MedDRA2Vec: Training Medical Graph Embeddings for Clinical NLP

Stanford CS224N Custom Project

Ayushi Tandel
Department of Computer Science
Stanford University
atandel@stanford.edu

Ariel Leong
Department of Computer Science
Stanford University
akl1@stanford.edu

Abstract

In recent years, advancements in natural language processing (NLP) and its applications have exploded across different domains, including speech recognition and translation. One domain that NLP has struggled to find a foothold in is the industry of biomedical and healthcare data. While vast amounts of data are available– encoded in texts, medical codes, ontologies, and patient data–researchers have struggled to develop reliable methods for applying deep learning to extract medical concepts and terms. Snomed2Vec suggested that deriving embeddings for medical terminology from biomedical knowledge graphs outputs better embeddings that are easier to reproduce and distribute [1]. Inspired by Snomed2Vec, our research project derives embeddings for medical terms from MedDRA, the Medical Dictionary for Regulatory Activities, using two different embedding methods, Poincare and Node2Vec. We evaluate these embeddings using cosine concept similarity and the accuracy of these embeddings in patient diagnosis tasks. We find that our MedDRA embeddings achieve, for the four relationships recorded in the MedDRA ontology, statistical power values of $-0.114, -0.062, 0.362$ and 0.158 for concept similarity for the Poincare embeddings, and values of $0.933, 0.712, 0.709$, and 0.524 for Node2Vec embeddings. For the patient diagnosis task, our Poincare embeddings achieve a highest accuracy of 0.343. The Node2Vec embeddings achieve a highest accuracy of 0.348. Our research shows that MedDRA embeddings are comparable, and even better in some cases, to BioBert embeddings in concept similarity and patient diagnosis tasks. Concept similarity comparisons to Snomed2Vec were difficult, but in regards to the patient diagnosis task, MedDRA performed slightly worse than SNOMED. We also provide our code base and resulting embeddings for re-creation and further research with the biomedical NLP.

1 Key Information to include

External Mentor: Marie Humbert-Droz

2 Introduction

Research on the text-based representation learning of medical concepts began with the use of skip-gram based models, such as word2vec, on medical text corpora [2] and clinical text [3]. More recent advances use co-occurrence information from electronic health records and doctor visits to learn concept embeddings. Med2Vec, the embeddings created via these methods, are available today almost for 27,000 ICD-9 codes, which are used to classify diseases and other patient symptoms. Other methods, such as the CUI2Vec algorithm, learn embeddings from a combination of medical corpora, electronic health record datasets, and clinical notes. Despite efforts to integrate deep learning methods

into healthcare, the use of embeddings derived from these research efforts has not been widespread. Prior research is often difficult to reproduce and access, since its main data sources–patient records, claims, and medical corpora–are confidential and limited to those with healthcare clearances. In response to these data challenges, Snomed2Vec proposed a new approach for the use of NLP in the biomedical and healthcare domain, using expert-curated knowledge graphs that already exist and are open to the public, such as SNOMED-CT. Snomed2Ved sought to tackle the issues of data accessibility and research reproducibility while also comparing these methods back to research that uses raw medical corpora and patient data [1]. Snomed2Vec tests three different knowledge graph embedding methods, node2vec, metapath2vec, and Poincare embeddings, on four evaluation tasks, namely multi label classification (predicting the type of a node), link prediction (predicting the existence or absence of a relationship between nodes), concept similarity, and patient diagnosis prediction. Concept similarity involves determining whether two nodes have a significant cosine similarity by comparing against a bootstrapped distribution of cosine similarities from node pairs that belong to the same categories as the nodes of interest. Patient diagnosis prediction involves predicting the diagnoses a patient receives during one visit given the diagnoses they received at their previous visits. One of the limitations of Snomed2Vec was its focus on a specific subset of SNOMED-CT, which is one of many biomedical ontologies. Our main goal was to derive embeddings from a different ontology, namely MedDRA, the Medical Dictionary for Regulatory Activities terminology, and to compare their efficiency to embeddings from past research, namely BioBert. In our research, we derive embeddings for the entire MedDRA ontology using two of Snomed2Vec's methods, Poincare and node2vec, and evaluate these embeddings based on concept similarity and patient diagnosis tasks. We find that our MedDRA embeddings achieve, for the four relationships recorded in the MedDRA ontology, statistical power values of $-0.114, -0.062, 0.362$ and 0.158 for concept similarity for the Poincare embeddings, and values of $0.933, 0.712, 0.709$, and 0.524 for Node2Vec embeddings. For the patient diagnosis task, our Poincare and Node2Vec embeddings achieve a highest accuracy of 0.343 and 0.348, respectively. Our research compares MedDRA embeddings to BioBert embeddings, showing that word embeddings derived from ontologies are comparable to other prominent word embeddings in the biomedical field. We also compared MedDRA2Vec results with Snomed2Vec, finding that MedDRA embeddings are almost as good on the patient diagnosis task. As mentioned previously, access and reproducibility have been serious challenges in the biomedical NLP field. Thus, we not only verify the methods and results of Snomed2Vec with a different ontology as part of our research but also give open access to the code base and embeddings derived.

3 Related Work

The biomedical community has spent a lot of time, money, and effort developing ontologies, such as SNOMED-CT and MedDRA. The potential utility of domain ontologies, specifically their ability to model and infer key relationships between key terms and concepts, is both widely known and acknowledged. However, there are a variety of barriers preventing the widespread use of ontologies amongst the biomedical community. In order for domain ontologies to be widely used, they must have a high degree of coverage of terms and relationships, be updated regularly, and have open access to the public. The development and refinement of ontologies is often a manual, expensive, and error-prone process, making it difficult to fulfill all three of these conditions. In response to these challenges, researchers are exploring methodologies in the Natural Language Processing (NLP) field to aid with information extraction and retrieval in the biomedical field. NLP methods have been used for entity extraction [4], entity linking, ontology building, and more thus far [5].

One up-and-coming area of biomedical NLP is the the field of word embeddings, since vector representations can be used to represent semantic meanings and relationships between biomedical terms. A variety of word embeddings have emerged, trained on different text corpora, like Wikipedia, biomedical literature, and clinical data [6]. Early on, researchers believed word embeddings trained on clinical notes and patient data were better than those trained on Wikipedia or other public text corpora. However, access to restricted clinical data for training word embeddings presents challenges of reproducibility and distributability. Thus, most derived biomedical word embeddings have not been widely used past the research stage. In hopes of responding to these data challenges and utilizing medical domain ontologies to their fullest, Snomed2Vec proposed using open-access medical ontologies to derive word embeddings. While Snomed2Vec achieved impressive results–performing 5 to 6 times better on concept similarity than previous embeddings from clinical

text and improving patient diagnosis accuracy [1]–the research posed the question of whether word embeddings derived from other biomedical ontologies would also compare to previous biomedical embeddings trained on medical text and corpora, or if these results were limited to the SNOMED-CT ontology. There have been multiple research projects that derive and evaluate word embeddings using the SNOMED-CT [7, 8], but not for other biomedical ontologies. In response to this gap, we derive and evaluate word embeddings from MedDRA, a smaller ontology, and explore the question of how source ontology impacts the quality of word embeddings.

Another challenging aspect of the biomedical NLP field for word embeddings is the lack of standardized evaluation methods. Unlike other NLP embeddings, there is no ranking or way to consistently compare biomedical word embeddings [6]. Thus, another gap we seek to investigate is why current evaluation methods are so hard to standardize.

4 Approach

4.1 Poincare Embeddings

Poincare embeddings are located in hyperbolic space rather than normal Euclidean space; this space has been found to better represent hierarchical structure [9]. The objective function used is

$$\Sigma_{(u,v)} \log \frac{e^{-d(u,v)}}{\Sigma_{v_1 \in N(u)} e^{-d(u,v_1)}},$$

where $N(u)$ is the set of negative samples for an entity u and $d(u, v)$ is the distance between points u and v in hyperbolic space and is given by

$$cosh^{-1}(1 + 2\frac{||u - v||^2}{(1 - ||u||^2)(1 - ||v||^2)}).$$

Following Snomed2Vec's approach, we used our list of MedDRA relations to create a Poincare model. We referenced the code at `https://gitlab.com/agarwal.khushbu/Snomed2Vec/-/blob/master/src/embedding_learning/poincare.py` and built a similar Poincare model with the exact same parameters but with MedDRA relations. We initialized the model, trained for 50 epochs, then exported the embeddings as a .txt file. Using these methods, we derived 200-dimensional Poincare embeddings for the x unique terms in the MedDRA ontology.

4.2 Node2Vec Embeddings

Node2vec is a random walk method, meaning that training involves taking paths through the graph for which nodes are randomly selected (in these cases, according to a probability distribution) [10]. The objective function for this method takes the form:

$$\Sigma_{v_i, v_j \in V(G)} - log(p_L(v_j|v_i)),$$

where v_i, v_j are graph nodes and $p_L(v_j|v_i)$ is a softmax whose argument is the dot product of the feature representations of a node of interest and a node in its context.

We adapted the node2vec implementation found at `https://github.com/aditya-grover/node2vec` to derive the node2vec embeddings for our MedDRA data. This implementation maximizes efficiency while also preserving important information regarding neighborhoods of nodes in a graph [10]. We first extracted all edge data from the MedDRA hierarchy file, which displays relationships between different levels of the ontology. After reading in all edges into an networkx object, we learned embeddings by optimizing the objective above using stochastic gradient descent. Using these methods, we derived 200-dimensional node2vec embeddings for approximately 108,000 unique terms in the MedDRA ontology.

4.3 Baseline

We initially considered comparing our MedDRA embeddings to the Snomed2Vec embeddings. However, the specificity of the subset of terms the Snomed2Vec focused on and the differences

between the SNOMED-CT and MedDRA ontologies made it difficult to so for the concept similarity task. Thus, MedDRA and Snomed2Vec embeddings are only compared for the patient diagnosis task. Additionally, we opted to use the readily available BioBert embeddings, which are the output of the BioBert model. BERT, initially published by Google, is an extremely versatile and compact model that delivers state-of-the-art results on eleven natural language processing tasks. The DMIS-lab expanded a BERT model with further training on PubMed articles to create 768-dimensional word embeddings adapted to the biomedical domain [11]. We downloaded the BioBert embeddings at `https://pypi.org/project/biobert-embedding/` to use as our baseline for both cosine concept similarity and patient diagnosis tasks.

5 Experiments

5.1 Data

Our external mentor, Marie Humbert-Droz, provided us with the full MedDRA terminology files. There are .asc files corresponding to each level of terms in the MedDRA ontology, namely Lowest Level Terms (llt), Preferred Terms (pt), High Level Terms (ht), High Level Group Terms ($hlgt$), and System Organ Classification (soc), as well as a mdhier.asc file that contains the relationships between terms from different levels. The MedDRA file relationship hierarchy is as follows: $llt \rightarrow pt \rightarrow hlt \rightarrow hlgt \rightarrow soc$. The first level, llt, are synonyms or sub-concepts of pt, which are basic medical concepts. The other levels sequentially encompass each one before it and are related by anatomy, pathology, physiology, etiology, or function. Table 1 shows the number of relationships between each level in the MedDRA hierarchy.

Table 1: Number of Relationships in MedDRA Hierarchy

Relationship Type	Count
LLT to PT	81885
PT to HLT	35406
HLT to HLGT	1756
HLGT to SOC	354

To pre-process the MedDRA data, we adapted the BioSyn python script for extracting all terms from the MedDRA ontology, creating a dictionary mapping all ontology terms to their corresponding codes. The reference code can be found at `https://github.com/dmis-lab/BioSyn/blob/master/preprocess/meddra_preprocess.py`. Next, we visualized the medhier.asc file using a pandas dataframe. We extracted all direct MedDRA relations, creating a list of relationships represented as pairs of strings. We outputted the unique pairs in the list, which can also be considered edge data for a network, as our final pre-processed data.

Computing the evaluation metric of patient diagnoses accuracy requires another dataset called MIMIC-III, which contains information on stays at the critical care units of the Beth Israel Deaconess Medical Center for over forty thousand patients [12]. The dataset, which includes information such as the date and time of each patient's hospital stays and the diagnoses associated with each stay, is structured as twenty-six tables.

5.2 Evaluation Methods

Cosine Concept Similarity
Following Snomed2Vec's approach, for every relationship in MedDRA between two terms from different levels, we calculated the cosine similarity between our embeddings for each term. We reported the statistical power, the 5th percentile cosine similarity value, for each pair of levels. The R code we adapted to implement the concept similarity evaluation method can be found at `https://gitlab.com/agarwal.khushbu/Snomed2Vec/-/blob/master/src/concept_similarity/Benchmarking_script.R`. A higher statistical power was interpreted to mean that many relationships between the two MedDRA levels achieved a high cosine similarity value.

Patient Diagnoses

For easier comparison, we evaluated our embeddings using a patient diagnosis method used by Snomed2Vec. The diagnoses from one patient visit were used to predict the diagnoses for the next visit. An embedding vector was generated for each patient visit: the diagnoses from the visit were recorded using a standard set of medical codes, the ICD-9 (International Classification of Diseases, Ninth Revision). The embeddings of the corresponding MedDRA terms were obtained and added together. Following Snomed2Vec, this input is run through an LSTM layer with 32 units. Dot-product attention is optionally applied. After running through a fully-connected layer, the model outputs a 284-dimension multi-hot vector. Each of the 284 classes represents a subset of ICD-9 diagnosis codes. (The mapping was done according to a standard ontology called Clinical Classifications Software). The value at each location in the vector is between 0 or 1, indicating the model's prediction of whether the patient was given a diagnosis belonging to that category during the next visit.

The evaluation metrics were total accuracy and the accuracies of predicting the top 5 and top 20 most rare and most frequent diagnoses. Total accuracy was calculated by taking at most 10 (out of 284) classes with the highest values in the output vector (fewer classes were used if the diagnoses from the next patient visit belonged to fewer than 10 classes, which was often the case) and counting the number of correct classes that were predicted. Determining the top 5 and top 20 most rare/most frequent diagnoses accuracies involved nearly identical processes. The accuracy was calculated by dividing the number of classes belonging to the desired category that were in the model's top 5 (or 20) predictions by the true number of classes from the next patient visit that are in the desired category.

The model was trained for 20 epochs using sigmoid cross-entropy loss and Adam optimizer. The batch size is 128. The train and test data was taken from MIMIC-III; an 80-20 split was employed. There were 12456 samples.

5.3 Experimental Details

We initialized our Poincare model for extracting embeddings with the same specifications as Snomed2Vec, including creating 200-dimensional embeddings for each term in the ontology and setting the number of negative samples equal to 2. We also trained the Poincare model for 50 epochs, like Snomed2Vec, which took around 20 minutes to run on Google Colab.

To derive the node2vec embeddings, we use the same specifications as used in the reference code, which can be found at `https://github.com/aditya-grover/node2vec/tree/master/src`. We did change the dimension of the word embeddings to 200 to match the dimension of our Poincare embeddings. The model took around 10 minutes to run and return the final word embeddings.

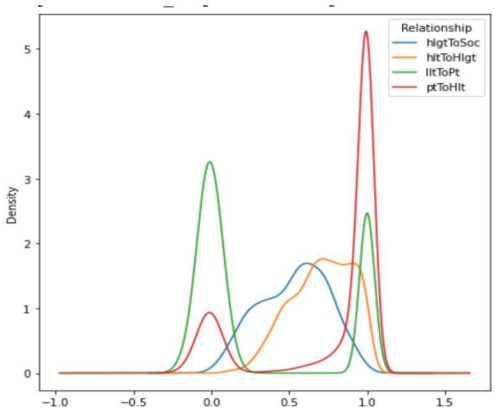
BioBert embeddings for each MedDRA term were obtained by taking each word in the term (examples of terms include "Adult respiratory distress syndrome" and "Whooping cough, unspecified organism"), obtaining its embedding, and averaging them together. Obtaining these embeddings took 3.5 hours.

5.4 Results

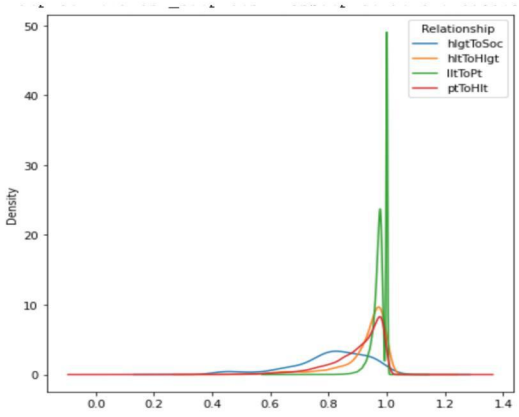
Table 2: Statistical Power Comparisons

Method	LLT to PT	PT to HLT	HLT to HLGT	HLGT to SOC
Node2Vec	0.933	0.712	0.709	0.524
Poincare	-0.114	-0.062	0.362	0.158
BioBERT	1.0	0.808	0.845	0.799
Snomed2Vec Poincare	0.7 (D1)	-	-	-
Snomed2Vec Node2Vec	0.79 (D1)	-	-	-

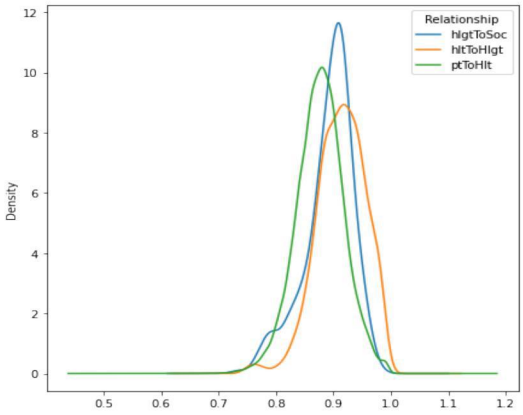
Table 2 shows the concept similarity results achieved by the Node2Vec and Poincare embeddings. The Snomed2Vec results were not directly comparable since they were evaluated on a different ontology, so we used BioBert embeddings as a baseline.



(a) Poincare



(b) Node2Vec



(c) BioBert

Figure 1: Concept Similarity Distributions

Table 3: Patient Diagnosis Results

Method	Attention	All Diagnoses Accuracy	Top 5 Rare	Top 20 Rare	Top 5 Frequent	Top 20 Frequent
Snomed2Vec Poincare	No	0.432	0	0.0121	0.376	0.823
Snomed2Vec Poincare	Yes	0.414	0.0172	0.0414	0.345	0.812
Snomed2Vec Node2Vec	No	0.376	0	0.008	0.304	0.785
Snomed2Vec Node2Vec	Yes	0.374	0	0	0.368	0.863
Poincare	No	0.343	0.029	0.032	0.319	0.903
Poincare	Yes	0.343	0.013	0.031	0.345	0.908
Node2Vec	No	0.342	0.025	0.028	0.331	0.913
Node2Vec	Yes	0.348	0.039	0.051	0.337	0.897
BioBert	No	0.317	0	0	0.319	0.879
BioBert	Yes	0.328	0.025	0.025	0.336	0.866

Figure 1 contains density plots for each embedding method showing the distribution of cosine similarity values for pairs of relationships. There is no curve for the "lltToPt" relationship on the BioBERT plot because every cosine similarity value was 1.0 and a density curve was unable to be drawn.

Table 3 shows the accuracies achieved by the three embedding methods on the patient diagnosis task. We also ran Snomed2Vec's Poincare and Node2Vec embeddings and record their results for comparison.

We also release access to our full code base and embeddings at https://drive.google.com/drive/folders/1VpYOhgzWFwGmR6YMtSJR-7Dpbdy5TuLr?usp=sharing.

5.5 Analysis

Cosine Concept Similarity
In general, the results show that while the concept similarity task is a common ontology evaluation task, it may not be the most relevant to MedDRA. Concept similarity measures how similar the terms in two categories are. MedDRA is a solely hierarchical ontology, with five tiers and one class of terms at each tier.

Similar to the Snomed2Vec paper, Node2Vec achieved high statistical power values while Poincare achieved lower ones. It is difficult to make more detailed comparisons to the Snomed2Vec concept similarity values as the relationships they were evaluating are not the same. Only one of the Snomed2Vec relationships, D1, solely contained hierarchical relationships (MedDRA relationships are also only hierarchical). But this group only contains terms related to disorders while our terms, which come from the whole MedDRA ontology, can be less similar.

In general, the concept similarity performance of the Poincare embeddings was unusual. For two of the relationship types, LLT and PT and PT and HLT, the statistical power values were negative. Looking at the distribution of cosine similarity values in Figure 1 shows why this is the case: the distribution for these two relationships is bimodal, and one of the peaks is centered at 0. The two curves for the other relationships are also unusual; in addition to the peak they plateau once or twice. It is unclear why the LLT and PT relationship has the lowest cosine similarity value of all of the relationships for the Poincare embeddings. It seemingly should have the highest because some LLT terms are synonyms of PT terms whereas for the other relationships, the terms at the lower hierarchy level are sub-concepts of the higher level terms. The unusualness of the Poincare results could be due to its usage of hyperbolic space instead of Euclidean space. Cosine similarity may not be a good way to capture similarity in this space.

The other two embedding methods, Node2Vec and the BioBert baseline, achieve higher statistical power values. For both methods, the value for the LLT-PT relationship is the highest, which is a desired result for the reasons explained in the above paragraph. BioBert achieves higher statistical power values than Node2Vec for every relationship. However, this is likely because terms share many words in common (one HLGT-SOC pair, for example, is "Gastrointestinal signs and symptoms" and "Gastrointestinal disorders"). Since embeddings for terms are calculated by averaging the embeddings of the words that comprise the term, the embeddings for each term in a pair could be very similar. As we see from its performance on the patient diagnosis task, however, successfully capturing similarities does not mean that the BioBert embeddings successfully capture hierarchical relationships or other information that could be helpful for other tasks.

Patient Diagnosis
In contrast to the concept similarity task, Node2Vec and Poincare embeddings seem to perform similarly on the patient diagnosis task. They perform better than the BioBert embeddings on most of the accuracy measures. All of the MedDRA embedding methods have lower total diagnosis accuracy than the Snomed2Vec ones. This is likely a result of the MedDRA ontology's smaller size and lesser focus on clinical terms.

Using attention does not seem to improve task performance. Some of the accuracies improved after attention is applied and some worsened. There seems to be no benefit to specially weighing specific parts of the input vector (an aggregation of the embeddings for each diagnosis). This makes sense; there is no sequential component to the input or output that would seem to clearly benefit from attention.

Additionally, the top 5 and top 20 rare/frequent diagnoses accuracy did not seem to correlate with total accuracy. The SNOMED embeddings achieved the lowest top 20 frequent diagnoses accuracy, for example, but earned the highest total accuracies. Analyzing the top 5 and top 20 diagnoses led to a hypothesis about model performance. The Snomed2Vec methods achieve a higher total accuracy but lower (for the most part) top 5 and top 20 accuracies. Several of these methods achieved 0% accuracy on the top 5 and top 20 rare diagnoses tasks. This led us to hypothesize that the model was optimizing to the task by learning which diagnoses were frequent and then consistently predicting

Table 4: Patient Diagnosis Results Without Using Embeddings

Method	Attention	All Diagnoses Accuracy	Top 5 Rare	Top 20 Rare	Top 5 Frequent	Top 20 Frequent
Poincare	No	0.324	0	0	0.358	0.914
Poincare	Yes	0.319	0	0	0.357	0.967
Node2Vec	No	0.324	0	0	0.343	0.857
Node2Vec	Yes	0.330	0	0	0.349	0.857

those. To test this, we ran the model without using embeddings (the input vector only contained 0s). The results, which are recorded in Table 4, show that the model was able to obtain total accuracies that are only a little lower than the accuracies of the Node2Vec and Poincare methods when embeddings are used. Top 5 and top 20 frequent diagnoses accuracies are higher than before. And as expected, both methods achieved 0% accuracy for top 5 and top 20 rare diagnoses.

This seems to suggest that the embeddings are not substantially contributing to model predictions. One reason could be the small number of ICD-9 terms for which there are corresponding MedDRA terms. For SNOMED, 2,418,977 of the 2,454,405 ICD-9 terms in the MIMIC-III data had corresponding SNOMED terms. Only 1,202,744 terms had corresponding MedDRA ones. This likely contributes to the improved performance of SNOMED embeddings compared to MedDRA ones. The ICD-9 terms that had no associated MedDRA terms could not be used. Thus the set of diagnoses for a patient visit could not be fully represented, and the model may not have received information that might have been important for the prediction.

5.6 Conclusion

MedDRA2Vec expands on the findings of Snomed2Vec, showing that ontologies other than SNOMED-CT can be used to derive medical embeddings and that these embeddings are comparable to models trained on vast amounts of clinical corpora, such as BioBert. On the patient diagnosis task, our MedDRA embeddings perform better than BioBERT ones, but worse than SNOMED-CT ones. We achieve a highest accuracy of $0.348, 0.328$, and 0.432, respectively. The decreased performance compared to SNOMED-CT is likely due to MedDRA having fewer terms pertaining to diagnoses and diseases. Reasons behind this include the smaller size of the MedDRA ontology and its focus on terms relevant to pharmaceuticals testing. On our other evaluation task, concept similarity, we find that the Node2Vec embeddings perform better than the Poincare ones. For the four relationships recorded in the MedDRA ontology, the statistical power values are $0.933, 0.712, 0.709$, and 0.524 for the former and $-0.114, -0.062, 0.362$ and 0.158 for the latter.

In our research, we, like other members of the biomedical community, encountered challenges with evaluating our embeddings and comparing them to previous work done in the field. Especially since previous work in the ontology field mainly derives embeddings from SNOMED-CT, it was challenging to find standardized tasks to use for evaluation across ontologies with different domains and sets of terms. Furthermore, we also found it challenging to use the MedDRA ontology since it is a lot smaller than SNOMED-CT and has a stricter hierarchy of terms. Nonetheless, we release two sets of word embeddings for biomedical terms to the community and share our code base in hopes of making our research both accessible and reproducible.

Our research highlights the difficulties of comparing derived embeddings with past research, so an avenue for future work could be to compile prominent medical terms embeddings in the field, evaluate them using standardized tasks, and report a global ranking. Another interesting avenue could be to derive medical term embeddings from a combination of ontologies. Since ontologies often have a specific domain and can have gaps in knowledge, it would be fascinating to study the effects of multiple ontology sources on word embeddings.

Overall, we learn that the MedDRA ontology can be used as a source for comparable word embeddings in the biomedical field. Like Snomed2Vec, we believe that ontologies are extremely useful for biomedical NLP and encourage further researchers to continue to explore word embeddings in the ontology domain.

References

[1] Khushbu Agarwal, Tome Eftimov, Raghavendra Addanki, Sutanay Choudhury, Suzanne Tamang, and Robert Rallo. Snomed2vec: Random walk and poincaré embeddings of a clinical knowledge base for healthcare analytics. *CoRR*, abs/1907.08650, 2019.

[2] José Antonio Miñarro-Giménez, Oscar Marín-Alonso, and Matthias Samwald. Applying deep learning techniques on medical corpora from the world wide web: a prototypical system and evaluation. *CoRR*, abs/1502.03682, 2015.

[3] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter 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, page 1819–1822, New York, NY, USA, 2014. Association for Computing Machinery.

[4] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, Roger G. Mark, and et al. Mimic-iii, a freely accessible critical care database, May 2016.

[5] Kaihong Liu, William R. Hogan, and Rebecca S. Crowley. Natural language processing methods and systems for biomedical ontology learning. *Journal of Biomedical Informatics*, 44(1):163–179, 2011. Ontologies for Clinical and Translational Research.

[6] Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12–20, 2018.

[7] I. Martinez Soriano, J. L. Castro Peña, J. T. Fernandez Breis, I. San Román, A. Alonso Barriuso, and D. Guevara Baraza. Snomed2vec: Representation of snomed ct terms with word2vec. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 678–683, 2019.

[8] David Chang, Ivana Balazevic, Carl Allen, Daniel Chawla, Cynthia Brandt, and Richard Andrew Taylor. Benchmark and best practices for biomedical knowledge graph embeddings, Jun 2020.

[9] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[10] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks, 2016.

[11] Jitendra Jangid. biobert-embedding: Embeddings from biobert.

[12] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, Roger G. Mark, and et al. Mimic-iii, a freely accessible critical care database, May 2016.