CS230 - Word Embeddings Quantify Exogenous Cultural Change

Paul Vicinanza *
Graduate School of Business
Stanford University
pvicinan@stanford.edu

Abstract

Longstanding theories of cultural evolution in society have emphasized the dualistic roles of endogenous and exogenous forces. Endogenous (internal or intrinsic to the system) encapsulates fads and fashions, incremental technological innovation, and linguistic drift while exogenous change refers to external shocks to the system such as war, economic crisis, or radical technological change such as the industrial revolution. Yet these theories remain empirically untested in large part due to the difficulty in quantifying cultural change. A renaissance in word embedding models, catalyzed by the recent developments of skip-gram negative sampling (SGNS) and GloVe, has provided researchers novel tools to examine cultural evolution through language. I study cultural evolution in the business community during the 2007/2008 financial crisis, by training separate word embedding models on a collection of over 100,000 quarterly earnings call segmented by time period. In doing so, I demonstrate that the rates of cultural change accelerate during periods of exogenous shocks and highlight specific words, such as "bankruptcy" and "housing," whose meaning evolves as a direct result of the crisis.

1 Introduction

How society grows, evolves, and thrives has long been a topic of formal inquiry in the natural and social sciences. It is not our superior cognitive ability that separates humans from other species but rather our propensity towards social learning, our capacity to build upon the advancements of others (Richerson and Boyd, 2005; Boyd et al., 2011). The development of deep neural network architectures are not attributable to any one individual. Without the development of mathematics and calculus, the invention of the first computer, or modern computational processing power, deep learning would not exist. This gradual accumulation of information aggregated over millennia and passed on from generation to generation is part of the monolithic construct sociologists term culture.

This paper presents the core results of my joint research project between CS230 and CS229a. In this paper, I discuss the main results from training embeddings over multiple time periods and comparing pairwise word embeddings between different models. I find that the correlations between models are highest during periods of economic stability, with the lowest levels of stability associated with the 2007/2008 financial crisis. In my other paper, I examine some properties of model stability, the impact of various hyper-parameters on model success, and identify a new set of hyper-parameters relating to inter-embedding comparison.

*Doctoral Student, Macro-organizational Behavior

,

2 Related work

While researchers have long sought to model cultural processes (for one example see Bikhchandani et al. (1992)), given the glacial pace at which social scientists adopt new methods word embeddings have yet to permeate mainstream cultural studies. Research to date has been conducted exclusively by computer scientists and computational linguists (e.g. Hamilton et al. (2016a)). This suite of papers studies the Google N-Grams corpus by training a separate embedding for each decade between 1800 and 1990. Hamilton et al. (2016b), for example, study the evolution of words in an embedding space over time and characterize the movement of tokens in the space to either be natural processes of linguistic drift or cultural change. In the 1900s the nearest neighbors for the word "gay" in the embedding space are "joyous", "daft", "merry", and "witty" but by the 1990s the nearest neighbors are "lesbian", "homosexual", and "heterosexual". The authors conclude that shifts in the nearest neighbors of a word over time accurately capture cultural change, as both the meaning of the word gay and the meaning *of being* gay has evolved over the last century.

In another study with the same corpus and methodology, Garg et al. (2018) trace the evolution of gender and ethnic stereotypes over time. They calculate the gender bias of a word in a decade by taking the average embedding distance between words that represent women (e.g. women, woman, female) and a target token and words that represent men. The difference between the similarity for men and women approximates the gender bias for the given word. In their analysis of occupational categories, the authors find that nurse, librarian, and housekeeper exhibit strong female biases in the embedding space while carpenter, engineer, and mechanic display strong male biases. For each word i in each decade j the authors can calculate the bias for this term b_{ij}. They then calculate the Pearson correlation coefficient between two decades for each bias value. Neighboring decades have relatively high bias correlations (0.7) while distant decades have lower correlations (0.5). Importantly, the authors find a discontinuity between the 1960s and the 1970s, corresponding to the rise of the woman's movement in the United States.

3 Data

For my project I obtained the full text transcripts of the quarterly earning calls of publicly traded firms between 2006 and 2016 from the crowd-sourced financial services website seekingAlpha. There are over 170,000 calls from nearly 6,000 firms totaling 170 million tokens. The quarterly earning calls are broken up into two sections. The first section comprises a prepared statement read by an executive (usually the CEO) or several executives for the firm. The second half is a question and answer (Q&A) section where analysts ask questions and the firm's executives respond. I restricted my analysis to the Q&A section because I sought to capture how conversational discourse in the business community evolves over time, not how the prepared remarks change.

Several pre-processing steps were taken to ensure data quality (See example 1). Statements were split into sentences using NLTK's sentence tokenizer. I stripped all punctuation from the strings, replacing them with space characters, and converted all characters to lowercase. Afterwards, the processed string was passed to a word tokenizing function to generate a list of tokens. I built a vocabulary to identify the most common tokens and removed all tokens below a certain threshold. I experimented with several vocabulary sizes, ranging from 5,000 to 20,000, with no difference in the results. I also experimented with removing stop words which made no difference as well, replicating previous findings in the literature (Lison and Kutuzov, 2017).

$$\begin{aligned} &Raw \\ &\text{"Sales are down in S. Africa and Ghana. Eddie, do you expect them to rise?"} \\ &Processed \\ &\text{[sales, are, down, in, s, africa, and][do, you, expect, them, to, rise]} \end{aligned} \tag{1}$$

4 GloVe

I segmented the data into yearly intervals and trained a separate embedding model for each interval. I experimented with different methods of producing word embeddings including word2vec and GloVe. I ultimately deciding upon GloVe for reasons of computational efficiency, because the package

enabled me to initialize the embedding from a pre-trained vector, and because GloVe has been show to have higher stability in small corpora (Mikolov et al., 2013; Pennington et al., 2014; Dingwall and Potts, 2018; Wendlandt et al., 2018).

GloVe trains the embedding vectors by first collecting the global co-occurrence matrix of words. Entry X_{ij} in the matrix X denotes the number of times that j occurs in the context of word i. Let P_{ij} be the probability of j occurring in the context of i, mathematically defined as $X_{ij}/\sum_k X_k$. At it's most basic form GloVe, seeks to learn the word embeddings $w \in \mathbb{R}^d$ and context vectors $\tilde{w} \in \mathbb{R}^d$ which accurately recall the ratio of co-occurrence probabilities between three words, i, j, and k.

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}} \tag{2}$$

Due to mathematical requirements, such that F be homomorphic, the authors reach the following equation for F by introducing bias terms b_i and $\tilde{b}_i$:

$$w_i^T \tilde{w}_j + b_i + \tilde{b}_j = log(X_{ij}) \tag{3}$$

This equation has several limitations. The model diverges when $log(X_{ij}) = 0$ and weighs all co-occurrences equally in a data set where the majority of co-occurrences never happen. As a result, Pennington et al. (2014) restate the objective function as a problem least squares with an additional weighting function, $f(X_{ij})$, where $f(x) = (x/x_{max})^\alpha$ when $x < x_{max}$ and 1 otherwise.

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log(X_{ij})^2 \tag{4}$$

Unlike previous work with embedding comparison (Hamilton et al., 2016a,b; Garg et al., 2018), I initialize all the models from one pretrained embedding (Wikipedia 2014 + Gigaword 5). This ensures that the embedding vectors can be directly compared without needing to solve the Orthogonal Prosecutes problem to align the embedding matrices.

5 Study Design

The core assumption of my methodology is that systematic differences between embeddings trained over different time periods quantifies meaningful latent semantic differences. Under this operating framework, I train separate word embeddings for documents split by date and assess overall similarity between the two models. Given two embeddings e_1 and e_2 and a word contained in the vocabulary $w \in v$, I define word similarity as the cosine similarity between the two embedding vectors for w: $s(w) = cossim(w_{e_1}, w_{e_2})$. A single measure of model alignment, M is mean word similarity.

$$M(e_1, e_2) = \frac{1}{len(v)} \sum_{w \in v} s(w) = \frac{1}{len(v)} \sum_{w \in v} \frac{w_{e_1} \cdot w_{e_2}}{\|w_{e_1}\| \|w_{e_2}\|} \tag{5}$$

Initially, I split the data by year, bucketing all of 2006, all of 2007, ... all of 2016. While the findings with this methodology are consistent with my conclusions, substantial differences in the quantity of text available for each year confounded the results. For example, there are approximately 3 million tokens available for training in 2006, while there are 25 million tokens for each year 2013 onward. To the extent that the quantity of text improves model accuracy (which it assuredly does), systematic differences between embedding similarity could be driven by text quantity.

To overcome this obstacle, I first identified the quantity of text available for the period before the 2007/2008 financial crisis. Depending on the exact start date of the crisis (results are robust to this specification), there are between 8 and 15 million tokens available for training, T_0. Using T_0 as the token count for each bin of documents, I construct n bins of documents ordered by date such that each bin has approximately T_0 tokens. While the length of time for each bin varies, the quantity of text remains constant. This methodology proved more robust to alternative embedding specifications so I report the results of this procedure.

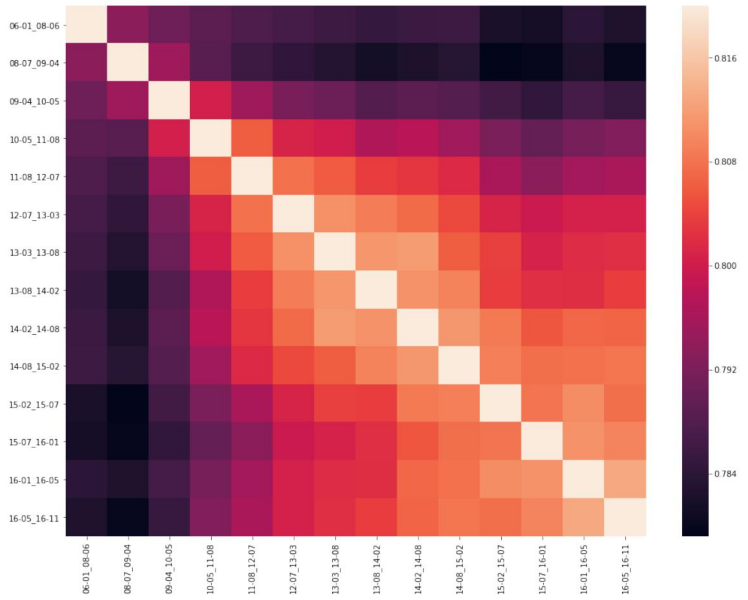

Figure 1: Heatmap of pairwise $M(e_1, e_2)$ computed for all embedding pairs

The process of hyperparameter selection is discussed in my final report for CS229a. I train an embedding for each year using a 100-dimensional word vector, vocab size of 15,000, x_{max} of 100, context window size of 10, window weighting function defined as $1/(x+1)$ where x is the distance in the vector between words i and j, learning rate of 0.05, α of 0.75, and 250 model iterations (the default in the literature is 100 iterations).

6 Results

Figure 1 presents the core results of the study. Analysis of inter-embedding similarity is restricted to the top 4,000 words in the vocabulary (robust to alternative specifications). Dates are presented in the format startYear-startMonth_endYear-endMonth. The top/left-most bar represents the embedding trained on the pre-crisis period, while the embedding trained during the height of the crisis (the collapse of Lehman Brothers onward) is the second bar. We observe as a general trend that embeddings trained on neighboring corpora have the highest similarity scores. We also see that the overall stability of the model increases as the time period proceeds. The period of time least correlated with other time periods is *not* the first time period but rather the time period corresponding to the height of the crisis. This finding supports sociological theorizing of cultural change during periods of crisis, which postulates that crisis promotes radical experimentation which subsides as normalcy is restored (Swidler, 1986).

Importantly, this finding is consistent with the argument that exogenous shocks accelerate the rate of cultural change. The correlations between embeddings is lowest during and immediately after the financial crisis and these differences persist in time. Not until the recovery is in full swing in 2011 does a stable rate of cultural change emerge. I replicate this map exactly if I use nearest neighbor similarity instead of cosine similarity, demonstrating that the results are not driven by any measurement choice.

One advantage of this methodology is that it enables identifying words whose meaning evolves during the period of observation. I find words which display high levels of stability in the later periods of the embedding (2015/2016) but whose location in the embedding space is substantially different from the first period of the model (2006/2007). While confidence in the vector position of any given is word is low given the data limitations (see 229a report for a discussion), selecting only those words which display high stability in the later periods alleviates much of the concern.

Table 1 reports a selection of these words. As evidence for exogenous shock theory, "bankruptcy" and "housing" appear to have changed meaning as a direct result of the crisis. "Housing" is strongly

Table 1: Evolving Words Between Early and Late Embeddings

Word	Early NNs	Late NNs
Housing	economy, home, residential, economic, recession	senior, starts, home, urban, funding, economy
Bankruptcy	filed, filing, retirement, severe, lenders, pool	court, filing, patent, filed, insurance, recession
Border	zone, forces, river, stirp, territory	cross, territory, zone, men, forces, agents, federal
Learning	curve, learn, centers, experience, schools	learn, machine, technology, training, software, creative
Analytics	presentation, map, deflation, reinvestment, web	tools, data, technology, solutions, capabilities, cloud
Social	networking, political, reform, education, physical	media, political, internet, digital, facebook, web

tied to economic logics in the early period but referenced more broadly in the later period. Likewise, the terms surrounding bankruptcy are linked to the financial crisis while this distinction erodes in the later period. For example, "pool" refers to a mortgage pool - the financial security issued by Fannie Mae and Freddie Mac responsible for exacerbating the crisis.

Consistent with early cultural embeddings research, I also observe evidence for endogenous change. I find a distinct shift in the meaning of terms associated with the data analytics revolution and identify several words in Table 2, such as "learning" and "analytics," which highlight this finding. I also observe a similar shift for the word "deep," but the word lacked adequate stability in the embedding space. Importantly, this is the first study of its kind to identify meaningful change over such a restricted time horizon.

7 Limitations/Future Work

While this project is an admirable first attempt at quantifying exogenous cultural shocks through word embeddings, it has several severe limitations I am working to resolve. Principally, because the data set begins in 2006, there is insufficient time before the crisis to establish a baseline rate of cultural change. Figure 1 would be much more convincing if the black bar occurring in the middle, flanked by orange on both sides. My team has found a similar data set of quarterly earnings call dating back to 2001 and are currently in negotiation for data access.

Additionally, I am expanding the analysis to an additional data set to demonstrate further robustness. I have recently acquired the complete set of U.S. Senate and House Committee meetings transcripts from 1995 to 2006 and am currently processing the data. My theory suggests that we should observe similar discontinuities in the pairwise embedding similarities as a result of 9/11.

8 Contributions

The work for this project was done entirely by me. My faculty advisers on this project are Dr. Amir Goldberg of Stanford GSB and Dr. Sameer Srivastava at the Berkeley Haas School of Business. Data scrapping was performed by the CIRCLE Research Support Team at the GSB. Software to construct weighted co-occurrence matrices and train embeddings was obtained from (Roam Analytics, 2018). Github code link: https://github.com/pvicinanza/cs230-class-project.

References

Boyd, Robert, Peter J. Richerson, and Joseph Henrich. 2011. The Cultural Niche: Why Social Learning is Essential for Human Adaption." *PNAS*, 108 (supplement 2): 10918-10925.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1992. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, 100(5): 992-1026.

Dingwall, Nicholas, and Christopher Potts. 2018. "Mittens: An Extension of GloVe for Learning Domain-Specialized Representations." *ArXiv:1803.09901v1*.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word embeddings quantify 100 years of gender and ethnic stereotypes." PNAS, 115(16): E3635-3644.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pg. 1489-1501.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. "Cultural Shift or Linguistic Drift? Comparing to Computational Measures of Semantic Change." Proc Conf Empir Methods Nat Lang Process: 2116-2121.

Lison, Pierre and Andrei Kutuzov. 2017. "Redefining Context Windows for Word Embedding Models: An Experimental Study." *Proceedings of the 21st Nordic Conference of Computational Linguistics*: 284-288.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26: *27th Annual Conference on Neural Information Processing Systems 2013.* Proceedings, 3111–3119.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532-1543.

Roam Analytics. 2018. Mittens. GitHub repository, https://github.com/roamanalytics/mittens.

Richerson, Peter J. and Robert Boyd. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution*. The University of Chicago Press: Chicago.

Swidler, Ann. 1986. "Culture in Action: Symbols and Strategies." *American Sociological Review,* 51(2): 273-286.

Wendlandt, Laura, Jonathan K. Kummerfled, and Rada Mihalcea. 2018. Factors Influencing the Surprising Instability of Word Embeddings. arXiv:1804.09692v1.

CS229a - Inter-embedding Comparison and Latent Hyper-parameter Evaluation

Paul Vicinanza *
Graduate School of Business
Stanford University
pvicinan@stanford.edu

1 Introduction

Word embedding models offer great promise for scholars seeking to quantifying culture and its evolution. While early embeddings are trained on billions of documents compiled over decades, modern applications have relied upon corpora which are orders of magnitude smaller, and it remains unclear how the stability of the embedding space responds to the diminution. Particularly when the comparison of different embeddings is at the heart of the analysis, how can researchers be certain that the difference between two embedding vectors for a given word highlight meaningful differences between the two corpora and not artificial noise induced by a smaller sample size?

I study cultural evolution in the business community during the 2007/2008 financial crisis in my write-up for CS230. By training separate word embedding models on a collection of over 100,000 quarterly earnings call segmented by time period, I demonstrate that the rates of cultural change accelerate during periods of exogenous shocks. I also highlight specific words, such as "bankruptcy" and "housing," whose meaning evolves as a direct result of the crisis. In my final project for CS229a, I focus instead on process of hyper-parameter selection and highlight a unique set of hyper-parameters needed to compare different embedding models currently unmentioned in the literature. I highly recommend briefly skimming the CS230 write-up before reading this report for added context.

While the CS230 paper highlights the main results, this paper instead focuses on the process used to generate them. I identify three distinct buckets for hyper-parameter selection which may influence the results. The first two, processing of the raw data and embedding model parameters, are being discussed in the literature. I could find no paper discussing the third set of hyper-parameters, those related to inter-embedding comparison, and thus elected to focus my discussion on this third bucket.

2 Related work

Research accessing the stability of word embeddings on corpora of limited size remains in its infancy. Wendlandt et al. (2018) evaluate the stability of word embeddings across the New York Times (NYT) corpus for word2vec, GloVe, and PPMI by predicted word stability through regression analysis. They predict the cosine similarity of two embedding vectors for a given word trained on different sections of the New York Times (e.g. Business vs. Arts vs. Sports vs. all NYT). For order variant models such as word2vec, which stream in the data, the order with which the words first appear in the training data significantly impact the stability of the embedding space. They also find that numerals, verbs, determiners, adjectives, and nouns have the highest stability while punctuation has the lowest stability. Overall, they find that word frequency is **not** a major factor in word stability and that GloVe provides the highest level of stability.

Antoniak and Mimno (2018) evaluate the stability of words in an embedding space by bootstrapping documents across a variety of corpora, including the New York Times, Reddit, and U.S. Circuit

*Ph.D. Student, Macro-Organizational Behavior

,

Court decisions, and compare word stability *within* a given corpus. They also compare the relative stability of word2vec, GloVe, and PPMI. In a major departure from Wendlandt et al. (2018), the authors evaluate stability along 20 keywords identified through a LDA topic model, rather than all words in the corpus. They ultimately reach a rather pessimistic conclusion: "The use of embeddings as sources of evidence needs to be tempered with the understanding that fine-grained distinctions between cosine similarities are not reliable and that smaller corpora and longer documents are more susceptible to variation" (pg. 116).

3 Dataset and Pre-processing

For my project I obtained the full text transcripts of the quarterly earning calls of publicly traded firms between 2006 and 2016 from the crowd-sourced financial services website seekingAlpha. There are over 170,000 calls from nearly 6,000 firms totaling 170 million tokens. The quarterly earning calls are broken up into two sections. The first section comprises a prepared statement read by an executive (usually the CEO) or several executives for the firm. The second section is a question and answer (Q&A) section where analysts ask questions and the firm's executives respond. I restricted my analysis to the Q&A section because I sought to capture how conversational discourse in the business community evolves over time, not how the prepared remarks change.

Several pre-processing steps were taken to ensure data quality. Statements were split into sentences using NLTK's sentence tokenizer. I stripped all punctuation from the strings and replaced them with space characters and converted all characters to lowercase. Afterwards, the processed string was passed to a word tokenizing function to generate a list of tokens. I built a vocabulary to identify the most common tokens and removed all tokens below a certain threshold. I experimented with several vocabulary sizes, ranging from 5,000 to 20,000, with no difference in the results. I also experimented with removing stop words which made no difference as well, replicating previous findings in the literature (Lison and Kutuzov, 2017).

I did find one preprocessing step that significantly influenced inter-embedding stability. Filtering vocabulary *after* splitting the documents into groups (as to before) lead to widely diverging word embeddings. For example, I could either filter the text to tokens which occur a minimum of 100 times in the entire corpus, or I could split the corpus in two and filter each set of documents for tokens which occur at least 50 times. I believe the latter procedure introduced substantial error because it meant important context words were missing in one corpus but present in another. I believe this explanation to be correct because the *greater* the minimum word frequency X_{min} (tokens which occur less than X_{min} are stripped) the *greater* the disparity between embeddings. Previous studies have not discussed the discrepancy and filter terms using the latter method. My finding suggests that inter-embedding stability could be improved by filtering text to a common set of terms before training the model.

4 GloVe

I segmented the data into yearly intervals and trained a separate embedding model for each interval. I experimented with different methods of producing word embeddings, including word2vec and GloVe, ultimately deciding upon GloVe for reasons of computational efficiency, because the package enabled me to initialize the embedding from a pre-trained vector, and because GloVe has been show to have higher stability in small corpora (Mikolov et al., 2013; Pennington et al., 2014; Dingwall and Potts, 2018; Wendlandt et al., 2018).

GloVe trains the embedding vectors by first collecting the global co-occurrence matrix of words. Entry X_{ij} in the matrix X denotes the number of times that j occurs in the context of word i. Let P_{ij} be the probability of j occurring in the context of i, mathematically defined as $X_{ij} / \sum_k X_k$. At it's most basic form GloVe, seeks to learn the word embeddings $w \in \mathbb{R}^d$ and context vectors $\tilde{w} \in \mathbb{R}^d$ which accurately recall the ratio of co-occurrence probabilities between three words, i, j, and k.

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}} \tag{1}$$

Due to mathematical requirements, such that F be homomorphic, the authors reach the following equation for F by introducing bias terms b_i and $\tilde{b}_i$:

$$w_i^T \tilde{w}_j + b_i + \tilde{b}_j = log(X_{ij}) \tag{2}$$

However, this equation has several limitations. The model diverges when $log(X_{ij}) = 0$ and it weighs all co-occurrences equally in a dataset where the majority of co-occurrences never happen. As a result, Pennington et al. (2014) restate the objective function as a problem least squares with an additional weighting function, $f(X_{ij})$, where $f(x) = (x/x_{max})^\alpha$ when $x < x_{max}$ and 1 otherwise.

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log(X_{ij})^2 \tag{3}$$

Unlike previous work with embedding comparison (Hamilton et al., 2016a,b; Garg et al., 2018), I initialize all the models from one pretrained embeddings. This ensures that the embedding vectors can be directly compared without needing to solve the Orthogonal Prosecutes problem to align the matrices.

5 Model Evaluation

Given a partition of the data into two equally sized bins, what GloVe parameters provide the highest level of inter-embedding similarity between words? Given two embeddings e_1 and e_2 and a word contained in the vocabulary $w \in v$, I define word similarity as the cosine similarity between the two embedding vectors for w: $s(w) = cossim(w_{e_1}, w_{e_2})$. A single measure of model alignment, M is mean word similarity.

$$M(e_1, e_2) = \frac{1}{len(v)} \sum_{w \in v} s(w) = \frac{1}{len(v)} \sum_{w \in v} \frac{w_{e_1} \cdot w_{e_2}}{\|w_{e_1}\| \|w_{e_2}\|} \tag{4}$$

6 Embedding Hyper-parameter Tuning

Table 1 describes the process of hyper-parameter selection, where I informally tuned to minimize model distance. Because training one embedding model takes several hours, even when parallelized on a remote server, rapid iteration is not feasible. As a result, only several values were experimented with for each parameter. For the sake for brevity I'll only discuss the two parameters which had a counter intuitive impact on model stability. Firstly, a smaller context window size was associated with higher inter-embedding similarity. Smaller context windows capture syntactical meaning while larger context windows measure semantic meaning (Lison and Kutuzov, 2017). Thus while smaller windows provide higher model alignment, they do so because they neglect the broader cultural connotations of the word and are ill-suited for my research project. The most surprising result was that model similarity diverged as the number of iterations increased. Two embeddings were more similar if trained for 100 iterations than 1000 iterations. I attribute this effect to my choice of a fixed initialization at the pretrained embedding. More iterations allows the model to drift further from the initial embedding (potentially overfitting insufficient data). Neither the number of iterations nor the window size impacted the core finding of my project (highlighted in the CS230 report).

7 Model Comparison Hyper-parameters

Independent of the hyper-parameters associated with data processing and training the word embedding models, I identify a third set of parameters related to model comparison. The first choice a researcher must make is whether to evaluate word stability by the word's nearest neighbors or cosine similarity. A word can move in the embedding space while retaining the same nearest neighbors or remain stable in the embedding space while it's neighbors change. For example, I find that the nearest neighbors of words such as car, train, plane, bus, etc. remain stable throughout the time period while the global position of this cluster shifts. Figure 1 reports all word self-cosine similarities. We observe a highly

Table 1: GloVe Hyper-parameter Assessment

Parameter	Final Value	Description	Details
# Dimensions	100	Number of dimensions for the word embedding	Tried 50, 100, 200, and 300 dimensional vectors. Major improvement between 50 and 100 with little difference afterwards.
x_{max}	100	Maximum x value for window weighting function in GloVe	Default for GloVe. Tried 1,000 and 100,000. Improves model consistency at the cost of overfitting to common words
Window size	10	Context window size for building co-occurrence matrix	Smaller windows provide higher stability. Tried 3, 10, and 20.
Window weighting function	$\frac{1}{d+1}$	Function to build weighted co-occurrence matrix.	Default. Untested.
Learning rate	0.05	Learning rate for Adam Optimization	Default. Untuned.
α	0.75	Exponential parameter for weighting function	Default. Untuned.
Number of iterations	250	Number of epochs through the data	Default in GloVe is 100. Embeddings are rarely run to convergence given the immense computational demands. I ran 100, 250, and 1000 iterations.

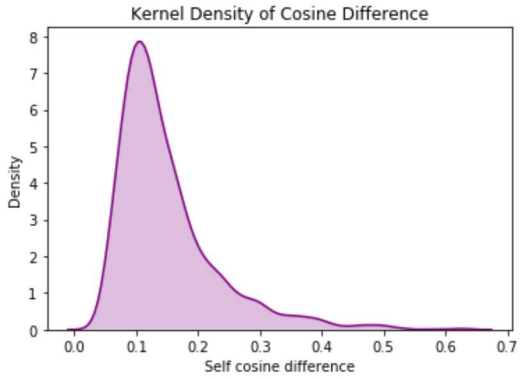

Figure 1: Kernel Density of Word-Self Cosine Similarities

left-skewed distribution with a mean of approximately 0.15. The tail of the distribution is comprised of words with highly imprecise meaning ("annualizing", "reacceleration") and typos ("wehave"). Interestingly, the words in the tail of distribution remain constant over random segmentations of the data, suggesting that high levels of dissimilarity are properties of the words themselves, rather than how the documents are split.

If nearest neighbors are chosen as the metric of model alignment (what percentage of the n-nearest neighbors of word w in e_1, NN_{we_1} are also in NN_{we_2}), the additional hyper-parameter of the size of the nearest neighbors must also be tuned. Smaller nearest neighbor sizes capture the most proximate words and closest meaning, but also display the highest variance. As the size of the nearest neighbor set increases, the variance of difference in sets also increases. However, the mean difference (40%) remains constant, suggesting that smaller set sizes are best at discrimination between the (in)-stability of certain words in an embedding. While most authors rely upon cosine similarity, (Hamilton et al.,

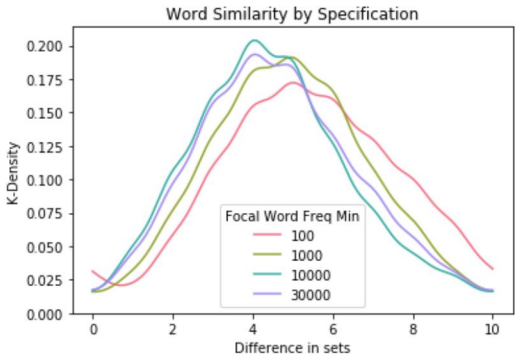

Figure 2: Difference in Sets for 10 Nearest Neighbors

2016b) find that changes nearest neighbors perform better at capturing shifts in meaning than cosine similarity, suggesting that nearest neighbors similarity may be better metric.

A third hyper-parameter which must be selected is the vocabulary used for inter-model comparison. I demonstrated earlier that the corpora should be trimmed to a shared vocabulary before training the embedddings, but it is not obvious what words should be included in the comparison. Wendlandt et al. (2018) use all tokens in the vocabulary while Antoniak and Mimno (2018) use a curated set of just twenty words, highlighting a wide divergence in the current literature. Figure 2 shows the kernel density for the difference in the 10 nearest neighbors for various cut-points in the vocabulary. Words which occur at least 100 times are shown in the red bar. I find that word stability does increase for tokens with frequency between 100 and 1,000, but that after 1,000 there is little increase in stability. This suggests that word frequency is an important attribute for stability in the model, but only up to a certain threshold.

8 Conclusion / Future Work

Inter-model reliability with smaller corpora is a function of the seemingly innocuous decisions researchers make in evaluation. Of the two papers discussing embedding stability, Wendlandt et al. (2018) finds that GloVe is the *most* stable embedding model while Antoniak and Mimno (2018) finds that it produces the *least* stable embedding vectors. I attribute their diverging findings to differences in what I term model comparison hyper-parameters, which are independent of the parameters used to train the embedding models themselves. I identify some of the parameters, such as the choice of model evaluation metrics, the sampling of the vocabulary, and the source of corpora variation and demonstrate their impact on inter-model reliability.

I have begun developing code to bootstrap the embeddings (randomly re-sample documents in the corpus to generate variance estimates for word vectors). Minimizing computational cost while limiting memory usage has proven tricky. Future work will focus on accessing the stability of my findings discussed in the CS230 project report. My team has recently acquired the complete transcripts of U.S. House and Senate Committee Hearings from 1995 to 2006 and I will look to replicate my findings in this new data set by treating the 9/11 attacks as the exogenous shock.

9 Contributions

The work for this project was done entirely by me. My faculty advisers on this project are Dr. Amir Goldberg of Stanford GSB and Dr. Sameer Srivastava at the Berkeley Haas School of Business. Data scrapping was performed by the CIRCLE Research Support Team at the GSB.

References

Antoniak, Maria and David Mimno. 2018. Evaluating the Stability of Embededing-based Word Similarities. *Transactions of the Association for Computational Linguistics*, 6: 107-119.

Dingwall, Nicholas, and Christopher Potts. 2018. "Mittens: An Extension of GloVe for Learning Domain-Specialized Representations." *ArXiv:1803.09901v1*.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word embeddings quantify 100 years of gender and ethnic stereotypes." PNAS, 115(16): E3635-3644.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pg. 1489-1501.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. "Cultural Shift or Linguistic Drift? Comparing to Computational Measures of Semantic Change." Proc Conf Empir Methods Nat Lang Process: 2116-2121.

Lison, Pierre and Andrei Kutuzov. 2017. "Redefining Context Windows for Word Embedding Models: An Experimental Study." *Proceedings of the 21st Nordic Conference of Computational Linguistics*: 284-288.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26: *27th Annual Conference on Neural Information Processing Systems 2013.* Proceedings, 3111–3119.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532-1543.

Roam Analytics. 2018. Mittens. GitHub repository, https://github.com/roamanalytics/mittens.

Wendlandt, Laura, Jonathan K. Kummerfled, and Rada Mihalcea. 2018. Factors Influencing the Surprising Instability of Word Embeddings. arXiv:1804.09692v1.