How does the Performance of Embeddings Trained on General Medical Text Vary by Field of Medicine?

John-Jose Nunez
Depts. of Psychiatry and Computer Science, UBC
jjnunez11@gmail.com

1 Introduction

1.1 Background

The application of natural language processing and machine learning to medicine presents an exciting opportunity for tasks requiring prediction and classification, such as predicting the risk of suicide after a patient is discharged from hospital (McCoy et al., 2016). A common approach is to convert the unstructured text produced by clinical interactions into low-dimension vector representations which can fed into these algorithms. These vectorizations are produced by training models on large unlabelled corpora. For example, the popular *word2vec* system (Mikolov et al., 2013) initially trained embeddings using a skip-gram model, training a vector for a target word based on what words are found within a window near it. It was initially trained on a Google News corpus containing around six billion tokens. Due to considerable differences between the language of medical text and general English writing, prior work has trained medical embeddings using specific medical sources.

Recent approaches in this vein include De Vine et al (2014) which trained embeddings for medical concepts in the Unified Library Management System (ULMS) (Bodenreider, 2004) using journal abstracts from MEDLINE as well as with clinical patient records. They then used these embeddings to compare predicted word similarity against human-judgements. Minarro-Gimenez et al (2014) trained embeddings using medical manuals, articles, and Wikipedia articles, comparing predicted vector similarity between medications against the National Drug File - Reference Terminology (NDF-RT) ontology. Choi et al (Choi et al., 2016) improved on this work by learning on large-scale health record data consisting of raw text from clinical notes mapped to concepts from UMLS. In their yet unpublished work, Beam et al (Beam et al., 2018) use an “extremely large” database of clinical notes, insurance claims, and full journal texts, and develop a new system termed “cui2vec”, mapping concepts into a set of unique identifiers based on UMLS, and then training vectors for these identifiers based on the occurrences of other identifiers within a certain window length.

All of the above examples were both trained and evaluated on general medical data, from all fields of medicine. It is unclear how these models perform in specific fields of medicine. For example, we can consider the medical speciality of psychiatry, the field of medicine concerned with mental illness such as depression or schizophrenia. Prior work has shown that psychiatric symptoms are often described in a long, varied, and subjective manner (Forbush et al., 2013 3 18) which may present a particular challenge for NLP.

Prior work has explored whether domain adaptation (DA), techniques to adapt data from other domains to work on a target, can improve performance when applied to this sub-domain of psychiatry. Lee et al (Lee et al., 2018) used these techniques to improve the task of de-identifying psychiatric notes. Zhang et al (Zhang et al., 2018) then applied DA to word embeddings trained from general language and medical sources, showing some improvements when targeting a psychiatric dataset.

1.2 !Contribution

In this work, I seek to start understanding how NLP performance may vary when applied to the difference fields of medicine. Specifically, I compare the quality of embeddings trained on general medical data by the field of medicine they are related to, using a variety of metrics previously described in the literature. As NLP is applied to medicine, field-specific applications will become increasingly popular. If this work finds little difference between the fields, future will be assured that embeddings trained from general medical text will be sufficient. Conversely, if differences are found for specific fields, future work may want to address this shortfall by using techniques like DA, or even creating embeddings specifically trained for this field.

2 Proposed Methodology

In order to determine which psychiatric and non-psychiatric terms should be compared, the most common concepts shall be used. For instance, we will compare the most commonly prescribed psychiatric and non-psychiatric drugs, or the most common diagnoses, based on prior epidemiology, in order to compare common, well described concepts.

For top diagnoses: Could access fancy Canadian Data with a data request per emails from librarians. Or, can use top diagonses cards from ACP from ICD 10 here or ICD 9 version which seems to skip mental disorders. Or could just do a "top 10" and show quality for all of those.

3 Methods

3.1 Obtaining Embeddings

Table 1 contains details of the embeddings compared in this project, all of which are based on *word2vec*. We obtained DeVine200 (De Vine et al., 2014), ChoiClaims300, and ChoiClinical300 (Choi et al., 2016) all from the later's website. We downloaded Beam-Cui2Vec (Beam et al., 2018) from this site.

3.2 Evaluation Methods

Determining a Concept Embedding's Field of Medicine A clinical concept could be understood to be a part of different medical fields depending on who is asked. In order to have an objective and unambiguous classification, we utilized the ninth revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-9). This is a widely used system to classify medical diseases and disorders, and classifies such into seventeen chapters representing medical systems or other such clusters, such as mental disorders, or disease of the respiratory system.

Medical Relatedness Property (MRP)

This metric from (Choi et al., 2016) is based on quantifying whether concepts with known relations are located near each other. They use, separately, known relationships between drugs and the diseases they treat or prevent, and the relations between diseases that are group together in the CCS hierarchical groupings, a classification from the Agency for Healthcare Research and Quality. The scoring utilizes Discounted Cumulative Gain, which attributes a diminishing score if a relationship is found within *k* neighbours the further it is.

In our implantation, we calculate the MRP based on the 'course' groupings from CCS drug hierarchies. Scores are calculated for CUIs that represent diseases with a known ICD9 code. The mean MRP is then calculated for all CUIs within a given ICD9 system.

DISCUSS THIS WAS IMPLEMNTED AND OTHERS FROM SCRATCH.

Medical Conceptual Similarity Property (MCSP) The other metric from Choi et al's work evaluates whether embeddings known to be of a particular set are clustered together. They use conceptual sets from the UMLS such as 'pharmacologic substance' or 'disease or syndrome'. Discounted cumulative gain is again used, based on whether a cui has other cuis of its set within *k* neighbours.

We reimplement this method, but instead of using the UMLS conceptual sets, we create sets from the ICD9 systems. We consider CUIs that represent diseases with ICD9 codes, and CUIs that represent drugs which treat or prevent ICD9 diseases, and attribute score if drugs or diseases are near others of the same medical system.

Correlation with UMNSRS Similarity (SimCor) (Yu et al., 2017) investigate whether the cosine similarity of embedding vectors are correlated with human judgements of similarity. The UMNSRS-Similarity dataset (Pakhomov, 2018) contains around 500 similarity ratings between medical concepts as rated by eight medical residents. A Spearman rank correlation is then computer between the cosine similarities and the UMNSRS-Similarity ratings for pairs.

Our implementation repeats the above. A medical system's mean correlation is calculated from all pairs that contain at least one disease with an ICD9 code in its system, or a drug that treats or prevents a disease in the system. Soearman rank correlation is again used.

Significance in Bootstrap Distribution (Bootstrap) Beam et al (2018) also evaluate how well known relationships between concepts are observed between embeddings, such as whether diseases are co-morbid, or a drug treats a condition. For a given type of relation, they generate a bootstrap distribution by randomly calculating cosine similarities of embedding vectors of the same class (eg. a random drug and disease when evaluating drug may-treat disease relations). For a given known relation, they consider the embeddings producing an accurate prediction if their cosine similarity is within the top 5%, the equivalent of p¡0.05 for a one-sided t-test.

Our implementation considers the *may-treat* or *may-prevent* known relationships from the NDF-RT dataset. The percentage of known relations for drug-disease pair within each medical system is calculated.

Centroid Prediction We implement a new method to compare against the others. A centroid is calculated for each medical system by averaging the normed embeddings for diseases with relevant ICD9 codes. We then calculate the percentage of drugs known to treat or prevent a disease in each system whose vectors are most similar

Table 1: Charectoristics of the embeddings compared, including the name referred, the embedding dimensions, the number of embeddings in the dataset, and the type of data used to train them.

Name	Dimension	Number	Data Used to Train
DeVine200	200	52,102	clinical narratives
			journal abstracts
ChoiClaims300	300	14,852	health insurance claims
ChoiClinical300	300	22,705	clinical narratives
BeamCui2Vec500	500	109,053	health insurance claims
			full journal text

(by cosine similarity) to the relevant centroid. For example, we would expect the CUI for ‘fluoxetine’, an antidepressant, to be most similar to the Mental Disorders centroid.

Some drugs treat or prevent diseases in *n* multiple system. For a medical system, such a drug is considered being accurately predicted if it the system’s centroid is amongst the *n* most similar centroids.

Analysis Our work seeks to determine if embeddings for one medical system are worse or better than others. Additionally, we week to determine whether there are similar differences between the different sets of embeddings. To do this, we must consider the scores generated using five different metrics, four different embeddings, and seventeen different medical systems. However, the scores from five metrics are not obviously convertible, at least not all together.

For now, we will assume our results are normally distributed; this may not be unreasonable as the processes underlying the quality of embeddings - the use of words representing clinical concepts in texts - stem from a natural process (human writing word choice).

To compare the embeddings from the different medical systems, for each evaluation method we calculate the mean score from each embedding. We then conducted a paired two-way t-test for five pairs of system’s score with a given embedding vs the mean score across all systems for an embedding. We then observe the relative difference vs the mean, and whether this is significant at $p < 0.05$.

We repeat the same steps to compare the embeddings themselves.

4 Results

Comparing the embeddings from medical systems across the five evaluation methods reveals that some systems have scores significantly above the mean across multiple methods (Table 2). Embeddings related to cancers and muskoskeletal system are significantly above the mean in two methods, while those of mental disorders, the nervous system, and the cardiovascular system are significantly above in three methods. No systems have scores below the mean on more than one metric; those that do include those related to diseases of pregnancy, the perinatal period, and skin disorders.

Evaluating the sets of embeddings (Table 2) shows some stark differences. The *cui2vec* embeddings from Beam et al are above the mean across all evaluation methods, significantly three times. Those from DeVine et al, and those from Choi et al based on the clinical narratives, do significantly worse. The remaining embeddings, those based on health insurance codes from Choi et al, are more middling.

5 !Discussion

5.1 !Lessons Learned

5.2 !Was Project Succesful?

5.3 !Strengths and weaknesses

6 !Future Work

Multiple avenues exist for improving this work. Currently, the dictionary between ULMS CUIs and ICD9CM codes only has around 40,000 entries. As such, many CUIs could not be used, despite representing concepts very related. Generating a larger dictionary automatically would be useful, and could possible be feasible given the requirement that they only be fit into large disease categories.

In this work, all known CUIs repersenting an ICD9 code are used for the calculations. An idea initially proposed was to repeat these analysis only using a list of the most frequent diseases. This was not done due to the difficulty of quantifying such lists, as it would require access to sensitive health systems data which takes time to access. REWORD.

Evaluating Zhang et al’s domain adaptation-trained embeddlings in this project would help quantify how much improvement this technique may lead, if their embeddings could be obtained.

Table 2: Score of embeddings from a given medical system according to a given evaluation method expressed as relative to the mean score for that method. Significant (paired t-test p <0.05) scores above are shown in orange, below blue. See Methods section for method abbreviations. Blank values represent no scores could be calculated for a given combo.

	MRP	MCSP	SimCor	Bootstrap	Centroid
Infections	-0.206	0.408	-0.341	0.348	0.119
Cancers	0.194	0.019	0.514	0.382	-0.04
Cndocrine	0.029	-0.116	0.199	0.229	-0.1
Blood diseases	-0.071	-0.285	-0.02	-0.063	0.097
Mental disorders	0.23	0.666	-0.449	0.25	0.325
Nervous system	0.61	0.638	-0.028	0.174	0.406
Cardiovascular	0.117	0.641	0.189	0.218	0.384
Respiratory	-0.13	0.066	-0.242	0.402	-0.07
Digestive	0.239	-0.072	-0.421	-0.067	-0.385
Genitourinary	0.285	0.08	-0.149	0.073	0.178
Pregnancy	-0.219	-0.56			
Skin	-0.276	-0.203	-0.312	0.18	0.075
Muskoskeletal	0.046	0.13	0.467	0.064	0.305
Congenital	-0.15	-0.128		0.106	0.524
Perinatal	-0.382	-0.567			
Ill-defined	-0.176	-0.281	-0.244	-0.183	0.053
Injury and poisoning	-0.138	-0.437	0.835	-0.115	0.128

Table 3: Score of set of embeddings according to a given evaluation method expressed as relative to the mean score for that method. Significant (paired t-test p <0.05) scores above are shown in orange, below blue. See Methods section for embedding set abbreviations.

	MRP	MCSP	SimCor	Bootsrap	Centroid
DeVine200	-0.345	-0.121	0.065	-0.464	-0.042
ChoiClaims300	0.097	0.036	-0.055	-0.041	0.018
ChoiClinical300	-0.135	-0.155	-0.047	-0.019	-0.005
BeamCui2Vec500	0.384	0.24	0.037	0.524	0.029

References

[Beam et al.2018] Andrew L. Beam, Benjamin Kompa, Inbar Fried, Nathan P. Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2018. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. *arXiv:1804.01486 [cs, stat]*, April.

[Bodenreider2004] Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, January.

[Choi et al.2016] Youngduck Choi, Chill Yi-I Chiu, and David Sontag. 2016. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41–50, July.

[De Vine et al.2014] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. 2014. 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, pages 1819–1822, New York, NY, USA. ACM.

[Forbush et al.2013 3 18] Tyler B. Forbush, Adi V. Gundlapalli, Miland N. Palmer, Shuying Shen, Brett R. South, Guy Divita, Marjorie Carter, Andrew Redd, Jorie M. Butler, and Matthew Samore. 2013 -3- 18. "Sitting on Pins and Needles": Characterization of Symptom Descriptions in Clinical Notes". *AMIA Summits on Translational Science Proceedings*, 2013:67–71.

[Lee et al.2018] Hee-Jin Lee, Yaoyun Zhang, Kirk Roberts, and Hua Xu. 2018. Leveraging existing corpora for de-identification of psychiatric notes using domain adaptation. *AMIA Annual Symposium Proceedings*, 2017:1070–1079, April.

[McCoy et al.2016] Thomas H. McCoy, Victor M. Castro, Ashlee M. Roberson, Leslie A. Snapper, and Roy H. Perlis. 2016. Improving Prediction of Suicide and Accidental Death After Discharge From General Hospitals With Natural Language Processing. *JAMA psychiatry*, 73(10):1064–1071, October.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, January.

[Minarro-Giménez et al.2014] José Antonio Minarro-Giménez, Oscar Marín-Alonso, and Matthias Samwald. 2014. Exploring the application of deep learning techniques on medical text corpora. *Studies In Health Technology And Informatics*, 205:584–588.

[Pakhomov2018] Serguei Pakhomov. 2018. Semantic Relatedness and Similarity Reference Standards for Medical Terms, May.

[Yu et al.2017] Zhiguo Yu, Byron C. Wallace, Todd Johnson, and Trevor Cohen. 2017. Retrofitting Concept Vector Representations of Medical Concepts to Improve Estimates of Semantic Similarity and Relatedness. *arXiv:1709.07357 [cs]*, September.

[Zhang et al.2018] Yaoyun Zhang, Hee-Jin Li, Jingqi Wang, Trevor Cohen, Kirk Roberts, and Hua Xu. 2018. Adapting Word Embeddings from Multiple Domains to Symptom Recognition from Psychiatric Notes. *AMIA Summits on Translational Science Proceedings*, 2017:281–289, May.