Pacific Symposium on Biocomputing 25:295-306(2020)

Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data

Andrew L. Beam[1*†], Benjamin Kompa[2*], Allen Schmaltz[1], Inbar Fried[3], Griffin Weber[2], Nathan Palmer[2], Xu Shi[1], Tianxi Cai[1], Isaac S. Kohane[2]
[1]*Harvard T.H. Chan School of Public Health Boston, MA 02115, USA*
[2]*Harvard Medical School Boston, MA 02115, USA*
[3]*University of North Carolina School of Medicine Chapel Hill, NC 27516, USA*
[*]*Denotes equal contribution* [†]*E-mail: andrew_beam@hms.harvard.edu*

Word embeddings are a popular approach to unsupervised learning of word relationships that are widely used in natural language processing. In this article, we present a new set of embeddings for medical concepts learned using an extremely large collection of multimodal medical data. Leaning on recent theoretical insights, we demonstrate how an insurance claims database of 60 million members, a collection of 20 million clinical notes, and 1.7 million full text biomedical journal articles can be combined to embed concepts into a common space, resulting in the largest ever set of embeddings for 108,477 medical concepts. To evaluate our approach, we present a new benchmark methodology based on statistical power specifically designed to test embeddings of medical concepts. Our approach, called *cui2vec*, attains state-of-the-art performance relative to previous methods in most instances. Finally, we provide a downloadable set of pre-trained embeddings for other researchers to use, as well as an online tool for interactive exploration of the *cui2vec* embeddings.

Keywords: machine learning; electronic health records; claims data; natural language processing

1. Introduction

Word embeddings have become an extremely popular way to represent sparse, high-dimensional data in machine learning and natural language processing (NLP). Modern notions of word embeddings based on neural networks have their roots in the neural language model of Bengio et al.,[1] though the idea is closely related to many other approaches, notably latent semantic analysis (LSA)[2] and hyperspace analogue to language (HAL).[3] Word embeddings are motivated by the observation that traditional representations for words, such as a one-hot encoding, are high dimensional and inefficient, since such an encoding captures none of the similarity or correlation information between words in the source text. The central idea is that a word can be characterized by "the company it keeps,"[4] thus context words which appear around a given word encode a large amount of information regarding that word's meaning. Word embeddings model this contextual information by creating a lower-dimensional space such that words that appear in similar contexts will be nearby in this new space.

The embedding approach in *word2vec*[5] has become quite popular since its introduction, and embeddings are now standard components in many NLP tasks. The main application has been in the use of "transfer learning," where embeddings are first learned using extremely large sources of unlabeled text (from web-crawls, Wikipedia dumps, etc.), and the embeddings are then used in a supervised task as components of a model (e.g., a recurrent neural network) which accepts the pre-trained embeddings as inputs. It has been shown that transfer learning can work as well as it does for image data,[6] opening up numerous possibilities to exploit transfer learning in many NLP applications. Within the context of medical data, recent examples have shown that transfer learning works very well for imaging tasks,[7,8] due in large part to the availability of pre-trained computer vision models[9–11] that were pre-trained on the ImageNet database.[12]

Machine learning has enormous potential in healthcare;[13] however, many researchers lack access to large sources of non-imaging healthcare data due to privacy concerns. This has resulted in a lack of pre-trained resources for applications in healthcare and medicine relative to other areas of machine learning and NLP. Moreover, because healthcare data come in a variety of forms, popular word embedding algorithms like *word2vec* and *GloVe*,[14] which were originally developed for text, cannot be directly applied to many kinds of healthcare data.

The primary goal of this work is to construct a comprehensive set of embeddings for medical concepts, which we refer to as *cui2vec*, by combining extremely large sources of multimodal healthcare data.

2. Overview of word2vec and GloVe

2.1. *word2vec*

The original work that introduced *word2vec*[5] actually contains a collection of models and algorithms including the continuous bag of words (CBOW) model and the skip-gram model. The CBOW model predicts the probability of the target word given its context defined within a window, while the skip-gram model predicts the surrounding context given the target word. Specifically, the skip-gram model[5] seeks to construct vector representations of a target word w and a context word c such that the conditional probability $p(w|c)$ is high for $<w, c>$ pairs that co-occur frequently in the source text. For the remainder of this paper we will use w and c to refer to the target word and context word respectively, and use and $\vec{w}$, $\vec{c}$ to refer to the $1 \times d$ dimensional target word and context embeddings. Under the skip-gram model, the conditional probability of observing context word c within a fixed window given the target word w is proportional to the dot-product of their corresponding vectors, and is given by the *softmax* function below:

$$p(w|c) = \frac{\exp(\vec{w}\vec{c}^T)}{\sum_j \exp(\vec{w}\vec{c}_j^T)} \tag{1}$$

where the sum in the denominator is over all unique context words in the source corpus. Note that this sum is generally intractable and requires approximations to estimate efficiently. Thus, the vectors $\vec{w}, \vec{c}$ encode information about how likely word w is to appear in a randomly selected piece of text, given word c has been observed.

A key feature of *word2vec* are techniques that enable efficient training on large corpora. For example, negative sampling approximates the sum in the denominator by randomly sampling k context words which do not appear in the current window. This allows the algorithm to be run with bounded memory requirements and in a parallel fashion, which improves the training speed and enables training on very large corpora.[15] Indeed, the key point of Mikolov et al. was that training a simple and scalable model with more data results in better accuracy than a complex non-linear model on a variety of benchmarks.

2.2. *GloVe*

Global Vectors for Word Representations (GloVe)[14] was introduced shortly after Mikolov et al. and differs in several important ways. *GloVe* produces word embeddings by fitting a weighted log-linear model to co-occurrence statistics. Given that a target word w and a context word c co-occur y times, *GloVe* solves the following least-squares optimization problem:

$$\underset{\vec{w},\vec{c},b_w,b_c}{\operatorname{argmin}} f(y)(\vec{w}\vec{c}^T + b_w + b_c - \log(y))^2 \tag{2}$$

where b_w, b_c are word and context biases, respectively and $f(y)$ is a weighting function and is given by:

$$f(y) = \begin{cases} \left(\frac{y}{y_{max}}\right)^{\alpha} & y < y_{max} \\ 1 & y \geq y_{max} \end{cases} \tag{3}$$

The final embedding for word i is the sum of the resulting word and context vectors for that word. This is repeated for all w,c pairs and is trained iteratively using stochastic gradient descent. The most expensive step is the construction of the term-term co-occurrence matrix, which is necessary before training can begin.

2.3. *Embeddings as a Factorization of a Modified Co-occurrence Matrix*

Previous work[16] by Levy and Goldberg showed that the skip-gram model with negative sampling (SGNS), which is often considered to be state-of-the-art,[17] is implicitly factorizing a shifted, positive pointwise mutual information (PMI) matrix of word-context pairs. Pointwise mutual information (PMI) is a measure of association between a word and a context word, and can be computed from the counts of word-context pairs in the corpus, given by:

$$\text{PMI}(w, c) = \frac{p(w, c)}{p(w) * p(c)} \tag{4}$$

where $p(w, c)$ is the number of times word w and context-word c occur in the same context window divided by the total number of word-context pairs, whereas $p(w)$, $p(c)$ are the singleton frequencies of w and c, respectively. If we shift the PMI by some constant $\log(k)$ (where k is the number of negative samples in the original *word2vec* paper[5]) and set all negative entries to 0, and factor the resulting *shifted positive pointwise mutual information matrix* (SPPMI) we recover the implicit objective of *word2vec*'s SGNS model. The element wise SSPMI transformation is shown below:

$$\text{SPPMI}(w, c) = \max(\text{PMI}(w, c) - \log(k), 0) \tag{5}$$

Therefore, one can simply factorize the SSPMI matrix using any factorization method, such as a singular value decomposition (SVD), to obtain a lower-dimension embedding of the words. This finding is critical as it links *word2vec* to traditional count-based methods that are based on co-occurrence statistics.

GloVe was originally presented in terms of explicit matrix factorization and provides an algorithm to perform this factorization (stochastic gradient descent to minimize sum-of-squared error). Thus, under this unified framework the starting point for both *word2vec* and *GloVe* is the construction of a term-term co-occurrence matrix. This insight is what allows us to use these algorithms on problems which may contain non-textual data sources, as we can materialize a co-occurrence matrix using any data where such co-occurrences can be computed. Then we simply use the *GloVe* algorithm to directly factor this matrix or use SVD to factor the SSMPI matrix to create *word2vec* style embeddings.

2.4. *Overview of cui2vec*

Medical data are multi-modal by nature and come in many forms including free text (in medical publications and clinical notes) and billing codes for diagnoses and procedures in the electronic healthcare record (EHR). The *cui2vec* system works by first mapping all of these concepts into a common concept unique identifier space (CUI) using a thesaurus from the Unified Medical Language System (UMLS). Next, a CUI-CUI co-occurrence matrix is constructed, but the way a co-occurrence is counted depends on the source data. For non-clinical text data (e.g., journal articles), it is first preprocessed (see Section 3) and chunked into fixed length windows of 10 words, and a co-occurrence is counted as the appearance of a CUI-CUI pair in the same window. For claims data, ICD-9 codes are mapped to UMLS CUIs and a co-occurrence is counted as the number of patients in which two CUIs appear in any 30-day period. Finally, for the clinical notes, we counted a co-occurrence as two CUIs appearing in the same 30-day 'bin' in a similar fashion to previous work,[18] but see the original publication[19] for the precise definition. Once the master co-occurence matrix is created, it can be directly factored by *GloVe* or transformed into a SSPMI matrix and factored using SVD to create *word2vec* embeddings.

Related Work

There is a long history of machine learning and natural language processing for clinical uses, but for the purposes of this paper we confine our review to papers that are directly seeking to create low dimensional representations of clinical concepts, in the spirit of *word2vec* and *GloVe*. The first investigations[20–22] using *word2vec* for medical concepts were performed shortly after the original *word2vec* paper appeared in 2013 and reported mixed results, though De Vine et al.[21] reported state-of-the-art performance with respect to human assessments of concept similarity and relatedness. Recently, transformer based models[23–26] have demonstrated state of the art performance on many NLP tasks. Alsentzer et al.[26] used clinical notes to fine-tune BERT.

Liu et al.[27] used embeddings jointly trained on Wikipedia and ICU notes to perform automatic expansion of abbreviations which are common in clinical notes. Lastly, Choi et

al.[18] performed the work that is most comparable to this study, which used similar sources of data to create embeddings for UMLS CUIs. Choi et al. used a claims database of 4 million patients and a novel methodology to create a set of clinical embeddings as well as the notes from Finlayson et al.[19]

2.5. *Contributions of this work*

The work presented here differs in several important ways from existing works. First, we have access to a much larger claims database of 60 million patients and a larger set of 1.7 million full text articles (not restricted to abstracts), which should enable both a much larger and higher quality set of embeddings. Secondly, the embeddings produced by Choi et al. are different for each data source, whereas we map all concepts into a common co-occurrence space to produce a single set of embeddings that can be used on tasks with different kinds of clinical data. This co-occurrence space mapping also allows us to use multimodal data that would be difficult to integrate using transformer-based models. We also present a new and expanded evaluation methodology that is both more interpretable and, we believe, a more natural way to benchmark sets of clinical embeddings that will be of general use for future medical embedding work. Finally, we believe that our approach incorporates many of the best practices with respect to tuning parameters (see Section 3) which also results in increased performance. In summary, this work presents results in a new set of embeddings for 108,477 medical concepts, the largest ever such collection, which are derived from three sources of clinical data and are equal to or exceed the existing state of the art on nearly all benchmarks.

3. Materials and Methods

3.1. *Data Sources*

The data come from the following three independent sources: an un-identifiable claims database from a nationwide US health insurance plan with 60 million members over the period of 2008-2015, a dataset of concept co-occurrences from 20 million notes at Stanford,[19] and an open access collection of 1.7 million full text journal articles obtained from PubMed Central. For the purposes of this study, the insurer has asked not to be named.

3.2. *Text Normalization and Preprocessing*

For text data it is important to first normalize against some standard vocabulary or thesaurus. Word embeddings operate on tokens, and many medical concepts can span multiple tokens. To collapse multi-word concepts into a single token, we used the Narrative Information Linear Extraction (NILE)[28] system normalized against the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT)[29] reference thesaurus. SNOMED-CT IDs were then mapped to concept unique identifiers (CUIs) from the UMLS.[30] The pipeline converts all letters to lowercase, removes punctuation, and replaces all medical concepts with their CUI representation (e.g., 'bronchopulmonary dysplasia' with C0006287 and 'resulting from' with C0678226). For example, our pipeline would transform the following sentence (taken from previous work[31]):

```
Bronchopulmonary Dysplasia was first described by Northway and colleagues in
1967 as a lung injury in a preterm infant resulting from oxygen and
mechanical ventilation.
```

into the following normalized representation:

```
C0006287 was first described by northway and colleagues in 1967 as a C0024109
C3263722 in a C0021294 C0678226 C0030054 and C0199470
```

Benchmarks and Evaluation

The benchmarking strategy leverages previously published 'known' relationships between medical concepts. We compare how similar the embeddings for a pair of concepts are by computing the cosine similarity of their corresponding vectors, and use this similarity to assess whether or not the two concepts are related. Cosine similarity between word vectors $\vec{w_1}, \vec{w_2}$ is given by:

$$\cos(\vec{w_1}, \vec{w_2}) = \frac{w_1 w_2^T}{\|w_1\|_2 \|w_2\|_2}$$

and is 1 if the vectors are identical and 0 if they are orthogonal. One approach would be to rank the cosine similarity for a known relationship against all others via a ranking metric such as mean-precision or discounted cumulative gain. However, such a strategy has several limitations. The primary issue is that many concepts may correctly be ranked higher than the query concept, but they may not be part of the database of known relationships. Thus, a ranking metric may incorrectly penalize a set of embeddings simply because some true relationships were ranked higher but were not included in the list of 'known' relationships.

Instead, we present a new approach based on the notion of statistical power. For a known relationship pair (x, y), we first compute the null distribution of scores by drawing 10,000 bootstrap samples (x^*, y^*) where x^* and y^* belong to the same category as x and y, respectively. For example, when assessing whether 'preterm infant' (which is a disease or syndrome) is associated with 'bronchopulmonary dysplasia' (also a disease or syndrome), we would randomly sample two concepts from the "disease or syndrom" class and compute their cosine similarity, and then repeat this procedure 10,000 times to create the bootstrap null distribution. We then compare the observed score between x and y and declare it statistically significant if it is greater than the 95th percentile of the bootstrap distribution (e.g., $p < 0.05$ for a one-sided test). Applying this procedure to the collection of known relationships, we calculate the statistical power to reject the null of no relationship which is the quantity we report in all experiments, except for the comparison to human assessments of similarity. This metric has the added benefit of being easy to interpret, as it is an estimate of the fraction true relationships discovered given a tolerance for a 5% false positive rate.

Below is a list of the benchmarks used in this study, along with details that are specific to each. We provide an example of a known relationship from each category to help the reader understand the types of relationships captured by each benchmark.

- **Comorbid Conditions**: A comorbidity is a disease or condition that frequently accompanies a primary diagnosis. We created a curated set of comorbid conditions for Addison's disease, autism, heart disease, obesity, schizophrenia, type 1 diabetes and type 2 diabetes. These comorbidities were extracted from the Mayo Clinic's Encyclopedia of Diseases and Conditions,[32] Wikipedia, and the Merck Manuals.[33]
  - *Example*: Primary condition: premature infant (CUI: C0021294) Comorbidity: bronchopulmonary dysplasia (CUI: C0006287)

- **Causative Relationships**: The UMLS contains a table (MRREL) of entities known to be the cause of a certain result. We extracted known instances of the relationships *cause of* and *causative agent*, and *induces* from the MRREL table. We computed the null distribution for these relationships by computing the similarity of randomly sampled concepts with the same semantic type as the cause and randomly sampled concepts with the same semantic type as the result.
  - *Example*: Cause: Jellyfish sting (CUI: C0241955) Result: Irukandji syndrome (CUI: C1655386)
- **National Drug File Reference Terminology (NDF-RT)**: The NDF-RT was created by the U.S. Department of Veterans Affairs, Veterans Health Administration. We extracted drug-condition relationships using the *may prevent* and *may treat* relationships. We assessed power to detect *may treat* and *may prevent* relationships using bootstrap scores of random drug-disease pairs.
  - *Example*: Drug: abciximab (CUI: C0288672) May Treat: Myocardial Ischemia (CUI: C0151744)
- **UMLS Semantic Type**: Semantic types are meta-information about which category a concept belongs to, and these categories are arranged in a hierarchy. We extracted the most specific semantic type available for each concept from the MRSTY file provided by UMLS. To assess power to detect if two concepts belonged to the same semantic type, we randomly sampled concepts from different semantic type classes and computed a marginal null distribution of scores.
  - *Example*: Concept: Metronidazole (CUI: C0025872, Semantic Type: Pharmacologic Substance) Concept: Clofazimine (CUI: C0008996, Semantic Type: Pharmacologic Substance)
- **Human Assessment of Concept Similarity**: Previous work[34] has assessed how resident physicians perceive relationships among 566 pairs of UMLS concepts. Each concept pair has an average measure of how similar or related two concepts are to be as judged by resident physicians. We report Spearman correlation between human assessment scores and cosine similarity from the embeddings for this benchmark.

3.3. *Implementation Details*

There are many hyper-parameters associated with both *word2vec* and *GloVe* that can have a dramatic effect on performance. In *word2vec* parameters such as the number of negative samples, the size of the context window, the amount of smoothing for the context singleton-frequencies, and whether or not the context vectors are used to construct the final embeddings are all options that the practitioner must choose. Levy and Goldberg[35] conducted a systematic set of experiments on the effects of these hyper-parameters on the performance of *word2vec*, and we follow their recommendations in this work. Specifically, we used the following settings for all *word2vec* experiments that are based on a singular value decomposition (SVD):

- Smoothing of singleton frequencies by a constant exponential term. Instead of using $p(w)$ in (4), we instead use $p(w)^\alpha$, where α is set to 0.75. In Levy and Goldberg, they

recommend only smoothing the context singleton frequencies, but our co-occurrence matrices are symmetric so there is no difference in the singleton frequency when it is a 'word' and when it is a 'context'.
- We set $k = 1$ in the SPPMI transformation (i.e., no shift).
- We construct the final embeddings using a symmetrically scaled sum of the word and context vectors resulting from the singular value decomposition. Given the first d singular vectors and singular values resulting from the SVD of a SPPMI matrix X, $SVD_d(X) = U_d \Sigma_d V_d$, the d-dimensional word embeddings W are constructed as follows:

$$\tilde{W} = U_d \sqrt{\Sigma_d}$$
$$\tilde{C} = V_d \sqrt{\Sigma_d}$$
$$W = \tilde{W} + \tilde{C}$$

- The SVD of the sparse SPPMI matrix was performed using the *augmented implicitly restarted Lanczos bidiagonalization algorithm* with the `irlba` package[36,37] in the R programming language.

For the comparison to the traditional *word2vec* algorithm on the articles from PubMed, we used the implementation available in the `gensim` python package.[38] We used the skip-gram algorithm, hierarchical softmax, 10 negative samples, and a window size of 10. We used the implementation of GloVe available in the R package `text2vec`.[39] We used the sum of target word and context vectors as the final embedding and set the $y_{max} = 100$. As a baseline, we performed a SVD on the raw co-occurence matrix, and we report these results as *PCA*.

4. Results

4.1. *Benchmark Results*

We compared embeddings created by *GloVe*, *word2vec*, and *PCA* on our benchmarks to determine which algorithm and embedding dimension produced the best results across each individual dataset and on the combined data. These results are shown in Table 1. The best configuration was *word2vec* with an embedding dimension of 500, as it achieved the highest performance across nearly all benchmarks. Interestingly, we saw only a modest effect of embedding dimension on the benchmarks based on power (see Supplement). Also of note, the most direct comparison we could make to the original *word2vec* algorithm was using PubMed articles. On this dataset, *word2vec* based on a SVD was better than the original algorithm, as shown in the second row group in Table 1.

The 500-dimensional *word2vec* style embeddings using the combined data are referred to as the *cui2vec* embeddings in all subsequent experiments.

4.2. *Comparison to previous results*

In total we were able to estimate embeddings for 108,477 unique concepts using the combined set of data, making this the largest set of embeddings for medical concepts to date. Figure 1 shows a visualization of the various intersections of the 108,477 concepts found across the different sources of data using the UpSet visualization method.[40,41]

Table 1: Comparison of *GloVe*, *PCA*, and *word2vec* for an embedding dimension of 500. Columns 1-4 report power to detect known relationships and column 5 reports the Spearman correlation between human assessments of concept similarity and cosine similarity from the embeddings. The best result for each each benchmark/dataset combination is shown in bold. The claims dataset contained only diagnosis codes and no drugs and so did not report results for the NDFRT benchmark.

Data Source	Algorithm	Causative	Comorbidity	Semantic Type	NDFRT	Human Assessment
Claims	GloVe	**0.56**	**0.73**	0.29	-	**0.45**
	PCA	0.40	0.15	0.32	-	0.19
	word2vec (SVD)	0.54	0.50	**0.40**	-	**0.45**
PMC Articles	GloVe	0.59	0.57	0.28	0.54	0.60
	PCA	0.30	0.24	0.24	0.29	0.29
	word2vec (SVD)	**0.83**	**0.59**	**0.49**	**0.84**	**0.67**
	word2vec (original)	0.75	0.51	0.48	0.74	0.59
Clinical Notes	GloVe	0.39	**0.73**	0.51	0.11	0.34
	PCA	0.36	0.31	0.47	0.14	0.53
	word2vec (SVD)	**0.75**	0.52	**0.74**	**0.49**	**0.59**
Combined Data	GloVe	0.40	**0.80**	0.37	0.50	0.39
	PCA	0.24	0.23	0.30	0.37	0.47
	word2vec (SVD)	**0.46**	0.52	**0.53**	**0.57**	**0.47**

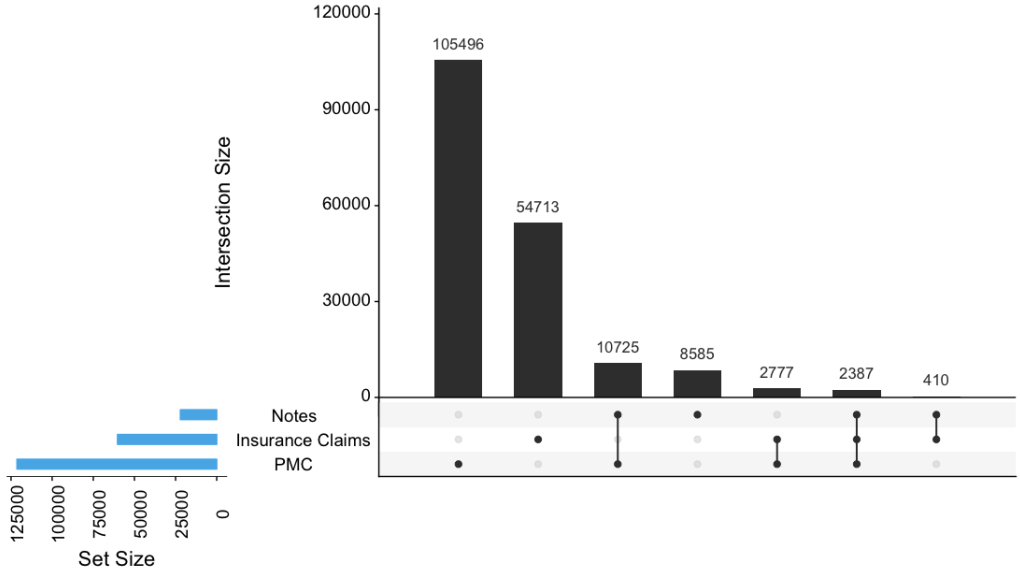

Fig. 1. *Upset* visualization of the intersection of medical concepts found in the insurance claims, clinical notes, and biomedical journal articles (PMC).

Most of the concepts appear in only one corpus, however 16,299 (14%) appeared in multiple sources. We evaluated previously published embeddings obtained through the *clinicalml* github repository (https://github.com/clinicalml/embeddings) for comparison to our *cui2vec* embeddings. Note that all three of the comparison embeddings come from different data sources and have very few concepts in common, so we were forced to perform pairwise comparisons between *cui2vec* and each set of embeddings.

The first comparison was against 300-dimensional embeddings for 15,905 concepts (of

which 12,568 were in common with *cui2vec*) derived from a claims database of 4 million patients. The results are shown in Table 2. We observed that *cui2vec* outperformed the reference embeddings in most tasks, in some instances by a substantial margin, though the embeddings from Choi et al. had the edge in the human assessment benchmark. Next, we compared 300-dimension embeddings for 28,394 concepts derived from the same set of clinical notes in Finlayson et al.[19] published as part of Choi et al.[18] In total, there were 21,789 concepts in common between *cui2vec* and this set of embeddings. Here *cui2vec* was again better in most benchmarks, in some cases by a large margin. Finally, we compared *cui2vec* against 200-dimensional embeddings for 59,266 concepts derived from 348,566 PubMed abstracts, first published in De Vine et al.[42] There were 33,376 concepts in common that were used for benchmarking. On this dataset we observed a huge relative improvement and *cui2vec* was uniformly better across all benchmarks, as shown in Table 2.

Table 2: Comparison of the performance of *cui2vec* to previously published embeddings. Columns 1-4 report power to detect known relationships and column 5 reports the Spearman correlation between human assessments of concept similarity and cosine similarity from the embeddings. The best result for each each comparison is shown in bold.

Source	Causative	Comorbidity	NDFRT	Semantic Type	Human Assessment
Choi et al. (claims)	0.25	**0.37**	0.63	0.24	**0.47**
cui2vec	**0.55**	0.31	**0.73**	**0.43**	0.35
Choi et al. (notes)	0.29	0.23	**0.52**	0.15	0.43
cui2vec	**0.42**	**0.25**	0.42	**0.36**	**0.51**
Devine et al. (PMC abstracts)	0.29	0.05	0.18	0.22	0.45
cui2vec	**0.48**	**0.31**	**0.46**	**0.48**	**0.50**

4.3. *Discussion*

In this study we have created the most comprehensive set of 108,299 clinical embeddings to date using extremely large and multi-modal sources of medical data. When compared to previous results, the *cui2vec* embeddings achieve state-of-the-art performance in many instances. Even though there is more healthcare data than ever, most of it is either unlabeled or weakly labeled, so the ability to extract meaningful structure in an unsupervised manner is extremely important. Another potential obstacle is that most sources of healthcare data are not easily shareable, which limits some researchers to small sources of local data. We hope to reduce both of these barriers by providing our *cui2vec* embeddings that were created using large and national sources of healthcare data. We believe that these embeddings will be generally useful for a variety of clinically oriented machine learning tasks.

Availability of Code and Data

An R package `cui2vec` implementing the *cui2vec* system can be found at the github repository. An interactive explorer of the embeddings can be found here.

Acknowledgements

ALB was supported by NIH/NHLBI 7K01HL141771-02, BK was supported by NIH T32HG002295.

References

1. Y. Bengio, R. Ducharme, P. Vincent and C. Janvin, A Neural Probabilistic Language Model, *The Journal of Machine Learning Research* **3**, 1137 (2003).
2. M. W. Berry, S. T. Dumais and G. W. O'Brien, Using Linear Algebra for Intelligent Information Retrieval, *SIAM Review* **37**, 573 (1995).
3. K. Lund and C. Burgess, Producing high-dimensional semantic spaces from lexical co-occurrence, *Behavior Research Methods, Instruments, and Computers* **28**, 203 (1996).
4. Z. S. Harris, Distributional Structure, *WORD* **10**, 146 (1954).
5. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, K. Chen, J. Dean, T. Mikolov and K. Chen, Distributed Representations of Words and Phrases and their Compositionality., in *NIPS'14* , 2013.
6. J. Howard and S. Ruder, Fine-tuned Language Models for Text Classification, *arXiv preprint arXiv:1801.06146* (2018).
7. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega and D. R. Webster, Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs., *Jama* **304**, 649 (2016).
8. A. L. Beam and I. S. Kohane, Translating Artificial Intelligence Into Clinical Care., *JAMA* **346**, 456 (2016).
9. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
10. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, Rethinking the inception architecture for computer vision, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
11. K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).
12. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009.
13. A. L. Beam and I. S. Kohane, Big data and machine learning in health care, *Jama* **319**, 1317 (2018).
14. J. Pennington, R. Socher and C. D. Manning, GloVe: Global Vectors for Word Representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* , 1532 (2014).
15. T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013).
16. O. Levy and Y. Goldberg, Neural Word Embedding as Implicit Matrix Factorization, *Advances in Neural Information Processing Systems (NIPS)* , 2177 (2014).
17. M. Baroni, G. Dinu and G. Kruszewski, Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014.
18. Y. Choi, C. Y.-I. Chiu and D. Sontag, Learning Low-Dimensional Representations of Medical Concepts, *AMIA* , 373 (2016).
19. S. G. Finlayson, P. LePendu and N. H. Shah, Building the graph of medicine from millions of clinical narratives., *Scientific data* **1**, p. 140032 (2014).
20. J. A. Minarro-Gimenez, O. Marin-Alonso and M. Samwald, Exploring the Application of Deep Learning Techniques on Medical Text Corpora, in *Studies in Health Technology and Informatics*, 2014.

21. L. De Vine, G. Zuccon, B. Koopman, L. Sitbon and P. Bruza, Medical Semantic Similarity with a Neural Language Model, in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14*, 2014.
22. S. Moen and T. S. S. Ananiadou, Distributional semantics resources for biomedical text processing (2013).
23. J. Devlin, M. Chang, K. Lee and K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, *CoRR* **abs/1810.04805** (2018).
24. I. Beltagy, A. Cohan and K. Lo, Scibert: Pretrained contextualized embeddings for scientific text, *CoRR* **abs/1903.10676** (2019).
25. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *CoRR* **abs/1901.08746** (2019).
26. E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann and M. McDermott, Publicly available clinical BERT embeddings, in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, (Association for Computational Linguistics, Minneapolis, Minnesota, USA, June 2019).
27. Y. Liu, T. Ge, K. Mathews, H. Ji and D. McGuinness, Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion, *Proceedings of BioNLP 15* , 92 (2015).
28. S. Yu and T. Cai, A short introduction to NILE, *arXiv preprint arXiv:1311.6063* (2013).
29. K. Donnelly, SNOMED-CT: The advanced terminology and coding system for eHealth, *Studies in health technology and informatics* **121**, p. 279 (2006).
30. O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic acids research* **32**, D267 (2004).
31. A. H. Jobe and E. Bancalari, Bronchopulmonary dysplasia, in *American Journal of Respiratory and Critical Care Medicine*, (7)2001.
32. M. C. Staff, Mayo Clinic: Diseases and Conditions.
33. M. H. Beers, R. Berkow *et al.*, The merck manual, *Disturbances in Newborns and Infants,* (1999).
34. S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen and G. B. Melton, Semantic similarity and relatedness between clinical terms: an experimental study, in *AMIA annual symposium proceedings*, 2010.
35. O. Levy, Y. Goldberg and I. Dagan, Improving Distributional Similarity with Lessons Learned from Word Embeddings, *Transactions of the Association for Computational Linguistics* **3**, 211 (2015).
36. J. Baglama and L. Reichel, Augmented Implicitly Restarted Lanczos Bidiagonalization Methods (2005).
37. J. Baglama, L. Reichel and B. W. Lewis, *irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices*, (2017).
38. R. Rehurek and P. Sojka, Software framework for topic modelling with large corpora, in *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.
39. D. Selivanov, text2vec: Modern text mining framework for r, *Computer software manual](R package version 0.4. 0). Retrieved from https://CRAN. R-project. org/package= text2vec* (2016).
40. J. R. Conway, A. Lex and N. Gehlenborg, UpSetR: An R package for the visualization of intersecting sets and their properties, *Bioinformatics* **33**, 2938 (2017).
41. A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot and H. Pfister, Upset: visualization of intersecting sets, *IEEE transactions on visualization and computer graphics* **20**, 1983 (2014).
42. L. De Vine, G. Zuccon, B. Koopman, L. Sitbon and P. Bruza, Medical semantic similarity with a neural language model, in *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, 2014.