The University of Iowa


Report for Qualifying Exam

# Understanding the flow of Temporal Context Shift in Data Voids

*Author:*
Manisha Keim
manisha-keim@uiowa.edu

*Advisor:*
Rishab Nithyanand
rishab-nithyanand@uiowa.edu

Department of Computer Science
July 11, 2024

# Contents

# 1 Abstract

In the digital age, online information seekers frequently encounter search results laden with misinformation, biased narratives, and conspiracy theories. This exposure can potentially lead users to accept and propagate false information. However, this phenomenon is not universal across all search queries. Certain searches, known as *data voids* are particularly susceptible to these issues.

*Data voids* occur when searching for terms that yield limited, non-existent, or highly problematic relevant information. Unlike common searches that produce abundant data, queries falling into data voids may return no results or present irrelevant and often inaccurate information. Data voids can be categorized into two main types: newly coined terms with no established information base, and existing terms whose meanings have evolved over time. This report focuses on the latter, examining data voids where context drift occurs. The following papers has been selected to explore this phenomenon.

- Statistically Significant Detection of Linguistic Change. ACM WWW 2014 [3]
- Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. ACL 2016 [1]
- Specious Sites: Tracking the Spread and Sway of Spurious News Stories at Scale. IEEE S&P 2024 [2]

# 2 Introduction

## 2.1 What is a Data Void?

## 2.2 Example of a data void?(Migrant Caravan)

## 2.3 Lifecycle of data void (from long tail to spike)

## 2.4 Types of Data Void

## 2.5 Technical Challenge

analyzing data void involves context drift over historical time. (How context drift happens with data void.)

# 3 Related Work

The evolution of word meanings over time has been a subject of significant interest. Researchers have employed a diverse array of methodologies, including word embeddings, neural language models, and diachronic analysis, to track and examine semantic shifts across historical periods. Much of this research has utilized the Google Books Ngram corpus, a vast collection spanning from 1900 to 2009, encompassing over 500 billion books in seven languages. This dataset provides n-grams with corresponding yearly occurrences and frequencies.

Diverse categories of semantic shifts exist, such as semantic, syntactic, or as Hamilton et al. (2016) noted, can be driven by cultural or linguistic factors. Certain studies concentrate on semantic changes, while others explore the broader evolution that words undergo, with some works focus on transformations in word usage and meaning.

Gulordava and Baroni (2011) introduce a novel method for finding semantic changes in words over time. They focus on detecting shifts between the 1960s and 1990s using the Google Books Ngram corpus. Their approach relies on comparing the similarity of a word's surrounding words in these two periods. They define context by considering the words that appear within two positions (a window size of 2) before and after the target word, essentially analyzing bigram statistics.

Kim et al. (2014) propose a distinct approach that leverages the power of neural language models. By analyzing how the model represents words at different points in time, they can capture how word meaning evolves. Unlike previous methods that focus on individual words or word frequency, this technique allows them to pinpoint not only words with shifting meanings but also the specific years when those changes occurred. Their approach goes beyond simple co-occurrence by using a window size of four and comparing the cosine similarity between the same word's vector representations across different time periods. This enables them to detect even subtle changes in how words are used. Additionally, by plotting the time series of a word's distance to its neighboring words in the model's vector space, they can visualize the exact period during which the semantic shift occurred. They identified words like 'cell' and 'gay' that have undergone significant semantic shifts over the analyzed period.

Kulkarni et al. (2015) introduced method for significant detection of lin-

guistic change, which laid out the foundation that utilized word embeddings for detecting semantic shifts. Building on the concept of distributed representations introduced by Hinton, they map symbolic data (words) into a continuous vector space where words with similar meanings are positioned close together. Their approach incorporates three methods to analyze these embeddings: frequency-based, syntactic, and distributional.

Furthering the research on word embeddings, Liao and Cheng (2016) compared the use of word embeddings with a clustering algorithm called DBSCAN to analyze semantic change. Their work investigated which method was more effective for identifying these semantic shifts.

Hamilton et al. (2016) address the challenge of distinguishing between cultural shifts and linguistic drift, both of which can contribute to semantic change. They propose two distinct computational measures based on word embeddings. One measure captures global semantic shifts, while the other, captures localized changes that reflect cultural influences. By analyzing these two measures together, researchers can know the causes behind semantic change.

Expanding on this concept, their subsequent work titled "*Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change*" (2016) investigates recurring patterns in semantic change across different languages. They identify two key principles: *the law of conformity*, which suggests that a word's frequency is inversely proportional to the rate of its semantic change. In other words, words that are used more frequently tend to experience slower shifts in meaning. *The law of innovation*, on the other hand, proposes that words with multiple meanings are more likely to undergo semantic transformations over time.

Furthermore, Kutuzov et al. (2018) conducted a survey on diachronic word embeddings and semantic shifts, consolidating the existing academic research in this domain. Their work provides a comprehensive overview of the methodologies and findings related to tracking semantic changes over time using computational techniques.

Giulianelli (2020) explores the computational analysis of lexical semantic change using contextualized word representations. Contextualized word representations, such as those derived from models like BERT or ELMo takes into account the context in which words are used. This enables a more detailed analysis of how semantic changes occur based on specific word contexts.

The impact of significant events like the COVID-19 pandemic on language and semantic change has also been a subject of study. Laurino et al. (2023)

explores tracking fast semantic changes through a large-scale word association task, aiming to understand how the collective mental lexicon evolves in response to such global events. Their research highlights the dynamic nature of language and how it incorporates new senses. For instance, words like "quarantine," "mask," and "social distancing" took on new and prominent meanings in everyday conversation.

Several studies from our literature review provide valuable insights for addressing data voids in our analysis.

- The first being the work by Kulkarni et al. which introduces us to time series construction. Their distributional methods focuses on finding subtle semantic shits to determine the context where a word in used. This concept aligns perfectly with our goal of uncovering shifts in word usage when encountering data voids.

- Second paper by Hamilton et al. introduces us to word emebeddings alignment. Since we want to compare word vectors from different time periods, vectors should be aligned in same coordinate axes. After aligning the embeddings for individual time periods, we can use the aligned word vectors to compute the semantic displacement that a word has undergone during a certain time-period.

- Third one by Laurino et al. investigates the impact of the COVID-19 pandemic on word meaning. One key finding from their work is that words directly related to the pandemic exhibited a greater difference in semantic similarity between pre-pandemic and pandemic time periods. This suggests that these words underwent a more significant and rapid semantic shift compared to control words not associated with the pandemic. They also employ semantic similarity analysis to quantify the shifts in meaning for pandemic-related words and provides evidence that the COVID-19 pandemic acted as a catalyst for rapid semantic change.

# 4 Research Papers

## 4.1 Statistically Significant Detection of Linguistic Change

### 4.1.1 About

This research aims to quantify linguistic shifts in word meaning and usage across time, focusing on semantic or contextual changes. The study utilizes diverse datasets, including Twitter posts, Amazon product reviews, and Google Book Ngrams, to construct time series for individual words. Three distinct methods are employed for statistical modeling:

- frequency analysis to capture sudden changes in word usage,
- syntactical analysis examining each word's part-of-speech tag distribution, and
- distributional analysis investigating contextual cues through word co-occurrence statistics.

To determine the statistical significance of word changes over time, the authors are developing a change point detection algorithm. Additionally, the research incorporates time series data representing the percentage of search queries according to Google Trends, providing further insight into evolving language patterns and word usage shifts.

### 4.1.2 Research Questions

1. What methods can quantify the statistical relevance of observed changes in a word's usage across different time periods?
2. Given that a word's usage has changed, how can the precise moment or period of this shift be determined?

# References

[1] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change, 2018.

[2] Hans W. A. Hanley, Deepak Kumar, and Zakir Durumeric. Specious sites: Tracking the spread and sway of spurious news stories at scale, 2024.

[3] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change, 2014.