Dynamic Word Embeddings for Evolving Semantic Discovery

Zijun Yao (Rutgers University), Yifan Sun (Technicolor), Weicong Ding (Amazon),
Nikhil Rao (Amazon), Hui Xiong (Rutgers University)

{zijun.yao, hxiong}@rutgers.edu, ysun13@cs.ubc.ca, {20008005dwc, nikhilrao86}@gmail.com

Language and Words Evolve over Time

- **CONTEXTS EVOLVE** president: Obama → Trump
- **MEANINGS EVOLVE** apple: fruit → technology
- **NEW WORDS ARISE** twitter, iphone, mp3

Goal

Learn time-aware vector representations (embeddings) of words to account for word evolution.

Challenges

- **Alignment**
  - The learned embeddings across time may not be placed in the same latent space, because the cost functions for training are invariant to rotations: $\widehat{U}(t)\widehat{U}(t)^T = U(t)RR^T U(t)^T = U(t)U(t)^T$
  - Alignment of embeddings across time is needed and challenging.
- **Sparsity**
  - Splitting the data across time → less training data per time slice.
  - Weakness of training separately across time → some words may have very few or no occurrences.
  - Making separate alignment (e.g., computing rotation matrix) problematic → "bad" time slices contaminate other time slices.

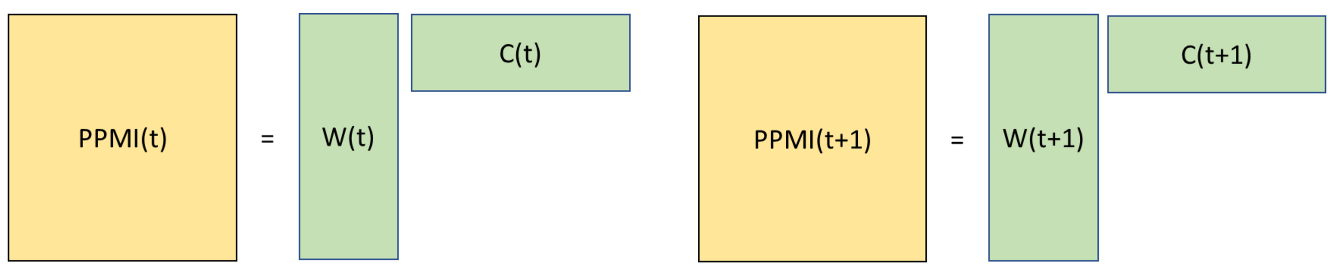
Main Novelty

- Learning the word embeddings across time jointly, thus obviating the need to solve a separate alignment problem.
- 1. This can be seen as an improvement over traditional, "single-time" methods such as word2vec.
- 2. Our experimental results suggest that enforcing alignment through regularization yields better results than two-step methods.
- 3. We share information across time slices: robust against data sparsity.

Dynamic Word Embedding

- **Our approach**

  Solving a composite problem with MF at each time point and a smoothing penalty across time.



- No need to find a rotation matrix.
- Smoothing aligns embeddings in successive time slices and makes embeddings more robust to missing data.
- **Model**
- Objective function.

PPMI factorization term for joint embedding over time

Smoothing term encourages the word embeddings to be aligned

$$\min_{U(1),\ldots,U(T)} \frac{1}{2}\sum_{t=1}^{T}\|Y(t)-U(t)U(t)^T\|_F^2 + \frac{\lambda}{2}\sum_{t=1}^{T}\|U(t)\|_F^2 + \frac{\tau}{2}\sum_{t=2}^{T}\|U(t-1)-U(t)\|_F^2.$$

- Positive Pointwise mutual information (PPMI).

$$\mathrm{PPMI}(t,L)_{w,c} = \max\{\mathrm{PMI}(\mathcal{D}_t,L)_{w,c},0\}. := Y(t)$$

$$\mathrm{PMI}(\mathcal{D},L)_{w,c} = \log\left(\frac{\#(w,c)\cdot|\mathcal{D}|}{\#(w)\cdot\#(c)}\right)$$

- Time-aware word embedding $U(t)$.
- Vocabulary size V, and total time slice T.

Reference

[1] Kulkarni *et al.* "Statistically significant detection of linguistic change." WWW 2015

[2] Hamilton *et al.* "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." ACL 2016

Optimization

- **A key challenge:** for large V and T , one might not be able to fit all the PPMI matrices in the memory.
- **A scalable solution**: decomposing the objective across time to solve for U(t).

Introduce W: quartic to quadratic

Enforces symmetry

$$\min_{U(t),W(t)} \frac{1}{2}\sum_{t=1}^{T}\|Y(t)-U(t)W(t)^T\|_F^2 + \frac{\gamma}{2}\sum_{t=1}^{T}\|U(t)-W(t)\|_F^2$$
$$+\frac{\lambda}{2}\sum_{t=1}^{T}\|U(t)\|_F^2 + \frac{\tau}{2}\sum_{t=2}^{T}\|U(t-1)-U(t)\|_F^2$$
$$+\frac{\lambda}{2}\sum_{t=1}^{T}\|W(t)\|_F^2 + \frac{\tau}{2}\sum_{t=2}^{T}\|W(t-1)-W(t)\|_F^2,$$

Solve using Block Coordinate Descent

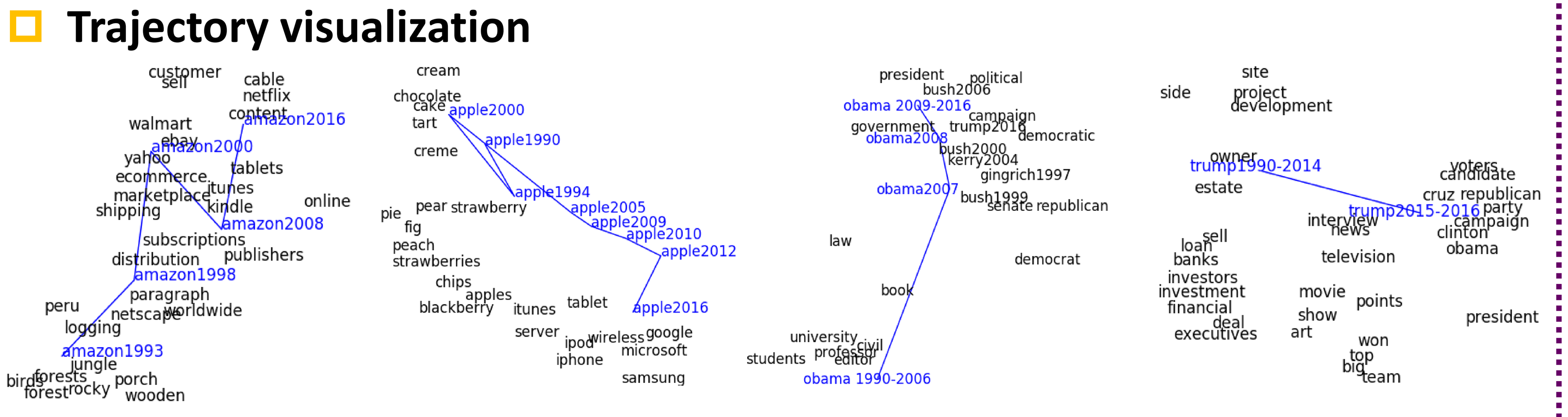
Enforces alignment

$$A = W(t)^T W(t) + (\gamma+\lambda+2\tau)I,$$
$$B = Y(t)W(t) + \gamma W(t) + \tau(U(t-1)+U(t+1))$$
$$U(t)A = B$$

For each single time *t* solve U(t)A = B (similarly for W)

Experimental Study

- **The New York Times:** 99872 articles from 1990 to 2016. T=27 time slices (one for each year). 59 news section (e.g., Business, Technology, Sports). V= 20936 unique words after removing stop words and rare (<200) words.
- **Trajectory visualization**



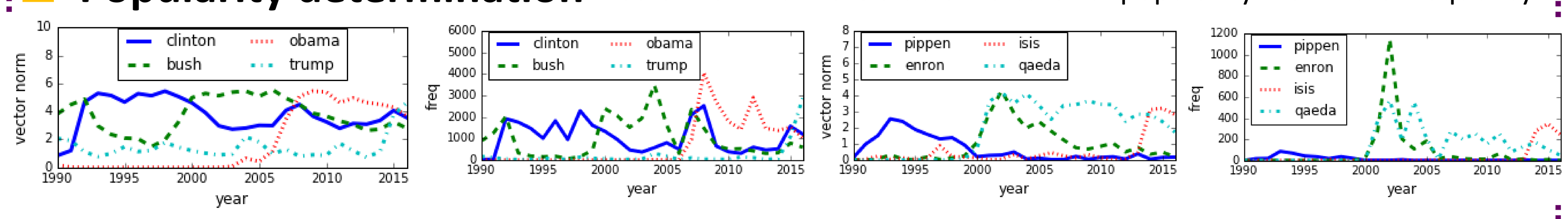
- **Equivalence searching** Closest word to query (word, year) in different years

Query	iphone, 2012	twitter, 2012	mp3, 2000
90-94	desktop, pc, dos, macintosh, software	broadcast, cnn, bulletin, tv, radio, messages, correspondents	stereo, disk, disks, audio
95-96			mp3
97		chat, messages, emails, web	
98-02	pc		
03			napster
04			mp3
05-06	ipod	blog, posted	itunes, downloaded
07-08	iphone		
09-12		twitter	
13-16	smartphone, iphone		

Question	US president	NYC mayor
Query	obama, 2016	blasio, 2015
90-92	bush	dinkins
93	clinton	
94-00		giuliani
01	bush	
02-05		bloomberg
06		n/a*
07		bloomberg
08		
09-10	obama	
11		cuomo*
12		bloomberg
13-16		blasio

year	1990	1991	1992	1993
word	**edberg**	lendl	sampras	**sampras**
1994	1995	1996	1997	1998
sampras	**sampras**	ivanisevic	**sampras**	**sampras**
1999	2000	2001	2002	2003
sampras	sampras	agassi	capriati	**roddick**
2004	2005	2006	2007	2008
federer	**federer**	roddick	**federer**	**nadal**
2009	2010	2011	2012	2013
federer	**nadal**	**djokovic**	federer	federer
2014	2015			
federer	**djokovic**			

- **Popularity determination** Vector norm: a more stable indicator of popularity than word frequency



- **Semantic similarity** Clustering analysis using section labels as ground-truth

Table 4: Normailized Mutual Information (NMI).

Method	10 Clusters	15 Clusters	20 Clusters
SW2V	0.6736	0.6867	0.6713
TW2V	0.5175	0.5221	0.5130
AW2V	0.6580	0.6618	0.6386
DW2V	**0.7175**	**0.7162**	**0.6906**

Table 5: F-measure (F_β).

Method	10 Clusters	15 Clusters	20 Clusters
SW2V	0.6163	0.7147	0.7214
TW2V	0.4584	0.5072	0.5373
AW2V	0.6530	0.7115	0.7187
DW2V	**0.6949**	**0.7515**	**0.7585**

- **Alignment quality** Query of equivalence words across time

Table 6: Mean Reciprocal Rank (MRR) and Mean Precision (MP) for Testset 1.

Method	MRR	MP@1	MP@3	MP@5	MP@10
SW2V	0.3560	0.2664	0.4210	0.4774	0.5612
TW2V	0.0920	0.0500	0.1168	0.1482	0.1910
AW2V	0.1582	0.1066	0.1814	0.2241	0.2953
DW2V	**0.4222**	**0.3306**	**0.4854**	**0.5488**	**0.6191**

Table 7: Mean Reciprocal Rank (MRR) and Mean Precision (MP) for Testset 2.

Method	MRR	MP@1	MP@3	MP@5	MP@10
SW2V	0.0472	0.0000	0.0787	0.0787	0.2022
TW2V	0.0664	0.0404	0.0764	0.0989	0.1438
AW2V	0.0500	0.0225	0.0517	0.0787	0.1416
DW2V	**0.1444**	**0.0764**	**0.1596**	**0.2202**	**0.3820**

- **Robustness**

Table 8: MRR and MP for alignment with every 3 years sub-sampling.

Method	r	MRR	MP@1	MP@3	MP@5	MP@10
AW2V	100%	0.1582	0.1066	0.1814	0.2241	0.2953
AW2V	10%	0.0884	0.0567	0.1020	0.1287	0.1727
AW2V	1%	0.0409	0.0255	0.0475	0.0605	0.0818
AW2V	0.1%	0.0362	0.0239	0.0416	0.0532	0.0690
DW2V	100%	0.4222	0.3306	0.4854	0.5488	0.6191
DW2V	10%	0.4394	0.3489	0.5036	0.5628	0.6292
DW2V	1%	0.4418	0.3522	0.5024	0.5636	0.6310
DW2V	0.1%	0.4427	0.3550	0.5006	0.5612	0.6299

Robustness against word removal with intensity r

Baselines:
SW2V = Static word2vec
TW2V = Transformed-Word2Vec [1]
AW2V = Aligned-Word2Vec [2]
DW2V = Proposed method

Conclusion

We proposed a model to learn time-aware word embeddings. Our proposed method simultaneously learns the embeddings and aligns them across time. We provided dynamic embeddings trained on the New York Times dataset, from which we discover evolving word semantics.