A bottom up approach to category mapping and meaning change

Haim Dubossarsky
The Edmond and Lily Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem 91904, Israel
haim.dub@gmail.com

Yulia Tsvetkov
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
ytsvetko@cs.cmu.edu

Chris Dyer
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
cdyer@cs.cmu.edu

Eitan Grossman
Linguistics Department and the Language, Logic and Cognition Center
The Hebrew University of Jerusalem
Jerusalem 91904, Israel
eitan.grossman@mail.huji.ac.il

Abstract

In this article, we use an automated bottom-up approach to identify semantic categories in an entire corpus. We conduct an experiment using a word vector model to represent the meaning of words. The word vectors are then clustered, giving a bottom-up representation of semantic categories. Our main finding is that the likelihood of changes in a word's meaning correlates with its position within its cluster.

1 Introduction

Modern theories of semantic categories, especially those influenced by Cognitive Linguistics (Geeraerts and Cuyckens, 2007), generally consider semantic categories to have an internal structure that is organized around prototypical exemplars (Geeraerts, 1997; Rosch, 1973).

Historical linguistics uses this conception of semantic categories extensively, both to describe changes in word meanings over the years and to explain them. Such approaches tend to describe changes in the meaning of lexical items as changes in the internal structure of semantic categories. For example, (Geeraerts, 1999) hypothesizes that changes in the meaning of a lexical item are likely to be changes with respect to the prototypical 'center' of the category. Furthermore, he proposes that more salient (i.e., more prototypical) meanings will probably be more resistant to change over time than less salient (i.e., less prototypical) meanings.

Despite the wealth of data and theories about changes in the meaning of words, the conclusions of most historical linguistic studies have been based on isolated case studies, ranging from few single words to few dozen words. Only recently though, have usage-based approaches (Bybee, 2010) become prominent, in part due to their compatibility with quantitative research on large-scale corpora (Geeraerts et al., 2011; Hilpert, 2006; Sagi et al., 2011). Such approaches argue that meaning change, like other linguistic changes, are to a large extent governed by and reflected in the statistical properties of lexical items and grammatical constructions in corpora.

In this paper, we follow such usage-based approaches in adopting Firth's famous maxim "You shall know a word by the company it keeps," an axiom that is built into nearly all diachronic corpus linguistics (see Hilpert and Gries, 2014 for a state-of-the-art survey). However, it is unclear how such 'semantic fields' are to be identified. Usually, linguists' intuitions are the primary evidence. In contrast to an intuition-based approach, we set out from the idea that categories can be extracted from a corpus, using a 'bottom up' methodology. We demonstrate this by automatically categorizing the entire lexicon of a corpus, using clustering on the output of a word embedding model.

We analyze the resulting categories in light of the predictions proposed in historical linguistics regarding changes in word meanings, thus providing a full-scale quantitative analysis of changes in the meaning of words over an entire corpus. This approach is distinguished from previous research by two main characteristics: first, it provides an exhaustive analysis of an entire corpus; second, it is fully bottom-up, i.e., the categories obtained emerge from the data, and are not in any way based on linguists' intuitions. As such, it provides an independent way of evaluating linguists' intuitions, and has the potential to turn up new, unintuitive or even counterintuitive

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

facts about language usage, and hence, by hypothesis, about knowledge of language.

2 Literature review

Some recent work has examined meaning change in large corpora using a similar bottom-up approach and word embedding method (Kim et al., 2014). These works analyzed trajectories of meaning change for an entire lexicon, which enabled them to detect if and when each word changed, and to measure the degree of such changes. Although these works are highly useful for our purposes, they do not attempt to explain why words differ in their trajectories of change by relating observed changes to linguistic parameters.

Wijaya and Yeniterzi (2011) used clustering to characterize the nature of meaning change. They were able to measure changes in meaning over time, and to identify which aspect of meaning had changed and how (e.g., the classical semantic changes known as 'broadening,' 'narrowing,' and 'bleaching'). Although innovative, only 20 clusters were used. Moreover, clustering was only used to describe patterns of change, rather than as a possible explanatory factor.

3 Method

A distributed word vector model was used to learn the context in which the words-of-interest are embedded. Each of these words is represented by a vector of fixed length. The model changes the vectors' values to maximize the probability in which, on average, these words could predict their context. As a result, words that predict similar contexts would be represented with similar vectors. This is much like linguistic items in a classical structuralist paradigm, whose interchangeability at a given point or 'slot' in the syntagmatic chain implies they share certain aspects of function or meaning.

The vectors' dimensions are opaque from a linguistic point of view, as it is still not clear how to interpret them individually. Only when the full range of the vectors' dimensions is taken together does meaning emerges in the semantic hyperspace they occupy. The similarity of words is computed using the cosine distance between two word vectors, with 0 being identical vectors, and 2 being maximally different:

$$(1) \qquad 1 - \frac{\sum_{i=1}^{d} W_i \times W'_i}{\sqrt{\sum_{i=1}^{d} (W_i)^2} \times \sqrt{\sum_{i=1}^{d} (W'_i)^2}}$$

Where d is the vector's dimension length, and W_i and W_i' represent two specific values at the same vector point for the first and second words, respectively.

Since words with similar meaning have similar vectors, related words are closer to each other in the semantic space. This makes them ideal for clustering, as word clusters represent semantic 'areas,' and the position of a word relative to a cluster centroid represents its saliency with respect to the semantic concept captured by the cluster. This saliency is higher for words that are closer to their cluster centroid. In other words, a word's closeness to its cluster centroid is a measure of its prototypicality. To test for the optimal size of the 'semantic areas,' different numbers of clusters were tested. For each the clustering procedure was done independently.

To quantify diachronic word change, we train a word vector model on a historical corpus in an orderly incremental manner. The corpus was sorted by year, and set to create word vectors for each year such that the words' representations at the end of training of one year are used to initialize the model of the following year. This allows a yearly resolution of the word vector representations, which are in turn the basis for later analyses. To detect and quantify meaning change for each word-of-interest, the distance between a word's vector in two consecutive decades was computed, serving as the degree of meaning change a word underwent in that time period (with 2 being maximal change and 0 no change).

Having two representational perspectives – synchronic and diachronic – we test the hypothesis that words that exhibit stronger cluster saliency in the synchronic model – i.e., are closer to the cluster centroid – are less likely to change over time in the diachronic model. We thus measure the correlation between the distance of a word to its cluster centroid at a specific point in time and the degree of change the word underwent over the next decade.

4 Experiment

We used the 2nd version of Google Ngram of fiction English, from which 10 millions 5-grams were sampled for each year from 1850-2009 to serve as our corpus. All words were lower cased.

Word2vec (Mikolov et al., 2013) was used as the distributed word vector model. The model was initiated to 50 dimensions for the word vectors' representations, and the window size for context set to 4, which is the maximum size giv-

en the constraints of the corpus. Words that appeared less than 10 times in the entire corpus were discarded from the model vocabulary. Training the model was done year by year, and versions of the model were saved in 10 year intervals from 1900 to 2000.

The 7000 most frequent words in the corpus were chosen as words-of-interest, representing the entire lexicon. For each of these words, the cosine distance between its two vectors, at a specific year and 10 years later, was computed using (1) above to represent the degree of meaning change. A standard K-means clustering procedure was conducted on the vector representations of the words for the beginning of each decade from 1900 to 2000 and for different number of clusters from 500 until 5000 in increments of 500. The distances of words from their cluster centroids were computed for each cluster, using (1) above. These distances were correlated with the degree of change the words underwent in the following ten-year period. The correlation between the distance of words from random centroids of different clusters, on the one hand, and the degree of change, on the other hand, served as a control condition.

4.1 Results

Table 1 shows six examples of clusters of words. The clusters contain words that are semantically similar, as well as their distances from their cluster centroids. It is important to stress that a centroid is a mathematical entity, and is not necessarily identical to any particular exemplar. We suggest interpreting a word's distance from its cluster's centroid as the degree of its proximity to a category's prototype, or, more generally, as a measure of prototypicality. Defined in this way, *sword* is a more prototypical exemplar than *spear* or *dagger*, and *windows*, *shutters* or *doors* may be more prototypical exemplars of a *cover of an entrance* than *blinds* or *gates*. In addition, the clusters capture near-synonyms, like *gallop* and *trot*, and level-of-category relations, e.g., the modal predicates *allowed, permitted, able*. The very fact that the model captures clusters and distances of words which are intuitively felt to be semantically closer to or farther away from a category prototype is already an indication that the model is on the right track.

sword, 0.06 *spear, 0.07* *dagger, 0.09*	*allowed, 0.02* *permitted, 0.04* *able, 0.06*
shutters, 0.04 *windows, 0.05* *doors, 0.08* *curtains, 0.1* *blinds, 0.11* *gates, 0.13*	*hat, 0.03* *cap, 0.04* *napkin, 0.09* *spectacles, 0.09* *helmet, 0.13* *cloak, 0.14*
gallop, 0.02 *trot, 0.02*	*handkerchief, 0.14* *cane, 0.15*

Table 1: Example for clusters of words using 2000 clusters and their distance from their centroids.

Figure 1 shows the analysis of changes in word meanings for the years 1950-1960. We chose this decade at random, but the general trend observed here obtains over the entire period (1900-2000). There is a correlation between the words' distances from their centroids and the degree of meaning change they underwent in the following decade, and this correlation is observable for different number of clusters (e.g., for 500 clusters, 1000 clusters, and so on). The positive correlations ($r>.3$) mean that the more distal a word is from its cluster's centroid, the greater the change its word vectors exhibit the following decade, and vice versa.

Crucially, the correlations of the distances from the centroid outperform the correlations of the distances from the prototypical exemplar, which was defined as the exemplar that is the closest to the centroid. Both the correlations of the distance from the cluster centroid and of the distance from the prototypical exemplar were significantly better than the correlations of the control condition (all *p's < .001 under permutations tests*).

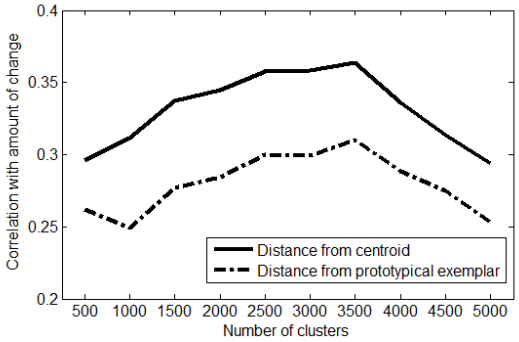

Figure 1. Change in the meanings of words correlated with distance from centroid for different numbers of clusters, for the years 1950-1960.

In other words, the likelihood of a word changing its meaning is better correlated with the distance from an abstract measure than with the distance from an actual word. For example, the likelihood of change in the *sword-spear-dagger* cluster is better predicted by a word's closeness

to the centroid, which perhaps could be conceptualized as a non-lexicalized 'elongated weapon with a sharp point,' than its closeness to an actual word, e.g., *sword*. This is a curious finding, which seems counter-intuitive for nearly all theories of lexical meaning and meaning change.

The magnitude of correlations is not fixed or randomly fluctuating, but rather depends on the number of clusters used. It peaks for about 3500 clusters, after which it drops sharply. Since a larger number of clusters necessarily means smaller 'semantic areas' that are shared by fewer words, this suggests that there is an optimal range for the size of clusters, which should not be too small or too large.

4.2 Theoretical implications

One of our findings matches what might be expected, based on Geeraert's hypothesis, mentioned in Section 1: a word's distance from its cluster's most prototypical exemplar is quite informative with respect to how well it fits the cluster (Fig. 1). This could be taken to corroborate Roschian prototype-based views. However, another finding is more surprising, namely, that a word's distance from its real centroid, an abstract average of the members of a category by definition, is even better than the word's distance from the cluster's most prototypical exemplar.

In fact, our findings are consonant with recent work in usage-based linguistics on attractors, 'the state(s) or patterns toward which a system is drawn' (Bybee and Beckner, 2015). Importantly, attractors are 'mathematical abstractions (potentially involving many variables in a multidimensional state space)'. We do not claim that the centroids of the categories identified in our work are attractors – although this may be the case – but rather make the more general point that an abstract mathematical entity might be relevant for knowledge of language and for language change.

In the domain of meaning change, the fact that words farther from their cluster's centroid are more prone to change is in itself an innovative result, for at least two reasons. First, it shows on unbiased quantitative grounds that the internal structure of semantic categories or clusters is a factor in the relative stability over time of a word's meaning. Second, it demonstrates this on the basis of an entire corpus, rather than an individual word. Ideas in this vein have been proposed in the linguistics literature (Geeraerts, 1997), but on the basis of isolated case studies which were then generalized.

5 Conclusion

We have shown an automated bottom-up approach for category formation, which was done on an entire corpus using the entire lexicon.

We have used this approach to supply historical linguistics with a new quantitative tool to test hypotheses about change in word meanings. Our main findings are that the likelihood of a word's meaning changing over time correlates with its closeness to its semantic cluster's most prototypical exemplar, defined as the word closest to the cluster's centroid. Crucially, even better than the correlation between distance from the prototypical exemplar and the likelihood of change is the correlation between the likelihood of change and the closeness of a word to its cluster's actual centroid, which is a mathematical abstraction. This finding is surprising, but is comparable to the idea that attractors, which are also mathematical abstractions, may be relevant for language change.

Acknowledgements

We thank Daphna Weinshall (Hebrew University of Jerusalem) and Stéphane Polis (University of Liège) for their helpful and insightful comments. All errors are, of course, our own.

Reference

Joan Bybee. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Joan Bybee and Clay Beckner. 2015. Emergence at the cross linguistic level. In B. MacWhinney and W. O'Grady (eds.), *The handbook of language emergence*, 181-200. Wiley Blackwell.

Dirk Geeraerts. 1997. *Diachronic prototype semantics. A contribution to historical lexicology*. Oxford: Clarendon Press.

Dirk Geeraerts. 1999. Diachronic Prototype Semantics. A Digest. In: A. Blank and P. Koch (eds.), *Historical semantics and cognition*. Berlin & New York: Mouton de Gruyter.

Dirk Geeraerts, and Hubert Cuyckens (eds.). 2007. *The Oxford handbook of cognitive linguistics*. Oxford: Oxford University Press.

Dirk Geeraerts, Caroline Gevaerts, and Dirk Speelman. 2011. How Anger Rose: Hypothesis

Testing in Diachronic Semantics. In J. Robynson and K. Allan (eds.), *Current methods in historical semantics*, 109-132. Berlin & New York: Mouton de Gruyter.

Martin Hilpert. 2006. Distinctive Collexeme Analysis and Diachrony. *Corpus Linguistics and Linguistic Theory*, 2 (2): 243–256.

Martin Hilpert and Stefan Th. Gries. 2014. Quantitative Approaches to Diachronic Corpus Linguistics. In M. Kytö and P. Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics*. Cambridge: Cambridge University Press, 2014.

Yoon Kim, Yi-I Chiu, Kentaro Haraki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 61-65. Baltimore, USA.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. *Proceedings of NAACL-HLT 2013:* 746–751. Atlanta, Georgia.

Eleanor H. Rosch. 1973. Natural Categories. *Cognitive Psychology* 4 (3): 328–350.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with latent semantic analysis. In K. Allan and J.A. Robinson (eds.), *Current methods in historical semantics*, 161-183. Berlin & New York: Mouton de Gruyter.

Derry T. Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web (DETECT '11*) 35-40. Glasgow, United Kingdom.