Identification of Semantic Shifts in English Using Word Embeddings

Peter Schoener

August 3, 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 first 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, but in order to get more accurate results this was later increased to 100.

Although it might have limited the issues caused by the size of the dataset, the text was not lemmatized for two reasons. Some words have vastly different meanings in one particular form than in the rest, meaning the algorithm could be confused by the overly long neighbor list, and secondly this is generally caused by one particular form drifting away from the others rather than by the whole group drifting.

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 under 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).

The top 100 words according to the metric are as follows:

word	drift
separate	499.00
liverpool	499.22
restriction	499.73
private	500.12
here	500.85
actual	503.23
natural	506.12
thus	506.33
sir	512.19
relative	514.25
district	521.20
chairman	522.18
light	522.73
began	523.44
continuance	523.90
propriety	532.08
directors	539.12
temporary	540.50
crisis	544.71
respectable	546.91
except	550.70
afterwards	550.98
none	555.18
particulars	556.38

word	drift
discipline	556.41
above	557.75
)	558.83
attendance	562.22
regular	562.70
disposition	564.47
precisely	564.93
parliamentary	565.15
:	567.60
forces	567.97
?	569.46
usual	575.83
whenever	576.18
lie	578.79
exemption	582.68
renewal	585.45
arms	588.48
materially	591.85
young	594.89
line	596.70
latter	606.63
censure	606.69
lieutenant	609.40
deficiency	619.78
enemy	620.24
notes	621.75
fell	626.59
ways	628.69
king	634.82
attending	638.10
residence	640.48
whereas	645.87
mischief	646.62
near	648.92
mode	649.69
w.	654.73
principal	655.09
corps	668.18
rising	671.17
directly	673.41
clergy	684.38
forth	686.01
merchants	687.78

word	drift
mr.	690.19
resumed	690.69
voluntary	718.86
(	720.04
throne	721.10
honourable	724.88
dublin	728.00
follows	728.56
accordingly	741.85
attorney	774.53
notwithstanding	775.09
field	784.19
respecting	821.67
lately	832.44
francis	834.03
mr	837.30
regularly	840.65
distinctly	851.35
rose	887.50
approbation	907.94
c.	908.77
volunteer	937.33
!	965.13
concurred	968.07
ad	970.48
&	983.99
commander	1017.29
begged	1020.11
contest	1030.57
pursuant	1033.49
tending	1039.51
species	1048.04
principally	1195.64

It is worth noting that among the highest ranked words are some with multiple senses, such as "tending," "pursuant," "contest," "rose," etc. For many of these it can be seen in the neighbor lists that the neighbor lists have changed, shifting between the senses, but for some there just happens to be a large amount of noise. Although present, words relating to the senses of "pursuant" as in "pursuing" and "in accordance with" are not very prominent in the neighbor lists, with numbers, tokens garbled by encoding errors or typos, and seemingly unrelated words accounting for a large portion of the variation. Indeed this word must have undergone some degree of shift, but the fact that it was picked up by the algorithm seems largely to have been a coincidence caused by the noise

in the limited dataset.

Another category featuring prominently is punctuation and abbreviations, e.g. "& ," "!," "mr.,"; these tokens are extremely common and appear in a variety of contexts, being noisy by nature, rather than as a result of the size of the dataset.

Though there are some words in the top list that seem to have undergone drift, it really does seem to be mainly those which have even at present two senses, of which one became more dominant over time. Counterexamples might include "mischief", which is ranked very highly and has retained largely the same denotation, but with a considerably softened connotation.

One unintended consequence of the metric is that as a word comes into or falls out of use, a period of instability arises because an embedding can not be accurately trained, leading to marking of words that gain or lose relevance in a big way, such as "propriety."

However, on the other end of the scale, it can be seen that particles such as "a" with a value of 325.01, "no" with 195.71, and "if" with 151.87 are among the lowest ranked by this metric. Clearly, function words such as these should be stable, and this is captured by the metric. That being said, there are some words at the bottom which intuitively seem to have undergone drift; "would" does not have the same sense today that it does in much archaic text, but it is assigned a value of only 136.61.

5 Conclusion

One immediately apparent property of the metric is that it assigns astronomically high values to the top few most drifted items, with the top percentile being twice as highly marked as the tenth. This is not in and of itself a problem, but would suggest that limiting one or more of the factors, for example with a logarithm, could prevent runaway values caused by one outlying feature of a candidate.

The observation that many words flagged simply have two meanings may simply be a result of the dataset going back no further than the beginning of the 19th century; current English is entirely mutually intelligible with the English of then; indeed, much of the most famous literature today is at least that old; if people can still understand that language, clearly the old senses of common words are still known, possibly leading to the perception that a word has not changed meaning and merely tipped in favor of one. Still, this is clearly a form of drift.

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. As mentioned above another expansion would be in time range; some people already have trouble understanding Early Modern English, so a corpus going back an additional three hundred years would be useful. Any earlier than that and the text would predate the rough standardization of English leading to further difficulties, but already such a

broad corpus would be difficult to find or assemble, especially with enough data to get meaningful embeddings. This might limit even further the set of words considered, since many would not appear toward one end of the spectrum or the other.

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.

All in all the metric is a qualified success; it marks words for which dominance of senses shifts or which go in and out of fashion, while ignoring those that clearly do not change meaning. However, it does seem to have some trouble identifying clear drift, that is, a word losing one meaning and gaining another. The improvements outlined would be a good starting point, but there is clearly far more work to be done in order to get a clear signal out of the noise.