Comparing the Intrinsic Performance of Clinical Concept Embeddings by their Field of Medicine

Anonymous EMNLP-IJCNLP submission

Abstract

Pre-trained word embeddings are becoming increasingly popular for natural language-processing tasks. This includes medical applications, where embeddings are trained for clinical concepts using specific medical data. Recent work continues to improve on these embeddings. However, no one has yet sought to determine whether these embeddings work as well for one field of medicine as they do in others. In this work, we use intrinsic methods to evaluate embeddings from the various fields of medicine as defined by their ICD-9 systems. We find significant differences between fields, and motivate future work to investigate whether extrinsic tasks will follow a similar pattern.

1 Introduction

The application of natural language processing (NLP) 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 method across NLP for such tasks is to use high-dimensional vector word representations. These word embedding include the popular *word2vec* system (Mikolov et al., 2013) which was initially trained on general English text, using a skip-gram model on a Google News corpus.

Due to considerable differences between the language of medical text and general English writing, prior work has trained medical embeddings using specific medical sources. Generally, these approaches have trained embeddings to represent medical concepts according to their 'clinical unique identifiers' (CUIs) in the Unified Library Management System (ULMS) (Bodenreider, 2004). Words in a text can then be mapped to these CUIs (Yu and Cai, 2013). Various sources have been used, such as medical journal articles, clinical patients records, and insurance claims (De Vine et al., 2014), (Minarro-Giménez et al., 2014), (Choi et al., 2016).

Prior authors have sought to improve the quality of these embedding, such as using different training techniques or more training data (Beam et al., 2018). In order to judge the quality of these embeddings, they have primarily used evaluation methods quantifying intrinsic qualities, such as their ability to predict drug-disease relations noted in the National Drug File - Reference Terminology (NDF-RT) ontology (Minarro-Giménez et al., 2014), or whether similar types of clinical concepts had cosine similiar vectors (Choi et al., 2016).

To date, these embeddings have been both trained and evaluated on general medical data, from all fields of medicine. It is unclear how well such embeddings perform for a specific field of medicine. For example, we can consider psychiatry, the field of medicine concerned with mental illnesses 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) which may present a particular challenge for training these embeddings and NLP tasks generally.

As these pre-trained embeddings may be increasingly be used for down-stream NLP tasks in specific fields of medicine, we seek to determine whether embeddings from one field perform relatively well or poorly relative to others. Specifically, we aim to follow prior work using intrinsic evaluation methods, comparing the geometric properties of embeddings vectors against others

given known relationships. This will offer a foundation for future work that may compare the performance of extrinsic NLP tasks in different medical fields. Finding relative differences may support that certain medical fields would benefit from embeddings trained on data specific to their field, or using domain adaptation techniques as sometimes used in the past (Yu et al., 2017).

2 Methods

2.1 Sets of Embeddings

We sought to compare a variety of clinical concept embeddings trained on medical data. Table 1 contains details of the sets 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 Github. We downloaded BeamCui2Vec500 (Beam et al., 2018) from this site. We were unable to obtain other sets of embeddings mentioned in the literature (Minarro-Giménez et al., 2014), (Zhang et al., 2018) (Xiang et al., 2019).

2.2 Determining a Field of Medicine's Clinical Concepts

A clinical concept's corresponding field of medicine is not necessarily obvious. 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) (Slee, 1978). This is a widely used system of classifying medical diseases and disorders, dividing them into seventeen chapters representing medical systems/categories such as mental disorders, or disease of the respiratory system. While the 10th version is available, we chose this version based on prior work using it, and the pending release of the 11th version. We will use these ICD9 systems to define the different medical fields.

We determined a CUI's field of medicine according to a CUI-to-ICD9 dictionary available from the UMLS (Bodenreider, 2004). We consider pharmacological substance related to a field of medicine system if it treats or prevents a disease with an ICD9 code within a particular ICD9 system. We determine this by using the NDF-RT dictionary, which maps CUIs of substances to the CUIs of conditions they treat or prevent, and then convert these CUIs to the ICD9 systems as before. As such, A CUI repersenting a drug may have multiple ICD9 systems and therefore medical fields.

2.3 Evaluation Methods

We sought to compare multiple methods for evaluating the quality of a medical field's embeddings based on prior work. We were unable to use Yu et al's (2017) method, based on comparing the correlation of vector cosine similarity against human judgements from the UMNSRS-Similarity dataset (Pakhomov, 2018) due to there being too few examples across many medical fields.

Medical Relatedness Measure (MRM)

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

In our implantation, we calculate the MRM based on the 'course' groupings from CCS drug hierarchies. Scores are calculated for CUIs that represent diseases with a known ICD9 code. The mean MRM is then calculated for all CUIs within a given ICD9 system. The implementation was adapted from Python 2.7 code available from the original author's Github.

Medical Conceptual Similarity Measure (MCSM) The other method used by 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, again giving a score to neighbours that are diseases or drugs from the same field of medicine/ICD9 system. Again, this was adapted from code from Choi et al's Github.

Name	Dimension	Number	Number of Training Data	Type of Training Data
DeVine200	200	52,102	17k + 348k	clinical narratives, journal abstracts
ChoiClaims300	300	14,852	4m	health insurance claims
ChoiClinical300	300	22,705	20m	clinical narratives
BeamCui2Vec500	500	109,053	60m + 20m + 1.7m	claims, narratives, full journal texts

Table 1: Characteristics 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.

Significance against Bootstrap Distribution (Bootstrap) Beam et al (2018) also evaluate how well known relationships between concepts are represented by embedding vector similarity. For a given known relation, they generate a bootstrap distribution by randomly calculating cosine similarities between embedding vectors of the same class (eg. a random drug and disease when evaluating drug-disease relations). For a given known relation, they consider that the embeddings produced 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. We calculate the percentage of known relations for drug-disease pair within each medical field. Beam et al have not yet made their code publicly available, so we reimplemented this technique in Python.

System Vector Accuracy (SysVec) We implement a new, simple method to evaluate a medical field's embeddings. A representative vector is calculated for each medical field/ICD9 system by calculating the mean of normalized embeddings of a field's diseases. We then calculate the percentage of drugs known to treat or prevent a disease in each system whose vectors are most similar (by cosine similarity) to the relevant system vector. For example, we would expect the CUI for 'fluoxetine', an anti-depressant, to be most similar to the Mental Disorders system vector.

Some drugs treat or prevent diseases in *n* multiple systems. For a given medical field, such a drug is considered being accurately predicted if the system's centroid is amongst the *n* most similar centroids. We implemented this in Python.

2.4 Comparing Scores

Comparing Sets of Embeddings : We calculated the mean scores for an embedding set, only including embeddings with corresponding ICD9 values and present in all of the compared sets. For the MCSM and MRM scores, we conducted two-way paired t-tests between the scores from each embedding set, adjusted with the Boneferroni correction. For the binary Bootstrap and SysVec scores, we judged statistical significance by calculating z-scores and their corresponding Boneferroni corrected p-values.

A negative control set of embeddings was constructed by taking the embeddings from Beam et al (2018) and randomly arranging which clinical concepts an embedding corresponds to.

Comparing Fields of Medicine : As the embeddings from Beam et al (2018) are most recent, trained on the most data, and have substantially higher mean scores than the other embeddings compared, we used these embeddings to compare scores from the different fields of medicine. This set also contained the most embeddings, allowing more embeddings from each field to be compared.

We sought to determine whether a field of medicine's embeddings were significantly worse or better than the average. As such, for each field of medicine we calculated the mean score from each evaluation method. We then used statistical tests to compare a field's scores from a given evaluation method with the same scores from all other fields. For MCSM and MRM scores we used two-tailed t-tests, and for Bootstrap and Sysvec, z-scores, all corrected with the Bonferroni correction.

To aggregate a medical field's results, we calculated a 'net score' by taking how many of the four method's scores were significantly above the mean, minus how many were significantly below. We found this more interpretable than other aggregate methods such as combining normalized scores.

3 Results

3.1 Differences Between Sets of Embeddings

Comparing the sets of embeddings (Table 3) shows consistent differences. BeamCui2Vec500's scores are the highest across all methods, and this difference is very significant, with p-value $\ll 10^{-}5$ after Bonferonni correction. The ChoiClaims300 embeddings seem next best, and the remaining sets still have much higher scores than those of the negative control.

3.2 Differences Between Medical Systems

Differences are also observed between embeddings from the various fields of medicine as represented by the ICD-9 systems 3. For instance, embeddings related to Infections and Parasitic Disease have scores significantly above the mean for all four evaluation methods, while those of Symptoms, Signs, and Ill-defined Conditions are significantly below for all three. Due to a smaller number of documented drug-disease relationships across two medical fields, scores were not calculated with those methods based on such.

4 Discussion and Future Direction

To our knowledge, this is the first investigation into whether clinical concept embeddings from a given field of medicine relatively good or bad compared to others. We conducted this investigation comparing available sets of such embeddings, using a variety of previously described intrinsic evaluation methods in addition to a new one. Given that one set of embeddings performed better than others, we used this set to compare the different fields of medicine, and found significant results between various fields.

The superior performance of one set of embeddings - those from Beam et al (2018) - are consistent with the depth and breadth of data used to train these embeddings. Training used three different types of data, including that from health insruance claims, clinical narratives, and full texts from medical journals. The number of data was also much larger the the other sets. Our work validates their findings that their embeddings offer the best performance. However, it would be interesting to also consider the recent clinical concept embeddings developed by (Xiang et al., 2019). They use a similar number of data (50 million) as Beam et al, using a large dataset from an electronic health records, and use a novel method to incorporate time-sensitive information. At the time of submission, we were unable to obtain their embeddings, and so leave this comparison to future work.

Examining the differences between fields of medicine, we note that the poor performance of embeddings from the system "Symptoms, Signs, and Ill-defined Conditions" may support validity of the results. This collection of miscellaneous medical conditions would not be expected to have the intrinsic vector similarity and cohesion evaluated by our evaluation methods.

Further work may explore why the other systems have varied performance. We wonder if the observed results correlate with possible distinctiveness of the various medical fields. For example, the best performing system was "Infectious and Parasitic Diseases". The conditions in this field are often unambiguous - a pathogen like *influenza virus* has little other meaning - and the drugs used for these diseases tend to be similarly specific. On the other hand, poorly performing systems such as "Diseases of the Skin and Subcutaneous Tissue" and "Diseases of the Musculoskeletal Systems and Connective Tissue" often utilize immunosuppressant medications that are used across many fields of medicine. Future work could investigate this conjecture by comparing scores when restricting what clinical concepts are compared, such as only common or distinct medications.

This work evaluated embeddings using intrinsic measures of embedding quality. This presents some advantages, but also the most obvious limitations and direction for future work. These intrinsic methods allowed a consistent evaluation to be carried out between medical fields, and allowed a wide variety of embeddings from a given field to be compared. The methods all evaluate qualities that well-trained embeddings should have, though still represent artificial use-cases. Evaluating these embeddings on extrinsic, down-stream tasks may provide more applicable comparisons. However, these tasks will need to be comparable and available for multiple medical fields. For instance, the recent work by Xiang et al (2019) compared embeddings trained by different methodologies on a task predicting the onset of heart failure (Rasmy et al., 2018). This would be an appropriate task to judge embeddings from "Diseases of the Circulatory System"; others would be needed for other

Embedding Set	MRM	MCSM	Bootsrap	SysVec
Negative Control	0.02	1.24	0.05	0.35
DeVine200	0.24	5.14	0.27	0.79
ChoiClaims300	0.43	5.34	0.42	0.80
ChoiClinical300	0.33	4.49	0.42	0.74
BeamCui2Vec500	0.52	6.39	0.67	0.90

Table 2: Mean scores for embedding sets for each evaluation method. See Methods section for abbreviations

ICD-9 Systen	MRM	MCSM	Bootstrap	SysVec	Net Score
All Systems (Negative Control)	0	1.08	0.04	0.25	-
All Systems	0.55	8.07	0.89	0.63	-
Infectious and Parasitic Diseases	**0.45**	**7.72**	**0.93**	**0.92**	+4
Neoplasms	**0.62**	**9**	0.94	0.55	+2
Endocrine, Nutritional and Metabolic Diseases, and Immunity Disorders	**0.44**	**5.64**	0.89	0.53	-2
Diseases of the Blood and Blood-forming Organs	**0.31**	**4.36**	0.82	0.79	-2
Mental Disorders	0.53	**9.34**	0.96	**0.83**	+2
Diseases of the Nervous System and Sense Organs	**0.76**	**8.44**	0.87	**0.33**	+1
Diseases of the Circulatory System	0.59	8.12	**0.96**	**0.72**	+2
Diseases of the Respiratory System	**0.36**	**5.85**	0.94	**0.82**	+1
Diseases of the Digestive System	**0.61**	7.93	**0.77**	0.62	0
Diseases of the Genitourinary System	**0.61**	**6.82**	0.86	0.58	0
Complications of Pregnancy, Childbirth, and the Puerperium	**0.51**	**10.27**	-	-	0
Diseases of the Skin and Subcutanous Tissue	**0.37**	**5.1**	0.81	0.58	-2
Diseases of the Musculoskeletal System and Connective Tissue	**0.47**	8.22	0.88	**0.29**	-2
Congenital Anomalies	0.5	**6.24**	0.73	0.73	-1
Certain Conditions Originating in the Perinatal Period	**0.48**	**9.84**	-	-	0
Symptoms, Signs, and Ill-defined Conditions	**0.26**	**2.68**	**0.77**	0.56	-3
Injury and Poisoning	**0.59**	**9.09**	**0.75**	**0**	0

Table 3: Comparison of mean scores using different evaluation methods for the fields of medicine as represented by their ICD-9 system. Significant differences below or above the mean of all other systems are bold (p-value <0.05 after Boneferroni correction). Net score is the number of these significant differences above that mean minus the number below. A system's score is not calculated if there are fewer than ten examples for a method. See Methods section for evaluation method abbreviations

systems. We also plan to investigate the validity of these intrinsic evaluation methods by comparing them to extrinsic results.

Another future direction could be to investigate what could be done to improve performance in the fields with lower scores. For instance, Zhang et al (2018) used domain adaptation techniques for psychiatric embeddings, and this could instead be carried out for those systems we identified as doing poorly. Alternatively, one could also train embeddings solely on data from one field of medicine and investigate how this affects performance.

References

Clinical Classifications Software (CCS), 2003. page 54.

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]*.

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

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.

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.

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. "Sitting on Pins and Needles": Characterization of Symptom Descriptions in Clinical Notes". *AMIA Summits on Translational Science Proceedings*, 2013:67–71.

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.

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

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.

Serguei Pakhomov. 2018. Semantic Relatedness and Similarity Reference Standards for Medical Terms.

Laila Rasmy, Yonghui Wu, Ningtao Wang, Xin Geng, W. Jim Zheng, Fei Wang, Hulin Wu, Hua Xu, and Degui Zhi. 2018. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *Journal of Biomedical Informatics*, 84:11–16.

Vergil N. Slee. 1978. The International Classification of Diseases: Ninth Revision (ICD-9). *Annals of Internal Medicine*, 88(3):424.

Yang Xiang, Jun Xu, Yuqi Si, Zhiheng Li, Laila Rasmy, Yujia Zhou, Firat Tiryaki, Fang Li, Yaoyun Zhang, Yonghui Wu, Xiaoqian Jiang, Wenjin Jim Zheng, Degui Zhi, Cui Tao, and Hua Xu. 2019. Time-sensitive clinical concept embeddings learned from large electronic health records. *BMC Medical Informatics and Decision Making*, 19(2):58.

Sheng Yu and Tianxi Cai. 2013. A Short Introduction to NILE. *arXiv:1311.6063 [cs]*.

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]*.

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.