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

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

1 Introduction and Proposed Contribution

1.1 Background

Recent work has applied natural language processing and machine learning to various predictive and classification tasks in medicine, such as EXAMPLE and EXAMPLE. Various methods are used to convert words to low-dimension vector representations which can then be used in these applications. These vectorizations are produced by training models on large corpora. For example, the popular *word2vec* system (Mikolov et al., 2013) initially trained embeddings using a skip-gram model 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 medical sources.

Recent approaches in this vein include De Vine et al (2014) who trained embeddings for medical concepts in the Unified Library Management System (ULMS) using journal abstracts from MEDLINE and clinical patient records. They then used these embeddings to compare predicted word similarity against human-judgements. Minarro-Gimenez et al (2014) trained using medical manuals, articles, and Wikipedia articles, comparing predicted word similarity between medications against the National Drug File - Reference Terminology (NDF-RT) ontology. Choi et al (Choi et al., 2016) improve on this work by learning on 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 In their yet unpublished work, Beam et al (Beam et al., 2018) use an "extremely large" database of clinical notes, insurance claims, and full journal texts, and develop a new system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 identifiers based on UCMLS, and then training on occurrences on these concepts with each other in a certain window length.

All of the above examples were both trained and tested on general medical data, from all fields of medicine. It is unclear whether these models will perform as well when applied to a specific field 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 tasks.

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 psychitric notes. Zhang et al (Zhang et al., 2018) then applied DA to word embeddings trained from various sources: general language (Wikipedia), medical articles (MEDLINE), non-psychiatric medical notes, and data from an online forum used by mental health patients, testing on a dataset of psychiatric notes, showing some improvement by using DA.

1.2 Contribution

This project seeks to build on these prior works advancing the application of word embedding techniques in psychiatry. Specially, 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. From our knowledge, this has not been examined previously. We will compare various prior techniques and embeddings, to determine if techniques such as DA, or training on larger data-sets, improve applicability to psychiatry. This will contribute to future applications of NLP to psychiatry by investigating whether or not embeddings trained on domain-specific data may be required. This will either give impetus to the collection of a large psychiatric dataset and the training of our own embeddings, or will reassure future work that general medical embeddings are sufficient.

2 Proposed Methodology

Generally, the project will to deploy the embeddings of prior projects, reusing their evaluation methods to compare these metrics when applied to psychiatric content compared to other fields of medicine. Psychiatric content (eg drugs, diagnoses) will be compared to those of general medical applicability, and other fields of medicine such as general practice, a general speciality like internal medicine, and another speciality that may have very specific terminology, ophthalmology. Performance metrics that can be compared against all embeddings and techniques will be used to determine if any improve domain applicability.

The embeddings/techniques to be replicated include:

- 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 used to compare psychiatric outcomes vs others:

- De Vine et al's (2014)'s evaluation framework, comparing predicted vector similarity against human judgements, comparing psychatric concepts vs others judged.
- Minarro-Gimenez et al's (2014)'s metric of predicting relationships between drugs based on the NDF-RT. This easily allows comparison between psychiatric drugs against those of other fields.
- Choi et al's (2016) Conceptual Similarity Property, comparing predicted vector similarity with whether concepts are neighbouring in UMLS. We will extended the concepts tested to include psychiatric and non-psychiatric ones
- 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". Subsets can again be specified.
- 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.

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 psychatric and non-psychatric drugs, or the most common diagnoses, based on prior epidemiology, in order to compare common, well described concepts.

2.1 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.

2.2 Project Flexibility and Extensibility

At a minimum, this project will use the available embeddings, and implement the 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. 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 based on prior work, such as using the matrix from Choi et al's work (2016) but only 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 psychiatric vs non-psychiatric documents. For instance, this could be done on articles from Wikipedia describing popular illnesses in psychiatry and non-psychiatry, or articles from the medical practice manual UpToDate.

3 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 of the various embeddings/techniques will follow a similar pattern with their general performance. 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

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.

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.

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 -3- 18. "Sitting on Pins and Needles": Characterization of Symptom Descriptions in Clinical Notes". *AMIA Summits on Translational Science Proceedings*, 2013:67–71.

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.

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

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.

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.