A Comparison of Word Embeddings for the Biomedical Natural Language Processing

Yanshan Wang*, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, Hongfang Liu*

Department of Health Sciences Research

Mayo Clinic

Rochester, USA

Email: {Wang.Yanshan, Liu.Sijia, Afzal.Naveed, Mojarad.Majid, Wang.Liwei, Shen.Feichen, Kingsbury.Paul1, Liu.Hongfang}@mayo.edu

Abstract

Background Neural word embeddings have been widely used in biomedical Natural Language Processing (NLP) applications as they provide vector representations of words capturing the semantic properties of words and the linguistic relationship between words. Many biomedical applications use different textual resources (e.g., Wikipedia and biomedical articles) to train word embeddings and apply these word embeddings to downstream biomedical applications. However, there has been little work on evaluating the word embeddings trained from these resources.

Methods In this study, we provide an empirical evaluation of word embeddings trained from four different resources, namely clinical notes, biomedical publications, Wikipedia, and news. For the former two resources, we trained word embeddings utilizing clinical notes available at Mayo Clinic and articles from the PubMed Central (PMC), respectively. For the latter two resources, we used publicly available pre-trained word embeddings, GloVe and Google News. We performed the evaluation qualitatively and quantitatively. In qualitative evaluation, we manually inspected five most similar medical words to a given set of target medical words, and then analyzed word embeddings through the visualization of those word embeddings. In quantitative evaluation, we conducted both intrinsic and extrinsic evaluation. Intrinsic evaluation directly tested semantic relationships between medical words using four published datasets for measuring semantic similarity between medical terms, i.e., Pedersen, Hliaoutakis, MayoSRS, and UMNSRS. In extrinsic evaluation, we applied word embeddings to downstream biomedical NLP applications, including clinical information extraction (IE), biomedical information retrieval (IR), and relation extraction (RE), using data from shared tasks.

Results Qualitative evaluation shows that the word embeddings trained from EHR and PubMed can find more relevant similar medical terms than these from GloVe and Google News. In the intrinsic quantitative evaluation, the semantic similarity captured by the word embeddings trained from EHR are closer to human experts' judgments on all four tested datasets. In the extrinsic quantitative evaluation, the word embeddings trained from EHR has the best F1 score of 0.900 for the clinical IE task; no word embeddings improve the performance in terms of inference average precision and mean average precision for the biomedical IR task; and the word embeddings trained on Google News has the best overall F1 score of 0.790 for the RE task.

*Corresponding authors.

Conclusion Based on the evaluation results, we can draw the following conclusions. First, the word embeddings trained on EHR and PubMed can capture the semantics of medical terms better than those trained on GloVe and Google News and find more relevant similar medical terms, and are closer to human experts' judgments, compared to these trained on GloVe and Google News. Second, there does not exist a consistent global ranking of word embedding quality for downstream biomedical NLP applications. However, adding word embeddings as extra features will improve results on most downstream tasks. Finally, the word embeddings trained on biomedical domain corpora do not necessarily have better performance than those trained on other general domain corpora for any downstream biomedical NLP tasks.

Index Terms

word embeddings; natural language processing; biomedical application; deep learning;

I. INTRODUCTION

Neural word embeddings have been widely used in Natural Language Processing (NLP) applications as they provide vector representations of words capturing the semantic properties of words and the linguistic relationship between words [1], [2], [3]. There has been an increasing number of studies applying word embeddings in common NLP tasks, such as information extraction (IE) [4], [5], [6], information retrieval (IR) [7], sentiment analysis [8], [9], question answering[10], [11], and text summarization [12], [13]. Recently in the biomedical domain word embeddings have been remarkably utilized in applications like biomedical named entity recognition (NER) [14], [15], medical synonym extraction[16], relation extraction (RE) including chemical-disease relation [17], drug-drug interaction [18], [19] and protein-protein interaction [20], biomedical IR [21], [22] and medical abbreviation disambiguation [23].

Many biomedical applications use task corpora to train word embeddings or use external data resources such as Wikipedia [21], [24] to train word embeddings based on an implicit assumption that the external resources contain the knowledge that could be used to enhance domain tasks [25], [26], [27]. A number of pre-trained word embeddings are publicly available, such as Google News embeddings [1] and GloVe embeddings [2]. These embeddings could capture semantics of general English words from a large corpus. However, one question remains unanswered: Do we need to train word embeddings for a specific NLP task since there are a number of public pre-trained word embeddings? This question becomes more significant for biomedical applications, and particularly more important for the clinical domain. The reason is that few electrical health records (EHRs) data are publicly available due to Health Insurance Portability and Accountability Act (HIPAA) rule while a big volume of biomedical literature data is available online. However, there has been little work on evaluating the word embeddings trained from these textual resources for biomedical applications, to the best of our knowledge.

In this study, we provide an empirical evaluation of word embeddings trained from four different resources, namely clinical notes, biomedical publications, Wikipedia, and news. For the former two resources, we utilized

---

[1]https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit

[2]https://nlp.stanford.edu/projects/glove/

clinical notes from the EHR system at Mayo Clinic and biomedical publications from the PubMed Central (PMC)[3] to train word embeddings separately. For the latter two resources, we used publicly available pre-trained word embeddings, GloVe and Google News. We performed the evaluation qualitatively and quantitatively. In qualitative evaluation, we manually inspected five most similar medical words to a given set of target medical words, and then analyzed word embeddings through the visualization of those word embeddings. In quantitative evaluation, we conducted both intrinsic and extrinsic evaluation. Intrinsic evaluation directly tested semantic relationships between medical words using four published datasets for measuring semantic similarity between medical terms, i.e., Pedersen [28], Hliaoutakis [29], MayoSRS [30], and UMNSRS [31], [32]. In extrinsic evaluation, we applied word embeddings to downstream NLP applications in the biomedical domain including clinical IE, biomedical IR, and RE, and measured the performance of word embeddings.

II. RELATED WORK

Due to the successful usage of word embeddings in a variety of NLP applications, there exists recent work on evaluation of word embeddings in representing word semantics. Most of the previous work focuses on evaluating word embeddings generated by different approaches. Baroni et al [33] presented the first systematic evaluation of word embeddings generated by count models (using DISSECT[4] on a corpus of 2.8 billion tokens constructed by concatenating ukWaC[5] the English GloVe[6] and the British National Corpus[7]), CBOW [1] (using word2vec[8] on the same corpus as DISSECT), Distributional Memory model[9] (on the same corpus as DISSECT), and Collobert and Weston model[10] (on the GloVe), and tested them on fourteen benchmarks in five categories: semantic relatedness (a dataset of semantic benchmarks constructed by asking human subjects to rate the degree of semantic similarity or relatedness between two words on a numerical scale, such as Rubenstein and Goodenough's dataset [34] and WordSim353 [35]), synonym detection (a dataset containing 80 multiple-choice questions that pair a target term with 4 synonym candidates, such as TOEFL dataset [36]), concept categorization (given a set of nominal concepts, the task is to group them into natural categories, such as Almuhareb-Poesio dataset [37]), selectional preferences (a dataset containing verb-noun pairs that were rated by subjects for the typicality of the noun as a subject or object of the verb, such as Ulrike Pado's dataset [38]), and analogy (a dataset containing semantic and syntactic analogy questions, such as Mikolov's dataset [1]). They found that the word2vec model, CBOW, performed the best for almost all the tasks. Schnabel et al [39] trained the CBOW model of word2vec [1], C&W embeddings [40], Hellinger PCA [41], GloVe [42], TSCCA [43] and Sparse Random Projections [44] on a 2008 GloVe dump, and

---

[3]http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[4]http://clic.cimec.unitn.it/composes/

[5]http://wacky.sslmit.unibo.it/

[6]http://en.GloVe.org/

[7]http://www.natcorp.ox.ac.uk/

[8]https://code.google.com/p/word2vec/

[9]http://clic.cimec.unitn.it/dm/

[10]http://ronan.collobert.com/senna/

tested on the same fourteen datasets. They also found the CBOW outperformed other embeddings on 10 datasets. In addition to this intrinsic evaluation, they conducted extrinsic evaluation by using the embeddings as input features to two downstream tasks, namely noun phrase chunking and sentiment classification, and found the results of CBOW were also among the best. Ghannay et al [45] conducted a similar intrinsic evaluation, they additionally evaluated the skip-gram models of word2vec [1], CSLM word embeddings [46], dependency-based word embeddings [3], and combined word embeddings on NLP tasks (i.e., Part-Of-Speech Tagging, chunking, named entity recognition, mention detection) and linguistic tasks using Mikolov's dataset [1] and the WordSim353 dataset[35]. They trained these word embeddings on the Gigaword corpus composed of 4 billion words and found that the dependency-based word embeddings gave the best performance on the NLP tasks and combination of the embeddings yielded significant improvement. In Nayak et al's study [47], they recommended that the evaluation should test both syntactic and semantic properties of the word embeddings and that the tasks should be closer to real-word applications. However, none of these studies evaluate word embeddings in the biomedical domain and none of these datasets focus on biomedical data.

As most of the aforementioned studies evaluate word embeddings in the general (i.e., non-medical) NLP domain, only one recent paper by Pakhomov et al [32] is about evaluating word embeddings in the biomedical domain, to the best of our knowledge. They trained the CBOW model on two biomedical corpora, namely clinical notes and biomedical publications, and one general English corpora, namely GloVe. The word embeddings were evaluated on subsets of UMNSRS dataset, which consisted of pairs of medical terms with the similarity of each pair assessed by medical experts, and on a document retrieval task and a word sense disambiguation task. They found that the semantics captured by the embeddings computed from biomedical publications were on par with that from clinical notes. Inspired by this work, we would like to conduct a complementary study to extend their evaluation of word embeddings by 1) utilizing four datasets to evaluate word embeddings on capturing medical term semantics; 2) conducting a qualitative evaluation; and 3) examining word embeddings on more biomedical application.

In this work, we provide a comparison of the quality of word embeddings trained separately from different resources. Specifically, our contributions are:

1) We performed qualitative evaluation where we manually inspected five most similar medical words to a given set of target medical words and plotted a visualization of selected medical words from those word embeddings.

2) We performed quantitative evaluation, including extrinsic and intrinsic evaluation. In the intrinsic evaluation, we used four published datasets for measuring semantic similarity between medical terms.

3) In extrinsic evaluation, we evaluated word embeddings by applying them to three publicly shared biomedical tasks, including biomedical IR, NER, and RE, and one institutional clinical NLP task.

III. Word Embeddings

Since it has been shown that the word2vec outperforms other approaches in generating good embeddings in general NLP tasks [33], [39], the skip-gram model of word2vec is utilized as the approach for generating word embeddings in this paper. Since no evidence shows the CBOW outperforms the skip-gram model or vice versa, we arbitrarily chose the skip-gram model of word2vec.

A. Brief Introduction

Word embeddings can be represented as a mapping $V \to \mathbb{R}^D : w \mapsto \theta$ which maps a word w from a vocabulary V to a real-valued vector θ in an embedding space with a dimension of D. The skip-gram model is an architecture proposed by Mikolov et al [1], which trains embeddings using the negative-sampling procedure. It constructed with the focus word as the single input vector, and the target contextual words are at the output layer. Negative-sampling updates a sample of output words per iteration, and the target output words should be kept in the sample and gets updated while a few non-target words are added as negative samples. Mathematically, given a target word w and its contextual word h, the goal is to maximize the log-likelihood on the training data, i.e.,

$$\max J = \log P(h|w),$$

where $P(h|w)$ is the conditional probability in the neural probabilistic language model that is usually defined as:

$$P(h|w) = \sigma(score(w,h)) = \frac{e^{score(w,h)}}{\sum_{w \in V} e^{score(w,h)}},$$

where $\sigma(\dot)$ is a softmax function that normalize real vector into a probability vector. Accordingly, the log-likelihood function can be written as:

$$J = score(w,h) - \log(\sum_{w \in V} e^{score(w,h)}).$$

Negative-sample is adopted here to avoid expensive computation over $|V|$ words, i.e.,

$$J' = \sum_{w,h \in D} \log Q_\theta(D=1|w,h) + \sum_{w,h \in D'} \log Q_\theta(D=0|w,h),$$

where D is the observed data, D' is the unobserved data, θ is the embedding vector, and $Q_\theta(D=1|w,h)$ is the probability of w and h being observed. The word embeddings can be computed by maximizing the log-likelihood function.

B. Parameter Settings

We tested dimensions of 20, 60 and 100 for word embeddings trained on EHR and PubMed and chose 100 for EHR and 60 for PubMed according to their performance in our intrinsic evaluation. Similarly, we chose the dimension of 100 for GloVe, and that of 300 for Google News since only 300 dimension was publicly available for Google News. The whole results of using different dimensions for word embeddings are provided in Appendix A. For training word embeddings on the EHR and PubMed, we set the window size to 5 words, the minimum word frequency to 7 (i.e., the words occurred less than 7 times in the corpus were ignored), and the negative sampling to 5. These parameters were selected based on previous studies [1], [3], [19].

IV. DATA

A. Text Corpora

We compared word embeddings computed from four different kinds of corpora. The first corpus, denoted as EHR, was from the Electronic Health Record system at Mayo Clinic. It contains textual clinical notes for a cohort

of 113k patients receiving their primary care at Mayo Clinic, spanning a period of 15 years from 1998 to 2013. The vocabulary size $|V|$ of this corpus is 103k. The second corpus, denoted as PubMed, is from a snapshot of the Open Access Subset of PMC in 2016. PMC is an online digital database of freely available full-text biomedical literature. The PubMed contains 1.25 million biomedical articles and 2 million distinct words (i.e., $|V|$). As comparisons, two additional public pre-trained word embeddings from two general English resources, i.e., Google News embeddings [11] and GloVe embeddings[12], were also considered in the evaluation. The Google News embeddings have embeddings for 3 million words from Google News, trained using the wor2vec [1]. The GloVe embeddings have embeddings for 400k words from a snapshot of Wikipedia in 2014 and Gigaword Fifth Edition[13], trained using the GloVe model [42].

B. Pre-processing

The PubMed was pre-processed minimally by removing punctuations (one exception is that we replaced '-' by '_' if two words were connected by '-' and we treated them as one word), lowercasing, and replacing all digits with "7". We conducted additional pre-processing on the EHR since the narratives written by physicians are more sparse than research articles. Specifically, the section of "Family history" in the corpus was removed if it was semi-structured [48]. See an example of the "Family history" section in Table I

TABLE I: An example of the "Family history" section.

MOTHER
Stroke/TIA
BROTHERS
4 brothers alive 1 brother deceased
SISTERS
2 sisters alive
DAUGHTERS
1 daughter alive
Heart disease

The section of "Vital Signs" was totally removed since it did not contain contextual information for training word embeddings. See an example of the "Vital Signs" section in Table II.

TABLE II: An example of the "Family history" section.

Height: 149.1 cm. Weight: 44.5 kg. BSA(G): 1.3573 M2. BMI: 20.02 KG/M2.

Note that these pre-processing strategies are specific for the clinical notes at Mayo Clinic. Moreover, we replace all text contractions with their respective complete text. For example, "can't" is replaced with "can not". For clinical

---

[11]https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing

[12]http://nlp.stanford.edu/data/glove.6B.zip

[13]https://catalog.ldc.upenn.edu/LDC2011T07

corpus, we removed all the clinical notes metadata and note sections headers, dates, phone numbers, weight and height information and punctuations from clinical corpus. For PubMed articles, we removed all the websites url, email addresses, twitter handler and punctuations.

V. QUALITATIVE EVALUATION

The first experiment is qualitative evaluation by manually inspecting the five most similar words to a given target word. We used the commonly employed cosine similarity to calculate the most similar words. Suppose w_1 and w_2 are two words, the similarity between w_1 and w_2 is defined as

$$\text{similarity}(w_1, w_2) = \frac{\theta_1 \cdot \theta_2}{\|\theta_1\|\|\theta_2\|}, \tag{1}$$

where θ_1 and θ_2 are vector representations for w_1 and w_2 in the embedding space, respectively. If the target word is a medical phrase s_1 consisting of multiple words, i.e., $s_1 = w_1, w_2, ..., w_n$, the similarity function becomes

$$\text{similarity}(s_1, w_2) = \frac{\Theta_1 \dot{\theta}_2}{\|\Theta_1\|\|\theta_2\|} \tag{2}$$

where $\Theta_1 = \frac{1}{n}\sum_i^n \theta_i$ is the representation for s_1 in the embedding space. This is different from Pakhomov et al's study [32] where only single word terms were considered.

Table III lists eight target words in three medical categories, i.e., disorder, symptom and drug, and their five most similar words induced by different word embeddings.

For the first target word describing an disorder, *diabetes*, EHR and PubMed find its synonym, *mellitus*, in the most similar words while GloVe and Google News fail to find it. EHR finds two terms related to co-morbidities of *diabetes*, *cholesterolemia* and *dyslipidemia*, and a most common modifier term, *uncontrolled*. PubMed finds terms relevant to co-existing conditions for *diabetes*, such as *cardiovascular* (very possibly from *cardiovascular disease*), *nonalcoholic* (very possibly from *nonalcoholic fatty liver disease*), *obesity*, and *polycystic* (very possibly from *polycystic ovary syndrome* which is a hyperandrogenic disorder that is associated with a high-risk of development of Type 2 diabetes). Most of these terms are from medical research topics and thus occur in the PubMed articles. GloVe finds two related terms, *hypertension* and *obesity*, while other three terms, i.e., *arthritis*, *cancer* and *alzheimer*, are less relevant disease names. Google News finds two morphological terms, *diabetics* and *diabetic*, relevant to the target words, one synonym, *diabetes_mellitus*, and one related disease name, *heart disease*. We can draw similar conclusions for the second and third disorder words.

The *dyspnea* example in the symptom category demonstrates the advantage of EHR and PubMed. EHR finds *palpitations*, a common cause of *dyspnea*, and *orthopnea*, *exertional*, and *doe* (dyspnea on exertion) are synonyms or specific conditions for *dyspnea*. PubMed finds related symptoms, *sweats* and *orthopnea*, a synonym *breathlessness*, a relevant disorder *hypotension*, and a term relevant to the symptom *rhonchi*. Wikipedia finds synonyms *shortness* and *breathlessness*, and less relevant symptoms *cyanosis* and *photophobia*. Google News finds less relevant symptoms *pruritus* and *rhinorrhea* and less relevant disease *nasopharyngitis*. Similar observations can be found for *sore throat* and *low blood pressure* as well.

The drug category further differentials the word embeddings. In *opioid*, EHR finds *opiate*, *benzodiazepine*, *sedative*, *polypharmacy*, which are very relevant medications. PubMed finds *nmda_receptor*, *affective_motivational*,
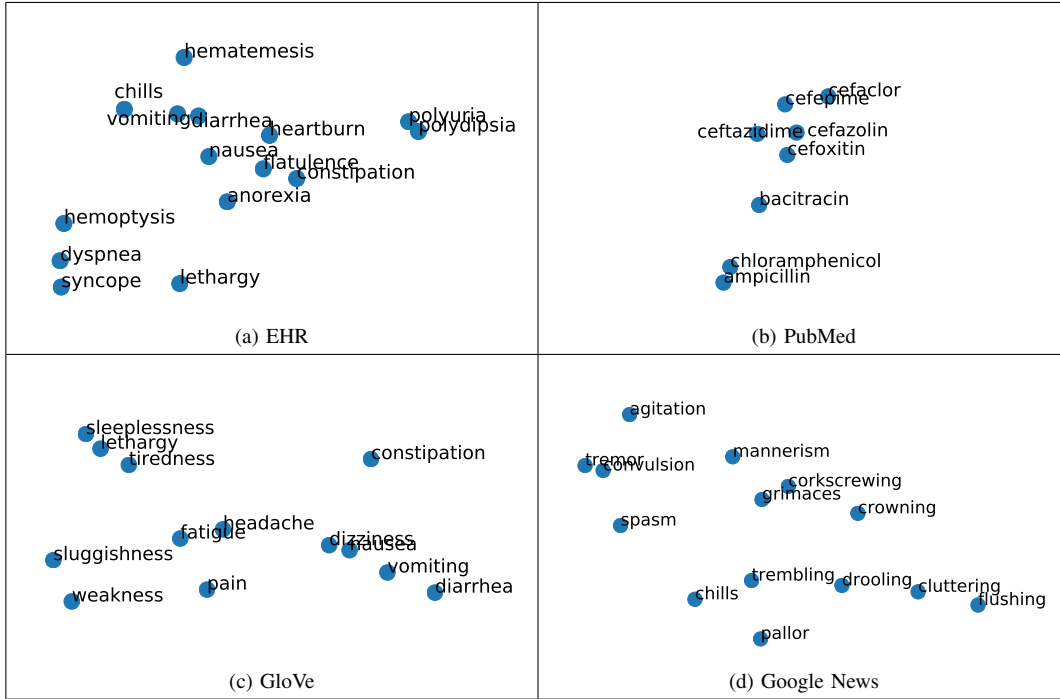
Fig. 1: Example clusters in different word embeddings.

naloxone_precipitated, *hyperlocomotion*, which are related to the mechanism of action of *opioid*. GloVe finds *analgesic* and less relevant *anti-inflammatory*, and Google News finds *opioid*-related phrases and relevant term *antipsychotics*. In *aspirin*, EHR also finds very clinically relevant used terms and PubMed finds relevant terms in research articles while GloVe and Google News only find medication names.

It is obvious shown from these target words that EHR and PubMed can capture the semantics of medical terms better than GloVe and Google News and find more relevant similar medical terms. However, EHR and PubMed find similar medical terms from different perspectives due to their different focuses. EHR contains clinical narratives and thus it is closer to clinical language. Yet, it contains terms with different morphologies and even typos, such as *melitis*, *caner* and *thraot* as listed in Table III. Differently, PubMed contains more medical terminology used in medical articles, and finds similar words mostly from a medical research perspective.

In order to visualize the semantics of medical terms captured by different word embeddings, we extracted 377 medical terms from the UMNSRS dataset [31], [32] and plotted the word embeddings for these medical terms using t-SNE [49]. Example clusters of medical terms in the word embeddings are shown in Figure 1. Figure 1a depicts a cluster of symptoms in word embeddings trained on EHR, such as "heartburn", "vomiting" and "nausea". As shown in Figure 1b, a cluster of an antibiotic medications can be obviously observed from PubMed embeddings, such as "bacitracin", "cefoxitin", and "chloramphenicol". Figures 1c and 1d illustrate clusters of symptoms in word embeddings trained from GloVe and Google News, respectively. Since we did not employ any clustering method, these clusters were intuitively observed from the visualization. The visualization of the whole 377 medical terms can be found in Appendix B.

TABLE III: Selected medical words and their five most similar words induced by different word embeddings.

Semantic Category	Target Word	EHR	PubMed	GloVe	Google News
Disorder	diabetes	mellitus, uncontrolled, cholesterolemia, dyslipidemia, melitis	cardiovascular, nonalcoholic, obesity, mellitus, polycystic	hypertension, obesity, arthritis, cancer, alzheimer	diabetics, hypertension, diabetic, diabetes_mellitus, heart_disease
	peptic ulcer disease	scleroderma, duodenal, crohn, gastroduodenal, diverticular	gastritis, alcoholism, rheumatic, ischaemic, nephropathy	ulcers, arthritis, diseases, diabetes, stomach	ichen_planus, Candida_infection, vaginal_yeast_infections, oral_thrush, dermopathy
	colon cancer	breast, ovarian, prostate, postmenopausally, caner	breast, mcf, cancers, tumor_suppressing, downregulation	breast, prostate, cancers, tumor, liver	breast, prostate, tumor, pre_cancerous_lesion, cancerous_polyp
Symptom	dyspnea	palpitations, orthopnea, exertional, doe, dyspnoea	sweats, orthopnea, breathlessness, hypotension, rhonchi	shortness, breathlessness, cyanosis, photophobia, faintness	dyspnoea, pruritus, nasopharyngitis, symptom_severity, rhinorrhea
	sore throat	scratchy, thoat, cough, runny, thraot	runny, rhinorrhea, myalgia, swab_fecal, nose	shoulder, stomach, nose, chest, neck	soreness, bruised, inflammed, contusion, sore_triceps
	low blood pressure	readings, pressue, presssure, bptru, systolically	dose, cardio_ankle, ncbav, preload, gr	because, result, high, enough, higher	splattering_tombstones, Zapping_nerves_helps, pressue, Marblehead_Swampscott_VNA, pill_Norvasc
Drug	opioid	opiate, benzodiazepine, opioids, sedative, polypharmacy	opioids, nmda_receptor, affective_motivational, naloxone_precipitated, hyperlocomotion	analgesic, opiate, opioids, anti-inflammatory, analgesics	opioids, opioid_analgesics, opioid_painkillers, antipsychotics, tricyclic_antidepressants
	aspirin	ecotrin, uncoated, nonenteric, effient, onk	chads, vasc, newer, cha, angina	ibuprofen, tamoxifen, pills, statins, medication	dose_aspirin, ibuprofen, statins, statin, calcium_supplements

VI. QUANTITATIVE EVALUATION

We tested word embeddings with quantitative evaluation to show the qualitative differences between them. We utilized extrinsic and intrinsic evaluation, where the former used four published datasets for measuring semantic similarity between medical terms and the latter used downstream biomedical tasks to evaluate word embeddings.

A. Intrinsic Evaluation

We tested word embeddings on four published biomedical datasets commonly used to measure semantic similarity between medical terms. The first is Pedersen's dataset [28] that consists of 30 medical term pairs that were scored by physician experts according to their relatedness. The second is Hliaoutakis's dataset [29] consisting of 34 medical term pairs with similarity scores obtained by human judgments. The third one is MayoSRS dataset, developed by Pakhomov et al [30], which consists of 101 clinical term pairs whose relatedness was determined by nine medical coders and three physicians from the Mayo Clinic. The relatedness of each term pair was assessed based on a four point scale: (4.0) practically synonymous, (3.0) related, (2.0) marginally related and (1.0) unrelated. We evaluated word embeddings on the mean score of the physicians and medical coders. The fourth is UMNSRS similarity dataset, also developed by Pakhomov et al [31], which consists of 566 medical term pairs whose semantic similarity was determined independently by eight medical residents from the University of Minnesota Medical School. The similarity and relatedness of each term pair was annotated based on a continuous scale by having the resident touch a bar on a touch sensitive computer screen to indicate the degree of similarity or relatedness.

For each pair of medical terms in the testing datasets, we used Equations 1 and 2 to calculate the semantic similarity for each pair in the embeddings space. Since some medical terms might not exist in the vocabulary of word embeddings, we used fastText [50] to compute word vectors for these out-of-vocabulary medical terms. Pearson correlation coefficient was employed to calculate the correlation between similarity scores from human judgments and those from word embeddings. Table IV lists the results for the four datasets. Overall, the semantic similarity captured by the word embeddings trained on EHR are closer to human experts' judgments, compared with other word embeddings. PubMed performs worse than EHR but has a comparative result for the UMNSRS dataset. GloVe and Google News are inferior to EHR and PubMed and perform similarly in representing medical semantics. Note that the four datasets and corresponding semantic similarity scores from both human experts and word embeddings are provided in the supplementary Excel file.

TABLE IV: Pearson correlation coefficient for four datasets.

Dataset	EHR	PubMed	GloVe	Google News
Pedersen's	**0.632**	0.569	0.403	0.357
Hliaoutakis's	**0.482**	0.311	0.247	0.243
MayoSRS	**0.412**	0.300	0.082	0.084
UMNSRS	**0.440**	0.404	0.177	0.154

B. Extrinsic Evaluation

Extrinsic evaluations are used to measure the contribution of word embeddings to specific biomedical tasks. In this evaluation, we applied word embeddings to three prevalent shared tasks: clinical IE, biomedical IR, and RE.

1) Clinical Information Extraction: We evaluated word embeddings on two clinical IE tasks. The first experiment is a shared task while the second is an institutional task. We would like to examine whether our local institutional word embeddings are better than external pre-trained word embeddings.

In the first experiment, we evaluated the word embeddings on an institutional information extraction task at Mayo Clinic. In this task, a set of 1000 radiology reports was examined to detect whether a hand and figure/wrist fracture could be identified. Reports were drawn from a cohort of residents of Olmsted County, aged 18 or older, who experienced fractures in 2009-2011. Each report was annotated by a medical expert with multiple years of experience abstracting fractures by assigning "1" if a hand and figure/wrist fracture was found, or "0" otherwise.

In our experiment, word embeddings were employed as features for machine learning models and evaluated by precision, recall and F1 scores [52]. For a clinical document $d = \{w_1, w_2, .., w_M\}$ where $w_i, i = 1, 2, ..., M$ is the ith word and M is the total number of words in this document, the feature vector x of document d is defined by

$$x = \frac{1}{M} \sum_i^M x_i,$$

where x_i is the embedding vector for word w_i from the word embedding matrix. Then x was utilized as input to the machine learning models. A prevalent machine learning model, Support Vector Machine (SVM), was tested in this experiment. We performed 10-fold cross validation on the dataset and reported means of precision, recall and F1 scores from the cross validation. As comparison, the baseline method used term frequency features as input. Table V reports the means of precision, recall, and F1 score from the cross validation.

The word embeddings trained from EHR are superior to other word embeddings on all metrics (precision: 0.974, recall: 0.972, F1 score: 0.972). The fracture dataset in this experiment is curated from the same EHR system as the EHR corpus used to train word embeddings, and thus they have identical sublanguage characteristics embedded in word embeddings. The word embeddings trained on PubMed also have comparable results (precision: 0.946, recall: 0.943, F1 score: 0.942). Since this task is a medical task with specific medical terminologies, word embeddings trained on Google News have the worst performance. However, the word embeddings trained on GloVe are close to those trained on EHR with 0.02 difference on F1 score without statistically significance (p<0.01). This experiment shows that word embeddings trained on local corpus have the best performance for a local task but those trained on external Wikipedia corpus also have comparable performance.

TABLE V: Results on the institutional fracture extraction task.

Metric	baseline	EHR	PubMed	GloVe	Google News
Precision	0.612	**0.974**	0.946	0.951	0.809
Recall	0.612	**0.972**	0.943	0.950	0.856
F1 score	0.609	**0.972**	0.942	0.950	0.823

We secondly tested word embeddings on the i2b2 (Informatics for Integrating Biology to the Bedside) 2006 smoking status extraction shared task [51]. Participants of this task were asked to develop automatic NLP systems to determine the smoking status of patients from their discharge records in Partners HealthCare. For each discharge record, an automatic system should be able to categorize it into five pre-determined smoking status categories: past smoker, current smoker, smoker, non-smoker, and unknown, where a past and current smoker are distinguished based on temporal expressions in the patient's medical records. The dataset contains a total of 389 documents, including 35 documents of current smoker, 66 of non-smoker, 36 of past smoker, and 252 of unknown.

The experimental results are shown in Table VI. First, it is obviously shown that word embedding features perform better than term frequency features due to the semantics embedded in word embeddings. From Table VI, we can observe that using word embeddings trained on EHR has the best performance with a F1 score of 0.900. This result is interesting since the EHR corpus used to train the word embeddings is from a different institution. The reason might be that the smoking dataset has similar sublanguage characteristics as our EHR corpus. This result indicates that effective word embeddings can be shared across institutions for clinical NLP tasks. Another interesting observation is that the performance of word embeddings trained on Google News is close to that on EHR corpus with a comparable F1 score and a better recall. The performance difference is not statistically significant ($p < 0.01$). This implies that word embeddings trained on a public dataset may not be definitely inferior to these trained on a medical specific dataset for a medical IE task. The likely cause is that the terminology used in smoking status extraction task also appears frequently in news, such as medications and advice for smokers.

TABLE VI: Results on the i2b2 2006 smoking status extraction task.

Metric	baseline	EHR	PubMed	GloVe	Google News
Precision	0.692	**0.919**	0.878	0.893	0.910
Recall	0.486	0.903	0.871	0.889	**0.905**
F1 score	0.539	**0.900**	0.867	0.884	0.897

2) Biomedical Information Retrieval: We utilized Text REtreival Conference 2016 Clinical Decision Support (TREC 2016 CDS) track as the experimental data sets for evaluating word embeddings for biomedical IR. TREC 2016 CDS track focuses on biomedical literature retrieval that helps physicians find the precise literature information and make the best clinical decision at the point of care [53]. EHRs from MIMIC-III data set [54] were utilized to generate the query topics. Those topics are categorized into three most common types, namely *Diagnosis*, *Test* and *Treatment*, according to physicians' information needs, and 10 topics are provided for each type. Each topic is comprised of a *note* field (admission note), a *description* field (jargons and clinical abbreviations are removed) and a *summary* field (simplified version of the description). The participants are required to use only one of these three fields in their submissions and at least one submission must utilize the *note* field. Submitted systems should retrieve relevant biomedical articles from a given PMC article collection for each given query topic to answer three corresponding clinical questions: *What is the patient's diagnosis? What tests should the patient receive? How should the patient be treated?*. Each IR system can retrieve up to 1000 documents per query.

In order to make comparison as fair as possible, we first implemented an IR system as a baseline system following [22] and then employed the simplest query expansion using word embeddings. That is, we expanded each query term with five most similar terms from word embeddings. Indri [55] was utilized as our indexing and retrieval tool. The preprocessing included stopword removal and Porter stemming. The stopword list was based on the PubMed stopwords [14]. The *article-id*, *title*, *abstract* and *body* fields of each document were indexed. Language models with two-stage smoothing [56] was used to obtain all the retrieval results. Official metrics, Inferred Normalized Discounted Cumulated Gain (infNDCG) [57], Inferred Average Precision (infAP) [57], Precision at 10 (P@10) and Mean Average Precision (MAP), were utilized to measure the performance.

Table VII shows the results on the TREC 2016 CDS track. It is interesting that query expansion using word embeddings almost does not improve retrieval performance and even worsen the performance when infAP and MAP were used as metrics. By comparing the retrieval performance of different word embeddings, we observe that EHR and PubMed perform slight better than GloVe and Google News without statistical significance ($p<0.01$). This result implies that applying different word embeddings trained from different resources has almost no significant difference for biomedical IR tasks. We also note that this may also be due to the query expansion method used in our evaluation.

TABLE VII: Results on the TREC 2016 CDS track.

Metric	baseline	EHR	PubMed	GloVe	Google News
infNDCG	0.249	**0.250**	0.249	0.249	0.238
infAP	**0.058**	0.056	0.055	0.051	0.052
P@10	0.247	0.243	**0.248**	0.233	0.243
MAP	**0.067**	0.063	0.065	0.063	0.059

3) Relation Extraction: Drug-drug interaction (DDI) extraction is a specific RE task in the biomedical domain. DDI is an unexpected change in a drug's effect on the human body when the drug and a second drug are co-prescribed and taken together. Automatic extracting DDI information from literature is a challenging and important research topic since the volume of the published literature grows rapidly and greatly. In this experiment, we evaluate word embeddings on DDIExtraction 2013 challenge corpus [58]. The dataset for DDIExtraction 2013 was composed of sentences describing DDIs from the DrugBank database and MedLine abstracts. In this dataset, drug entities and DDIs were annotated at the sentence level and each sentence could contain two or more drugs. An RE system should be able to automatically extract DDI drug pairs from a sentence. We leveraged the baseline system introduced in [19] and evaluated word embeddings by concatenating to the baseline features. We utilized Random Forest [59] as the classifier since it had the best performance in [19]. F1 score was used as the evaluation metric.

Table VIII shows the results on the DDIExtraction 2013 challenge. We can see that the overall performance of word embeddings trained on Google News is the best. The reason is that the semantics of general English terms in the context of drug mentions are more important for determining the drug interactions. For example, in the

---

[14]http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/

sentence "Acarbose may interact with metformin", the term "interact" is crucial to classify the relation. Since these crucial terms are generally not medical terminology, word embeddings trained on Google News where the corpus represents general English are able to capture the semantics of these terms. However, the superiority of Google News to other corpora is minor for this task. Another interesting observation is that word embeddings trained from PubMed have the best performance for the DrugBank corpus while these from Google News perform the best for the MedLine corpus. Though MedLine abstracts are from PubMed articles, this result shows that word embeddings trained from the same corpus are not necessarily superior to other embeddings.

TABLE VIII: F1 scores on the DDIExtraction 2013 challenge.

Category	baseline	EHR	PubMed	GloVe	Google News
DrugBank (5265 pairs)	0.590	0.708	**0.715**	0.714	0.705
MedLine (451 pairs)	0.690	0.696	0.690	0.699	**0.708**
Total (5716 pairs)	0.760	0.789	0.788	0.787	**0.790**

VII. CONCLUSION AND DISCUSSION

In this study, we provided an empirical evaluation of word embeddings trained from four different resources, namely clinical notes, biomedical publications, Wikipedia, and news. We performed the evaluation qualitatively and quantitatively. In qualitative evaluation, we selected a set of medical words and evaluated the five most similar medical words. We then analyzed word embeddings through the visualization of those word embeddings. We conducted extrinsic and intrinsic evaluation for quantitative evaluation. Intrinsic evaluation directly tested semantic relationships between medical words using four published datasets for measuring semantic similarity between medical terms while extrinsic evaluation evaluated word embeddings in three downstream biomedical NLP applications, i.e., clinical IE, biomedical IR, and RE.

Based on the evaluation results, we can draw the following conclusions. First, the word embeddings trained on EHR and PubMed can capture the semantics of medical terms better than those trained on GloVe and Google News and find more relevant similar medical terms. However, EHR finds similar terms vis a vis clinical language while PubMed contains more medical terminology used in medical articles, and finds similar words mostly from a medical research perspective. Second, the medical semantic similarity captured by the word embeddings trained on EHR and PubMed are closer to human experts' judgments, compared to these trained on GloVe and Google News. Third, there does not exist a consistent global ranking of word embedding quality for downstream biomedical NLP applications. However, adding word embeddings as extra features will improve results on most downstream tasks. Finally, word embeddings trained from biomedical domain corpora do not necessarily have better performance than those trained on other general domain corpora. That is, there might be no significant difference when word embeddings trained from an out-domain corpus were employed for a biomedical NLP application. However, the performance of word embeddings trained from a local corpus is better for local NLP tasks.

In a biomedical NLP application, our experiment implicitly show that applying word embeddings trained from corpora in a general domain, such as Wikipedia and news, is not significantly inferior to applying those obtained

from biomedical or clinical domain, which is usually difficult to access due to privacy. This result is consistent with but more general than the conclusion drawn in [32]. Thus, lack of access to a domain-specific corpus is not necessarily a barrier to the use of word embeddings in practical implementations.

As a future direction, we would like to evaluate word embeddings on more downstream biomedical NLP applications, such as medial named entity recognition and clinical note summarization. We will investigate whether different word embeddings represent language characteristics differently for a corpus, such as term frequency and medical concepts. We also want to assess word embeddings across health care institutions using different EHR systems and investigate how sublanguage characteristics affect the portability of word embeddings. Moreover, we want to apply clustering methods on word embeddings and compare the word-level and concept-level difference between clusters of medical terms.

VIII. Acknowledgement

References

[1] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations." in *hlt-Naacl*, vol. 13, 2013, pp. 746–751.

[2] F. Liu, J. Chen, A. Jagannatha, and H. Yu, "Learning for biomedical information extraction: Methodological review of recent advances," *arXiv preprint arXiv:1606.07993*, 2016.

[3] O. Levy and Y. Goldberg, "Dependency-based word embeddings." in *ACL (2)*, 2014, pp. 302–308.

[4] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn *et al.*, "Clinical information extraction applications: A literature review," *Journal of biomedical informatics*, 2017.

[5] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao *et al.*, "Relation classification via convolutional deep neural network." in *COLING*, 2014, pp. 2335–2344.

[6] T. H. Nguyen and R. Grishman, "Employing word representations and regularization for domain adaptation of relation extraction." in *ACL (2)*, 2014, pp. 68–74.

[7] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones, "Word embedding based generalized language model for information retrieval," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 795–798.

[8] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification." in *ACL (1)*, 2014, pp. 1555–1565.

[9] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 142–150.

[10] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Advances in neural information processing systems*, 2015, pp. 2953–2961.

[11] L. Dong, F. Wei, M. Zhou, and K. Xu, "Question answering over freebase with multi-column convolutional neural networks." in *ACL (1)*, 2015, pp. 260–269.

[12] D. Yogatama, F. Liu, and N. A. Smith, "Extractive summarization by maximizing semantic volume." in *EMNLP*, 2015, pp. 1961–1966.

[13] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.

[14] B. Tang, H. Cao, X. Wang, Q. Chen, and H. Xu, "Evaluating word representation features in biomedical named entity recognition tasks," *BioMed research international*, vol. 2014, 2014.

[15] S. Liu, B. Tang, Q. Chen, and X. Wang, "Effects of semantic features on machine learning-based drug name recognition systems: word embeddings vs. manually constructed dictionaries," *Information*, vol. 6, no. 4, pp. 848–865, 2015.

[16] A. N. Jagannatha, J. Chen, and H. Yu, "Mining and ranking biomedical synonym candidates from wikipedia," in *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi)*, 2015, pp. 142–151.

[17] Z. Jiang, L. Jin, L. Li, M. Qin, C. Qu, J. Zheng, and D. Huang, "A crd-wel system for chemical-disease relations extraction," in *The fifth BioCreative challenge evaluation workshop*, 2015, pp. 317–326.

[18] 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, 2016.

[19] Y. Wang, S. Liu, M. Rastegar-Mojarad, L. Wang, F. Shen, F. Liu, and H. Liu, "Dependency embeddings and amr embeddings for drug-drug interaction extraction from biomedical texts," in *Proceedings of the 8th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017.

[20] Z. Jiang, L. Li, and D. Huang, "A general protein-protein interaction extraction architecture based on word representation and feature selection," *International Journal of Data Mining and Bioinformatics*, vol. 14, no. 3, pp. 276–291, 2016.

[21] S.-H. Jo and K.-S. Lee, "Cbnu at trec 2016 clinical decision support track," in *Text REtrieval Conference (TREC 2016)*, 2016.

[22] Y. Wang, M. Rastegar-Mojarad, R. K. Elayavilli, S. Liu, and H. Liu, "An ensemble model of clinical information extraction and information retrieval for clinical decision support." in *TREC*, 2016.

[23] Y. Wu, J. Xu, Y. Zhang, and H. Xu, "Clinical abbreviation disambiguation using neural word embeddings," in *Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP)*, 2015, pp. 171–176.

[24] H. Gurulingappa, L. Toldo, C. Schepers, A. Bauer, and G. Megaro, "Semi-supervised information retrieval system for clinical decision support." in *TREC*, 2016.

[25] F. Shen and Y. Lee, "Knowledge discovery from biomedical ontologies in cross domains," *PloS one*, vol. 11, no. 8, p. e0160005, 2016.

[26] F. Shen, H. Liu, S. Sohn, D. W. Larson, and Y. Lee, "Predicate oriented pattern analysis for biomedical knowledge discovery," *Intelligent Information Management*, vol. 8, no. 03, pp. 66–85, 2016.

[27] F. Shen, H. Liu, S. Sohn, D. Larson, and Y. Lee, "Bmqgen: Biomedical query generator for knowledge discovery," *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pp. 1092–1097, 2015.

[28] T. Pedersen, S. V. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *Journal of biomedical informatics*, vol. 40, no. 3, pp. 288–299, 2007.

[29] A. Hliaoutakis, "Semantic similarity measures in mesh ontology and their application to information retrieval on medline," *Master's thesis*, 2005.

[30] S. V. Pakhomov, T. Pedersen, B. McInnes, G. B. Melton, A. Ruggieri, and C. G. Chute, "Towards a framework for developing semantic relatedness reference standards," *Journal of biomedical informatics*, vol. 44, no. 2, pp. 251–265, 2011.

[31] 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*, vol. 2010. American Medical Informatics Association, 2010, p. 572.

[32] S. V. Pakhomov, G. Finley, R. McEwan, Y. Wang, and G. B. Melton, "Corpus domain effects on distributional semantic modeling of medical terms," *Bioinformatics*, vol. 32, no. 23, pp. 3635–3644, 2016.

[33] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors." in *ACL (1)*, 2014, pp. 238–247.

[34] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, vol. 8, no. 10, pp. 627–633, 1965.

[35] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 406–414.

[36] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104, no. 2, p. 211, 1997.

[37] A. Almuhareb, "Attributes in lexical acquisition," Ph.D. dissertation, University of Essex, 2006.

[38] S. Padó and M. Lapata, "Dependency-based construction of semantic space models," *Computational Linguistics*, vol. 33, no. 2, pp. 161–199, 2007.

[39] T. Schnabel, I. Labutov, D. M. Mimno, and T. Joachims, "Evaluation methods for unsupervised word embeddings." in *EMNLP*, 2015, pp. 298–307.

[40] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[41] R. Lebret and R. Collobert, "Word emdeddings through hellinger pca," *arXiv preprint arXiv:1312.5542*, 2013.

[42] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[43] P. Dhillon, J. Rodu, D. Foster, and L. Ungar, "Two step cca: A new spectral method for estimating vector models of words," *arXiv preprint arXiv:1206.6403*, 2012.

[44] P. Li, T. J. Hastie, and K. W. Church, "Very sparse random projections," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 287–296.

[45] S. Ghannay, B. Favre, Y. Esteve, and N. Camelin, "Word embedding evaluation and combination." in *LREC*, 2016.

[46] H. Schwenk, "Cslm-a modular open-source continuous space language modeling toolkit." in *INTERSPEECH*, 2013, pp. 1198–1202.

[47] N. Nayak, G. Angeli, and C. D. Manning, "Evaluating word embeddings using a representative suite of practical tasks," *ACL 2016*, p. 19, 2016.

[48] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Liu, F. Shen, and H. Liu, "Systematic analysis of free-text family history in electronic health record," *AMIA Summits on Translational Science Proceedings*, vol. 2017, p. 104, 2017.

[49] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[50] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[51] Ö. Uzuner, I. Goldstein, Y. Luo, and I. Kohane, "Identifying patient smoking status from medical discharge records," *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 14–24, 2008.

[52] Y. Wang, E. Atkinson, S. Amin, and H. Liu, "A distant supervision paradigm for clinical information extraction," 2018.

[53] K. Roberts, D. Demner-Fushman, E. M. Voorhees, and W. R. Hersh, "Overview of the trec 2016 clinical decision support track." in *TREC*, 2016.

[54] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, 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, 2016.

[55] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft, "Indri: A language model-based search engine for complex queries," in *Proceedings of the International Conference on Intelligent Analysis*, vol. 2, no. 6. Citeseer, 2005, pp. 2–6.

[56] C. Zhai and J. Lafferty, "Two-stage language models for information retrieval," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002, pp. 49–56.

[57] E. Yilmaz, E. Kanoulas, and J. A. Aslam, "A simple and efficient sampling method for estimating ap and ndcg," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 603–610.

[58] I. Segura Bedmar, P. Martínez, and M. Herrero Zazo, "Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)." Association for Computational Linguistics, 2013.

[59] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

APPENDIX A

RESULTS OF USING DIFFERENT DIMENSIONS FOR WORD EMBEDDINGS.

TABLE IX: Pearson correlation coefficient for four datasets.

Dataset	EHR (d=20)	EHR (d=60)	EHR (d=100)	PubMed (d=20)	PubMed (d=60)	PubMed (d=100)	GloVe (d=50)	GloVe (d=100)	Google News (d=300)
Pedersen's	0.390	0.542	0.632	0.304	0.569	0.363	0.334	0.403	0.357
Hliaoutakis's	0.333	0.417	0.482	0.117	0.311	0.164	0.159	0.247	0.243
MayoSRS	0.192	0.296	0.412	0.177	0.300	0.154	0.001	0.082	0.084
UMNSRS	0.310	0.375	0.440	0.295	0.404	0.396	0.190	0.177	0.154

Fig. 2: Visualization of word embeddings trained on EHR.

APPENDIX B
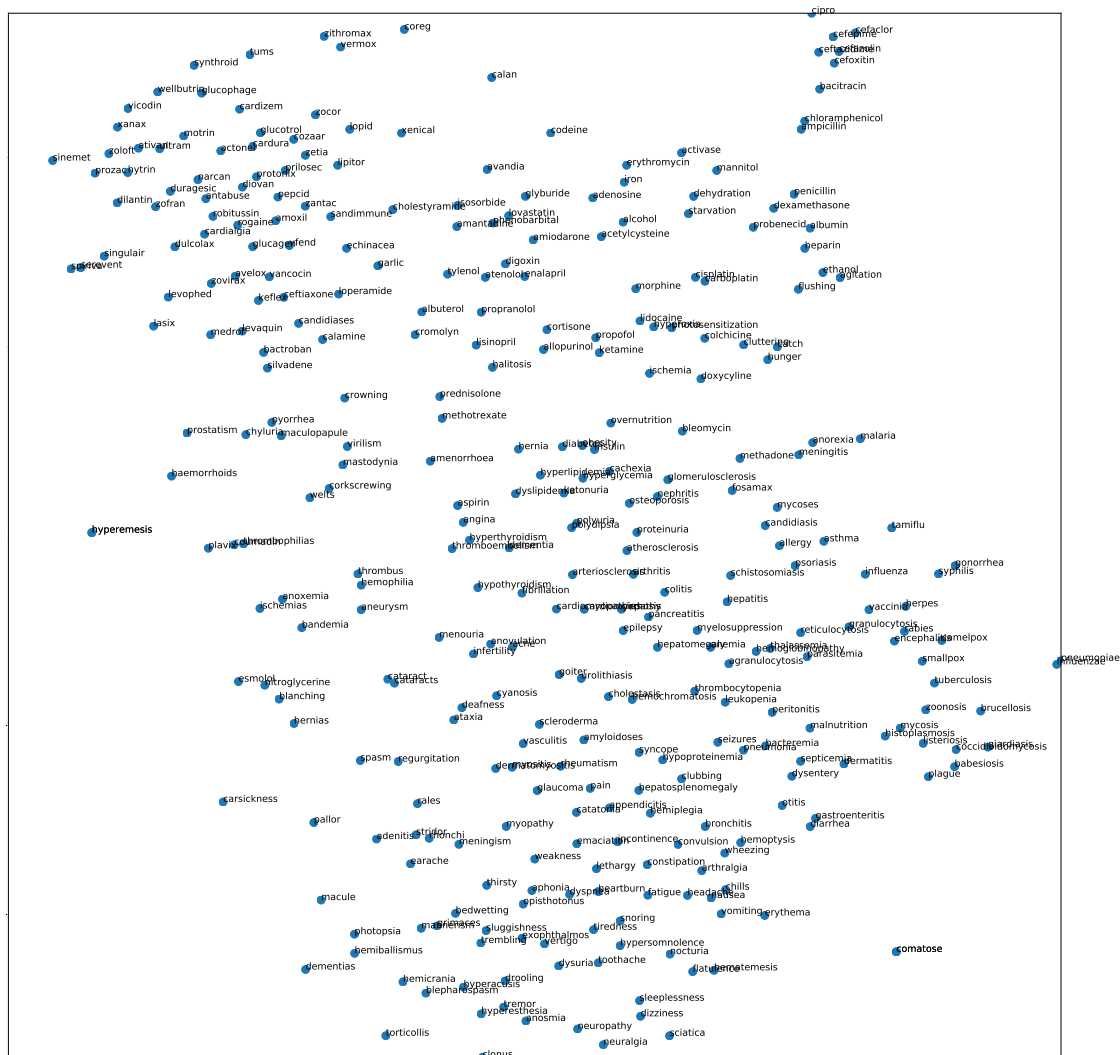
VISUALIZATION OF WORD EMBEDDINGS.

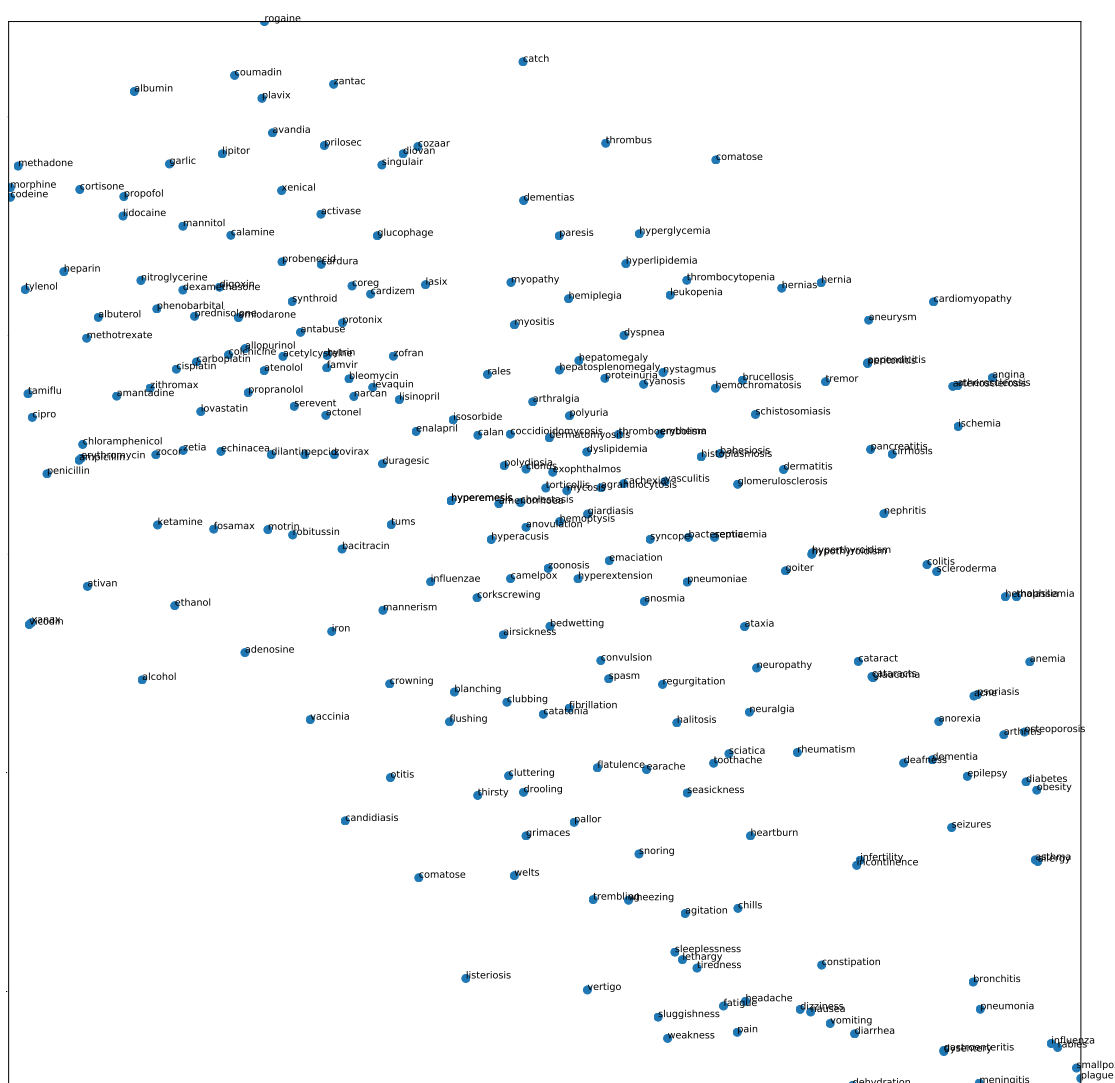Fig. 3: Visualization of word embeddings trained on PuBMed.

Fig. 4: Visualization of word embeddings trained on GloVe.

Fig. 5: Visualization of word embeddings trained on Google News.