CCS Coding of Discharge Diagnoses via Deep Neural Networks

Chadi Helwe * [1] Shady Elbassuoni [1] Mirabelle Geha [2]
Eveline Hitti [2] Carla Makhlouf Obermeyer [3]

[1]Department of Computer Science, American University of Beirut

[2]Department of Emergency Medecine, American University of Beirut

[3]Department of Epidemiology and Population Health, American University of Beirut

July 5, 2017

*cth05@aub.edu.lb

Outline

Outline

Introduction

Discharge Diagnosis

- Is a short piece of free-text that describes the final diagnosis provided to a patient by health-care professionals upon their release from hospitals

International Classification of Diseases Code (ICD Code)

- Used for billing purposes and for statistical analysis and reporting

Clinical Classification Software Code (CCS Code)

- CCS collapses over 14,000 ICD codes into 285 mutually exclusive categories, known as the single-level CCS codes

Example

Diagnosis	ICD Code Description	CCS Code Description
benzodiazepine overdose	poisoning by other sedatives and hypnotics	poisoning by other medications and drugs
left wrist laceration	open wound of wrist, with tendon involvement	open wounds
complete heart block	atrioventricular block, complete	conduction disorders

Goal

We aim to bypass the stage of ICD coding and *directly* map discharge diagnoses into CCS codes

Why?

- ICD coding is a tedious error-prone *manual* process
- Automating ICD coding is difficult due to the large number of possible ICD codes

How?

- We investigated the applicability of deep learning to automatically predict CCS codes for discharge diagnoses

Outline

Procedure

CCS Coding using Deep Learning:

- Step 1: We trained word embeddings using CBOW on a corpus of over 10,000 medical documents
- Step 2: We preprocessed the data using MetaMap [2] to retrieve medical concepts and fold-out abbreviations
- Step 3: We transformed the preprocessed data to vectors using our embeddings
- Step 4: We trained a Long Short-Term Memory (LSTM) followed by dense neural networks

Example: MetaMap Preprocessing

Diagnosis	MetaMap Concepts
interior myocardial infarction	cardiac infarction
hypertension	hypertensive disease
uti	urinary tract infections

Deep Learning Model

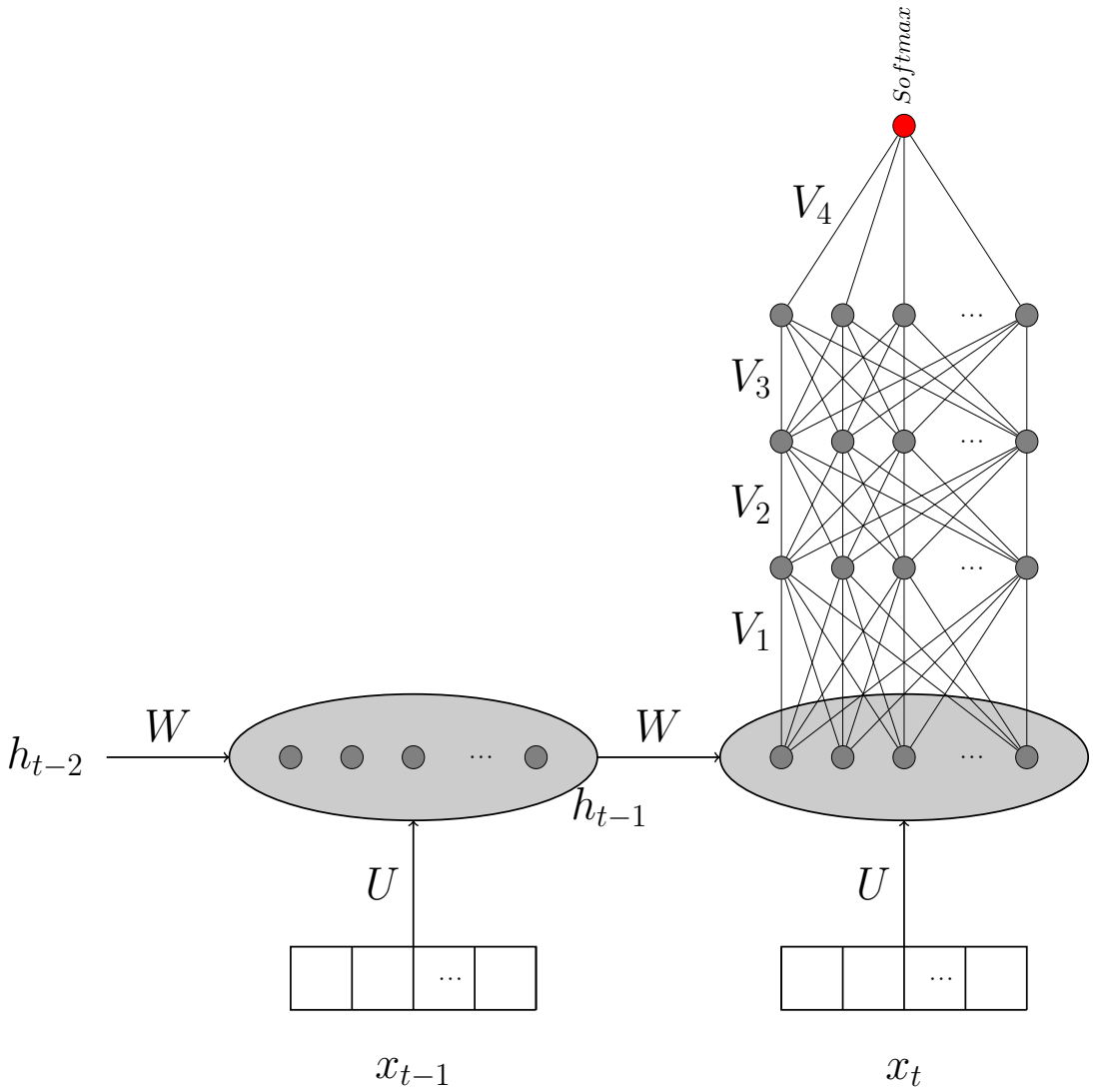

Figure 1: LSTM + Dense Neural Networks

Outline

Dataset

We used the MIMIC-III [8] dataset:

- Consists of discharge diagnoses for over 58000 deidentified patients
- ICD-9 codes were manually assigned
- CCS codes were retrieved using the CCS tool
- We used a *majority vote* to obtain a single CCS code per diagnosis
- The dataset was split as follows:
  - 80% training
  - 10% validation
  - 10% test

Hypotheses

We investigated three hypotheses:

- First, predicting the CCS codes directly from discharge diagnoses is more effective than predicting ICD codes first then using the CCS tool
- Second, using deep learning based on word embeddings is more effective than using SVM that relies on tf-idf
- Third, preprocessing discharge diagnoses using the MetaMap tool improves the coding process

Experiment Setup

Deep learning model:

- We used stochastic gradient descent as the optimizer
- We trained for 100 epochs with a batch size of 5

SVM model:

- We used tf-idf as input features
- We used the linear soft-margin SVM

Example: CCS Coding

Diagnosis	esophageal cancer sda	motor vehicle accident
DNN Direct	cancer of esophagus	crushing injury or internal injury
DNN ICD + CCS Tool	cancer of stomach	other fractures
SVM Direct	cancer of esophagus	intracranial injury
SVM ICD + CCS Tool	cancer of stomach	other fractures
Actual CCS Code	cancer of esophagus	crushing injury or internal injury

Results

Model	Raw Data (F1 Score)	MetaMap Data (F1 Score)
DNN Direct	0.863	0.959
DNN ICD + CCS Tool	0.788	0.845
SVM Direct	0.838	0.954
SVM ICD + CCS Tool	0.778	0.843

Outline

Conclusion

Conclusion:

- We proposed an approach to automatically code medical records such as discharge diagnoses
- Our approach uses deep learning
- All of three approaches were validated:
  - Predicting the CCS codes directly from discharge diagnoses is more effective than predicting ICD codes first then using the CCS tool
  - Using deep learning based on word embeddings is more effective than using SVM that relies on tf-idf
  - Preprocessing discharge diagnoses using the MetaMap tool improves the coding process

Future Work:

- We plan to test the validity of our hypotheses on more datasets
- We plan to extend our model to handle the case of multilabels

Thank You

Additional Slides

Related Work

ICD coding	MetaMap in the biomedical domain	Deep learning in the biomedical domain
Lita et al. [11]	Sanchez et al. [13]	Cernazanu et al. [4]
Goldstein and Arzumtsyan [7]	Lana et al. [9]	Al Rahhal et al. [1]
Farakas and Szarvas [6]	Aronson and Rindflesch [3]	Lipton et al. [10]
Perotte et al. [12]	Goldstein and Arzumtsyan [7]	Choi et al. [5]
Yan et al. [14]		

MetaMap Preprocessing

MetaMap:

- A tool that automatically maps biomedical texts into the Unifided Medical Language System (UMLS) concepts
- Uses a knowledge-intensive approach based on symbolic, natural-language processing and computational linguistic techniques

Data Preprocessing:

- We passed each diagnosis to MetaMap as input and retrieved a set of Concept Unique Identifiers (CUIs)
- Each CUI represents a UMLS concept
- We used the preferred names of the concepts identified and these represent the new instances

Word Embeddings

Word embeddings:

- Are vector representations of words that are used as input features to deep learning neural networks
- Were trained using the Continuous Bag of Words (CBOW) model on a large medical corpus of 10,403 documents
- The large medical corpus consists of:
  - Articles of all the diseases described in Wikipedia
  - Web pages describing diseases from medical websites like MedlinePlus, MayoClinic, John Hopkins Medicine, and National Health Service

References I

MM Al Rahhal, Yakoub Bazi, Haikel AlHichri, Naif Alajlan, Farid Melgani, and RR Yager.
Deep learning approach for active classification of electrocardiogram signals.
Information Sciences, 345:340–354, 2016.

Alan R Aronson.
Effective mapping of biomedical text to the umls metathesaurus: the metamap program.
In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

Alan R Aronson and Thomas C Rindflesch.
Query expansion using the umls metathesaurus.
In *Proceedings of the AMIA Annual Fall Symposium*, page 485. American Medical Informatics Association, 1997.

References II


Cosmin Cernazanu-Glavan and Stefan Holban.
Segmentation of bone structure in x-ray images using convolutional neural network.
Adv. Electr. Comput. Eng, 13(1):87–94, 2013.


Youngduck Choi, Chill Yi-I Chiu, and David Sontag.
Learning low-dimensional representations of medical concepts.
AMIA Summits on Translational Science Proceedings, 2016:41, 2016.


Richárd Farkas and György Szarvas.
Automatic construction of rule-based icd-9-cm coding systems.
BMC bioinformatics, 9(3):S10, 2008.

References III


MBA Ira Goldstein and MLS Anna Arzumtsyan.
Three approaches to automatic assignment of icd-9-cm codes to radiology reports.
2007.


Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark.
Mimic-iii, a freely accessible critical care database.
Scientific data, 3, 2016.


Sara Lana-Serrano, Daniel Sanchez-Cisneros, Leonardo Campillos, and Isabel Segura-Bedmar.
Recognizing chemical compounds and drugs: a rule-based approach using semantic information.
In *BioCreative Challenge Evaluation Workshop*, volume 2, page 121. Citeseer, 2013.

References IV

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell.
Learning to diagnose with lstm recurrent neural networks.
arXiv preprint arXiv:1511.03677, 2015.

Lucian Vlad Lita, Shipeng Yu, Radu Stefan Niculescu, and Jinbo Bi.
Large scale diagnostic code classification for medical patient records.
In *IJCNLP*, pages 877–882. Citeseer, 2008.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad.
Diagnosis code assignment: models and evaluation metrics.
Journal of the American Medical Informatics Association, 21(2):231–237, 2014.

References V

Daniel Sanchez-Cisneros, Paloma Martínez, and Isabel Segura-Bedmar.
Combining dictionaries and ontologies for drug name recognition in biomedical texts.
In *Proceedings of the 7th international workshop on Data and text mining in biomedical informatics*, pages 27–30. ACM, 2013.

Yan Yan, Glenn Fung, Jennifer G Dy, and Romer Rosales.
Medical coding classification by leveraging inter-code relationships.
In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 193–202. ACM, 2010.