STA260 Consulting Project

Word Semantic Shift in Early Modern English

Group Members: Doudou Zhou, Yi Han, Wancheng Cai, Yidong Zhou, Yishan Huang

Department of statistics, UC Davis

March 10, 2021

Outline

Introduction and Background

Word Semantic changes over time

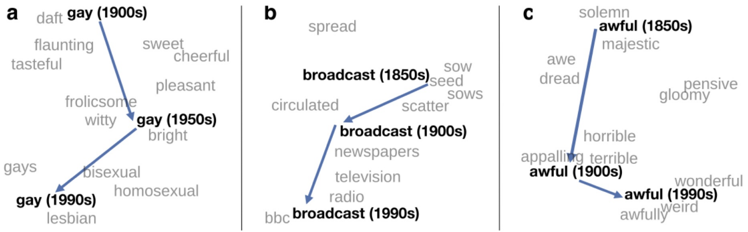


Figure 1: Examples of semantic change

Introduction and background

Two quantitative laws of semantic changes

- the *law of conformity*: The rate of semantic change scales with an inverse power-law of word frequency;
- the *law of innovation*: Independent of frequency, words that are more polysemous have higher rates of semantic changes

Object

EEBO-TCP Dataset

The *Early English Books Online TCP* (EEBO-TCP) is a corpus consisting of 60000 transcriptions from 1470 to 1700.

- large range in document length
- reprints
- transcription errors

Goal

Verify the validity of the two laws in EEBO datasets

Data preprocessing

Solve the undesirable properties in EEBO Datasets

- large range of document length: Separating datasets to equal number of words according to the print time.
- reprints: Compare the author set, title and main text of each pair of books to determine pairs of duplicate/reprinting books. We delete the earlier book in all such pairs.
- transcription errors: Delete all the non-ASCII character.

Word Embedding Methods

PPMI matrix

Positive Pointwise Mutual Information (PPMI) is a characteristic that can be used to represent the correlation of two words. The word vectors correspond to the rows of the matrix $\mathsf{M}^{\mathrm{PPMI}} \in \mathsf{R}^{|V| \times |V_c|}$ with entries given by

$$\mathsf{M}_{i,j}^{\mathrm{PPMI}} = \max\left\{\log\left(\frac{\hat{p}\left(w_i, c_j\right)}{\hat{p}(w)\hat{p}\left(c_j\right)}\right) - \alpha, 0\right\} \tag{1}$$

where $\alpha > 0$ is a negative prior which provides the smoothing bias, and the $\hat{p}$ corresponds to the smoothed empirical probabilities of word (co-)occurrences within fixed-size sliding windows of text.

Word Embedding Methods

SVD of the PPMI matrix

SVD embeddings correspond to low-dimensional approximations of the PPMI embeddings learned via singular value decomposition. The vector embedding for word wi is given by

$$\mathsf{w}_i^{\mathrm{SVD}} = (\mathsf{U}\Sigma^{\gamma})_i$$

where $\mathsf{M}^{\mathrm{PPMI}} = \mathsf{U}\Sigma \boldsymbol{V}^{\top}$ is the truncated singular value decomposition of $\mathsf{M}^{\mathrm{PPMI}}$ and $\gamma \in [0, 1]$ is an eigenvalue weighting parameter. Setting $\gamma < 1$ has been shown to dramatically improve embedding qualities.

Word Embedding Methods

word2vec

Example: *'I like to eat apple'*

- Skip-gram (SG): Uses the target word to predict the context word within fixed-size sliding windows of text. It actually builds a one-layer neural networks.
- Negative Sampling: Only consider a subset of the '0' output in the one-hot representation which highly decrease the dimension of the weights that need to be updated each time, reducing the calculation burden.
- Skip-gram with negative sampling (SGNS): Combine these two.

word2vec

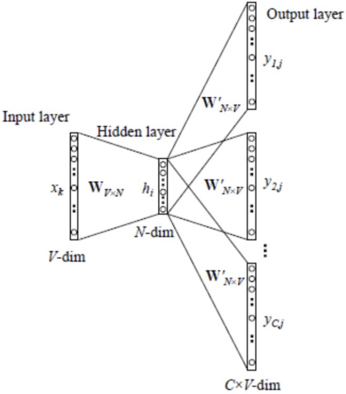


Figure 2: Skip-gram method

Quantifying characteristics

Quantifying polysemy

We measure a word's contextual diversity, and thus polysemy, by examining its neighborhood in an empirical co-occurrence network. The polysemy of a word is defined as the reciprocal of its local clustering coefficient within this network:

$$d\left(w_i\right) = 1 / \frac{\sum_{c_i, c_j \in N_{\mathrm{PPMI}}\left(w_i\right)} \mathbb{I}\left\{\mathrm{PPMI}\left(c_i, c_j\right) > 0\right\}}{\left|N_{\mathrm{PPMI}}\left(w_i\right)\right| \left(\left|N_{\mathrm{PPMI}}\left(w_i\right)\right| - 1\right)} \tag{2}$$

where $N_{\mathrm{PPMI}}\left(w_i\right) = \left\{w_j : \mathrm{PPMI}\left(w_i, w_j\right) > 0\right\}$.

Quantifying semantic change

Embedding alignment

We use orthogonal Procrustes to align the learned low-dimensional embeddings across the consecutive time points.
Defining $\mathrm{W}^{(t)} \in \mathbb{R}^{d\times|\mathcal{V}|}$ as the matrix of word embeddings learned at year t, we align across time-periods while preserving cosine similarities by optimizing:

$$\mathrm{R}^{(t)} = \arg\min_{\mathrm{Q}^\top\mathrm{Q}=\mathrm{I}} \left\| \mathrm{Q}\mathrm{W}^{(t)} - \mathrm{W}^{(t+1)} \right\|_F$$

with $\mathrm{R}^{(t)} \in \mathbb{R}^{d\times d}$. The solution corresponds to the best rotational alignment and can be obtained efficiently using an application of SVD.

Quantifying semantic change

Quantifying semantic change

The measurements of a word's rate of semantic change is defined as:

$$\Delta^{(t)}(w_i) = \cos - \text{dist}\left(\mathrm{w}_i^{(t)}, \mathrm{w}_i^{(t+1)}\right) = 1 - \frac{\mathrm{w}_i^{(t)} \cdot \mathrm{w}_i^{(t+1)}}{\|\mathrm{w}_i^{(t)}\| \|\mathrm{w}_i^{(t+1)}\|} \quad (3)$$

depends on its frequency, $f^{(t)}(w_i)$ and a measure of its polysemy, $d^{(t)}(w_i)$

$\Delta^{(t)}(w_i)$ is log-transformed and normalized and referred as $\tilde{\Delta}^{(t)}(w_i)$, which is the response we consider.

Linear Mixed Model

Linear Mixed Model

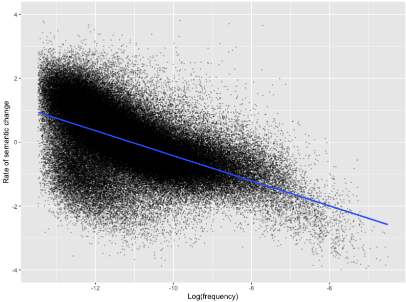
$$\tilde{\Delta}^{(t)}(w_i) = \beta_f \log\left(f^{(t)}(w_i)\right) + \beta_d \log\left(d^{(t)}(w_i)\right) + \beta_t + z_{w_i} + \epsilon_{w_i}^{(t)} \quad \forall w_i \in \mathcal{V}, t \in \{t_0, \ldots, t_n\} \tag{4}$$

where $\beta_f, \beta_d,$ and β_t correspond to the fixed effects for logarithm of frequency, logarithm of polysemy and the decade t, respectively. $z_{w_i} \sim \mathcal{N}(0, \sigma_{w_i})$ is the random intercept for word w_i and $\epsilon_{w_i}^{(t)} \in \mathcal{N}(0, \sigma)$ is an error term.

Then we can study the relationship between rates of semantic change and word frequency/polysemy depending on the values of β_f, β_d.

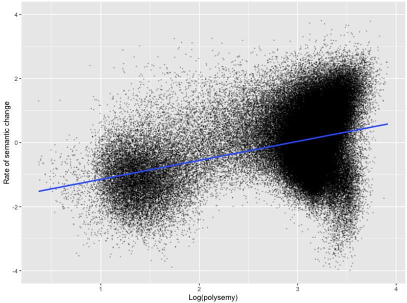
Law of conformity: Frequently used words change at slower rates

- the logarithm of a word's frequency, $\log(f(w_i))$, has a significant and substantial negative effect on rates of semantic change.
- rates of semantic change are proportional to a negative power (β_f) of frequency, i.e., $\Delta(w_i) \propto f(w_i)^{\beta_f}$ with $\beta_f = -0.4365$ and p-value$< 2 \times 10^{-16}$



Law of innovation: Polysemous words change at faster rates

- the logarithm of the polysemy score exhibits a strong positive effect on rates of semantic change.
- As with frequency, the relation takes the form of a power law $\Delta(w_i) \propto d(w_i)^{\beta_d}$ with $\beta_d = 0.6993$ and p-value$< 2 \times 10^{-16}$



Conclusions

- for EEBO corpus, rates of semantic change obey a scaling relation of the form $\Delta(w_i) \propto f(w_i)^{\beta_f} \times d(w_i)^{\beta_d}$ with $\beta_f = -0.4365 < 0$ and $\beta_d = 0.6993 > 0$.
- frequent words change at slower rates while polysemous words change faster, and that both these relations scale as power laws, which confirms that the laws of conformity and innovation [Hamilton et al., 2016] are true for EEBO corpus.

Acknowledgement

The authors thank Bala Rajaratnam and Benjamin Roycraft for their helpful comments and discussions. The authors also thank Arthur Koehl for his insightful ideas and suggestions.

Contribution

- Doudou Zhou: Construct the PPMI matrices, train the embeddings using SVD and compute the rotation matrices.
- Yi Han: Train the SGNS word2vec model and get the word-context matrices.
- Yishan Huang: Data cleaning and pre-processing.
- Wancheng Cai: Summarize all word embedding results for the final linear mixed model.
- Yidong Zhou: Run linear mixed model and analyze the result.

All the team members contribute equally to the final report and presentation slides.

Bibliography

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016).
Diachronic word embeddings reveal statistical laws of semantic change.
arXiv preprint arXiv:1605.09096.