Word Embedding and the Beyond

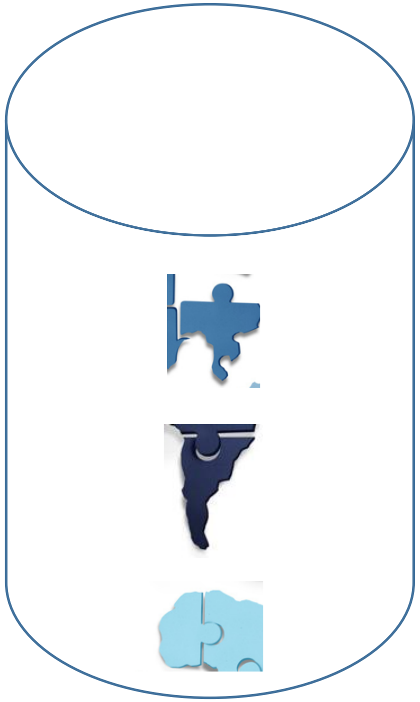
Benyou Wang

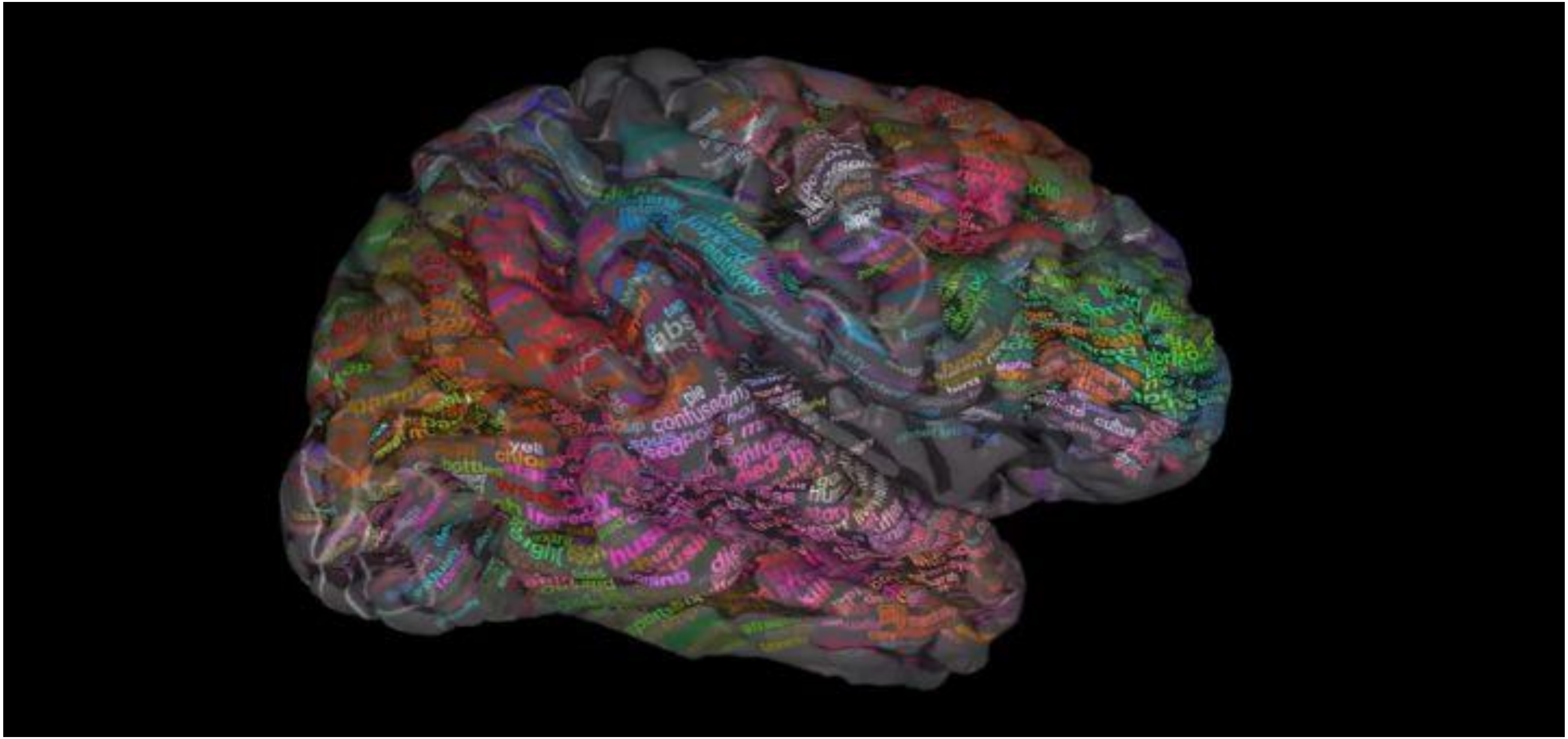
University of Padova

Contents

- **What** embedding is and **why**
- **Trends** of word embedding
- Word embedding in **dynamics**
  - Examples
  - Our ideas

What does “embed” means?





Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458. doi:10.1038/nature17637

http://gallantlab.org/huth2016/

Trends for Neural IR

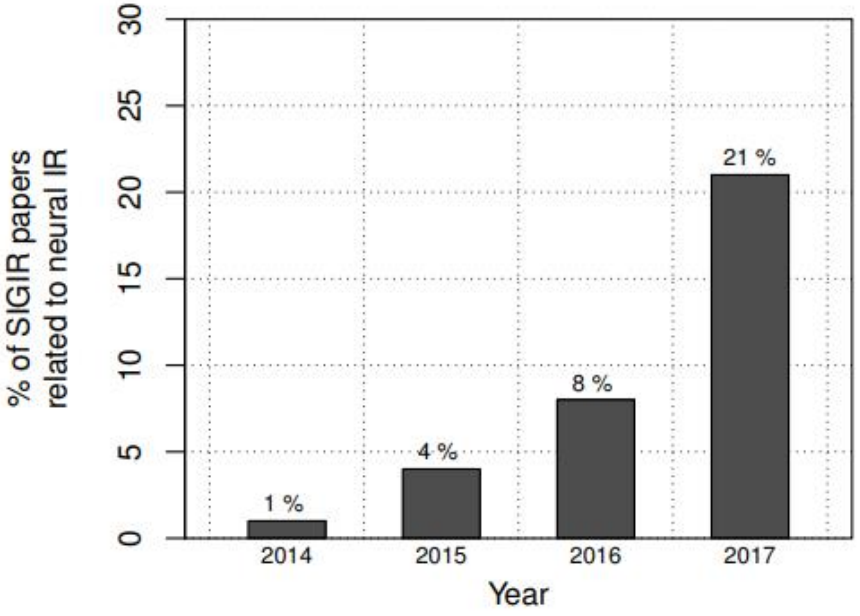


Figure 1: The percentage of neural IR papers at the ACM SIGIR conference—as determined by a manual inspection of the paper titles—shows a clear trend in the growing popularity of the field.

Mitra B, Craswell N. Neural Models for Information Retrieval[J]. arXiv preprint arXiv:1705.01509, 2017.

Example of **Mismatch**

Query	Document	Term Matching	Semantic Matching
seattle best hotel	seattle best hotels	no	yes
pool schedule	swimmingpool schedule	no	yes
natural logarithm transformation	logarithm transformation	partial	yes
china kong	china hong kong	partial	no
why are windows so expensive	why are macs so expensive	partial	no

Hang li, http://www.hangli-hl.com/uploads/3/4/4/6/34465961/tsinghua_opportunities_and_challenges_in_deep_learning_for_information_retrieval.pdf

Localist representation

Size color … unknown

- BMW [1, 0, 0, 0, 0] [.3, .7, .2, .1, .5]
- Audi [0, 0, 0, 1, 0] [.5, .3, .2, .1, .0]
- Benz [0, 0, 1, 0, 0] [.2, .0, .31, .03, .01]
- Polo [0, 0, 0, 1, 0] [.1, .1, .5, .5, 0.2]

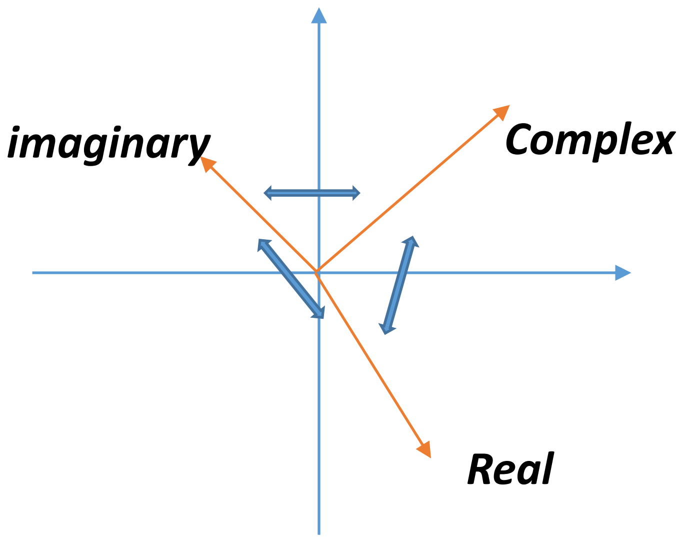
http://www.cs.toronto.edu/~bonner/courses/2014s/csc321/lectures/lec5.pdf

Distributed representation

			Size color … unknown
• BMW	[1, 0, 0, 0, 0]		[.3, .7, .2, .1, .5]
• Audi	[0, 0, 0, 1, 0]		[.5, .3, .2, .1, .0]
• Benz	[0, 0, 1, 0, 0]		[.2, .0, .31, .03, .01]
• Polo	[0, 0, 0, 1, 0]		[.1, .1, .5, .5, 0.2]

Embedding

Distributional hypothesis   *linguistic items with similar distributions have similar meanings*



*Life is **complex**. It has both **real** and **imaginary** parts*

https://en.wikipedia.org/wiki/Distributional_semantics

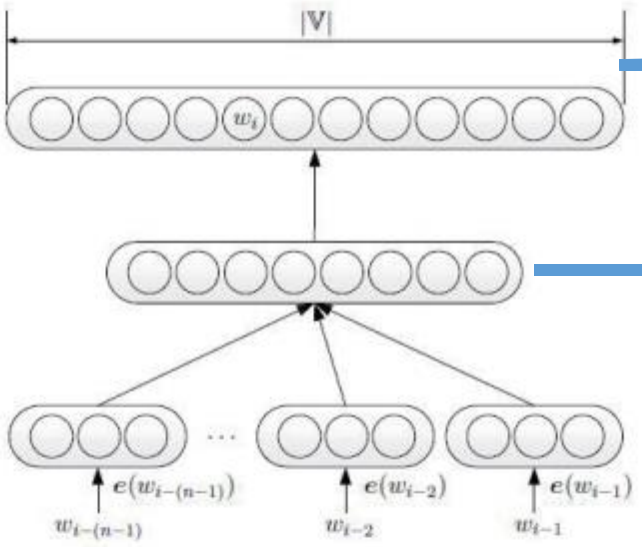
How to get Distributed representation

- Matrix Factorization
  - Word-word Matrix
  - Document-word Matrix
    - PLSA
    - LDA
- Sample-based Prediction
  - NNLM
  - C & W
  - Word2vec

Glove is a combination between these two schools of approaches

Levy, Omer, and Yoav Goldberg. "Neural word embedding as implicit matrix factorization." *Advances in neural information processing systems*. 2014.

NNLM to Word2vec



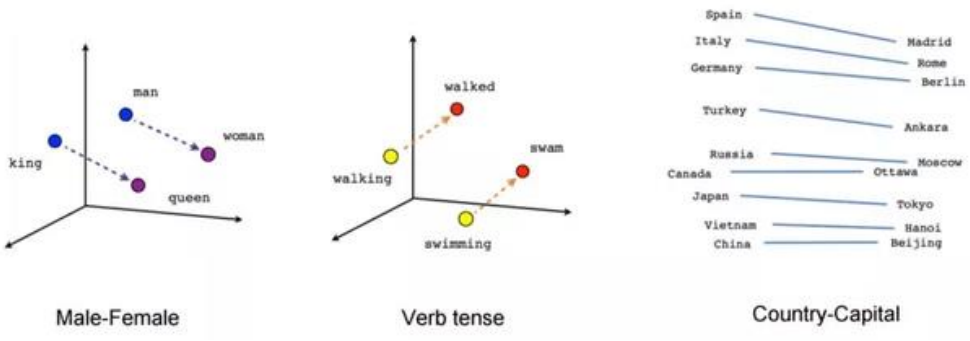
Negative sampling
Hierarchical softmax

Concat -> mean

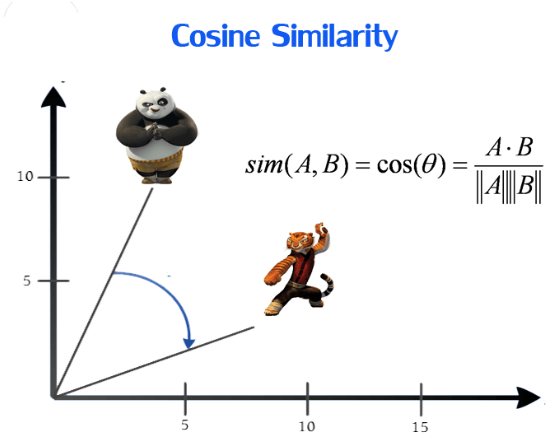
Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(Feb): 1137-1155.
Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

Advantage of word embedding

- Linguistic regulation
  - $\overrightarrow{king} - \overrightarrow{man} = \overrightarrow{queen} - \overrightarrow{woman}$



- Semantic matching
  - As the initial input Feature/**Weight** for NN



Side effect – Additivity compositionality

Examples:

- $\overrightarrow{king} - \overrightarrow{man} = \overrightarrow{queen} - \overrightarrow{woman}$
- $\overrightarrow{Italy} - \overrightarrow{Rome} = \overrightarrow{China} - \overrightarrow{Beijing}$
- $\overrightarrow{go} - \overrightarrow{went} = \overrightarrow{take} - \overrightarrow{took}$

Sexism implicit/stereotypes

$$\overrightarrow{man} - \overrightarrow{woman} = \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

What should this be in the case of ***complex-valued word*** *embedding?*

Gittens A, Achlioptas D, Mahoney M W. Skip-gram-zipf+ uniform= vector additivity[C]// **ACL**. 2017, 1: 69-76.

Problems

- Noise
    - Trained with neural network
- OOV
    - Subword information
- Multi-sense/Polysemy
- Semantic composition
- Beyond real-valued vector
    - Complex-valued
    - Gaussian Embedding
- Bias
- Corpus-sensitive
    - Elmo/Bert
- **Non-static**

http://ruder.io/word-embeddings-2017/

ACL 2016

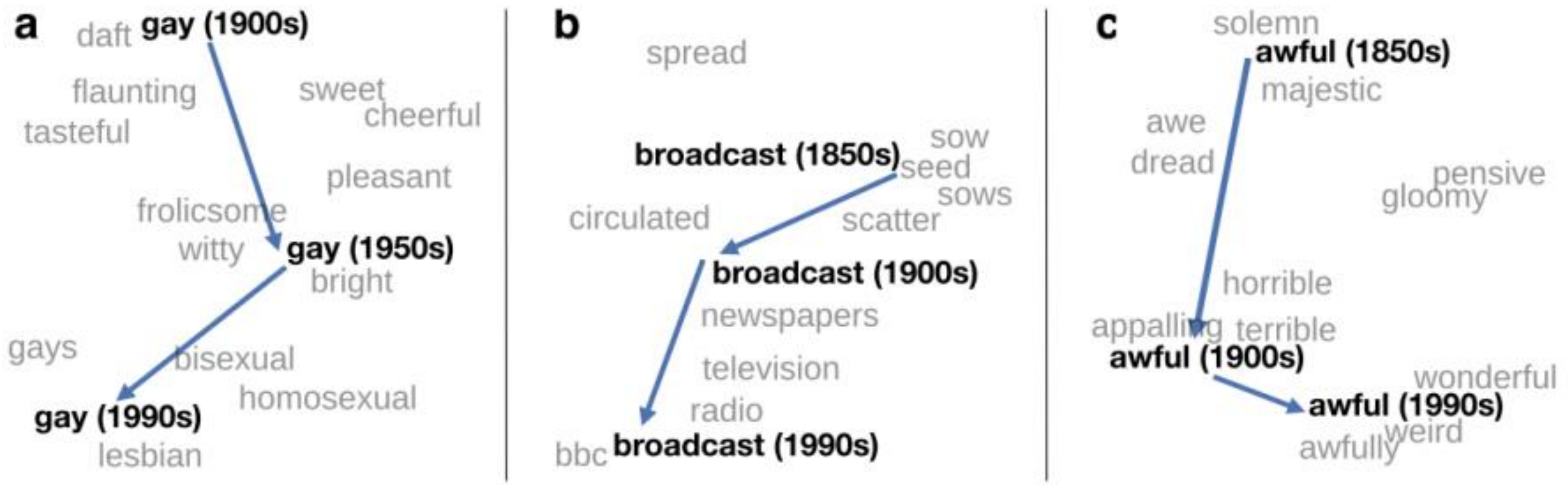


Figure 1: Two-dimensional visualization of semantic change in English using SGNS vectors.[2] **a**, The word *gay* shifted from meaning "cheerful" or "frolicsome" to referring to homosexuality. **b**, In the early 20th century *broadcast* referred to "casting out seeds"; with the rise of television and radio its meaning shifted to "transmitting signals". **c**, *Awful* underwent a process of pejoration, as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Simpson et al., 1989).

Hamilton W L, Leskovec J, Jurafsky D. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change[C]// **ACL**. 2016, 1: 1489-1501.

https://github.com/williamleif/histwords

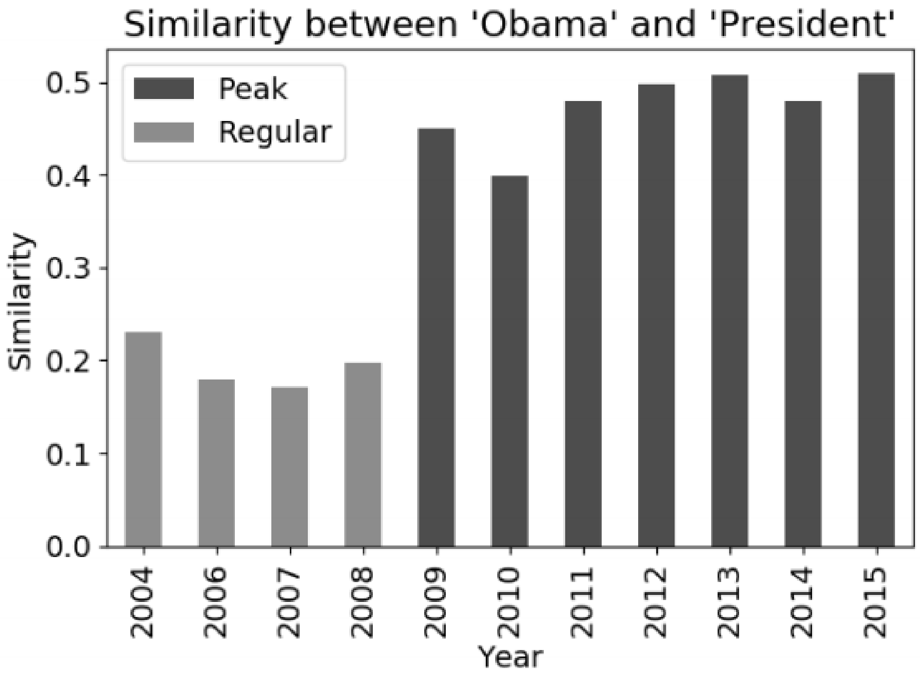
ACL 2017

1987	reagan	koch	soviet	iran_contra	navratilova	yuppie	walkman
1988	reagan	koch	soviet	iran_contra	sabatini	yuppie	tape_deck
1989	bush	koch	soviet	iran_contra	navratilova	yuppie	walkman
1990	bush	dinkins	soviet	iran_contra	navratilova	yuppie	headphones
1991	bush	dinkins	soviet	iran_contra	navratilova	yuppie	cassette_player
1992	bush	dinkins	russian	iran_contra	sabatini	yuppie	walkman
1993	clinton	dinkins	russian	iran_contra	navratilova	yuppie	cd_player
1994	clinton	mr_giuliani	russian	iran_contra	sanchez_vicario	yuppie	walkman
1995	clinton	giuliani	russian	white_house	graf	yuppie	cassette_player
1996	clinton	giuliani	russian	whitewater	graf	yuppie	walkman
1997	clinton	giuliani	russian	iran_contra	hingis	yuppie	headphones
1998	clinton	giuliani	russian	lewinsky	hingis	yuppie	headphones
1999	clinton	mayor_giuliani	russian	white_house	hingis	yuppie	buttons
2000	clinton	giuliani	russian	white_house	hingis	yuppie	headset
2001	bush	giuliani	russian	iran_contra	capriati	yuppie	headset
2002	bush	bloomberg	russian	white_house	hingis	gen_x	mp3_player
2003	bush	bloomberg	russian	white_house	agassi	hipsters	walkman
2004	bush	bloomberg	north_korean	iran_contra	federer	gen_x	headphones
2005	bush	bloomberg	north_korean	white_house	roddick	geek	ear_buds
2006	bush	bloomberg	iranian	white_house	hingis	teen	headset
2007	bush	bloomberg	iranian	capitol_hill	federer	dads	ipod

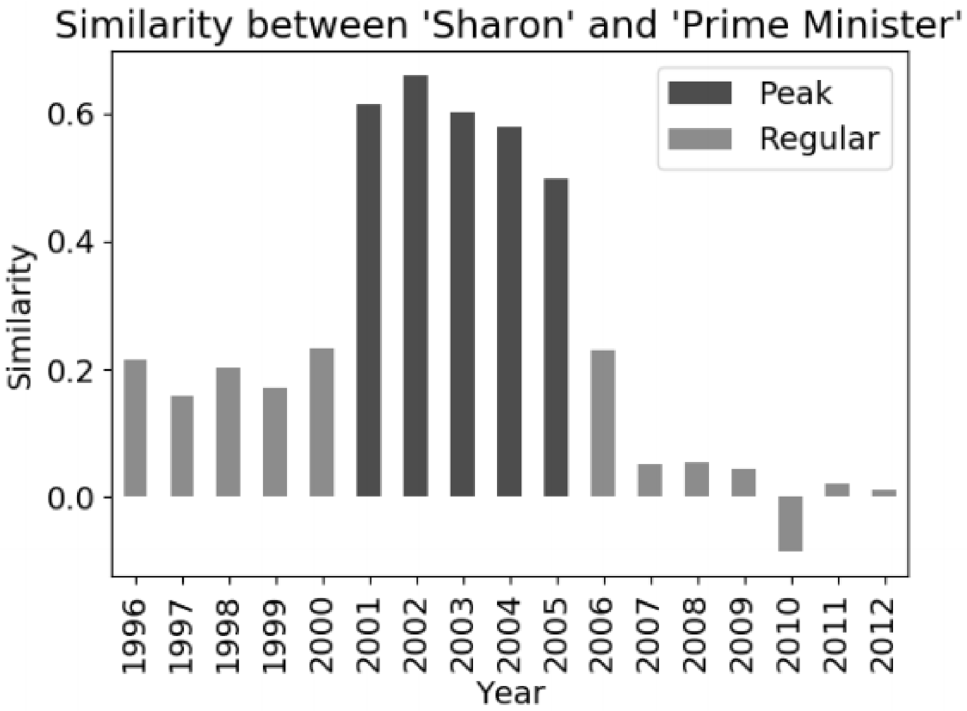
Table 1: Examples of words from 1987 and their analogues over time. Each column corresponds to a single point in vector space, and each row shows the word closest to that point in a given year.

Szymanski, T. (2017). Temporal Word Analogies : Identifying Lexical Replacement with Diachronic Word Embeddings. ACL (pp. 448–453).

EMNLP 2017 Word Relatedness over Time



(a)



(b)

Similarity identified by the algorithms between words over time. Dark gray indicates high similarity whereas light gray indicates nonsignificant similarity

Rosin, G., Radinsky, K., & Adar, E. (2017). Learning Word Relatedness over Time. EMNLP.

EMNLP 2017-Laws of semantic change

- The Law of Conformity
  - which frequency is negatively correlated with semantic change
- The Law of Innovation
  - which polysemy is positively correlated with semantic change
- The Law of Prototypicality
  - which prototypicality is negatively correlated with semantic change

Dubossarsky, H., Grossman, E., & Weinshall, D. (2017). Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. EMNLP(pp. 1147–1156).

ICML 2017 –dynamic Word embedding

Dynamic Word Embeddings

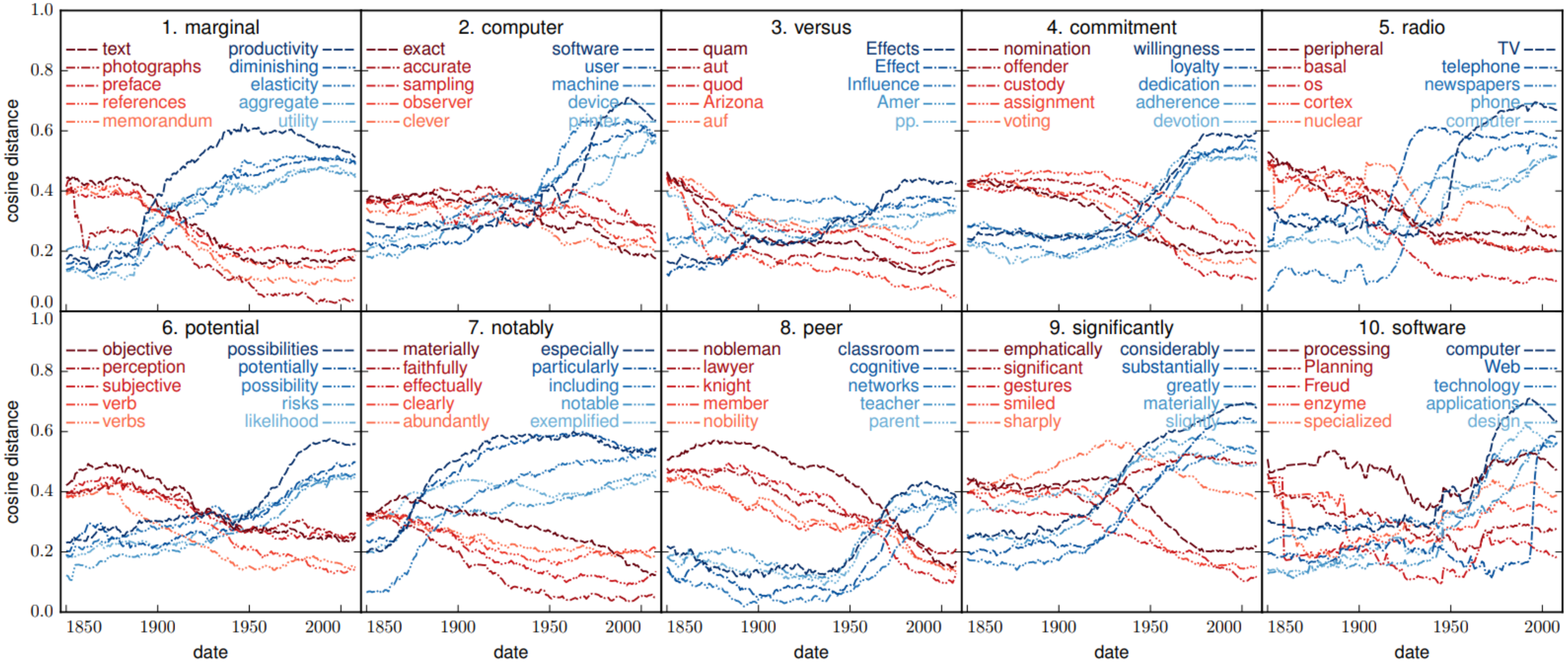
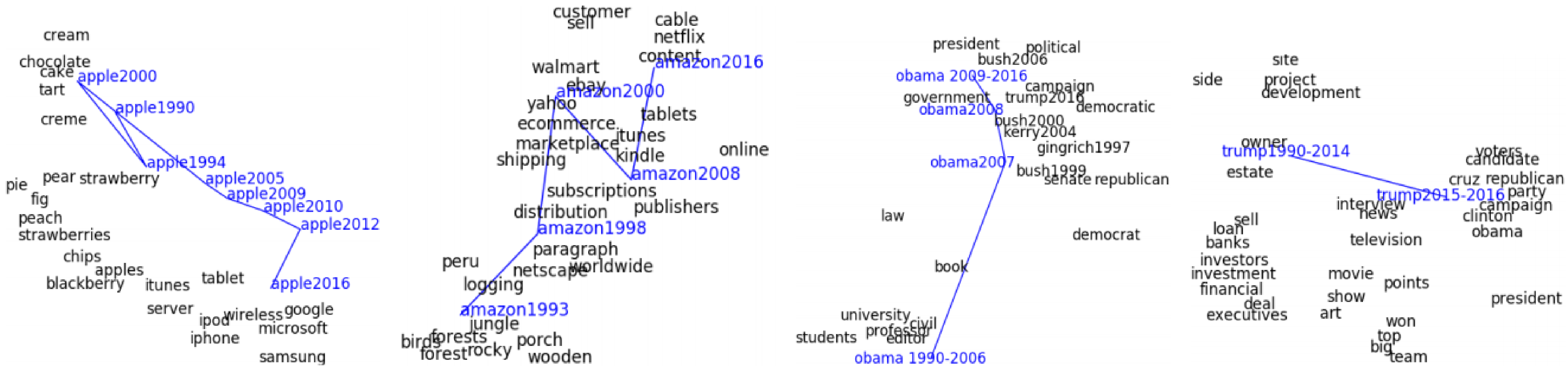


Figure 1. Evolution of the 10 words that changed the most in cosine distance from 1850 to 2008 on Google books, using skip-gram filtering (proposed). Red (blue) curves correspond to the five closest words at the beginning (end) of the time span, respectively.

Bamler R, Mandt S. Dynamic Word Embeddings[C]// ICML. 2017: 380-389.

WSDM 2018 - evolving semantic discovery

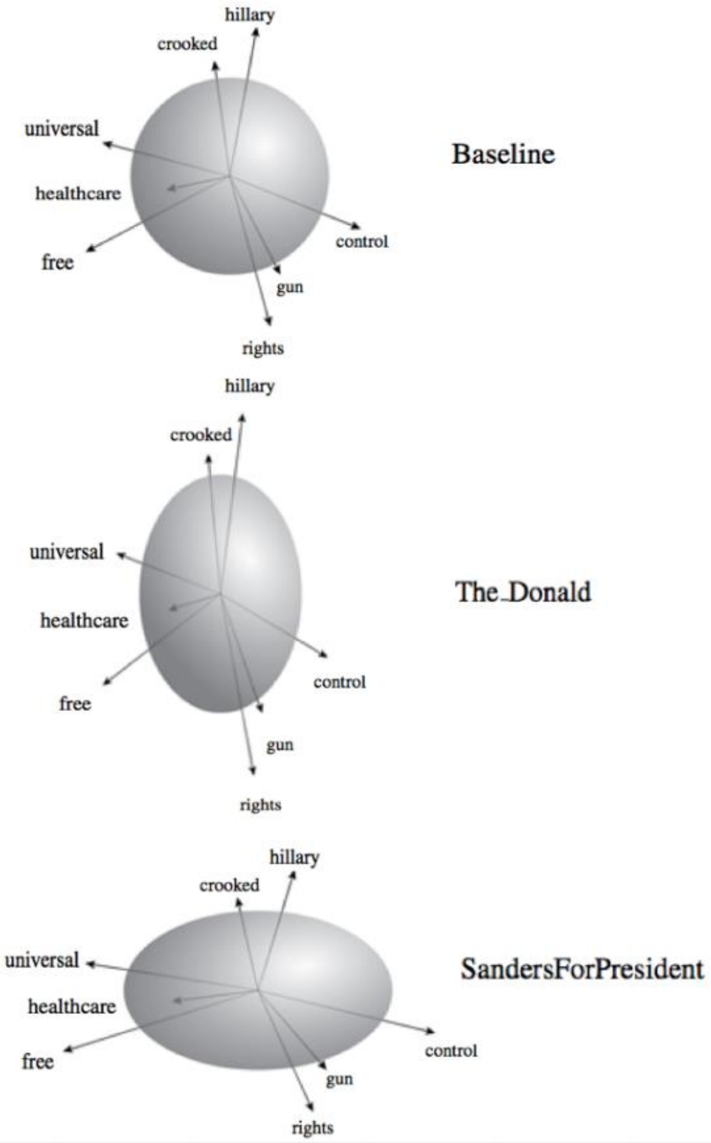


(a) apple (b) amazon (c) obama (d) trump

Figure 1: Trajectories of brand names and people through time: `apple`, `amazon`, `obama`, and `trump`.

Yao Z, Sun Y, Ding W, et al. Dynamic word embeddings for evolving semantic discovery[C]// WSDM. ACM, 2018: 673-681.

ICML



Tian K, Zhang T, Zou J. CoVeR: Learning Covariate-Specific Vector Representations with Tensor Decompositions[J]. ICML 2018, 2018.

Lack of reasonable benchmark for evaluation

- Reason:
  - Hard to find annotators **from 10 years ago**
  - The ground truth is a little bit subjective

The changing stereotype over time

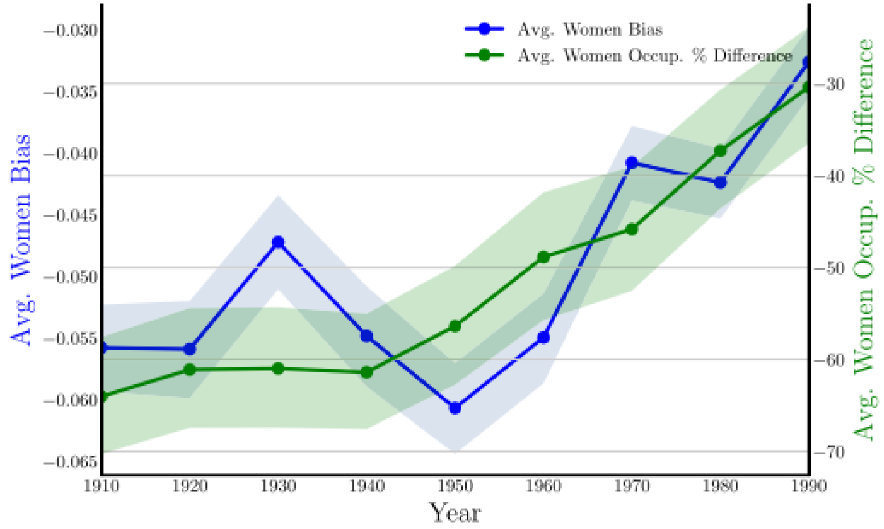


Fig. 2. Average gender bias score over time in COHA embeddings in occupations vs. the average percentage of difference. More positive means a stronger association with women. In blue is relative bias toward women in the embeddings, and in green is the average percentage of difference of women in the same occupations. Each shaded region is the bootstrap SE interval.

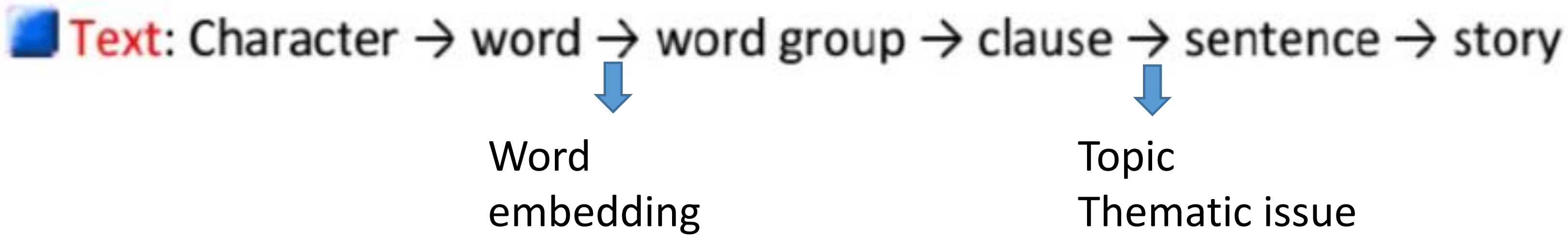
Bolukbasi T, Chang K W, Zou J Y, et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings[C]//**NIPS**. 2016: 4349-4357.
Garg N, Schiebinger L, Jurafsky D, et al. Word embeddings quantify 100 years of gender and ethnic stereotypes[J]. Proceedings of the National Academy of Sciences, 2018, 115(16): E3635-E3644.

Linking embedding with topic/thematic issue

- For a topic, it is usually considered as a distribution of words
  - $p^{(i)} = p(p_{w_1}, p_{w_1}, \dots p_{w_{|v|}})$

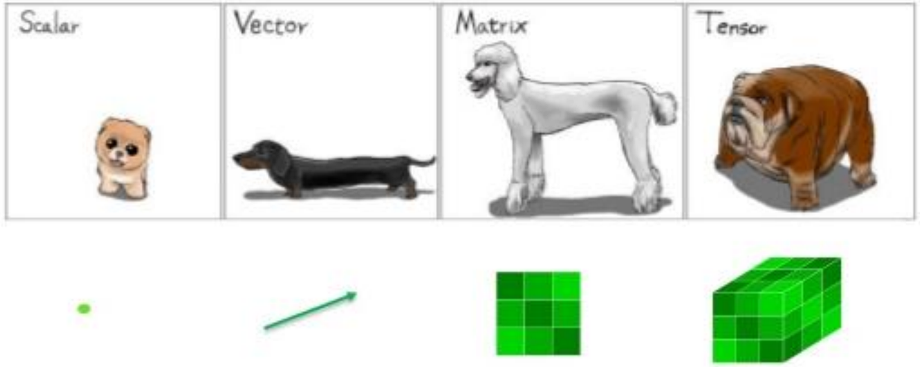
- For a word embedding, its neighbor has a well-designed distance, we could also get a distribution as $p_{w_j} = \frac{e^{d_{ij}}}{\sum e^{d_{ij}}}$.

- In a sense, word embedding is considered **lower-level topic**



Dynamics

- Concatenate the Document-Term or Term-Term Co-occurrence as a Tensor
  - $[M_{t_1}, M_{t_2}, \ldots, M_{t_T}]$ as $T_{t,d,w}$, 3-d Tensor, where M_{t_1} is the D-W matrix.



  - Tensor composition/factorization machine for **time-aware word embedding**
    - *Obtain the neighbor words of "nuclear" in different time stamp.*