Temporal Analysis of Language through Neural Language Models

Yoon Kim* **Yi-I Chiu*** **Kentaro Hanaki*** **Darshan Hegde*** **Slav Petrov◇**

*New York University, New York
◇Google Inc., New York

{yhk255, yic211, kh1615, dh1806}@nyu.edu
slav@google.com

Abstract

We provide a method for automatically detecting change in language across time through a chronologically trained neural language model. We train the model on the Google Books Ngram corpus to obtain word vector representations specific to each year, and identify words that have changed significantly from 1900 to 2009. The model identifies words such as *cell* and *gay* as having changed during that time period. The model simultaneously identifies the specific years during which such words underwent change.

1 Introduction

Language changes across time. Existing words adopt additional senses (*gay*), new words are created (*internet*), and some words 'die out' (many irregular verbs, such as *burnt*, are being replaced by their regularized counterparts (Lieberman et al., 2007)). Traditionally, scarcity of digitized historical corpora has prevented applications of contemporary machine learning algorithms—which typically require large amounts of data—in such temporal analyses. Publication of the Google Books Ngram corpus in 2009, however, has contributed to an increased interest in *culturomics*, wherein researchers analyze changes in human culture through digitized texts (Michel et al., 2011).

Developing computational methods for detecting and quantifying change in language is of interest to theoretical linguists as well as NLP researchers working with diachronic corpora. Methods employed in previous work have been varied, from analyses of word frequencies to more involved techniques (Guolordava et al. (2011); Mihalcea and Nataste (2012)). In our framework, we train a Neural Language Model (NLM) on yearly corpora to obtain word vectors for each year from 1900 to 2009. We chronologically train the model by initializing word vectors for subsequent years with the word vectors obtained from previous years.

We compare the cosine similarity of the word vectors for same words in different years to identify words that have moved significantly in the vector space during that time period. Our model identifies words such as *cell* and *gay* as having changed between 1900–2009. The model additionally identifies words whose change is more subtle. We also analyze the yearly movement of words across the vector space to identify the specific periods during which they changed. The trained word vectors are publicly available.[1]

2 Related Work

Previously, researchers have computationally investigated diachronic language change in various ways. Mihalcea and Nastase (2012) take a supervised learning approach and predict the time period to which a word belongs given its surrounding context. Sagi et al. (2009) use a variation of Latent Semantic Analysis to identify semantic change of specific words from early to modern English. Wijaya and Yeniterzi (2011) utilize a Topics-over-Time model and K-means clustering to identify periods during which selected words move from one topic/cluster to another. They correlate their findings with the underlying historical events during that time. Gulordava and Baroni (2011) use co-occurrence counts of words from 1960s and 1990s to detect semantic change. They find that the words identified by the model are consistent with evaluations from human raters. Popescu and Strapparava (2013) employ statistical tests on frequencies of political, social, and emotional words to identify and characterize epochs.

Our work contributes to the domain in sev-

[1] http://www.yoon.io

eral ways. Whereas previous work has generally involved researchers manually identifying words that have changed (with the exception of Gulordava and Baroni (2011)), we are able to automatically identify them. We are additionally able to capture a word's yearly movement and identify periods of rapid change. In contrast to previous work, we simultaneously identify words that have changed and also the specific periods during which they changed.

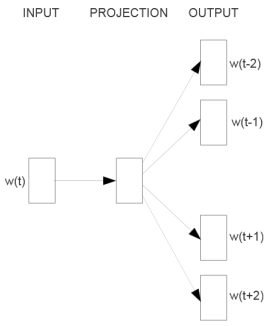

Figure 1: Architecture of a Skip-gram model (Mikolov et al., 2013b).

3 Neural Language Models

Similar to traditional language models, NLMs involve predicting a set of future word given some history of previous words. In NLMs however, words are projected from a sparse, 1-of-V encoding (where V is the size of the vocabulary) onto a lower dimensional vector space via a hidden layer. This allows for better representation of semantic properties of words compared to traditional language models (wherein words are represented as indices in a vocabulary set). Thus, words that are semantically close to one another would have word vectors that are likewise 'close' (as measured by a distance metric) in the vector space. In fact, Mikolov et al. (2013a) report that word vectors obtained through NLMs capture much deeper level of semantic information than had been previously thought. For example, if x_w is the word vector for word w, they note that $x_{apple} - x_{apples} \approx x_{car} - x_{cars} \approx x_{family} - x_{families}$. That is, the concept of pluralization is learned by the vector representations (see Mikolov et al. (2013a) for more examples).

NLMs are but one of many methods to obtain word vectors—other techniques include Latent Semantic Analysis (LSA) (Deerwester et al., 1990), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and variations thereof. And even within NLMs there exist various architectures for learning word vectors (Bengio et al. (2003); Mikolov et al. (2010); Collobert et al. (2011); Yih et al. (2011)). We utilize an architecture introduced by Mikolov et al. (2013b), called the Skip-gram, which allows for efficient estimation of word vectors from large corpora.

In a Skip-gram model, each word in the corpus is used to predict a window of surrounding words (Figure 1). To ensure that words closer to the current word are given more weight in training, distant words are sampled less frequently.[2] Training is done through stochastic gradient descent and backpropagation. The word representations are found in the hidden layer. Despite its simplicity—and thus, computational efficiency—compared to other NLMs, Mikolov et al. (2013b) note that the Skip-gram is competitive with other vector space models in the Semantic-Syntactic Word Relationship test set when trained on the same data.

3.1 Training

The Google Books Ngram corpus contains Ngrams from approximately 8 million books, or 6% of all books published (Lin et al., 2012). We sample 10 million 5-grams from the English fiction corpus for every year from 1850–2009. We lower-case all words after sampling and restrict the vocabulary to words that occurred at least 10 times in the 1850–2009 corpus.

For the model, we use a window size of 4 and dimensionality of 200 for the word vectors. Within each year, we iterate over epochs until convergence, where the measure of convergence is defined as the average angular change in word vectors between epochs. That is, if $V(y)$ is the vocabulary set for year y, and $x_w(y, e)$ is the word vector for word w in year y and epoch number e, we continue iterating over epochs until,

$$\frac{1}{|V(y)|} \sum_{w \in V(y)} \arccos \frac{x_w(y,e) \cdot x_w(y,e-1)}{\|x_w(y,e)\| \|x_w(y,e-1)\|}$$

is below some threshold. The learning rate is set to 0.01 at the start of each epoch and linearly decreased to 0.0001.

[2] Specifically, given a maximum window size of W, a random integer R is picked from range $[1, W]$ for each training word. The current training word is used to predict R previous and R future words.

Most Changed		Least Changed	
Word	Similarity	Word	Similarity
checked	0.3831	*by*	0.9331
check	0.4073	*than*	0.9327
gay	0.4079	*for*	0.9313
actually	0.4086	*more*	0.9274
supposed	0.4232	*other*	0.9272
guess	0.4233	*an*	0.9268
cell	0.4413	*own*	0.9259
headed	0.4453	*with*	0.9257
ass	0.4549	*down*	0.9252
mail	0.4573	*very*	0.9239

Table 1: Top 10 most/least changed words from 1900–2009, based on cosine similarity of words in 2009 against their 1900 counterparts. Infrequent words (words that occurred less than 500 times) are omitted.

Once the word vectors for year y have converged, we initialize the word vectors for year $y+1$ with the previous year's word vectors and train on the $y + 1$ data until convergence. We repeat this process for 1850–2009. Using an open source implementation in the `gensim` package, training took approximately 4 days on a 2.9 GHz machine.

4 Results and Discussion

For the analysis, we treat 1850–1899 as an initialization period and begin our study from 1900.

4.1 Word Comparisons

By comparing the cosine similarity between same words across different time periods, we are able to detect words whose usage has changed. We are also able to identify words that did not change. Table 1 has a list of 10 most/least changed words between 1900 and 2009. We note that almost all of the least changed words are function words. For the changed words, many of the identified words agree with intuition (e.g. *gay*, *cell*, *ass*). Others are not so obvious (e.g. *checked*, *headed*, *actually*). To better understand how these words have changed, we look at the composition of their neighboring words for 1900 and 2009 (Table 2).

As a further check, we search Google Books for sentences that contain the above words. Below are some example sentences from 1900 and 2009 with the word *checked*:

1900: "However, he *checked* himself in time, saying —"

1900: "She was about to say something further, but she *checked* herself."

2009: "He'd *checked* his facts on a notepad from his back pocket."

2009: "I *checked* out the house before I let them go inside."

Word	Neighboring Words in	
	1900	2009
gay	*cheerful* *pleasant* *brilliant*	*lesbian* *bisexual* *lesbians*
cell	*closet* *dungeon* *tent*	*phone* *cordless* *cellular*
checked	*checking* *recollecting* *straightened*	*checking* *consulted* *check*
headed	*haired* *faced* *skinned*	*heading* *sprinted* *marched*
actually	*evidently* *accidentally* *already*	*really* *obviously* *nonetheless*

Table 2: Top 3 neighboring words (based on cosine similarity) specific to each time period for the words identified as having changed.

At the risk of oversimplifying, the resulting sentences indicate that in the past, *checked* was more frequently used with the meaning *"to hold in restraint"*, whereas now, it is more frequently used with the meaning *"to verify by consulting an authority"* or *"to inspect so as to determine accuracy"*. Given that *check* is a highly polysemous word, this seems to be a case in which the popularity of a word's sense changed over time.

Conducting a similar exercise for *actually*, we obtain the following sentences:

1900: "But if ever he *actually* came into property, she must recognize the change in his position."

1900: "Whenever a young gentleman was not *actually* engaged with his knife and fork or spoon —"

2009: "I can't believe he *actually* did that!"

2009: "Our date was *actually* one of the most fun and creative ones I had in years."

Like the above, this seems to be a case in which the popularity of a word's sense changed over time (from *"to refer to what is true or real"* to *"to express wonder or surprise"*).

4.2 Periods of Change

As we chronologically train the model year-by-year, we can plot the time series of a word's distance to its neighboring words (from different years) to detect periods of change. Figure 2 (above) has such a plot for the word *cell* compared to its early neighbors, *closet* and *dungeon*, and the more recent neighbors, *phone* and *cordless*. Figure 2 (below) has a similar plot for *gay*.

Such plots allow us to identify a word's period of change relative to its neighboring words,

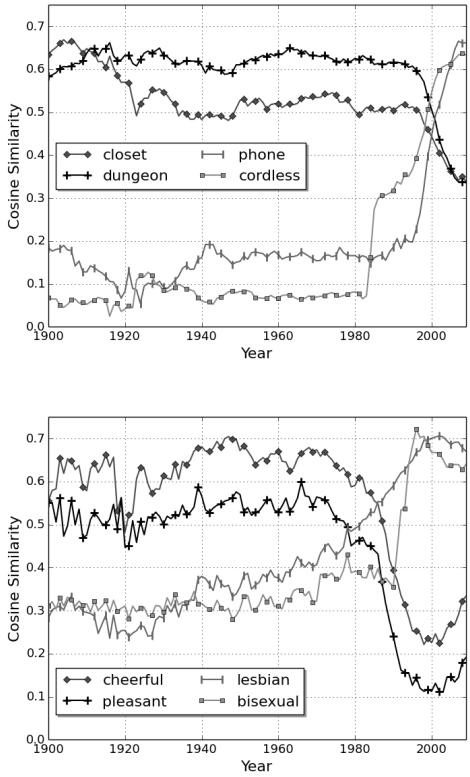

Figure 2: (Above) Time trend of the cosine similarity between *cell* and its neighboring words in 1900 (*closet*, *dungeon*) and 2009 (*phone*, *cordless*). (Below) Similar plot of *gay* and its neighboring words in 1900 (*cheerful*, *pleasant*) and 2009 (*lesbian*, *bisexual*).

and thus provide context as to how it evolved. This may be of use to researchers interested in understanding (say) when *gay* started being used as a synonym for *homosexual*. We can also identify periods of change independent of neighboring words by analyzing the cosine similarity of a word against itself from a reference year (Figure 3). As some of the change is due to sampling and random drift, we additionally plot the average cosine similarity of all words against their reference points in Figure 3. This allows us to detect whether a word's change during a given period is greater (or less) than would be expected from chance. We note that for *cell*, the identified period of change (1985–2009) coincides with the introduction—and subsequent adoption—of the cell phone by the general public.[3] Likewise, the period of change for *gay* agrees with the gay movement which began around the 1970s (Wijaya and Yeniterzi, 2011).

4.3 Limitations

In the present work, identification of a changed word is conditioned on its occurring often enough

[3]http://library.thinkquest.org/04oct/02001/origin.htm

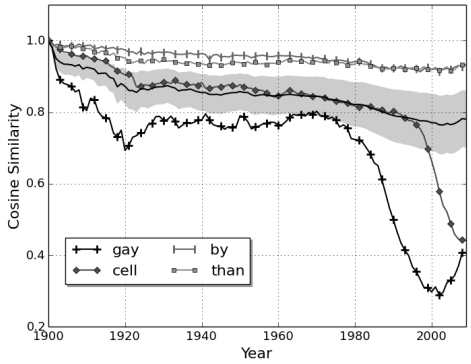

Figure 3: Plot of the cosine similarity of changed (*gay*, *cell*) and unchanged (*by*, *than*) words against their 1900 starting points. Middle line is the average cosine similarity of all words against their starting points in 1900. Shaded region corresponds to one standard deviation of errors.

in the study period. If a word's usage decreased dramatically (or stopped being used altogether), its word vector will have remained the same and hence it will not show up as having changed. One way to overcome this may be to combine the cosine distance and the frequency to define a new metric that measures how a word's usage has changed.

5 Conclusions and Future Work

In this paper we provided a method for analyzing change in the written language across time through word vectors obtained from a chronologically trained neural language model. Extending previous work, we are able to not only automatically identify words that have changed but also the periods during which they changed. While we have not extensively looked for connections between periods identified by the model and real historical events, they are nevertheless apparent.

An interesting direction of research could involve analysis and characterization of the different types of change. With a few exceptions, we have been deliberately general in our analysis by saying that a word's *usage* has changed. We have avoided inferring the *type* of change (e.g. semantic vs syntactic, broadening vs narrowing, pejoration vs amelioration). It may be the case that words that undergo (say) a broadening in senses exhibit regularities in how they move about the vector space, allowing researchers to characterize the type of change that occurred.

References

Y. Bengio, R. Ducharme, P. Vincent. 2003. Neural Probabilitistic Language Model. *Journal of Machine Learning Research 3:1137–1155.*

D. Blei, A. Ng, M. Jordan, J. Lafferty. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3:993–1022*.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuglu, P. Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research 12:2493–2537*.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman. 2011. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41(6):391–407*.

K. Gulordava, M. Baroni. 2011. A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus. *Proceedings of the GEMS 2011 Workshop*.

E. Lieberman, J.B. Michel, J. Jackson, T. Tang, M.A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature, 449: 716–716, October*.

Y. Lin, J.B. Michel, E.L. Aiden, J. Orwant, W. Brockman, S. Petrov. 2012. Syntactic Annotations for the Google Books Ngram Corpus. *Proceedings of the Association for Computational Linguistics 2012*.

J.B Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S.Pinker, M.A. Nowak, E.L. Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science, 331(6014): 176–182, January.*

R. Mihalcea, V. Nastase. 2012. Word Epoch Disambiguation: Finding How Words Change Over Time. *Proceedings of the Association for Computational Linguistics 2012*.

T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, S. Khudanpur. 2010. Recurrent Neural Network Based Language Model. *Proceedings of Interspeech.*

T. Mikolov, W.T Yih, G. Zweig. 2013a. Linguistic Regularities in Continuous Space Word Representations. *Proceedings of NAACL-HLT 2013, 746–751*.

T. Mikolov, K. Chen, G. Corrado, J.Dean. 2013b. Efficient Estimation of Word Representations in Vector Space *arXiv Preprint*.

O. Popescu, C. Strapparava. 2013. Behind the Times: Detecting Epoch Changes using Large Corpora. *International Joint Conference on Natural Language Processing, 347–355*

E. Sagi, S. Kaufmann, B. Clark 2009. Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. *Proceedings of the EACL 2009 Workshop on GEMS: 104–111*.

D.T. Wijaya, R. Yeniterzi. 2011. Understanding semantic change of words over centuries. *Proceedings of the 2011 international workshop on DEtecting and Exploiting Cultural diversiTy on the social web: 35–40*.

W. Yih, K. Toutanova, J. Platt, C. Meek. 2011. Learning Discriminative Projections for Text Similarity Measures. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning, 247–256*.