言葉の意味変化をとらえる

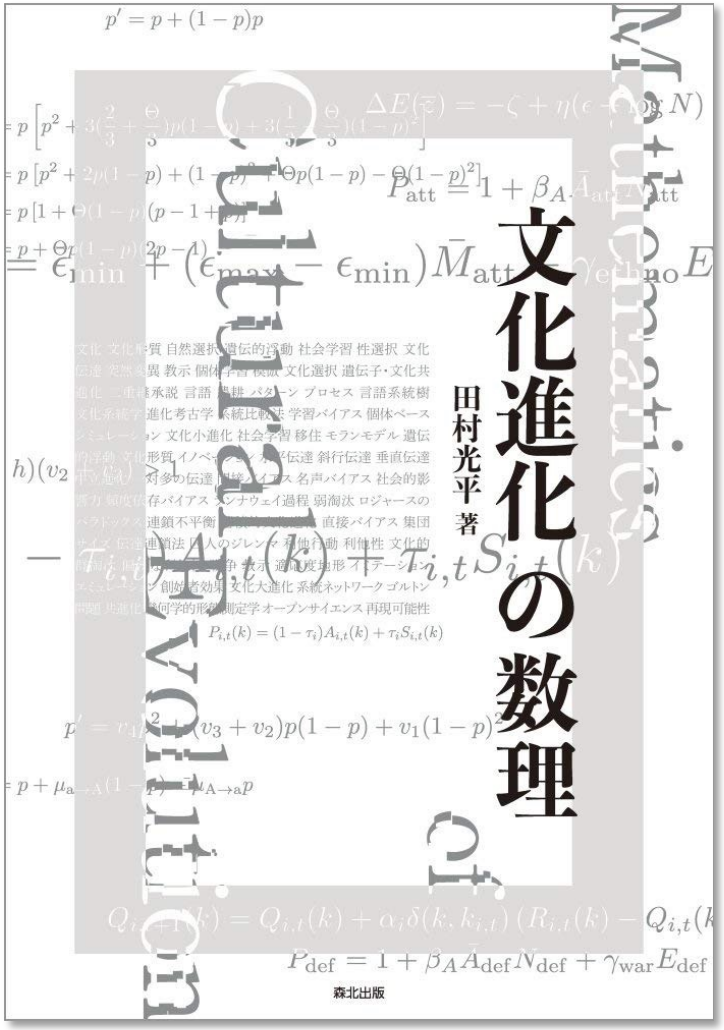
持橋大地
統計数理研究所/国立国語研究所
daichi@ism.ac.jp

日本統計学会春季集会「文化とデータサイエンス」
2026-3-7(土)

文化進化の数理 (1)

- https://sites.google.com/view/icetseminar/home
- オンラインで、どなたでも参加可能です

文化進化の数理 (2)



- 田村光平(東北大学)著、森北出版、2020年
- https://www.amazon.co.jp/dp/4627062710

言語の時間的変化

"*Language is a dynamic system, constantly evolving and adapting to the needs of its users and their environment*" (Aitchison, 2001)

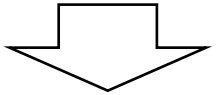
- 音韻変化: 単語の発音の変化
- 文法的変化: 使われる統語構造の変化
- 意味的変化: 単語の意味の時間的変化
  - "cute": 賢い (18世紀) → 悪賢い (19世紀) → かわいい (現在)
  - 意味の変化した単語は多数あり、その変化は未知
  - どうやって自動的にこうした変化を推定すればよい？

今日の話題

- 単語の意味 (離散) の数とその変化を追跡する統計モデル
  - 潜在トピックモデルの拡張
- ある単語に何個の意味があるのか、その割合がどう変化したのかを追跡できる
  (Inoue+, EMNLP 2022 Long)
- 他にも、単語の意味変化についてさまざまな研究をしています (本日は時間の関係で割愛)

単語の意味とは

- 単語の「意味」とは何か？
  → 分布仮説 (Harris 1954)に従うとする
  - 単語の意味は、前後の文脈に現れる単語の分布に反映される



- 通時コーパスから、目的の単語を含むスニペットをまず抽出する
- 通時コーパス: 英語では COHA (Corpus of Historical American English)、日本語では国語研のCHJ (日本語歴史コーパス) などの言語リソースがある

スニペットの抽出

目標単語: *“coach”*

通時コーパス

タイムスタンプ＋前後のスニペット

1853	The driver made room for the trunk on the top of the **coach**.
1900	The chair passed the **coach**, the horses proceeding at a walk.
1949	Tell him if I start **coaching**, it′ ll be as a head **coach** at a top school.
1995	available for inspection, like the **Coach** bags offered on Canal Street
2003	Football **coach** and other top school officials have been interviewed.

単語の意味変化の先行研究

(Hamilton+ 2016)

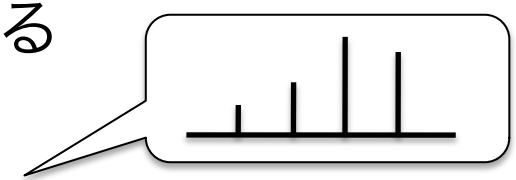
- 単語の前後の文脈(スニペット)から、単語の**埋め込みベクトル**を計算 (word2vecなど)
  →時間的な変動を観察する
- 変化したことはわかるが、何がどう変化したかはわからない！

得られた“coach”のスニペット

トピック z

1853	[driver, make, room, trunk]	→馬車
1900	[chair, pass, horse, proceed, walk]	→馬車
1949	[tell, start, coach, head, top, school]	→スポーツ
1995	[available, inspection, bags, offer]	→ブランド
2003	[football, top, school, official, interview]	→スポーツ

- 各スニペットは、いずれかの意味(トピック) z に対応していると考えられる
- 生成モデル：
  - ある確率分布 $\phi(z)$ に従って、z を選択
  - トピック z から、トピック毎の単語分布 $\psi(w|z)$ に従って実際のスニペットが生成される
- もちろん、トピックは未知



トピックの教師なし学習

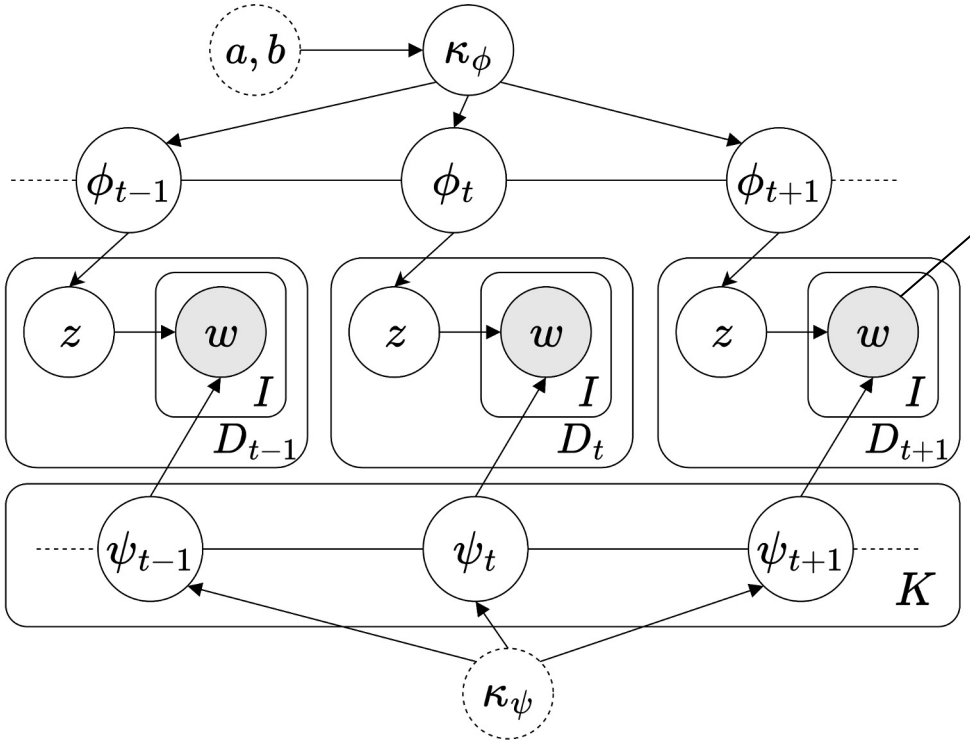
- アルゴリズム：
  - 各スニペットを、確率的に*K*個のトピックzのどれかに割り当てる ($\phi(z)$がわかる)
  - 各トピック　に割り当てられたスニペットの文章全体から、$\psi(w|z)$を更新
  - 以上を繰り返す
- スニペットのクラスタリングに相当する

ただし...

- 単語の意味は変化する
  → 時刻 *t* での意味の割合 $\phi_t(z)$ は、少しずつ変化する
  → 時刻 *t* でのトピック-単語確率 $\psi_t(w|z)$ も変化する
- 観測値がない年もあるので、うまく平滑化する必要
- 意味の総数*K*は、事前にはわからない
  → 本研究で、ディリクレ過程によって自動推定

先行研究: SCAN (Frermann+ TACL 2016)

- SCAN: Dynamic Bayesian Model of Sense ChANge
  - スニペット時系列の動的トピックモデル

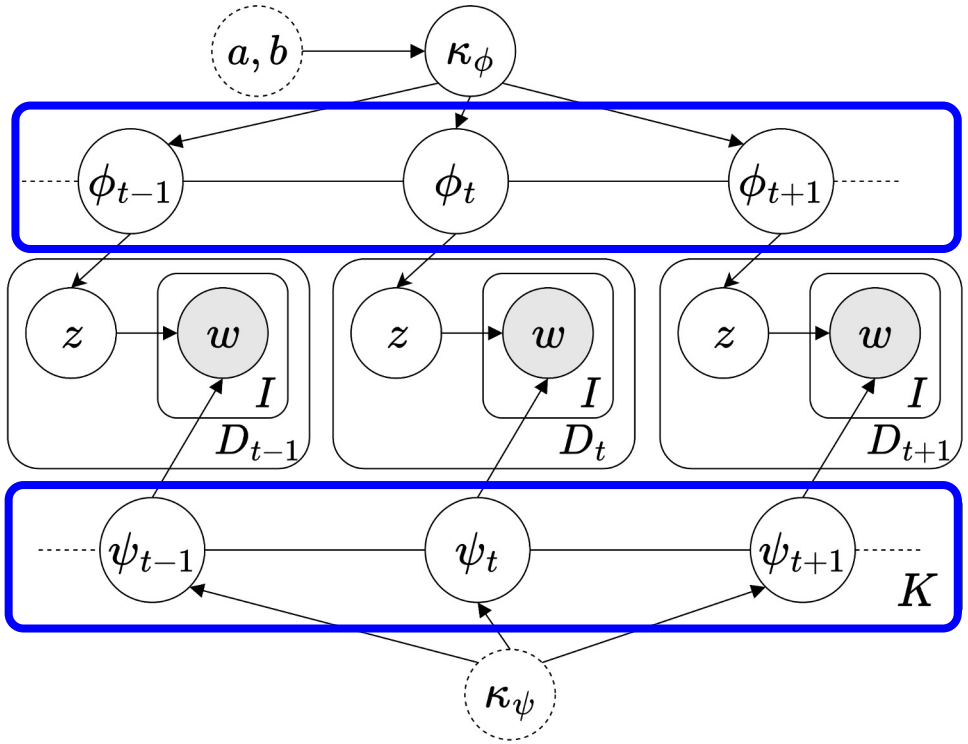


各スニペット

```
[chair, pass,
 horse,
 proceed,
 walk]
```

意味分布の時間発展

- 意味分布 ϕ_t と意味-単語分布 ψ_t は両方とも、時間に従って変化する (ガウス分布のマルコフ確率場)
  - トピック z が決まったら、そこから単語 w を生成



Draw $\kappa^\phi \sim Gamma(a, b)$
for time interval $t = 1..T$ **do**
  Draw sense distribution
  $\phi^t | \phi^{-t}, \kappa^\phi \sim \mathcal{N}(\frac{1}{2}(\phi^{t-1} + \phi^{t+1}), \kappa^\phi)$
  for sense $k = 1..K$ **do**
    Draw word distribution
    $\psi^{t,k} | \psi^{-t}, \kappa^\psi \sim \mathcal{N}(\frac{1}{2}(\psi^{t-1,k} + \psi^{t+1,k}), \kappa^\psi)$
  for document $d = 1..D$ **do**
    Draw sense $z^d \sim Mult(\phi^t)$
    for context position $i = 1..I$ **do**
      Draw word $w^{d,i} \sim Mult(\psi^{t,z^d})$

SCANの生成モデル

トピック z

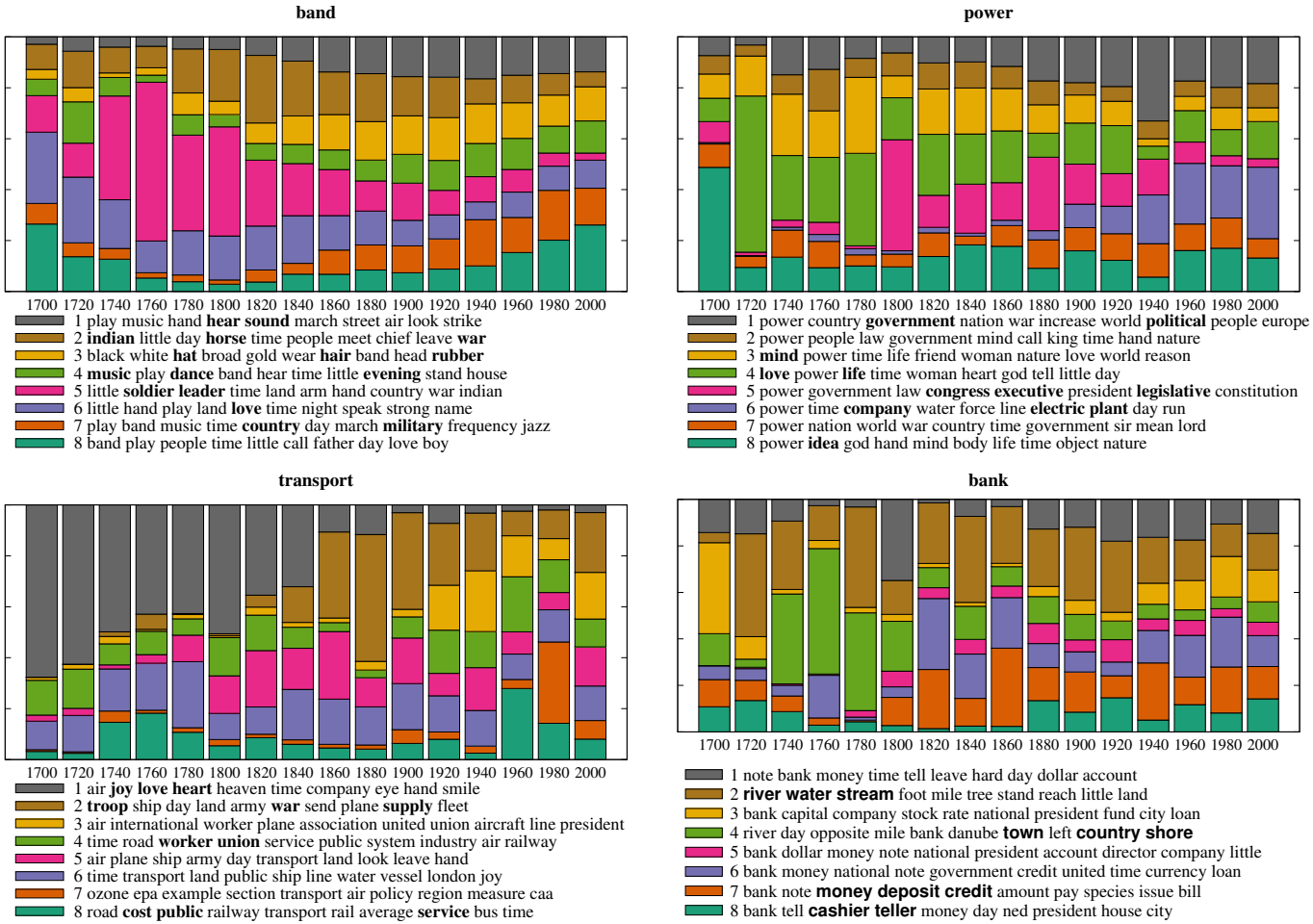
1853	[driver, make, room, trunk]	→馬車
1900	[chair, pass, horse, proceed, walk]	→馬車
1949	[tell, start, coach, head, top, school]	→スポーツ
1995	[available, inspection, bags, offer]	→ブランド
2003	[football, top, school, official, interview]	→スポーツ

- For *t* = 1..*T* do
  - $\phi_t|\boldsymbol{\phi}_{-t} \sim \mathcal{N}\left(\frac{1}{2}(\phi_{t-1}+\phi_{t+1}), \kappa_\phi\right)$ (トピックの事前確率)
  - For *k* = 1..*K* do (トピック-単語分布)

    $\psi_{t,k}|\boldsymbol{\psi}_{-t} \sim \mathcal{N}\left(\frac{1}{2}(\psi_{t-1,k}+\psi_{t+1,k}), \kappa_\psi\right)$
  - For snippet 1..*N(t)* do (時刻tのスニペット集合)
    - Draw $z \sim \text{Logistic-Mult}(\phi_t)$ (各スニペットのトピック)
    - For *i* = 1..*L* do
      - Draw $w_i \sim \text{Logistic-Mult}(\psi_{t,z})$ (単語を生成)

SCANによる推定例 (1700-2010)

- “band”, “power”, “transport”, “bank” の場合



意味の数の推定

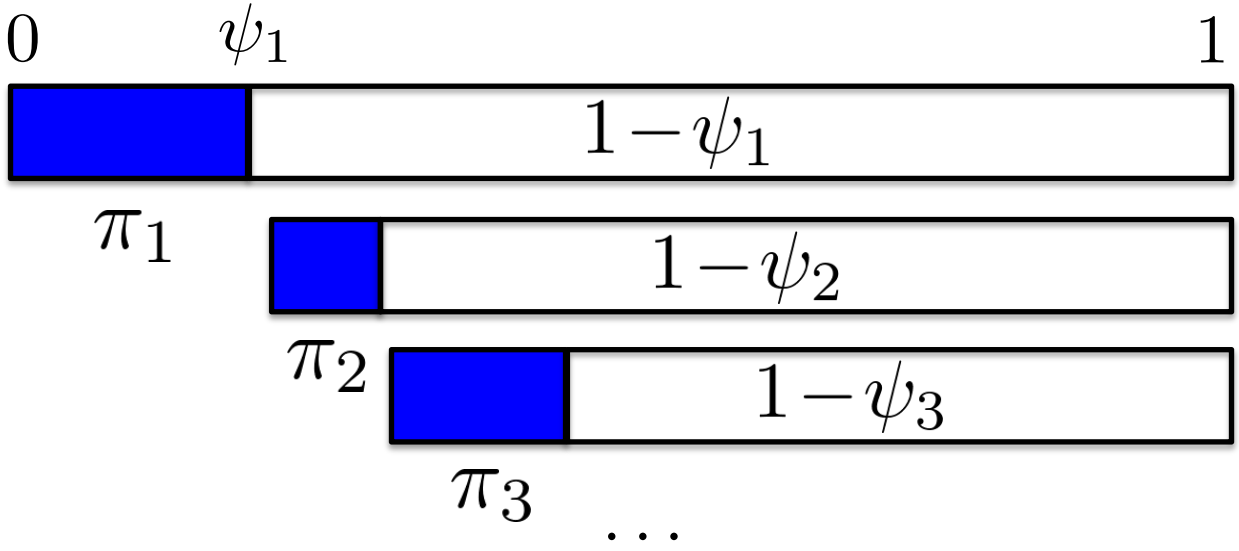
- SCANでは、“意味の総数” K は単語ごとにあらかじめ指定する必要がある
  - 一般には未知！& K によって結果が変わる
- 時間によって変わる意味の総数を、どうやって推定するか？
  - 使えるデータは、下のようなスニペット集合だけ

“coach”

```
1853  [driver, make, room, trunk]
1900  [chair, pass, horse, proceed, walk]
1949  [tell, start, coach, head, top, school]
2003  [football, top, school, official, interview]
```

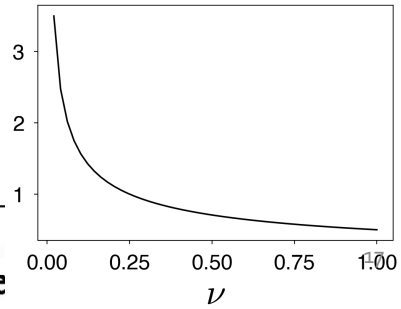
棒折り過程 (stick-breaking process)

- 無限次元の多項分布 $\pi = (\pi_1, \pi_2, \pi_3, \cdots)$ を生成する確率過程 (ディリクレ過程と等価 (Sethuraman 1994))

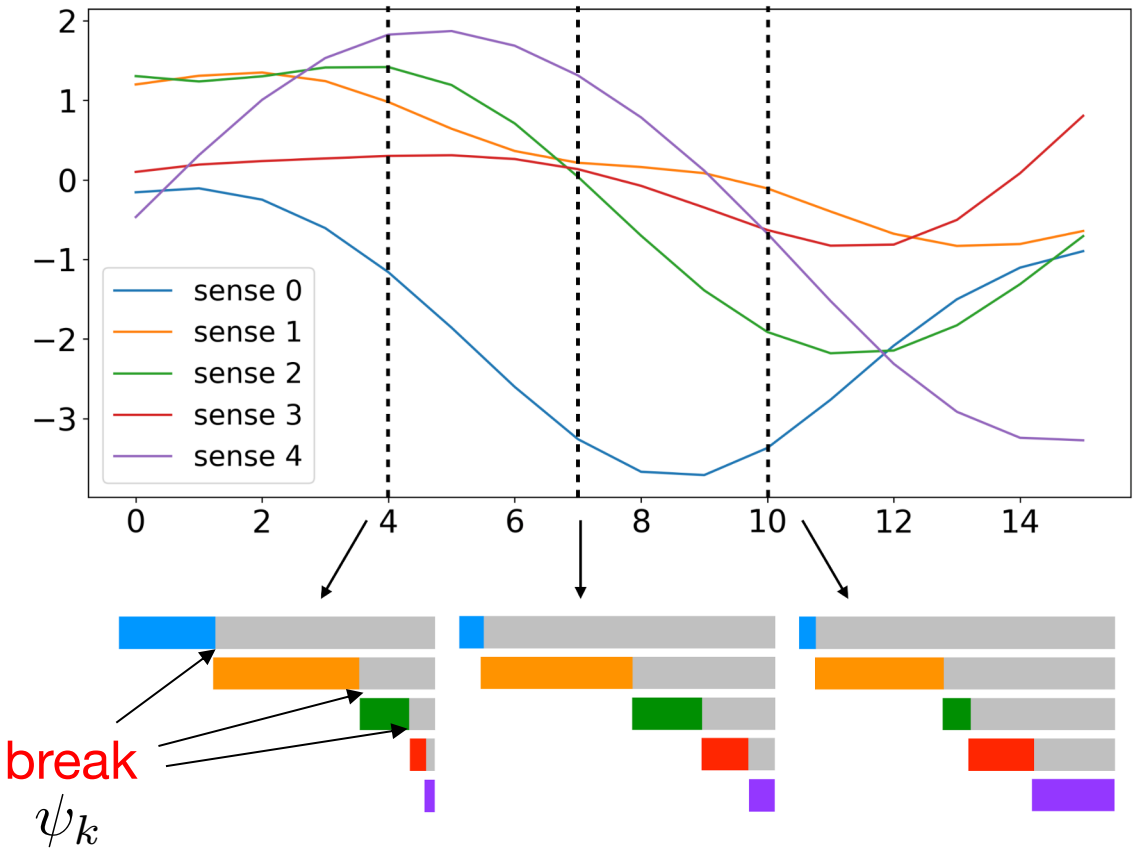


$$\pi_k = \psi_k \prod_{j=1}^{k-1}(1-\psi_j), \quad \psi_j \sim \mathrm{Be}(1, \gamma)$$

$\mathrm{Be}(\alpha, 1)$



ロジスティック棒折り過程 (Ren+ JMLR2011)



- 棒を折る割合を、新しく以下で定義

$$\psi_k = \frac{1}{1 + e^{-x_k}}$$

- ガウス確率変数をシグモイド関数で確率に変換
  → 本研究ではガウス過程を用いる

- 時間変化する意味の割合(の対数) x_k から、各時刻ごとに無限次元の分布を仮定して推定

推論

- All the inference is done by Gibbs sampling.
- Because Gaussian is not conjugate to multinomial, we employ Pólya-Gamma auxiliary variables ω to sample α_t (and transform into ϕ_t):

$$
\begin{aligned}
&p(\alpha_t \mid z, \boldsymbol{\alpha}_{-t}, \omega) \\
&\quad \propto \mathcal{N}\left(\omega^{-1} f(c) \mid \alpha_t\right) \mathcal{N}\left(\alpha_t \mid \boldsymbol{\alpha}_{-t}, \kappa_\phi^{-1}\right) \\
&\quad \propto \mathcal{N}\left(\alpha_t \mid \tilde{\mu}, \tilde{\kappa}_\phi^{-1}\right).
\end{aligned}
$$

- Precision parameter $\kappa_\phi^{(k)}$ should be estimated for each topic *k*:

$$
\begin{aligned}
&p(\kappa_\phi^{(k)} \mid \alpha_k, a, b) \\
&\quad = \mathrm{Ga}\left(a + \frac{T}{2}, b + \frac{1}{2}\sum_{t=1}^{T}(\alpha_{t,k} - \bar{\alpha}_k)\right)
\end{aligned}
$$

実際のテキストでの実験

- データ: Corpus of Historical American English (COHA)
  - 先行研究に基づいて、分析に使う単語を選択

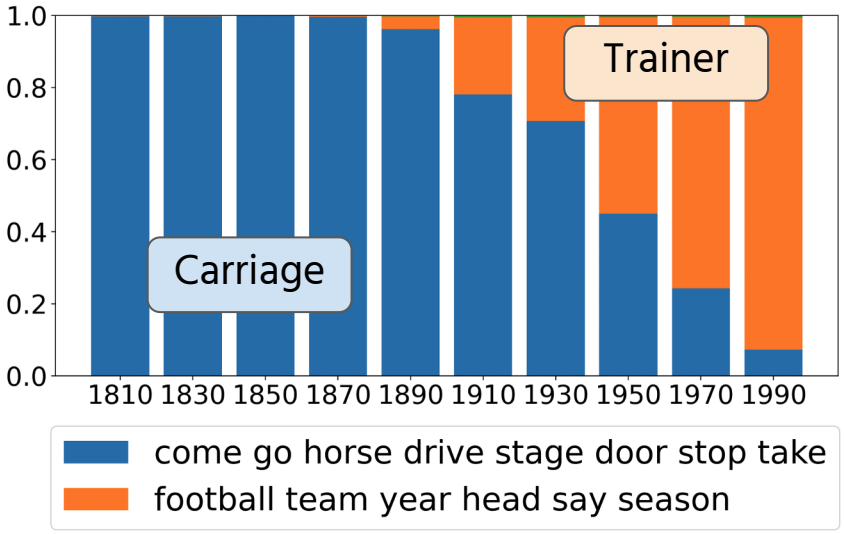
Word	Years	Samples	#Vocab
coach	1811–2009	9,758	11,962
record	1815–2009	33,992	23,886
power	1810–2009	142,527	42,932

- 数値的評価:
  - ベースライン: SCAN, HDP-LDA, BERT
  - 時間変化するトピック分布のcoherenceとdiversity を調べて評価

実験結果 (1)

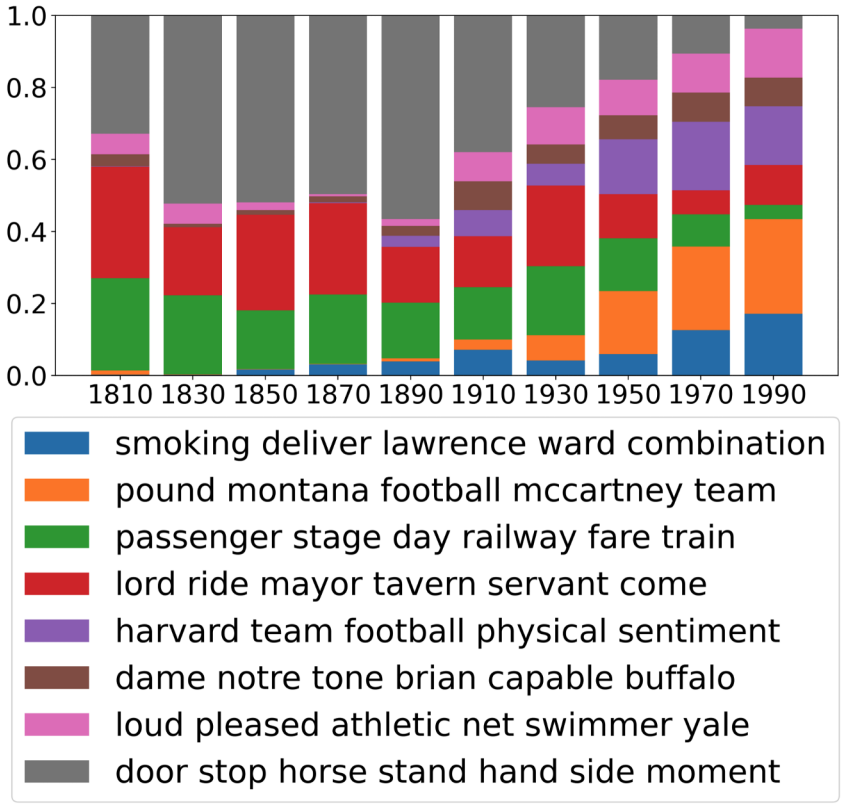
- “coach” : estimated senses=2

Infinite SCAN



Semantic shift: “carriage” → “trainer”

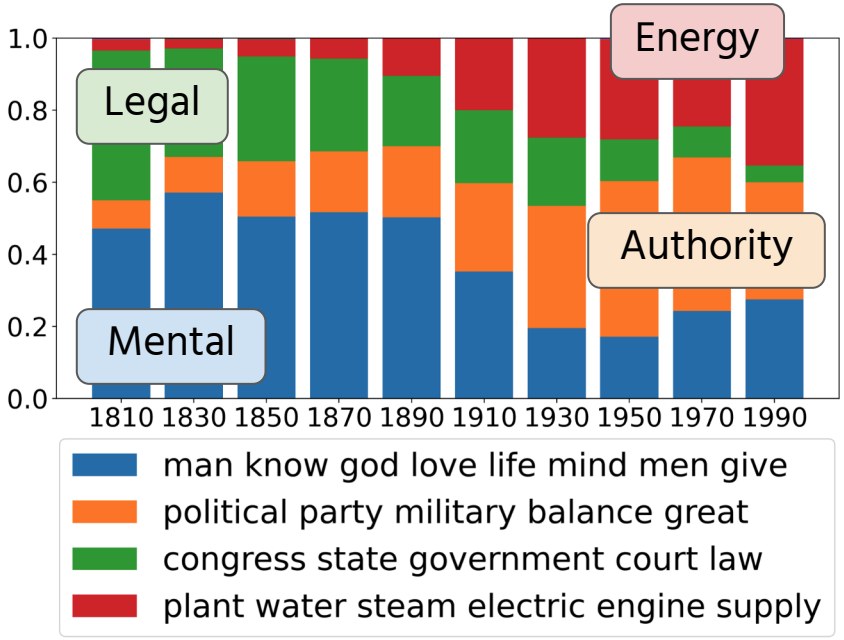
SCAN



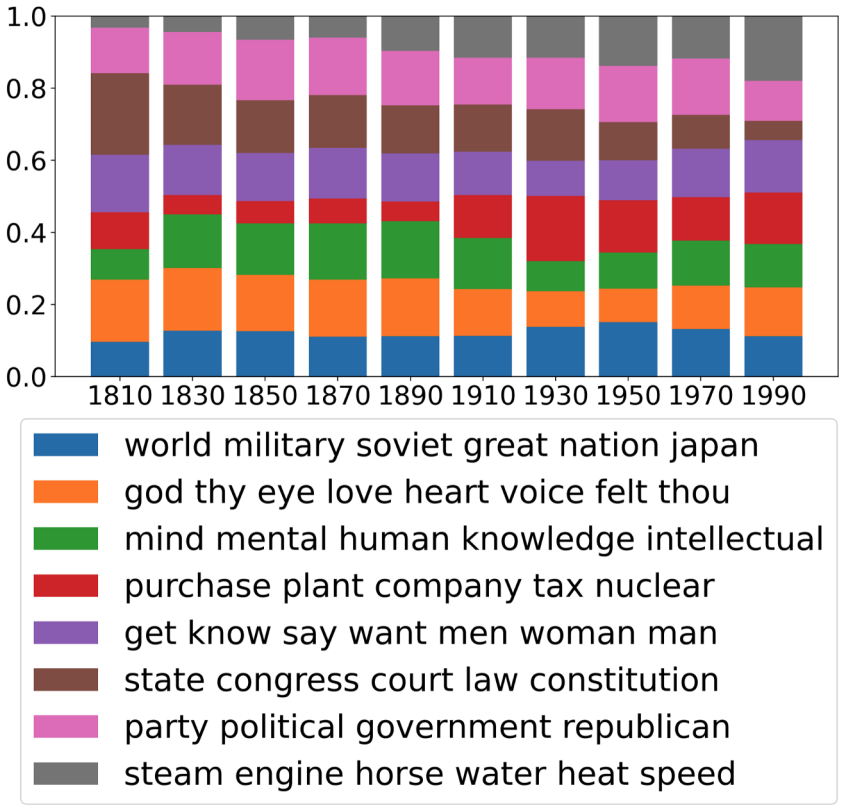
実験結果 (2)

- “power” : estimated senses=6

Infinite SCAN



SCAN



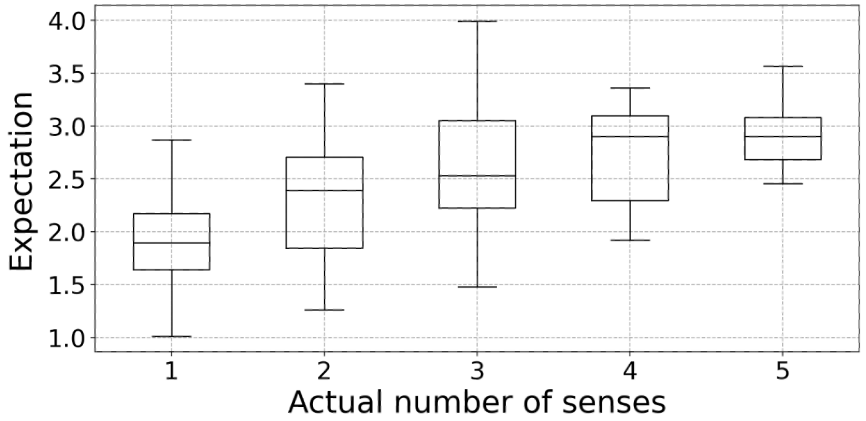
Sense birth: “energy”

実験結果 (3)

- モデルは意味の数を正しく発見できているのか？
- OntoNotesデータセットからランダムに120語を選んでモデルを計算し、正解率を計算

Model	Accuracy	PCC
HDP-LDA	0.258	0.019
BERT + K-means	0.217	0.026
BERT + DBSCAN	0.125	-0.070
SCAN (K = 5)	0.158	0.141
SCAN (K = 8)	0.000	0.087
Infinite SCAN	**0.358**	**0.474**

Prediction results for the number of senses.



Correlation between actual and estimated number of senses by Infinite SCAN.

まとめ

- 言葉の意味変化を解釈可能な形で扱うために、
  無限次元の離散分布であるディリクレ過程が時間的に
  変化する統計モデルを導入した
  - トピック事前分布、トピック-単語分布のいずれも
    時間発展する
  - Stick-breaking process+ガウス過程+ロジスティック回帰
  - 推定はPolya-GammaでGibbs samplerが作れる
- 対象の単語を含むスニペット集合について適用し、
  意味の数を事前に決めずに適切な意味変化が推定可能な
  ことを示した
- "Infinite SCAN: An Infinite Model of Diachronic Semantic Change".
  Seiichi Inoue, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura,
  Daichi Mochihashi. EMNLP 2022 (Long, oral), pp.1605-1616.