Every Word has its History: Interactive Exploration and Visualization of Word Sense Evolution

Adam Jatowt[1], Ricardo Campos[2], Sourav S Bhowmick[3], Nina Tahmasebi[4] and Antoine Doucet[5]

[1]Kyoto University
Kyoto, Japan
adam@dl.kuis.kyoto-u.ac.jp

[2]Polytechnic Institute of Tomar,
LIAAD–INESC TEC
Portugal
ricardo.campos@ipt.pt

[3]Nanyang Technological University
Singapore
assourav@ntu.edu.sg

[4]University of Gothenburg
Gothenburg, Sweden
nina.tahmasebi@svenska.gu.se

[5]University of La Rochelle
La Rochelle, France
antoine.doucet@univ-lr.fr

ABSTRACT

Human language constantly evolves due to the changing world and the need for easier forms of expression and communication. Our knowledge of language evolution is however still fragmentary despite significant interest of both researchers as well as wider public in the evolution of language. In this paper, we present an interactive framework that permits users study the evolution of words and concepts. The system we propose offers a rich online interface allowing arbitrary queries and complex analytics over large scale historical textual data, letting users investigate changes in meaning, context and word relationships across time.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

Word sense change, diachronic word change, historical linguistics

ACM Reference format:

Adam Jatowt, Ricardo Campos, Sourav S Bhowmick, Nina Tahmasebi and Antoine Doucet. 2018. Every Word has its History: Interactive Exploration and Visualization of Word Sense Evolution. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM'18). ACM, Torino, Italy, 4 pages.* https://doi.org/10.1145/3269206.3269218

1. INTRODUCTION

Every word has its own history. In particular, words that have existed for long time are likely to have undergone several changes on a semantic level. As language is our most important communication tool, we believe that properly understanding the nature of changes in the meaning and usage of words – the basic elements of the language – is important for professionals working with historical texts, such as linguists, historians, librarians and social scientists. Many prior approaches relied on manual analysis of old texts. Laborious work was required to trace word occurrences across past texts and to compare their contexts and usage for drawing general inferences. As a result, manual approaches covered a relatively small set of words or time periods. Providing detailed overview of the evolution of any word over an entire timeline has however become currently possible.

CIKM'18, October 22-26, 2018, Torino, Italy.

ACM ISBN 978-1-4503-6014-2/18/10...$15.00.
DOI: https://doi.org/10.1145/3269206.3269218

Computational approaches applied on large diachronic corpora have significant potential to advance evolutionary linguistic studies and researchers have already started proposing methods for studying word evolution [1-7]. However, few effective tools to interactively explore word change over time were proposed. Furthermore, they tend to offer simple options such as a term frequency graph and keyword-in-context view [1], rely on prepared animated scatterplots [5] or apply hierarchical clustering of frequency counts of target word in adjacent time periods [6]. Others like [7] focus on relatively short time periods (e.g., 20 years based on news articles from The New York Times Annotated Corpus) by tracking senses over time represented through LDA topics.

Equipped with effective tools, scientists would be able to freely investigate any desired word to see how it evolved over time. Furthermore, such a tool could be useful for non-expert users, as word etymology analysis attracts significant interest of the public. The recent enhancement of automatic definitions generated by the Google search engine for "definition queries" is one example. For an input word, standard word definition is complemented with a brief description of the word's origin as well as its frequency plot over time. Although users can see the count of word over time, they are however left at their own to reason about word meaning change over time.

With this in mind, we propose a multi-perspective system designed for the analysis of semantic change of words and concepts over time which is derived from our previous work [3]. Using word representations from distributional semantics, we allow evolutionary word investigation at several levels: *word analysis*, *contrastive word pair analysis, multi-word analysis* and *temporal context analysis*. Altogether these four core modes provide a user with a visual explanation of semantic change for any query word. Their synergy permits storytelling of the word evolution based on visual analytics. We use two datasets as underlying data: COHA and Google Books n-gram dataset. Due to the very large data size of the latter, our methods must be scalable to be able to promptly respond to user queries. We release an online working service that enables anyone to investigate arbitrary words in real time and is accessible at: `http://tinyurl.com/WordEvolutionStudy`

2. DATASETS & PREPROCESSING

Large scale data is the basis of effective and reliable analysis. In our approach, we utilize the largest available historical corpus that offers a comprehensive representation of English language in the past, the Google Books 5-gram dataset[1]. This dataset has been compiled over about 5% of ever published books which were scanned and subject to Optical Character Recognition (OCR). It covers the time frame from before 1600 to 2010. The data is organized as n-gram counts for every year. As an additional corpus, we use the Corpus of Historical American English (COHA) [1] which has been compiled with a decade level granularity. COHA contains over 400 million words in the form of 4-grams collected from about 107k documents written from 1810 to 2010. An important advantage of COHA is that it is based on a stable rate of diverse document genres for each decade. Comparison of results obtained from these two datasets should allow for more informed judgments. Although COHA contains carefully selected and balanced prose texts based on a stable rate of different genres across decades, it is many orders of magnitude smaller. For efficiency and for facilitating result comparison across both the corpora, we converted the Google 5-gram dataset to the decade granularity by summing 5-grams within each decade. Decade level should be sufficient to reason about word evolution as, typically, diachronic semantic changes materialize over longer time spans. This also provides smoothing effect to decrease the impact of short-term fluctuations. Other preprocessing steps included converting words to lowercase, removing punctuation and discarding rare n-grams.

3. SYSTEM DESCRIPTION

3.1 Word Representation

We adopt a common approach used in NLP for representing words, the distributional semantics, according to which, a word's meaning is captured by co-occurring words (hereafter called context terms). For a given target word w in a decade d, we collect all n-grams that contain w. We then sum the counts of all context terms. The word representation in d is then given by a vector, whose size is the number of unique words found in the dataset. The weights in this vector are calculated as the normalized counts of context terms co-occurring with the word w in d. Note that neural network based word embeddings have been also used for diachronic sense detection [2,4]. We decided to start with simpler and intuitive word representation leaving the addition of other solutions for later stage.

3.2 Word Analysis

In the first analysis, *word analysis*, the user issues a query word and the system evaluates the degree of its context change across time. In particular, the vector representation of the query at a reference decade d_r is compared with the one of any other decade. If the similarity between the word's vector at decade d_i and the one at reference decade d_r is low (i.e. $sim(d_i[w], d_r[w]) \rightarrow 0$), then a semantic change is likely to have occurred between these two decades. We use the following measures for computing context self-similarity over time: cosine similarity, Pearson Correlation Coefficient and Jaccard Coefficient[2]. A user can choose either one of them and can compare views based on different measures.

Fig. 1 shows the snapshot of the word analysis. The user starts by selecting a particular reference decade (i.e., by default the latest decade) with which the representation of the query in other decades is to be compared. The resulting similarity plot is then drawn for the entire time frame[3]. Fig. 1 demonstrates the result obtained for the word `mail`. The top left-hand side graph shows by the thick blue line the similarity plot between the query's semantics in the reference decade and the ones in each past decade using cosine similarity. It permits understanding when the meaning of the query became similar to the meaning present in the reference decade. The curve with the high and steep increase characterizes a word which acquired the present meaning in a relatively short time period (as in the example of Fig.1). On the other hand, a flat curve indicates words with stable or slowly changing meaning over time. The right-hand side plot in Fig. 1 (blue line) adds more evidence to the word evolution analysis by outputting similarities between two consecutive decades. It is then possible to examine if the input word underwent other significant changes over time including ones not related to the current meaning. Finally, the bottom graph in Fig. 1 shows both the raw word count and its normalized frequency over time. It offers complementary information about the *interplay of word popularity and its semantic change*. Additionally, the plot helps to detect periods characterized by low frequencies of query which may poorly reflect query's pattern of semantic change[4].

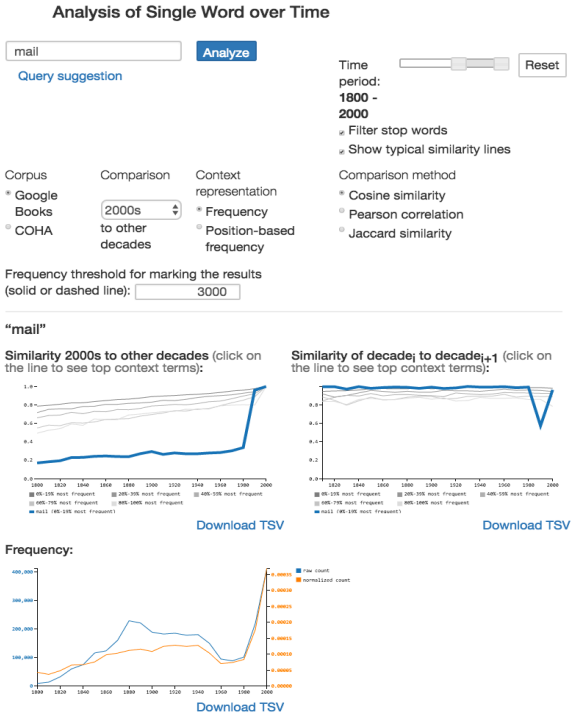

Figure 1. Word analysis example (word: `mail`).

Frequency-based referencing. Reasoning about the sense shift of a word should be always done with proper context. We then draw average similarity curves for all words belonging to different frequency segments (see gray lines in the top-left graph plot of Fig. 1). For example, the shape of `mail` in Fig. 1 differs greatly from the graph of its corresponding frequency bin (the grayest color curve representing the 0%-19% of the most frequent terms). Note that the curves for less frequent words exhibit on average greater

[1] http://books.google.com/ngrams/datasets

[2] An option is provided for removing stop words from computation.

[3] The system allows freely setting the time frame for an optimal view.

[4] To alert users about potential non-representativeness, periods with frequency below a pre-specified threshold are indicated by dashed lines.

change over time than ones for more frequent words, which confirms the observation made in prior studies [4].

Detailed context exploration. This mode permits fine-grained investigation of query at each decade to let users understand the reasons behind the similarity fluctuations. Users can click on any decade to see the list of the 50 top context terms co-occurring with the query word and their frequencies at that decade. They can then contrast such terms with those in other decades by clicking on any other decade to open up new lists with differing words being automatically colored. Fig 2 shows the joint side-by-side comparison of `mail`'s context over three different decades.

At 1830s

word	frequency
coat	3,710
coach	2,021
coats	923
shirt	727
black	451
transportation	406
clad	296
complete	277
plate	263
coaches	252

At 1980s

word	frequency
globe	4,873
daily	3,448
order	2,897
toronto	1,225
registered	1,189
class	886
rand	784
coat	697
certified	585
business	530

At 1910s

word	frequency
order	5,20
daily	4,22
net	2,97
coat	2,78
pacific	2,28
steamship	2,12
company	2,07
service	1,92
class	1,91
steam	1,76

Figure 2. The color-coded comparison of top co-occurring terms at three different decades selected by clicking on the view shown in Fig 1 (the lists have been truncated).

3.3 Contrastive Word Pair Analysis

Contrastive word pair analysis is another mechanism for quantifying the change in word meaning which works through comparison. Analyzing the similarity of two related words (e.g., synonyms) can shed light on the evolution of each of them.

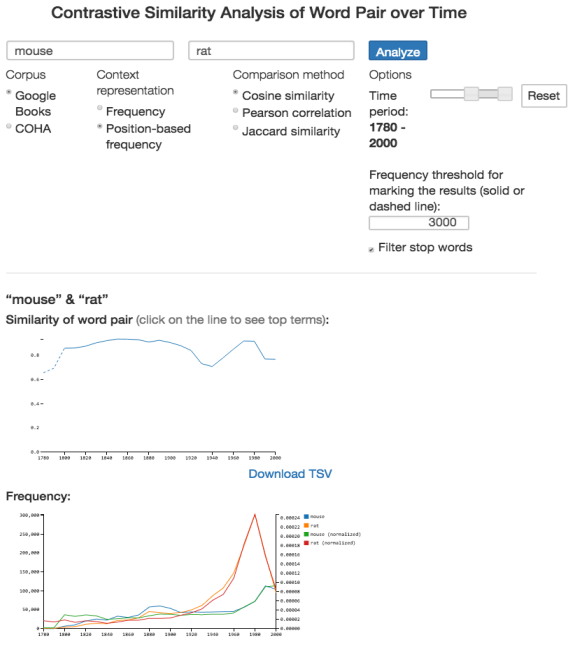

Figure 3. Contrastive word pair analysis (`mouse` vs. `rat`).

For example, it is known that the meanings of `nice` and `pleasant` gradually converged. Using the word level analysis, one could then first observe that `pleasant` was characterized by relatively stable meaning across past decades, while, in contrast, `nice` assumed variety of meanings in the past including "foolish" or "silly". Then, the comparison of the usage of both words over time becomes helpful for better gauging the changes in their meaning. Fig. 3 shows partial view of the results obtained for comparing example words `mouse` and `rat`.

Contrastive exploration of context. While the above-described word-to-word similarity curve quantifies how much the two input words were similar in the past, we also need a way to portray the commonalities and differences of the words at particular time points (similar to the option provided in *word analysis* described in Sec. 3.2). By clicking on the top plot in Fig. 3 the two side-by-side lists are opened with the top context terms, one for each input query word (see Fig. 4). The lists are color-coded for highlighting the same and different context terms. A user can click on any other decade to analyze variations of the pair's commonalities and differences across different decades.

At 1940s

mouse		rat	
word	frequency	word	frequency
cat	1,467	white	2,406
house	544	albino	1,643
trap	463	tat	1,232
rat	454	rat	1,186
church	391	bite	1,159
lion	355	fever	938
country	352	behavior	821
quiet	343	jour	693
mouse	320	comp	551

At 1890s

mouse		rat	
word	frequency	word	frequency
cat	1,359	trap	874
trap	1,046	water	647
country	721	hole	623
field	699	tat	610
church	450	musk	496
quiet	405	tail	368
mouse	394	smell	320
lion	371	poisoned	306
white	347	mouse	291

Figure 4. Visualization of top context terms and their differences for word pair at two decades selected from the view shown in Fig. 3 (the lists have been truncated).

3.4 Multi-word Analysis

Investigating concept evolution is complementary to analyzing the meaning change of a single term. For example, while one may visualize the meaning change of a word `car` based on the previously discussed *word analysis* mode, a more complete view of its evolution can be achieved when viewing it "through the lens" of the evolution of the entire concept represented by the set of similar or strongly related words. In the case of the word `car`, one could then also observe changes of `auto`, `automobile`, `vehicle` or `truck` to compare the evolution of the word `car` with the broader concept of an autonomous object driven by a human. Our system allows combining temporal characteristics of words in the word set into a single joint representation. The *word analysis* and the *contrastive word pair analysis* described above are then extended here to the *multi-word analysis* scenario where a word set of up to *n* terms is tracked over time to reveal the inter-word similarities. Fig. 5 demonstrates the results obtained for the `car` example. The top left graph in Fig. 5 shows the across-time similarity curves of each word separately by thin lines. Their aggregate is denoted by a thick blue line. The line is computed by combining the context of each individual term from each decade.

Contrasting the plots of individual words with their aggregate allows for understanding if a word's sense followed, or deviated, from the meaning of its corresponding concept over time, offering in this way additional pieces of information for solving "the word's evolution puzzle". The chart on the top right of Fig. 5 portrays decade-to-decade changes of each word as well as the one of the aggregate curve. Next, the bottom left graph portrays the frequency over time of each word and of the word set, which is computed as the sum of the individual term frequencies. Finally, the right bottom

graph of Fig. 5 plots word-to-word similarities for each combination of two words in the input word set.

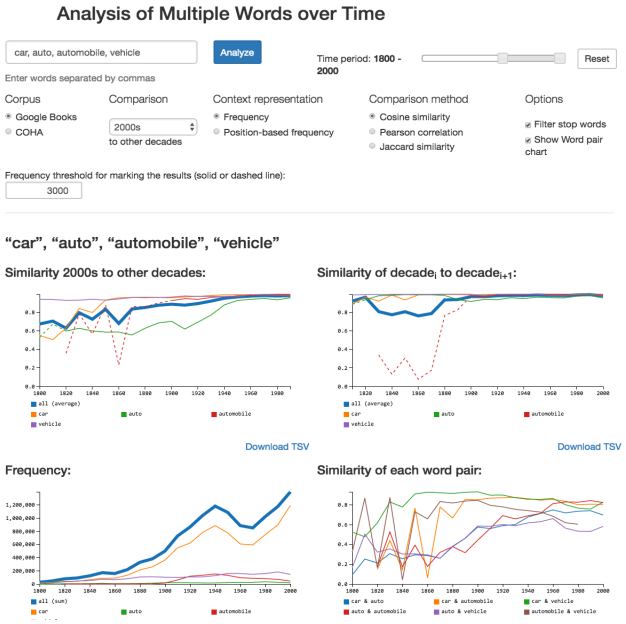

Figure 5. Multi-word analysis example (concept: `car`).

3.5 Temporal Context Analysis

Given a query word, a natural question for a careful investigator is how its context terms changed over time. To answer this question, we develop an approach for *temporal context analysis* by temporal summarization of query context. We use term clouds to represent the context of a query over time. A term cloud is an effective visualization style for quickly grasping contents of large document collections. A standard term cloud is however suited for synchronic text collections. There seems to be no ready solution for effectively showing contents of diachronic collections using standard term clouds. Displaying the cloud of context terms over the entire period in a single view would obviously not allow for a time-focused investigation. One could of course simply display separate clouds for each decade. This would however require side-by-side, sequential comparisons placing cognitive burden on users.

We propose extending the standard word cloud by enhancing it with temporal views of each context term. This helps not only to spot the frequent context terms overall (by comparing their font size) but also to observe how they occurred over time. In particular, we propose two modes for capturing term occurrence patterns over time: *time series view* and *colored matrix view* (not shown here due to space limitation). The first one allows for more detailed quantitative analysis while the second is for a quick overview of the data through colored scheme. In both the modes, it is possible to flexibly adjust the amount of terms to be shown and to zoom-in or out the clouds and to switch anytime between the time-focused or standard term cloud ("show/hide details" button). Finally, there is an option of clicking on any context term in the cloud to see the examples of containing it n-grams and the query word, hence, connecting the bird's eye (temporal summary) view with the raw data to allow users detailed investigation or verification of hypotheses.

Word pair context summary. The time-enhanced views of context terms provide an overview of what context terms co-appeared with the query and what were their occurrence patterns over time. We next extend this analysis to consider also contrastive type analysis of context. This corresponds to the *contrastive word pair analysis* described in Sec. 3.3. The input is then a pair of words. Recall that in *contrastive word pair analysis* users could select any decade to compare the context terms of query words at particular decades (as shown in Fig. 4). Now however we wish to contrast the top context terms over the entire time in a single view.

When two words are input, the output cloud contains combined context terms for the two words, each shown in a different color. The shared context terms for both the input query words are aligned one next to another allowing immediate comparison of their saliency. Their temporal plots are also merged to facilitate time-focused comparison. Finally, if the same query is entered twice (e.g., `car`, `car`) the system contrasts the summary of context terms occurring before and the ones appearing after the query word. Such a view can help determine similarities and differences between context before and after query.

4. IMPLEMENTATION

To accommodate the large size of data, we have used MapReduce framework, Apache Spark and PostgreSQL 9.3.9 with default indexing algorithm (B-tree). Scala 2.11.6 was used for data preprocessing and server-side programming together with a Web application framework: Play Framework 2.3.8. TypeScript 1.5 (JavaScript) was applied for client-side programming, while for UI we used the following libraries: D3.js 3.5.6, Bootstrap 3.3.2 and jQuery 2.1.3. We have also pre-computed results for the most common 100k English words. Thanks to these, results are returned relatively fast, typically, within a second.

As mentioned before, the system is available online for arbitrary queries offering then a discovery tool for the audience. Users can also download the results in tsv format for offline analysis. Furthermore, it provides query suggestions computed based on the distinctive shapes of terms' time series plots from Sec. 3.2. For example, it suggests words that retained stable senses over long time or underwent high semantic shift over short time.

ACKNOWLEDGMENTS

This work is partially funded by JSPS kakenhi grants #17H01828, #18K19841, MIC SCOPE (#171507010) and the ERDF through the COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT as part of project UID/EEA/50014/201.

5. REFERENCES

[1] M. Davies. The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *LLC*, 25(4):447–464, 2010.

[2] Y. Kim, *et al.* Temporal Analysis of Language through Neural Language Models. In *Proc. of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pp. 61-65, 2014.

[3] A. Jatowt, K. Duh. A Framework for Analyzing Semantic Change of Words across Time. In *Proc. of JCDL2014*, pp. 229-238. IEEE, 2014.

[4] W. Hamilton *et al.* Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proc. of ACL2016*, pp. 1489-1501.

[5] M. Hilpert, F. Perek. Meaning Change in a Petri Dish: Constructions, Semantic Vector Spaces, and Motion Charts. *Linguistics Vanguard*, 1(1), 339-350, 2015.

[6] M. Hilpert, S. T. Gries. Assessing Frequency Changes in Multistage Diachronic Corpora: Applications for Historical Corpus Linguistics and the Study of Language Acquisition. *Literary and Linguistic Computing*, 24(4), 385-401, 2008

[7] C. Rohrdantz, *et al.* Towards Tracking Semantic Change by Visual Analytics. In *Proc. of ACL2011*, 2011.