UMLS Metathesaurus Atom Semantic Grouping with Context and Lexical Neural Networks Summer Experience Abstract

Will Hibbard

Lister Hill National Center for Biomedical Communications,

U.S. National Library of Medicine, National Institute of Health, Bethesda, Maryland

Mentor: Dr. Olivier Bodenreider

7/21/22

The Diversity in Data Science and Informatics (DDSI) summer internship program was a challenging and rewarding educational experience. Participating in this prestigious program allowed me to deepen my understanding of data science while also allowing me to in my weaker areas like ontology and neural network structures. In truth, I have never worked with a neural network before this program, but this internship allowed me the time and tools to learn and work with them. Overall, the experience was pleasantly outside of my comfort zone, and allowed me to learn valuable skills which I will incorporate into future research projects.

The purpose of our project was to build and train a neural network that could take atoms of information from the UMLS (Universal Medical Language System) database and accurately predict their correct semantic group. The motivation behind this project is that using a purely lexical approach has resulted in the models finding false synonymies between atoms. By using an embedding method in the model that also takes contextual information into account as well as lexical, this will yield accurate more accurate semantic predictions of atoms.

My supervisors Dr. Kin Wah Fung and Yuqing Mao gave me the papers, resources, and training I needed to understand neural networks enough to work with them. Kin Wah showed me papers from previous years of the program on methods previously tried for the sorting problem. A neural network is an algorithm that mimics the way a human brain learns. It breaks down information into quantifiable values and makes predictions based on those values. In this case, the information is turned embedded into vectors that the algorithm can work with. Since I had no prior experience with neural networks, I spent the first few weeks of the program learning what they were by reading papers and going through example code in blogs. I based my own neural network model on one blog which shows a neural network that classifies wine types (Verma, 2020). Yuqing Mao walked me through the blog and explained all the facets and features of a neural network. My model ended up being a classification type network with 5 layers and used SAP BERT as the embedding method. The 5 layers were used to let the model guess the sematic groups of the embedded atoms and correct its guesses on following pass throughs. I decided on using SAP BERT for the embedding method because it maintained the sentence and paragraph structure when embedding the atoms, which I thought was good contextual information. I had heard that other embedding methods lose that context when taking a “bag of words” approach, and I wanted to avoid that purely lexical approach for this project. I also chose SAP BERT because it was developed and trained on UMLS data (Cambridgeltl, 2021). This did present the

risk of possible bias, but I believed that its exposure to UMLS would help it make predictions better suited to the information I was working with.

The datasets used to experiment with the model were obtained using SQL code in TOAD, a remote server the NIH uses to specify and gather information from the UMLS. The whole dataset was composed of atoms only in English and from active source voca bularies. Since the entire dataset was too large to run in Biowulf, subsets of this dataset were created to run as proxies until I could figure out how to run the whole thing. The first data set was made of 10% of every semantic group in the whole dataset, thus allowing the model to show results that are in line with the proportional representation of the dataset. The second subset was made of 7000 atoms from each semantic group and was ran to test the model's ability to predict semantic groups without weighing in favor of larger groups. For the actual experimentation with the neural network, the datasets were run multiple time with different parameters such as epoch, batch size, learning rate, number of hidden layers, and number of nodes per hidden layer. Each value was incrementally increased or decreased and then run through Biowulf. Biowulf is a remote server in the NIH which allows researchers to process and analyze massive amounts of data.

The results obtained from these trails depicted two sets of values, those being category values and average values. The category values are stratified by the different semantic groups and denote the precision the model had in predicting the right group, the recall of how many assignments were correct, and the F-1 score of how well the model did for that group. The average values are derived from the whole dataset and denote the average accuracy of prediction, the macro average of how well that prediction was on its own, and the weighted average of how well prediction went with weights. The general average of the precision among runs was ~54%, which was not practical in application to our stated goal.

Though I could not get the whole dataset to run during my 9 weeks in the program, the silver lining is that I have some future directions for any future continuations of this project. The first direction that could be taken is creating a confusion matrix from the distribution of F-1 scores. The confusion matrix would depict which atoms got misclassified and what groups they got sorted into instead. This could show what groups are more likely to have atoms misattributed to them. A similar result may also be found with a binary prediction strategy in which the model just predicts which of all the atoms would get sorted into a particular semantic group. The data on groups the model has low prediction confidence with could feed into a third future direction, fine tuning the model. Fine tuning is when the weights of the network are adjusted so that they can account for a type of data or dataset being fed to it. Information about which groups may be more likely to get incorrect atoms can help adjust the weights of future neural networks for this task.

References

-

- Cambridgeltl, N. A. (2021, May 10). *CAMBRIDGELTL/Sapbert: [NAACL & ACL 2021] Sapbert: Self-alignment pretraining for Bert & XL-Bel: Cross-lingual biomedical entity linking.* GitHub. Retrieved July 22, 2022, from https://github.com/cambridgeltl/sapbert
- Verma, A. (2020, March 18). PyTorch [Tabular] —Multiclass Classification. Retrieved June 22, 2022, from https://towardsdatascience.com/pytorch-tabular-multiclass-classification-9f8211a123ab.