|  |
| --- |
| **eMERGE Network: Proposal for Analysis**Project/Manuscript Concept Sheet |
| **Reference Number** | NT367 |
| **Submission Date** | Nov 11, 2019 |
| **Project Title** | Using Federated Learning to Train Embeddings for Phenotyping Across a Nationwide Research Network |
| **Tentative Lead Investigator** *(first author)* | Yuan Luo |
| **Tentative Senior Author** *(last author)* | Wei-Qi Wei |
| **All Other Authors**  | Jennifer Pacheco, Luke Rasmussen, NLP WG, and any other eMERGE NLP collaborators explicitly expressing interest |
| **Sites Involved** | Northwestern, Vanderbilt, any other interested institutions |
| **Background / Significance** | As more NLP-based phenotypes have been proposed, developed and implemented using deep learning methods, it becomes critical to generalize such methods to across multiple institutions of eMERGE (and beyond). A key to such generalization is the extension of word embeddings to truthfully reflect their semantics and meanings across multiple institutions. The word or code embeddings are meaningful real-valued vectors where semantically similar words or medical codes usually have close embedding vectors. The word and code embeddings learned by neural networks often capture linguistic/conceptual regularities and patterns that are useful in computational phenotyping [1].Embeddings for medical concepts in structured EHR and unstructured clinical notes is the critical catalyst in driving biomedical deep learning systems. A portable embeddings that apply to multiple institutions is therefore critical in building a portable phenotyping system. Multiple systems have been proposed to learn embeddings in both structured and unstructured data settings [2, 3]. However, one challenge in learning good embeddings is the lack of large scale cross-institutional datasets, as such datasets are usually confined within institutional boundaries due to privacy concerns. In this work, we will leverage the recently developed tool of privacy preserving deep learning via additively homomorphic encryption [4] in order to train portable embeddings for unstructured words in clinical text. We will evaluate the performance using MIMIC-III (Medical Information Mart for Intensive Care III) are evaluated. We will compare the embedding from a single-site vs. the embedding from multiple sites on various concept extraction corpora, e.g., i2b2 2010, i2b2 2012, and SemEval 2014. Regarding the project to combine notes from multiple sites for a better word embedding representation, eMERGE is in a very unique position to leverage the close collaboration across several leading medical centers quickly generating a robust tool for public research.Our plan is to start with two sites, i.e. VUMC and NW. So it won't hurt the schedule of our other projects bad. We will demonstrate our preliminary results from the two sites and develop the pipeline in case other sites are willing to join (and ensure they can implement the training process quickly and easily). We believe the output will be an important product from the NLP group.[1] Luo Y, Cheng Y, Uzuner Ö, Szolovits P, Starren J. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. Journal of the American Medical Informatics Association. 2017 Aug 31;25(1):93-8.[2] Yao L, Mao C, Luo Y, editors. Graph Convolutional Networks for Text Classification. AAAI; 2019.[3] Choi E, Bahadori MT, Searles E, Coffey C, Thompson M, Bost J, Tejedor-Sojo J, Sun J, editors. Multi-layer representation learning for medical concepts. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016: ACM.[4] Aono Y, Hayashi T, Wang L, Moriai S. Privacy-preserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security. 2018;13(5):1333-45. |
| **Outline of Project** | 1. Randomly select a corpus of 200,000 clinical notes from each site across multiple note types, various specialties and temporal windows
2. Implement the federated learning neural networks for training word embeddings
3. Analyze task performance by comparing the federatedly learned embedding with the embedding learned on single institution data
4. Prepare and submit the manuscript.
 |
| **Desired Data - Common Variables\*** *(Available from the CC)* | [ ] Demographics [ ] ICD9/10 codes[ ] CPT codes[ ] Phecodes[ ] BMI | [ ] Common Variable Labs[ ] Common Variable Meds[ ] Other: Case/Control status on Phase I and [ ] Phase II phenotypes |
| **Other Desired Data *(Available from participating sites)*** | *Please specifically list out any data elements that participating sites would collect or extract from clinical or other sources for this project (i.e. not common variables above)* * Clinical notes (all type)
 |
| **Planned Statistical Analyses** | Descriptive summary of the number of tasks and the agreement among participants. |
| **Ethical Considerations** | None |
| **Target Journal** | JAMIA |
| **Milestones\*\*** | Total Duration of the study: 9 monthsCompletion of study design/approvals: December 2019Prepare the clinical notes: March 2020Implement the federated learning neural networks for training word embeddings: March- April 2020Draft of manuscript to authors: June-July 2020 (with a possible earlier design paper)First submission: Aug-Sept 2020 |

***\*\**** *This section should include the timeline for completion of project, including: approval, project duration, first and second draft of the paper and submission.*

**\*Common Variables available across all datasets:**

* Demographics: sex, year of birth, decade of birth, race, ethnicity
* Codes: (repeated values & age at event): ICD, CPT, Phecodes
* BMI: (repeated value & age at event) height, weight, BMI
* Labs: (lab name, repeated lab value & age at event) Serum total cholesterol, LDL, HDL, Triglycerides, Glucose fasting/non-fasting/unknown, & White Blood Cell count
* Medications: (medication name, repeated, & age at event) Cerivastatin sodium, Rosuvastatin, Simvastatin, Fluvastatin, Pravastatin, Lovastatin, Atorvastatin, & Pitavastatin
* Other: Case/Control status on Phase I and Phase II phenotype: only on GWAS dataset participants