Comparing General and Locally-Learned Word Embeddings for Clinical Text Mining

Jidapa Thadajarassiri
Data Science Program
Worcester Polytechnic Institute
Worcester, USA
jthadajarassiri@wpi.edu

Cansu Sen
Computer Science Department
Worcester Polytechnic Institute
Worcester, USA
csen@wpi.edu

Thomas Hartvigsen
Data Science Program
Worcester Polytechnic Institute
Worcester, USA
twhartvigsen@wpi.edu

Xiangnan Kong
Computer Science Department
Worcester Polytechnic Institute
Worcester, USA
xkong@wpi.edu

Elke Rundensteiner
Computer Science Department
Worcester Polytechnic Institute
Worcester, USA
rundenst@wpi.edu

***Abstract*—Dense vector-representations of words, referred to as word embeddings, capture word relations differently depending on the type and the size of the corpus they are trained on. Choosing which word embeddings to use poses a problem when applying machine learning on clinical text which contains many specialized words. In this work, we explore the effects of different embedding sources for clinical text classification using various cohort sizes on three medical prediction tasks: Clostridium Difficile infections, MRSA infections, and in-hospital mortality. We compare three embedding sources: pre-trained embeddings from large and general corpora, pre-trained embeddings from large and domain-related corpora, and locally-learned embeddings trained on the task-specific training data. We experiment with several cohort sizes ranging from 20 patients to 2,500 patients. Our results indicate that pre-trained domain-related embeddings are superior for medium-sized cohorts greater than 150 patients, while locally-learned embeddings become increasingly competitive as cohort size grows.**

***Index Terms*—Word embeddings, Clinical text classification**

I. INTRODUCTION

Clinical notes contain valuable information that quantifies health statuses of patients. These documents, written by clinicians, contain patient-specific information that is only described through expert observations. Building machine learning models that can extract such content when performing prediction tasks allows for automatic processing of vast amounts of unstructured text to support clinical decision making [4], [5]. When using machine learning models for text classification, words first need to be transformed into numerical representations. *Word embeddings* have recently become a popular approach to this challenge as they capture semantic relationships between words. However, with the rising popularity of publicly-available word embeddings, a natural question arises: *Should we train our own word embeddings or use a set of publicly-available embeddings?*

The success of word embeddings in capturing important relationships between words is affected by both the size and the type of the text they are trained on. Utilizing which embeddings is commonly considered across downstream textual-related tasks [10], [11]. Training embeddings on a large corpus (*e.g.*, Wikipedia) provides vast amounts of information, allowing for general embeddings that are broadly useful across many tasks in many domains. However, such generality comes at the cost of ignoring word meanings that are specific to domains (*e.g.*, "patient" may typically be an adjective in Wikipedia text, but a noun in clinical text). Another option is one of the few publicly-available embedding sources that are trained on medical text (*e.g.*, medical journal abstracts), partially solving the problem of generality in large popular resources. However, writing styles and guidelines can still differ between regions and facilities. Thus, training word embeddings on task-specific data has potential to capture highly-specialized word meanings. However, a common challenge in the clinical domain is *small cohort sizes*, which make it difficult to train complex machine learning models, such as those that generate word embeddings. The correct choice of embedding source remains a challenging problem present in any machine learning for healthcare task that involves text.

In this study, we explore the effects of embedding sources and cohort sizes on model performance in three clinical prediction tasks - in-hospital mortality, Clostridium Difficile (C. Diff.) infections, and Methicillin-resistant Staphylococcus aureus (MRSA) infections. Our results indicate that the best choice for embedding source often depends on the available cohort size, demonstrating that there is no one-size-fits-all solution to embedding choice.

II. METHODS

A. Data Extraction

MIMIC-III [3] is a publicly available database containing de-identified Electronic Heath Records of over $58,000$ patient admissions between 2001 and 2012 in the intensive care units (ICU) of Beth Israel Deaconess Medical Center. A major component of MIMIC-III is the vast number of clinical notes, recorded by physicians throughout each patient's stay. We extract cohorts from MIMIC for the following prediction tasks:

Mortality: In-hospital mortality prediction is a common baseline task in machine learning for healthcare. From the *Admissions* table, we extract patients who perish during their stay in the ICU, indicated by *hospital expire flag* = 1. Of 5,854 patients who have this flag set, 5,000 have notes taken prior to their death.

C. Diff.: Clostridium Difficile infection is a dangerous in-hospital acquired bacterial infection which can be fatal in patients with weakened immune systems [8]. Using the microbiology test *80139*, we extract 1,035 patients who have confirmed cases of C. Diff. during their stays.

MRSA: Methicillin-resistant Staphylococcus aureus, a pervasive antibiotic-resistant bacteria, has been a leading cause of in-hospital infections and is a growing concern [9]. We again use the *Microbiology Events* table to extract patients who test positive for MRSA, using the code *80293*. There are 1,240 MRSA patients with note events.

For all three conditions, negative cohorts (with equal sizes as their corresponding positive cohorts) are randomly sampled from the patients with no records of each corresponding condition. Following data extraction, a patient's notes are combined into one document. Finally, we remove all punctuation and convert all letters to lowercase.

B. Word Embeddings

Word embeddings are prevalently used to transform words into numeric vectors. These vectors preserve relationships between words by considering the context in which each word appears. Learning word embeddings is time-consuming, especially when the source text is large. For this reason, many pre-trained word embeddings are published to help researchers bypass the learning process and commonly used as the baselines for text classification [10], [11]. However, vector representations can differ dramatically between embedding sources depending on both the size and content of the training corpora. Thus, locally-learned embedding is another appealing choice to use for clinical machine learning.

In this work, we study the choice between these three embeddings sources:

1) Global embeddings are pre-trained from general corpus such as Wikipedia, Twitter, or news [6], [7]. Word vectors from these pre-trained embeddings show high performance in capturing word meaning and relation since they are trained on large corpora with a vast amount of training words. However, the general corpora they are trained on may not preserve domain-related meanings of words. For example, the vector representation of "patient" in *global embeddings* is closer to the vector for "tolerate".

2) Domain embeddings are pre-trained from a large domain-related corpus such as PubMed, PMC texts, or biomedical literatures. Compared to *global embeddings*, these are trained on a more related vocabulary to clinical notes. Thus, *domain embeddings* could better capture the medical meaning of words. For example, the vector representation of "patient" in *domain embeddings* is closer to the vector for "sufferer".

3) Local embeddings are learned based on a specific cohort of patients which will be classified by a machine learning algorithm (*e.g.*, mortality prediction). Since *local embeddings* are learned directly from the task data, the task-dependent relations of words are preserved. However, if a cohort size that used to train is too small, *local embeddings* might not have enough data to create a good quality embedding in capturing this implicit relations.

We use the following sources in our experiments: 1) *Global embeddings*: pre-trained embedding learned from Wikipedia 2014 and Gigaword 5 [6]. 2) *Domain embeddings*: pre-trained embedding learned from PubMed and PMC texts [1]. 3) *Local embeddings*: learned on our local cohort data. The embedding dimensions are 200 in all three sources.

C. Word2Vec Skip-Gram Model

To train *local embeddings*, we use the Word2Vec Skip-Gram model [2] with minimum-count = 15, negative-sampling = 15 and contextual-window-size = 1.

Skip-Gram assumes that words appear in similar context are likely to have similar meanings. Thus, pairs of context-target words are created from an input corpus. The model predicts context words from a given target word, i.e., for a target word w_t and a given number of contextual-window-size c, Skip-Gram will predict the probability of c words before and c words after w_t. $\mathbf{x}_{w_t}$, a one-hot encoding of w_t, is fed into a 2-layer neural network. The final weight matrix of the 1^{st} layer is the desired word embedding.

Let $\mathbf{w} = \{w_1, w_2, ..., w_T\}$ be the T training words and d be the number of hidden units. The size of weight matrices of 1^{st} layer ($\mathbf{W}$) and 2^{nd} layer ($\mathbf{W}'$) are $T \times d$ and $d \times T$ respectively. The t^{th} row of $\mathbf{W}$, which is the d-dimensional vector representation $\mathbf{v}_{w_t}$ of w_t, is copied to the hidden layer.

$$\mathbf{h}_t = \mathbf{x}_{w_t}^T \mathbf{W} \text{ and } \mathbf{u} = \mathbf{W}'^T \mathbf{h}_t^T \quad (1)$$

where $u_j \in \mathbf{u}$ is the predicted score of word w_j being the context word. A softmax activation function is lastly applied to obtain the probability distribution in the output layer:

$$p(w_j|w_t) = \hat{y}_j = \frac{exp(u_j)}{\sum_{j'=1}^{T} exp(u_{j'})} \quad (2)$$

where $\hat{y}_j$ is the probability of w_j being the context word.

Model is trained using the objective function:

$$\max \frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{j=t-c \\ j \neq t}}^{t+c} \log p(w_j|w_t) \quad (3)$$

D. Experiments

We empirically address the effects of three embedding sources (*global, domain, local*) on different cohort sizes.

1) Experimental Setting: From the cohorts extracted from MIMIC, 10% of the data is set aside to use as the test set. This leads to test sets of size 300 for mortality, 200 for C. Diff., and 200 for MRSA. Remaining data is used to form cohorts ranging from a very small size with only 20 patients to a large size with 2,500 patients. Constrained

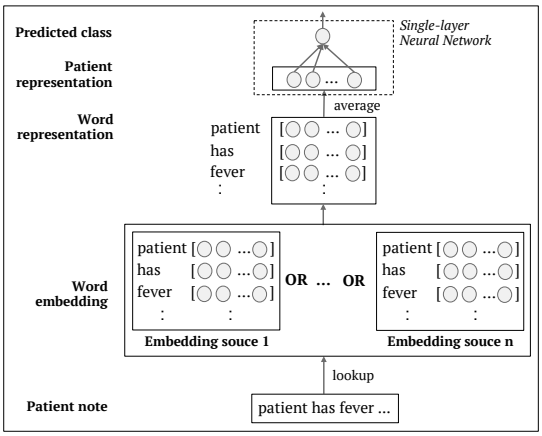


Fig. 1. Patient classification scheme used in our empirical evaluation.

by the total amount of available data, we experiment with cohort sizes ranging from 20 to 1,000 patients for C. Diff. and MRSA infections, and from 20 to 2,500 for mortality. We repeat experiments 10 times for each cohort size in each task by sampling with replacement. The model performance are evaluated by comparing accuracy computed on the test set.

2) Patient Classification: Our patient classification methodology is outlined in Fig. 1. For each embedding source (*global*, *domain*, and *local*), we transform words in each patient's clinical note to vector representations (a 200-dimensional dense vector for each word). Word vectors within a note are then averaged element-wise to form one vector representation per patient, which is then input to the classifier.

A single-layer neural network with a sigmoid activation function serves as our classifier. This model generates probabilities of each patient representation being labeled as each class. We pick the maximum class probabilities to be the predicted label. The classifier is trained using stochastic gradient descent with learning rate = 0.01 for 20,000 epochs.

III. Results and Discussion

Vocabulary coverage by embedding source. Training set vocabulary sizes and how much of the vocabulary is covered by each embedding source are presented in Table I for all cohort sizes. These are average numbers from 10 patient folds. Although *local embedding* is learned from the training data, infrequent words are removed by the minimum-count parameter, which controls how many times a word must appear in order for an embedding to be generated. This leads to different vocabulary appearing in the training set and the *local embeddding*. Among the three embedding sources, we observe that *domain embedding* has the highest vocabulary coverage rate in all cohort sizes. It is due to *domain embedding* being trained on a large, related corpora. *Local embedding* with our parameter setting covers, on average, only one-fifth of the vocabulary in the training data. The ratio of vocabulary coverage by *global* and *domain embeddings* interestingly decreases gradually when the cohort sizes grow. This shows that the ratio of specialized words in clinical notes that do not appear in external corpus increases with cohort size.

TABLE I
Average numbers of unique words from 10 experiments of each cohort size on three data sets.

(a) Mortality

Cohort size	Vocabulary size	Vocabulary coverage		
		global emb.	domain emb.	local emb.
20	7 k	5 k (76%)	5 k (82%)	1 k (16%)
50	11 k	7 k (70%)	8 k (77%)	2 k (22%)
150	18 k	11 k (60%)	13 k (69%)	4 k (24%)
250	24 k	13 k (55%)	15 k (64%)	6 k (24%)
500	33 k	16 k (48%)	19 k (57%)	8 k (24%)
750	40 k	18 k (44%)	22 k (54%)	9 k (23%)
1000	47 k	19 k (41%)	24 k (50%)	11 k (22%)
1250	53 k	20 k (38%)	26 k (48%)	12 k (22%)
1500	59 k	21 k (37%)	27 k (46%)	13 k (22%)
1750	64 k	22 k (35%)	28 k (44%)	14 k (21%)
2000	68 k	23 k (34%)	30 k (43%)	14 k (21%)
2250	73 k	24 k (33%)	31 k (42%)	15 k (21%)
2500	77 k	24 k (32%)	31 k (41%)	15 k (20%)

(b) C. Diff. Infection

Cohort size	Vocabulary size	Vocabulary coverage		
		global emb.	domain emb.	local emb.
20	7 k	5 k (77%)	6 k (83%)	1 k (20%)
50	11 k	7 k (70%)	8 k (78%)	3 k (24%)
150	19 k	11 k (60%)	13 k (68%)	5 k (26%)
250	24 k	13 k (55%)	16 k (64%)	6 k (26%)
500	34 k	16 k (47%)	19 k (57%)	9 k (25%)
750	42 k	18 k (43%)	22 k (53%)	10 k (25%)
1000	48 k	20 k (40%)	24 k (50%)	12 k (24%)

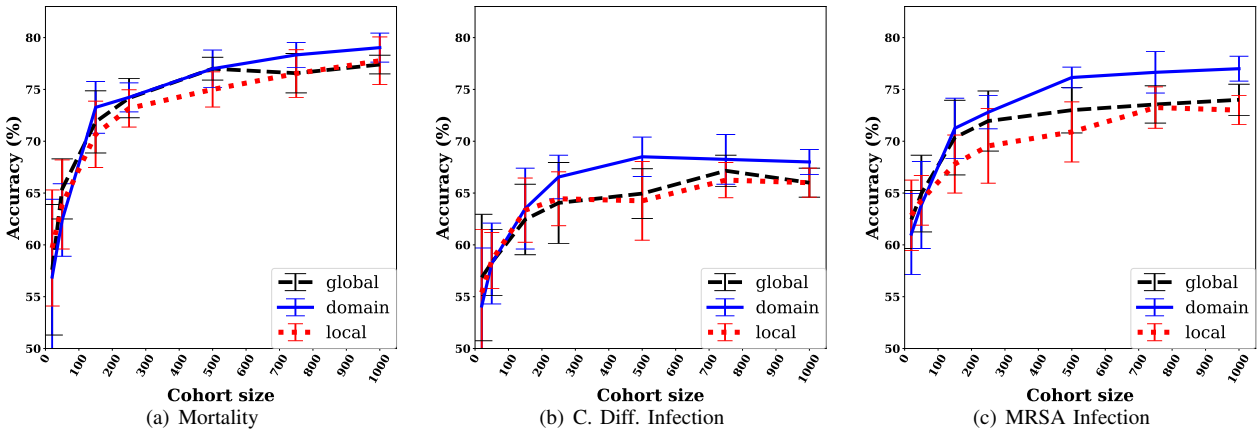
(c) MRSA Infection

Cohort size	Vocabulary size	Vocabulary coverage		
		global emb.	domain emb.	local emb.
20	6 k	5 k (80%)	6 k (93%)	1 k (23%)
50	10 k	7 k (73%)	8 k (86%)	2 k (25%)
150	16 k	10 k (64%)	13 k (81%)	4 k (25%)
250	20 k	12 k (59%)	16 k (77%)	5 k (25%)
500	28 k	15 k (53%)	19 k (69%)	7 k (26%)
750	35 k	17 k (48%)	22 k (64%)	9 k (25%)
1000	39 k	18 k (45%)	24 k (62%)	10 k (25%)

Classification performance. Classification accuracies are presented in Fig. 2 and 3. Accuracy consistently improves as the cohort size increases, since larger cohort sizes provide more training data that allows more information for modeling. Overall, the best performing embedding changes by cohort size. This is because the model utilizes information containing in each embedding differently when the available number of words changes.

Small cohort sizes refer to the datasets that contain less than 20k words (patients $\leq$ 150 in our experiment). Fig. 2 shows that there is no consistently optimal embedding for these cohort sizes. For the smallest cohort size of 20 patients, *domain embedding* has the lowest performance in all tasks. In this case, *global embedding* performs better by transferring external knowledge to enhance the task performance compared to *domain embedding*. However, when cohort size increases to 150 patients, the knowledge transferred from *domain embedding* becomes more beneficial than *global embedding*.

Medium cohort sizes refer to the datasets that contain between 20k and 50k words (150 $<$ patients $\leq$ 1,000 in our experiment). We consistently observe that using *domain embedding* performs equally or better than the others as shown



(a) Mortality (b) C. Diff. Infection (c) MRSA Infection

Fig. 2. Classification results for small and medium cohorts on three data sets.

in Fig. 2. Although *global embedding* overall shows a slightly better performance than *local embedding* on mortality and MRSA infection tasks, both embeddings perform similarly on C. Diff. infection task. It shows that medium cohort sizes are still not large enough to train a good *local embedding*. Instead, cohorts in this range highly benefit from the knowledge transfer from *domain embedding*.

Large cohort sizes refer to the datasets that contain greater than 50k words (patients $>$ 1,000 in our experiment) which is only investigated for mortality task due to the data availability. As shown in Fig. 3, *domain embedding* still performs well while *local embedding* becomes more comparable to the *domain embedding* as cohort size grows, even with less than half vocabulary coverage. Additionally, *local embedding* now performs, on average, better than *global embedding*, proving *local embedding* benefits from the larger training data greatly.

IV. CONCLUSIONS

In clinical text classification, choosing word embeddings is selective among available cohort sizes. Pre-trained *global embeddings* from general corpora are not a universally good choice when working with clinical text. According to our results, we recommend pre-trained *global embedding* when working with a very small cohort size while *domain embedding* consistently outperforms for medium cohort sizes. For a large cohort size, while using *domain embedding* is still a good choice, learning *local embeddings* quickly becomes highly competitive and we speculate that as cohort sizes continue to grow, performance of *local embedding* will improve as well.

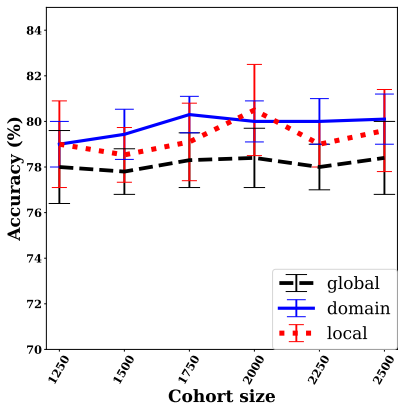


Fig. 3. Classification results for large cohorts on mortality data.

Embeddings learned from different corpora preserve different knowledge and characteristics of words. As a future work, we plan to integrate multiple embeddings acquiring the best parts of each individual option.

REFERENCES

[1] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou, "Distributional semantics resources for biomedical text processing," *Proceedings of Languages in Biology and Medicine*, pp. 3944, 2013.

[2] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, pp. 3111-3119, 2013.

[3] W. Johnson, J. Pollard, L. Shen, L. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, A. Celi, and G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, 2016.

[4] S. Liu, B. Tang, Q. Chen, and X. Wang, Drug-Drug Interaction Extraction via Convolutional Neural Networks, *Computational and Mathematical Methods in Medicine*, vol. 2016, Article ID 6918381, 2016.

[5] K. L. C. Barajas, and R. Akella, "Dynamically Modeling Patient's Health State from Electronic Medical Records: A Time Series Approach," *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[6] J. Pennington, R. Socher, C. and C. Manning, "Glove: Global vectors for word representation," *Proceedings of the Empiricial Methods in Natural Language Processing*, pp. 1532-1543, 2014.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *ICLR Workshop*, 2013.

[8] C. T. Evans, and N. Safdar, "Current Trends in the Epidemiology and Outcomes of Clostridium difficile Infection," *Clin. Infect. Dis.*, vol. 60 (suppl 2), pp. s66-S71, 2015.

[9] C. L. Maree, R. S. Daum, S. Boyle-Vavra, K. Matayoshi, and L. G. Miller, "Community-associated Methicillin-resistant Staphylococcus aureus Isolates and Healthcare-Associated Infections," *Emerging Infectious Diseases*, vol. 13(2), pp. 236242, 2007.

[10] V. Major, A. Surkis, and Y. Aphinyanaphongs, "Utility of general and specific word embeddings for classifying translational stages of research," *AMIA Annu Symp Proc. 2018*, pp. 1405-1414, 2018.

[11] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, and H. Liu, "A comparison of word Embeddings for the biomedical natural language processing," *Journal of Biomedical Informatics.* Vol. 87, pp. 1220. 2018.