Comparing High Dimensional Word Embeddings Trained on Medical Text to Bag-of-Words For Predicting Medical Codes

Comparing High Dimensional Word Embeddings Trained on Medical Text to Bag-of-Words For Predicting Medical Codes

Vithya Yogarajan[1][0000−0002−6054−9543], Henry Gouk[2][0000−0002−0924−2933], Tony Smith[0000−0003−0403−7073][1], Michael Mayo[1], and Bernhard Pfahringer[1][0000−0002−3732−5787]

[1]Department of Computer Science, University of Waikato
[2]School of Informatics, University of Edinburgh
vy1@students.waikato.ac.nz

Abstract. Word embeddings are a useful tool for extracting knowledge from the free-form text contained in electronic health records, but it has become commonplace to train such word embeddings on data that do not accurately reflect how language is used in a healthcare context. We use prediction of medical codes as an example application to compare the accuracy of word embeddings trained on health corpora to those trained on more general collections of text. It is shown that both an increase in embedding dimensionality and an increase in the volume of health-related training data improves prediction accuracy. We also present a comparison to the traditional bag-of-words feature representation, demonstrating that in many cases, this conceptually simple method for representing text results in superior accuracy to that of word embeddings.

Keywords: word embeddings, binary classification, machine learning for health

1 Introduction

Recent years have seen significant growth in the use of machine learning techniques to better understand health care and improve quality of service —primarily due to the increase in the availability of large quantities of electronic health records (EHRs). Secondary analysis of EHRs has the potential to improve a variety of healthcare aspects, including patient care, medical outcomes, surgical outcomes, risk management, clinical decision support and medical diagnoses. However, the free-form text content of EHRs poses many challenges not typically addressed by conventional natural language processing (NLP). Due to the complexity and variations presented in the data, and the legal and ethical aspects associated with the use of this data, the analysis of EHRs has not seen benefits as significant as those enjoyed by more common application domains.

Word embeddings are often used for solving problems that involve extracting high-level knowledge from free-form text data. These word embeddings are

typically trained on corpora composed of general language, such as archives of English Wikipedia, that are unlikely to be representative of the way language is used in EHRs and other related healthcare data sources. Using embeddings trained on general text for tasks that involve specialised language results in a domain shift, which will typically cause suboptimal performance [21]. Ideally, if we want to classify documents derived from EHRs then we should train the embeddings on a large collection of free-form text extracted from EHRs. For various legal and ethical reasons, this is not possible: the collections of health records available for research purposes are not large enough to train high quality word embeddings.

The contributions of this work are two-fold: (i) we demonstrate that training word embeddings on health-related corpora provides an increase in accuracy compared to embeddings trained on general text—particularly when the dimesionality of the embeddings is increased; (ii) it is shown that the bag of words representation is often as effective, if not more so, than dense word embeddings when applied to medical code prediction.

2 Related Work

Many NLP tasks, health-related or otherwise, use word embeddings to represent text data, due to their ability to encode semantic similarity between words. Word embeddings represent a single word or sub-word as a vector based on the context in which it appears. Examples of use of word embeddings for health applications include: learning medical concepts such as diagnosis codes, medication codes, procedure codes [6], early detection of heart failure [5], and medical event detection [13]. Many previous techniques have used the word2vec [19, 20] or GloVe [22] packages for training embeddings. One issue with the methods employed by word2vec and GloVe is that they cannot produce embeddings for words that were not seen during training. In contrast, fastText [2, 16, 17] makes use of character level n-grams, which enables one to generate accurate embeddings for words that do not appear in the training vocabulary. The use of character-level n-grams is of particular importance in the medical domain, where a significant number of compound words are used [25, 4, 28].

Generally applications of NLP in health use general text to train word embeddings. In cases where the health-related text is used to train word embeddings, most published models only use between 200 and 400 dimensions [1, 18]. Recent studies show that the use of large corpora from more than one source can improve the performance of embeddings [3, 27]. Chen *et al.*, (2019) [4] and Zhang *et al.*, (2019) [28] provide embeddings on health-related texts, with word embeddings of 700- and 200- dimensional embeddings respectively. Zhang *et al.*, (2019) [28] make use of the sub-word information during the training of word embeddings. We also make use of sub-word information during the training of word embeddings; however, in contrast to Zhang *et al.*, (2019) [28], we present high dimensional word embeddings. Also, our word embeddings make use of large corpora of health-related text from multiple sources. We present comparisons of

F-measures using these recently published word embeddings for the prediction of medical codes to our word embeddings.

Purushotham *et al.*, (2017) [23] use Medical Information Mart for Intensive Care (MIMIC) III to present benchmark models on clinical prediction tasks such as mortality prediction, forecasting length of stay, and ICD-9 code group prediction. MIMIC III is one of the largest publicly available medical databases, containing both structured data and free-form text records [15, 9, 7]. We make use of the free-form text hospital discharge reports contained in MIMIC III, along with the corresponding ICD-9 diagnosis codes.

3 Representing Text

The bag of words (BOW) approach is a simple method for representing text that does not consider the order that words occur in a document. A document is represented as a sparse vector where each element stores either the number of occurrences of a word, or a binary value indicating that the word is present in the document. BOW is considered to be a relatively simple yet effective method [8, 17].

Embedding words in vector spaces that encode semantics has become popular in recent years. In general, for many NLP tasks, continuous word representations trained on large unlabelled datasets have been shown to improve performance relative to other representations [2, 16, 17]. Figure 1 provides a pictorial example of how these vector spaces may be organised. The use of word embeddings is motivated by the distributional hypothesis [12], which states that there is a higher chance that words with similar meaning will occur in similar contexts. By examining a large corpus, it is possible to learn embeddings that capture the semantic similarity between words, as inferred by the contexts they are seen in. Word embeddings provide a means for effective representation learning without the complexity of deep neural networks, and can be trained efficiently on large datasets [19].

FastText [2] is one popular system for learning word embeddings. It supports both the skip-gram with negative sampling (SGNS) and continuous bag of words (CBOW) methods for training word embeddings. In contrast to word2vec, where distinct word embeddings are learnt directly from words, fastText represents each word as a bag of character n-grams, and word embeddings are obtained by summing these character n-gram representations. More information on fastText is provided by Bojanowski *et al.*, (2016) [2]. For example, the tri-grams for the word "apple" are "app", "ppl", and "ple". The resulting word embedding vector for "apple" will be the sum of the vectors of each of these three tri-grams. This modelling choice enables fastText to produce vectors even for novel words that were not present in the training data, as long as at least some of the n-grams have been seen before. It has been shown that fastText can achieve accuracies similar to deep learning classifiers, while being a lot more efficient to train [17].

The classification problems encountered in natural language processing typically involve predicting labels for entire documents, rather than individual words.

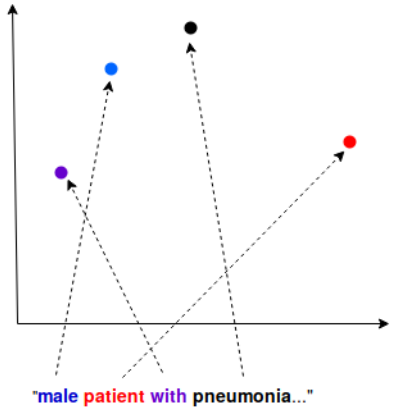

Fig. 1: Visual representation of word embeddings, where each word is mapped to a vector. For simplicity only a 2-D representation is used for embeddings.

As such, one must define a representation for documents that can be easily constructed using the embeddings learned for words. In this work, we obtain document embeddings by computing the vector sum of the embeddings for each word in the document. This vector sum is then normalised to have length one, to ensure that documents of different lengths have representations of similar magnitudes.

4 Data

MIMIC III is used in this study both for classification experiments and for training word embeddings. The most recent version of this dataset, MIMIC III, is one of the most comprehensive publicly available medical databases [7, 9, 15]. It contains de-identified health records of 49,785 adult patient admissions (age $>$ 15) and 7,870 neonatal admissions to critical care units. The data was collected at the Beth Israel Deaconess Medical Center between 2001 and 2012. It includes information such as demographics, laboratory test results, procedures, medications, and physician notes. For this research, we are interested in the discharge summaries of patients admitted to the hospital.

The TREC precision medicine/clinical decision support track 2017 (TREC 2017) [24] provides a considerable corpus of health-related free-form text. This includes 26.8 million published abstracts of medical literature listed on PubMed Central, 241,006 clinical trials documents, and 70,025 abstracts from recent proceedings focused on cancer therapy from AACR (American Association for Cancer Research) and ASCO (American Society of Clinical Oncology). The dataset from the TREC 2017 competition is used here for training word embeddings.

Medical codes, such as the International Classification of Diseases (ICD-9) codes, are widely used to describe diagnoses of patients [14]. Most hospitals

Table 1: Percentage of occurrence of ICD-9 code groupings in unique hospital admissions in MIMIC III. The total number of hospital admissions with a recorded discharge summary is 52,710. E and V codes are referring to external causes of injury and supplemental classification.

ICD-9	%	ICD-9	%	ICD-9	%
Circulatory (circ)	78.40	Digestive (diges)	38.80	Muscular (musc)	17.99
E and V (e+v)	69.09	Blood (bld)	33.56	Prenatal (pren)	17.07
Endocrine (endo)	66.51	Symptoms (symp)	31.36	Neoplasms (neop)	16.37
Respiratory (resp)	46.63	Mental (ment)	29.66	Skin (skin)	12.02
Injury (inj)	41.42	Nervous (nerv)	29.10	Congenital (cong)	5.41
Genitourinary (gen)	40.29	Infectious (inf)	26.96	Pregnancy (preg)	0.31

manually assign the correct codes to patient records based on doctors' clinical diagnosis notes. Hence, the use of machine learning techniques to predict ICD codes from free-form medical text and thus automating the medical coding process has become an important research avenue.

MIMIC III contains ICD-9 annotations to indicate the diagnoses and diseases of admitted patients. There are 6,984 distinct ICD-9 codes reported in MIMIC III, among the more than 50,000 patient admission records found in this database. These can be grouped into 18 categories, as shown in Table 1 along with the frequencies of these groups. Records typically have more than one code assigned. This work focuses on the application of labelling discharge summaries, as these are the most readily available free-form text records in the MIMIC III dataset.

5 Experiments

We consider the 18 categories of medical codes, presented in Table 1, as 18 separate binary classification problems. That is, each group of ICD-9 codes from the MIMIC III discharge summaries is predicted in isolation. FastText is used for training word embeddings and representing documents, and the Waikato Environment for Knowledge Analysis (WEKA) [11, 26] framework is used to train classifiers on these documents. This section discusses the experimental setup in more detail.

5.1 Data Pre-processing

In order to maximise the use of free-form medical text "as is," we minimise preprocessing. One of the significant issues of data mining medical text in free-form is the use of acronyms and abbreviations. Simple changes such as converting uppercase letters to lowercase, or omitting full stops can result in a completely

Table 2: Word embeddings trained by us (top), from previous work (middle), or concatenations thereof (bottom). Dimension details are presented, as are training times and word embeddings model sizes.

Models	Dimensions	Source Data	Train Time	Model Size
M300	300	MIMIC	1 hour	5G
T300	300	TREC	7 hours	13G
TM300	300	TREC+MIMIC	9 hours	15G
T600	600	TREC	13 hours	23G
TM600	600	TREC+MIMIC	16 hours	30G
T900	900	TREC	19 hours	35G
TM900	900	TREC+MIMIC	23 hours	54G
W300 [10]	300	Wiki	-	7G
BWV200 [4, 28]	200	PubMed[1]+MIMIC	-	26G
BSV700 [4, 28]	700	PubMed+MIMIC	-	21G
T300+M300	600	TREC+MIMIC	8 hours	18G
W300+T300+M300	900	Wiki+TREC+MIMIC	8+ hours	25G
T900+W300	1,200	Wiki+TREC	19+ hours	42G
TM900+W300	1,200	Wiki+TREC+MIMIC	23+ hours	61G

different meaning. For example, "Ab" is used to refer to an antibody, while "AB" is used to refer to abortion. As word embeddings are case sensitive, we keep the text as is for both training and experiments to capture maximum meaning. We do not perform downcasing, nor do we remove special characters or full stops. We only remove extra spaces and unwanted newline characters, as fastText uses new line characters to separate examples.

5.2 Training Word Embeddings

Our embeddings are trained to the exact same specifications as the Wikipedia and common crawl fastText models in [10]. We make use of both MIMIC III and TREC 2017 datasets to train our word embeddings. The sizes of these datasets are 4GB and 24GB, respectively. The word embeddings are trained using the CBOW method, character n-grams of length 5, a window of size 5, ten negative samples per positive sample, and with various settings for the number of dimensions. The learning rate used for training these models is 0.05, with the exception of M300 (see Table 2), where the learning rate is 0.03. We also include two very recently published (2019) medical text trained word embeddings of dimension size 200 and 700 [4, 28] for comparison.

Table 2 presents details of the embedding trained by us, previously published word embeddings, and the concatenated word embeddings. Concatenated

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

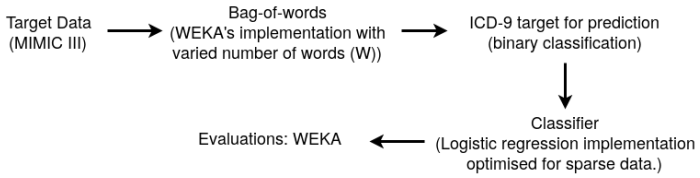

Fig. 2: Flow chart of using bag-of-words for prediction.

embeddings are word embeddings formed by concatenating multiple word embeddings. For example, in the T300+M300, the first 300 elements are the word vectors obtained using the TREC dataset, and the second 300 elements are taken from the embeddings trained on MIMIC III. The table includes details on dimensions, input data, training time[2] and the size of the model. Both the size of the input data and the number of dimensions influence the training times and model sizes.

5.3 Experimental Process and Classification

Figures 2 and 3 present flowcharts of using BOW and word embeddings for predicting ICD-9 groups from MIMIC III discharge summaries. We use a total of 52,710 discharge summaries, with text length ranging from a few sentences to close to twenty pages. WEKA's implementation of BOW is used with a varied number of words. We use ten-fold cross-validation and classifiers as implemented in WEKA.

We use logistic regression with ridge value of 1 for word embeddings experiments. We experimented with the use of random forests, with various parameter choices, as well as other ridge values for logistic regression. However, we found logistic regression was performing well and was providing consistent F-measures across a range of different ridge values. The purpose of this research is not to achieve the highest possible F-measures, but to show that the high dimensional word embeddings trained on medical text do provide advantages in health applications relative to low dimensional embeddings trained on general text. Hence, we only present results for logistic regression.

For BOW due to the sparsity of the data for large dictionary sizes (such as 100,000 or 600,000 words) we use an implementation of logistic regression optimised for sparse data.

6 Results

This section presents an overview of our experimental results. Table 3 provides a comparison of F-measure for predicting ICD-9 groups from free-form MIMIC III discharge summaries for 300-dimensional and 600-dimensional embeddings.

[2] Training was run on a 4 core Intel i7-6700K CPU @ 4.00GHz with 64GB of RAM.

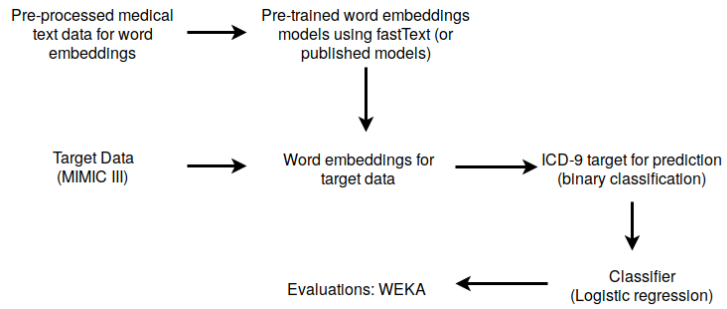

Fig. 3: Flow chart of using word embeddings for prediction.

For 300-dimensional embeddings, W300 are word embeddings that are trained by fastText on Wikipedia and other common crawl text. W300 embeddings are readily available for use in any application. Except for the circulatory label, which is the most frequent one (78.4%), word embeddings specially trained on medical corpora have better F-measures. Overall T300 provides better F-measures than other 300-dimensional word embeddings for most ICD-9 groups. When compared to the recently published BWV200, we found that our 300-dimensional word embeddings performed better for all categories, and on par for E and V.

For 600-dimensional word embeddings, Table 3 presents comparisons across embeddings obtained in a single training phase (T600 and TM600), and word embeddings obtained via concatenation (T300 + M300). We compare our 600 dimension word embeddings to the published 700 dimensional word embeddings (BSV700). F-measures of our 600 dimensional word embeddings are on par with or better than those of the recently published high dimensional word embeddings.

Table 4 presents a comparison for predicting ICD-9 code from free-form discharge summaries in MIMIC III with various dimensions of word embeddings, different number of words for BOW, and between word embeddings and BOW. For word embeddings, the best F-measures for 600-, 900- and 1200- dimensional embeddings are presented for each ICD-9 group. We also indicate which model produced the best 900-dimensional and 1,200-dimensional embeddings (see Table 2 for details of word embeddings, input data and model dimensionality). Generally, the higher the dimensionality, the better the F-measures are for predicting the ICD-9 groups.

For BOW, F-measures of dictionary sizes 1,000, 10,000, 100,000 and 600,000 are presented for all 18 ICD-9 groups. BOW with 600,000 number of words is the largest possible number of features. There is an increase in F-measure as the size of the dictionary increases. A dictionary size of 600,000 results in the best F-measures for all ICD-9 category except pregnancy.

In comparison to word embeddings, F-measures of BOW is consistently better for all ICD-9 groups that occur for less than 42% of the examples. In terms of BOW performance, pregnancy is the most interesting ICD-9 group. Its frequency

Table 3: A comparison of F-measures for predicting ICD-9 groups using 200, 300, 600 and 700-dimensional word embeddings are presented. BWV200 and BSV700 are both published word embeddings and are compared with 300-dimensional word embeddings and 600-dimensional word embeddings, respectively. We use bold to indicate the best F-measures among low dimensional groups of word embeddings (200-300) and the higher dimensional word embeddings (600-700). The best F-measure across all presented word embeddings is underlined for each category.

ICD-9	BWV200	W300	M300	T300	TM300	T600	TM600	T3+M3	BSV700
circ	0.931	**0.932**	**0.932**	**0.932**	0.931	**0.935**	0.934	0.924	0.931
e+v	**0.829**	0.828	**0.829**	**0.829**	0.828	**0.832**	**0.832**	**0.832**	0.831
endo	0.847	0.845	0.846	**0.849**	0.846	**0.851**	0.849	0.850	0.847
resp	0.774	0.774	0.774	**0.778**	0.772	**0.789**	0.788	0.787	0.776
inj	0.660	0.649	**0.663**	0.662	0.660	0.675	0.676	0.677	**0.682**
gen	0.721	0.716	0.724	**0.731**	0.725	**0.740**	**0.740**	**0.740**	0.732
diges	0.679	0.692	0.693	**0.696**	0.692	**0.712**	0.705	0.710	0.696
bld	0.557	0.566	**0.573**	0.570	0.570	0.593	0.589	**0.594**	0.586
symp	0.475	0.486	0.482	**0.487**	0.483	0.504	0.502	0.500	**0.505**
ment	0.533	0.530	0.530	**0.542**	0.539	**0.577**	0.576	**0.577**	0.559
nerv	0.530	0.534	0.527	**0.543**	0.531	**0.571**	0.558	0.564	0.553
inf	0.634	0.634	0.641	**0.647**	**0.647**	0.663	0.659	**0.664**	0.648
musc	0.254	0.274	0.258	**0.294**	0.267	**0.338**	0.314	0.319	0.315
pren	0.589	0.590	0.588	**0.594**	0.587	0.601	0.597	0.598	**0.603**
neop	0.693	0.688	0.702	**0.705**	0.690	0.728	0.721	**0.732**	0.727
skin	0.343	0.335	0.344	**0.346**	0.344	0.389	0.384	0.386	**0.397**
cong	0.365	0.371	0.369	**0.391**	0.350	**0.438**	0.406	0.435	0.424
preg	0.525	0.502	0.543	**0.565**	0.512	0.579	0.566	**0.599**	0.586

is only 0.31% in the MIMIC III dataset. Text in this group is very specific, including uniquely identifying words such as delivery, labour and birth, and is probably one possible explanation to the success of BOW for predicting the pregnancy label.

7 Discussion

In this paper, we investigate how the source domain used for training word embeddings impacts the performance of medical text classification. We also demonstrate the effect that embedding dimensionality plays in determining the accuracy of the resulting classifiers. The prediction of ICD-9 codes from discharge summaries of MIMIC III is used as an example health application to show that high dimensions, especially trained on health-related corpora, have better F-measures compared to word embeddings with lower dimensions or are trained

Table 4: A comparisons of F-measure for ICD-9 groups between word embeddings with varied dimensions (left) and BOW with varied number of words (right). For word embeddings the best F-measure across 900-dimensional and 1200-dimensional word embeddings for each category are presented. Corresponding best 900- and 1,200- dimensional models are also listed. For details of the models see Table 2. We use bold to indicate the best F-measures among varied dimensional word embeddings and varied number of words for BOW. The best F-measure across all is underlined for each category.

ICD-9	Word Embeddings					BOW			
	600	900	best 900-dim model	1,200	best 1,200-dim model	1,000	10,000	100,000	600,000
circ	0.935	**0.937**	T900	0.936	T9W3	0.930	0.920	0.931	**0.932**
e+v	0.832	**0.833**	W3T3M3	**0.833**	T9W3	0.801	0.788	0.808	**0.812**
endo	0.851	0.853	T900	**0.854**	T9W3	0.814	0.825	0.840	**0.845**
resp	0.789	0.792	W3T3M3	**0.794**	TM9W3	0.763	0.775	0.788	**0.792**
inj	0.677	0.684	WTM+T900	**0.689**	TM9W3	0.642	0.675	0.693	**0.697**
gen	0.740	0.748	TM9	**0.751**	TM9W3	0.733	0.751	0.769	**0.775**
diges	0.712	0.724	T900	**0.730**	T9W3	0.701	0.728	0.748	**0.752**
bld	0.594	0.601	TM9	**0.607**	T9W3	0.558	0.595	0.608	**0.614**
symp	0.504	0.514	W3T3M3	**0.517**	T9W3	0.476	0.507	0.524	**0.528**
ment	0.577	0.592	TM9	**0.606**	TM9W3	0.567	0.616	0.635	**0.639**
nerv	0.571	0.577	T900	**0.586**	T9W3	0.491	0.594	0.628	**0.629**
inf	0.664	0.671	T900	**0.677**	T9W3	0.606	0.667	0.693	**0.698**
musc	0.338	0.354	T900	**0.372**	T9W3	0.344	0.476	0.488	**0.489**
pren	0.601	0.607	W3T3M3	**0.608**	Both	0.574	0.557	0.620	**0.623**
neop	0.732	0.741	W3T3M3	**0.746**	T9W3	0.665	0.713	0.766	**0.773**
skin	0.389	0.418	T900	**0.435**	T9W3	0.438	0.483	0.526	**0.530**
cong	0.438	**0.465**	T900	0.463	T9W3	0.348	0.485	0.519	**0.529**
preg	0.599	0.593	W3T3M3	**0.605**	TM9W3	**0.737**	0.705	0.709	0.726

on general text such as Wikipedia. We also compare our word embeddings with recently published word embeddings and show that our embeddings perform better for most ICD-9 groups and are very similar for others. Reasons for such differences include pre-processing of input data, parameter selection, and source of the datasets used for training the embeddings. We also present comparisons with BOW, where we observe that F-measures obtained using BOW are consistently better that word embeddings for all ICD-9 groups that occur for less than 42% of the examples. In general, word embeddings are favoured over BOW as word embeddings are known to capture the meaning of text content, and can better utilise a range of classifiers compared to BOW. However, the results also indicate that for some categories, such as pregnancy, where the data is rather imbalanced, and very specific vocabulary is used, BOW may be the better option.

We also present the sizes and training times required for training word embeddings. Model sizes and training times are both influenced by the input data size and the number of dimensions generated, and can become quite large. The main reason for the large model sizes is the use of hash tables for storing character n-gram information. FastText does provide ways to reduce the final word embeddings model sizes, however such compression necessarily also impacts on accuracy.

For this research, we considered ICD-9 groupings as a set of binary classification problems. Alternatively they could also be represented as a single multi-label classification problem. Each unique patient admitted to the hospital can have more than one diagnosis and be categorised into different groups or have more than one diagnosis from the same ICD-9 group. Our main aim for this research was to investigate high dimensional word embeddings trained in the medical text and hence treating ICD-9 grouping as a binary classification problem was sufficient. However, to optimise the accuracy of predicting ICD-9 code from the free-form medical text, it will be essential to also investigate it as a hierarchical multi-label classification problem in future work.

References

1. Beam, A.L., Kompa, B., Fried, I., Palmer, N.P., Shi, X., Cai, T., Kohane, I.S.: Clinical concept embeddings learned from massive sources of multimodal medical data. arXiv preprint arXiv:1804.01486 (2018)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
3. Cao, Y., Huang, L., Ji, H., Chen, X., Li, J.: Bridge text and knowledge by learning multi-prototype entity mention embedding. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1623–1633 (2017)
4. Chen, Q., Peng, Y., Lu, Z.: Biosentvec: creating sentence embeddings for biomedical texts. The 7th IEEE International Conference on Healthcare Informatics (2019)
5. Choi, E., Schuetz, A., Stewart, W.F., Sun, J.: Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association : JAMIA **24**(2), 361–370 (2017), http://doi.org/10.1093/jamia/ocw112
6. Choi, Y., Chiu, C.Y.I., Sontag, D.: Learning low-dimensional representations of medical concepts. AMIA Summits on Translational Science Proceedings p. 4150 (2016)
7. Data, M.C.: Secondary Analysis of Electronic Health Records. Springer (2016)
8. Goldberg, Y.: Neural network methods for natural language processing. Synthesis Lectures on Human Language Technologies **10**(1), 1–309 (2017)
9. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
10. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018)

11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter **11**(1), 10–18 (2009)
12. Harris, Z.S.: Distributional structure. WORD **10**(2-3), 146–162 (1954), DOI:10.1080/00437956.1954.11659520
13. Jagannatha, A.N., Yu, H.: Bidirectional RNN for medical event detection in electronic health records. Association for Computational Linguistics. North American Chapter. Meeting p. 473482 (2016)
14. Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics **13**(6), 395 (2012)
15. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific data **3**, 160035 (2016)
16. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016)
17. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
18. Mencía, E.L., De Melo, G., Nam, J.: Medical concept embeddings via labeled background corpora. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). pp. 4629–4636 (2016)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
21. Pakhomov, S.V., Finley, G., McEwan, R., Wang, Y., Melton, G.B.: Corpus domain effects on distributional semantic modeling of medical terms. Bioinformatics **32**(23), 3635–3644 (2016)
22. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
23. Purushotham, S., Meng, C., Che, Z., Liu, Y.: Benchmark of deep learning models on large healthcare mimic datasets. arXiv preprint arXiv:1710.08531 (2017)
24. Roberts, K., Demner-Fushman, D., Voorhees, E.M., Hersh, W.R., Bedrick, S., Lazar, A.J., Pant, S.: Overview of the trec 2017 precision medicine track. NIST Special Publication pp. 500–324 (2017)
25. Shi, H., Xie, P., Hu, Z., Zhang, M., Xing, E.P.: Towards automated ICD coding using deep learning. arXiv preprint arXiv:1711.04075 (2017)
26. Witten, I., Frank, E., Hall, M., Pal, C.: Data Mining: Practical Machine Learning Tools and Techniques, 4th Edition. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2016)
27. Yamada, I., Shindo, H., Takeda, H., Takefuji, Y.: Joint learning of the embedding of words and entities for named entity disambiguation. arXiv preprint arXiv:1601.01343 (2016)
28. Zhang, Y., Chen, Q., Yang, Z., Lin, H., Lu, Z.: BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific data **6**(1), 52 (2019)