Project Proposal: Do Word Embeddings Trained on General Medical Data Work for Psychiatric Concepts?

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

1 Introduction and Proposed Contribution

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 In their yet unpublished work, Beam et al (Beam et al., 2018) use an "extremely large" database of clinical notes, insurance claims, and full journal texts, and develop a new system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. This may be especially true in 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

This project aims to advance the application of word embedding techniques in psychiatry. Specifically, we will seek to determine whether embeddings trained on general medical data perform as well on psychiatric content as they do on other domains within medicine. We are unaware of prior work investigating this. We will compare multiple techniques for embeddings and evaluation. This will help determine generally how well these performance on psychiatric concepts, and whether various attributes may help or hinder this applicability, such as embeddings trained on larger training sets, or the use of DA. This may impact future work by suggesting if psychiatric applications should use general-medicine trained embeddings, or those trained only on domain-specific data.

2 Proposed Methodology

Generally, the project will deploy the embeddings of prior projects, using their evaluation methods to compare performance on psychiatric concepts with those from other fields of medicine. The comparison will be made with broader fields of medicine such as internal medicine, and those that are similarly specialized like ophthalmology.

We will compare the following embeddings/techniques, all of implement or are based upon word2vec:

- De Vine et al's (2014) embeddings trained on medical records and abstracts.
- Minarro-Gimenez et al's (2014) embeddings trained on medical manuals and articles, Wikipedia.
- Choi et al's (2016)'s two sets of embeddings trained differently using raw data mapped to a matrix based on UMLS techniques.
- Zhang et al's (2018) best performing embeddings using domain-adaptation techniques.
- Beam et al's cui2vec embeddings trained on health insurance claims and full journal texts.

The evaluation techniques to be replicated and used to determine psychatry-specific performance:

- De Vine et al's (2014)'s evaluation framework, comparing predicted vector similarity against human judgements, using the evaluation from (Koopman et al., 2012) which compares predicted similarity against human judgements from (Pedersen et al., 2007) and (Caviedes and Cimino, 2004).

  I remember seeing these, where did we put them? Enough psych to matter?

- Comparison against the UMNSRS benchmark per (Yu et al., 2017).

  About 500 relations. However, include drugs, disorders, symptoms. Could do a cui2icd? Or hand annotate? Downloaded the file into data. Maybe use 'may treat into icd9's to figure out what system'? Or perhaps symptoms are related as well

- Minarro-Gimenez et al's (2014)'s metric of predicting relationships between drugs based on the NDF-RT.

  Figure out if this is different enough from the first Choi one

- Choi et al's (2016) Conceptual Similarity Property, comparing predicted vector similarity with whether concepts are neighbouring in UMLS.

  Look at more; weird performance and unsure how to consider vs others

- Choi et al's (2016) Medical Relatedness Property, comparing predicted vector similarity with relatedness according to NDF-RT and the ICD9 groupings, based on these database's item relations such as "may-treat" and "may-prevent".
- Done, still look at top 10 maybe
- Beam et al's (2018) statistical score based on whether known similarities in UMLS, NDF-RT and other work are predicted correctly in at least 95% of bootstrapped samples of pairs of concepts.

  No code yet, sigh. Work on others first. Will need way to systemize drugs, non-disorders.

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 Data

The embeddings from De Vine et al's (2014) and both from Choi et al were obtained from the later's website. Beam et al's cui2vec embeddings were downloaded from this site. The corresponding author for the remaining embeddings from Minarro-Gimenez and Zhang were contacted, but a response was not received. Miarro-Gimenez's project files are available at this code archive site but lack documentation and could not be accessed.

TODO: Try to extract the embeddings. Probably accessed by TestClientThreadWord2vec.java, try opening it up in Eclipse, maybe can modify it and retrieve them.

3.2 Evaluation

We reimplemented Choi et al's Medical Relatedness Property metric by extending their code to calculate this property for specific domains of medicine. First, we calculate this propertry for all concepts repersented by ICD-9 CM codes with each system. TODO: we then repeated this amongst the most popular diagnoses within each category. TODO ANOVA was performed to determine the significance of differences between systems and embeddings.

3.3 Current Data Availability

Of the five works mentioned above, two have their data publicly available for download, one does not but has previously shared data with other authors, one is fully published so will likely share, and one is planning to share, but only when they are published. Relevant authors have been or will be contacted.

3.4 Project Flexibility and Extensibility

At a minimum, this project will use the available embeddings, and implement the evaluation metrics whose code is available, or whose description is sufficient to allow replication. An extensible system will be used such that future embeddings, when available, can be easily incorporated and compared. It is expected that, even if not all embeddings are available by the project due data, the implementation of the embeddings and evaluation metrics available will be the majority of the work for the total project, and will yield a sizeable contribution.

If the proposed methodology is implemented easily and quickly, a possible extension will be determine the feasibility of training new embeddings based only on psychiatric data, such as using a subset of the matrix used by Choi et al's (2016); we could try only using the portion of the matrix with terms related to psychiatry.

Alternatively, it may be interesting to use the embeddings from prior work to carry out various document-level summarization techniques, and compare doing so for psychiatric vs non-psychiatric documents. For instance, this could be done on articles from Wikipedia describing popular illnesses in and outside of psychiatry, or a similar set of articles from the medical practice manual and learning resource UpToDate.

In the longer term, this project may be applicable to a separate project applying NLP and ML techniques to a large BC Cancer clinical dataset consistency of the medical records of around 50,000 patients and their free text medical documents, numbering in the 100,000's. This dataset may allow both evaluation or training when available in the future.

4 Results

Using Choi et al's Medical Relatedness Property, we see the embeddings performance on all codes within each ICD-9 system in Table **??**. TODO ANOVA. We then repeat this for the most common diagnoses, finding the following results in Table. In this case, ANOVA shows:

5 Expected Results

Due to the uniqueness of psychiatry, we expect the various embeddings will generally perform worse when used for psychiatric concepts than those not in this speciality. We expected that the performance the various embeddings/techniques that work better generally will also work better for psychiatric content. However, it would not be overly surprising if the embeddings trained on larger dataset may perform worse for psychiatric terms, as the psychiatric-specific meaning of a word may get "drowned-out" more in larger datasets.

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.

[Caviedes and Cimino2004] Jorge E. Caviedes and James J. Cimino. 2004. Towards the development of a conceptual distance metric for the UMLS. *Journal of Biomedical Informatics*, 37(2):77–85, April.

[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.

[Koopman et al.2012] Bevan Koopman, Guido Zuccon, Peter Bruza, Laurianne Sitbon, and Michael Lawley. 2012. An Evaluation of Corpus-driven Measures of Medical Concept Similarity for Information Retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2439–2442, New York, NY, USA. ACM.

[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.

Table 1: Medical Relatedness Property Across All ICD-9 Diagnoses Within Systems.

ICD9 System	DeVine200	ChoiClaims300	ChoiClinical300	BeamCui
all	24.55	47.56	37.42	51.
infectious and parasitic disease	25.49	30.13	25.7	36.
neoplasms	30.51	50.3	35.97	52.
endocrine nutritional and metabolic diseases and immunity disorders	27.41	43.87	32.76	45.
diseases of the blood and blood-forming organs	21.67	36.76	34.65	39.
mental disorders	30.98	49.1	42.35	56.
diseases of the nervous system and sense organs	34.8	67.47	57.57	72.
diseases of the circulatory system	27.72	43.17	36.51	52.
diseases of the respiratory system	23.01	38.93	26.32	38.
diseases of the digestive system	21.11	57.95	35.6	59.
diseases of the genitourinary system	15.96	57.17	42.63	68.
complications of pregnancy childbirth and the puerperium	16.17	32.84	32.64	40.
diseases of the skin and subcutanous tissue	12.69	37.36	21.51	43.
diseases of the musculoskeletal system and connective tissue	17.72	48.18	38.46	44.
congenital anomalies	22.95	29.1	22.92	43.
certain conditions originating in the perinatal period	25.93	17.08	9.17	36
symptoms signs and ill-defined conditions	20.08	38.14	25.57	34.
injury and poisoning	9.44	35.06	37.38	37.

[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.

[Pedersen et al.2007] Ted Pedersen, Serguei V. S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, June.

[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.