Identification of Semantic Shifts in English Using Word Embeddings

Peter Schoener

April 12, 2018

Abstract

In this paper, I attempt to identify words which have undergone significant semantic shifts over time. I do this using the deviations in the cosines of the most similar words.

1 Introduction

Semantic shift, the change in meaning of a word over domain or time, is an important phenomenon to take into account when working with narrow corpora or individual texts. It can have a profound effect on the meaning of what is being read, leading to misinterpretation, or may simply look unfamiliar, in which case it is useful to analyze the meanings it may have taken in that particular context.

The ability to automatically flag drifting words and and examine them more closely would also be helpful for understanding the causes, rate, and markers of shifts, which would help not only with language reconstruction, but with prediction and identification of emerging changes.

Perhaps most importantly, shifts must be taken into account when working with broad corpora. The distributional hypothesis, for example, is widely accepted, but when working with a corpus that extends far into the past and attempting to apply it to current language, one might find that the learnt embeddings of certain words are inaccurate for the present, being an average of their current and past meanings.

There is no natural or direct way to quantify the extent of drift, and so my method uses a metric internally and attempts externally only to rank or generate a list of words which should be examined for shifts.

2 Related Work

This paper is a more or less direct extension of a paper by Kutuzov and Kuzmenko (c. 2016, not published in this form). That paper looks at shifts as marked by deviation from an averaged set by a specific set in the list of embeddings most similar to that being examined. By counting the number of overlapping words in the lists of the ten closest, they determine whether or not a substantial shift has occurred.

The experiment was originally planned to be more along the lines of Leeuwenberg et al., 2016, but after changing the domain only their methods for evaluating cosines were considered. The relative cosine idea did not, however, affect the usefulness of the final metric.

3 Method

This experiment used the British Parliament's Hansard Corpus since it is large enough to have substantial amounts of data even for the earliest years. It totals 1.6 billion tokens, though it seems not all of these are indexed by date; only those indexed could be used since the goal was to separate the models by year. The extracted raw text contains about 350 million tokens, with about 17,000 in the smallest year-specific set.

The text contains many abbreviations, so rather than tokenizing directly or navely it made sense the NLTK sentence splitter and then the NLTK tokenizer. The text was normalized to lowercase due to irregular and/or frequent capitalization in certain parts of the record. The embeddings used were Word2Vec with 50 dimensions, a rather low number but necessary due to computing power constraints. Still, the normal intuitions around embedding models held.

In order to account for rare tokens and sparse individual years, the embedding models were smoothed with a window of five years. This window was chosen so as to keep even fast drifts noticeable while making the embeddings dense enough to use. Only the words that appear in all windows could be considered by the drift metric, so it is important that rare words, being the most prone to drift, not be entirely filtered out.

The metric works by creating the set union of the nearest ten neighbor sets of a word over all windows. This is why it is important that the word be present in all windows; its absence would skew toward a smaller union, and moreover this would be an indicator that it is an uncommon token with imprecise embedding values. The metric is as follows:

$$f(w) = \sum_{w' \in W} \underset{y}{\sigma}(cos(w_y, w'_y))$$

that is, the sum over all words in the neighbor list union W of the standard deviation over all years y in cosine to that word. For a word not present in a given year, the cosines will clearly all be zero, as will therefore the deviation, skewing the sum against marking drift as having occurred. However, this metric has the advantage, given sufficient data, of marking both words which move relative to their neighbors and those which get new neighbors altogether.

The list of all words to which the metric can be applied are then ranked according to it, and the words can be evaluated by hand in descending order. Since this paper is mostly about a method for identifying drift rather than specific instances of it, I did not exhaustively examine the drift of the words returned beyond what was necessary to check the validity of the ranking.

4 Results

Filtering for only words which occur in all windows, I arrived at a list of just over a thousand. While this is not an incredibly long list and does not include many uncommon words — which might have been subject to more severe drift — it does rank several common words which have drifted, as well as some more or less technical vocabulary which has markedly higher frequency and/or different meaning in the context of the corpus (legislative discussions).

Among the highest flagged words were some which underwent substantial meaning change, such as "before," which seems to have steadily gained its locative sense ("out," opposite of "within," etc.) starting in the 1860s. "say" also appears to have started introducing postulates (similarly to "think," "imagine," "conceive," etc.) around the 1890s, its closeness to the original meaning remaining the same. The versatile preposition "on" seems to have become more commonly used to introduce a topic ("concerning," "regarding," etc.) although in the sense of "to be on about sth." ("discussing") it first lost popularity and then fluctuated, never fully losing this meaning.

There are other, less direct changes that can be seen: "could" and "would" are supposed to have become less tentative, closing in on factives. For example, they both moved closer to various forms of "is" and lost similarity to such words as "think". However, I would speculate that this is due to their always having been found in the same sorts of contexts, which is exactly what the model captures, and these factives becoming more common in later years.

There were also notable fluctuations in the cooccurence of "ought" and "to," showing that the stylistic choice of "X ought Y" versus "X ought to Y" tipped back and forth a few times during the examined period. This could also be an indicator of the use of "ought" in its nonmodal sense, or of the verb being implied.

However, there were many words which clearly underwent no change and nonetheless topped the ranking. Stable particles such as "a," "if," and "no" were in the top few despite clearly having no actual meaning change.

5 Conclusion

One explanation of the strange inclusion of very stable words at the top of the list is that they are not similar enough to any words to have stable correspondences; while "the" is in the list of similar words for "a," so are 65 others which have no such similarity, at least not in terms of meaning.

An obvious continuation of this experiment would be to rerun it on a larger dataset, possibly one with more general domain, to fill in the gaps that limit the list of words that can be evaluated. It would also be interesting to see how well the metric works on words that do not appear in all years, but it should be borne in mind that the current setup, especially evaluating the words on the metric, was already fairly computationally expensive and allowing short gaps would vastly increase the space to be evaluated.

Also, it would be interesting to see the effects of using a lemmatizer during preprocessing. Many of the similar words captured were alternate forms of each other, and consolidating them would hopefully make the effects on them more visible. It would also free up space in the top ten for more words, which could greatly affect the final outcome since the new additions would be more distant and possibly more mobile.

This is not necessarily desirable; the current setup already captures dozens of neighbors for each word, obscuring the closer meanings and possibly skewing the metric with changes in cosine that do not actually represent changes in meaning. It may therefore be better to reduce the number of neighbors captured for the union, which could in turn more greatly reduce the size of that union; less similar and therefore less stable neighbors would not be counted.