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
jjnunez@cs.ubc.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 then be 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, producing a target word's vector based on what words surround 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.

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), which words in text can then be mapped to (Yu and Cai, 2013). Examples include De Vine et al (2014) who trained embeddings using journal abstracts from MEDLINE as well as with clinical patient records. They evaluated these embeddings by comparing vector cosine similarity against human-judged similarity. Minarro-Gimenez et al (2014) trained embeddings using medical manuals, articles, and Wikipedia articles, judging quality by their ability to predict drug-disease relations noted in the National Drug File - Reference Terminology (NDF-RT) ontology. Choi et al (Choi et al., 2016) improved on this work by learning two sets of embeddings, from health insurance claims and clinical narratives. They evaluated their embeddings by determining their ability to predict known relations including those in the NDF-RT, disease hierarchies, and medical concept type. In their yet unpublished work, Beam et al (Beam et al., 2018) learn embeddings on the largest dataset, combining health insurance claims, clinical narratives and full journal texts. They developed a new system termed "cui2vec", training CUI embeddings based on the occurrences of other identifiers within a certain window length. They use an assortment of aforementioned known relations to compare the quality of these embeddings.

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 training these embeddings and NLP tasks generally.

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, we seek to start understanding how NLP performance may vary when applied to the difference fields of medicine. Specifically, we 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.

Additionally, this project also contributes to the evaluation of medical embeddings. These methods generally evaluate whether the embedding vectors for a given concept are similar to other embeddings given a known relation. This work is the first to use concepts' related fields of medicine as such a relation, which we believe may be more clinically relevant than some prior relations.

Lastly, by comparing multiple sets of embeddings using different evaluation metrics, this work seeks to also evaluate the relative performance of embeddings trained with different methods.

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 BeamCui2Vec (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

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	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

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 (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-tailed 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 sets of embeddings. This time, we calculate the mean scores from each medical system for each evaluation method, and conduct the paired two-tailed t-tests on seventeen pairs for each medical system. The pairs are the mean score from all the embedding sets, against the score from the given embedding set. Again, we calculate the score relative to mean, and significance as above.

After observing that the embeddings from Beam et al well outperform the other embedding sets, we then evaluate the embeddings by medical system one last time using only MCSP. This is chosen as it resulted in the highest number of significant differences, and is able to calculate scores for all systems. Choosing one embedding set allows more embeddings to be compared, as scores can only be calculated for each method on embeddings shared by all embedding sets. Also choosing one evaluation method then allows statistical analysis to be performed on the individual embedding scores. As such, for each medical system we then perform a one-sided two-tailed t-test between its embeddings and the scores from all embeddings. We then repeated the same but only for the embeddings representing CUIs that were overlapping between the four embedding sets so that the effect of this restriction could be determined.

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

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

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.

In Table 4 we focus on MCSP scores from the best performing set of embeddings, those from Beam et al. We observe results that are somewhat similiar to the joint analysis. Embeddings related to mental disorders, the nervous system, and the cardiovascular system again perform when all relevant embeddings are observed, in line with the general results. When only the overlapping embeddings are considered, muskoskeletal embeddings are below average, while genitourinary are above, though the other three are again better. Of note, these results do become weight on the number of examples found per system, unlike prior results. While these results descriptively seem to match, the mean MCSPs are not well correlated between those using all embeddings vs only the overlapping ones. Assuming normality, Pearson's coefficent is only 0.16, while not assuming this and using Spearman's we get only 0.01.

5 Discussion

5.1 Embedding Quality by Field of Medicine

In this project, we seek to determine whether embeddings for clinical concepts learned from general medical text well well for the various fields of medicine. We investigate this by using different metrics of embedding quality and different sets of embeddings, and use ICD9 systems to determine the relevant medical fields. Based on the methods used so far, our results suggest a consistent pattern the embeddings from certain medical systems perform better than others- namely, those of mental disorders, the nervous system, and those of the cardiovascular system, and possibly the muskoskeletal system. It is less clear if any systems perform particularly poorly, though some of the systems had few or no relevant embeddings to for some metrics.

Why some of these systems seem to perform better are unclear. If we take the numbers of compared embeddings as a proxy for frequency, the well performing systems do not stand out as either popular nor unpopular. The frequency of concepts found in the training data may be a cause, and presents an option to examine quantitatively in the future.

Strengths of the project include an objective means to relate a clinical concept to a field of medicine, and the combination of multiple evaluation metrics and sets of embeddings to well survey relevant work. However, these also present weakness. First and foremost, the combination of metrics and embeddings and systems creates a daunting task for statistical analysis, and this could likely be improved by considering advanced techniques such as non-parametric equivalents to ANOVA and ensuring post-hoc analysis.

Additionally, we see that there is a difference in the

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

Table 4: Comparison of MCSP scores using the embeddings from Beam et al when considering all relevant embeddings, and only those that are overlapping with the other sets of embeddings. The MCSP column repersents the mean MCSP score, Examples the number of embeddings per medical system, Relative the relative performance against the mean for all systems, orange/blue if significantly above/below at p <0.05. The final column is whether the scores for a system are significantly different all embeddings vs overlapping.

	All Relevent Embeddings			Overlapping Embeddings			
ICD9 Systems	MCSP	Examples	Relative	MCSP	Examples	Relative	Different?
Infections	7.72	2261	-0.043	6.84	334	0.223	Yes
cancers	9.00	1194	0.116	4.47	116	-0.202	Yes
endocrine	5.64	545	-0.301	3.98	193	-0.29	Yes
blood diseases	4.36	199	-0.459	3.67	81	-0.345	Yes
mental disorders	9.34	662	0.158	7.67	165	0.371	Yes
nervous system	8.44	1787	0.046	7.27	434	0.298	Yes
cardiovascular	8.12	869	0.007	7.74	307	0.383	No
respiratory	5.85	405	-0.274	4.64	132	-0.17	Yes
digestive	7.93	852	-0.016	4.62	210	-0.175	Yes
genitourinary	6.82	606	-0.154	5.75	210	0.027	Yes
pregnancy	10.27	1325	0.273	2.47	10	-0.559	Yes
skin	5.10	305	-0.368	4.01	102	-0.283	Yes
muskoskeletal	8.22	1041	0.019	5.24	168	-0.064	Yes
congenital	6.24	457	-0.227	5.05	101	-0.098	Yes
perinatal	9.84	310	0.22	2.49	11	-0.555	Yes
ill-defined	2.68	558	-0.668	2.81	224	-0.498	No
injury and poisoning	9.09	2975	0.127	2.42	86	-0.569	Yes

results when fewer or greater numbers of embeddings are considered. The initial analysis only considered embeddings that were common to all embedding sets, significantly restrict the numbers of embeddings considered compared to what could be considered when only using the *cui2vec* embeddings from Beam et al. This score difference may very well be due to whether less common medical terms are selected; the set of overlapping embeddings likely represent more common clinical concepts if all four training methods found them.

This last people is important to consider when assessing the project's results. The differences we find between medical systems could be largely due to how many rare clinical concepts a system has; these rare terms may have had fewer chances to be trained well due to seldomly coming up in the training data. This may or may not be a confounder. A future researcher using the embeddings in a particular field of medicine may be using embeddings from a variety of concepts in her field, including rare ones, in which case worse performance due to a field containing rare terms would be desired. Conversely, applications may only consider common terms in a field, in which case these rare terms would indeed be confounding. This limitation can be addressed, as will be discussed below.

5.2 Evaluating Sets of Embeddings

Another goal of the project was to asses the quality of embeddings from each set. This can be difficult to truly asses, as there is no gold standard to compare ratings. However, we believe the results indicated that the various evaluation methods used did do a good job of comparing qualities. Namely, Beam et al's *cui2vec* embeddings clearly come out ahead. This is expected, as these embeddings are the most modern, are trained on the largest amount of data, are trained on both journal articles and insurance data, and contain the largest number of embeddings. Our main contrbution in this line is the use of ICD9 systems as a new source of known relationships/clusters, which we then implemented in some of the methods. We believe these relations may be more clinically relevant than some used in the past, such as whether a clinical concept is a drug or symptom, though this is up for future work to decide.

ANY MORE IN THIS SECTION?

5.3 Comparing Evaluation Methods

Our work also allows us to observe differences between the evaluation metrics used. Again, this is difficult to concretely judge, as do we not know what the methods should result in. Yu et al's method of comparing correlation with similiarity as judged by resident physicians was hampered by having a small number of judged similarities (around 500) which led to few comparisons per system in our deployment. Our new method attempted, while simple, found smaller differences between both systems and sets of embeddings. Neither method finds the *cui2vec* embeddings to be significantly better than the others. Taken together, this may suggest these two methods are inferior. However, our new method does find similar results to the other methods, for instance also finding the significantly better systems, so it may not be worse. The method from Yu et al could likely be slight improved; some of the human simiarlity comparisons are not used as the CUIs chosen are not found within our embeddings, but could be changed to very related ones that are, eg Diabetes –¿ Diabetes Mellitus.

The remaining methods are quite similar. The two methods from Choi et al use discounted cumulative gain to judge whether neighbouring vectors are part of known relations, while Beam et al's method evaluates these relationships compared to a bootstrapped sample. The three methods could all use different sets of relation metrics, further muddling their relative performance. For instance, Beam et al's method as implement in our project evaluates known relations between drugs and the diseases they treat or prevent, but it could instead be changed to evaluate whether two concepts are related to the same medical system. This represents yet another degree of freedom inherent in our results, which challenges the interpretability. THIS IS KINDA WEIRD.

5.4 !Lessons Learned

The main lesson I learned from this project was to consider the evaluation ahead of time. The many variables - which set of embeddings, which system, which evaluation method, what the evaluation method uses for comparisons - make it a challenge to sort out definitive results. However, I am not sure what I would've changed retrospectively, and much of this lesson may simply be "time to learn some more stats".

5.5 !Was Project Succesful?

5.6 !Strengths and weaknesses

6 !Future Work

Improving the current project : Multiple avenues exist for improving the current project. Multiple methods exist for further statistical analysis, including those that are non-parametric, and those that would be able to consider all raw scores, without needing to work on means and disregard some data such as counts.

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.

Further consideration of the combination of variables could also be helpful. For instance, all of the evaluation methods could be adapted to use as their known-relation whether an embedding's CUI is within a system; this could reduce as a variable.

As discussed previously, a possible confounded is whether embeddings from a given medical system performance worse due to containing rare diseases. While not necessarily an unintended feature, this could be ameliorated by only using common CUIs. This was intended to be a part of this current project, but was deferred. Frequencies of ICD9 codes are not readily available except within public health databases which requires formal applications for, which likely could not be acquired in the timeframe of this project, though they represent an interesting avenue for future work.

Possible extension Evaluating Zhang et al's domain adaptation-trained embeddings in this project would help quantify how much improvement this technique may lead, if their embeddings could be obtained. DA could be carried out on embeddings from a poorly performing medical system to see if it is improved. Or, for a larger project, medical-field-specific embeddings could be trained, and compared to the ones used in this project, in order to determine the scale of possible improvement. Finally, our methods for evaluation consider metrics for embedding quality. While a basic component of possible NLP applications, a more 'real-world' evaluation could be attempted carrying out actual NLP tasks on documents from different medical systems to understand their relative performance from this perspective. For instance, NLP tasks using the embeddings could be carried out on articles on conditions from the various medical systems from UpToDate, a medical encyclopaedia, or Wikipedia. Or, even more applied, NLP tasks could be evaluated on medical documents produced by physicians of different specialities, an opportunity I may soon have with a personal project.

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 and Cai2013] Sheng Yu and Tianxi Cai. 2013. A Short Introduction to NILE. *arXiv:1311.6063 [cs]*, November.

[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.