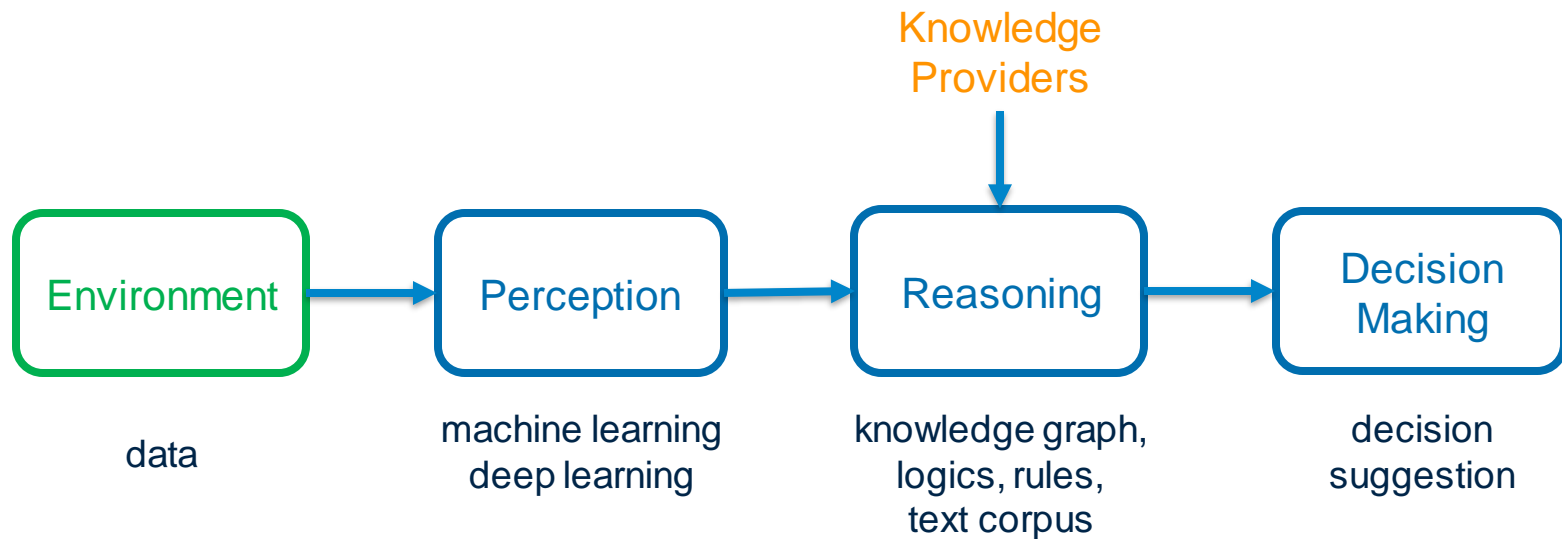


Integrating Prior Knowledge with Learning in Natural Language Processing

Unsupervised Phenotype Annotation via Semantic
Latent Representations on Electronic Health Records

Deep Learning with Prior Knowledge



Prior Knowledge

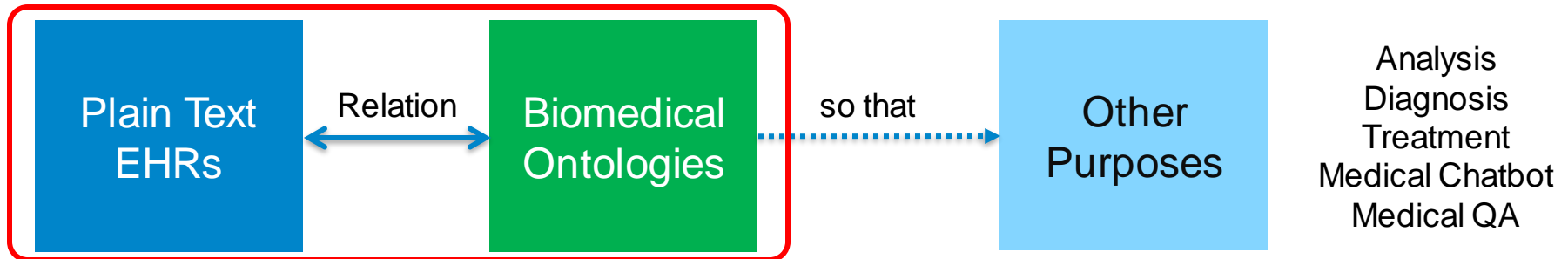
- **Unstructured** knowledge
 - Which can be implicitly contained in **large text corpus**.
 - Great success being used together with **pre-training techniques**.
 - **Structured** knowledge
 - Which can be explicitly defined by knowledge graph, ontologies.
 - Proven to be effective in tasks that require reasoning and understanding.
-

Our Research at Imperial DSI

- Leveraging structured and unstructured data as prior knowledge to improve deep learning models in natural language processing.
- Using structured data
 - Integrating Semantic Knowledge to Tackle Zero-shot Text Classification. NAACL 2019.
 - Unsupervised Annotation of Phenotypic Abnormalities via Semantic Latent Representations on Electronic Health Records. IEEE BIBM2019.
- Using unstructured data
 - PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. Submitted to ICML 2020.

Motivation

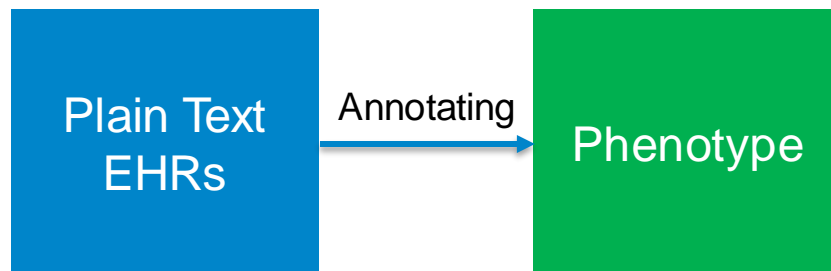
- Electronic health records (EHRs) are widely and increasingly adopted.
- Understanding the professional and natural languages in medical domain with manually pre-defined biomedical ontologies.



- International Classification of Diseases
- Human Phenotype Ontology
- Medical Subject Headings
- Online Mendelian Inheritance in Man
- ...

Motivation

- We aim to annotate EHRs with pre-defined **phenotypes**.
- EHRs serve as a rich source of phenotype information.
- We find the patients who were diagnosed as the same disease could be further classified into **sub-groups by phenotypes**.
- Phenotype annotation on EHRs can help disease diagnosis, and genomic diagnostic, towards precision medicine.



- **Human Phenotype Ontology (HPO)**
- Phenotype: observable characteristics
- Examples:
 - Abnormality of the digestive system
 - Abnormality of the immune system
 - Abnormality of the nervous system

Datasets

- MIMIC-III
- A public EHRs database with 52k notes from 40k patients

- Human Phenotype Ontology (HPO)
- Standardize 13k HPO terms and each with description.

Admission Date: **2117-8-11** Discharge Date: **2117-8-11**
Date of Birth: **5053-3-21** Sex: F
Service: MEDICINE
Admission: **First Name(L) 2117**
Chief Complaint: nausea, vomiting
Major Surgical or Invasive Procedure:
NONE

History of Present Illness
S/P of acute myocardial Type 1 diabetes mellitus w/ neuropathy, nephropathy, HTN, gastropancreas, CAD and nephropathy, recently hospitalized for catheter-directed angioplasty 720 7/17/17.
Neuropathy: **Date onset 12/2008** CAD hospitalizations in **4/12/17** and **7/13/17** with inferior MI, HTN, and diabetes on treatment with insulin, glimepiride, lisin, and diuretic as assessed. Last office blood glucose 115, 130, and 140. On 2/17/17, she also had a low HbA_{1c} which raised concern for hypoglycemia. Patient has been on DKA with A1c 10 and bicarb 11.

In the ED initial vitals were RR 20, HR 111, BP 159/22, SpO2 94, 4.7, WBC 11.0, Anion Gap 18.0, 2.0 Bicarb 12.0, 2.0, 2.0 on her 2nd, 1st, insulin drip at 4 units/hr. On home at 2300, bicarb was on a 2 with difficulty to control sugar 180, have been High. Given 30 cc intravenous insulin in ED.
She was called on an insulin drip at 1 unit/hr until 8:00. Insulin also up to 22mg PO and Metformin 500 TID for past 20 days (low 180 MG).

Review of systems: otherwise negative.
Past Medical History
Type 1 diabetes mellitus w/ neuropathy, nephropathy, and nephropathy 720 7/17/17 **2117-8-11**
HTN, 5 years
Atherosclerosis 1.5 years
CAD: MI, Coronary A. B. S. stenosis
11 vertebral fracture **2117-3-17**
Nephrotic proteinuria
Social History
Patient lives at home w/ **Last Name(S) ** with her 6th daughter and boyfriend. She has no history of STD, tobacco, or alcohol use. She is currently unemployed and awaiting disability.

Family History
Both parents have HTN and T2DM. Grandfather had an MI in his
Physical Exam
GDM: awake, alert, and oriented
HEENT: PMMA, MMMA, no PD, neck supple. No normal LAD
Chest: RR 22, normal, 0/0 crackles (normal), normal heart
Bowel: at 1200. Normal bowel sounds
Pain: CMA, with no radiation or other
Skin: BP, HR, TC, not red, not itchy, not dry
Spleen: not enlarged, no splenomegaly, DOP, PLS
Diabetes: no edema, no xanthelasma
Sclera: no xanthelasma or bruising, no skin-healing
Nails: thin, brittle, slight onychomycosis, Paster 5/4/17
Bilirubin: 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4
Hemoglobin: 10.1, 10.1, 10.1, 10.1, 10.1, 10.1, 10.1, 10.1
Neurology: normal sensation distal to ankles.

Posttest Results:

Admission Labs **2117-8-11 (2117-8-11)
WBC (1.0) 16.40, HB (13.0) 16.30, HCT (3.5) 46.00, PCT (12.6) 12.60, UPOB (2) 4.00, ALT (69) 14.00, AST (69) 14.00, PHOS (3) 107.00, BUN (5) 10.00, CREA (1.0) 1.40, URIC (4) 6.40, CRP (3) 2.50, SODIUM (127) 127.00, POTASSIUM (4.0) 3.00, GLU (161) 161.00
Discharge Labs **2117-8-11 (2117-8-11)
WBC (4) 16.40, HB (13.0) 16.30, HCT (3.5) 46.00, PCT (12.6) 12.60, UPOB (2) 4.00, ALT (69) 14.00, AST (69) 14.00, PHOS (3) 107.00, BUN (5) 10.00, CREA (1.0) 1.40, URIC (4) 6.40, CRP (3) 2.50, SODIUM (127) 127.00, POTASSIUM (4.0) 3.00, GLU (161) 161.00****

Radiology
CXR: No evidence of pneumonia or other pulmonary abnormalities. No pulmonary edema. Normal size of the cardiac silhouette.
MRA: No evidence of stenosis. Mild atherosclerotic changes in the aorta, noted at C4/5/6.

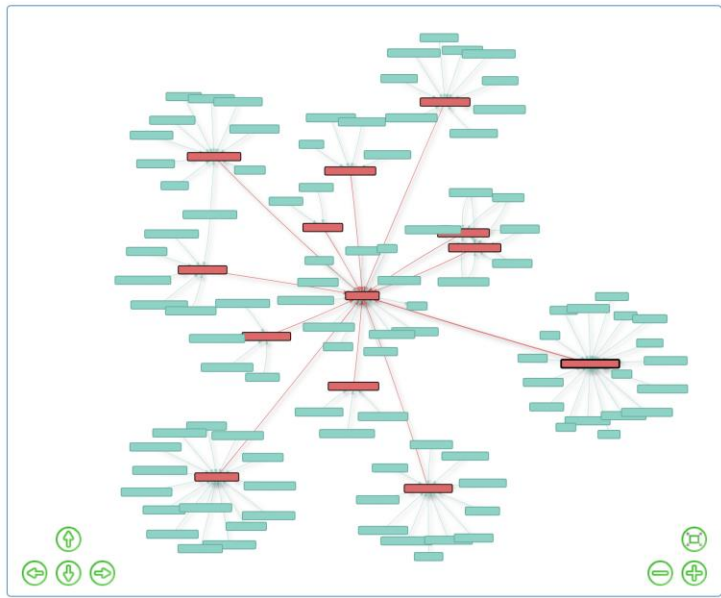
Brief Hospital Course
S/P of HTN & newly corrected type 1 DM, with neuropathy, gastropancreas, nephropathy 77777 (CAD), nephropathy presents with DKA and hypernatremia S/P T2DM.
A Diabetic ketoacidosis: Patient admitted at home with vomiting and tingling in lower extremities. Sugar at home recently has been 300. On 2/17/17, glucose was 300. On her 2nd, 1st, insulin drip at 4 units/hr. On home at 2300, bicarb was on a 2 with difficulty to control sugar 180, have been High. Given 30 cc intravenous insulin in ED.
She was called on an insulin drip at 1 unit/hr until 8:00. Insulin also up to 22mg PO and Metformin 500 TID for past 20 days (low 180 MG).

Medications on Admission
1. Chlorzoxazone 250mg Tablet Sig: One (1) Tablet PO Q2H (CAD)
2. Lantus 100 units/10mL Solution Sig: Twenty Two (22) units Subcutaneous every 8H
3. Lantus 100 units/10mL Solution Sig: Twelve (12) units Subcutaneous every 8H
4. Humalog 100 units/10mL Solution Sig: Sliding scale as directed Subcutaneous before meals and 1hr before bedtime
5. Metformin 500 mg PO Tablet Sig: Sliding scale as directed Subcutaneous before meals and 1hr before bedtime
6. Insulin 100 units/10mL Solution Sig: Sliding scale as directed Subcutaneous before meals and 1hr before bedtime
7. Insulin 100 units/10mL Solution Sig: Sliding scale as directed Subcutaneous before meals and 1hr before bedtime

**7 glibenclamide 500 mg Capsule Sig: One (1) Capsule PO Q2H every 12 hours
Discharge Medications
1. Chlorzoxazone 250mg Tablet Sig: One (1) Tablet PO DAILY (CAD)
2. glibenclamide 500 mg Capsule Sig: One (1) Capsule PO Q2H every 12 hours
3. Chlorzoxazone 250mg Tablet Sig: One (1) Tablet PO Q2H (CAD)
4. Humalog 100 units/10mL Solution Sig: Sliding scale as directed Subcutaneous before meals and 1hr before bedtime
5. Metformin 500 mg PO Tablet Sig: Three (3) Tablets PO DAILY (CAD)
6. Lantus 100 units/10mL Solution Sig: As directed by **Last Name(S) ** units**

Discharge Disposition
Home
Discharge Diagnosis
Diabetic ketoacidosis
Hypernatremia (Blood test values)
Insulin resistance
Chronic renal insufficiency
Discharge Condition
Manual Return, Clear and coherent.
Level of Consciousness: Alert and oriented.
Activity Status: Ambulatory - independent.

Discharge Instructions
You were admitted to the hospital with DKA, hypernatremia, and blood in your vomit. You were admitted to the ED with an insulin drip, and your blood sugars improved. Your blood glucose medications were adjusted to better control your blood glucose while you were in DKA, but you were not started on your home insulin at discharge. This blood in your vomit was likely related to the insulin drip. You should continue to follow up with your primary care doctor in 3-4 weeks to discuss whether you should change your insulin regimen. You should follow up with your primary care doctor in 3-4 weeks to discuss whether you should change your insulin regimen. You should follow up with your primary care doctor in 3-4 weeks to discuss whether you should change your insulin regimen. You should follow up with your primary care doctor in 3-4 weeks to discuss whether you should change your insulin regimen.



Previous Works

- Information retrieval based approaches
 - E.g. OBO Annotator, NCBO Annotator, Bio-LarK, MetaMap, etc.
 - Suffers from computational inefficiency.
- Deep learning models
 - E.g. CNN [1]
 - Requires gold standards which is hard and expensive to acquire.

Our Work

- We propose a novel **unsupervised** deep learning framework to exploit supportive phenotype knowledge in HPO and annotate general phenotypes from EHRs semantically.
 - We demonstrate that our proposed method achieves state-of-the-art annotation performance and computational efficiency compared with other methods.
-

Problem Formulation (Example)

EHRs snippets

“Your blood pressure medications were adjusted to better control your **blood pressure** while you were in **DKA**”

“Given your complaints of **chronic cough** and **heartburn**, you should also discuss beginning a trial of a proton pump inhibitor such as Nexium or Prilosec to see if this helps your symptoms”

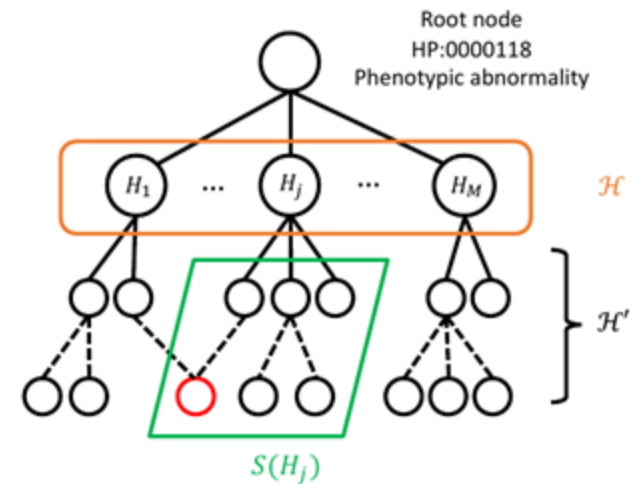
Potential annotation with HPO terms

- **Abnormality of the cardiovascular system**
- **Abnormality of the digestive system**

- **Abnormality of the respiratory system**
- **Abnormality of the digestive system**

Problem Formulation

- There are two types of data sources.
 - $\mathcal{X} = \{X_1, \dots, X_N\}$: a collection of EHRs and each EHR consists of textual notes written by clinicians.
 - $\mathcal{H} = \{H_1, \dots, H_M\}$: a standardized general category of human phenotypes provided by HPO.
 - The HPO also provides additional subclasses \mathcal{H}'



Problem Formulation

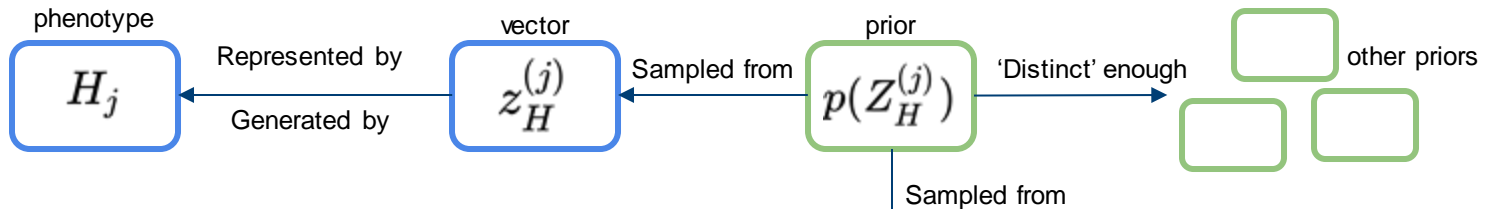
- The EHR can include either multiple phenotypes or a single or none.
Therefore, learning the annotation of phenotypes from EHRs is essentially learning the conditional probability:

$$p(\mathbf{1}_{H_j} | X_i)$$

- I.e. a binary classification for each H_j
- As a whole, a multi-label classification on \mathcal{H}

Semantic Latent Representations

- Assumptions
 - The semantics of a general **phenotype** is represented by a **prior distribution**. The prior distribution of each phenotype should be 'distinct' enough from each other.



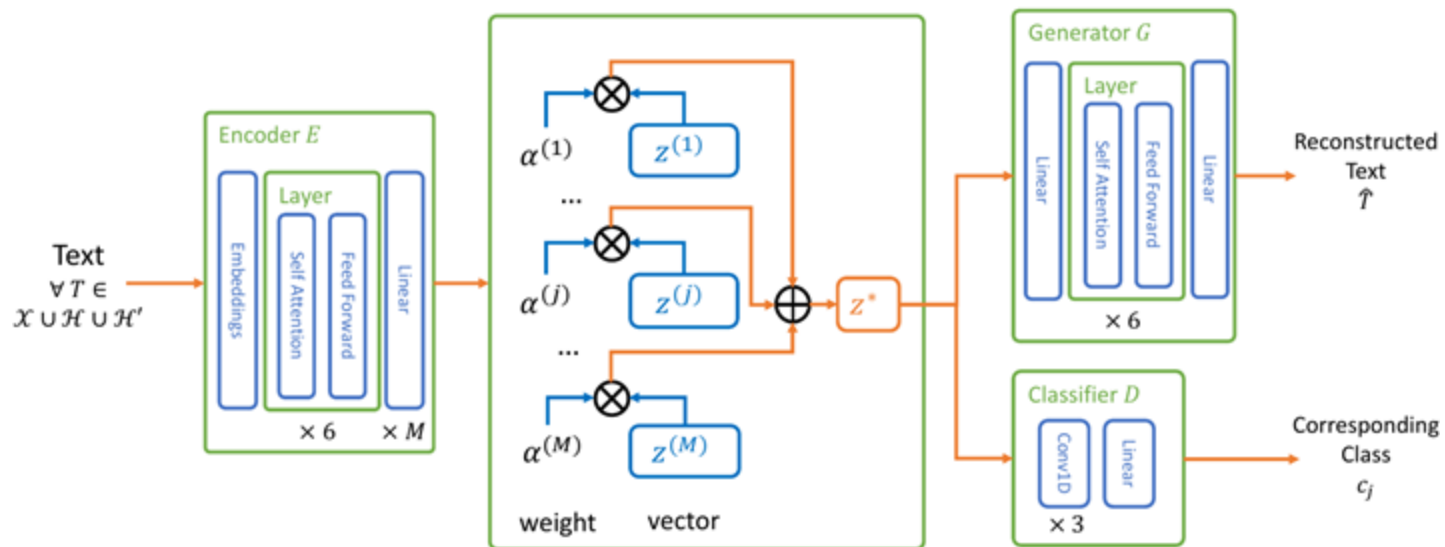
- The semantics of EHR is a composition of the semantics of phenotypes.

Represented by
Generated by

$$z_i^* = \sum_{j=1}^M \alpha_i^{(j)} z_i^{(j)}$$

weight

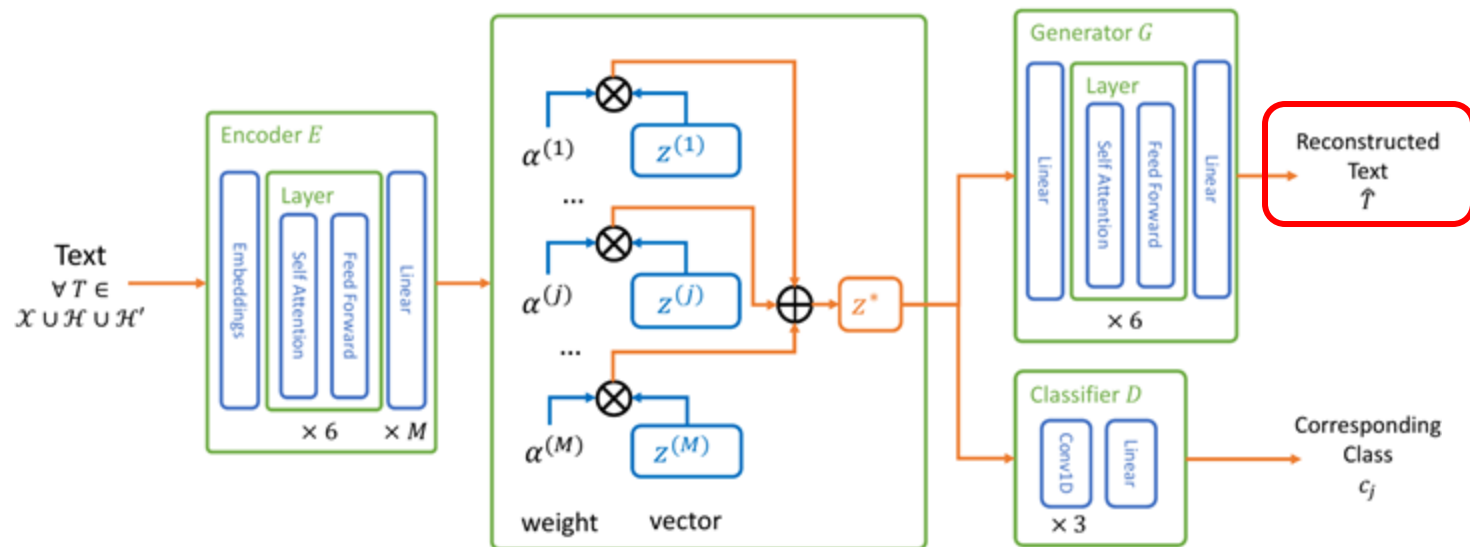
An Auto-Encoder Model



- General text reconstruction

$$\max \mathbb{E}_{Z \sim p(z^*|T)} [p(T|Z)] = \max_{\theta_E, \theta_G} \mathbb{E}_{Z \sim p_E(T; \theta_E)} [p_G(Z; \theta_G)]$$

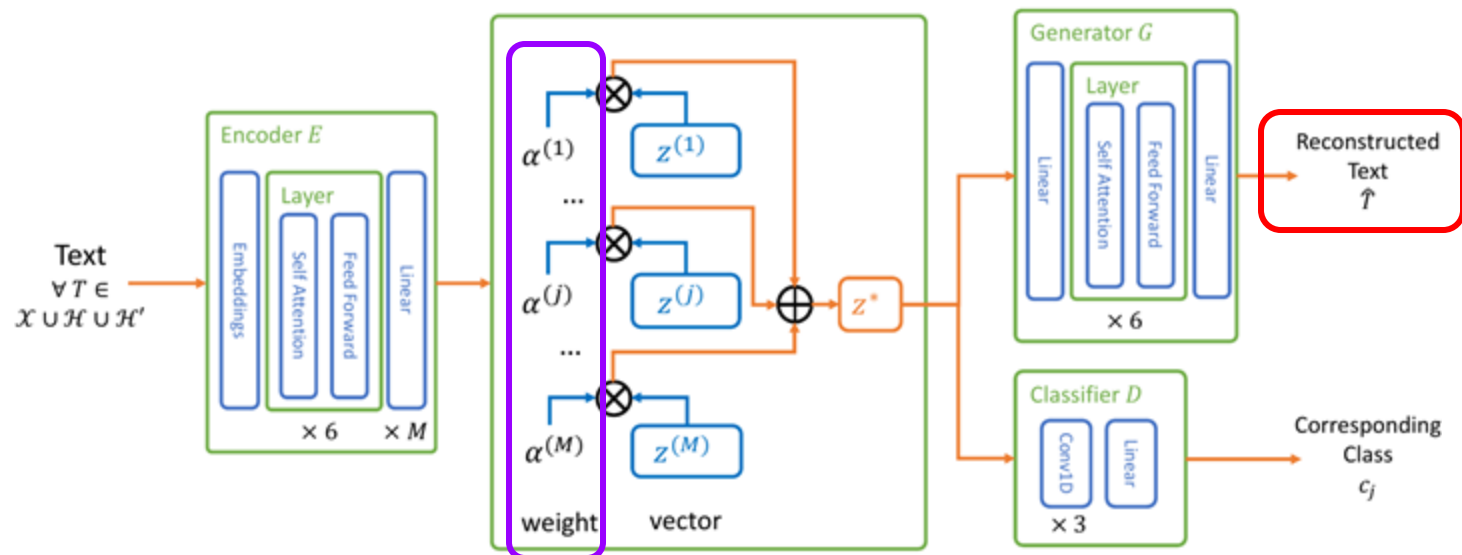
An Auto-Encoder Model



- Loss 1: text reconstruction of EHRs.

$$\mathcal{L}_{\text{rec}}^X = \frac{1}{N} \sum_{i=1}^N \left[-\log p_G(X_i | E(X_i)) \right]$$

An Auto-Encoder Model



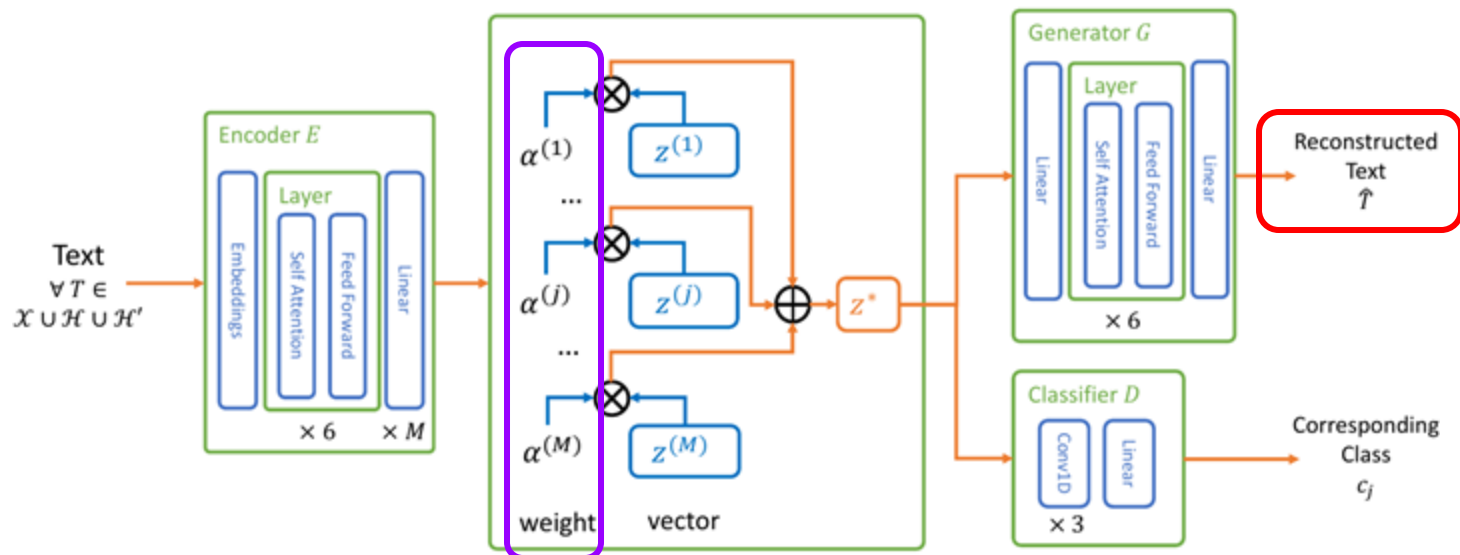
- Loss 2: text reconstruction of the general phenotypes' description.

$$\mathcal{L}_{\text{rec}}^H = \frac{1}{M} \sum_{j=1}^M \left[-\log p_G(H_j | E(H_j)) + \frac{1}{M} \left[-\log(\alpha^{(j)}) - \sum_{k=1, k \neq j}^M \log(1 - \alpha^{(k)}) \right] \right]$$

reconstruction

constraint

An Auto-Encoder Model



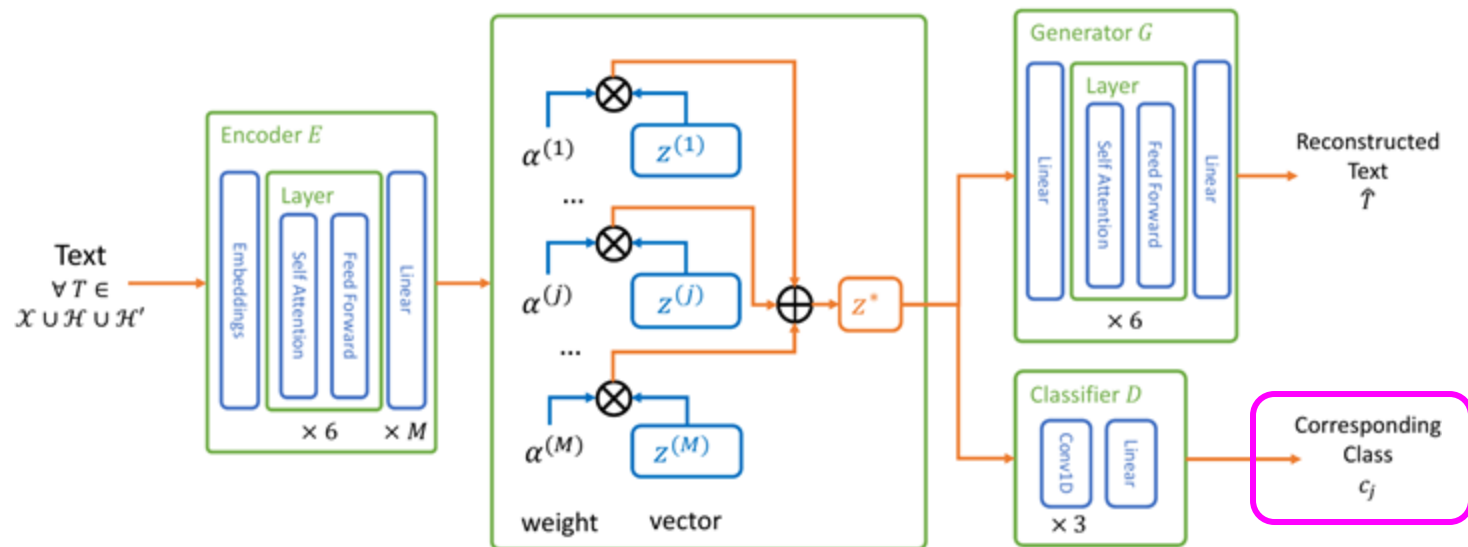
- Loss 3: text reconstruction of the phenotype subclasses' description.

$$\mathcal{L}_{\text{rec}}^{H'} = \frac{1}{|H'|} \sum_{H' \in H'} \left[-\log p_G(H'|E(H')) \right] + \frac{1}{M} \left[-\sum_{H' \in S(H_j)} \log(\alpha^{(j)}) - \sum_{H' \notin S(H_j)} \log(1 - \alpha^{(j)}) \right]$$

reconstruction

constraint

An Auto-Encoder Model



- Loss 4: the latent vectors sampled from different priors can be classified to different classes, then the priors are thought to be 'distinct' enough.

$$\mathcal{L}_{\text{pr}} = \frac{1}{\#_T} \sum_T \sum_{j=1}^M \left[-\log p_D(c_j | z^{(j)} \in E(T)) \right]$$

Algorithm

Algorithm 1: The training algorithm.

Input: EHRs \mathcal{X} (training set), general phenotypic abnormalities \mathcal{H} , and additional subclasses \mathcal{H}' .

1 Initializing $\theta_E, \theta_G, \theta_D$;

2 **repeat**

3 Sample a mini-batch of B textual examples

$$\{T_{(i)}\}_{i=1}^B \subseteq \mathcal{X} \cup \mathcal{H} \cup \mathcal{H}' ;$$

4 Get $z_{(i)}^*$ and $\{z_{(i)}^{(j)}\}_{j=1}^M$ by $E(T_{(i)})$;

5 Reconstruct $\hat{T}_{(i)}$ by $G(z_{(i)}^*)$;

6 Calculate $\mathcal{L}_{rec}^X, \mathcal{L}_{rec}^H, \mathcal{L}_{rec}^{H'}$ respectively ;

7 Classify $z_{(i)}^{(j)}$ by $D(z_{(i)}^{(j)})$;

8 Calculate \mathcal{L}_{pr} by Equation 7 ;

9 Update $\theta_E, \theta_G, \theta_D$ by gradient descent on:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rec}^X + \lambda_2 \mathcal{L}_{rec}^H + \lambda_3 \mathcal{L}_{rec}^{H'} + \lambda_4 \mathcal{L}_{pr} \quad (6)$$

10 **until** convergence;

Output: The encoder E .

Experiments - Datasets

- Discharge summaries from MIMIC-III as EHR.
 - #EHR: 52,722
 - 70% as training, 30% as testing (random split)
- Human Phenotype Ontology (HPO).
 - General phenotype $M = |\mathcal{H}| = 24$
 - Phenotype subclasses $|\mathcal{H}'| = 13795$

Experiments - Time Efficiency

A COMPARISON OF DIFFERENT METHODS. THE #RECORDS REFERS TO THE NUMBER OF TEXTUAL RECORDS USED IN THE ORIGINAL WORKS. THE TIME WAS MEASURED BY THE DURATION OF ANNOTATING 52,722 EHRs IN INFERENCE STAGE WITH A SINGLE THREAD INTEL I7-6850K 3.60GHZ AND A SINGLE NVIDIA TITAN X.

Method	Available (A) Open source (O)	#Records	Time to annotate 52,722 EHRs
OBO	A, Not O	515	1.0 hour
NCBO	A, Not O	/	36.7 hours
MetaMap	A, O	/	~ 22 days
Bio-LarK	Not A, Not O	228	/
CNN [16]	Not A, Not O	1,610	/
Ours	A, O	52,722	40.2 min

Experiments - Accuracy

THE PERFORMANCE OF ANNOTATION RESULTS COMPARED WITH THE SILVER STANDARD. ALL THE NUMBERS ARE AVERAGED ACROSS EHRs IN THE TESTING SET.

Method	Precision	Recall	F1
Random	0.5541	0.5401	0.5108
Keyword	0.6732	0.4982	0.5194
OBO	0.6817	0.5917	0.5775
NCBO	0.6782	0.5724	0.5659
MetaMap	0.7425	0.5231	0.5576
Ours	0.7113	0.6805	0.6383

- Silver standard



[1] <https://hpo.jax.org/app/download/annotation>

[2] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," Proceedings of the National Academy of Sciences, vol. 104, no. 21, pp. 8685-8690, 2007.

[3] J. Park, D.-S. Lee, N. A. Christakis, and A.-L. Barabási, "The impact of cellular networks on disease comorbidity," Molecular systems biology, vol. 5, no. 1, p. 262, 2009.

Conclusion & Future Works

- A novel unsupervised deep learning framework to annotate phenotype from EHRs.
- The experiments have shown the effectiveness and efficiency of our method.
- We believe our method can provide a better indication for disease diagnosis.

- Integrate external biomedical literature from PubMed, Elsevier, etc.
- Extend to annotate all the 13k specific phenotype in HPO.
- Improve embeddings of HPO by taking both semantics and hierarchy into consideration.
- Apply to general domains.

Imperial College
London

Thanks! QA?

Jingqing Zhang
Data Science Institute
Imperial College London
jz9215@imperial.ac.uk
