Language Shift Odyssey and Dynamic Word Embedding

Xiaolei Huang, Michael J. Paul
Department of Information Science, University of Colorado Boulder

Language Shift Overtime

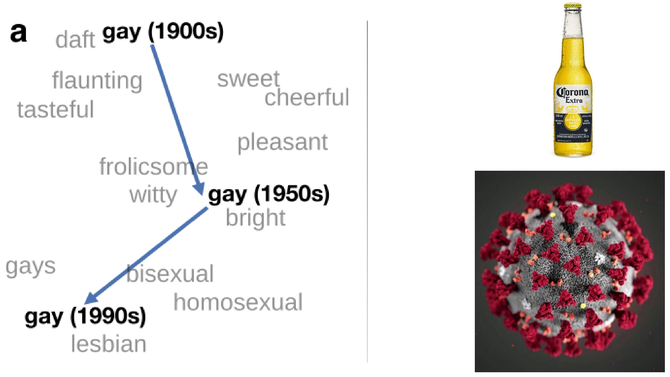


Figure 1. Language shift examples[1].

Data Overview

Dataset	Time intervals	Size
Amazon	1997-99, 2000-02, 2003-05, 2006-08, 2009-11, 2012-14	829K
Dianping	2009, 2010, 2011, 2012	2.98M
Twitter	2013, 2014, 2015, 2016	9.83K
Yelp-Hotel	2005-08, 2009-11, 2012-14, 2015-17	171K
Yelp-Rest.	2005-08, 2009-11, 2012-14, 2015-17	1.32M
Eco. News	1950-70, 1971-85, 1986-2000, 2001-14	6.29K
NYTimes	1990 – 2016 / 3	1.90 M

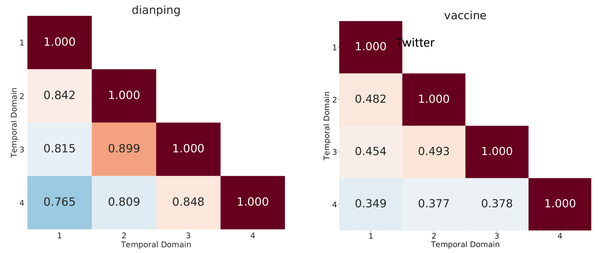
Impacts on ML Models

Classifiers perform best when applied to the same interval they were trained. Performance diminishes when applied to other time interval.

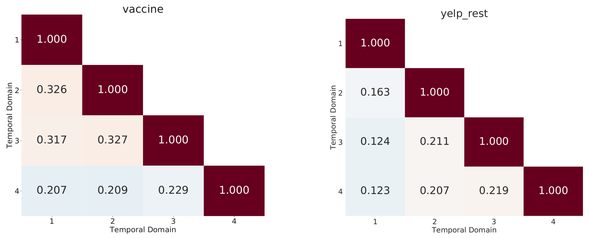
Figure 2. Document classification performance when training and testing on different years.

Why? Three perspectives of language shifts

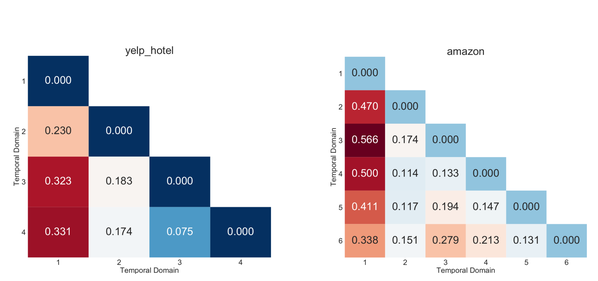
Word Usage Overlaps



Word Context Overlaps



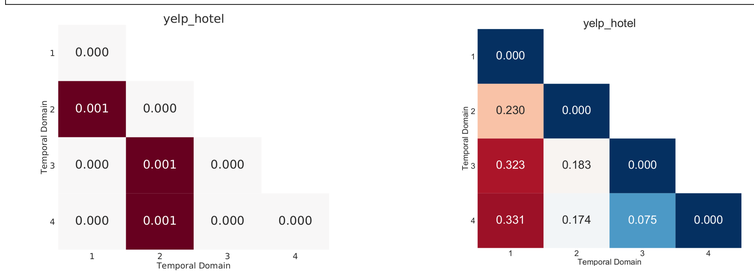
Semantic distance



Generally, closer time intervals share higher overlap and have smaller semantic distance shifts, and vice versa.

Does every word shift same speed?

Frequentist words **vs.** Polysemous words

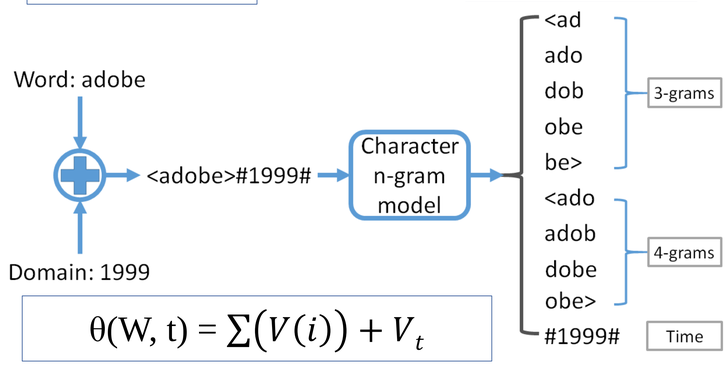


slow **fast**

Dynamic Word Embedding (DWE)

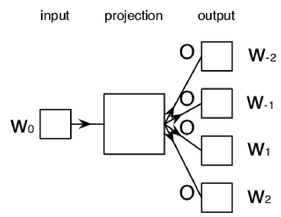
$$W_t = \theta(W, t)$$

Characters-level



$$\theta(W, t) = \sum(V(i)) + V_t$$

Skip-Ngram



Predict 2 sampled words co-occurred or not.

Understanding Temporality by DWE

90-92	93-95	96-98	99-01	02-04	05-07	08-10	11-13	14-16
pie	mac	pie	pie	mac	mac	app	chipset	intel
mac	pie	mac	oreo	pie	pie	mac	mac	mac
nut	apples	nut	nut	oreo	nut	aol	intel	aol
apples	nut	oreo	mac	apples	aol	sony	sony	sony
cake	3com	3com	cake	nut	idisk	pie	oreo	app

Table 1. : Top 5 closest words to “apple” across time on NYTimes.

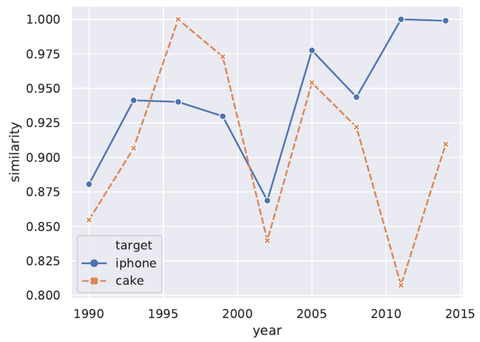


Figure 3. Dynamic semantic similarity between iphone and apple (blue) & cake and apple (orange) on the NYTimes dataset.

University of Colorado Boulder

Computational Language and EducAtion Research
C L E A R

Acknowledgement

The presenter would thank for poster payment from Computational Linguistics, Analytics, Search and Informatics MS Program. This work was supported in part by the National Science Foundation under award number IIS-1657338.

References

1. https://nlp.stanford.edu/projects/histwords/
2. Huang, Xiaolei, and Michael J. Paul. "Examining Temporality in Document Classification." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vol. 2. 2018.
3. Kutuzov, Andrey, et al. "Diachronic word embeddings and semantic shifts: a survey." Proceedings of the 27th International Conference on Computational Linguistics. 2018.
4. He, Yu, et al. "Time-evolving Text Classification with Deep Neural Networks." IJCAI. 2018.

© POSTER TEMPLATE BY GENIGRAPHICS® 1.800.790.4001 WWW.GENIGRAPHICS.COM