Prediction of drug-disease association by mining Pubmed abstracts using deep learning with application in drug repositioning

Methods

From Drugbank.com, drugs and their associated diseases were downloaded. The disease names were mapped to the UMLS disease concepts, for standardization. The combination of the drug and disease names were searched in the Pubmed for abstracts which contain both the drug and the disease names. The abstracts were rearranged that each drug-disease pair would have one line of text of abstracts, which were then fed into the neural network language model. The disease names were one-hot coded which were fed into the neural network as true responses.

The neural network language model was composed of seven layers. The input text matrix would go through the embedding layer first, then one-dimensional convolutional layer, followed by the one-dimensional max pooling layer. The fourth and fifth layers ware the same bi-directional LSTM layers. The last two layers are the fully connected dense layers with the second to last layer having twice the number of nodes as the last. The NN outputs of predicted responses were compared to the true responses and the accuracy was calculated to measure the prediction performance.