CAPP 30123 Project Report

Group name: SSSP

Group members: Ling Dai, Yilun Dai, Jie Heng, Fangfang Wan

Introduction

Our project is aimed to investigate the semantic change in written English, and its potential driving factors. That is, how meanings of vocabularies in written English change over time, and what factors can influence the magnitude of the changes. To do that, we analyzed an abundance of textual documents in written English using several popular approaches in semantics and big data methods such as MapReduce. Through our analysis, we successfully identified two factors that were statistically associated with the semantic change of a vocabulary: relative frequency, and polysemy.

Description of Dataset

Two corpus were used in our project: (1) Corpus of Late Modern English (CLMET 3.1) and (2) the Gutenberg Project Corpus.

The Corpus of Late Modern English is about 0.2 GB in size. Created by De Smet, Diller and Tyrkkö, this corpus is a mixed-genre British English Corpus, with fiction, narrative non-fiction, drama, letters and treatise, covering the period 1710 - 1920. Due to its genre-balanced nature, CLMET can be a good representation of late modern written English, and thus an excellent data source for our analysis. However, CLMET also brings certain disadvantages. While the corpus itself is already of limited size, the fact that some documents are not accurately

dated further accentuate such problem, making it hard to perform time-series analysis with this corpus.

To make up for the disadvantage of CLMET, we also used the Gutenberg Project corpus as a secondary data source to further validate our results. The Gutenberg Project is a relatively large corpus (3.6 GB). Nevertheless, because it is not carefully mixed in genres, the corpus itself may not be as representative as CLMET.

Hypotheses

There were mainly two hypotheses that we were trying to test throughout or project:

(1) The magnitude of semantic change is negatively correlated with relative usage frequency

(2) The magnitude of semantic change is positively correlated with polysemy (number of meanings for a vocabulary)

Our reasoning behind the first hypothesis is that, the more frequently used words in written language are more likely to be treated as "cornerstones" of this language. That is, these vocabularies are so essential to written English that people define the meanings of other vocabularies based on these most frequently used ones. Therefore, the meanings of these highly frequently used vocabularies are less likely to change throughout time.

Our reasoning behind the second hypothesis is that, if a vocabularies have many different meanings, it is likely that the relative frequency of each meaning being used will change over time, causing the vocabulary to have quite different embedding in different periods.

Algorithm

1. Cosine Similarity/Distance

In our project, we used cosine distance as the measurement of semantic change. Cosine distance is defined as one minus cosine similarity, and cosine similarity can essentially be viewed as the "angle" between two vectors of the same dimensions. The cosine similarity between two vectors is computed by dividing their dot product by the product of their magnitudes (Figure 1) [1]. The value of cosine similarity can range from -1 to 1, and is independent of the lengths of the vectors.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figure 1: definition of cosine similarity

While there are many important applications of cosine similarity in various fields, in linguistics, it has mainly been used to calculate the similarity score between two vocabularies using their word embedding vectors. Our project uses an extension of this concept: instead of computing the similarity between meanings of two vocabularies, we compute the similarity/distance between meanings of a single vocabulary in two different periods of time. Using this approach, we can then quantify how dramatically the meaning of a vocabulary has changed over time.

2. Word embedding (collocation)

Word embedding is a common technique in natural language processing (NLP) and machine learning: it converts information about a word into vectors. In our project, we mainly

focus on the information of collocation, or conditional frequency, of each vocabulary. In another word, we measure how frequently does a vocabulary appear together with other words in a sentence and store that information as a vector. MapReduce is used to complete this task. Using a reference frame of 1000 vocabularies, the MapReduce would divide the co-occurrence counts of one word in the text corpus and one word in the reference frame by the total counts of the word in the corpus, generating a vector of 1000 length for each unique vocabulary in the text files.

In particular, the mapper goes through each sentence in the corpus and yields a list for each word. The list has one more element than the reference frame. The first element is the word count, '1', and other element is the times that words in the reference frame co-occur with the key word in the same line(sentence). If an identical word is found, the count of co-occurrence remains '0'. The combiner and reducer would sum up all elements in the lists with the same keyword. The reducer, then, divide the second-to-last elements by the first element, the total word counts and remove the first number, yielding a list with the same length with the reference frame. The collocation vectors generated are then used for calculating cosine similarity.

3. Word count and relative frequency

The relative frequency is computed as the count of a specific vocabulary divided by the total word count. We first calculate the word count for every unique word. The mapper counts each word, the combiner adds the count corresponding to the unique word, and the reducer yields the count of each unique word, saved in a csv document. We then import the csv to a pandas dataframe and divide each unique word count by the total word count to obtain the relative frequency.

Relative frequencies are used in two tasks: (1) to test our first hypothesis, and (2) to select high frequency vocabularies that are used as the reference frame for computation of collocation vectors.

Challenges

1. Estimating polysemy:

    Unlike relative frequency that is relatively easy to measure, polysemy (the number of meanings of a vocabulary) is very hard to accurately quantify.

    In a study conducted by scholars at Stanford University, the researches indirectly measured polysemy of vocabulary using the diversity of context in which the vocabulary appears. However, due to the complexity and indirectness of that approach, we decided to use a more straight-forward method: count the number of meanings using the meaning() method of the PyDictionary library. While our approach might not be complete accurate, it is overall a good estimate of polysemy of a vocabulary.

Results

1. Magnitude of semantic change is negatively correlated with relative frequency:

We first fitted an OLS linear regression model to investigate the correlation between magnitude of semantic change and relative frequency using the CLMET corpus. The linear model shows that there is a negative association (coefficient = -555.1462) between semantic change and relative frequency, which is consistent to our first hypothesis. The relative frequency is significant at the 0.01 level (P>|t| = 0.000).

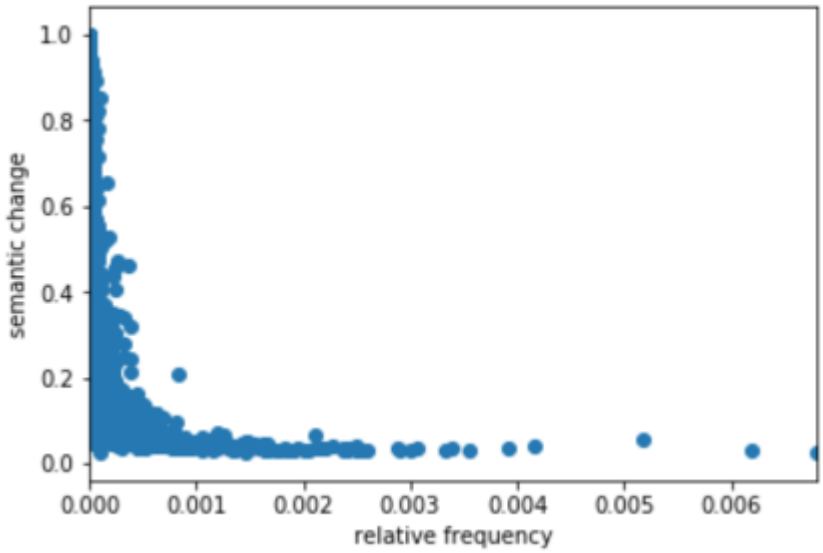

Figure 2: scatterplot of relative frequency vs. semantic change (CLMET)

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5374	0.002	339.611	0.000	0.534	0.540
relative_freq	-555.1462	9.265	-59.921	0.000	-573.306	-536.987

Figure 3: estimated coefficients of linear regression model (CLMET)

The observed negative correlation between semantic change and relative frequency was also confirmed using the Gutenberg corpus. For the Gutenberg corpus, the OLS model shows a negative correlation (coefficient = -511.2515) at the 0.01 level (P>|t| = 0.000). Moreover, because the scatter plot suggests a highly non-linear relationship between the magnitude of semantic change and relative frequency, we also fitted a model using the log scale of relative frequency. The scatterplot of log relative frequency vs. semantic change shows that the relationship between these two variables is approximately linear.

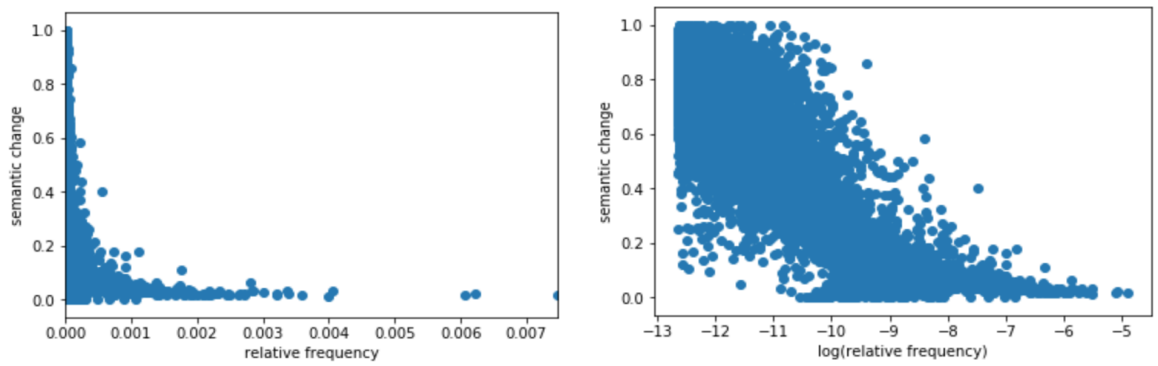

Figure 4: left: scatterplot of relative frequency vs. semantic change (Gutenberg)
right: scatterplot of log relative frequency vs. semantic change (Gutenberg)

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5250	0.002	333.738	0.000	0.522	0.528
relative_freq	-511.2515	8.726	-58.592	0.000	-528.354	-494.149

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.3646	0.007	-187.617	0.000	-1.379	-1.350
relative_freq	-0.1680	0.001	-258.096	0.000	-0.169	-0.167

Figure 5: Top: estimated coefficients of relative frequency (Gutenberg)
Bottom: estimated coefficients of log relative frequency (Gutenberg)

2. Magnitude of semantic change is negatively correlated with polysemy:

Contrary to our original hypothesis, we found that polysemy had a positive correlation with the magnitude of semantic change. Our OLS regression result for CLMET yielded a negative estimated coefficient of -0.0168 and a p-value of 0. The results were also similar for the Gutenberg corpus: estimated coefficient is -0.0171 and the p-value is 0.

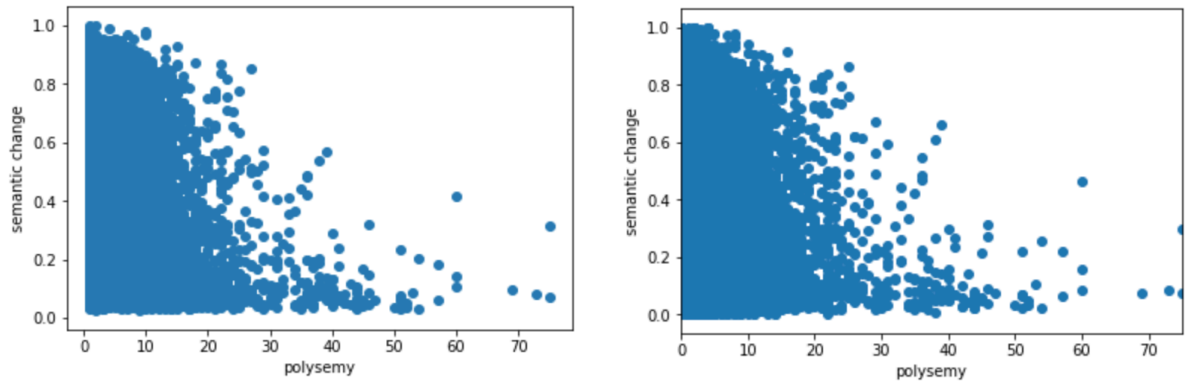

Figure 6: Scatterplot of polysemy vs. semantic change (left: CLMET, right: Gutenberg)

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5800	0.002	272.431	0.000	0.576	0.584
polysemy	-0.0168	0.000	-52.080	0.000	-0.017	-0.016

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5737	0.002	281.871	0.000	0.570	0.578
polysemy	-0.0171	0.000	-54.311	0.000	-0.018	-0.016

Figure 7: estimated coefficient of polysemy (top: CLMET, bottom: Gutenberg)

From the scatterplots, we can observe that while low-polysemy vocabularies have huge variation in their magnitude of semantic change, all high-polysemy vocabularies seem to have relatively low semantic change.

While the results for polysemy were not consistent to our original hypothesis, we do not currently have a concrete explanation for the observed phenomenon. In order to further investigate the underlying mechanism, more analyses need to be done on polysemy and diversity of context of vocabularies.

References:

1. Wikipedia. (May 22, 2018). *Cosine Similarity*. https://en.wikipedia.org/wiki/Cosine_similarity.