國立臺灣大學電機資訊學院暨中央研究院

資料科學學位學程

碩士論文

Data Science Degree Program

College of Electrical Engineering and Computer Science

National Taiwan University and Academia Sinica

Master Thesis

整合階層種類知識於多標籤診斷文字理解

Leveraging Hierarchical Category Knowledge for Multi-Label Diagnostic Text Understanding

蔡尚錡

Shang-Chi Tsai

指導教授：陳縕儂博士、古倫維博士

Advisor: Yun-Nung Chen, Ph.D.、Lun-Wei Ku, Ph.D.

中華民國 108 年 8 月

August, 2019

doi:10.6342/NTU201903870

doi:10.6342/NTU201903870

誌謝

兩年來的碩士歷程，讓我從一個完全不懂機器學習與深度學習的研究新人慢慢蛻變成可以想出有趣的研究方向與題目，進而提出合適的算法與模型用來解決問題。能有這樣的成長，我最感謝陳縕儂教授的指導，帶我一步一步地進行研究，透過設計實驗與撰寫學術論文的過程，我學到了很多也過得非常有成就感。同時，我也感謝實驗室每一位學長姐與學弟妹，大家在我的碩士生活中，總是會陪我聊天舒壓也會給我很多研究過程中的建議，讓我收穫很多不同方向的知識，也因爲看到你們的努力與成就，進而督促自己能變得更強。最後，希望實驗室的大家都能過得很好，朝自己的目標邁進。

doi:10.6342/NTU201903870

doi:10.6342/NTU201903870

摘要

電子病歷中記錄了病人相關的症狀描述、看診的歷史紀錄等文字資料，每一筆紀錄都有其對應的診斷代碼，代表著該次就醫時醫師所下的診斷結果及其治療方案等資訊。這篇論文主要想利用醫院中這類型的文字描述資料結合醫學上的專家知識來做診斷代碼的預測。為了能讓模型有效利用診斷代碼的階層關係這類額外的專家知識，我們提出了各種不同的方式去計算卷積神經網路的損失函數以此來取得同一種類別的診斷中所共享的語義資訊。這樣的資訊不只讓模型有額外的醫學知識作為學習方向，也幫助解決訓練資料中樣本數量不平衡的問題。根據我們做在 MIMIC3 這份國際通用的資料集的結果顯示，我們提出的方法確實能夠有效利用階層種類的知識並提供模型有意義的資訊來幫助改善現階段最好的預測結果。而這樣的討論與研究也顯示了結合額外的專家知識於機器學習的模型中是有一定的好處與重要性，能啓發未來更多的研究方向。

關鍵字：診斷文字, 醫療診斷, 多標籤預測

doi:10.6342/NTU201903870

doi:10.6342/NTU201903870

Abstract

Clinical notes are essential medical documents to record each patient's symptoms. Each record is typically annotated with medical diagnostic codes, which means diagnosis and treatment. This paper focuses on predicting diagnostic codes given the descriptive present illness in electronic health records by leveraging domain knowledge. We investigate various losses in a convolutional model to utilize hierarchical category knowledge of diagnostic codes in order to allow the model to share semantics across different labels under the same category. The proposed model not only considers the external domain knowledge but also addresses the issue about data imbalance. The MIMIC3 benchmark experiments show that the proposed methods can effectively utilize category knowledge and provide informative cues to improve the performance in terms of the top-ranked diagnostic codes which is better than the prior state-of-the-art. The investigation and discussion express the potential of integrating the domain knowledge in the current machine learning based models and guiding future research directions.

Keywords: clinical notes, multi-label classification, ICD prediction

doi:10.6342/NTU201903870

doi:10.6342/NTU201903870

Contents

誌謝 iii

摘要 v

Abstract vii

1 Introduction 1

1.1 Motivation and Problem Description 1

1.2 Main Contribution 2

1.3 Thesis Structure 2

2 Proposed Approach 3

2.1 Convolutional Models 3

2.1.1 TextCNN 3

2.1.2 Convolutional Attention Model (CAML) 4

2.2 Knowledge Integration Mechanism 4

2.2.1 Cluster Penalty 4

2.2.2 Multi-Task Learning 5

2.2.3 Hierarchical Learning 5

2.3 Training 6

3 Evaluation 7

3.1 Experimental Setup 7

3.2 Results 8

doi:10.6342/NTU201903870

3.3 Qualitative Analysis 9

4 Conclusion 11

Bibliography 13

doi:10.6342/NTU201903870

List of Figures

2.1 The architecture with the proposed category knowledge integration. . 6

doi:10.6342/NTU201903870

List of Tables

3.1 Dataset comparison and statistics. From the full set of the internal data (1495 labels) to 200, only 6.0% of data points are discarded. . . 8

3.2 The results on internal data. 9

3.3 The results on MIMIC-3 data (%). † indicates the improvement over the baseline. 9

3.4 The case study from MIMIC3-Full using CAML with and without average meta-labels. 10

doi:10.6342/NTU201903870

Chapter 1

Introduction

1.1 Motivation and Problem Description

Electronic health records (EHR) usually contain clinical notes, which are free-form text generated by clinicians during patient encounters, and a set of metadata diagnosis codes from the International Classification of Diseases (ICD), which represent the diagnoses and procedures in a standard way. ICD codes have a variety of usage, ranging from billing to predictive modeling of the patient state [1]. Automatic diagnosis prediction has been studied since 1998 [2]. [3] pointed out the main challenges of this task: 1) the large label space, with over 15,000 codes in the ICD-9 taxonomy, and over 140,000 codes in the newer ICD-10 taxonomies [4], and 2) noisy text, including irrelevant information, misspellings and non-standard abbreviations, and a large medical vocabulary. Several recent work attempted at solving this task by neural models [5, 3].

However, most prior work considered the output labels independently, so that the codes with few samples are difficult to learn [5]. Therefore, [3] proposed an attentional model to effectively utilize the textural forms of codes to facilitate learning. In addition to textual definitions of codes, the category domain knowledge may provide additional cues to allow the codes under same category to share parameters, so the codes with few samples can benefit from it. To effectively utilize the category knowledge from the ICD codes, this paper proposes several refined category losses

doi:10.6342/NTU201903870

and incorporate them into convolutional models and then evaluate the performance on both MIMIC-3 [6] and our internal datasets. The experiments on MIMIC shows that the proposed knowledge integration model significantly improves the previous methods and achieves the state-of-the-art performance, and the improvement can also be observed in our internal dataset. The idea is similar to the prior work [7], which considered the keyword hierarchy for information extraction from medical documents, but our work focuses on leveraging domain knowledge for clinical code prediction.

1.2 Main Contribution

Our contributions are three-fold:

- This paper first leverages external domain knowledge for diagnostic text understanding.
- The paper investigates multiple ways for incorporating the domain knowledge in an end-to-end manner.
- The proposed mechanisms improve all prior models and achieves the state-of-the-art performance on the benchmark MIMIC dataset.

1.3 Thesis Structure

In the following chapters, we elaborate the proposed models in Chapter 2 and shows the experiments and analysis in Chapter 3. Finally, the thesis will be concluded in Chapter 4.

Chapter 2

Proposed Approach

Given each clinical record in EHR, the goal is to predict the corresponding diagnostic codes with the external hierarchical category information. This task is framed as a multi-label classification problem. The proposed mechanism is built on the top of various convolutional models to further combine with the category knowledge. Below we introduce the previously proposed convolutional models and detail the mechanism that leverages hierarchical knowledge.

2.1 Convolutional Models

2.1.1 TextCNN

Let $x_i \in \mathbb{R}^k$ be the k-dimensional word embedding corresponding to the i-th word in the document, represented by the matrix $X = [x_1, x_2, ..., x_N]$, where N is the length of the document. TextCNN [8] applies both convolution and max-pooling operations in one dimension along the document length. For instance, a feature c_i is generated from a window of words $x_i, x_{i+1}, ..., x_{i+h}$, where h is the kernel size of the filters. The pooling operation is then applied over $c = [c_1, c_2, ..., c_{n-h+1}]$ to pick the maximum value $\hat{c} = \max(c)$ as the feature corresponding to this filter. We implement the model with kernel size = 3,4,5, considering different window sizes of words.

doi:10.6342/NTU201903870

2.1.2 Convolutional Attention Model (CAML)

Because the number of samples of each code is highly unbalanced, it is difficult to train each label with very few samples. To resolve this issue, the CAML model utilizes the descriptive definition of diagnosis codes, which additionally applies a per-label attention mechanism, where the additional benefit is that it selects the n-grams from the text that are most relevant to each predicted label [3].

2.2 Knowledge Integration Mechanism

Considering the hierarchical property of ICD codes, we assume that using the higher level labels could learn more general concepts and thus improve the performance. For instance, the definitions of ICD-9 codes 301.2 and 307.1 are “Schizoid personality disorder” and “Anorexia nervosa” respectively. If we only use the labels given by the dataset, they are seen as two independent labels; however, in the ICD structure, both 301.2 and 307.1 belong to the same high-level category “mental disorders”. The external knowledge shows that category knowledge provides additional cues to know code relatedness. Therefore, we propose four types of mechanisms that incorporate hierarchy category knowledge to improve the ICD prediction below.

2.2.1 Cluster Penalty

Motivated by [9], we compute two constraints to share the parameters of the ICD codes under the same categories. The between-cluster constraint, $\Omega_{between}$, indicates the total distance of parameters between mean of all ICD codes and the mean of each category.

$$\Omega_{between} = \sum_{k=1}^{K} \left\| \bar{\theta_k} - \bar{\theta} \right\|^2 , \tag{2.1}$$

where $\bar{\theta}$ is the mean vectors of all ICD codes, $\bar{\theta_k}$ is the mean vector of the k-th category. The within-cluster constraint, Ω_{within}, is the distance of parameters

doi:10.6342/NTU201903870

between the mean of each category and its low-level codes.

$$\Omega_{within} = \sum_{k=1}^{K} \sum_{i \in \mathcal{J}(k)} \left\| \theta_i - \bar{\theta_k} \right\|^2 , \tag{2.2}$$

where $\mathcal{J}(k)$ is a set of labels that belong to the k-th category. $\Omega_{between}$ and Ω_{within} are formulated as additional losses to enable the model to share parameters across codes with the same categories.

2.2.2 Multi-Task Learning

Considering that the high-level category can be treated as another task, we apply a multi-task learning approach to leverage the external knowledge. This model focuses on predicting the low-level codes, y_{low}, as well as its high-level category, y_{high}, individually illustrated in Figure 2.1.

$$y_{high} = W_{high} \cdot h + b_{high} \tag{2.3}$$

where $W_{high} \in \mathbb{R}^{N_{high} \times d}$, N_{high} means the number of high-level categories, and d is the dimension of hidden vectors derived from CNN.

2.2.3 Hierarchical Learning

We build a dictionary for mapping our low-level labels to the corresponding high-level categories illustrated in Figure 2.1. To estimate the weights for high-level categories, y_{high}, two mechanisms are proposed:

- Average meta-label: The probability of the k-th high-level category can be approximated by the averaged weights for low-level codes that belong to the k-th category.

$$y_{high} = \frac{1}{k} \sum y_{low}^{k} \tag{2.4}$$

- At-least-one meta-label: Motivated by [9], meta labels are created by examining whether any disease label for the k-th category has been marked as tagged,

doi:10.6342/NTU201903870

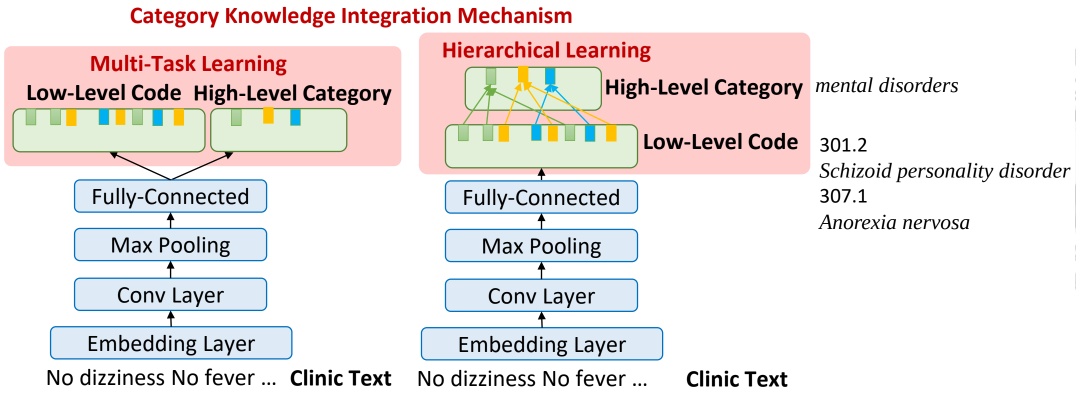

Figure 2.1: The architecture with the proposed category knowledge integration.

where the high-level probability is derived from the low-level probability of disease labels.

$$y_{high} = 1 - \prod_{k}(1 - y^{k}_{low}) \tag{2.5}$$

2.3 Training

The proposed hierarchy category knowledge integration mechanisms are built on top of the multi-label convolutional models, which treat each ICD label as a binary classification. The predicted values for high-level categories come from the proposed mechanisms. Considering that learning low-level labels directly is difficult due to the highly imbalanced label distribution, we add a loss term indicating the high-level category in order to learn the general concepts in addition to the low-level labels, and train the model in an end-to-end fashion. Note that the high-level loss is set as $loss_{high} = \Omega_{between} + \Omega_{within}$ for cluster penalty and the binary log loss for other methods.

$$loss = loss_{low} + \lambda \cdot loss_{high}, \tag{2.6}$$

where λ is the parameter to control the influence of the knowledge category.

Chapter 3

Evaluation

3.1 Experimental Setup

We evaluate our model on two datasets, one is the benchmark MIMIC-3 data and another is our internal dataset. MIMIC-3 [6] is a benchmark dataset, where the text and structured records from a hospital ICU. We use the same setting as the prior work [3], where 47,724 discharge summaries is for training, with 1,632 summaries and 3,372 summaries for validation and testing, respectively. Another medical dataset is obtained from an internal hospital, where each record includes narrative notes describing a patient's stay and associated diagnostic ICD-9 codes. There are total 1,495 ICD-9 codes in the data, and the distribution is highly imbalanced. Our data is noisy due to typos and different writing styles, where the OOV rate is 0.373 based on the large vocabulary obtained from PubMed and PMC. As shown in Table 3.1, our data is more challenging due to much shorter text inputs and higher OOV rate compared with the benchmark MIMIC-3 dataset. We split the whole set of 25,375 records into 17,762 as training, 2,537 as validation, and 5,076 as testing.

All models use skipgram word embeddings trained on PubMed[1] and PMC[2] [10]. We evaluate the model performance using metrics for the multi-label classification task, including precision at K, mean average precision (MAP), and micro-averaged,

[1] https://www.ncbi.nlm.nih.gov/pubmed

[2] https://www.ncbi.nlm.nih.gov/pmc

doi:10.6342/NTU201903870

	MIMIC-3		Internal
	Full	50	200
# training documents	47,424	8,067	17,762
mean length of texts	1,485	1,530	50.35
vocabulary size	51,917	51,917	25,654
OOV rate	0.137	0.137	0.373
# labels	8,922	50	200
mean number of labels	15.9	5.7	1.7

Table 3.1: Dataset comparison and statistics. From the full set of the internal data (1495 labels) to 200, only 6.0% of data points are discarded.

macro-averaged F1 and AUC.

3.2 Results

The baseline and the results of adding the proposed mechanisms are shown in Table 3.3. For MIMIC3-50, all proposed mechanisms achieve the improvement for almost all metrics, and the best one is from the hierarchical learning with average meta-label. The consistent improvement indicates that category knowledge provides informative cues for sharing parameters across low-level codes under the same categories. For MIMIC3-Full, our proposed mechanisms still outperform the baseline CNN model, and the best performance comes from the one with multi-task learning. The reason may be that multi-task learning has more flexible constraints compared with hierarchical learning, and it is more suitable for this more challenging scenario due to data imbalance. In addition, the proposed knowledge integration mechanisms using multi-task learning or hierarchical learning with average meta-label are able to improve the prior state-of-the-art model, CAML [3], demonstrating the superior capability and the importance of domain knowledge.

To further investigate the model effectiveness, we perform the experiments on the internal dataset in Table 3.2. Due to shorter clinical notes and higher OOV rate, this dataset is more challenging and the results are lower than the ones in MIMIC-3. Nevertheless, the proposed methods still improve the performance by integrating category knowledge using multi-task learning or hierarchical learning with average meta-label. In sum, our proposed category knowledge integration mechanisms are

doi:10.6342/NTU201903870

Data-200	Macro-F1	Micro-F1
CNN	7.6	39.8
+ Multi-Task	11.7†	41.6†
+ Hierarchical (avg)	9.2†	44.1†
CAML	6.2	42.6
+ Multi-Task	14.5†	44.7†
+ Hierarchical (avg)	18.4†	45.7†

Table 3.2: The results on internal data.

MIMIC3-50	P@1	P@3	P@5	MAP	Macro-F	Micro-F	Macro-AUC	Micro-A
CNN [5]	82.8	71.2	61.4	72.4	57.9	63.0	88.2	91.2
+ Cluster Penalty	83.5†	71.9†	62.4†	73.1†	58.3†	63.7†	88.5†	91.3†
+ Multi-Task	83.5†	71.3†	61.9†	72.5†	57.6	62.8	88.1	91.1
+ Hierarchical avg	84.5†	72.1†	62.4†	73.5†	58.6†	64.3†	88.9†	91.4†
at-least-one	83.4†	72.1†	62.4†	73.4†	58.5†	63.8†	88.4†	91.3†
MIMIC3-Full	P@1	P@3	P@8	P@15	Macro-F	Micro-F	Macro-AUC	Micro-A
CNN [5]	80.5	73.6	59.6	45.4	3.8	42.9	81.8	97.1
+ Cluster Penalty	80.9†	74.0†	59.5	45.2	3.3	40.5	82.1†	97.0
+ Multi-Task	82.8†	75.8†	61.5†	46.6†	3.6	43.9†	83.3†	97.3†
+ Hierarchical avg	79.0	73.1	59.2	45.2	4.3†	42.7	83.0†	97.1
at-least-one	82.1†	74.3†	59.7†	44.9	2.6	42.0	80.3	96.7
CAML [3]	89.6	83.4	69.5	54.6	6.1	51.7	88.4	98.4
+ Cluster Penalty	88.4	82.4	68.8	54.0	5.4	51.2	87.5	98.3
+ Multi-Task	89.7†	83.4	69.7†	54.8	6.9†	52.3†	88.8†	98.5†
+ Hierarchical avg	89.6	83.5†	70.9†	56.1†	8.2†	53.9†	89.5†	98.6†
at-least-one	89.4	83.3	69.5	54.8†	6.2†	51.7	88.3	98.4

Table 3.3: The results on MIMIC-3 data (%). † indicates the improvement over the baseline.

capable of improving the text understanding performance by combining the domain knowledge with neural models and achieve the state-of-the-art results.

3.3 Qualitative Analysis

From our prediction results, we find that our proposed mechanisms tend to predict more labels than the baseline models for both CNN and CAML. Specifically, our methods can assist models to consider more categories from shared information in the hierarchy. The additional codes often contain the right answers and sometimes are in the correct categories but not exactly matched. Moreover, our mechanisms have the capability of correcting the wrong codes to the correct ones which are under the same category. The appendix provides some examples for reference.

(a) Clinical notes admission date discharge date date of birth sex m service surgery allergies no drug allergy information on file attending first name3 lf chief complaint fall from bike major surgical or invasive procedure n a history of present illness 71m who was brought to the hospital1 ed after a fall from his bike past medical history seizure disorder bph spinal stenosis sleep apnea social history n a family history n a physical exam no brainstem reflexes pertinent results n a brief hospital course mr known lastname was admitted after a fall from his bicycle he was seen getting up from the accident and then collapsed shortly thereafter he then was noted to be in asystole when ems arrived the total amount of time the patient was in asystole is not known upon arrival to the ed he had regained a pulse a neuro exam was performed and he had no brainstem reflexes an mri confirmed a c2 level spinal cord injury and changes consistent with an anoxic brain injury the neob was contact name ni but due to unknown circumstances surrounding his cardiac arrest he did not meet donation criteria the family elected to withdraw care he was extubated and expired shortly thereafter medications on admission n a discharge medications n a discharge disposition expired discharge diagnosis odontoid fracture spinal cord injury respiratory failure discharge condition n a discharge instructions n a followup instructions n a
Baseline: 327.23 345.90 348.1 518.81 E826.1
Proposed: 327.23 33.24 345.90 348.1 401.9 518.81 600.00 780.39 780.57 806.01 96.04 96.6 96.71 96.72 E826.1
Ground truth: 288.50 345.90 348.1 356.9 427.5 518.81 600.00 780.57 806.01 807.01 96.04 96.71 E826.1
(b) Clinical notes admission date discharge date date of birth sex f service neurosurgery allergies wellbutrin lipitor flagyl levaquin attending first name3 lf chief complaint decline in mental status major surgical or invasive procedure angiogram with embolization of aneurysm history of present illness 63f who began to have mental staus decline dysarthria at home brought to needhan hospital1 where had head ct showing large l parietal hemorrhage was transferred to hospital1 for further treatment upon arrival there was concern for airway safety and she was intubated was reportedly moving all extremities prior to intubation past medical history ccy multiple ercp for biliary strictures benign breast tumor l aneurysm clip no deficit chronic autoimmune hepatitis on steroids osteoporosis social history married she smokes to cigarettes a day does not drink any alcohol she is a retired hospital3 manager she watches her grandson a couple times a week participates in book clubs walks and traveling family history thyroid disease is positive in the family as is rheumatoid arthritis her sister died at years of age of liver disease of unknown cause it is not known whether that also was autoimmune hepatitis there is also cirrhosis in the family physical exam hunt and doctor last name doctor last name gcs 6t e v 1t motor o t afeb bp hr r16 o2sats intubated sedated examined in ed just after intubated heent pupils 4mm reactive neck supple extrem warm and well perfused no c c e neuro no eye opening all to nox pertinent results cta redemonstrated ip ic sah worsened mass effect with 10mm rightward mls and effacement of the basal cisterns there is downward herniation aneurysms ruptured left mca partially calcified right m1 origin aneurysm the latter is amenable to coiling possible third small left mca trifurcation aneurysm await reformations brief hospital course pt was admitted to the icu for close neurological observation in the afternoon of admission the patient s mental status declined including loss of cough and gag brain test testing was initiated by the icu and concluded that she was brain dead preparations were made for organ donation per the families request medications on admission all flagyl levaquin statins wellbutrin discharge medications n a discharge disposition expired discharge diagnosis n a discharge condition deceased discharge instructions n a followup instructions n a name6 md name8 md md md number completed by
Baseline: 571.5 733.00 96.04 96.72
Proposed: 305.1 38.91 38.93 39.72 401.9 431 518.81 571.42 571.5 733.00 733.09 88.41 96.04 96.6 96.72
Ground truth: 276.3 276.8 348.4 348.89 38.93 39.72 430 571.42 733.00 88.41 96.04 96.71 V49.86 V58.65

Table 3.4: The case study from MIMIC3-Full using CAML with and without average meta-labels.

Chapter 4

Conclusion

This paper proposes multiple mechanisms using the refined losses to leverage hierarchical category knowledge and share semantics of the labels under the same category, so the the model can better understand the clinical texts even if the training samples are limited. The experiments demonstrate the effectiveness of the proposed knowledge integration mechanisms given the achieved state-of-the-art performance and show the great generalization capability for multiple datasets.

doi:10.6342/NTU201903870

doi:10.6342/NTU201903870

Bibliography

[1] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, “Doctor ai: Predicting clinical events via recurrent neural networks,” in Machine Learning for Healthcare Conference, pp. 301–318, 2016.

[2] L. R. de Lima, A. H. Laender, and B. A. Ribeiro-Neto, “A hierarchical approach to the automatic categorization of medical documents,” in Proceedings of the seventh international conference on Information and knowledge management, pp. 132–139, ACM, 1998.

[3] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, “Explainable prediction of medical codes from clinical text,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 1101–1111, 2018.

[4] W. H. Organization et al., “International statistical classification of diseases and related health problems: tenth revision-version for 2007,” http://apps.who. int/classifications/apps/icd/icd10online/, 2007.

[5] H. Shi, P. Xie, Z. Hu, M. Zhang, and E. P. Xing, “Towards automated icd coding using deep learning,” arXiv preprint arXiv:1711.04075, 2017.

[6] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, “Mimic-iii, a freely accessible critical care database,” Scientific data, vol. 3, p. 160035, 2016.

doi:10.6342/NTU201903870

[7] G. Singh, J. Thomas, I. Marshall, J. Shawe-Taylor, and B. C. Wallace, "Structured multi-label biomedical text tagging via attentive neural tree decoding," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2837–2842, Association for Computational Linguistics, 2018.

[8] Y. Kim, "Convolutional neural networks for sentence classification," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751, 2014.

[9] A. Nie, A. Zehnder, R. L. Page, A. L. Pineda, M. A. Rivas, C. D. Bustamante, and J. Zou, "Deeptag: inferring all-cause diagnoses from clinical notes in under-resourced medical domain," arXiv preprint arXiv:1806.10722, 2018.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, pp. 3111–3119, 2013.

doi:10.6342/NTU201903870