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 22, 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 [6]
- 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. [8]

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 meanings 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* [6], *neural language models* [5], and *diachronic analysis* [3, 7], to track and examine semantic shifts across historical periods. Much of this research utilized the *Google Books Ngram corpus* [1–3, 5–7, 9], 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.

Words can change meaning over time through several linguistic processes. 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, 6]
  - Refers to the evolution of word *meanings* over time.
  - Example: The word 'gay' originally meant 'happy' or 'carefree', but now it predominantly means 'homosexual'.
- *syntactic alterations* [2, 6**?**]
  - Syntax focuses on the *structure* of language. This type of change pertains to modifications in the arrangement of words in sentences.
  - Example: The word 'apple', 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.
- *broadening* [2, 6**?**]
  - A word's meaning becomes more general than its earlier meaning, also known as *generalization.*
  - Example: 'Holiday' originally meant a religious festival, but now it can refer to any day of celebration or time off.
- *narrowing* [6**?**]
  - A word's meaning becomes more specific than its earlier meaning, also known as *specialization.*
  - Example: 'Literally' used to mean 'figuratively' or 'symbolically'. Now, it is used to emphasize the truthfulness of a statement.
- *amelioration*
  - A word takes on a more *positive* meaning over time.
  - Example: The word 'thrifty' once meant 'cheap' but now suggests responsible use of resources.
- *pejoration* [1, 8]
  - A word takes on a more *negative* meaning over time.
  - Example: The word 'awful' meant 'full of awe' which transitioned to 'terrible' or 'appalling'.

Hamilton et al. (2016) highlighted that *cultural* or *linguistic* factors could have driven these transformations. Certain studies concentrated on semantic changes, some worked on syntactic changes, [2, 6, 8] and others explored the broader evolution that words underwent [1, 3, 5, 6]. Table 1 outlines various types of linguistic changes and the approaches employed to identify them.

Building on the extensive research on *semantic* change. Semantic change refers to any change in the meaning(s) of a word over time or acquiring a new sense [1, 8]. 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.

Linguistic Change	Approach	TCD	Example		
			Word	Former Usage	New Usage
Semantic	**FREQUENCY [1, 3, 6]**				
	Log Ratio	1960 & 1990	disk	-	-
	Word Frequency	1900–2005	bitch	female dog	slang
	DISTRIBUTIONAL [1–3, 5, 6, 8? , 9]				
	Local Mutual Information	1960 & 1990	sleep	deep sleep	sleep disorder
	Continuous Word Embeddings	1900–2009	cell	closet, dungeon	phone, cordless
	Change Point Algorithm	1900–2005	gay	cheerful, dapper	lesbian, homosexual
	Clustering (DBSCAN)	1900–2000	mouse	mice, rat	cursor, pointer
	Cultural and linguistic Drift	1800–2000	virus	infected with the virus	spreading computer virus
	Diachronic Word Embeddings	1800–1999	awful	full of awe	terrible or appalling
	Contextual Word Embeddings	1910 – 2009	tenure	short term leases, insecurity of tenure	tenure of office, employment and tenure
	Semantic Similarity Analysis	2014–2022	immunity	politics ('legislator', 'representative').	health-related ('prevention', 'AIDS')
	PART OF SPEECH [2, 6]				
Syntactic	POS Tags	1900–2000	windows	doors and windows of a house	Microsoft Windows
	Mixed-model Regressions	1800–2000	actually	originally, nominally	presumed, believe

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

- **Investigating Semantic Shifts Through Word Contexts.** Gulordava and Baroni (2011) detected semantic change by focusing on words used in the 1960s and 1990s. They compared the similarity of the surrounding words (words that co-occur with the target word) in these two time periods.

  - To assess similarity, the researchers calculated the *Local Mutual Information (LMI)* score between the central word and its surrounding words. A low LMI score between the target word and its surrounding words across the two time periods indicated a potential semantic shift.
  - The study demonstrated that their distributional similarity models were effective in capturing *cultural shifts* in word meaning. For example, they found that the word 'sleep' acquired more negative connotations related to sleep disorders when comparing its contexts in the 1960s to those in the 1990s.

- **Tracking Meaning Evolution Through Neural Nets.** Kim et al. (2014) focused on analyzing how word meanings evolved between 1900 and 2009.
  - They developed the first method that employed *prediction-based model* to trace semantic shifts. This involved training a model on data from a specific year y_i and then using the resulting word vectors as the starting point for training the model on the next year's data $y_i + 1$.
  - Their method analyzed *global shifts* in a word's vector semantics. Additionally, by plotting the time series of a word's distance to its neighboring words in the model's vector space, they visualized the period during which the semantic shift occurred.
  - They demonstrated this for the word 'cell' compared to its early neighbors, 'closet' and 'dungeon,' and the more recent neighbors, 'phone' and 'cordless.' For 'cell,' the identified period of change (1985-2009), which interestingly coincides with the introduction and widespread adoption of cell phones by the public.

- **Pinpointing Significant Shifts Statistically.** Kulkarni et al. (2015) proposes a novel computational approach to identify and quantify the semantic and usage changes in words across various media (new products, movies and books).
  - Building on the concept of distributed representations proposed by Hinton [4], they map words into a continuous vector space where words with similar meanings are positioned close together.
  - The approach hinges on constructing *property time series* for each word. They propose three methods for constructing these time series (§5).:
    * **Frequency:** This method analyzes changes in a word's overall frequency of use, assuming a sudden shift in frequency might indicate a semantic shift.
    * **Syntactic:** This method examines the distribution of a word's part-of-speech tags (e.g., noun, verb) across different time periods, aiming to capture changes in how the word functions grammatically.
    * **Distributional:** This approach leverages word embeddings which are created for each year, and then alignment is done to represent them in joint embedding space, and it's utilized to construct distributional time series for a word's displacement.
  - Finally, they employed statistically sound *change point detection algorithm* to identify significant moments in these time series, pinpointing the periods where word meaning or usage likely underwent a shift. Their results indicated that computational methods for the detection of semantic shifts can be robustly applied to time spans less than a decade.

- **Word Semantic Modelling of Polysemant.** Liao and Cheng (2016) explored the semantic changes of words. Their approach built on the understanding that word meaning is closely tied to its context. When a word's meaning changes, the surrounding words used with it *(context words)* are likely to change as well. They focused on *polysemous* words (words with multiple meanings) and aimed to detect when new meanings emerged.

  - They used the *skip-gram architecture with negative sampling* [10] to obtain word embeddings. This technique helped in capturing the contextual meaning of words by representing them in a continuous vector space.
  - They employed *DBSCAN* to group word embeddings. DBSCAN helped in identifying clusters of similar word contexts and distinguishing them from noise, which could indicate semantic changes.
  - To find similar words, they used a nearest neighbor search method called *Random Project Forest*. This method helped in identifying words that are contextually similar to a given word.
  - Finally, they compared the stability of *similar words* with the stability of their *context words*.
- **Distinguishing Cultural Shifts from Linguistic Drift.** Hamilton et al. (2016) addressed the challenge of distinguishing between *cultural shifts* and *linguistic drift*, both of which can contribute to semantic change.
  - They proposed two distinct measures based on distributional semantics to distinguish between these two types of semantic change:
    * **Local Neighborhood Measure:** This measure focused on the closest neighbors (most similar words) in a word's embedding. A drastic shift in these nearest neighbors suggests a significant change in core meaning, potentially driven by a *cultural shift* (e.g., 'gay' changing from carefree to referring to homosexuality).
    * **Global Measure:** This measure considered the overall distribution of a word's surrounding words in a larger context window. Gradual changes in this broader distribution are more likely to reflect *linguistic drift*, the natural evolution of language due to regular processes (e.g., 'promise' expanding from a declaration to also suggesting a likelihood).
  - Prior research often treated semantic change as a single phenomenon. Hamilton et al. offered a novel approach by distinguishing between cultural shifts and linguistic drift using their two measures.
- **Diachronic Analysis on Historical Data:** The primary goal of Hamilton et al. (2016) is to track semantic changes and understand how the meanings of words shift in different historical contexts.
  - They created word embeddings for different time periods using both the PPMI matrix with SVD and the SGNS model. These embeddings were generated from historical text corpora to capture the contextual usage of words in each period.
  - After creating the embeddings, they aligned the embeddings. Once the embeddings are aligned, they calculate the *semantic displacement* of a word. This essentially measured how much a word's vector representation has moved in the embedding space between two time periods.
  - The study demonstrates that their method effectively identified semantic change in words and uncovered two statistical 'laws' of semantic change:
    * **Law of Conformity**, which suggests that the rate of semantic change is inversely proportional to a word's frequency. High-frequency words tend to change meaning more slowly.
    * **Law of Innovation**, on the other hand, proposes that words with multiple meanings are more likely to undergo semantic transformations over time.
  - They leveraged the rich information within word embeddings to quantify the degree of semantic change (semantic displacement) and identified potential patterns.

- **Semantic Shifts with Contextual Embeddings:** Giulianelli et al. (2020) explored the phenomenon of lexical semantic change using an unsupervised approach using *BERT*.
  - They utilized the *BERT* model to generate contextual word embeddings, which captured the meaning of a word based on its surrounding context.
  - The extracted word usage vectors are then clustered into different *usage types* using the *k-means clustering algorithm*. This helped to identify distinct senses or meanings of the words as they appear in various contexts.
  - They analyzed these clusters for a specific word across different time periods. The study proposed three metrics to quantify semantic change:
    * **Entropy Difference (ED)**: Measures the change in uncertainty (entropy) of a word's usage distribution over time.
    * **Jensen-Shannon Divergence (JSD)**: Compares the similarity of word usage distributions across time intervals.
    * **Average Pairwise Distance (APD)**: Computes the average distance between word usage vectors from different periods, indicating shifts in word meaning.
  - The qualitative analysis indicated that the approach could capture various linguistic phenomena, including both synchronic (current usage) and diachronic (historical changes) aspects.
- **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.

Linguistic change studies have shifted to include *syntax* alongside semantics. The field of linguistic change has traditionally focused on *semantics*. However, recent studies have begun to explore the role of *syntax* in language evolution as well. 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 verbs 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. They calculated the likelihood of a word appearing in specific grammatical contexts over time.
  - To quantify temporal change, they compared the probability distributions of POS tags for a particular word across different time periods. This essentially measured the divergence between these distributions.
  - An example they cited was the word 'windows'. Its POS tag shifted from a common noun (referring to doors and windows of a house) to a proper noun ('Microsoft Windows'). This highlights how a word's grammatical function can change alongside its meaning.
- **Capturing Differences Between Nouns and 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* are more associated with *global measures* and are particularly responsive to variations in verbs.

  - To validate this hypothesis, the authors employed a statistical technique called a *linear mixed model* where word type (noun or verb) is a fixed effect, amount of change measured by each metric (local or global) treated as the dependent variable. By analyzing the model's results, they could assess whether there's a significant difference in the way nouns and verbs exhibit change.
  - 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.

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 RP1: Statistically Significant Detection of Linguistic Change

4.1 Introduction

This research aims to quantify linguistic shifts in word meaning and usage across time, focusing on semantic or contextual changes. To understand data voids, we are sure that a change is happening, but we need to understand the change that is happening over time. Also after knowing that a change has occurred, we would like to know when did that change really occur so that we can co-relate that the data void occured when a political event happened or how the COVID outbreak event, which could help us identify that the term were introduced due to that event, which can be a future direction we can look into. This paper is exciting because the ideas inspired me to pursue this project and get future directions as well.

Their approach focuses on three methods to construct time series for a word. The first one is based on frequency, which captures sudden changes in word usage. In our work, this corresponds to strategic new terms data voids or breaking news data void, where breaking news data void shows that there's comes a sudden change in a word activity whenever's there an event happen. For example, the term "filmyourhospital" was introduced in during 2020 when coronavirus took place, which corresponds to a sudden spike. And this term was never used before 2020, making it a new term that introduced to support the conspiracy theory regarding encouraged people to visit local hospitals to take pictures and videos of empty hospitals to help "prove" that the COVID-19 pandemic is an elaborate hoax.

Their second method is Syntactic method where they analyze each word's part of speech tag.

And last method is distributional method which captures overall usage of word using word co-occurence. This type of method can help us identify data voids which have changed their meaning over time, such as Fragmented Concepts or Outdated Terms DV.

They apply these techniques on three different domains: books, tweets and reviews.

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.

- **RQ1.** *What methods can quantify the statistical relevance of observed changes in a word's usage across different time periods?*
- **RQ2.** *Given that a word's usage has changed, how can the precise moment or period of this shift be determined?*

5 RP2.

5.1 Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change

1. Creating Word Embeddings: • They create word embeddings for different time periods using historical text corpora. 2. Using SVD, PPMI, and SGNS: • SVD (Singular Value Decomposition): They apply SVD to the Positive Pointwise Mutual Information (PPMI) matrices to reduce the dimensionality of the word vectors. • PPMI (Positive Pointwise Mutual Information): They use PPMI to measure the association between words and their contexts. PPMI values are computed for each time period to capture word-context relationships. • SGNS (Skip-Gram with Negative Sampling): They also use the SGNS model to create word embeddings, leveraging its effectiveness in capturing semantic relationships. 3. Aligning Word Embeddings: • They align word embeddings across different time periods to ensure that the same word in different periods is represented in a comparable way. This alignment allows for tracking changes in word meanings over time. 4. Analyzing Semantic Change: • By comparing word embeddings from different time periods, they analyze the semantic shifts. They use cosine similarity to measure changes in word meanings and identify words that have undergone significant semantic shifts.

• PPMI: They computed PPMI matrices for each time period, capturing the association between words and their contexts. • SVD: They applied SVD to the PPMI matrices to reduce the dimensionality of the vectors, making them more manageable and interpretable. • SGNS: They used the SGNS model to directly create word embeddings, leveraging its ability to capture semantic relationships in a low-dimensional space.

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] Geoffrey E. Hinton. 1986. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*. Amherst, MA, Amherst, MA, 1–12.

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.