Topical advection as a baseline model for corpus-based lexical dynamics

Topical advection as a baseline model for corpus-based lexical dynamics

Andres Karjus[1], Richard A. Blythe[1,2], Simon Kirby[1], Kenny Smith[1]
[1]Centre for Language Evolution, School of Philosophy, Psychology and Language Sciences, University of Edinburgh; [2]School of Physics and Astronomy, University of Edinburgh
akarjus@exseed.ed.ac.uk, {r.a.blythe, simon.kirby, kenny.smith}@ed.ac.uk

1 Introduction

An important question in the field of corpus-based evolutionary language dynamics research is concerned with distinguishing selection-driven linguistic change from neutral evolution, and from changes stemming from language-external factors (cultural drift). A commonly used proxy for the popularity or selective fitness of an element is its corpus frequency. However, a number of recent works have pointed out that raw frequencies can often be misleading. We propose a model for controlling for drift in contextual topics in corpora - the topical-cultural advection model - and demonstrate that this simple measure is capable of accounting for a considerable amount of variability in word frequency changes in a corpus spanning two centuries of language use.

2 Background and motivation

There have been various proposals to carry over the selection and neutral drift paradigm from evolutionary biology, (where drift stands for differential replication without selection, cf. Croft (2000)), and apply similar tests to language data (Reali and Griffiths, 2010; Blythe, 2012; Ahern et al., 2016; Sindi and Dale, 2016). While previous research has been mostly concerned with distinguishing selection from drift in terms of frequencies, ours is a model for controlling for topical drift (somewhat similarly to Hamilton et al. (2016b), who contrast cultural and linguistic change). Clearly no linguistic element exists in isolation, without context. In order to objectively model the success or decline of an element, its context (or topic) should be taken into account. The potential effect of cultural processes and hot media topics on language usage patterns, as attested in corpora, have been often noted in recent studies. However, the way such phenomena are viewed varies: while 'culturomics' and related approaches treat word frequency changes as a way to study historical real-world changes (Michel et al., 2011), both on their own and as effects on language dynamics (Bochkarev et al., 2014; Petersen et al., 2012), a number of linguists have voiced concerns about relying on frequencies for linguistic inference without controlling for corpus composition in terms of register, genre and topic (Chelsey and Baayen, 2010; Lijffijt et al., 2012; Hinrichs et al., 2015; Szmrecsanyi, 2016; Calude et al., 2017).

3 The cultural-topical advection model

The cultural-topical advection model formalizes the following intuition: if a topic becomes more prevalent, then the words describing it, relating to it and possibly giving rise to it, should become more frequent as well, and vice versa with decline (with a clearer effect on topic-specific words). The term *advection* is borrowed from physics, denoting transport of particles by bulk motion or flow.

We define the 'topic' of a word as the set of words that are most strongly associated with the target word in a given period. This is inspired by the recent proposal of the APSym distributional semantics similarity metric, which is based on the intersection of the most strongly associated (mutual information weighted) co-occurring context words (Santus et al., 2016). The advection value of a word in a given t period w_t is defined as the weighted mean of the (smoothed) log changes in frequencies $(log(w_{freq_t}+1)-log(w_{freq_{t-1}}+1))$ of the ordered set of associated words N, weighted by their association score (i.e.,
$wMean(\{logChange(N_{i_t}) \mid i = 1, ...m\}, W_{1:m})$,

Proceedings of the Society for Computation in Linguistics (SCiL) 2018, pages 186-188.
Salt Lake City, Utah, January 4-7, 2018

where set of weights W corresponds to the PPMI association scores of the words in the set N and m is the number of context words to use.

We also implemented the advection measure using Latent Dirichlet Allocation (Blei et al., 2003), a more traditional topic model. The results in terms of the descriptive power of the model were rather similar. In an LDA-driven advection model, each topic is assigned a frequency change value, based on the (weighted) frequency changes of the words in the topics; the topical advection value of a target word is the (topic-word association weighted) mean of the change values of its topics. In contrast with LDA, our PPMI-weighted top-relevant-context-words based model requires almost no optimization of parameters (only choosing the m), is considerably simpler (and thus faster), and the results are easily traceable and interpretable, as each "topic" of a target is just a short list of top context words.

4 Results

We used the Corpus of Historical American English (COHA) (Davies, 2010) in order to get a sense of how well the topical-cultural advection model performs. Since cultural effects are likely the most pronounced on nouns (cf. also Hamilton et al. (2016a)), we only model the advection of common nouns; we use only content words from the co-occurrence vectors (of window size 5; $m = 75$), and set a (rather conservative) threshold of 100 occurrences per period for words to be included in the advection model, to maintain reliable semantics. We also experimented with adding "smoothing" to the input data to the topic models, in the sense of concatenating text from a target period and its preceding period, in order to better capture diminishing topics and words.

To test the descriptive power of the advection model, we correlate the log frequency changes of nouns to their respective advection (topic log change) values. For the first test, we include data points on frequency change across 19 decades (1820-2000) of all nouns that occur above the chosen frequency threshold at least once. For the second test, we use only a subset of ("persistent") nouns that always remain above the threshold and in the frequency band of $[20, 1000]$pmw, in the decades 1900-2000.

We find that, as expected, frequency changes do correlate significantly positively with advection, and that the aforementioned smoothing operation further improves the correlation. Table 1 illustrates the amount of variability in frequency changes described by advection (topic changes). The results remained fairly consistent across separate periods.

	no smooth	smooth
n unique words	7539	10076
n data points	75494	107096
PPMI vectors	0.2	0.31
LDA	0.17	0.25
n unique words	2004	2004
n data points	38076	38076
PPMI vectors$'$	0.26	0.38
LDA$'$	0.17	0.34

Table 1: The R^2 values of the two methods with and without smoothing. Top: models using all words that occur above the threshold at least once; frequency change data points from 19 decades (more data points in the smoothed versions: concatenated data results in more words being above the minimal threshold). Bottom half, separated by double line, marked with $'$: models using the persistent subset.

5 Conclusions

We conclude that advection can be considered a reasonably strong baseline for describing changes in word frequency. Obviously, the implementation is open to improvements and experimentation with the parameters. It would be fairly straightforward to use this approach as time series decomposition, by subtracting the advection value from the frequency change value, and reforming the frequency time series as a cumulative sum of the resulting values. This has potential to be useful in carrying out more objective tests of linguistic selection akin to Ahern et al. (2016), by removing or controlling for the topical-cultural element. As a baseline, it could be useful to models incorporating further effects of language change, such as structural-phonological properties (Szmrecsanyi, 2016) and content biases (Tamariz et al., 2014), polysemy (Hamilton et al., 2016b), socially conditioned variation (Samara et al., 2017),

network properties (Pierrehumbert et al., 2014) and other sociolinguistic effects (Calude et al., 2017).

In principle, the advection approach also could be used in other domains of cultural evolution, where there is diachronic data available about the co-occurrence of traits or properties (in lieu of context words) of cultural elements (in lieu of words).

References

Christopher A. Ahern, Mitchell G. Newberry, Robin Clark, and Joshua B. Plotkin. 2016. Evolutionary forces in language change. *arXiv preprint arXiv:1608.00938*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Richard A. Blythe. 2012. Neutral evolution: A null model for language dynamics. *Advances in complex systems*, 15(3-4).

V. Bochkarev, V. Solovyev, and S. Wichmann. 2014. Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface*, 11(101).

Andreea Simona Calude, Steven Miller, and Mark Pagel. 2017. Modelling loanword success a sociolinguistic quantitative study of Mori loanwords in New Zealand English. *Corpus Linguistics and Linguistic Theory*, 0(0), January.

Paula Chelsey and Harald R. Baayen. 2010. Predicting new words from newer words: Lexical borrowings in French. *Linguistics*, 48(6):1343–1374.

W. Croft. 2000. *Explaining Language Change: An Evolutionary Approach*. Explaining Language Change : an Evolutionary Approach. Longman.

Mark Davies. 2010. *The Corpus of Historical American English (COHA): 400 million words, 1810-2009. Available online at http://corpus.byu.edu/coha/.*

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2016:2116–2121.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Lars Hinrichs, Benedikt Szmrecsanyi, and Axel Bohmann. 2015. Which-hunting and the Standard English relative clause. *Language*, 91(4):806–836.

Jefrey Lijffijt, Tanja Sily, and Terttu Nevalainen. 2012. CEECing the baseline: Lexical stability and significant change in a historical corpus. In *Studies in Variation, Contacts and Change in English*, volume 10. Research Unit for Variation, Contacts and Change in English (VARIENG).

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

Alexander M. Petersen, Joel Tenenbaum, Shlomo Havlin, and H. Eugene Stanley. 2012. Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death. *Scientific Reports*, 2, March.

Janet B. Pierrehumbert, Forrest Stonedahl, and Robert Daland. 2014. A model of grassroots changes in linguistic systems. *arXiv preprint arXiv:1408.1985*.

F. Reali and T. L. Griffiths. 2010. Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of the Royal Society B: Biological Sciences*, 277(1680):429–436, February.

Anna Samara, Kenny Smith, Helen Brown, and Elizabeth Wonnacott. 2017. Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*, 94:85–114.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache. 2016. Testing APSyn against Vector Cosine on Similarity Estimation. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation, PACLIC 30, Seoul, Korea, October 28 - October 30, 2016*.

Suzanne S. Sindi and Rick Dale. 2016. Culturomics as a data playground for tests of selection: Mathematical approaches to detecting selection in word use. *Journal of Theoretical Biology*, 405:140 – 149.

Benedikt Szmrecsanyi. 2016. About text frequencies in historical linguistics: disentangling environmental and grammatical change. *Corpus Linguistics and Linguistic Theory*, 12(1):153–171.

M. Tamariz, T. M. Ellison, D. J. Barr, and N. Fay. 2014. Cultural selection drives the evolution of human communication systems. *Proceedings of the Royal Society B: Biological Sciences*, 281(1788):20140488–20140488, June.