CS 224N: Language Dynamics analysis through Word2Vec Embeddings

Jeha Yang
Department of Statistics, Stanford University
jeha@stanford.edu

Claire Donnat
Department of Statistics, Stanford University
cdonnat@stanford.edu

Abstract

Recent breakthroughs in word representation methods have generated a new spark of enthusiasm amidst the computational linguistic community, with methods such as *Word2Vec* have indeed shown huge potential to compress insightful information on words' contextual meaning in low-dimensional vectors. While the success of these representations has mainly been harvested for traditional NLP tasks such as word prediction or sentiment analysis, recent studies have begun using these representations to track the dynamics of language and meaning over time. However, recent works have also shown these embeddings to be extremely noisy and training-set dependent, thus considerably restricting the scope and significance of this potential application. In this project, building upon the work presented by [1] in 2015, we thus propose to investigate ways of defining interpretable embeddings, and as well as alternative ways of assessing the dynamics of semantic changes so as to endow more statistical power to the analysis.

1 Problem Statement, Motivation and Prior Work

The recent success of Neural-Network-generated word embeddings (word2vec, Glove, etc.) for traditional NLP tasks such as word prediction or text sentiment analysis has motivated the scientific community to use these representations as a way to analyze language itself. Indeed, if these low-dimensional word representations have proven to successfully carry both semantic and syntactic information, such a successful information compression could thus potentially be harvested to tackle more complex linguistic problems, such as monitoring language dynamics over time or space. In particular, in [1], [5], and [7], word embeddings are used to capture drifts of word meanings over time through the analysis of the temporal evolution of any given word' closest neighbors. Other studies [6] use them to relate semantic shifts to geographical considerations.

However, as highlighted by Hahn and Hellrich in [3], the inherent randomness of the methods used to encode these representations results in the high variability of any given word's closest neighbors, thus considerably narrowing the statistical power of the study: how can we detect real semantic changes from the ambient jittering inherent to the embeddings' representations? Can we try to provide a perhaps more interpretable and sounder basis of comparison than the neighborhoods to detect these changes? Building upon the methodology developed by Hamilton and al [1] to study language dynamics and the observations made by Hahn and Hellrich [3], we propose to tackle this problem from a mesoscopic scale: the intuition would be that if local neighborhoods are too unstable, we should thus look at information contained in the overall embedding matrix to build our statistical framework.

In particular, a first idea is that we should try to evaluate the existence of a potentially "backbone" structure of the embeddings. Indeed, it would seem intuitive that if certain words –such as "gay" or "asylum" (as observed by Hamilton et al) have exhibited important drifts in meaning throughout the 20th century, another large set of words – such as "food","house" or "people" – have undergone very little semantic change over time. As such, we should expect the relative distance between atoms in this latter set (as defined by the acute angle between their respective embeddings) to remain relatively constant from decade to decade. Hence, one could try to use this stable backbone graph as a way to triangulate the movement of the other word vectors over time, thus hopefully inducing more interpretable changes over time.

Such an approach could also be used to answer the question of assessing the validity of our embeddings for linguistic purposes: how well do these embeddings capture similarity and nuances between words? A generally

wide-spread method consists in using reference datasets, annotated by humans, giving correlations between certain pairs of words as benchmarks values to assess the quality of the embeddings' representations. However, this seems to lead to a temporal bias, as shown by figure 1: only the most recent embeddings correlate well with the human annotated data. A stable dictionary of word vectors could thus potentially solve this question and provide a sounder way of benchmarking our embeddings.

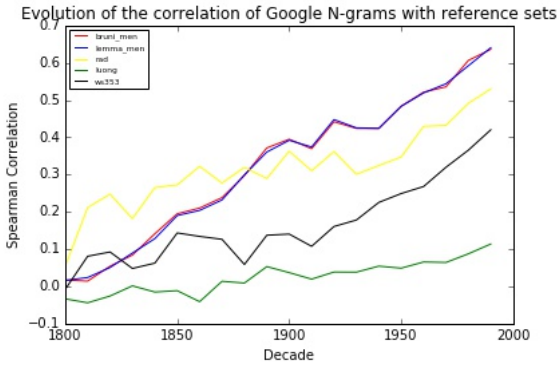


Figure 1: Evolution of the correlation between the reference sets and the Decadewise-agglomerated embeddings

2 The data

In this project, we choose to build upon the results provided by the analysis of Hamilton et al to analyze the feasibility of a "stable dictionary" construction. Following the footsteps of the authors, the data that we selected consists in the 300-dimensional *word2vec* embeddings obtained from the COHA 4-grams and 2009 Google 5-grams datasets, aggregated over periods of ten year. As prescribed by the authors in [1], we filter the 100,000 *word2vec* embeddings by getting rid of the proper nouns and retaining only $|\mathcal{V}| = 10,000$ most frequent ones over time.

Since the embeddings are inherently random and the orientations of the vectors are thus potentially likely to change from decade to decade, an additional alignment step is required – that is, we rotate the embeddings by R where R is the rotation matrix such that the overall displacement is minimal:

$$R = \text{ arg min}_{Q:Q^TQ=I}||W_t - RW_{t+1}||_F$$

where $W_t \in \mathbb{R}^{|\mathcal{V}|\times 300}$ denotes the matrix of the word embeddings for decade t, and $||\cdot||_F$ is the Frobenius norm.

The assumptions that will be used to develop our framework are summarized as follows:

- A large proportion of words are "stable" over time: we do not except most common words ("food", "car", etc..) to undergo important changes in meaning over time
- We focus on the word embeddings from the 20th century (1900+), which should also limit the variability of the vocabulary

3 First Approach: assessing the embeddings' stability

An natural model of the evolution of the embeddings could be given by:

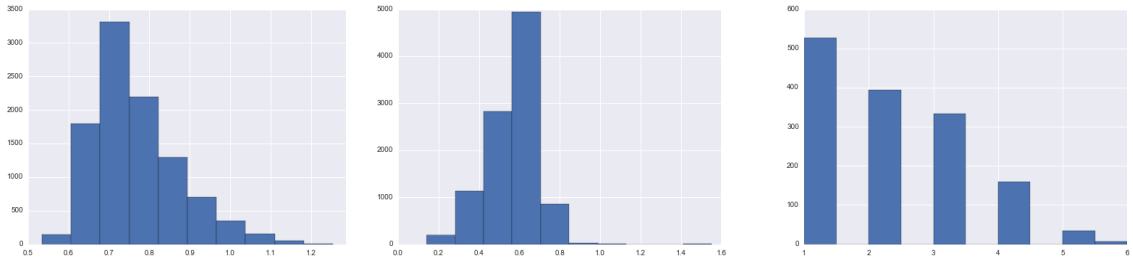
$$\forall t, \quad W_{t+1} = W_t + Z_{t+1} + I_{t+1}$$

where Z_{t+1} is a noise displacement matrix, and I_{t+1} denotes the innovation at time $t+1$, i.e, the true significant semantic changes from time t to $t+1$.

A first natural assumption is to hope for a high "signal-to-noise" situation, where we would expect a large proportion of the embeddings to remain quite stable over time, and a small proportion to exhibit large variations in orthogonal directions to the previous vector: the innovation matrix I_{t+1} should thus be 0 in most cases, and only record rare but truly significant semantic changes. We thus introduce the displacement ratio of each vector, denoted as $\Delta_{i,t}$:

$$\Delta_{i,t} = \frac{||\Pi^{\perp W_{t+1,i}}[W_{t+1,i} - W_{t,i}]||_2}{||W_{t,i}||_2}$$

where $\Pi^{\perp W_{t+1,i}}$ denotes the projector in the orthogonal complement of $W_{t+1,i}$,such that $\Delta_{i,t}$ could serve as benchmark to select a stable dictionary among the words that consistently score the lowest according to that metric. Histograms of the distribution of such ratios are given in Figure 3a and Figure 3b, for both normalized and unnormalized embeddings.



(a) 50s/60s decades: normalized Google 5-grams (b) 50s/60s decades: unnormalized COHA 4-grams (c) Persistence of the lowest scoring vectors

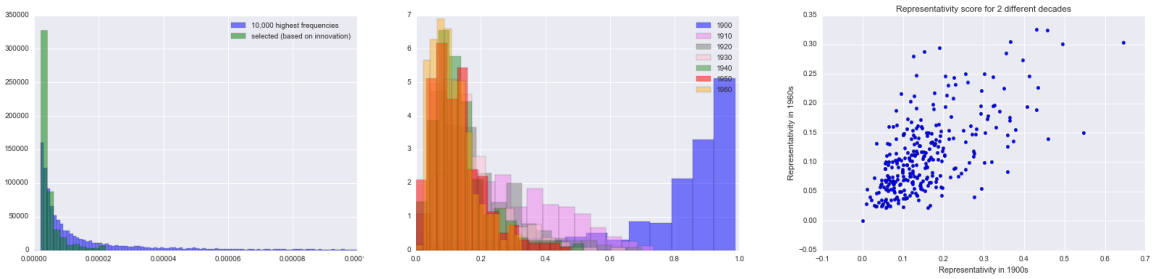
Figure 2: histograms for the distribution of the orthogonal displacement ratio $\frac{||\Pi^{\perp W_{t+1,i}}[W_{t+1,i}-W_{t,i}]||_2}{||W_{t,i}||_2}$ for each word vector i

A first naive approach to select a stable dictionary of words would be to select words that score persistently low according to the displacement-ratio metric. However, this task rapidly proves to be difficult, as shown by Figure 7: not only does the innovation+noise term account for 50% of each word vector's norm at time $t+1$ – thus indicating a high variability of the embeddings and a low signal –, rare are the vectors that "persistently" score low according to the orthogonal displacement metric. This also highlights the shortcomings of any purely local consideration, which will struggle to distinguish real signal from noise.

To push the study further, we nonetheless try an "innovation" driven procedure on the COHA dataset, for decades ranging from 1900 to 1970: for each decade, we filter the words, retaining only those with displacement ratio below 40%. Among these selected vectors, a "significant" persistence level was defined as the vectors that scored at least 3 times below the latter threshold, which we use as dictionary. This approach leads to the selection of 553 different words, and we test the relevance of our dictionary by evaluating its "representativity level": among all words that were not selected, we select a random subset of size 300, which we then try to express as a linear combination of the dictionary's atoms $\mathcal{D}$ (using l1-penalized linear regression). For each of these test words – here denoted as w– we then define a score by:

$$\text{score}_w = \frac{||\mathcal{D}\hat{\beta}_w||^2}{||w||^2}$$

where β_w are the recovered lasso coefficients. Intuitively, this score captures how much of the embedding w the dictionary is able to recover. The results are shown below.



(a) Comparison of the distribution of the word frequencies in the overall dictionary to the selected basis (b) Distribution of the representativity of the dictionary over decades (c) Comparison of the scores for each test word, in the 1930s vs 1960s

These plots show that our basis selection has failed in that it does not successfully represent the overall samples: the distribution of the scores averages in general around 20%, and although our dictionary exhibits stable angles over time and low correlation between atoms, it fails to correctly represent the information of the overall word embeddings. This can be explained that the words that were chosen rank among the rarest ones (figure 3a), and thus lack expressive power. Hence, we need to develop a less myopic dictionary search strategy, and guide our selection through the amount of information that each embedding can incorporate into the dictionary.

4 Towards an information-richer dictionary: a convex optimization approach

The quest for an interpretable and representative dictionary of words can be reformulated as follows:

$$\text{Find a subset } \mathcal{S} \text{ of } \{1..|V|\} \text{ s.t.} \forall i \in \{1..|V|\}, \quad \hat{w}_i = \sum_{j \in \mathcal{S}} \alpha_j w_j + z_i, \qquad z_i = \text{noise}$$

In other words, we are trying to find a diagonal matrix A such that $A_{ii} = 1 \iff i \in \mathcal{S}$ which should both provide good representation of the other words (i.e, minimize the mean square error between the reconstruction and the original embedding matrix W), and should select relatively uncorrelated atoms:

$$\min_A \min_\beta \sum_{i=1}^{|V|} (\hat{W}_i - A\hat{W}\hat{\beta})^2$$

$$\text{s.t} ||\beta|| \leq t \quad \text{and} \quad ||\hat{W}^T_{i \in \mathcal{S}} \hat{W}_{i \in \mathcal{S}}||_2 \leq u$$

(The two conditions impose (1) sparsity on the selected coefficients for each word reconstruction, and (2) uncorrelatedness of the dictionary.)

However, this is a hard combinatorial problem that could potentially induce us to compare all possible sets of dictionary of fixed size H, and compare how well they are able to reconstruct the original embedding matrix. An approximation to this problem could then be formulated as follows:

$$\min_p \min_A ||W - AW||_2^2 + \lambda \sum_{j=1}^{|V|} |A_{ij}| + \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} p_j^{A_{ij} \neq 0} (1 - p_j)^{A_{ij} = 0} + \lambda_2 \sum_{j=1}^{|V|} p_j$$

that is, each of the words has a probability p_j to be selected in the dictionary, and upon selection, the distribution of the $A_{\cdot,j}$ is Laplacian: this way, we encourage the selection of only a few atoms, and a sparse representation of each word embeddings in terms of the selected dictionary atoms.

To find a solution to this optimization problem, we propose the following approaches:

1. Compute the PCA for each word embedding matrix at time t. This gives us a "condensed" representation of H orthogonal directions required to explain the variance of the words, and we only retain the Principal Components that account for 95% of this variance. We denote as Y their representation in the original basis. Instead of reconstructing the original embedding matrix, we could now try to reconstruct the Principal Components.
2. Solving the previous optimization problem for each point in the dataset, enforcing each diagonal coefficient A_{ii} to be 0 .

In any of these approaches, we would have to solve a penalized Lasso regression problem sequentially over the different decades, by computing for each word vector:

$$\forall w \in [1, |V|], \quad \beta_{w,\cdot} = \arg\min_\beta (W_w^{(t)} - W^{(t)}\beta)^2 + \lambda \sum_{j=1}^{|V|} \text{pen}_j |\beta_j|$$

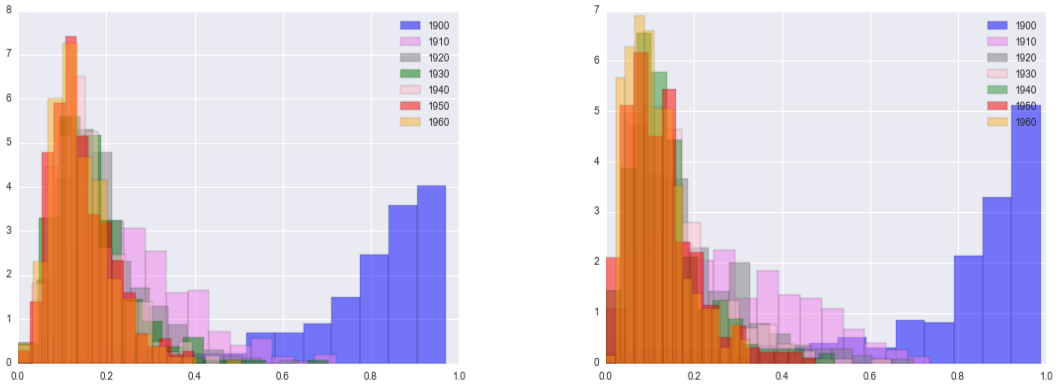
We can then define a "Local Linear Lasso Similarity " between words by comparing their representation -i.e, their recovered Lasso Regression Coefficient β_w from the regression against the basis of selected embeddings: indeed, our dictionary is constructed to express each word embedding as a linear combination of its atoms, thus capturing local linear similarities. These characteristics should constitute a more "inherent" property of the embeddings, and should transfer from decade to decade without suffering from as much noise variability

Algorithm 1: **Greedy Lasso with Exploration**

INPUT: objective matrix Y–i.e, the matrix to reconstruct, embeddings W,
Activation set $\mathcal{S} = \emptyset$,
Dirichlet prior distribution over the activated set $P = \frac{2}{|\mathcal{V}|}[1, \cdots 1]$ (uniform prior over the active set)
for decade=1900 **to** 1990 **do**

  for w in $\mathcal{V}$ **do**
    $\hat{P}_{dw+1} \leftarrow$ Dirichlet(P_{dw}) : randomly sample a new $\hat{P}$ –(to encourage exploration),
    pen $\leftarrow \frac{1}{1+\hat{P}}$ (update of the Lasso penalties),
    $\hat{\beta}_w = \arg\min_\beta ||w - W^{(t)}\beta||_2^2 + \lambda \sum_{j=1}^{|V|} \text{pen}_j |\beta_j|$,
    $P_{dw+1} \leftarrow P_{dw+1} + 1_{\hat{\beta} \neq 0}$,
  end for
end for$S \leftarrow 1_{P > \eta}$ where η is a pre-specified threshold

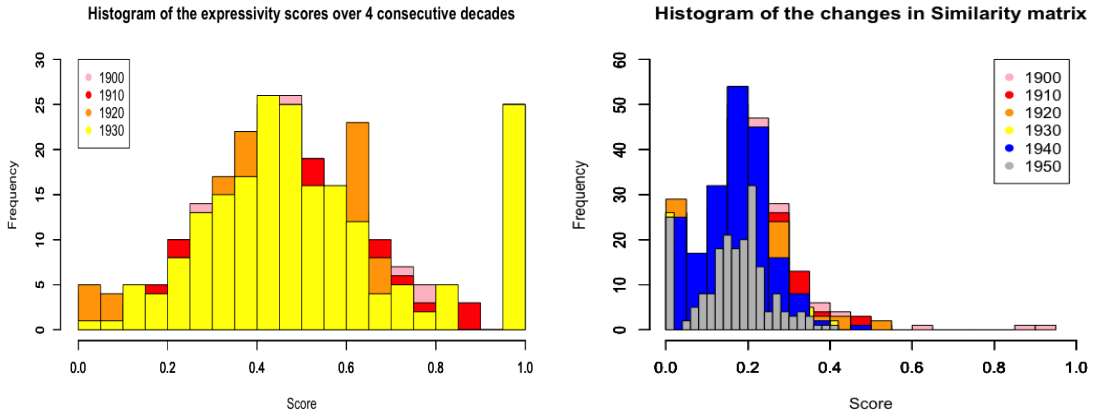
The results for this approach are displayed in Figure 4.



(a) Sequential Computation of the coefficients

(b) Activation set selected among among 4 different dictionaries

Figure 4: Lasso Reconstruction Scores (as defined by $s_w = \frac{||w\hat{\beta}_w||^2}{||w||^2}$ for the reconstruction level of a dictionary selected via Greedy Lasso for the COHA dataset



(a) Evolution of Lasso-Similarity Score distribution for the first 200 words

(b) Distribution of the changes in the Lasso Similarity

Figure 5: Lasso Expressivity Scores for the reconstruction of a dictionary selected via Greedy Lasso (scores are shown for the 200 first most frequent words of the Google 5-grams dataset)

Looking at these plots, it would seem that we are able to draw interesting first conclusions on the validity of our approach. For the Google 5-gram dataset, the selected dictionary seems to be able to reconstruct well the embeddings (given the fact that our previous analysis has highlighted the presence of really high noise levels). Moreover, this scoring seems to be consistent over time, and, as shown by figure 5b, the Lasso coefficients exhibit less variability than the innovation matrix previously studied. As such, it is possible to define a "sounder" null distribution to detect the semantic changes. For instance, in 5b, we see that a few words ("one", "kind", and "work") have exhibited important changes during the 1900 decade, thus pushing us to investigate these particular words.

These plots also underline the sensibility of the "Local Linear Neighbors" approach to correct initialization and the size of the dataset: as shown in Figure 4a, for the COHA dataset, the recovered dictionary (of size hovering 600) has high predictive power for the first dataset that it tries to explain (the 1900s decade in that case), but that expressive power does not translate at all to any of the other decades. 4b shows the reconstruction levels that are obtained when initializing with a different decade. This is an interesting result in its own right, at it seems to exclude the possibility of finding locally linear subspaces that could be robust over time: despite initializing the embeddings with the representation from the previous decade, the embeddings are torn apart at every new training phase, and locally linear relationships are shattered by the training process.

5 A nonparametric approach to test changes in word meanings : pooled drift statistic

As described by William Hamilton et al. in [1] and in [2], W_{t-1} was used as the word embeddings initialization for training over decade t when training W_t.

Inspired by this, if every word meaning remains unchanged(the "null hypothesis") from decade $(t-1)$ to decade t, one could expect that word embeddings trained by using *pooled* sentences in decades $(t-1)$ and t (with the same initialization W_{t-1}) are similar to W_t ; let's call this matrix of word embeddings $\tilde{W}_t$. Then we can use the *pooled drift statistic*, thedifferencebetweenW_t("original") and $\tilde{W}_t$("mixed") to test the null hypothesis. Note that rotations introduced above are no longer necessary for this approach, because we use the same initialization.

Possible benefits of this approach, compared to [1] and [2] are :

- Capturing per-decade changes, which is shorter than per-century results ;
- Reflecting differences in both lengths and angles of word vectors to conduct more general analysis than cosine-similarity based methods(although local neighborhood measure in [2] utilized lengths informationin some extent).

Having no backup theory on null distribution, we try to use the mean squared difference("MSD") between original and mixed word vectors as a measure of word meaning change. For convenience, we choose top 10,000 frequent words to analyze, which is large enough. Histograms of MSD are given as Figure 6.

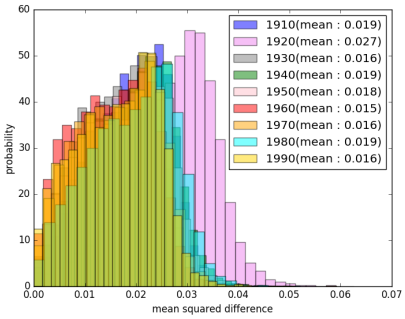


Figure 6: Histograms of MSD

It is observed that except for 1920s, histograms are overlapped pretty well and means are similar to each other, suggesting that there is a (right-skewed) null distribution of MSD under the assumption that there are only a few words changed in their meanings ; for 1920s, the disagreement seems mainly due to the imbalanced data set sizes between 1900s-1910s(∼3k) and 1920s-1990s(∼10k).

Also, clear outliers possibly corresponding to words with changed meanings are found, hence we list the 10 most changed words in terms of MSD in Table 1.

Decade	10 most changed words from (Decade - 10) to Decade
1910s	crow, th, peter, dick, lucy, **uncle**, duke, princess, temple, jack
1920s	dick, sue, p, hardy, sands, al, lucy, **interior**, imperial, ha
1930s	sue, title, dick, sheriff, ah, duke, peter, hugh, **zero**, hunt
1940s	hunt, tower, **pearl**, dick, **hammer**, fuller, graham, lucy, sterling, ah
1950s	peter, hunt, forest, tower, temple, admiral, uncle, lucy, aunt, **guest**
1960s	**bond**, admiral, sergeant, shooting, peter, camera, speaks, lucy, sing, pages
1970s	**bond**, lance, bloom, lieutenant, charlotte, **salt**, colonel, admiral, grandfather, peter
1980s	editor, m, re, lance, **bush**, p, grandfather, haven, harry, horn
1990s	horn, ray, l, chalk, m, colonel, **x**, dean, d, q

Table 1: Most changed words in terms of MSD

Although results seem quite random, we can observe some interesting things :

- There are 13 words that appear consecutively in this list : lance, grandfather, horn, peter, sue, dick, ah, hunt, m, lucy, admiral, tower, bond. This suggests that word vectors can change over several decades i.e can be pretty unstable.

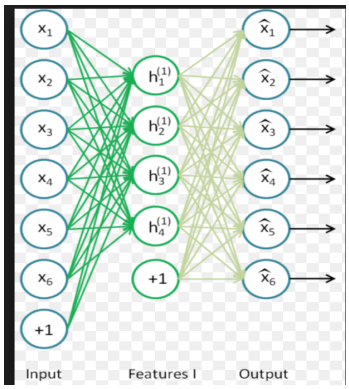
- Seemingly meaningless words(th, p, al, ha, ah, m, re, p, l, m, x, d, q) are selected over the century, perhaps due to lots of appearances rather than their semantic changes. Over 1950s - 1980s, selections of war-related words(admiral, sergeant, shooting, lieutenant, colonel) can be also explained in the same way, considering World War II, Vietnam War and Cold War. In fact, we can check such phenomena from small experiments along with several repeated (same) sentences ; due to the randomness in skip-gram negative sampling, even training on the same sentences results in word embedding changes.
- Despite many false discoveries, we can give plausible explanations for 10 bolded words' meaning changes from `http://www.etymonline.com/`), as follows :

Word	**Historical changes**	**Year**
uncle	*say uncle* ("admit defeat")	1909
zero	blood type *zero* $\rightarrow$ O	1926
interior	*interior design*	1927
pearl	*Pearl Harbor*	1942
hammer	"to defeat heavily"	1948
guest	*be my guest* ("go right ahead")	1955
salt	*Strategic Arms Limitation Talks*	1968
bond	of persons (originally of things)	1969
bush	George H. W. Bush	1981
x	*generation x*	1991

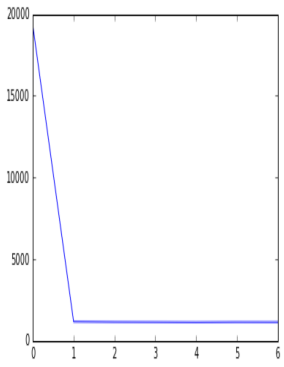
Table 2: Semantic changes parallel to Table 1

6 A neural network approach

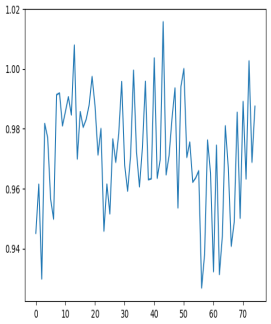
The previous sections have highlighted the difficulty of finding stability in word embedding representations. A next natural approach would thus be to take a step back and try to define classes – or clusters– of vectors, that could potentially exhibit less variability In particular, work by Xavier Glorot and al. on Sparse Autoencoder seems a natural and excellent candidate to be applied to this purpose: using the embeddings as input, the goal would be to find a sparse low dimensional hidden layer such that the reverse mapping (from the hidden layer to the output, which is in this case, the input embedding itself) is the most faithful. In that particular case, the hidden layer could be interpreted as a distribution over latent clusters, and thus, be used as a way to define larger neighborhoods.



(a) autoencoder



(b) Loss of the autoencoder over 6 different decades



(c) Test reconstruction error:

Figure 7: histograms for the distribution of the orthogonal displacement ratio $\frac{||\Pi^{\perp W_{t+1,i}}[W_{t+1,i}-W_{t,i}]||_2}{||W_{t,i}||_2}$ for each word vector i

This approach is more myopic as it uses the Deep Learning framework to train the classes, without training bias or variability considerations, and, as such, does not provide an "understanding" of the word embeddings themselves. We have thus began implementing this approach, using a first sigmoid layer and a second reLu layer to enforce sparsity, as prescribed by [8]. The results that we have yet so far have not been satisfactory, but this might be due to the very noisy nature of the COHA dataset that we have used for our experiments, and the nature of this Deep Learning framework, that requires considerable amounts of input to produce faithful results. The next step would thus be to use this approach on a larger dataset, such as the Google-n-grams.

7 Conclusion and Potential Future Directions

This project has proposed to push the analysis of the attractivity of word embeddings for studying language dynamics on a mesoscopic scale, considering information contained in the integrality of the embedding matrix to express each word as a sparse linear combination of a few others. This approach has shown interesting promises in the case where the dataset is of extremely large size (such as the Google n-grams data). However, it has not yet been shown to work on smaller dataset, thus narrowing the scope and interpretable power of this approach.

However, it could be interesting to tackle the problem on mesoscopic scale, using for instance an LDA type of approach, where we could define for each word embedding, a distribution over clusters, and assess their similarities an d stability by working on these distributions– the intuition being that real clusters should exhibit less variability than on the microscopic level.

Also, to improve the power of pooled drift statistic, one can subsample from pooled sentences to balance training sets for W_t and $\tilde{W}_t$, and/or find a better measure of changes using word counts just like GloVe) to control false discoveries.

References

[1] William L. Hamilton, Jure Leskovec, Dan Jurafsky *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change*. Acl 2016, p1489-1501

[2] William L. Hamilton, Jure Leskovec, Dan Jurafsky *Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change*. EMNLP 2016.

[3] Udo Hahn, Johannes Hellrich *Bad Company Neighborhoods in Neural Embedding Spaces Considered Harmful*, in Proceedings of the 26th International Conference on Computational Linguistics (COLING-16), p2785-2796

[4] William L. Hamilton, Kevin Clark, Jure Leskovec, Daniel Jurafsky *Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora*,in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16),p 595-605

[5] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, Steven Skiena, *Statistically significant detection of linguistic change*, in Proceedings of the 24th international conference on World Wide Web (WWW '15), 2015, p625-635

[6] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, Steven Skiena, *Freshman or Fresher? Quantifying the Geographic Variation of Internet Language*, in arXiv.org, 2015, p615-618

[7] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, Slav Petrov, *Temporal Analysis of Language through Neural Language Models*, in Proceedings of the Workshop on Language Technologies and Computational Social Science@ACL 2014, p 6165.

[8] Xavier Glorot, Antoine Bordes, Yoshua Bengio *Deep Sparse Rectifier Neural Networks.*, Aistats, 15,106,p275

Work Division for the project

- Data gathering, cleaning: Jeha and Claire
- Innovation based approach (Section 3): Jeha and Claire
- Stable Basis (Section 4): Claire
- alternative approach (Section 5): Jeha