Backprop to the Future: A Neural Network Approach to Linguistic Change over Time

Eun Seo Jo **Dai Shen** **Michael Xing**
Stanford University
{eunseo, dai2, huizex}@stanford.edu

Abstract

We introduce two RNN-based approaches to gauge moments of language change over time and apply them to a set of historical diplomatic documents for interpretative results. Our first model is an RNN acceptor model that predicts the year of a given document. We show that the RNN implementation outperforms a double-layer feed-forward at this task and use the softmax weights to demonstrate topical and semantic change over historical eras. Our second model is a standard RNN language model that is trained with feeds of temporally sequenced documents. We capture the changing perplexity of the language model over time to identify moments of dramatic shifts in language. The two models in the end produce similar results – the 1910s and 1940s in American diplomacy proved to be notable moments in linguistic change.

1 Introduction

Linguistic (topical and semantic) change over time can reflect critical shifts in historic context whether they be in discourse, ideology, or even access to communications technology. This project examines how these shifts are embedded in a set of declassified diplomatic documents. We introduce two RNN-based approaches to identify moments of linguistic change and continuity. The first model is a straight-forward RNN acceptor model that predicts the year bucket of a given document. This training generates weights per year bucket that we use to find trends in language. The second approach trains a language model over temporally sequential sets of training data. We observe the changes in perplexity over these eras and find that perplexity peaks at moments of dramatic language shifts. While our models successfully identify these moments, as with many deep learning models they fail to capture the exact features that evolved over time. We hope that our work will contribute to new discussions concerning linguistic change over time.

2 Related Work

To the best of our knowledge, there is relatively sparse existing literature on capturing linguistic change over time, especially using neural networks. One recent work introduces new methods of comparing language usage across time spans. Hamilton et al. analyze linguistic change with word vector embeddings generated from PPMI, SVD and word2vec. Using these results, they propose several laws of linguistic evolution[1]. However, Hamilton et al. do not propose any method of capturing general trends of language in a given corpus. We extend his work by vectorizing language per decade with neural network training. We also engage with existing literature on stylometry and authorship attribution. Ramnial et al. emphasize the strength of stylometric features in identifying authorship and in plagiarism detection [2]. Tweedie et al. show its neural network application [3]. Our work attempts to capture stylistic change using neural nets by masking out all non-stop words and show that the hidden layers are able to capture stylistic differences over the years rather successfully.

3 Data

We collected 238,097 historical diplomatic documents that span from 1860 to 1983. We process the data differently for the two models.

3.1 Data Preprocessing for RNN Classification Model

We performed stratified sampling to overcome the significant variance in frequency of documents over the year buckets. We also capped the number of documents at 5000 to smooth out the disproportionate representation of certain years. Figure 1 shows the distribution of the documents across years, the year bucket delineation, and our capping line.

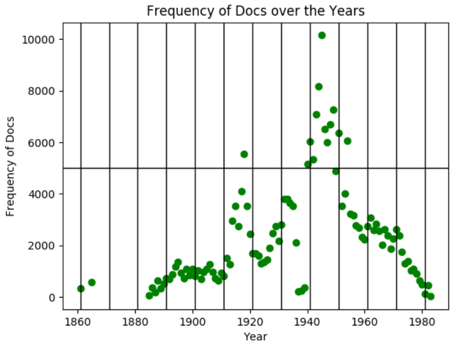

Figure 1: Frequency of Documents over the Years. Documents were capped at 5000 per year, for smoothing.

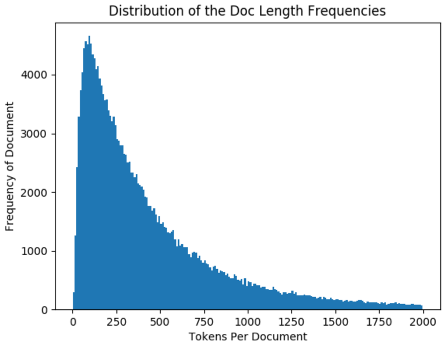

Figure 2: Histogram of the documents by token count. We chose to truncate all documents at length 600 and pad those shorter than.

We keep punctuation as individual tokens to preserve sentence structure meaning and style. We also truncate long documents to 600 words and pad shorter documents to 600 words. Since most of our documents are less than 600 words long 2, we do not consider this to be a significant loss of data. For documents that are padded to 600 words, we keep track of their actual length for dynamic RNN.

In one iteration of the RNN model, we also mask out all tokens except punctuation and stop words to capture pure semantic change.

3.2 Data Preprocessing for RNN Language Model

For the RNN Language Model we divide the corpus into buckets of half decades from 1890 and 1980. The smaller year ranges allow us to examine more granular level change. We divide the documents into 50 token length chunks.

3.3 GloVe Word Vectors

For both models, we use GloVe vectors with 400,000 vocab size of 100 dimensions. Table 1 shows that 2.7% of all tokens from the dataset were unidentified.

Missing Count	Total Count	Percentage
3255491	118855819	2.7

Table 1: Percentage of All Tokens Missing from Glove Vectors (Vocab Size 400K) and Masked Out

4 Experiments

4.1 Document Classification

We first build a neural network model to predict the year of a given document.

4.1.1 Baseline: Bag of Words and feed-forward model

As our baseline, we implement a two level feed forward neural network. Each layer applies RELU activation. The input is a bag of words (BOW) averaged vector. We used dropout to prevent over fitting. Figure 3 shows the feed forward network. This model extracts word features but none of their sequential information.

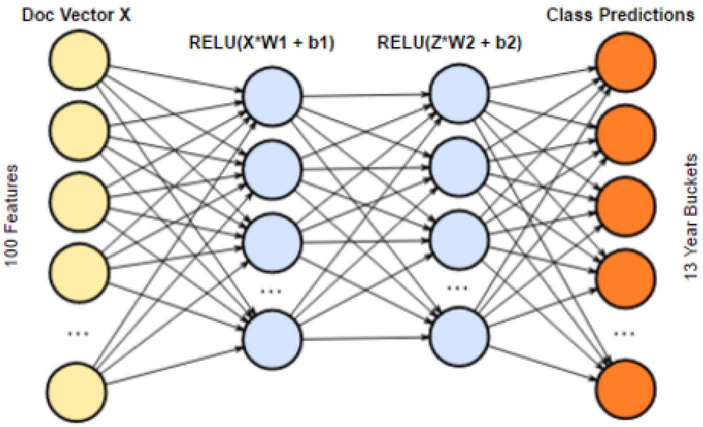

Figure 3: Baseline: A bag-of-words two-layer feedforward network with RELU activation.

4.1.2 Recurrent classification model with GRU cells

We then trained and tested on an acceptor model to classify documents into their corresponding year buckets. The acceptor model consists of a single layer of GRU cells and prediction is made from the output of the last cell. Figure 4 shows the general layout of the model. Each GRU cell has the following composition:

$$\begin{aligned}
z^t &= \sigma(x^t U_z + h^{(t-1)} W_z + b_z) \\
r^t &= \sigma(x^t U_r + h^{(t-1)} W_r + b_r) \\
o^t &= \tanh(x^t U_o + (r^t \otimes h^{(t-1)}) W_o + b_o) \\
h^t &= z_t \otimes h^{(t-1)} + (1 - z^t) \otimes o^t
\end{aligned}$$

Following the last time-step we make the year prediction as follows:

$$\text{pred} = \text{softmax}(h^{(last)} U + b_2)$$

where *pred* is the probability distribution of the document being in each decade. Similar to the baseline, we applied dropout to prevent overfitting. We expect GRU to have less problem with vanishing gradient and learn longer sequence data. To prevent exploding gradient, we integrate gradient clipping. We expect this model to better capture sentence structures.

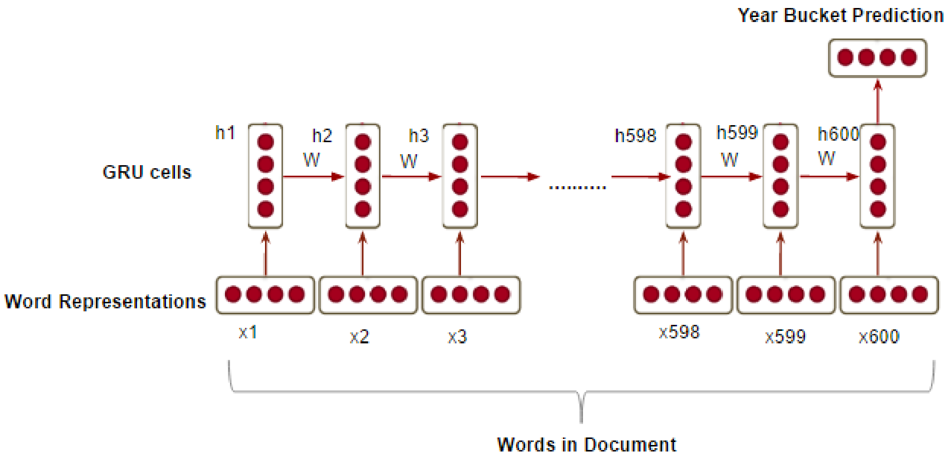

Figure 4: Document Classification Model Using GRU cells

4.2 Language Model

Our second model is a language model that trains by predicting the next word given the current and all previous words. We use the same RNN schematics as the previous model but make a prediction at each time step. The output of each GRU cell is a probability distribution over 400,000 classes, our vocabulary size. The GRU composition changes to the following:

$$
\begin{aligned}
z^t &= \sigma(x^t U_z + h^{(t-1)} W_z + b_z) \\
r^t &= \sigma(x^t U_r + h^{(t-1)} W_r + b_r) \\
o^t &= \tanh(x^t U_o + (r^t \otimes h^{(t-1)}) W_o + b_o) \\
h^t &= z_t \otimes h^{(t-1)} + (1 - z^t) \otimes o^t \\
\text{pred} &= \text{softmax}(h^{(t)} U + b_2)
\end{aligned}
$$

We train our language model data in chronological order so that it is able to learn in sequence. We divided them into half decade buckets. We make the first 85% of each decade as the training data and the last 15% of each decade as the development data (see figure 5). We train(85%)-dev(15%) in sequence over the eras.

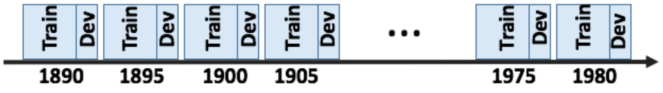

Figure 5: Diagram of Language Model Document Feeder: The 5 year corpora are fed into the model in sequence, starting from 1890.

5 Results

5.1 RNN Classification

With 30 epochs of training the RNN classification model, we were able to achieve f1 of over 80% as shown in figure 8. Figure 6 and 7 show the loss and accuracy respectively of the RNN compared to the baseline model. The RNN outperforms the bag of words baseline consistently throughout the epochs and reaches the peak performance after fewer epochs. This demonstrates that sentence structure played a significant role in predicting the year of a given document. We can also see the results of the RNN trained with the stop-words only corpus. The accuracy performs surprisingly well, surpassing 50%, indicating that stylistic changes were significant enough to be captured.

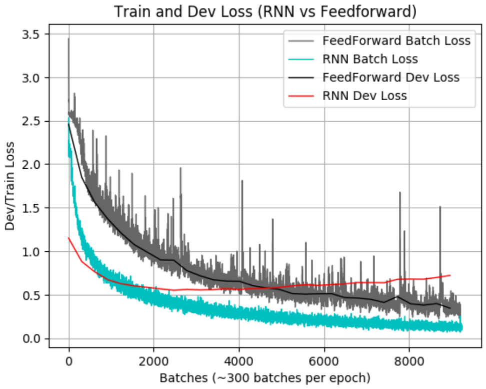

Figure 6: Comparison of Dev set loss over 30 epochs for RNN and Feedforward models

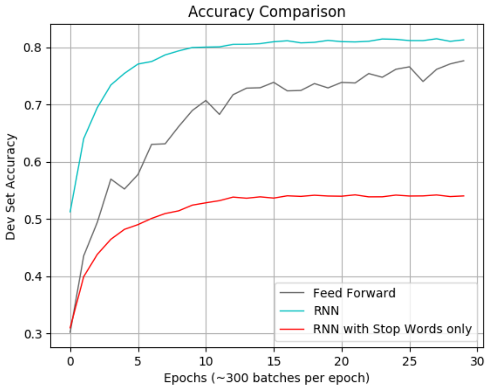

Figure 7: Comparison of dev set accuracy over 30 epochs for RNN, RNN with only stop words, and Feedforward models

Figure 8: Test F1 Scores of RNN, Feedforward

F1 Scores of Models	
Model	F1 Score
Feedforward RNN RNN-stop words	0.719617723243 0.802763569257 0.501838766611

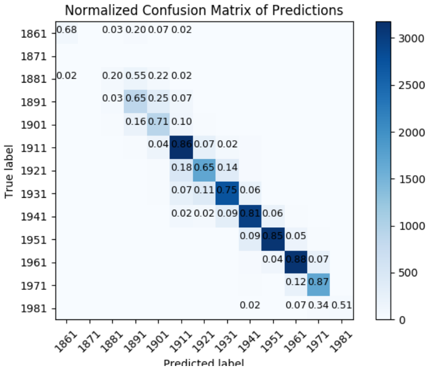

Figure 9: Confusion matrix of the feedforward baseline model's prediction of year classes after 30 epochs. The first class corresponds to the decade [1861,1870] and so on.

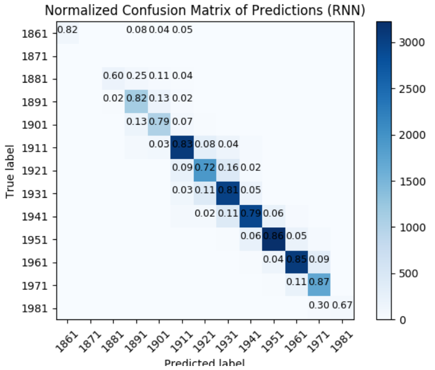

Figure 10: Confusion matrix of the RNN models' prediction of year classes after 30 epochs

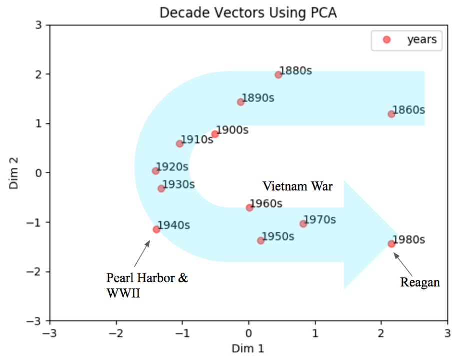

Figure 11: PCA visualization of the decade classes from the softmax weights. Dimensions reduced from 300 to 2. Note: 1870 is not graphed because the class bucket was missing data.

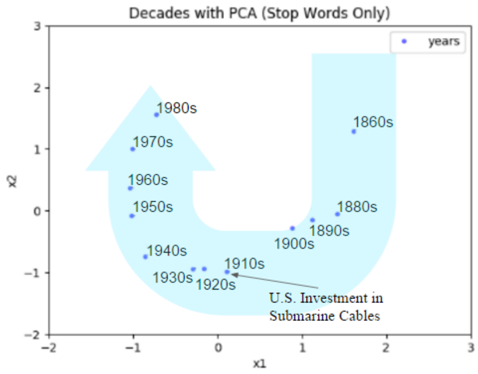

Figure 12: PCA visualization of decade classes from the softmax weights trained with non stop words and punctuation . Dimensions reduced from 300 to 2. Note: 1870 is not graphed because the class bucket was missing data.

To visualize the historical trends in language, we took the softmax weights, the U matrix of dimension [class $\times$ hidden size] in:

$$\text{pred} = \text{softmax}(h^{(t)}U + b_2)$$

and reduced the dimensions using PCA and charted the decade vectors in 11 and with only stop words in 12. The visualizations show overall trends in language through the century and a half. Figure 11 shows how general language, including ideologies and issues of concern, changed over time. Figure 12 maps the changes in purely stop word usages over time, an attempt to focus on stylistic change. It is notable that while 11 shows greater variance in changes from decade to decade, 12 appears more consistent, results which verify our expectations about topics and style. Given that the dataset is a collection of diplomatic documents, we labeled several notable historical events for interpretation. The distance between 1940 and 1950 in 11 is worth highlighting as post-WWII and the commencement of international organizations such as the UN(1945), WHO(1948), and IMF(1945) changed American diplomatic language to incorporate more vocabulary concerning global procedures. In figure 12 the jump in style from 1900 to 1910 aligns with historical literature on diplomatic language. The consensus is that eyeing the geopolitical and strategic potential of expanding the reach of telecommunications coverage, the U.S. State Department made significant investments in submarine cables during the first world war, which altered the constraints of telegraphic expression [4].

Finally, for hyperparameter tuning we ran 30 epochs on hidden sizes of 200,300, and 600 but observe little difference in performance. Figures 13 and 14 show our results.

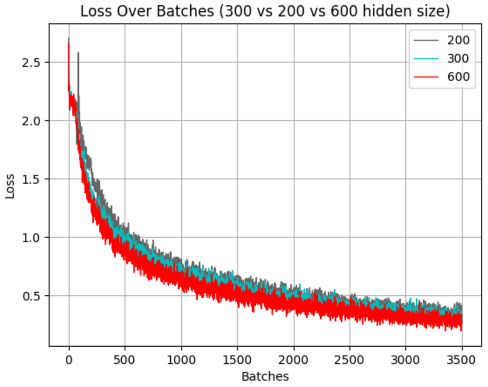

Figure 13: Comparison of batch loss of the RNN model at hidden size: 200,300, and 600

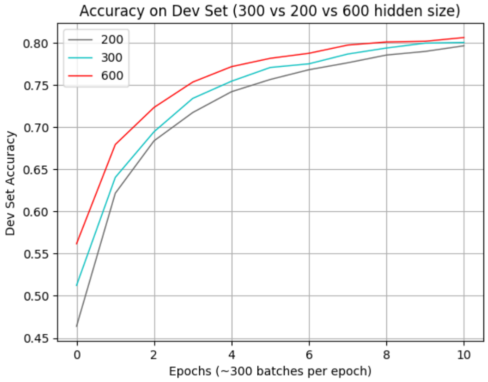

Figure 14: Comparison of accuracies at different levels of hidden size: 200,300,600

5.2 RNN Language Model

The results of the RNN Language Model corroborate our results from the RNN classification model. Figure 16 charts the changing dev set perplexity over the years at five-year intervals. If language is consistent enough throughout the years we would expect the perplexity to fall consistently as we feed more data. However, the results show perplexity spiking at certain historical moments, in 1915 and 1945, indicating the change in language was dramatic enough to confuse the model trained on the previous years' language. These two moments of spike align with the our results from above and can be attributable to the same historical events.

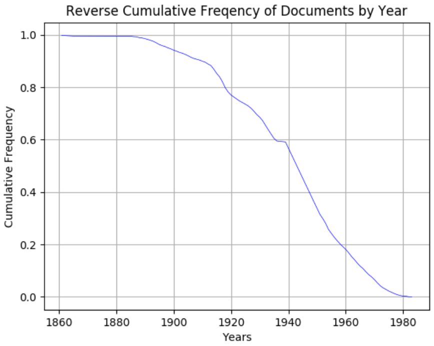

Figure 15: Reverse cumulative frequency of documents by year showing most of the documents in the dataset are from the mid-twentieth century.

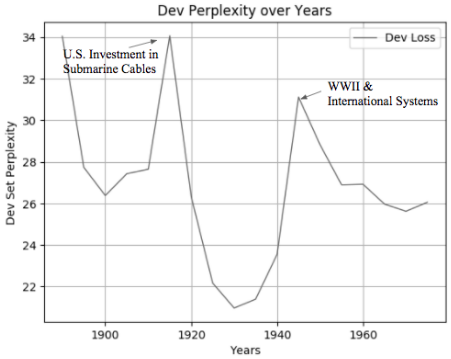

Figure 16: Dev set perplexity over the five-year corpora. The perplexity peaks at two places in the graph, 1915 and 1945.

Figure 17: Example Excerpts of Documents from Dataset

Document Example Excerpts	
Date	Excerpt
November 26, 1855	*sir , in my preceding despatch of to-day's date. i have replied only to the latter portion of mr. olney's despatch of the 20th july last, which treats of the application of the monroe doctrine to the question of the boundary dispute between venezuela and the colony of british guiana. but it seems desirable, in order to remove some evident misapprehensions as to the main features of the question, that the statement of it contained in the earlier portion of mr. olney's despatch should not be left without reply.*
May 29, 1953	*walter hallstein called on me this afternoon to discuss edc treaty. After conversation on this subject he told me that upon his return to germany tonight he and chancellor would deliberate upon following topic that had been under consideration by them and, if they decide go ahead, will be taking it up with you almost immediately. i thought it might be useful tell you of its possible reference to you although it may not take the form which hallstein now contemplates.*

Table 2: The two documents chosen above evince change in linguistic style.

6 Conclusion

The two RNN-based approaches ultimately produce similar results, showing that the 1910s and 1940s were moments of significant linguistic change in American diplomacy. Our experiments show that RNN models can be useful tools for measuring linguistic and topical change over time. Finally, there are still many avenues to expand on the task of linguistic evolution. One direction would be to engineer creative methods of interpreting the features of linguistic change.

References

[1] William L. Hamilton & Jure Leskovec Dan Jurafsky *ACL 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.*

[2] Ramnial& Panchoo *Authorship Attribution Using Stylometry and Machine Learning Techniques.*

[3] Tweedie& Singh& Homes *Neural Network Applications in Stylometry: The "Federalist Papers"*

[4] Nickles, David Paull *Under the Wire: How the Telegraph Changed Diplomacy* Harvard University Press, 2003.