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.* [1, 2] and other investigators [3, 4], successful word representation methods inspired from neural-network language modeling have been demonstrated and explained. These representations are nonlinear [5] and the semantic spaces in which they are embedded are randomly oriented as the cost functions used for training are invariant under rotation [6]. 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.* [6] 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 τ controls how fast the embedding may change between time slices and thus the time resolution. Obviously, even though alignement 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 [7].

In a recent article, Kulkarni *et al.* [8] 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.*[9] 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.*[10] using word embeddings derived from various sources.

These nonlinear approaches rely on a foundational assumption in Natural Language Processing (NLP), the *distributional hypothesis*[11], 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[12,13,14] relies on the fact that in $\mathbb{R}^d$ an exponentially large number $N \gg d$ of random vectors quasi-orthogonal to each other[15] 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

$$\dots \qquad t_{i-3} \qquad t_{i-2} \qquad t_{i-1} \qquad t_i \qquad t_{i+1} \qquad t_{i+2} \qquad t_{i+3} \qquad \dots$$

will increment term vector $|\mathcal{T}_i\rangle$ by the sum of four seed vectors $|\mathcal{T}^s\rangle$:

$$|\mathcal{T}_i\rangle \mathrel{+}= \rho_{i-2}|\mathcal{T}^s_{i-2}\rangle + \rho_{i-1}|\mathcal{T}^s_{i-1}\rangle + \rho_{i+1}|\mathcal{T}^s_{i+1}\rangle + \rho_{i+2}|\mathcal{T}^s_{i+2}\rangle \tag{1}$$

where the $\rho_{i+\delta}$ are multiplicative coefficients, for example weights. If term t_i occurs N times, vector $|\mathcal{T}_i\rangle$ 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 $|\mathcal{T}_i\rangle$ 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 $\mathcal{S}$, the similarity σ_{ij} between terms t_i and t_j is the scalar product of the associated, normalized term vectors:

$$\sigma_{ij} = \langle \mathcal{T}_i | \mathcal{T}_j \rangle$$

(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 $|\mathcal{T}^k_i\rangle$ are embedded in semantic space $\mathcal{S}_k$ and one can define symbolically the semantic space $|\mathcal{S}_K\rangle$ of the whole corpus

$$\mathcal{S}_K = \sum_{k \in K} \mathcal{S}_k \tag{2}$$

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[7] 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 $\mathcal{S}_K$ was also created by combining all 27 partial spaces $\mathcal{S}_k$.

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.* [6]: 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.

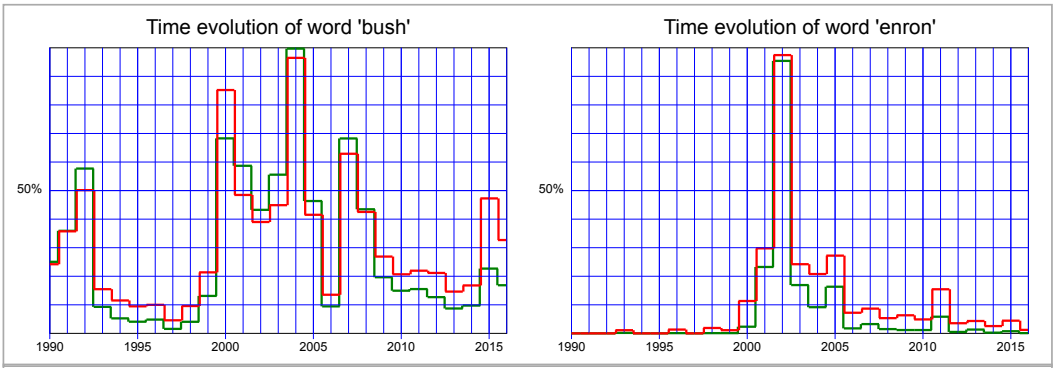

Figure 1 -- Time evolutions of words *bush* and *enron*
Norms are in red, frequencies in green and both are normalized: for *bush* the maximum frequency occurs in 2004 (3490 occurrences), as well as the maximum norm (1833); for *enron*, the maximums are respectively 1145 and 877. The dips in 2006 are 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.* [6] 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 $|\mathcal{T}_t^K\rangle$ of term t from the *total* semantic space $\mathcal{S}_K$
2 : Find a set R of representative neighbors of term t in $\mathcal{S}_K$
3 : **for each epoch** $k \in K$
4 : Extract and normalize the vector $|\mathcal{T}_r^k\rangle$ of each member $r \in R$ from the *partial* semantic space $\mathcal{S}_k$
5 : Compute the similarity $\sigma_r^k = \langle \mathcal{T}_r^k | \mathcal{T}_t^K \rangle$ 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 $\mathcal{S}_K$ is compared to vectors extracted from partitions $\mathcal{S}_k$; 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 items 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
deforestation																											
rainforest																											
iquitos																											
manaus																											
bezos																											
streaming																											
online																											
downloads																											
netflix																											
Nr. of occurrences	**18**	**36**	**7**	**19**	**5**	**1**	**10**	**13**	**26**	**76**	**154**	**75**	**62**	**54**	**70**	**75**	**5**	**73**	**91**	**76**	**71**	**101**	**196**	**102**	**399**	**131**	**83**

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. Items 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.)