Can We Reliably Create New Variables for Social Scientists by Measuring Changes in Model Predictions?

Adrian Apaza (apaza@stanford.edu)

Motivation

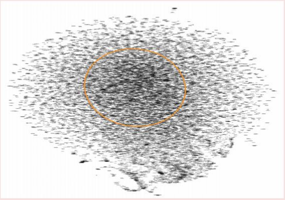
Machine Learning and Deep Learning are relatively new and just being applied to social science. One of the ways they are applied is in variable creation in measuring cultural change. This is done by seeing how the predictive power of a model changes over time and has been applied to issues such as political polarization (Gentzkow et al., 2017). However we will apply this to job postings to see if a model can reliably predict IT jobs and could capture changes. For example we could then get a variable as the rate of change of accuracy in the model over time as a proxy for the rate of technical and cultural change of IT jobs.

Models

I train three variant models as described for each of the possible GloVe choices. Each model is tuned for hyper parameters, but for each job posting's vector$_i$, I feed it into a standard neural network that utilizes logistic loss (cost =Σ and L2 regularization (with differing parameters λ for the models found after hyper parameter search). T-SNE dimensionality reductions of word embedding vectors are presented with annotated shapes and density (Maaten & Hinton).

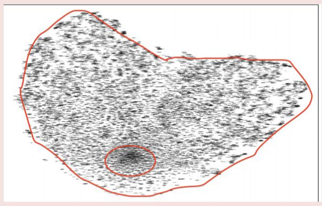
Model 1: GloVe from Job Postings Only

Model input is from GloVe using only job posting data.
3 layers with 400 neurons each.
$\lambda = 2*10^{-3}$



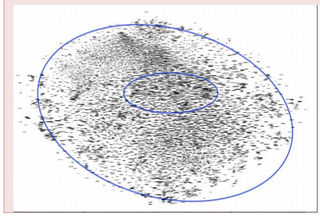
Model 2: GloVe from Job Postings + Wikipedia

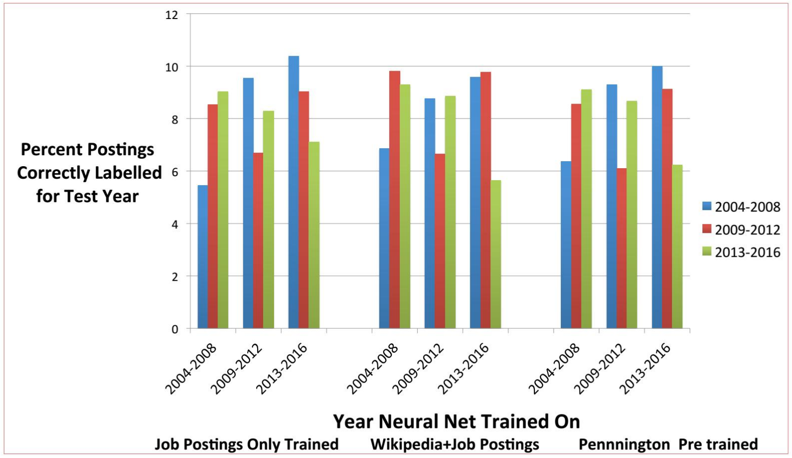
Model input is from GloVe using job postings + Wikipedia data of roughly same size.
3 layers with 400 neurons each.
$\lambda = 8*10^{-3}$



Model 3: GloVe from Pennington et al.

Model input is from GloVe (Wikipedia 2014 + Gigaword 5 data)
3 layers with 300 neurons each.
$\lambda = 8*10^{-3}$





Data and Features

I use online job board posting data from the country of Armenia through the yahoo mailing group of the website careercenter.am. This data has certain jobs labeled as IT or not, and collection was done by previous researchers funded by an American University of Armenia's research grant. The dataset includes multiple attributes like location and salary, but I only focus on job description, date posted, and whether or not it is an IT job. This dataset is extremely valuable due to its wide time range (2004-2016).

I use GloVe to extract features from the dataset. I use pre-trained word vectors from Pennington et al., training on solely data from the job postings dataset, and training combined from the job postings dataset and Wikipedia articles (since Wikipedia was also used for Pennington et al.) All are 300 dimensional. Since each job description has many words, I then find the average vector of each posting 'i' as follows:

$\text{Vector}_i = (1/\text{Number or Words})* (\Sigma_{\text{All Words}} \text{Vector}_{\text{word}})$

Models used on All Years Simultaneously

	Job Posting Only Embeddings	Job Postings + Wikipedia Embeddings	Pennington et al. Embbeddings
Train Error% (N=14251)	1.64199	2.561224	2.526139
Dev Error% (N=2750)	6.231579	6.568421	6.357895
Test Error% (N=2750)	6.273684	6.989474	6.778947

Results +Discussion

The goal of this project was to see if we can generate new variables for social science research based upon change of modeling predictions over time. With a variable of rate of change equal to the difference of predictive power for a given trained model used on different years. However we can see that modeling decisions do matter a lot. GloVe training choices can greatly influence the model, and surprisingly models trained on both Wikipedia data and Job postings data performed worse. Furthermore, as we can see from the graphs, the rate of change in predictive power is not the same across models. This is concerning as many social scientists are starting to use machine learning techniques such as deep learning, yet model specification can greatly alter their results.

For future work I would try to resolve these differences between models. And look to see why the GloVe model using Wikipedia and job posting data performs relatively poorly; we can already see that the T-SNE projections of the word vectors are very different.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy (2017). "Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech" NBER Working Paper 22423.

Jeffrey Pennington, Richard Socher, Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation

L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.