Analyzing Syntactic Change in American English Using a Neural Network Part-of-Speech Tagger

Anonymous ACL submission

Abstract

We train a diachronic long short-term memory (LSTM) part-of-speech tagger on a large corpus of American English from the 19th, 20th, and 21st centuries. In addition to evaluating the overall performance of the model, we analyze whether the embeddings are able to learn the temporal structure between years, and the extent to which the model can be used to predict the year of composition of a novel sentence. Our network achieves a validation accuracy of 93.5%. Additionally, our analysis of the learned year embeddings revealed a strong linear correlation between the principal component of the year embeddings and time.

1 Introduction

We define a diachronic language task as a standard computational linguistic task where the input includes not just text, but also information about when the text was written. In particular, diachronic part-of-speech tagging is the task of finding a sequence of part-of-speech tags for a sequence of words dated to a specific year. Thus, a well performing diachronic part-of-speech tagger trained on American English must learn some representation of the evolution of American English during the 19th, 20th and early 21st centuries. Our goal is to determine how well such a model can perform and whether the model can be used to analyze syntactic changes in American English during the 19th, 20th and 21st centuries.

Our method approaches this problem using neural networks, which have seen considerable success in a diverse array of natural language processing tasks over the last few years. Prior work using deep learning methods to analyze diachronic language change has focused more on lexical change than syntactic development. Through statistical analysis of several types of word embeddings, William L. Hamilton, Jure Leskovec, and Dan Jurafsky (2016) found that there is an inverse power relationship between word frequency and the rate of semantic change, and that the rates of semantic change for words with multiple meanings are higher. Eun Seo Jo and Xing (2017) trained different language models on a corpus of American diplomatic documents and found that the 1910s and the 1940s marked sudden changes in the language of American diplomacy. Niyogi and Berwick (1995) attempted to develop a mathematical model of French syntactic change and found evidence for both gradual and sudden syntactic innovations.

Building on this prior research, we will use the Corpus of Historical American English (COHA) (Davies, 2010-), an LSTM part-of-speech tagger, and dimensionality reduction techniques to investigate syntactic change in American English during the 19th through 21st centuries. Our project takes the part-of-speech tagging task as a proxy for diachronic syntax modeling and has three main goals:

1. Achieve high accuracy on the task of diachronic part-of-speech tagging.

2. Determine whether our model can be used to predict the year of composition of novel sentences.

3. Assess whether a temporal progression is encoded in the network's learned year embeddings, and if so, whether this temporal progression is gradual or spiked around specific years.

2 Related Work

William L. Hamilton, Jure Leskovec, and Dan Jurafsky (2016) evaluated changes in word meaning over time, suggesting many rules that govern semantic change. In each time period, they created singular-value decomposition (SVD) embeddings, positive pointwise mutual information (PPMI) embeddings, and skip-gram with negative sample (SGNS) embeddings, and built models with six data sets from the COHA corpus and Google n-grams. William L. Hamilton, Jure Leskovec, and Dan Jurafsky (2016) measured synchronic accuracy, or how well the different embeddings captured similarity between words in different time periods. They also measured diachronic validity, or how well the embeddings represented shifts in word meaning over time. The research indicated that there is a negative power relationship between a words frequency in the documents and the words rate of semantic change, and that when word frequency is not taken into account, the rate of semantic change for words with multiple meanings is higher.

The language models trained by Eun Seo Jo and Xing (2017) suggested that the 1910s and 1940s were periods of sudden linguistic change in American diplomatic documents. The models were trained on American diplomatic documents from 1860 to 1983. Eun Seo Jo and Xing (2017) divided the documents into half decade groups and trained separate GRU language models for each group, using an 85% train data and 15% test data split. The sudden linguistic changes during the 1910s and 1940s were illustrated by peaks in the perplexities of these language models.

Niyogi and Berwick (1995) attempted to build a mathematical model of syntactic change with dynamics motivated by language contact and language acquisition. They found that their model predicted both gradual and sudden changes to a parameterized grammar depending on the properties of the languages in contact. In particular, they used their simulation to study how V2 was gained and lost throughout the history of the French language. For several toy languages, their model found that contact between +V2 languages and -V2 languages would lead to a gradual adoption of V2 syntax by the population. However, for a +V2 language with VOS order and -V2 language with SVO order, the model predicted that the adoption of V2 would be very sudden (occurring in < 4 generations).

Whereas William L. Hamilton, Jure Leskovec, and Dan Jurafsky (2016) and Eun Seo Jo and Xing (2017) were primarily concerned with lexical change, we hope to focus on syntactic change by choosing the task of part-of-speech tagging. We will use the same COHA corpus as William L. Hamilton, Jure Leskovec, and Dan Jurafsky (2016), and a similar probabilistic modeling approach to Eun Seo Jo and Xing (2017), except that word prediction will be replaced by part-of-speech prediction. One other key difference is that we will train one diachronic part-of-speech model, whereas Eun Seo Jo and Xing (2017) trained several different language models for different time periods. This network architecture will concatenated the word vector representation at each time step with the document's year embedding. An interesting feature of our approach is that the activation values for sentences from different time frames will have a shared representation. In principle, learning from sentences in any year can inform predictions about sentences in neighboring years. We will be able to analyze the year embedding space to see if the embeddings encode a temporal progression. By analyzing the nature of this progression, we may be able to probe whether the syntactic changes detected by our model are gradual or sudden in a similar vein to the questions asked by Niyogi and Berwick (1995).

3 Data

3.1 Corpus of Historical American English

We used the Corpus of Historical American English for our work. This corpus is composed of documents dating from 1810 to 2009 and contains a total of over 400 million words. The genre mix of the texts is balanced in each decade, and includes fiction works, academic papers, newspapers, and popular magazines. Because of computational constraints, we selected 300 documents from each decade, except from the 1810s, for which only 63 documents were available. We also cut off all sentences at a maximum length of 32 words. Across all decades, we had a total of 2,054,902 sentences. Because of memory constraints, we used a random sample of 750,000 of these sentences. We put 70% of these into a training set, 15% into a development set, and 15% into a test set.

Texts in COHA are annotated with word,

lemma, and part-of-speech information. The part-of-speech labels come in three levels of specificity, with the most specific level containing several thousand part-of-speech classes. We chose to use the least-specific label for our model, which still had 423 part-of-speech classes. Note that in the code, the number of classes is stated as 424, but one index is unused.

3.2 Word Embeddings

Our model utilized pre-trained 300-dimensional Google News (Mikolov et al., 2013) word embeddings that were learned using a standard word2vec architecture. When no embedding existed for a word in the corpus, we assigned the word an embedding vector of random normally distributed numbers. We set the embeddings for unknown words in this way, so that the embeddings for different unknown words would be treated differently. Because of memory constraints, we stored the embeddings for all of the words in all the documents' vocabulary (given to us by a lexicon file) in an embedding matrix. In the input tensor, we encoded each sentence in the document with indices that would redirect us to the appropriate embeddings for each word. Due to computational limits, we were unable to load the entire embedding matrix into our network graph, so we only included embeddings for the 600,000 most common words in the vocabulary. The other words were replaced by a special symbol `UNK`.

4 Network Architecture

We used a single-layer LSTM model in which the input at time i was the concatenation of w_i, the embedding of word i, and t, the embedding of the document's year of composition. A diagram of this architecture can be seen in Figure 1. The word embeddings were loaded statically, and we assigned embeddings chosen uniformly at random to words whose embeddings were unknown. In contrast, the year embeddings were randomly initialized via Xavier initialization and learned dynamically by our network. Thus, we did not explicitly enforce that the year embeddings should encode any temporal structure between years.

We gave both the word embeddings and year embeddings a dimensionality of 300. We picked the size of our LSTM layer to be 512.

Due to the size of our training set and our limited computational resources, we ran our network for just 1 training epoch. Manual tweaking of learning rate and batch size revealed that the network's performance was not particularly sensitive to their values. Ultimately, we set our learning rate to 0.001 and our batch size to 100. We did not incorporate dropout or regularization into our model.

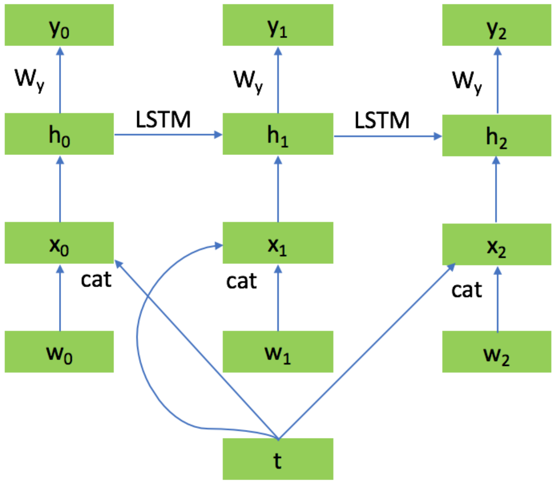

Figure 1: Network architecture. The input to the LSTM layer at each time step is the concatenation of each word with the document's year. The output at the corresponding time step is a predicted part-of-speech tag.

A public implementation of our part-of-speech tagger is linked in the references section (Stark and Merrill, 2018).

5 Results

5.1 Tagger Performance

Our network achieves 93.5% validation accuracy after training for 1 epoch. Given the specificity of the part-of-speech space ($m = 423$), we are pleased with this performance. The training loss curve can be seen in Figure 2.

5.2 Temporal Prediction

We randomly selected four sentences from the test data and relabeled their year with all possible values in the 1810 to 2009 range. We predicted the perplexity of the sentence at each possible year and fit a curve to these points using locally weighted scatterplot smoothing (LOWESS). In Figures 4 and 6, perplexity was low during the time period around the correct year. However, we did not observe this trend in Figures 3 and 5. Interestingly, the sentence in Figure 3 had V2 word

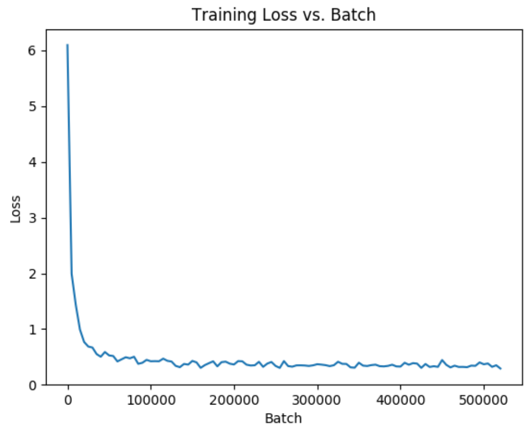

Figure 2: Network learning curve during the single training epoch.

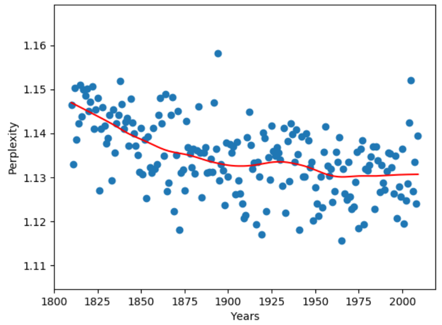

Figure 3: Perplexity over time for: no longer hold i the sublime communion of my youth , with them that have departed .
Sentence originally from 1822.

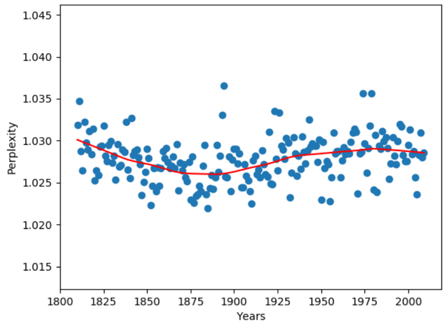

Figure 4: Perplexity over time for: meanwhile , though the squire was entirely unconscious of it , there was a sword hanging over his own head .
Sentence originally from 1864.

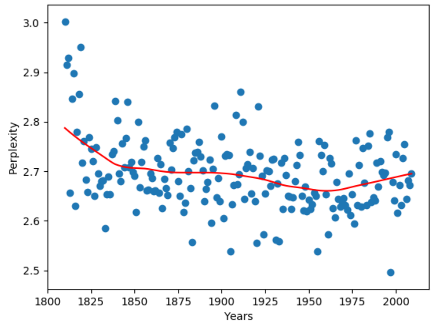

Figure 5: Perplexity over time for: nelson , UNK people and eightyone loads , are left behind , and their chances of being relieved are daily growing less and less .
Sentence originally from 1891.

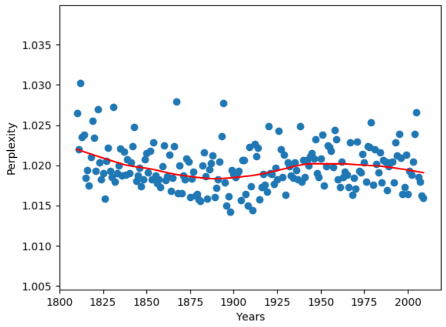

Figure 6: Perplexity over time for: she went to the bed and began smoothing the sheets deftly .
Sentence originally from 1902.

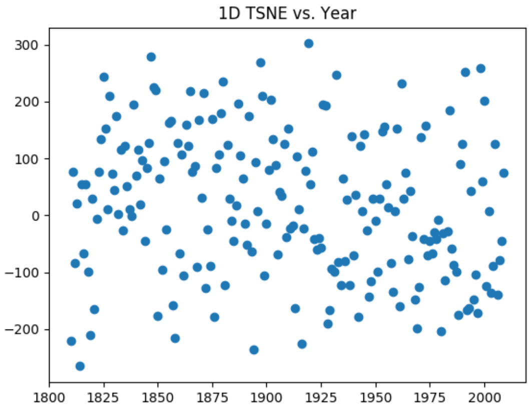

Figure 7: Learned year embeddings using a TSNE 1D reduction display no obvious temporal trend.

order, which is a qualitatively archaic syntactic construction, but our model gave lower perplexity scores for that sentence in more recent years relative to the actual year of composition (1822). The problematic sentence in Figure 5 had an UNK token, which may explain the inaccurately predicted perplexity curve. Overall, the inaccuracy of the perplexity curve as a prediction method is not surprising, since finding the precise year of composition for a single sentence is a difficult task.

We also noted that the perplexity curve is not very smooth, which suggests either that the learned trends in perplexity are not continuous from year to year, or that the variance of perplexity around a specific year is very high. Based on our findings in 5.3, it seems that the word embeddings are linearly structured with respect to time, so the second interpretation seems more likely.

5.3 Year Embeddings

To analyze whether our model learned to detect the temporal relationship between nearby years, we aimed to see whether the individual year embeddings learned temporal proximity. To do this, we reduced the year embeddings to one and two-dimensional space using both t-Distributed Stochastic Neighbor Embedding (t-SNE) and principal component analysis (PCA) dimensionality reduction. Using t-SNE is a more standard method of dimensionality reduction for neural network embeddings because it can incorporate a cosine distance metric. However, we also tried using PCA because PCA is constrained to reparameterize the data along principal components, and we

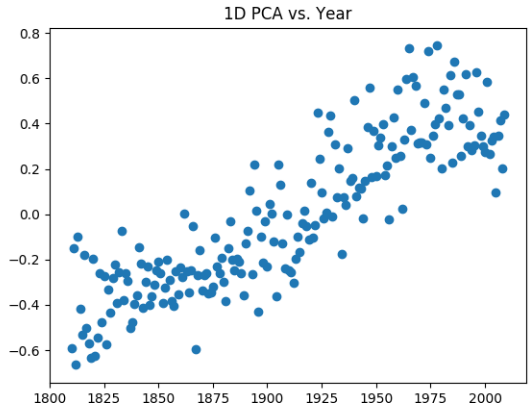

Figure 8: Learned year embeddings using a PCA 1D reduction display a clear temporal trend.

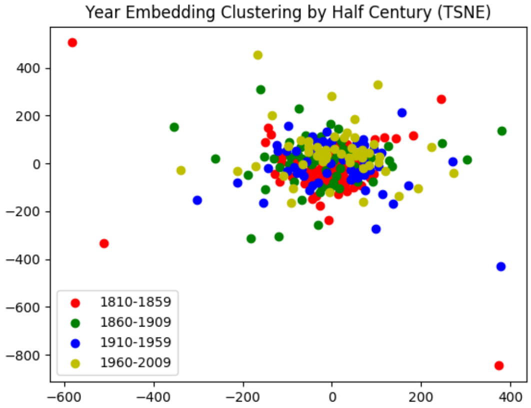

Figure 9: Learned year embeddings using TSNE 2D clustering display no obvious temporal clustering.

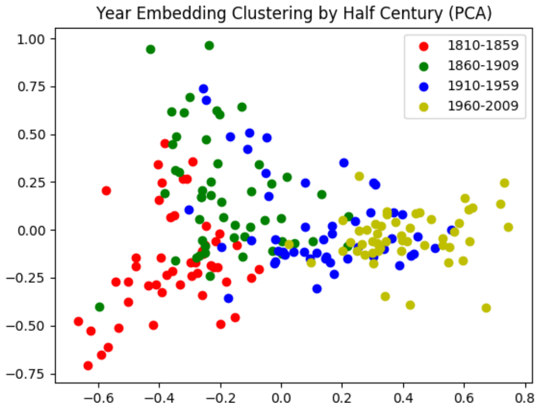

Figure 10: Learned year embeddings using PCA 2D clustering display clear temporal clustering.

know that temporal structure could be encoded by giving each vector a position along a temporal axis.

Plotting the year embeddings using 1D and 2D t-SNE dimensionality reduction did not reveal any intuitive ways in which the temporal ordering of the years was encoded in the year embeddings (see Figure 7 and Figure 9). However, when we ran the same analysis with PCA, a clear linear relationship (linear regression $R^2 = 0.76$) between the years and the principal component of the embedding vectors (see Figure 8) emerged. We also observed clustering of temporally proximal years in the 2D case (see Figure 10). The 1D result shows that the most significant component of the year representation learned by our network was encoding the relative position of each year within the chronological sequence. The seemingly linear relationship between year embedding and year in Figure 8 might be construed as evidence that our model is finding gradual and not sudden syntactic change in American English.

6 Conclusion

With a validation accuracy of 93.5%, we achieved our first goal of obtaining good performance on diachronic part-of-speech tagging. In our analysis of the perplexity curves for specific sentences, we found that in two out of four instances, the model assigns low perplexity to the labeled part-of-speech sequence during the time period around the correct year. Through our PCA analysis of the year embeddings, we found that the network learned to embed years according to a chronological sequence without any initial constraints to do so. Overall, our findings suggest that our model effectively incorporates some continuous notion of time into its modeling of the syntax of 19th, 20th and 21st century American English. The strong linear relationship between time and the year embedding may suggest that most of the syntactic change that our model learned was gradual as opposed to sudden.

7 Further Work

One way to improve our architecture might be to pick a representation for the year that enforces the temporal ordering between years (we saw that this was learned naturally by our model, but perhaps enforcing it would lead to better results). Some possible ways to do this would be representing the year as a continuous value between 0 and 1, or using a temperature-slider encoding where each bit i in the year vector represents whether the encoded year is greater than or equal to i. Enforcing the ordering of years in this way might yield smoother perplexity curves.

Another direction for further study would be experimenting with the number and size of the LSTM layers, as well as the size of the year embeddings.

Acknowledgments

We obtained access to COHA through our affiliation with Yale University. Thank you to Kevin Merriman for his help in expediting this process for us.

References

Mark Davies. 2010-. The corpus of historical american english: 400 million words, 1810-2009.

Dai Shen Eun Seo Jo and Michael Xing. 2017. Backprop to the future: A neural network approach to linguistic change over time.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Partha Niyogi and Robert C. Berwick. 1995. The logical problem of language change.

Gigi Stark and William Merrill. 2018. Diachronic part-of-speech tagger. `https://github.com/viking-sudo-rm/DiachronicPOSTagger`.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal 133 statistical laws of semantic change. *ACL 2016*.