Unsupervised detection of diachronic word sense evolution

Jean-François Delpech
WebGlyphs, Inc.
1515 N. Colonial Ct
Arlington, VA 22209-1439
United States
jfdelpech.webglyphs@gmail.com

Most words have several senses and connotations which evolve in time due to semantic shift, so that closely related words may gain different or even opposite meanings over the years. This evolution is very relevant to the study of language and of cultural changes, but the tools currently available for diachronic semantic analysis have significant, inherent limitations and are not suitable for real-time analysis. In this article, we demonstrate how the linearity of random vectors techniques enables building time series of congruent word embeddings (or semantic spaces) which can then be compared and combined linearly without loss of precision over any time period to detect diachronic semantic shifts. We show how this approach yields time trajectories of polysemous words such as Amazon or Apple, enables following semantic drifts and gender bias across time, reveals the shifting instantiations of stable concepts such as hurricane or president. This very fast, linear approach can easily be distributed over many processors to follow in real time streams of social media such as Twitter or Facebook; the resulting, time-dependent semantic spaces can then be combined at will by simple additions or subtractions.

1. Introduction

Over the last few years, following the work of Mikolov et al. and other investigators, successful word representation methods inspired from neural-network language modeling have been demonstrated and explained. These representations are nonlinear and the semantic spaces in which they are embedded are randomly oriented as the cost functions used for training are invariant under rotation. These two properties make it very difficult to combine the semantic spaces obtained from several corpora and thus to compare their semantic neighborhoods: the word alignment required for diachronic investigations is thus a difficult problem.

Approximate, time-dependent word embeddings have nevertheless been proposed in a few recent publications, with encouraging results, but they are computationally intensive and have limited time resolutions.

Yao et al. show that a low-rank factorization of the pointwise mutual information matrix can be performed by adding to the usual minimization problem a smoothing term to encourage word embeddings to remain aligned between time slices at \( t_{i-1} \) and \( t_i \); a parameter \( \tau \) controls how fast the embedding may change between time slices and thus the time resolution. Obviously, even though alignment may be reasonable between consecutive time slices, it will usually be poor over larger time periods. They use as a dataset articles from the New York Times published from January 1990 to July 2016; they have made this corpus public and it was used in the present article.

In a recent article, Kulkarni et al. discuss three approaches for detecting semantic shifts of words:

- Their first approach involves simply counting the change in the relative number of occurrences (frequency) of a given word between time slices. Despite its simplicity, this method sometimes yields useful results.
- Their second approach involves tracking the functionality of a word; for example, they show that apple is almost always a common noun before 1976 but often becomes a proper noun after the creation of the eponymous company.
- Their third approach is more specific, as it actually enables tracking word senses across time slices (or epochs) by first learning embeddings for each epoch and then aligning them. However, their alignment process starts with two assumptions of uncertain validity, that (a) the semantic spaces are equivalent under a linear transformation and that (b) the meaning of most words do not shift over time.

They give a few interesting examples using a 24 month long sample from Twitter from September 2011 to October 2013 (resolution of 1 month), the Amazon Movie Reviews dataset from 2000 to 2012 (resolution of 1 year) and the Google Books Ngram corpus over the last 105 years, with a 5-year resolution.
Hamilton et al. have used similar approaches to demonstrate empirically that polysemous words change at a faster rate than non-polysemous words. Also, gender and ethnic stereotyping has been quantified over the past 100 years by Garg et al. using word embeddings derived from various sources.

These nonlinear approaches rely on a foundational assumption in Natural Language Processing (NLP), the distributional hypothesis, according to which linguistic items with similar distributions (i.e. occurring in the same contexts) have similar meanings, where "context" means in practice a moving window of width $w$ or perhaps sometimes a full sentence.

A linear approach is also obviously compatible with this assumption and a very simple, linear embedding transformation relies on the fact that in $\mathbb{R}^d$, an exponentially large number $N \gg d$ of random vectors quasi-orthogonal to each other can be created. These vectors can then be treated to a good approximation as if they were orthogonal and

a. to each distinct, significant term $t$ in a large set of documents is associated a normalized random seed vector belonging to $\mathbb{R}^d$ which is quasi-orthogonal to any other seed vector, and

b. to each term $t$ is then attached as term vector, i.e. a linear, weighted combination of the seed vectors of the terms co-occurring with $t$ in all windows of fixed size $w$.

With a window size $w = 5$, this means that the sentence fragment centered around term $t_i$

\[
\ldots \quad t_{i-3} \quad t_{i-2} \quad t_{i-1} \quad t_i \quad t_{i+1} \quad t_{i+2} \quad t_{i+3} \quad \ldots
\]

will increment term vector $|T_i\rangle$ by the sum of four seed vectors $|T^s\rangle$:

\[
|T_i\rangle = \rho_{i-2}|T^s_{i-2}\rangle + \rho_{i-1}|T^s_{i-1}\rangle + \rho_{i+1}|T^s_{i+1}\rangle + \rho_{i+2}|T^s_{i+2}\rangle
\]  

(1)

where the $\rho_{i+4}$ are multiplicative coefficients, for example weights. If term $t_i$ occurs $N$ times, vector $|T_i\rangle$ will be the sum of $N$ equations such as Equation 1, with terms which may be different or not, depending on the strength of their association with $t_i$. This is obviously independent of sentence order.

Each term vector $|T_i\rangle$ as well as any linear combination of term vectors such as documents are themselves embedded in $\mathbb{R}^d$. In this $d$-dimensional Euclidean semantic space noted $S$, the similarity $\sigma_{ij}$ between terms $t_i$ and $t_j$ is the scalar product of the associated, normalized term vectors:

\[
\sigma_{ij} = \langle T_i | T_j \rangle
\]

(In what follows, inner products will always be assumed to be evaluated between normalized vectors, unless otherwise noted.)

Consider a partition $k$ of corpus $K$; all term vectors $|T^k\rangle$ are embedded in semantic space $S_k$ and one can define symbolically the semantic space $|S_K\rangle$ of the whole corpus

\[
S_K = \sum_{k \in K} S_k
\]  

(2)

since the coordinates of each vector result from a combination of the same linear operations in a different order.

A marked advantage of this linear, transparent process is that it is by nature incremental, which is very important in diachronic studies; a small addition to (or deletion from) the data set involves only a small, finite number of words, at least to a first, very good approximation (over time, word frequencies do vary, and this may be reflected in slowly evolving logarithmic weight factors.) Seed vectors are derived from word strings by a hashing function: they are thus reproducible and semantic spaces from different epochs can be combined at will, as in Equation 2. In fact, despite the use of random vectors, the whole process is totally deterministic in that, even with sentences in random order (provided word order remains fixed within each sentence) the same corpus with the same seed vectors will always generate the same semantic space within rounding errors.

In this work, the embedding space dimensionality was $d = 300$ and the window size was $w = 11$ (5 words before and 5 after the central word.) Yao’s et al. dataset was used. The 27 time slices contain a total of 92,965 articles and 91,423,989 words; out of a total of 455,008 distinct words, 142,439 distinct words remain after removing the 100 most frequent words (which occur 42,703,951 times, or 47% of the time) and ignoring words occurring less than 5 times. A total semantic space $S_K$ was also created by combining all 27 partial spaces $S_k$. 

2/10
2. Linearity

The time evolutions of words *bush* and *enron* are shown in Figure 1. It can be seen that the norm (square of the length of the term vector for each year, in this linear process) and the frequency (number of occurrences) closely track each other, as would be expected. This is quite different from what is shown in Figures 2 and 3 of Yao *et al.*; between 1993 and 1998, they find that the frequencies for *bush* become almost zero but that the nonlinear norms remain substantial, which seems to indicate that the time resolution of their method is of the order of 2 or 3 years.

![Figure 1 — Time evolutions of words *bush* and *enron*](image)

Norms are in red, frequencies in green and both are normalized: for *bush* the maximum frequency occurs in 2004 (3490 occurrences), as well as the maximum norm (1833); for *enron*, the maximum norms are respectively 1145 and 877. The dips in 2006 are due to the fact that this year is an outlier with a shorter text (about 25% of neighboring years; 2016 text is also shorter by about half.) Note that norm and frequency track each other closely, which is to be expected for a linear process. *For bush,* the numbers track the election years and in 2015 — 2016, the focus had shifted to candidate Jeb Bush.

3. Time trajectories

In their figure 1, Yao *et al.* show time trajectories of brand names by plotting for each epoch the 2D t-SNE projections of the brand name and of its closest neighbors; the exact meaning of these plots is not obvious as the contribution of the slice-to-slice alignments and of the projection technique are difficult to unravel.

In a linear, random-vector approach, time evolution can easily be analyzed with the following algorithm:

**Algorithm 1** — Time evolution algorithm for term *t*

1. Extract and normalize the vector \( |\mathcal{T}_t^K| \) of term *t* from the total semantic space \( S_K \)
2. Find a set \( R \) of representative neighbors of term *t* in \( S_K \)
3. For each epoch \( k \in K \)
   4. Extract and normalize the vector \( |\mathcal{T}_t^k| \) of each member \( r \in R \) from the partial semantic space \( S_k \)
   5. Compute the similarity \( s_r = \langle \mathcal{T}_t^k, \mathcal{T}_t^K \rangle \) for each member \( r \) and keep closest terms
6. Return Top neighbors as a function of \( k \)
7. Tabulate top neighbors

Note that in this algorithm, a vector extracted from the total semantic space \( S_K \) is compared to vectors extracted from partitions \( S_k \); this does not require any approximation because Equation 2 is strictly valid for a linear process.

All references to *amazon* were close to *rainforest* and to related terms before 1998, as shown in Table 1. The company was founded in 1994 by Jeff Bezos; he appears in second place (pink rectangle) in 1997; terms related to *rainforest* being still dominant. After 1997, terms relating to the company and its competitors tend to be more frequent but both senses coexist.

| 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 00 | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| de|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| forest |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   | rainforest |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   | ignitos |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   | manus |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   | streaming |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   | online |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   | downloads |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   | netflix |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

**Table 1** — Time trajectory of *amazon*

This table shows the closest (in red) and the second closest neighbors (pink) of *amazon* from 1990 to 2016 according to the NYT corpus. The top four words relate to the Amazon region of South America; this meaning remains stable throughout the years. Terms related to the company (founded in 1994) begin to appear significantly in the corpus in 1998 and the two meanings (geography and business) are both present after 1998.

Apple’s behavior is different, which is not surprising since the company was already in existence in 1990 (Table 2.) The two meanings coexist from the start, separated by *blackberry*, which can itself be a fruit or a device/company name (all appearances of *apple* in the New York Times refer to a fruit before 2000.)
4. Semantic drift between epochs

To evaluate semantic drifts, we first built two semantic spaces, \( S_0 = \sum_k S_{k|1990, 1995} \) and \( S_1 = \sum_k S_{k|2010, 2015} \) covering the first and last six full years of the corpus by a linear combination of each year’s semantic space (cf. Equation 2). The algorithm to evaluate the drift between (0) and (1) is then very simple:

\[
\text{Algorithm 2 - Algorithm to evaluate the drift in meaning of terms between time periods 0 and 1}
\]

1. Input: Semantic spaces \( S_0 = \sum_k S_{k|1990, 1995} \) and \( S_1 = \sum_k S_{k|2010, 2015} \) for the two time periods
2. for each term \( t \) in corpus
3. Extract and normalize vectors \( |T_0^t| \) and \( |T_1^t| \) of \( t \) from \( S_0 \) and \( S_1 \)
4. Compute \( \sigma_{01} = \langle T_0^t | T_1^t \rangle \)
5. \( \sigma_{01} \) is an indicator of the drift
6. return The lists of terms closest to \( \nu_0 \) and \( \nu_1 \)

This algorithm is very fast and can easily handle all the terms, but in practice a term must be frequent enough for the result to be significant; only the 5210 non-stop words occurring more than 1500 times have been kept.

Words can be grouped into four categories, as shown in Tables 3, 4, 5 and 6. In these four tables, the reference words are followed by the similarity \( \sigma_{01} \) between their vectors \( \nu_0 \) and \( \nu_1 \) for the two time periods. The lists of the words closest to \( \nu_0 \) and \( \nu_1 \) in each year and the numbers between parentheses are their total number of occurrences. The words underlined in blue are common between the two periods.

a. Very stable words

Stable words exhibit a strong similarity (\( \sigma_{01} \geq 0.70 \)) between periods (Table 3). This is the case for words such as inches, feet, office, murder, which show little change with time.

| Term | \( \sigma_{01} \) | Year 1990–95 | Year 2010–15 |
|------|-----------------|---------------|--------------|
| inches | 0.88 | feet (5), pounds (4), diameter (3), crackle (1), bench-press (1), concours (1), cylinder (1), width (1), timbers (1), spacing (1), dices (1), mine (1), metalurgist (1), oxves (1), hurdles (1), record-holder (1), emaciation (1), thickness (1) | feet (5), pounds (5), diameter (2), high-tops (1), implements (1), right-thinking (1), rainbow (1), tailors (1), aleksey (1), e-class (1), facemask (1), milliville (1), danbury (1), clintonville (1), c-class (1), big-picture (1), lomghi (1) |
| office | 0.85 | state (6), house (4), district (4), yesterday (3), officials (3), money (3), department (3), percent (2), agency (2), city (2), services (2), states (1), government (1), deputy (1) | city (4), home (3), district (3), case (2), state (2), government (2), weeks (1), mukasey (1), part (1), lackawanna (1), deal (1), carmo (1), company (1), department (1), right (1), investigation (1) |
| feet | 0.77 | inches (5), square (5), pounds (3), square (2), garden (1), madison (1), damage (1), flood (1), maneh (1), cul-de-sac (1), throwers (1), millimeter (1), red-bearded (1), rigger (1), swooned (1), calves (1) | inches (4), foot (4), pounds (4), square (2), body (2), washes (1), cuts (1), capacity (1), legs (1), two-pull (1), size (1), baths (1), bahrain (1), extra (1), force (1), squeeze (1), madison (1) |
| murder | 0.70 | charges (4), trial (3), manslaughter (3), counts (3), criminal (2), case (2), kidnapping (2), defendants (2), prosecutors (2), lawyers (2), possession (1), rape (1), simpson (1), prosecution (1), riot (1) | counts (2), burglary (2), manslaughter (2), strangulation (2), robbery (2), prison (2), assadsaurian (1), benefactors (1), mother (1), rampart (1), frayr (1), shred (1), raze (1), gasst (1), solicito (1), children (1), rape (1) |

Table 3 -- Examples of stable words between 1990–1995 and 2010–2015

For higher similarity values, despite some minor time-related shifts, word meanings show little significant change with time.

b. Moderately stable words and food words

A second category is still reasonably stable and exhibits similarities \( \sigma_{01} \approx 0.5 \). This is the case for many food words shown in Table 4: eating habits have much changed during these 20 years, at least for readers of the New York Times: for example, the word sauce disappears in the second period. In the first period, most of the words correspond to standard international fare but they become much more ethnic and more dispersed in the second period (note that
apparently unrelated words may in fact be subtly related. For example, *sosinski* doesn’t refer to the basketball player but to a Montauk lobster fisherman who almost drowned...
d. Unstable words

Some words (see Table 6) are inherently unstable because they refer to fast-changing items, such as crisis, borders, moscow or muslims. More surprising is the unstability of lunch, from eating food to a much less specific connotation.

| crisis | σ_m = 0.29 |
|--------|------------|
| 1990–95 | minesweepers (1), persian (1), boredom (1), world (1), action (1), villanueva (1), marshaled (1), midrock (1), deployment (1), discussions (1), support (1), conflict (1), build-up (1), omit (1), midlife (1), oil-rich (1), big-think (1), golf (1), perennials (1), plenty (1), incursion (1), grammar (1) |
| 2010–15 | financial (6), economy (5), government (4), banks (4), market (4), debt (2), institutions (2), rescue (1), recession (1), growth (1), information (1), default (1), industry (1), executives (1), contraction (1), deal (1), bailout (1), services (1), sovereign (1), company (1), greece (1), eurozone (1) |
| borders | σ_m = 0.28 |
| 1990–95 | scrolled (1), yoga (1), riley (1), conacher (1), faceoff (1), economy (1), mccarter (1), chorske (1), phone (1), agreement (1) |
| 2010–15 | doctors (2), group (1), guinea (1), core (1), fronti (1), kraig (1), waldron (1), syria (1), sprecher (1), curbside (1), cargo (1), country (1), europe (1), company (1), africa (1), ebola (1), migrants (1), countries (1), nurses (1), decins (1), potboiler (1), china (1), vidor (1), annick (1) |
| moscow | σ_m = 0.27 |
| 1990–95 | russia (4), russian (4), yeltsin (2), boris (2), union (2), russia (1), estonia (1), latvia (1), criminality (1), blockade (1), megarussian (1), large (1), leaders (1), leaders (1), diplomat (1), ukrainian (1) |
| 2010–15 | snowden (1), leaker (1), debaltseve (1), kvasha (1), ecuador (1), comstock (1), latino (1), peat (1), novosibirsk (1), putin (1), benfica (1), contractor (1), rozehnal (1), bradsher (1), datsyuk (1), russian (1), oleg (1), vladimir (1), crimea (1), spartak (1), Pavel (1) |
| muslims | σ_m = 0.26 |
| 1990–95 | serbs (4), croat (4), bosnian (3), muslim (2), realpolitik (1), territory (1), cleansing (1), rajaf (1), non-arab (1), slaves (1), forefathers (1), sunni (1), saudi (1), kiseltak (1), croatia (1), croat (1), vares (1), slovenes (1), croatian (1), krajina (1), allists (1), christians (1), bosnia (1) |
| 2010–15 | hindus (2), clerics (1), hindu (1), jats (1), wiccans (1), jihadist (1), amredola (1), mohammed (1), andhra (1), quran (1), torqueing (1), swap (1), ahmadis (1), kumar (1), buddhists (1), fugana (1), sikhs (1) |
| lunch | σ_m = 0.18 |
| 1990–95 | entrees (6), dinner (6), range (6), appetizers (6), courses (4), fixe (3), carte (2), desserts (2), prix-fixe (2), pre-theater (2), pastas (1), mondays (1), price (1), prix (1), noon (1), three-course (1), brunch (1) |
| 2010–15 | fact (1), apartment (1), outing (1), wright-phillips (1), home (1), cronenberg (1), breakfast (1), meritocratic (1), dinner (1), bastille (1), ritz-carlton (1) |

Table 6 -- Some unstable words

5. Gender bias in the New York Times over the last twenty years

Garg et al. have used a number of sources in their evaluation of gender bias and we have downloaded from their GitHub site lists of adjectives of appearance and of intelligence, a list of occupations, and two lists of words related respectively to men and to women (man terms and woman terms, to use Garg et al. terminology).

Tables 7 and 8 have been constructed using the following algorithm:

Algorithm 3 - Find most frequent men and women qualifiers in time periods \( K_0 \) and \( K_1 \)

1: Input: The lists of qualifiers and of man/woman terms
2: for each year \( k \in \{ K_0, K_1 \} \)
3: for each qualifier \( q \)
4: Extract qualifier vectors \( \mathbf{q}^k \) from partial semantic space \( S_k \)
5: for each man term \( m \)
6: Extract male vectors \( \mathbf{M}_m^k \) from \( S_k \)
7: Compute \( \sigma_{mq}^k = \langle \mathbf{A}^k, \mathbf{M}_m^k \rangle \)
8: return largest \( \sigma_{mq}^k \)
9: for each woman term \( w \)
10: Extract female vectors \( \mathbf{W}_w^k \) from \( S_k \)
11: Compute \( \sigma_{wq}^k = \langle \mathbf{W}_w^k, \mathbf{Q}_w^k \rangle \)
12: return largest \( \sigma_{wq}^k \)
13: if \( \sigma_{mq}^k > \sigma_{wq}^k \) qualifier \( q \) is taken as preferentially female in year \( k \) and conversely
14: return The gender attached to qualifier \( q \) in year \( k \)
15: return The lists of man and female qualifiers in time periods \( K_0 \) and \( K_1 \)
16: Create table of man and female qualifiers for both time periods

Given its readership, one would expect the New York Times to avoid gender stereotyping; this is indeed the case:

Table 7 -- Top qualities by gender in the periods 1990 — 1995 and 2010 — 2015

| Year     | Male                   | Female                  |
|----------|------------------------|-------------------------|
| 1990–95  | healthy, slim, attractive, slender, genius, ugly, brilliant, muscular, smart, homely, fashionable, judicious, resourceful | strong, beautiful, wise, thoughtful, intelligent, handsome, weak, fat, thin, inventive, blushing, clever, reflective, gorgeous |
| 2010–15  | strong, healthy, ugly, brilliant, handsome, athletic, thoughtful, alluring, slim, clever, ingenious, imaginative, plump | intelligent, beautiful, smart, muscular, wise, attractive, logical, thin, fashionable, gorgeous, weak, genius, inventive, fat |

Table 8 -- Top occupations by gender in the periods 1990 — 1995 and 2010 — 2015

| Year     | Male          | Female                   |
|----------|---------------|--------------------------|
| 1990–95  | police, official, doctor, secretary, lawyer, artist, driver, professor, manager, guard, athlete, physician, surgeon, dancer | author, teacher, baker, sales, pilot, farmer, soldier, cook, judge, designer, nurse, attendant, engineer, psychologist |
| 2010–15  | police, official, guard, manager, soldier, architect, pilot, physician, driver, nurse, therapist, smith, mason, cook | doctor, lawyer, teacher, author, sales, designer, baker, professor, photographer, farmer, artist, operator, scientist, secretary |
6. Equivalences

While the concepts of hurricane or president are stable, their instantiations change with time. However, with the distributional hypothesis, we would expect most of their their neighboring vectors in different years to be close as they correspond to generic hurricanes or generic presidents. Table 9 was built with the following algorithm:

**Algorithm 4** - Algorithm to find in year $k$ the terms closest to term $t$ in year $k'$

1: Input: All partial semantic spaces $S_k$
2: Extract and normalize vector $|T_k^t|$ of term $t$ from the partial semantic space $S_k$
3: for each year $k \in K$
   4: return the two closest neighbors of $|T_k^t|$ in the partial semantic space $S_k$
5: Tabulate top two neighbors

The result is shown in Table 9: the vectors of an instantiation in the relevant year $k'$ from semantic space $S_k$ (say katrina$_{2005}$ or obama$_{2016}$) are indeed most often closest to each year's instantiation: for example, andrew$_{1992}$ or floyd$_{1999}$ are closest to katrina$_{2005}$ etc.

Note that andrew$_{1993}$ is only second closest to katrina$_{2005}$ and comes after fanfare. This is due to the fact that in this partition actual fanfares are associated with proper names such as andrew: this is an artefact of the distributional hypothesis, which is also one of the reasons why the catastrophic hurricane katrina casts a long shadow after 2005: the term does not necessarily refer to the hurricane, both before and after 2005, and this introduces some uncertainty. Similarly, bush or clinton do not always refer to a president, or hillary to Ms. Clinton (her first appearance in this New York Times corpus dates from 1992.) Even clinton in 1990 refers less than 50% of the time to Arkansas Governor and future U.S. President Bill Clinton.

| Year | Top term | Next term | Terms equivalent to katrina$_{2005}$ | Terms equivalent to obama$_{2016}$ | Terms equivalent to clinton$_{1996}$ | Terms equivalent to hillary$_{2015}$ |
|------|----------|-----------|-------------------------------------|----------------------------------|-------------------------------------|-------------------------------------|
| 1990 | —        | —         | bush                                | sort                             | marcos                               | edelman                             |
| 1991 | segment  | acres      | bush                                | took                             | bush                                | number                              |
| 1992 | andrew   | tasks      | bush                                | campaign                         | bush                                | clinton                             |
| 1993 | fanfare  | andrew    | country                             | television                       | clinton                             | support                             |
| 1994 | —        | —         | clinton                             | support                          | clinton                             | campaign                            |
| 1995 | felix    | luis      | clinton                             | resident                         | clinton                             | campaign                            |
| 1996 | hurricane| storms     | clinton                             | president                        | clinton                             | campaign                            |
| 1997 | —        | —         | clinton                             | violence                         | clinton                             | campaign                            |
| 1998 | bonnie   | wilmington| candidates                         | visitors                         | candidates                         | visitors                             |
| 1999 | floyd    | floods     | clinton                             | congress                         | clinton                             | congress                            |
| 2000 | —        | —         | business                            | message                          | business                            | message                             |
| 2001 | —        | —         | campaign                            | bush                             | campaign                            | bush                                |
| 2002 | —        | —         | bush                                | supporters                       | bush                                | others                              |
| 2003 | isabel   | wings     | bush                                | governor                         | bush                                | support                             |
| 2004 | ivan     | francis   | obama                               | clinton                          | congress                            | clinton                             |
| 2005 | katrina  | rita      | candidates                         | bush                             | candidates                         | bush                                |
| 2006 | katrina  | category  | message                             | masters                          | column                              | study                               |
| 2007 | katrina  | tomando   | obama                              | illinois                         | bush                                | democrats                           |
| 2008 | katrina  | orleans   | obama                              | president-elect                  | clinton                             | bush                                |
| 2009 | —        | —         | obama                               | michelle                         | clinton                             | obama                               |
| 2010 | —        | —         | obama                               | campaign                         | clinton                             | obama                               |
| 2011 | category | larkin    | obama                               | bush                             | clinton                             | survival                            |
| 2012 | katrina  | irene     | obama                               | something                        | clinton                             | government                          |
| 2013 | katrina  | aftermath | obama                               | something                        | obama                               | clinton                             |
| 2014 | katrina  | aftermath | obama                               | home                             | obama                               | clinton                             |
| 2015 | katrina  | joaquin   | obama                               | candidate                        | clinton                             | obama                               |
| 2016 | blessing | category  | obama                               | hillary                          | clinton                             | obama                               |

Table 9 -- Top two closest terms to katrina$_{2005}$, obama$_{2016}$, clinton$_{1996}$ and hillary$_{2015}$ from 1990 to 2016

The empty lines in the katrina column correspond to years without significant hurricanes (i.e. class 2 and higher), according to the National Hurricane Center reports. Terms in bold are the correct answers but other terms are often quite relevant (illinois for obama, or campaign for both presidents.) The year 2006 should be ignored as this partition is an outlier with a much shorter vocabulary.
7. Order encoding and positional neighbors

As shown by Sahigren *et al.* 19, permutations are a means to encode order in word space (see also Widdows *et al.* 13.) Let’s define permutation \( \Pi^0 \) as the natural sequence of the first \( d \) integers,

\[
\Pi^0 = (0, 1, 2, \ldots d - 1)
\]

and \( \Pi^1 \) as a random permutation of the first \( d \) integers, then \( \Pi^{-1}, \Pi^2 \) and \( \Pi^{-2} \) are defined by

\[
\Pi^{-1} \times \Pi^1 \equiv \Pi^1 \times \Pi^{-1} = \Pi^0, \quad \Pi^2 = \Pi^1 \times \Pi^1 \quad \text{and} \quad \Pi^{-2} = \Pi^{-1} \times \Pi^{-1}
\]

If we randomly select a vector \( |T\rangle \) in a space of large enough dimension \( d \), the vector \( |V\rangle = \Pi^1 |T\rangle \) obtained by permutation of the indices will clearly be quasi-orthogonal 19 to \( |T\rangle \), provided \( \Pi^1 \) has no fixed point.

If we replace Equation 1 with the following:

\[
|V_i\rangle = \Pi^{-2}(\rho_{i-2} |T_{i-2}\rangle^*) + \Pi^{-1}(\rho_{i-1} |T_{i-1}\rangle^*) + \Pi^0(\rho_{i} |T_{i}\rangle^*) + \Pi^1(\rho_{i+1} |T_{i+1}\rangle^*) + \Pi^2(\rho_{i+2} |T_{i+2}\rangle^*)
\]

where \( |T_{i+\delta}\rangle^* \) is the normalized seed vector of term \( t_{i+\delta} \) and \( \rho_{i+\delta} \) is, as above, a multiplicative coefficient (for example a weight) the scalar product of (unnormalized) \( |V_i\rangle \) with the permuted, normalized seed vector \( \Pi^0 |T_{i+\delta}\rangle \) will yield \( \rho_{i+\delta} + \epsilon \) where \( \epsilon \) is a zero-centered noise contribution. This fact can be used to retrieve the most likely immediate positional neighbors of any word, as long as the noise remains acceptable 20.

| Predict word in next position | Predict words in previous and next positions |
|-------------------------------|------------------------------------------|
| president ? | york ? | columbia ? | supreme ? | world ? | year | ? trade ? | ? status ? |
| president bush | york city | columbia university | supreme court | world war | 1990 | national trade center | united states government |
| president bush | york city | columbia university | supreme court | world war | 1991 | national trade center | united states american |
| president bush | york city | columbia university | supreme court | world war | 1992 | national trade center | united states government |
| president clinton | york city | columbia university | supreme court | world war | 1993 | world trade center | united states government |
| president clinton | york city | columbia university | supreme court | world war | 1994 | national trade center | united states government |
| president clinton | york city | columbia university | supreme court | world war | 1995 | world trade center | united states government |
| president clinton | york city | columbia university | supreme court | world war | 1996 | national trade center | united states party |
| president clinton | york times | columbia university | supreme court | world war | 1997 | world trade center | united states government |
| president clinton | york times | columbia university | supreme court | world war | 1998 | world trade center | united states government |
| president clinton | york times | columbia university | supreme court | world war | 1999 | world trade center | united states party |
| president clinton | york times | columbia university | supreme court | world series | 2000 | world trade center | united states party |
| president bush | york times | columbia university | supreme court | world trade | 2001 | world trade center | united states american |
| president bush | york times | columbia university | supreme court | world war | 2002 | world trade center | united states american |
| president bush | york times | columbia university | supreme court | world war | 2003 | world trade center | united states american |
| president bush | york city | columbia university | supreme court | world war | 2004 | world trade center | united states american |
| president bush | york city | columbia university | supreme court | world war | 2005 | world trade center | united states american |
| president clinton | york times | columbia university | supreme court | world theater | 2006 | east side street | united states park |
| president bush | york city | columbia university | supreme court | world war | 2007 | world trade center | united states american |
| president obama | york city | columbia university | supreme court | world war | 2008 | world trade obama | united states american |
| president obama | york times | columbia university | supreme court | world war | 2009 | senator obama | united states government |
| president obama | york times | columbia university | supreme court | world series | 2010 | world trade center | united states party |
| president obama | york times | columbia university | supreme court | world war | 2011 | world trade center | united states government |
| president obama | york city | columbia university | supreme court | world war | 2012 | world trade obama | united states government |
| president obama | york city | columbia university | supreme court | world war | 2013 | world trade center | united states government |
| president obama | york city | columbia university | supreme court | world war | 2014 | world trade center | united states government |
| president obama | york city | columbia university | supreme court | world series | 2015 | world trade center | united states government |
| president obama | york city | columbia university | supreme court | world war | 2016 | world trade center | united states party |

Table 10 -- Positional neighbors of selected words

Except for the year 2006, which should be ignored as it is an outlier with a much shorter vocabulary, the guessed words are coherent with what would be expected.
This is shown in Table 10: ignoring outlying year 2006, the term following president is almost always the name of the actual president in the relevant year, york is followed by either times or city, supreme is always followed by court, series and war may both occur after world. As should be expected, predictions are less effective in the trigram situation since, for example, the digrams world trade and trade center do occur in other trigrams than world trade center which occurs 77% of the time; united states government occurs only half the time because united states occurs much more frequently (47,463 times vs. 673 times.)

8. Conclusions

The linearity of random vectors techniques enables the building of time series of semantic spaces which can then be compared and combined linearly without loss of precision over any time period to detect diachronic semantic shifts. In this article, we have demonstrated several uses of this approach: visualization of time trajectories of polysemous words such as amazon or apple, detection of semantic drifts and of gender bias across time, revealing the shifting instantiations of stable concepts such as hurricane or president. This very fast, linear approach can easily be distributed over many processors to follow real time language evolutions and the resulting, time-dependent semantic spaces can then be combined at will by simple additions or subtractions. This being a statistical process, its time resolution will be dependent on sample size and thus on the volume (number of words) of each time slice: one year is a lower limit for the sample of the New York Times\textsuperscript{2} we used but new Wikipedia downloads can be followed as they are posted, new patents and innovations can be followed on a weekly basis and streams of social media such as those of Twitter or Facebook could probably be followed with a time resolution of a few seconds.

Bibliography

[1] Mikolov, T., Chen, K., Corrado, G. and Dean, J., Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781v3, 2013.
[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J., Distributed Representations of Words and Phrases and their Compositionality, arXiv:1310.4546v1, 2013.
[3] Goldberg, Y. and Levy, O., Neural Word Embedding as Implicit Matrix Factorization, In Advances in Neural Information Processing Systems 27 (NIPS), 2014.
[4] Levy, O., Segalard, A., Goldberg, Y., A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments, arXiv:1608.05426v2 [cs.CL], 2017.
[5] Arora, S. Word Embeddings: Explaining their properties, http://www.offconvx.org/2016/02/14/word-embeddings-2/ Downloaded on May 8, 2018.
[6] Yao, Z., Sun, Y., Ding, W., Rao, N., and Xiong, H., Dynamic Word Embeddings for Evolving Semantic Discovery, In WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining, February 5 — 9, 2018, arXiv:1703.00607v2, Feb. 13, 2018.
[7] Yao, Z., Sun, Y., Ding, W., Rao, N., and Xiong, H., The data is available at https://sites.google.com/site/zijunyaorutgers/ and was downloaded on February 23, 2018.
[8] Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S., Statistically Significant Detection of Linguistic Change, arXiv:1411:3315v1, Nov. 12, 2014.
[9] Hamilton, W., Leskovec, J., and Jurafsky, D., Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change, arXiv:1605.09096v4 [cs.CL], September 2016.
[10] Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes, Proceedings of the National Academy of Sciences (PNAS), March 2018. Data and code related to this paper are available on GitHub https://github.com/nikhgarg/EmbeddingDynamicStereotypes (downloaded April 17, 2018).
[11] Firth, J. R. stated that “You shall know a word by the company it keeps”, in A Synopsis of Linguistic Theory, 1930-1955, Studies in Linguistic Analysis, Special volume of the Philological Society, Oxford, UK, 1962.
[12] Sahlgren, M., The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces, PhD Dissertation, Stockholm University, Sweden, 2006.
[13] Widdows, D. and Cohen, T., The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics, Fourth IEEE International Conference on Semantic Computing (IEEE ICSC2010), 2010.
[14] Delpech, J.-F., Unsupervised word sense disambiguation in dynamic semantic spaces, arXiv:1802.02605v2, Feb. 8, 2018.
[15] Delpech, J.-F. Semantic Technology-Assisted Review (STAR): Document analysis and monitoring using random vectors, arXiv:1711.10307, 2017.
[16] The fact that sauce disappears from the neighborhood of food terms doesn’t mean that the term has disappeared (it is in fact only slightly less frequent in the second epoch) but simply that it occurs less frequently within windows containing the reference term.
After Garg et al. 10, the following terms have been used to check stereotypes:
- Appearance terms such as alluring, voluptuous, homely, muscular... from  
  https://github.com/nikhgarg/EmbeddingDynamicStereotypes/blob/master/data/adjectives_appearance.txt
- Intelligence terms such as precocious, inventive, adaptable, reflective... from  
  https://github.com/nikhgarg/EmbeddingDynamicStereotypes/blob/master/data/adjectives_intelligencegeneral.txt
- Occupation terms such as judge, dancer, optometrist... from  
  https://github.com/nikhgarg/EmbeddingDynamicStereotypes/blob/master/data/occupation_percentages_gender_occ1950.csv
- Terms exclusively relating to men such as son, his, him, father, man... from  
  https://github.com/nikhgarg/EmbeddingDynamicStereotypes/blob/master/data/male_pairs.txt
- Terms exclusively relating to women such as daughter, mother, woman, girl... from  
  https://github.com/nikhgarg/EmbeddingDynamicStereotypes/blob/master/data/female_pairs.txt

NOAA—HRS Complete list of continental U.S. landfalling hurricanes  
http://www.aoml.noaa.gov/hrd/tcfaq/E23.html downloaded on April 16, 2018.

Sahlgren, M., Holst, A., Kanerva, P., Permutations as a Means to Encode Order in Word Space  
Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci’08), Washington D.C., 2008.

In this experiment, we use vectors of dimension 300 while Sahlgren et al. 19 use vectors are of dimension 3000,  
which should reduce noise by a factor $\sqrt{10}$. 