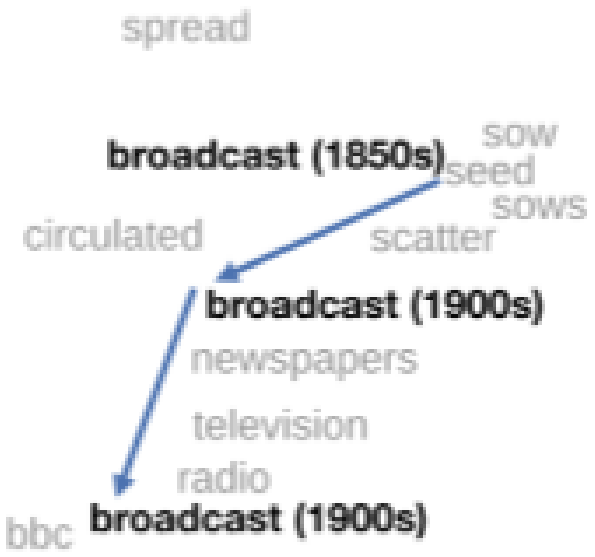
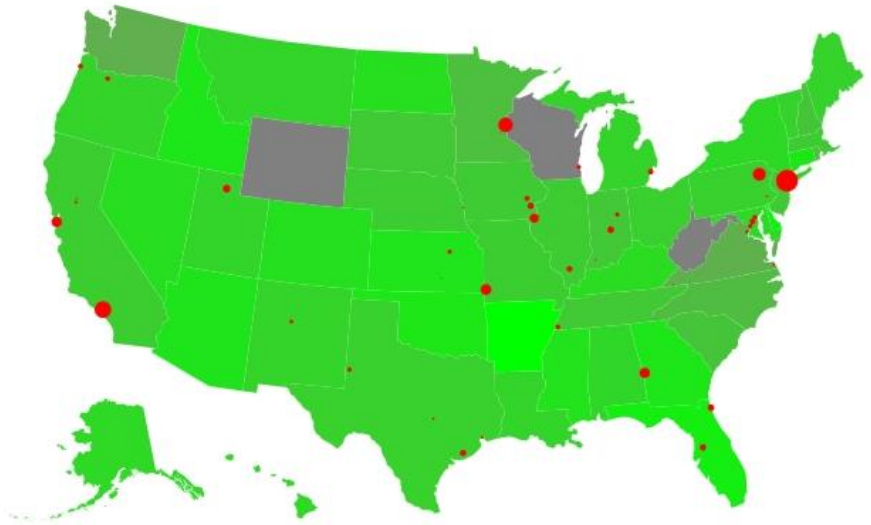
Exploring Semantic Change of Chinese Word Using Crawled Web Data

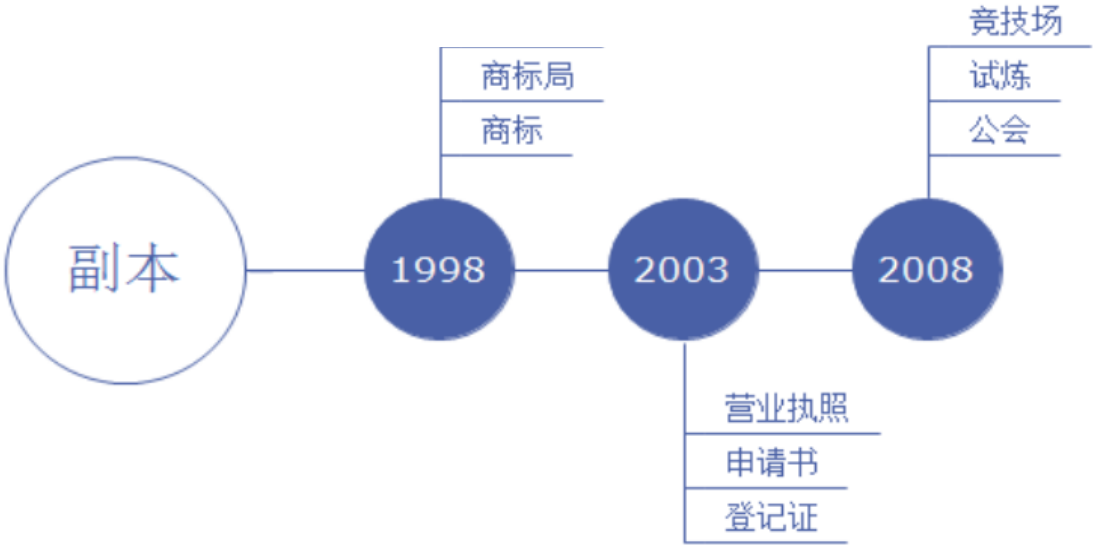
Xiaofei Xu, Yukun Cao and Li Li

Southwest University, China

Diachronic Analysis
+
Web Data


spread
broadcast (1850s)
sow
seed
sows
circulated
scatter
broadcast (1900s)
newspapers
television
radio
bbc
broadcast (1900s)




副本
1998
商标局
商标
2003
营业执照
申请书
登记证
2008
竞技场
试炼
公会

Methods

- PPMI
- PPMI-based SVD
- Word2vec

Methods

- In the PPMI, words are represented by constructing a high dimensional sparse matrix

$$M \in R^{|V_w| \times |V_c|}$$

- where each row denotes a word w, and each column represents a context c.
- The value of the matrix cell M_{ij} is obtained by

$$M_{ij} = max\left\{log\left(\frac{p(w_i, c_j)}{p(w_i)p(c_j)}\right), 0\right\}$$

Methods

- PPMI-based SVD is a method that applied SVD to the PPMI matrix. In SVD, we have $S = U \cdot V \cdot T$, we can apply this to a PPMI matrix and the word embedding can be approximated by U or $U \cdot V$.
- Word2vec is a typical neural network based method with three layers. Here we use the Skip-gram model with negative sampling.

Vector Alignment

- Word embedding in every time slice is trained separately, to compare the word vectors, we need to align vectors.
- Here we use orthogonal Procrustes to align the learned word representations.

Similarity measure

- We use cosine similarity between words to measure the similarity

Data

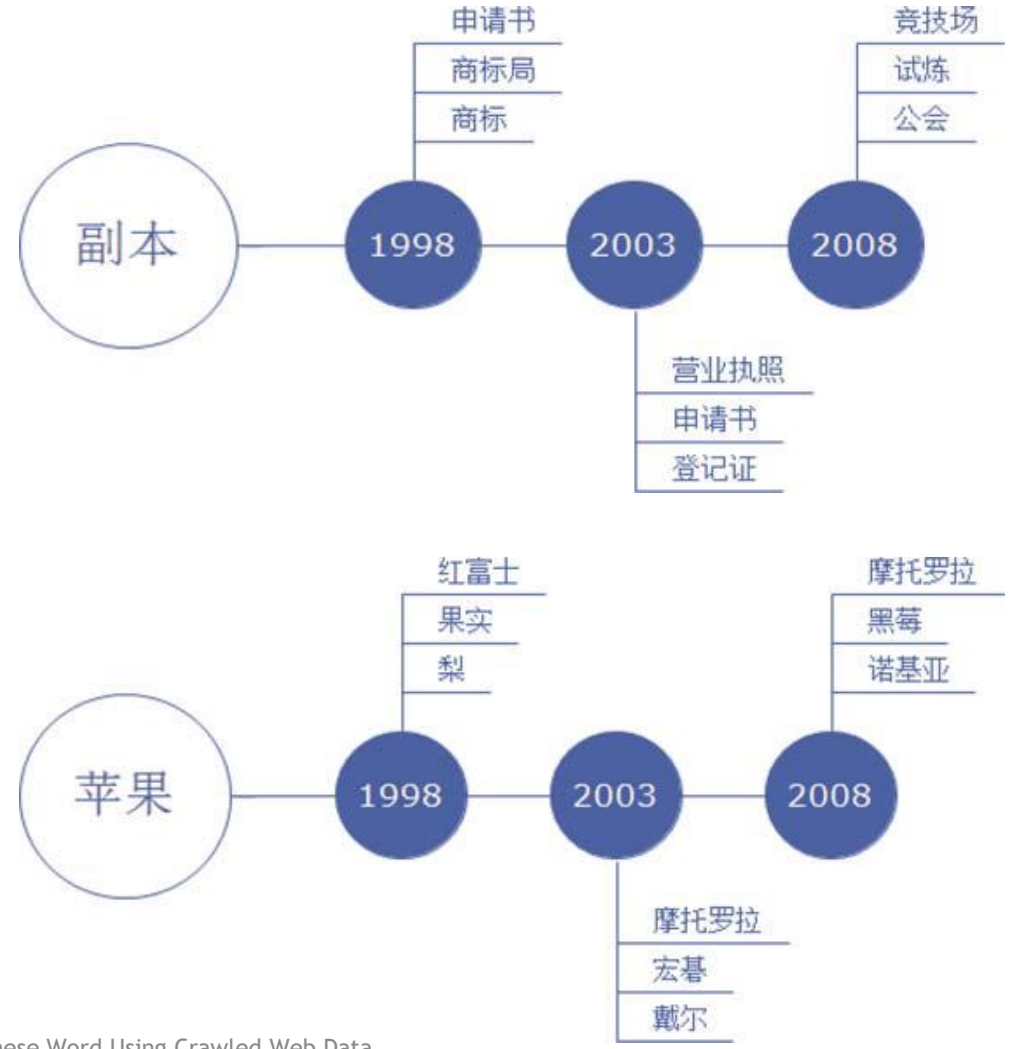
- Search engine crawled web pages provided by Sogou Lab
- Organized as XML
- Preprocess:
  - Extract the time information from document data
  - Clean corpus
  - Word segmentation
- Finally, 52,324,791 lines of data, 467,826,233 words and 225,182 unique words.

Evaluation

Table 1. Number of shifted words

Threshold	0.1	0.2	0.3	0.4	0.5
PPMI	12778	19157	19963	20146	20224
SVD	552	2371	6177	11048	15544
SGNS	9373	9382	9400	9604	11987

Evaluation


副本
1998
申请书
商标局
商标
2003
营业执照
申请书
登记证
2008
竞技场
试炼
公会
苹果
1998
红富士
果实
梨
2003
摩托罗拉
宏碁
戴尔
2008
摩托罗拉
黑莓
诺基亚

Discussions

- How to judge a historical word embedding method is effective quantitatively?

Discussions

- How to find a relationship between diachronic shift words, linguistic evolution and social changes?

Discussions

- Can we align the vectors during training?
- Can we align the vectors jointly?

Discussions

- Are recent pre-trained deep contextualized word representations applicable to this task?

Recent work

- Defines a evaluation method that not rely on human labeled “ground-truth” data.
- Defines a improved word embedding method based on word2vec to support naturally aligned embedding generation.

Recent work

- Apply word change point detection to recommender system
- We suppose that the meaning change of word need many people talk about this word differently. And we think this is the “Hot” word, and we should recommend these relevant items to new customers.

Recent work

- Apply word change point detection to ancient Chinese
  - More useful
  - More challenge

Thank you!

Website: swu-nlp.group
Contact: nakamura@email.swu.edu.cn