRM}_{\mathrm{CCS}}$.

	$\mathrm{MRM}_{\mathrm{CCS}}(-, \text{Fine-grained}, 40)$	$\mathrm{MRM}_{\mathrm{CCS}}(-, \text{Coarse-grained}, 40)$
$\mathrm{MCEMJ}_{\mathrm{r=5,d=200}}$ [?]	0.2293	0.2490
$\mathrm{MCEMC}_{\mathrm{month,ns20}}$	0.4127	0.4422
$\mathrm{MCEMC}_{\mathrm{month,hs}}$	**0.4536**	**0.4804**
$\mathrm{MCECN\text{-}SGD}_{\mathrm{1Bil,7d,ns20}}$	0.2966	0.3319
$\mathrm{MCECN\text{-}SGD}_{\mathrm{10Bil,7d,ns20}}$	0.3087	0.3420
$\mathrm{MCECN\text{-}SVD}_{\mathrm{7d,ns10}}$	0.3461	0.3776

Summary of the Analysis

Neighborhood Visualization: *MCEMC*

Table 6: The neighborhood of the diagnosis code 710.0 in the MCEMC. We display the top 5 neighbors for each type of code, filtering duplicates.

	Nearest Neighbors of ICD9 710.0 (Systemic lupus erythematosus) in MCEMC
	Diagnoses (ICD9)
1	695.4 (Lupus erythematosus)
2	710.9 (Unspecified diffuse connective tissue disease)
3	710.2 (Sicca syndrome)
4	795.79 (Other and unspecified nonspecific immunological findings)
5	443.0 (Raynaud's syndrome)
	Laboratory tests (LOINC)
1	4498-2 (Complement C4 in Serum or Plasma)
2	4485-9 (Complement C3 in Serum or Plasma)
3	5130-0 (DNA Double Strand Ab) in Serum)
4	14030-1 (Smith Extractable Nuclear Ab+Ribonucleoprotein Extractable Nuclear Ab in Serum)
5	11090-8 (Smith Extractable Nuclear Ab in Serum)
	Drugs (NDC)
1	00378037301 (Hydroxychloroquine Sulfate 200mg)
2	00024156210 (Plaquenil 200mg)
3	51927105700 (Fluocinolone Acetonide Miscell Powder)
4	00062331300 (All-flex Contraceptive Diaphragm Arcing Spring Ortho All-flex 80mm)
5	00054412925 (Cyclophosphamide 25mg)

Neighborhood Visualization: *MCECN*

Table 7: A few neighborhood examples from MCECN illustrating genotypic-phenotypic relations.

(cd52, C2733653)	**(bcl1, C2599665)**
(cd52 protein, human, C0376272) (mycosis fungoides/sezary syndrome nos, C0862196) (t-cell receptor, C0034790) (lymphoma, t-cell, cutaneous, C0079773) (pralatrexate, C1721300)	(cyclins, C0079183) (proliferating cell nuclear antigen, C0072108) (lymphoplasmacytic lymphoma, C2700641) (paired box 5 protein, C0167636) (cyclin d1, C0174680)
(jak2 mutation, C2827348)	**(kras mutation, C2747837)**
(refractory anemia with ringed sideroblasts, C1264195) (large platelets (finding), C1148412) (anagrelide, C0051809) (hypercellular bone marrow, C1334068) (myeloid metaplasia, C0027013)	(mesothelioma, C0025500) (cdx2 protein, human, C1505661) (cdx2 antigen, C1829706) (pleural mass, C1709576) (braf protein, human, C1259929)

the article under investigation

Future Work

- As mentioned earlier, looking at various parametrizations, arising from different medical ontologies.
- An iterative loop of updating the ontologies, and searching for new meaningful signals.