Weakly Supervised and Ontology Driven Representation Learning of Clinical Notes and Concepts

Ravikiran Chanumolu
International Institute of Information Technology, Hyderabad
ravikiran.chanumolu@research.iiit.ac.in

Mihir Shekhar
Microsoft India
mihir.shekhar@microsoft.com

Kamalakar Karlapalem
International Institute of Information Technology, Hyderabad
kamal@iiit.ac.in

***Abstract*— Clinical document representation learning is desirable in many applications such as information retrieval, patient clustering, and classification. Complexity of medical terminology and large size of clinical documents limit the performance of existing unsupervised representation learning techniques. While supervised approaches using deep learning overcome these barriers, the cost of labeling a large amount of data by medical experts is prohibitively high. In this paper, we propose a framework of learning representations of clinical notes using online medical ontologies and weak supervision from open-source tools. We further show that along with representations of clinical notes, our framework also gives meaningful embedding for underlying clinical concepts.**

I. Introduction

Clinical notes are a valuable resource to resolve diverse clinical queries. They provide elaborate patient information in addition to the hypothesis behind clinical reasoning and inference. Therefore learning numerical feature representations of clinical notes is desirable for several tasks which include patient understanding, classification, cohort identification, etc..

Neural network models have been widely applied for representation learning in recent years, and have delivered good results on a number of natural language processing (NLP) tasks such machine translation and question answering. Velupillai et al. [1], discusses the diverse applications of these models in clinical informatics. However, the authors note that procuring large annotated datasets for training these models is not feasible for most tasks in biomedical domain. This is due to the high cost of manual labeling. The well-known clinical NLP benchmarks such as Integrating Biology and the Bedside (i2b2) obesity comorbidity recognition, i2b2 smoking status detection, and the recent National NLP Clinical Challenges (https://n2c2.dbmi.hms.harvard.edu) have only hundreds of examples per phenotype which is insufficient for neural network training.

Transfer learning is one paradigm that has attracted widespread usage to tackle this issue. In this paradigm, the problem of insufficient data is addressed by means of pre-training a neural network on a large dataset and subsequently fine-tuning the model on a more specialized task.

In clinical domain researchers have exploited large unlabeled corpora such as PubMed and PMC to pretrain shallow representations of words and medical concepts using techniques such as word2vec. More recent works have made use of techniques like BioSent2vec [2] to learn representation of sentences and even short paragraphs. These works have been successful in dealing with complex medical vocabularies. However, large size of clinical notes limits their performance on document level clinical tasks. Few works in clinical informatics have also focused on learning representaions of clinical notes using autoencoders [3] [4]. However, we find that autoencoder based approaches fall short in capturing complex semantics and inter-relationship between medical entities. Alternatively, Escudi et al. [5], shows the merits of using an approach based on supervised deep learning for document level coding. They use a hybrid architecture made of a convolutional neural network on the text data, and a multi-layer perceptron on the structured data, both trained together to predict the ICD diagnoses labels related to each patient admission.

While, in this work we focus only on unstructured text data in patient EHR, we extend the work in [5] by jointly identifying ICD diagnoses codes, ICD procedure codes and medicine mentions associated with clinical notes. Additionally, an important argument made by the authors in [5] is that clinical encounters are always linked with corresponding ICD codes since they are required for billing purposes. Therefore, their work avoids cost and effort for labelling data to build machine learning models. While this is true in United States, in many countries like India, such labels associated with patient information are not available.

To mitigate this lack of annotated data, we make use of weak supervision. Researchers have utilized the weak supervision paradigm to leverage programmatically labeled datasets. In the biomedical domain, weak supervision has been used for clinical phenotyping and also to identify drug-drug interaction [6, 7]. In this paper, we employ open source clinical named-entity-recognition (NER) system to identify ICD codes and medicine mentions in clinical documents. We use these labels as a source of weak supervision to train our model. We empirically show that our approach gives semantically rich and robust embeddings.

We primarily focus on discharge summaries since they capture useful information about the patient which includes the reason for hospitalization, significant findings, treatment provided, patients discharge condition, etc.. Importantly, a large corpus of discharge summaries is freely accessible through MIMIC-III [8] database.

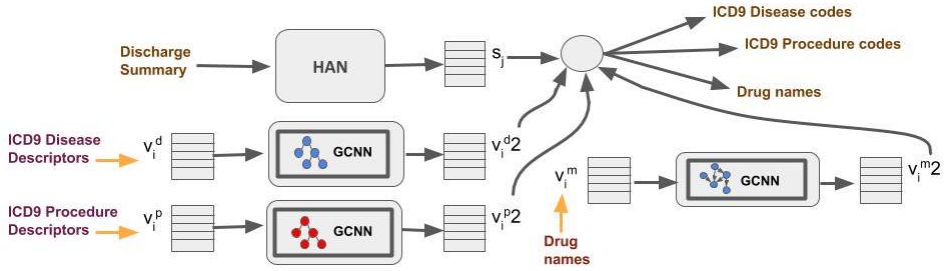

Fig. 1: Overview of our proposed architecture, weak-HAN-G

MIMIC-III [8] contains discharge summaries covering more than 5000 different diagnoses and procedure concepts. But a majority of these concepts occur infrequently. Data sparsity is a big challenge in our training scheme. To account for this, we expand upon the few-shot learning framework proposed in [9] to incorporate knowledge from online ontologies and websites.

To summarize, we propose a novel transfer learning approach to document modeling for clinical notes. During pertaining, our system jointly identifies ICD diagnoses codes, procedure codes, and medication mentions in discharge summaries. These labels are extracted using an off-the-shelf clinical NER system for supervision. Further, we make use of structured resources available online to leverage relationships between medical concepts. To demonstrate the quality and generic nature of the representations, we show results on two downstream tasks, namely **i2b2 Obesity Challenge** and **Mortality Prediction**. We compare our approach with strong baselines to emphasize the contributions of our work.

II. The Proposed Method

Figure1 shows the overall schematic of our architecture. The architecture has four main components. In the first component, we use Apache cTakes[10] to annotate discharge summaries by identifying the International Classification of Diseases(ICD) 9^{th} edition disease and procedure codes and medication. Second, we use the Hierarchical Attention Network (HAN) [11] to embed discharge summaries into a low dimensional space. Third, we use the description of the ICD9 codes and names of drugs to form a vector representation for each concept label. The label vectors are then, passed through a two-layer graph convolution neural network (GCNN) [12] to incorporate hierarchical information about the label space. Finally, the vectors returned from the GCNN are matched to the document vectors to jointly predict ICD9 disease and procedure codes and drugs in the input discharge summary. In the rest of the paper, we refer to our architecture as **weak-HAN-G**.

A. Weak Annotator

To annotate clinical concepts in discharge summaries, we adopted the NLP annotator and parser, Apache clinical Text Analysis and Knowledge Extraction System (cTakes) [10]. We use cTakes to extract ICD9 disease codes, ICD9 procedure codes and medication We use the named-entity recognition pipeline in cTakes to identify words or phrases in the text which align with disease, procedure and drug names. cTakes also checks if a given medical condition mentioned in the clinical note is relevant to the patient. We then map the identified words and phrases corresponding to diseases and procedures to ICD9 ontology. Wiegreffe et al. [13], showed that the error in ICD9 annotations given by cTakes is high. As the error in the weak labels increases, it makes proper model training difficult. To mitigate this problem, we reduce all codes given by cTakes to the more general categories i.e. the 3 digits before decimal point in an ICD9 code.

B. Embedding discharge summaries

The input layer in our model is an embedding lookup function that converts words in discharge summaries into d-dimensional word vectors using a lookup matrix $E \in \mathbf{R}^{V \times d}$ where V is the size of the vocabulary and d is the dimensionality of word vectors. We use BioWordVec [14] which provides word embeddings trained using PubMed and Medical Subject Headings (MeSH) databases. We then use HAN [11] which takes word vectors in discharge summary D_i as input and outputs a fixed-sized vector s_i. HAN models the hierarchical structure of documents and uses an attention mechanism that finds the most important words and sentences for the task under consideration.

C. GCNN network

We use the GCNN network as proposed in [9] to leverage the relationships between medical concepts i.e between ICD9 codes imposed by ICD9 taxonomy and between drugs imposed by drug ontology [1]. This allows for transferring learning from classes with sufficient data to classes with fewer data. To begin with, we compute feature vectors for each concept label by averaging the embeddings of words in its description.

$$\mathbf{v}_i = \frac{1}{|B|} \sum_{j \in B} \mathbf{w}_j \tag{1}$$

where $\mathbf{v}_i \in \mathbf{R}^d$ and B is the index set of words in the descriptor. For drug labels, starting with the label vectors

[1]In our drug ontology, drugs used to treat the same disease are adjacent to one another. In our experiments we found that this structure improves downstream task performance. Further, instead of using existing ontologies like RxNorm, Drugbank, NDFRT, etc. it was easier to enforce the above structure by scraping drugs.com where against each drug, the site lists all the alternative drugs that can be prescribed in its place.

$\mathbf{v}_i^m$, we combine the label vectors of its neighbors for the i^{th} label to form:

$$\mathbf{v}_{i1}^m = f\left(\mathbf{W}^m \mathbf{v}_i^m + \sum_{j \in \mathcal{N}_a^m} \frac{\mathbf{W}^{m,a}\mathbf{v}_j}{|\mathcal{N}_a^m|} + \mathbf{b}_m\right) \quad (2)$$

where $\mathbf{W}^m \in \mathbf{R}^{q \times d}$, $\mathbf{W}^{m,a} \in \mathbf{R}^{q \times d}$, $\mathbf{b}^m \in \mathbf{R}^q$ are parameters of the model that need to be optimized. f is the rectified linear unit function and N_a^m is the index set of the i-th label's adjacent drugs.

For disease and procedure labels however, the structure in label space is in the form of hierarchy. Therefore, we use different parameters to differentiate between parent and children adjacent nodes.

$$\mathbf{v}_{i1}^k = f\left(\mathbf{W}^k \mathbf{v}_i^k + \sum_{j \in \mathcal{N}_p^k} \frac{\mathbf{W}^{k,p}\mathbf{v}_j}{|\mathcal{N}_p^k|} + \sum_{j \in \mathcal{N}_c^k} \frac{\mathbf{W}^{k,c}\mathbf{v}_j}{|\mathcal{N}_c^k|} + \mathbf{b}_k\right) \quad (3)$$

where k represents disease, procedure concept categories and $\mathbf{W}^k \in \mathbf{R}^{q \times d}$, $\mathbf{W}^{k,p} \in \mathbf{R}^{q \times d}$, $\mathbf{W}^{k,c} \in \mathbf{R}^{q \times d}$ and $\mathbf{b}^k \in \mathbf{R}^q$ are all model parameters that need to be optimized.

We then attempt to match discharge summary vectors s_j with label vectors. Hence, to compare s_j to label vector v_{i1}^K, we transform it into,

$$\mathbf{e}_j^K = ReLU\left(\mathbf{W}_o^K \mathbf{s}_j + \mathbf{b}_o^K\right), \quad i = 1, \ldots, L^K \quad (4)$$

where K represents disease, procedure and drug concept categories. $\mathbf{W}_o \in \mathbf{R}^{d,q}$ and $\mathbf{b_o} \in \mathbf{R}^q$. This transformation is required to match the dimension to that of v_{i1}^K. Finally, the prediction for each label i is generated via

$$\hat{y}_i^K = sigmoid\left(\mathbf{e_j^K}^T \mathbf{v}_{i1}^K\right), \quad i = 1, \ldots, L^K \quad (5)$$

D. Joint Training

We train weak-HAN-G to simultaneously identify all three types of concepts i.e diseases, drugs and procedures using multi-label binary cross-entropy loss.

$$\mathcal{L}_T = \sum_{k \in K} \sum_{i=1}^{L^k} \left[-y_i^k \log\left(\widehat{y}_i^k\right) - \left(1 - y_i^k\right) \log\left(1 - \hat{y}_i^k\right)\right] \quad (6)$$

where $y_i^k \in \{0, 1\}$ is the ground truth for the i_{th} concept label from k_{th} category and $\widehat{y}_i^k$ is our sigmoid score for the i_{th} label.

III. Experiments and Results

Towards the goal of understanding and improving clinical text representations for downstream prediction performance, we evaluate them on the below outlined tasks. Further, to emphasize on the quality of representations, we opt for minimal fine-tuning using a linear classifier (logistic regression) on the downstream tasks.

A. Baseline Methods

We compare weak-HAN-G with **text variational auto-encoder (text-VAE)** proposed in [15] which was shown to learn generic text representations. We also compare with **BioSentVec** [2]. We take the average of the sentence level embeddings derived from BioSentVec to get document representation. Further, we use medical concept embeddings given by [16] to construct document embedding. We average the embeddings for concepts - in each concept category i.e disease, procedures, and medication - found in the discharge summary by cTakes and concatenate them to get document representation. We refer to this representation as **Concept2Vec**^{a}. Similarly, we construct document representation using the concept embeddings given by weak-HAN and call it **Concept2Vec**^{b}. Finally, for each downstream task, we also report the performance of weak-HAN-G without using GCNN and refer to this model as **weak-HAN**.

B. i2b2 Obesity Challenge

We evaluate our representation on the i2b2 2008 Obesity Challenge [17]-intuition task. The publicly available dataset contains approximately 1,230 clinical notes and fifteen frequently occurring obesity co-morbidities. To align with this competition, we report macro averaged precision (**MP**), recall (**MR**) and F1-score (**MF**). Due to limited space, we will only report the average performance across the 16 targets.

C. Mortality Prediction

Mortality Prediction plays an important role in evaluating the seriousness of the patient condition. For this task, we construct the dataset using MIMIC-III critical care database. From the retrieved patients, we filter neonates and patients with more than one hospital admission as done in [18]. Our focus in this work is on discharge summaries. Therefore, we also remove patients who died in the hospital as in-hospital death of a patient is directly indicated in a discharge report making the task trivial. The resulting dataset contains 22475 records. We split the dataset into 80-20% as training and test subsets. We create multiple tasks within mortality prediction based on the time frame within which a patient dies post discharge: within 1 month (1-M), 3 months(3-M), 6 months(6-M) and 1 year (1-Y). To account for the class imbalance [2] in these tasks, we use the area under the ROC curve (**AUC-ROC**) as performance metric.

D. Results and Discussion

Table1 shows the performance of the learned clinical note and concept representations on i2b2 obesity challenge. Further, the performance on mortality prediction tasks can be seen in Table2. Our results show the weak-HAN-G performs consistently better than self-supervised representation techniques like BioSent2Vec and text-VAE. Further, we note that across all tasks, our model performs better with GCNN

[2]positive-negative class ratio for each task: 1-M (0.051), 3-M (0.092), 6-M (0.13), 1-Y (0.17)

Method	MP	MR	MF
text-VAE	82.7	49.6	52.4
BioSent2Vec	87.7	53.6	56.8
Concept2Vec^a	90.8	56.9	60.3
Concept2Vec^b	91.3	58.0	61.2
Solt et al. [19]	74.8	**65.7**	67.4
Yao et al. [20]			**67.6**
weak-HAN	90.6	59.1	61.4
weak-HAN-G	**92.6**	60.0	62.7

TABLE I: Obesity Challenge

Method	1-M	3-M	6-M	1-Y
text-VAE	57.86	56.53	55.98	55.08
BioSent2Vec	85.30	82.86	81.61	80.07
Concept2Vec^a	86.86	85.60	85.63	85.89
Concept2Vec^b	88.91	87.03	86.87	86.88
weak-HAN	88.92	87.72	86.93	86.91
weak-HAN-G	**90.18**	**88.12**	**88.42**	**88.46**

TABLE II: Mortality Prediction

as compared to without GCNN (weak-HAN). This shows the usefulness of incorporating structure in label space with GCNN when the labels are inaccurate which is the case with weak supervision.

From Table1 we see that our approach shows comparable performance to the top i2b2 system [19]. In [19], Solt et al. uses strong handcrafted features capable of identifying important phrases showing presence or absence of obesity and related comorbidities. Yet, we achieve competitive results with minimal fine-tuning of the representations given by weak-HAN-G. Our results are also comparable to the current state of the art [20] which also used Solt et al.'s work to identify trigger phrases and predict classes with very few examples. Additionally, in Table1 and Table2 we also compare the note representation constructed using concept embeddings given by our model (Concept2Vec^b) and concept embeddings given by [16] (Concept2Vec^a). In [16], Choi et al. directly applies skip-gram to structured longitudinal visit records to learn the representation of medical concepts. For training, the medical codes annotated against each patient visit in MIMIC-III database are used. Even though we use cTakes [10] output instead of true labels for model training, concept embeddings given by weak-HAN-G performs better on i2b2 obesity challenge and mortality prediction.

Lastly, Weak-Han-G is trained by simultaneous prediction of diagnoses codes, procedure codes and identification of medicine names in discharge summaries. In this section, we explore the usefulness of this joint training. Table2 and Table3 show the performance of our model trained separately to predict ICD9 diagnoses codes (**weak-HAN-G$^{\mathbf{d}}$**), ICD9 procedure codes (**weak-HAN-G$^{\mathbf{p}}$**) and identify medicine names (**weak-HAN-G$^{\mathbf{m}}$**). The last row in both tables reiterates the results obtained using joint training. We observe that joint training produces significantly better results compared to individual training schemes.

IV. Conclusion

In this paper, we proposed **weak-HAN-G**, an integrated approach that jointly identifies medical concepts using weak supervision and simultaneously incorporates hierarchical relationships in medical concept space. Results on downstream predictive tasks demonstrate that our paradigm learns semantically rich representations of medical concepts and text. In the future, we would like to include more weak signals such as relations between medical problems, tests and procedures.

Method	1-M	3-M	6-M	1-Y
weak-HAN-G$^{\mathbf{d}}$	86.01	85.87	84.94	84.95
weak-HAN-G$^{\mathbf{p}}$	80.69	80.35	79.06	79.57
weak-HAN-G$^{\mathbf{m}}$	87.53	85.86	85.92	85.15
weak-HAN-G	90.18	88.12	88.42	88.46

TABLE III: Mortality Performance : Joint vs. Separate training

Method	MP	MR	MF
weak-HAN-G$^{\mathbf{d}}$	86.27	57.72	60.22
weak-HAN-G$^{\mathbf{p}}$	79.54	48.57	50.64
weak-HAN-G$^{\mathbf{m}}$	79.34	53.08	55.08
weak-HAN-G	92.60	60.04	62.73

TABLE IV: Obesity Performance : Joint vs. Separate training

References

[1] S. Velupillai, H. Suominen, M. Liakata, A. Roberts, A. D. Shah, K. Morley, D. Osborn, J. Hayes, R. Stewart, J. Downs, *et al.*, "Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances," *Journal of biomedical informatics*, vol. 88, pp. 11–19, 2018.

[2] Q. Chen, Y. Peng, and Z. Lu, "Biosentvec: creating sentence embeddings for biomedical texts," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 1–5, IEEE, 2019.

[3] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," *Scientific reports*, vol. 6, p. 26094, 2016.

[4] M. Sushil, S. Šuster, K. Luyckx, and W. Daelemans, "Patient representation learning and interpretable evaluation using clinical notes," *Journal of biomedical informatics*, vol. 84, pp. 103–113, 2018.

[5] J.-B. Escudié, A. Saade, A. Coucke, and M. Lelarge, "Deep representation for patient visits from electronic health records," *arXiv preprint arXiv:1803.09533*, 2018.

[6] D. Li, S. Liu, M. Rastegar-Mojarad, Y. Wang, V. Chaudhary, T. Therneau, and H. Liu, "A topic-modeling based framework for drug-drug interaction classification from biomedical text," in *AMIA Annual Symposium Proceedings*, vol. 2016, p. 789, American Medical Informatics Association, 2016.

[7] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E. J. Atkinson, S. Amin, and H. Liu, "A clinical text classification paradigm using weak supervision and deep representation," *BMC medical informatics and decision making*, vol. 19, no. 1, p. 1, 2019.

[8] 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.

[9] A. Rios and R. Kavuluru, "Few-shot and zero-shot multi-label learning for structured label spaces," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2018, p. 3132, NIH Public Access, 2018.

[10] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[11] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480–1489, 2016.

[12] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
[13] S. Wiegreffe, E. Choi, S. Yan, J. Sun, and J. Eisenstein, "Clinical concept extraction for document-level coding," *arXiv preprint arXiv:1906.03380*, 2019.
[14] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, "Biowordvec, improving biomedical word embeddings with subword information and mesh," *Scientific data*, vol. 6, no. 1, p. 52, 2019.
[15] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.
[16] Y. Choi, C. Y.-I. Chiu, and D. Sontag, "Learning low-dimensional representations of medical concepts," *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 41, 2016.
[17] Ö. Uzuner, "Recognizing obesity and comorbidities in sparse data," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 561–570, 2009.
[18] Y. Si and K. Roberts, "Deep patient representation of clinical notes via multi-task learning for mortality prediction," *AMIA Summits on Translational Science Proceedings*, vol. 2019, p. 779, 2019.
[19] I. Solt, D. Tikk, V. Gál, and Z. T. Kardkovács, "Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 580–584, 2009.
[20] L. Yao, C. Mao, and Y. Luo, "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks," *BMC medical informatics and decision making*, vol. 19, no. 3, p. 71, 2019.