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 16, 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 [5]
- Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. ACL 2016 [3]
- The Pandemic in Words: Tracking Fast Semantic Changes via a Large-Scale Word Association Task. [7]

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

Tracking how word meaning change over time. The evolution of word meanings over time has been a subject of significant interest. Researchers employed a diverse array of methodologies, including *word embeddings* [5], *neural language models* [4], and *diachronic analysis* [3, 6], to track and examine semantic shifts across historical periods. Much of this research utilized the *Google Books Ngram corpus* [1–6, 8], 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.

Various type of linguistic changes. Words have the ability to undergo shifts in meaning over the course of time as a result of various linguistic mechanisms, including but not limited to *semantic drift* [1, 5], *syntactic alterations* [2, 5], *broadening* as opposed to *narrowing* of meaning, and the processes of *amelioration* versus *pejoration.* These transformations, as highlighted by Hamilton et al. (2016), could be driven by *cultural* or *linguistic* factors. Certain studies concentrate on semantic changes, while others explored the broader evolution that words underwent [4], with some works focus on transformations in word usage and meaning [1, 2, 5]. The causes of linguistic changes and the methodologies employed to identify them are outlined in Table 1.

Semantic Change. Semantic change refers to any change in the meaning(s) of a word over time or acquiring a new sense [1, 7]. Kutuzov et al. (2018) conducted a survey on 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.

- **Co-occurrence Counts:** Gulordava and Baroni (2011) quantify how frequently words appear within the same context. 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.
- **Neural Language Models:** Kim et al. (2014) suggest a new method using *neural language models* to track word meaning evolution. They analyze word vectors over time to capture changes. Their method compares word vectors using a window size of four from 1900–2009 to detect subtle changes in word usage. 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. Words like 'cell' and 'gay' show significant changes over time.
- **Word Embeddings:** Kulkarni et al. (2015) introduced method for detecting linguistic change with statistical significance that utilized *word embeddings*. Building on the concept of distributed representations proposed 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*, which are later discussed in detail in next section.(§4)
- **Clustering:** 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.
- **Cultural vs. Linguistic:** 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. Cultural changes like new technologies are related to local neighbourhood change. Examples: gay, virus, cell These words gained new meanings due to uses in community-specific vernacular (gay) or technological advances (virus, cell).

- **Diachronic Word Embeddings:** Expanding on the above concepts, their subsequent work investigates recurring patterns in semantic change across different languages. They propose 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.
- **BERT:** 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 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.
- **Semantic Similarity:** 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.

Syntactic Change. The syntactic functionality of a word can evolve by transitioning into a new part-of-speech (POS) category. Nouns, due to their inherent flexibility in meaning, exhibit a greater tendency to undergo these changes driven by cultural shifts. While are more likely to participate in gradual semantic changes that follow established linguistic patterns.

- **Acquiring a new POS:** Kulkarni et al. assigned part-of-speech (POS) tags to a large collection of text (corpus) and then calculated how likely different words were to appear in certain grammatical contexts (probability distributions). They used the word 'apple' as an example, which transitioned from being used as a common noun (e.g., a fruit) to a proper noun (referring to the Apple company) after the company's rise in the 1980s.
- **Nouns vs. Verbs:** Hamilton et al. (2016) highlight how cultural changes, often influenced by new technologies, are closely tied to transformations within local neighborhoods, particularly sensitive to shifts in nouns. On the other hand, linguistic changes, exemplified by subjectification, are more associated with global measures and are particularly responsive to variations in verbs. The evolution of words like 'actually', 'must', and 'promise' demonstrate these changes. For instance, 'must' has transitioned from expressing obligation to indicating necessity, showcasing a common pattern seen in modal verbs. Similarly, 'promise' shows how performative speech acts undergo significant pragmatic and subjectification-related changes over time.
- **Acquiring a new sense:** Words are complex entities with multiple senses that can evolve over time while their form remains constant, as discussed by (Laurino et al., 2023). Their research studies the overall mental lexicon of words, focusing on quantifying the drift of

Causes	Methods	Technical Method	Time Period	Example Papers
Semantic	Frequency	Co-occurence	1900-2009	[1]
		Word Usage		
		Cosine Distance		
	Distributional	Word Vectors	1900-2020	[2]
		NLM		
		Cosine Similarity		
		Aligning Embeddings		
		DBSCAN		
		BERT		
Syntactic	POS	JS Divergence	1900-2000	[3]
		KL Divergence		
		Local Neighborhood		

Table 1. Various types of linguistic change and method used to detect them.

word concepts. The study highlights how most words possess a range of meanings that can be added, removed, or modified across different temporal contexts.

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?

4.1.3 Methodology.

4.1.4 Results.

4.2 Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change

4.2.1 About.

4.2.2 Research Questions.

4.2.3 Methodology.

4.2.4 Results.

4.3 Specious Sites: Tracking the Spread and Sway of Spurious News Stories at Scale

4.3.1 About.

4.3.2 Research Questions.

4.3.3 Methodology.

4.3.4 Results.

References

[1] Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus.. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, Sebastian Pado and Yves Peirsman (Eds.). Association for Computational Linguistics, Edinburgh, UK, 67–71. https://aclanthology.org/W11-2508

[2] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 2116–2121. https://doi.org/10.18653/v1/D16-1229

[3] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Katrin Erk and Noah A. Smith (Eds.). Association for Computational Linguistics, Berlin, Germany, 1489–1501. https://doi.org/10.18653/v1/P16-1141

[4] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, Cristian Danescu-Niculescu-Mizil, Jacob Eisenstein, Kathleen McKeown, and Noah A. Smith (Eds.). Association for Computational Linguistics, Baltimore, MD, USA, 61–65. https://doi.org/10.3115/v1/W14-2517

[5] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2014. Statistically Significant Detection of Linguistic Change. arXiv:1411.3315 [cs.CL] https://arxiv.org/abs/1411.3315

[6] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1384–1397. https://aclanthology.org/C18-1117

[7] Julieta Laurino, Simon De Deyne, Álvaro Cabana, and Laura Kaczer. 2023. The Pandemic in Words: Tracking Fast Semantic Changes via a Large-Scale Word Association Task. *Open Mind* 7 (06 2023), 221–239. https://doi.org/10.1162/opmi_a_00081 arXiv:https://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi_a_00081/2133848/opmi_a_00081.pdf

[8] Xuanyi Liao and Guang Cheng. 2016. Analysing the Semantic Change Based on Word Embedding. In *Natural Language Understanding and Intelligent Applications*, Chin-Yew Lin, Nianwen Xue, Dongyan Zhao, Xuanjing Huang, and Yansong Feng (Eds.). Springer International Publishing, Cham, 213–223.