CSC2611 Lab Assignment: Word embedding and semantic change
Winter 2021

Yoon A Park

Github: https://github.com/yoona3877/semantic_word_change

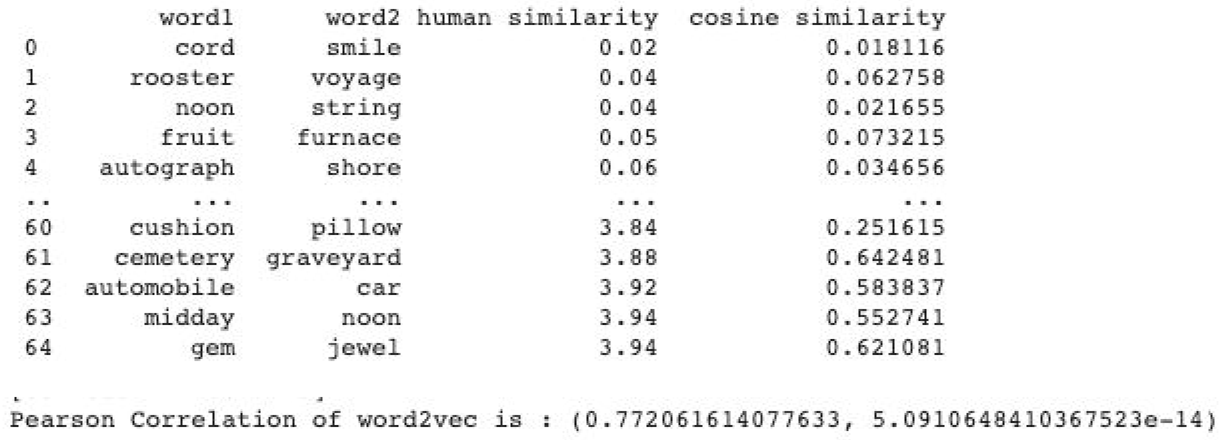
A. SYNCHRONIC WORD EMBEDDING

Pearson correlation of word2vec and LSA
Word2vec & Human judgement: 0.772
LSA & Human judgement: 0.143

Comment: Cosine similarity calculated with Word2Vec word embeddings has a much higher correlation with the human judgement than LSA. This is an expected result due to the complexities involved for building Word2Vec. Followings are some reasoning behind the higher similarities between word2vec and human judgements.

1) Word2vec is a context-free language model that is trained on large corpus of text data.
2) A deep neural network that was used for training can capture hidden linguistic (semantic/syntactic, pragmatic, etc.) patterns.
3) Word2vec is a prediction-based model which is better at generalization than LSA which is a frequency-based model.

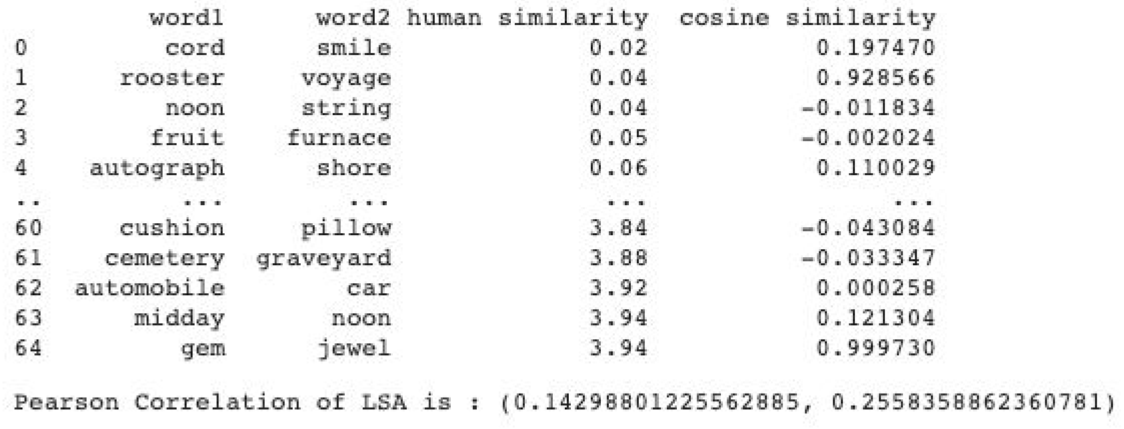
Table1: Word pairs and comparison between human judgement and word2vec cosine similarity



```
         word1      word2  human similarity  cosine similarity
0         cord      smile              0.02           0.018116
1      rooster     voyage              0.04           0.062758
2         noon     string              0.04           0.021655
3        fruit    furnace              0.05           0.073215
4    autograph      shore              0.06           0.034656
..         ...        ...               ...                ...
60     cushion     pillow              3.84           0.251615
61    cemetery  graveyard              3.88           0.642481
62  automobile        car              3.92           0.583837
63      midday       noon              3.94           0.552741
64         gem      jewel              3.94           0.621081

Pearson Correlation of word2vec is : (0.772061614077633, 5.0910648410367523e-14)
```

Table2: Word pairs and comparison between human judgement and LSA cosine similarity



```
         word1      word2  human similarity  cosine similarity
0         cord      smile              0.02           0.197470
1      rooster     voyage              0.04           0.928566
2         noon     string              0.04          -0.011834
3        fruit    furnace              0.05          -0.002024
4    autograph      shore              0.06           0.110029
..         ...        ...               ...                ...
60     cushion     pillow              3.84          -0.043084
61    cemetery  graveyard              3.88          -0.033347
62  automobile        car              3.92           0.000258
63      midday       noon              3.94           0.121304
64         gem      jewel              3.94           0.999730

Pearson Correlation of LSA is : (0.14298801225562885, 0.2558358862360781)
```

Semantic & Syntactic Analogy Test

Due to the limited vocabulary used for building LSA (only 5031 words), most of the analogy pairs in the data file were filtered out. **Table 3** shows that Word2vec significantly outperformed LSA for both semantic and syntactic analogy. Again, this is expected due to the complexity involved when constructing word2vec. Both models perform better at the semantic analogy test than another. This might be due to the fact that both models are trained with the co-occurence of a target word and surrounding words. Semantic pairs (e.g. king and queen) tend to occur more frequently than syntactic pairs (e.g. fast and faster).

Table 3: Semantic and Syntactic Analogy Accuracies of Word2Vec and LSA

Analogy	Number of pairs	word2vec	LSA
Semantic	163	0.345679	0.018518
Syntactic	2045	0.133007	0.006357

How to improve vector-based model

Defining word similarities is the first and foremost step before conducting downstream analysis. Word2Vec and LSA are based trained with co-occurence of words, which might be successful at capturing some linguistic features, but fail to capture others. There are numerous ways to improve the current word embeddings, and followings are a few examples.

a) Generate synthetic data to capture syntactic relationships. From Table 3, we can see that one of the most common word embeddings, word2vec, has difficulty capturing the syntactic information. This can be improved by training the model with the synthetic grammatic/synthetic pairs.

b) Sentiment similarity can also be well improved by careful pre- or post-processing. The simplest approach of sentiment analysis is a binary classification that distinguishes positive from negative. Oftentimes, even the simple sentiment classification fails with word2vec. For example, ‘good’ and ‘bad’ tend to be close together in word2vec due to their high co-occurrence. This pitfall can be improved by utilizing sentiment lexicons or taking contextual representation into account.

c) Also, word meanings tend to change depending on in which domain they are used. A word pair that has similar in meaning in one domain may become totally irrelevant in another domain. To remedy the confusion around the word meaning and its usage in different contexts, additional features can be utilized to specify the domain in which the target word is used.

B. DIACHRONIC WORD EMBEDDING

Three Measure of Semantic Change

Below are three proposed methods in an attempt to find words with the most semantic change.

1) Select words that have the least similarity between any two periods. For example, let's say there are two distinct words each diachronic with embeddings of A = [1,1,1,7] and B = [1,2,3,4] respectively. There are four time periods, each position representing a distinct time. The largest distance within A and B is 7-1=6 and 4-1=3 respectively. Therefore, A has a larger change in semantic.
2) This method is identical to the first method, but vector spaces in different time periods are aligned. The idea of vector space alignment has been proposed by Kulkarni et al. (2015). The author claims that aligning the embeddings in different time periods is essential due to the stochastic nature of model training. Once the embeddings are aligned, it selects words that have the least similarity between any two periods.
3) The last method relies on the 20-nearest-neighbors of a target word of the most recent word embeddings (i.e. word embeddings of 1990). After collecting the nearest neighbors, it computes the average cosine similarity of the target word and the nearest neighbors in the different time period. Then, it selects words that have the least average cosine similarity between the target word and the neighbors in different time periods.

Table 4: Top-20 and Bottom-20 of change in semantic using 3 different methods

	Largest distance		Aligned Largest distance		Aligned knn-based	
Rank	Top	Bottom	Top	Bottom	Top	Bottom
1	objectives	april	mcgraw	days	bar	months
2	computer	november	ministers	time	wall	summer
3	programs	february	relief	period	acts	november
4	sector	october	star	month	st	day
5	radio	january	acts	weeks	actions	december
6	goals	june	roof	months	organ	year
7	perspective	december	chest	duration	orders	july

8	shri	september	witnesses	date	pound	month
9	impact	century	bond	year	characters	january
10	approach	daughter	committee	day	beings	april
11	van	evening	rules	arrival	command	february
12	media	july	commons	trip	punishment	october
13	patterns	husband	grounds	visit	decisions	september
14	assessment	coast	electricity	couple	chamber	week
15	berkeley	trees	languages	outbreak	bones	june
16	princeton	river	corn	term	arrangement	season
17	shift	church	carbon	decades	stanford	spring
18	therapy	increase	affair	night	consumer	august
19	film	god	breach	summer	responsibility	centuries
20	j	mildes	wisconsin	war	fleet	morning

Pearson correlation

Table 5 shows the Pearson correlation between the three proposed methods. Note that the correlation between the first method and others is low due to the fact that the embeddings used in the first method are not aligned. The second and the third method shows moderate correlation with approximately 0.55 as the embeddings used are both aligned.

Table 5: Pearson Correlation between 3 proposed methods of change in semantics

	1st method	2nd method	3rd method
1st method	1.000000	0.254980	0.207737
2nd method	0.254980	1.000000	0.543528
3rd method	0.207737	0.543528	1.000000

Evaluation (using Pearson Correlation)

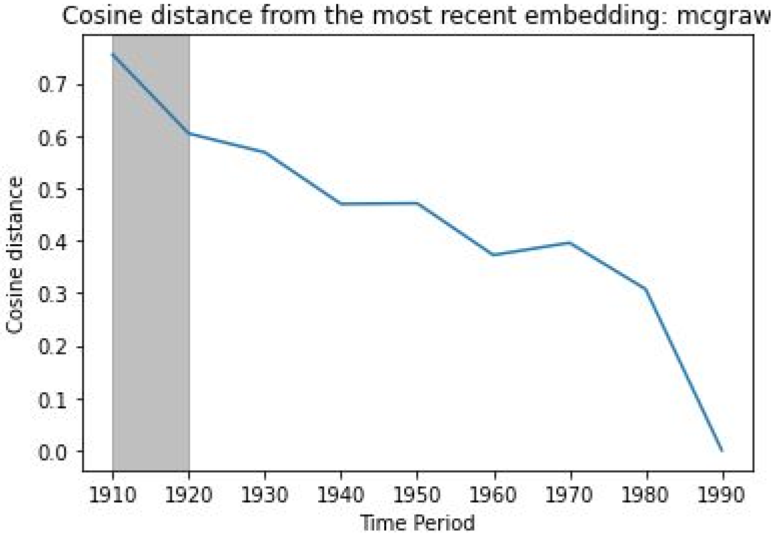
For evaluation, I came up with an algorithm that focuses more on the change in similarity in the subsequent time period. Kulkarni et al. (2015) claims that semantic change occasionally relies on the sudden change of new usage in a short time period rather than gradual change. With this claim, the evaluation algorithm focuses on finding the largest cosine distance between the subsequent time period for each word. As such, the word with the largest semantic change with such a method can be found by selecting the word with the largest distance between the subsequent time period. **Table 6** shows the Pearson correlation between the evaluation and the three proposed methods. Evaluation is done with both aligned and non-aligned embeddings. Only the one highlighted in red is relevant to the study. We can see from **Table 6** that the second method has the highest correlation with the evaluation. Thus, the second method will be used for further analysis.

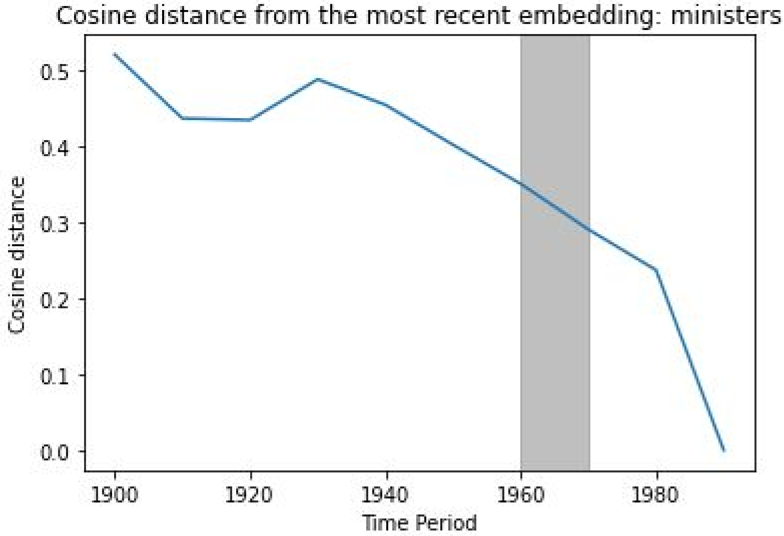
Table 6: Pearson Correlation between Evaluation and 3 different methods

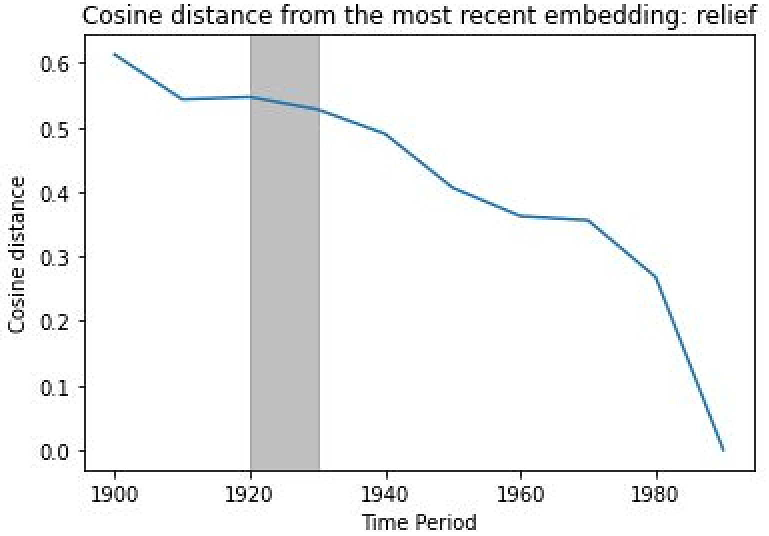
	Not Aligned	**Aligned**
First method	**0.715257**	0.187165
Second method	0.180171	**0.801019**
Third method	0.117423	**0.443575**

Detect the point of semantic change

From the evaluation, the second method is chosen to visualize semantic change. The top three words that are shown to have the highest semantic change based on the second method are ‘mcgraw’, ‘ministers’, and ‘relief’. The y-axis displays the cosine similarity of embeddings between a specific period and the most recent period (i.e. 1990). The grey-highlighted zone indicates the largest change in cosine similarity in the subsequence time period. As the y-axis displayed the distance away from the embeddings of 1990, the highlighted zone is not necessarily the steepest line in the graph.


Cosine distance from the most recent embedding: mcgraw
Cosine distance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1910
1920
1930
1940
1950
1960
1970
1980
1990
Time Period


Cosine distance from the most recent embedding: ministers
Cosine distance
0.0
0.1
0.2
0.3
0.4
0.5
1900
1920
1940
1960
1980
Time Period


Cosine distance from the most recent embedding: relief
Cosine distance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1900
1920
1940
1960
1980
Time Period

Reference

Kulkarni, Vivek et al. 2015. Statistically Significant Detection of Linguistic Change. The International World Wide Web Conference Committee.