

# Understanding Semantic Change of Words Over Centuries

Derry Tanti Wijaya  
Language Technologies Institute  
Carnegie Mellon University  
5000 Forbes Ave.  
Pittsburgh, PA 15213  
dwijaya@andrew.cmu.edu

Reyyan Yeniterzi  
Language Technologies Institute  
Carnegie Mellon University  
5000 Forbes Ave.  
Pittsburgh, PA 15213  
reyyan@cs.cmu.edu

## ABSTRACT

In this paper, we propose to model and analyze changes that occur to an entity in terms of changes in the words that co-occur with the entity over time. We propose to do an in-depth analysis of how this co-occurrence changes over time, how the change influences the state (semantic, role) of the entity, and how the change may correspond to events occurring in the same period of time. We propose to identify clusters of topics surrounding the entity over time using Topics-Over-Time (TOT) and k-means clustering. We conduct this analysis on Google Books Ngram dataset. We show how clustering words that co-occur with an entity of interest in 5-grams can shed some lights to the nature of change that occurs to the entity and identify the period for which the change occurs. We find that the period identified by our model precisely coincides with events in the same period that correspond to the change that occurs.

## Categories and Subject Descriptors

H.2.8 [Information Systems]: Database Management—*Database Applications, Data Mining*; H.3 [Information Systems]: Information Storage and Retrieval; I.2.7 [Natural Language Processing]: Text analysis—*topics over time, k-means clustering*

## General Terms

Algorithms, Experimentation.

## Keywords

Topic clustering, topic transition over time, semantic change, event detection

## 1. INTRODUCTION

Entities; words or real world entities change over time. Words change semantically and this change is reflected in the way the words are being used. As people use words in new

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

DETECT'11, October 24, 2011, Glasgow, Scotland, UK.  
Copyright 2011 ACM 978-1-4503-0962-2/11/10 ...\$10.00.

contexts, the meanings of the words change gradually, often to the point that the new meaning is radically different from the original usage. For example, awful<sup>1</sup> originally meant 'awe-inspiring, filling someone with deep awe', as in *the awful majesty of the Creator*. At some point it becomes something extremely bad, as in an awfully bad performance, but now the intensity of the expression has lessened and the word is now used informally to just mean 'very bad', as in *an awful mess*. Some words also change semantically, not in their original meanings but change in a way that they acquire additional meanings or are used to refer to other named entities over time. For example, mouse is used originally to refer to small long-tailed animal but it is now also used to refer to a device used to control cursor movement.

Automatically identifying changes to an entity over time is beneficial to many natural language applications. For example, for a macro-reader<sup>2</sup> that gathers 'background/common-sense' facts about entities from a large collection of input text, it is important for the reader to automatically identify temporal changes that occur to the entities since it will motivate a time-aware and hence a more precise micro-reader.

Another possible application of change identification is event extraction. This is a difficult problem in information extraction because unlike other named-entities, at the words/sentences (surface-) level, it is difficult to assign a precise label to an event. Furthermore, an event usually involves multiple entities and multiple relations and there are multiple ways to express the same event. By modeling an event at a meta-level as a sequence of topics change over time, we can extract more similar events from a collection of texts when a similar sequence of topics change is identified.

Tracking change over time in the frequency of entity mentions in a collection of texts can indicate that there is a change happening to the entity when its frequency changes. However, a change to an entity may not always be accompanied by a change in its frequency. Furthermore, knowing that there is a frequency change does not give further insight to the nature of the change or what causes it.

In this paper, we propose to conduct a deeper analysis than frequency change, by learning about the nature of the change itself from the change in the words that co-occur with the entity over time. In the next section we describe related works in this area and describe our proposed approach in Section 3. We describe our experimental setting in Section 4 and give detailed analysis of our experiment results in Section 5. We conclude with future works in Section 6.

<sup>1</sup><http://www.ruf.rice.edu/~kemmer/Words04/meaning>

<sup>2</sup>NELL: <http://rtw.ml.cmu.edu/rtw>

## 2. RELATED WORKS

A quantitative study on cultural trends: ‘culturomics’ that focuses on linguistic and cultural phenomena was done with the computational analysis of Google Books Ngram dataset [3]. By studying usage frequency over time of the n-grams that represent entities of interest, they study linguistic changes, e.g. lexicon and grammar changes; and cultural phenomena, e.g. how people and events are remembered.

In terms of grammatical trends, they study the frequency change in English irregular verbs. They find that high-frequency irregulars such as ‘found’ are less likely to be replaced by their regular forms than lower frequency irregulars such as ‘dwelt’. In terms of cultural phenomena, they track fame by measuring the frequency of a famous person’s name. They find that famous people in different periods of time follow the same kind of trajectories: pre-celebrity period, rapid rise to prominence, a peak and a slow decline in fame. By following such trajectories, one might be able to identify status of famous people in different periods of time. They conclude by highlighting that culturomics’ challenge lies in the interpretation of evidence (in this case, frequency) provided by the large Google dataset. Our paper intends to complement their interpretation further by using not only frequency but also actual words and word co-occurrences in the n-grams to study linguistic and cultural change. For example, the authors hypothesize that the change in the word ‘speed’ from its irregular form: ‘sped’ to its regular form: ‘speded’ might have been caused by the shift in meaning from ‘to move rapidly’ towards ‘to exceed the legal limit’, i.e. ‘to speed up’ [3]. The purpose of our paper is precisely to enable us to confirm or refute such hypothesis.

In another work, the authors study temporal changes in public opinion in tweets [1]. They identify a change (a breakpoint) in public opinion when there are both emotion pattern and word pattern change (measured with cosine and Jaccard similarity changes) in tweets from one point in time to another. If a pattern in a time period is less similar to a pattern in the preceding period but more similar to a pattern in the following period, a breakpoint is reported and events that cause this change are described by choosing keywords from all tweets in that period of change. Unlike this paper, our paper uses a change in topic surrounding an entity as an indication of change to the entity thus directly finds keywords that describe the change from the topics identified.

## 3. APPROACH

### 3.1 Pre-processing Step

For a given word  $w$  that represents an entity of interest (e.g. ‘awful’, ‘mouse’), we retrieve all 5-grams that contain  $w$  (case insensitive) as its third word. We are interested in the two words before  $w$  and the two words after  $w$  in the 5-grams. For each of these words  $v$  that co-occur with  $w$  (discarding stop words), we compute match count: the total number of co-occurrence of  $v$  with  $w$  in a year  $t$  ( $Match_t(v)$ ), and volume count: the total number of volumes where  $v$  co-occur with  $w$  in that year ( $Vol_t(v)$ ). Based on these counts, we compute *tfidf* score of  $v$  in the year  $t$  as:

$$tf_t(v) = Match_t(v) / \sum_x Match_t(x)$$

$$idf_t(v) = VolCount(t) / Vol_t(v)$$

$$tfidf_t(v) = tf_t(v) * \log(idf_t(v))$$

where  $x$  is all words that co-occur with  $w$  in year  $t$  and  $|x|$  is the number of distinct words that co-occur with  $w$  in

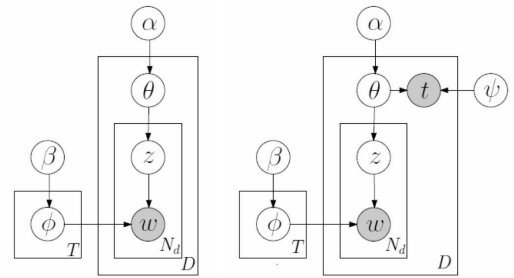


Figure 1: LDA (left) and TOT(right) models

year  $t$ .  $VolCount(t)$  is the number of volumes in year  $t$ . Since Google Books dataset does not provide the number of volumes published in a year, we approximate it by dividing the total number of words in the year by the average words-per-volume in the year:

$$VolCount(t) \approx totalWords_t / aveWordsPerVolume_t \\ \approx \sum_x Match_t(x) / [(1/|x|) * \sum_x Match_t(x) / Vol_t(x)]$$

We treat each year  $t$  as a document containing words that have co-occurred with  $w$ , our entity of interest, in that year. We compute *tfidf* scores of the words in each document. We cluster these ‘documents’ using Topics-Over-Time and  $k$ -means clustering.

### 3.2 Topics-Over-Time (TOT)

Latent Dirichlet Allocation (LDA) [2] is a generative model that represents each document in a corpus as a mixture of topics. Each topic is a distribution over all the words in the corpus. A representation of LDA is given in Figure 1.

Topics-Over-Time [5] is an LDA like topic model that captures not only the low dimensional structure of data, but also captures how that structure changes over time. TOT is a time-dependent extension of LDA which explicitly models time jointly with word co-occurrence patterns. The structure of TOT (Figure 1) is similar to LDA’s with only one difference in that TOT also parameterizes a continuous Beta distribution over time associated with each topic, and these topics generates both words and timestamps.

### 3.3 K-means Clustering

In  $k$ -means clustering, the documents are partitioned into  $k$  clusters. Each document belongs to the cluster with the nearest centroid. In this paper, we treat each document  $d$  as an  $n$ -dimensional vector where  $d_i$  is the *tfidf* score of the word  $i$  in  $d$ .  $n$  is the size of the vocabulary of all words that have ever co-occurred with  $w$  over time.

The algorithm finds  $k$  clusters in the data via an iterative procedure: (1) place  $k$  points in the  $n$ -dimensional space to represent the initial centroids, (2) assign each document to the cluster that has the closest centroid, (3) when all documents have been assigned to clusters, recalculate the positions of the  $k$  centroids, (4) repeat from step 2 until the centroids no longer move. At the end of this iterative procedure, each document (i.e. each year) will be assigned to a cluster. We indicate that there is a topic change when two consecutive years are assigned to different clusters. We also obtain  $k$   $n$ -dimensional vector centroids. The topic of each cluster is described by the top words (words with highest *tfidf* scores in the centroid) of the cluster.

## 4. EXPERIMENTAL SETTING

### 4.1 Dataset

We conduct an experiment of our proposed approach on Google Books Ngram dataset. This dataset is a corpus of about 5 million digitized books, roughly 4% of all books ever published. The books were scanned and their texts were digitized using optical character recognition (OCR) [3]. The resulting corpus contains over 500 billion words, in English, French, Spanish, German, Chinese, Russian and Hebrew. To make the release of the data possible, due to copyright constraints, the data is released in the form of ‘1-gram’ to ‘5-gram’ and how often these n-grams were used over time; their match counts, page counts and volume counts.

An interesting feature of this dataset is firstly, its time coverage. The books range from as early as the 1500s and as late as 2008, providing a great playground to capture changes that occur slowly, as is often the case with linguistic changes [3]. Secondly, the dataset is special in that it contains only n-grams. There is no notion of documents or sentence boundaries; what is available are in the form of bits and piece of the original books.

### 4.2 TOT and K-means Clustering

For each entity of interest, we treat each year as a document. The document contains words that have co-occurred with the entity in that year. For TOT, we use the match counts of the words for clustering while for  $k$ -means, we use the  $tfidf$  scores of the words.

We use  $k = 20$  for  $k$ -means clustering. After  $k$ -means clustering, each ‘document’ (year) is assigned to a cluster. We pick the top 10 words (with highest  $tfidf$  scores) from each cluster’s centroid to represent the cluster. However, some clusters are very small, consisting only of one year while some others consist of years that are not consecutive to one another. To focus on more persistent topics and their changes, we pick only clusters that exhibit topic consistency, i.e. clusters that have a consecutive run of four years or more to represent here.

In each year, we compute the topic density of each consistent cluster in that year as the proportion of words in that year belonging to the top 10 words of the cluster. We plot these topic densities over years and use locally weighted scatter plot smoothing (LOESS) in R to smooth the plot <sup>3</sup>.

## 5. EXPERIMENTS

For the experiments, we choose words that may have shown different behaviors over time. These behaviors can be either a total or a partial change in the meaning or a change in the network structure of the word. In this section we analyze some words that exhibit different behaviors using our proposed methods.

### 5.1 Change in Meaning

The first group of words that we analyzed is the words in which the semantic meaning is changing over time. Words may get additional meanings and these meanings may suppress the original meaning after a while. For instance, originally ‘gay’ word is used as an adjective for happiness, cheerfulness, pleasant etc. Around 1970s with the homosexuality

<sup>3</sup><http://stat.ethz.ch/R-manual/R-patched/library/stats/html/scatter.smooth.html>

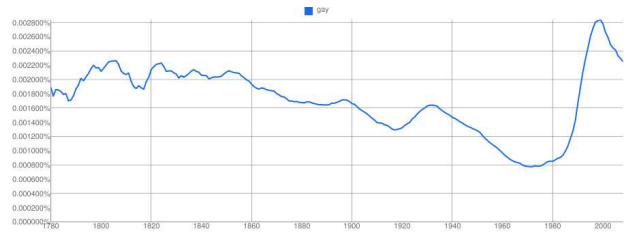


Figure 2: Google Books Ngram Viewer for the word ‘gay’

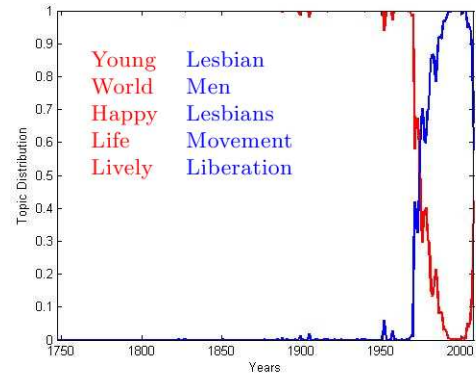


Figure 3: TOT with 2 topics for ‘gay’

movements the word is started to be used as a noun with the meaning homosexual man. Over the years the word has been used more and more with the latter meaning and now when it has been used the first meaning that comes to people’s mind is the homosexual man.

When the frequency of this word is analyzed with Google Books Ngram Viewer (Figure 2), a decrease followed by an increase can be seen around 1970s-1980s. Deeper analysis is required to understand this change. When TOT is applied (Figure 3) over the co-occurring words with 2 topics, the change of topics can be seen to occur around 1970s when the word gay started to be used to refer to the homosexual man. The top words for each cluster are a good indicator of the original and the latter meaning of the word.

Furthermore a similar graph can be obtained from  $k$ -means clustering (Figure 4). In this graph we can see that there is a break between two clusters in 1975. In order to see this change in the words co-occurrence network, we looked at the n-grams of the top 5 words from each cluster for time  $t$  (1975),  $t-1$  (1974),  $t-20$  (1954) and  $t+20$  (1995). In the graph (Figure 5) each word is a node and the edge between nodes is weighted by the number of n-grams in which these words co-occur together. In 1954, the node gay is mostly connected with the nodes young, life etc. but as the time goes on nodes such as liberation, lesbian get connected to gay and their weights increases over time. With these graphs we can easily see the change in the word usage over the years.

Another word in which the meaning is changing over time is awful. As mentioned in the introduction, awful originally meant full of awe, inspiring, which later became something very bad. When we apply TOT and  $k$ -means (Figure 6) we did not get any meaningful topics or clusters. The main rea-

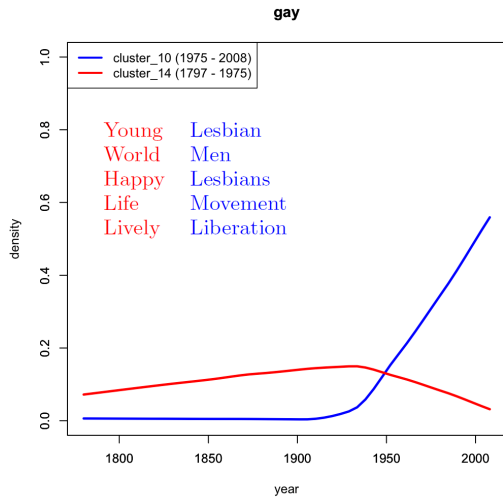


Figure 4: Most consistent clusters in k-means for ‘gay’

son for this is that awful is an adjective which can be used with any noun phrase. Sometimes it can be even used as an adverb. Therefore it is hard to get a meaningful cluster from all these different words. When we look at the words co-occurrence networks (Figure 7) produced from the n-gram co-occurrences we can see that the top words in each cluster are only connected with the word awful but there is no connection among themselves which means that they do not appear in the same n-grams and therefore they do not belong to the same topic. For adjective words like awful it is hard to see the meaning change by looking into topic models or clusters simply because the words that co-occur with it are very different from one another and hence cannot be clustered into meaningful topics.

## 5.2 Getting Additional Meaning

Another group of words that also change semantically are words that get additional meanings or used for other named entities over time. The difference between these words from the first group of words is that their original meaning is still widely used, but additional uses of the word are introduced over time. Within the words co-occurrence graph this can be seen as emergence of new sub graphs with new nodes.

An example in this group is the word mouse. It is originally used to refer to a small long-tailed animal but after 1970s it is also used to refer to pointing device used to control cursor movement. Similar to the gay word, we can see an increase in the frequency of mouse around 1970s. Applying TOT and k-means give topic models and clusters in which the new meaning of mouse can be seen clearly. In TOT the break is occurring around 1980s-1990s when the mouse device started to be used with personal computers. In k-means, we display 4 clusters in Figure 8. The first two clusters are similar both in density and top words, therefore decreasing the k in k-means (currently 20) may result in the merging of these two clusters. With k-means we can see the new meaning of mouse in the third cluster. The last cluster is from the original meaning of mouse as animal but used in the context of scientific research as test animal.

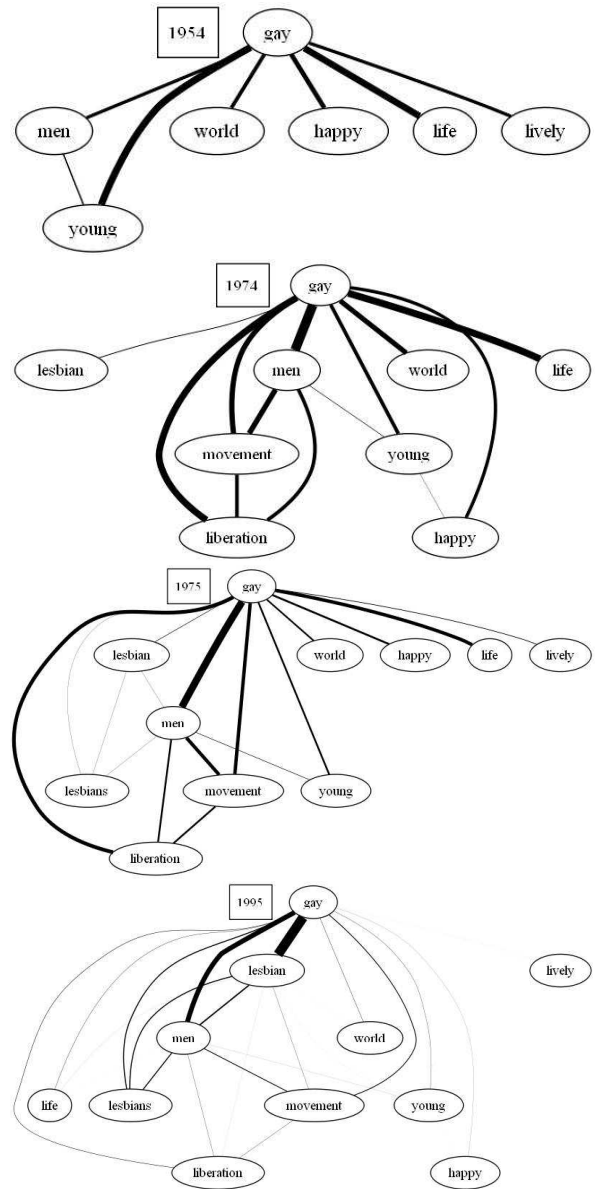


Figure 5: Words co-occurrence networks for ‘gay’

Another example in this category is the word king. King is originally used to refer to male sovereign or monarch. Around 1950s as Martin Luther King gained popularity, the word king is used many times in the context of his surname. In TOT (Figure 9), the top five words are similar in both clusters. These words are coming from the original meaning of king but words like Martin and Luther also appear in the second topic model as new words. Figure 10 shows the change in the words co-occurrence network over time. Around 1965 with the popularity of Martin Luther, a sub graph emerged which is not connected to other original top words. One thing to note in the graphs is that the weights of these sub graphs are lower than the weights of words used for the original meaning of the word. Over time these sub graphs may disappear while the nodes from the original meaning remain.

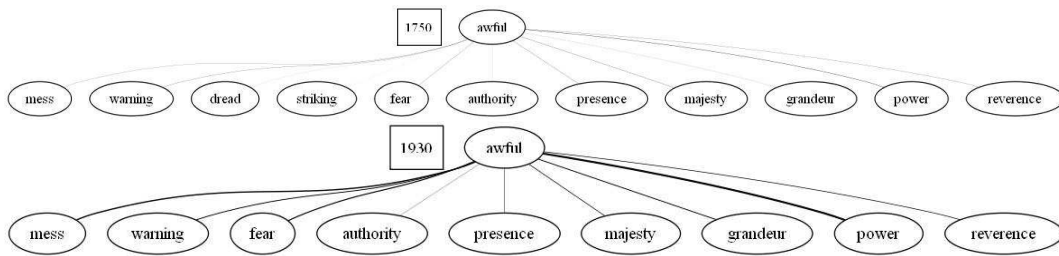


Figure 7: Words co-occurrence networks for ‘awful’

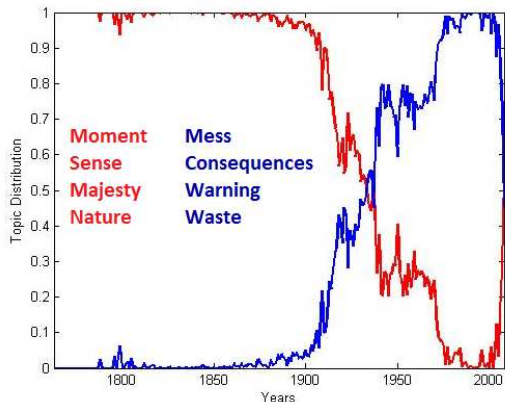


Figure 6: TOT with 2 topics for ‘awful’

### 5.3 Unchanged

In previous groups of words, we see words getting new additional meanings or losing their original meaning over time. However this is not the case for most words. Words such as woman and god are mostly unchanged. They are used with the same set of words for centuries. In Figure 11, k-means results show that the resulting clusters are very similar both in terms of the top words and the cluster densities.

## 6. CONCLUSION AND FUTURE WORK

Identifying changes that occur to an entity over time is important for many applications such as macro and micro reading and event extraction. Analyzing changes in usage frequency of an entity is often not enough to identify these changes, let alone describe or understand the nature of the changes. Some changes such as linguistic changes occur slowly and gradually over time. Such change may not always be accompanied by an identifiable surge/decline in usage frequency.

In this paper, we contribute methods that attempt not only to automatically identify when (at which year or at which period of time) changes occur but also what changes occur: i.e. what topics are in transition. The method achieves this by analyzing changes that happen to an entity based on changes observed in the words surrounding the entity over time.

For the experiment, Google Books Ngram dataset proves an excellent choice for a dataset as it provides coherent bits and pieces of information that dates back to as early as the 1500s. Hence the dataset contains useful information on changes that may have occurred slowly and gradually over

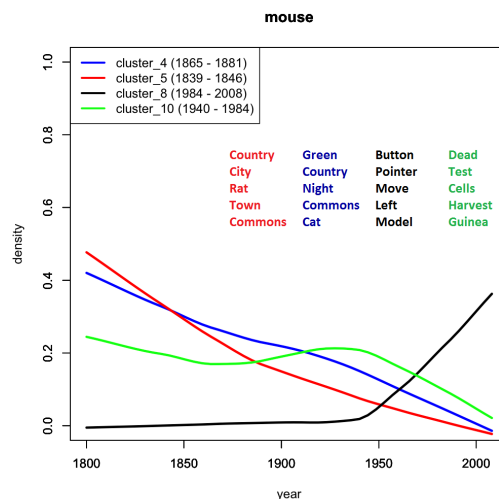


Figure 8: Most consistent clusters in k-means for ‘mouse’

time. From our experiments on this dataset, we find that our clustering methods (TOT and k-means) are both effective in uncovering hidden topics (clusters) and identifying the year for which topics change. Using these clustering methods, we are able to identify the exact period in which a change occurs. For all the entities we test, the period that our method identifies coincides precisely with events in the same period that correspond to the change. For example, for the word Iran, our k-means approach is able to identify the country’s change from monarchy to an Islamic republic with a new cluster related to words such as republic and revolution emerging after 1978 (1979 is the year of the Islamic revolution). Another similar example can be seen with the word Kennedy. K-means was able to identify the John F. Kennedy the senator before the election (one cluster) and him being the president after the election (another cluster). The break between the two clusters is at 1961, the exact year Kennedy was elected. Similar state changes are observed for word Clinton, from governor to president. The breakpoint occurs at 1993, the exact year he was elected. Seeing similar clusters and similar changes between the two presidents brings up the question whether this method can be applied to other presidents as well, to time scope their period of presidency, and whether we can use similar state change on these entities to indicate election event, for example.

Another advantage of our approach is that by using the

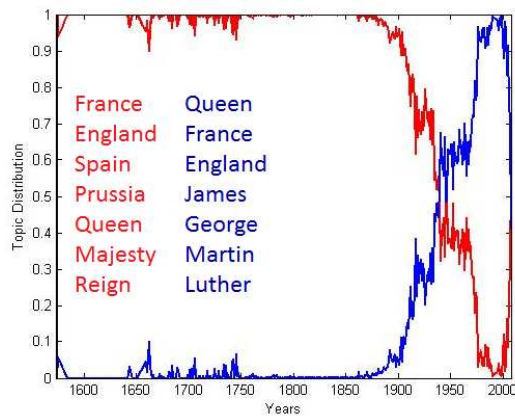


Figure 9: TOT with 2 topics for ‘king’

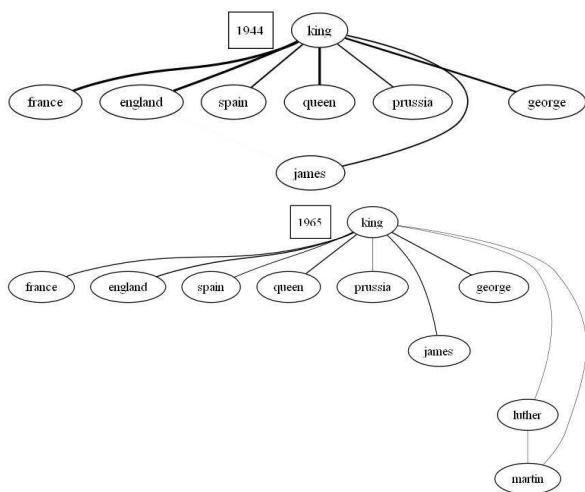


Figure 10: Words co-occurrence networks for ‘king’

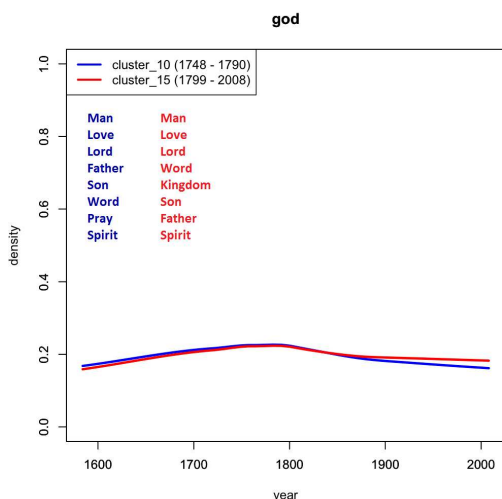


Figure 11: Most consistent clusters in k-means for ‘god’

top words of the topics identified, we are able to automatically describe the changes that occur. In future, we would like not only to describe, but to also understand the nature of the change further. This can be achieved by digging further into the words co-occurrence network and how it evolves over time: to see whether topics change can be explained in terms of the change in co-occurrence links between the words that surround an entity. Indeed, in our preliminary study, we observe that some links appear while some others disappear from this words co-occurrence network. Based on this observation, could the word awful, at some point in the past, have a co-occurrence link with a word that can both refer to something majestic and something terrifying; hence kick starting its gradual shift to its current ‘very bad’ meaning? Methods that study evolution of dynamic networks such as temporal exponential random graph model (ERGM) [4] might be used to study this network of co-occurrence over time.

Another potential for this work is to build a web service that offers this change-over-time analysis for any input word. To do so, searching for 5-grams that co-occur with the input word from Google Books Ngram can be a bottleneck due to the large size of this dataset. However, advanced techniques of indexing can be used to make this search for 5-grams process more efficient.

## 7. ACKNOWLEDGMENTS

The authors would like to thank William W. Cohen for his valuable feedback and helpful discussions. This work was supported in part by Carnegie Mellon University’s iLab and Center for the Future of Work and in part by DARPA (under contract numbers FA8750-08-1-0009 and AF8750-09-C-0179) and by Google and Fulbright. We also gratefully acknowledge the contributions of our colleagues on the CMU NELL project. Any opinions, findings, conclusions and recommendations expressed in this paper are the authors’ and do not necessarily reflect those of the sponsors. Finally, we thank the anonymous reviewers for their helpful comments.

## 8. REFERENCES

- [1] C. G. Akcora, M. A. Bayir, M. Demirbas, and H. Ferhatosmanoglu. Identifying breakpoints in public opinion. In *Proceedings of the First Workshop on Social Media Analytics, SOMA '10*, pages 62–66, New York, NY, USA, 2010. ACM.
- [2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.
- [3] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, January 14 2011.
- [4] G. L. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p\*) models for social networks. *Social Networks*, Jan. 2007.
- [5] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '06*, pages 424–433, New York, NY, USA, 2006. ACM.