「現代語の意味の変化に対する計算的・統計力学的アプローチ」シンポジウム

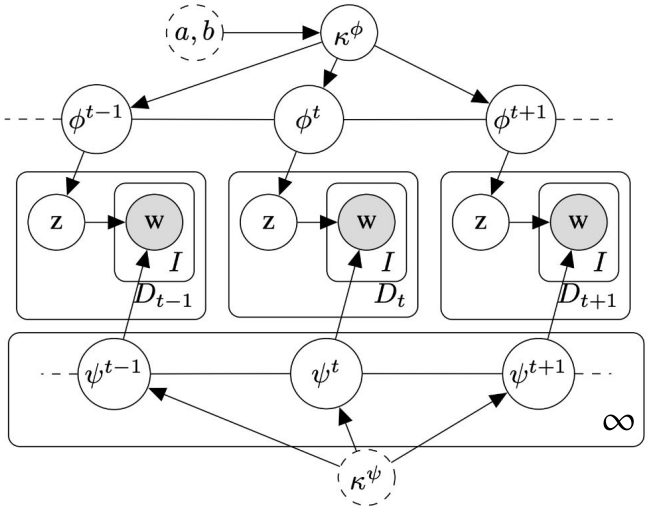
動的トピックモデルを用いた
単語の通時的な意味変化のモデル化とその応用

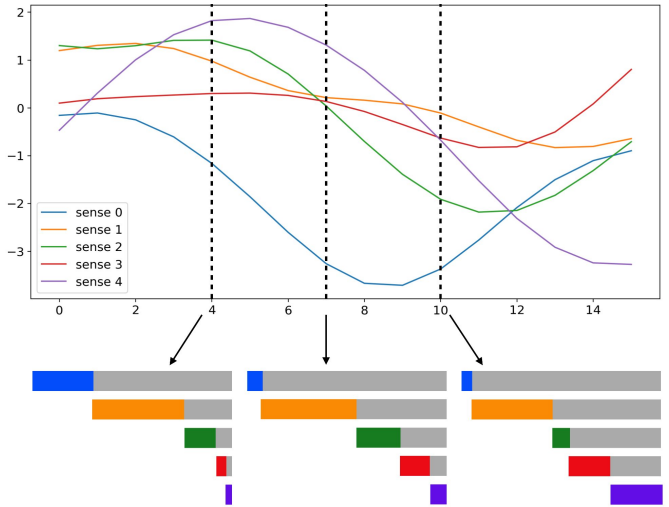
2022/03/09

東京都立大学 井上誠一
自然言語処理研究室 (小町研究室)

概要

- 動的トピックモデルを用いた意味変化のモデル化
- 単語の<u>意味変化</u>と<u>語義数</u>を同時に推定できる動的トピックモデルの提案
  - 擬似データを用いた実験で，擬似的な **意味変化と語義数を正しく推定できる** ことを検証
  - 英語 / 日本語データを用いて単語の意味変化を推定 / 分析





背景: 単語の意味変化

「言語は動的なシステムであり，常に進化し，話者とその環境の需要に適応している」[Aitchison, 2001]

- 音韻: 音のパターンの変化
- 統語: 文の組み立てのパターンの変化
- **意味: (文) / 単語の意味のパターンの変化**

e.g.) “cute” [Stevenson, 2010] の意味変化

18世紀: “賢い” → 19世紀後半: “狡猾な” → 現代: “魅力的な”

→ **このような意味変化を自動的に捉えたい**

背景: 単語の意味変化

自然言語処理的 (≠ 言語学的)なアプローチ

- 単語の「意味」を分布仮説 [Harris, 1954] に従って定義
  - 単語の意味はその文脈（周辺単語）によって特徴付けられるとする
- 通時的な意味変化をコーパスからモデル化
  - コーパスに現れない要因（社会的・歴史的な影響等）は無視

手法は様々

- 単語分散表現を用いた手法 [Kulkarni+, 2015; Hamilton, 2016; Bamler+, 2017]
  - モデルの推定が容易だが精緻さに欠ける
- トピックモデルを用いた手法 [Emms+, 2016, Frermann+, 2016]
  - モデルの推定が若干面倒だが意味変化を詳細に捉えることができる

単語の意味変化のモデル化

意味をどのように表現するか？

→ 分布仮説（単語の意味は単語の周辺単語によって特徴づけられる）

入力の設計

→ 対象単語の用例ごとに周辺単語を抽出 = スニペット

年代	用例	スニペット
1853	The driver made room for the trunk on the top of the **coach**.	{driver, make, room, trunk}
1900	The chair passed the **coach**, which immediately fell in behind it, the horses proceeding at a walk.	{chair, pass, fell, horse, proceed, walk}
1949	Tell him if I start **coaching** , it'll be as a head **coach** at a top school.	{tell, start, coach, head, top, school}
2003	The head football **coach**'s absolute dictatorship of the football field was reproduced.	{head, football, absolute, dictatorship, football, field, reproduce}

先行研究: SCAN

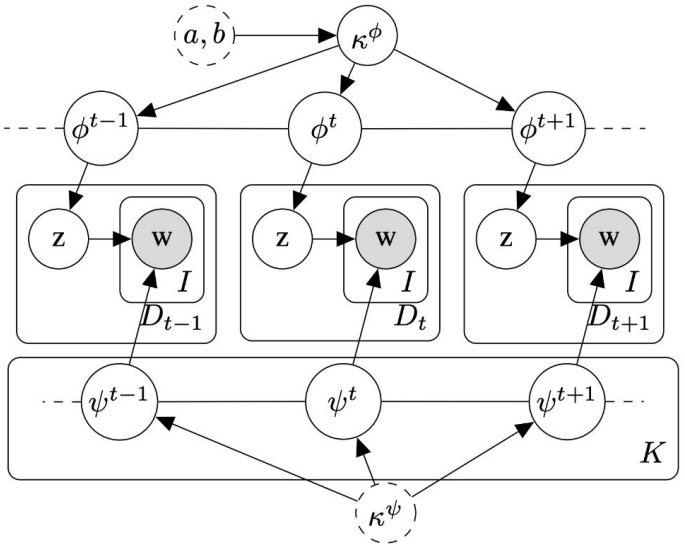
SCAN: Dynamic Bayesian Model of **S**ense **C**h**AN**ge [Frermann+, 2016]

対象単語の意味変化のモデル化:

① 対象単語の文脈単語（スニペット）集合を

② 自動的に分類する生成モデルを与え（=トピックモデル）

③ 時系列拡張をすることでモデル化（=動的トピックモデル）



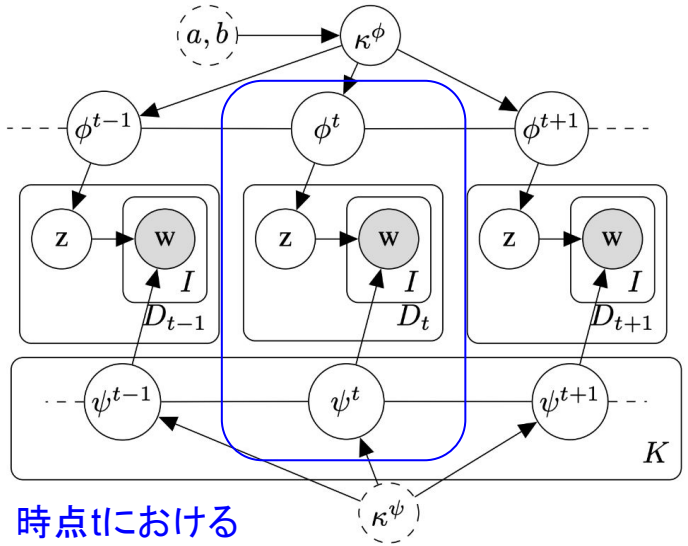
先行研究: SCAN

スニペットの生成過程（混合ユニグラムモデル）

Draw $\kappa^{\phi} \sim Gamma(a, b)$
for time interval $t = 1..T$ **do**
  Draw sense distribution
  $\phi^t | \boldsymbol{\phi}^{-t}, \kappa^{\phi} \sim \mathcal{N}(\frac{1}{2}(\phi^{t-1}$
  for sense $k = 1..K$ **do**
    Draw word distribution
    $\psi^{t,k} | \boldsymbol{\psi}^{-t}, \kappa^{\psi} \sim \mathcal{N}(\frac{1}{2}(\psi^{t-1,k} + \psi^{t+1,k}), \kappa^{\psi})$

for document $d = 1..D$ **do**
  Draw sense $z^d \sim Mult(\boldsymbol{\phi}^t)$
  for context position $i = 1..I$ **do**
    Draw word $w^{d,i} \sim Mult(\boldsymbol{\psi}^{t,z^d})$

混合ユニグラムモデル



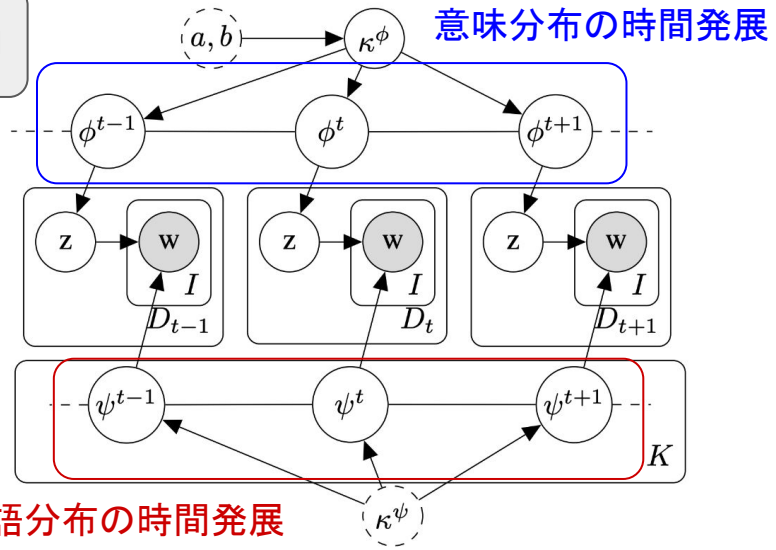
時点tにおける
スニペット集合の生成

先行研究: SCAN

GMRF（Gaussian Markov Random Field）を用いた時系列拡張

Gaussian Markov Random Field

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

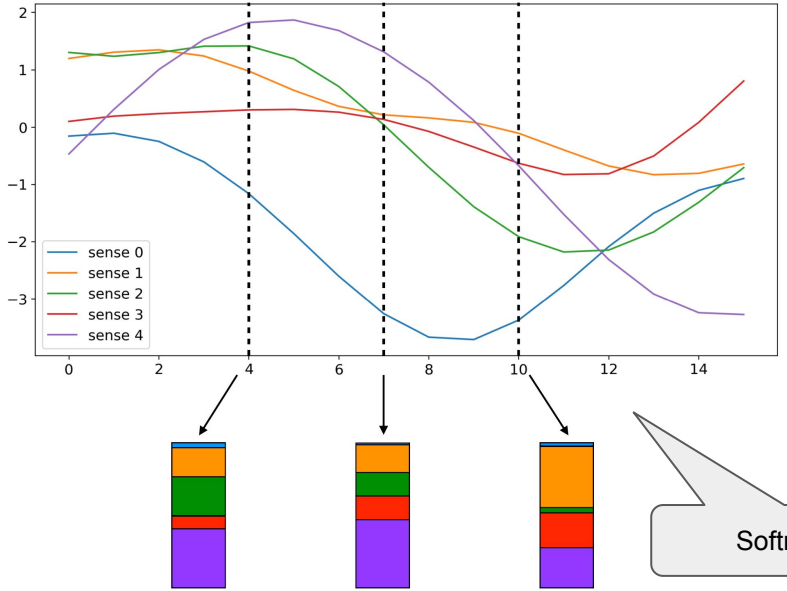


意味分布の時間発展

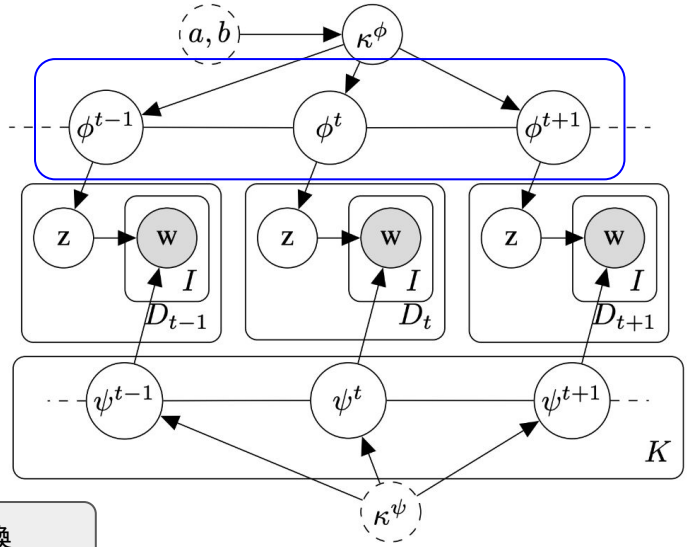
語義-単語分布の時間発展

先行研究: SCAN

ガウス事前分布とSoftmax変換



Softmax変換



先行研究: SCAN

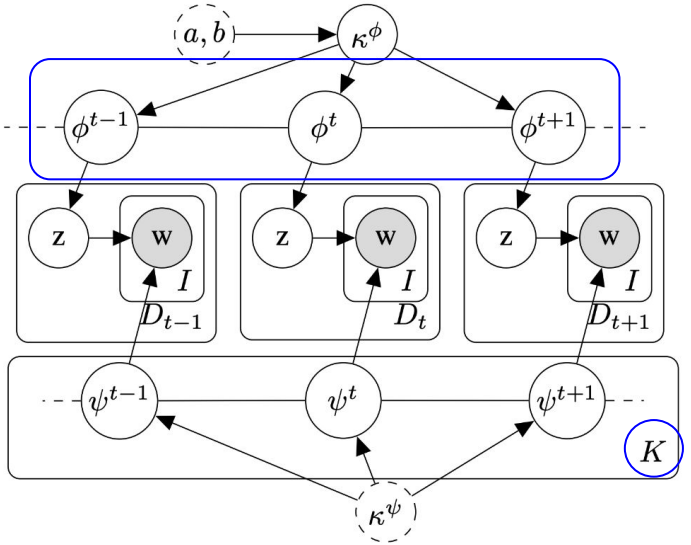
SCAN: スニペットに対し語義によるクラスタリングを行う確率的生成モデル

問題点:

実際に解析を行う単語の語義数は自明でないことが多いにも関わらず…

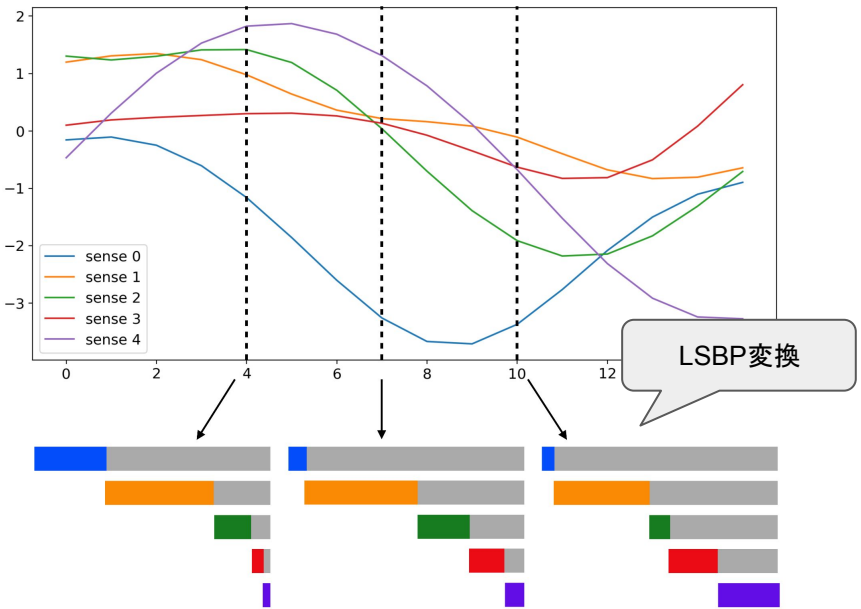
解析の対象単語の**語義数 = Kを事前に設定する必要**がある

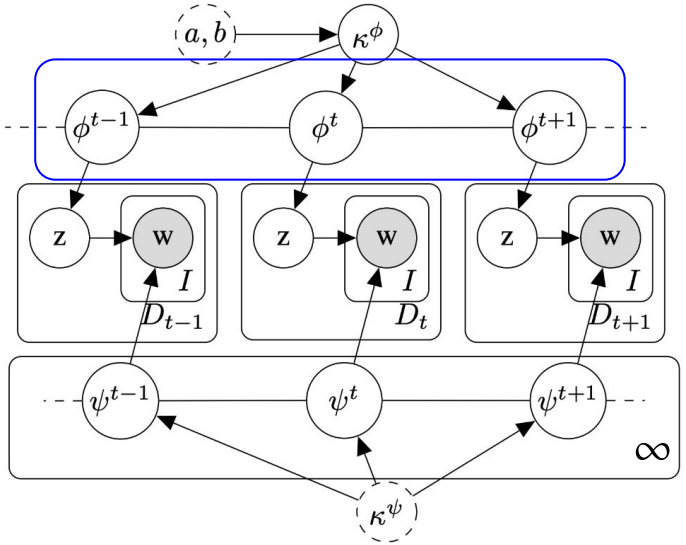
→ **語義数Kもデータから自動推定したい**



提案手法: Infinite SCAN

意味の分布の上に**ディリクレ過程** (Logistic Stick-Breaking process [Ren+, 2011])を考えることで**対象単語によって異なる語義数**を自動で推定できるモデル





提案手法: Infinite SCAN

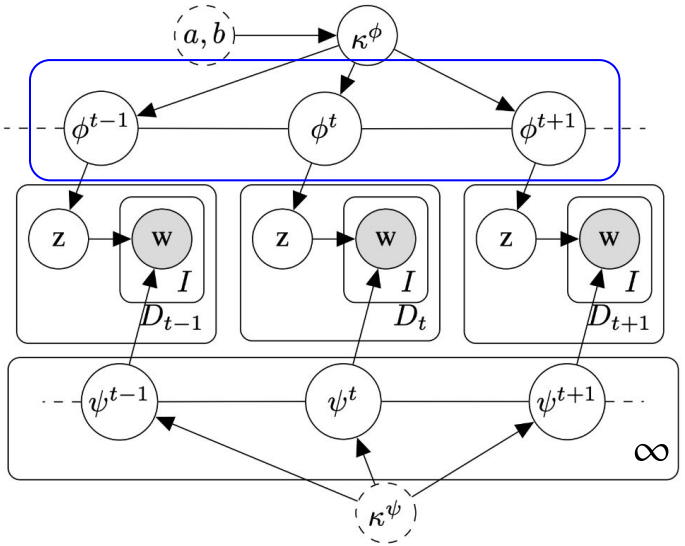
Infinite SCANの推定: ブロック化Gibbs Sampler

Stick-breaking表現のGaussian Prior φの推定:

- Pólya-Gamma分布 [Polson+, 2013] を用いたGibbsサンプラー [Linderman+, 2015]

Gaussianの精度パラメータκの推定:

- 各語義kごとに推定

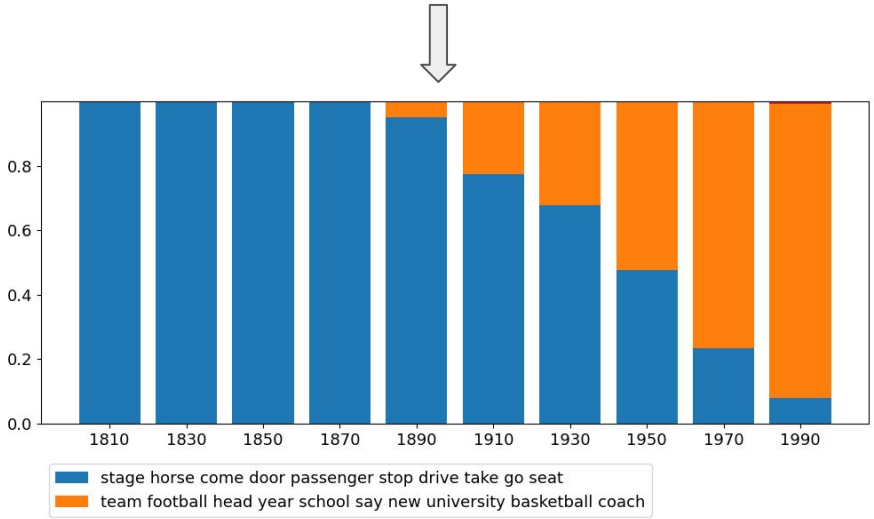


提案手法: Infinite SCAN

意味変化のモデル化

- 入力
  - スニペット集合
- 推定パラメータ
  - 意味分布
  - 語義-単語分布
  - （意味分布の変化速度）
- 出力（観察したいもの）
  - 意味分布
  - 各語義を代表する単語（確率上位単語; NPMI上位単語）

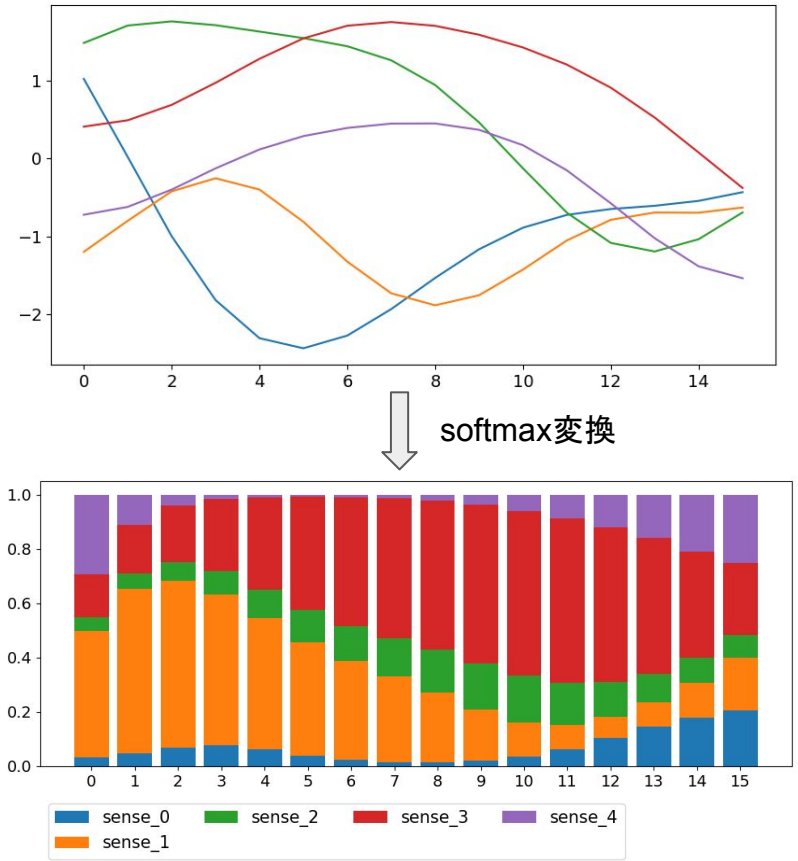
年代	用例	スニペット
1853	The driver made room for the trunk on the top of the **coach**.	{driver, make, room, trunk}
1900	The chair passed the **coach**, which immediately fell in behind it, the horses proceeding at a walk.	{chair, pass, fell, horse, proceed, walk}
1949	Tell him if I start **coaching** , it'll be as a head **coach** at a top school.	{tell, start, coach, head, top, school}
2003	The head football **coach**'s absolute dictatorship of the football field was reproduced.	{head, football, absolute, dictatorship, football, field, reproduce}



実験1: 実験設定

① 擬似データを用いた検証

- 擬似データの生成
  - ガウス過程からサンプルした曲線を Softmax 変換することで意味分布を生成
  - 各語義kにおける語義-単語分布はディリクレ分布からサンプリング
- 評価
  - ベースライン（SCAN）と比較
  - 真の意味分布と推定した意味分布の Kullback–Leibler距離
  - 語義数を 2 … 5と動かして検証



実験1: 擬似データによる検証

語義数 2 … 5の擬似データに対する推定でSCANを超える性能

- 意味変化と同時に語義数も推定しているInfinite SCANのほうが，推定された意味分布と真の分布のKullback–Leibler距離が小さい

→ さまざまな語義数の単語に対して意味変化を推定できそう

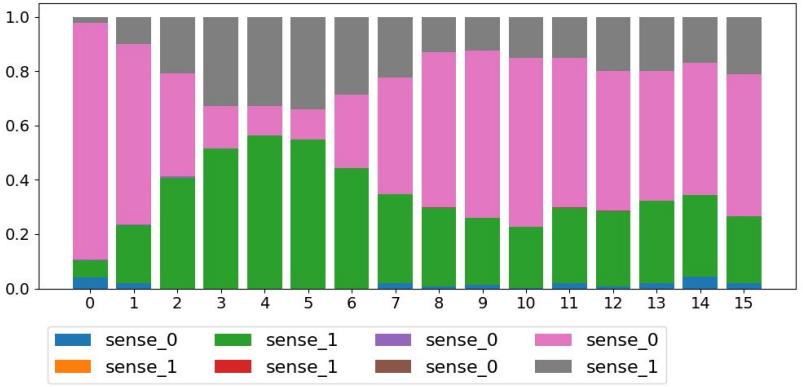
語義数 S_w	2	3	4	5
SCAN ($K = 8$)	0.2612	0.2788	0.1031	0.0085
Infinite SCAN	**0.0009**	**0.0016**	**0.0044**	**0.0043**

実験1: 擬似データによる検証

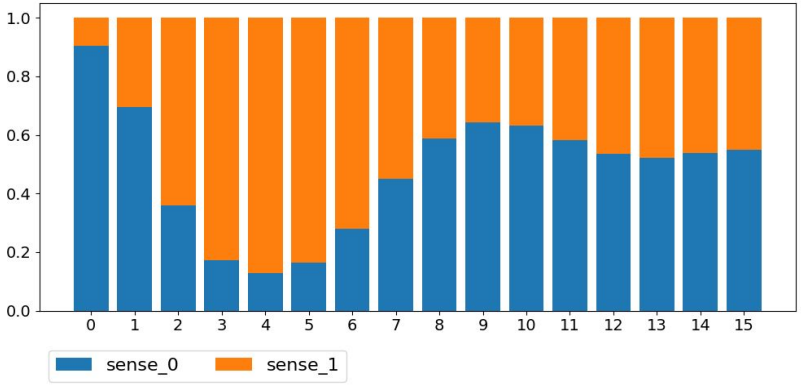
語義数 = 2の擬似データに対する推定

KL divergence:

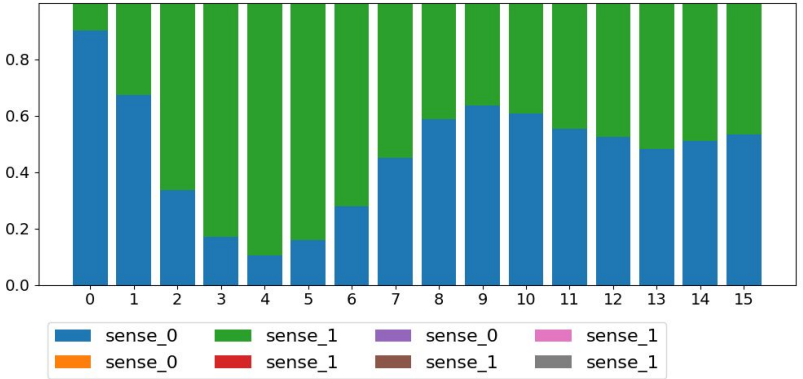
SCAN = 0.2612, 提案手法 = **0.0009**



SCANの推定結果



正解データ



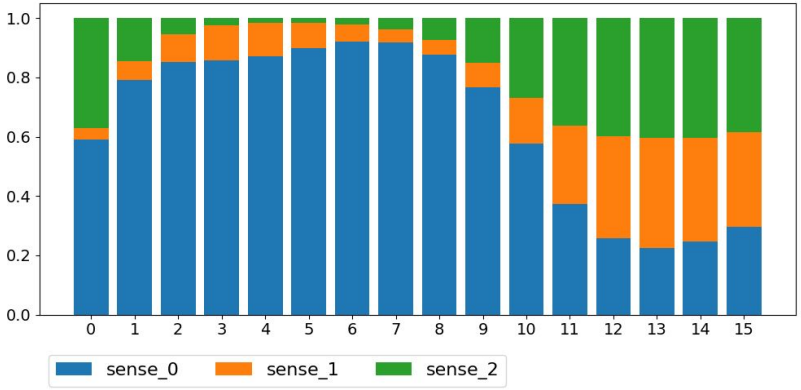
Infinite SCANの推定結果

実験1: 擬似データによる検証

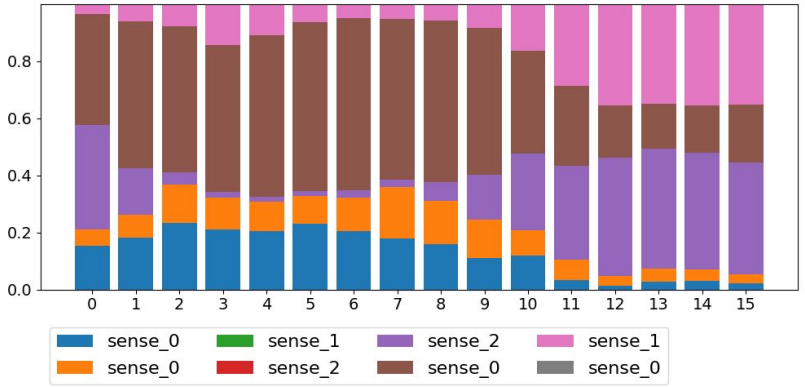
語義数 = 3の擬似データに対する推定

KL divergence:

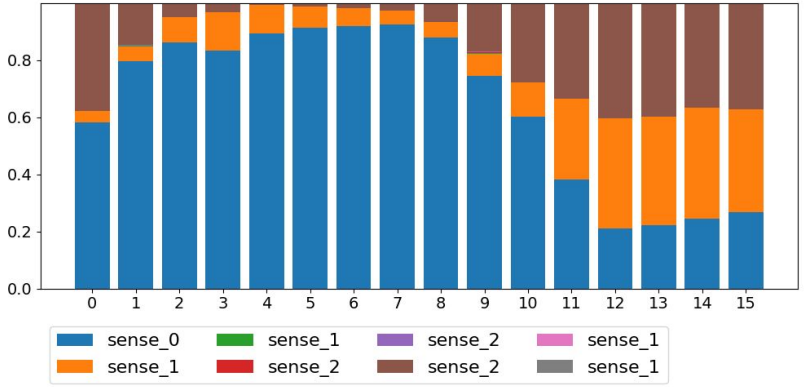
SCAN = 0.2788, 提案手法 = **0.0016**



正解データ



SCANの推定結果



Infinite SCANの推定結果

実験2: 実験設定

② 実データを用いた評価と分析

- データセット
  - 英語: Corpus of Historical American English (COHA) (1810-2009)
  - 日本語: 日本語歴史コーパス (CHJ) (1874-1997)
  - 対象単語を評価用 / 分析用に選定しスニペットを作成
- 評価
  - 英語コーパスを用いる
  - ベースライン（SCAN）と比較
  - 意味のcoherenceとdiversity
- 分析
  - 英語 / 日本語コーパスを用いる

コーパス	年代	単語数
COHA （英語）	1810–2009	142,587,656
CHJ（日本語）	1874–1997	36,701,284

単語	年代	スニペット数	語彙サイズ
image	1810–2009	19,499	19,104
nature	1810–2009	83,188	33,375
pass	1810–2009	36,605	24,555
record	1815–2009	33,992	23,886
coach	1811–2009	9,758	11,962
power	1810–2009	142,527	42,932
団塊	1895–1997	92	634
取り組む	1887–1997	647	2,083

実験2: 実データを用いた評価 / 分析

評価指標

- **Coherence**: 同一語義内の確率上位単語ペアの意味的な類似度から計算

e.g.) f(political, party) + f(political, government) …

- **Diversity**: 全ての確率上位単語のうちどれくらいユニークな単語が含まれるか

e.g.) 20 / 21 = 0.952

Sense 1	Sense 2	Sense 3
political	congress	plant
party	constitution	water
europe	government	electric
war	executive	purchase
government	law	supply
country	court	nuclear
economic	legislative	steam

powerの語義-単語分布

実験2: 実データを用いた評価 / 分析

Coherence（意味の一貫性）とDiversity（語義の多様性; 非冗長性）においてSCANを超える性能

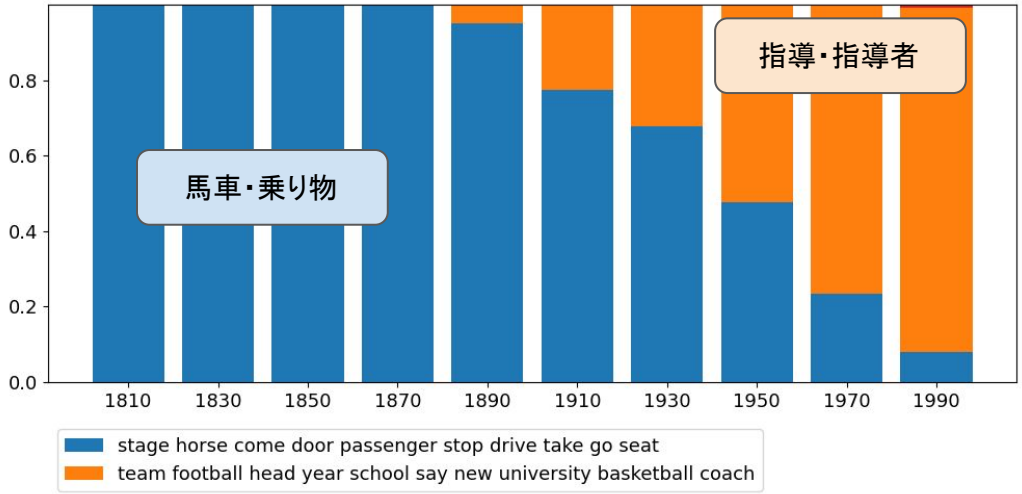
- SCANは語義数を自動的に推定できないため，“無駄な”語義が出来てしまう
- “無駄な”語義を作らないInfinite SCANはCoherence, Diverstyともに良い

	Coherence	Diversity
SCAN ($K = 8$)	0.124	0.275
Infinite SCAN	**0.146**	**0.398**

実験2: 実データを用いた評価 / 分析

対象単語 = ‘coach’の推定結果

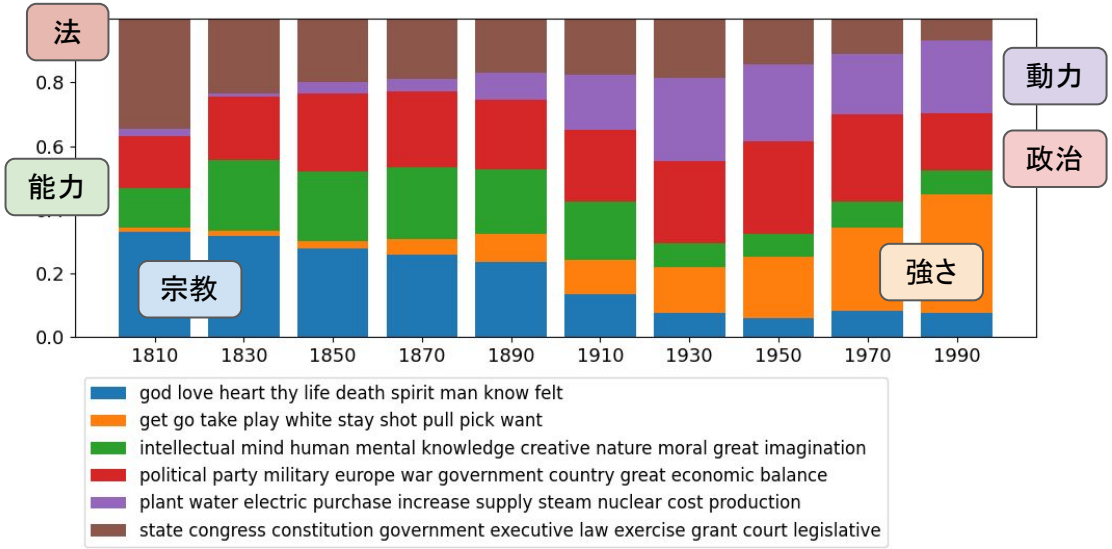
- 推定された語義数 = 2
- “馬車”や“乗り物”という意味から，“指導”や“指導者”といった意味に変化



実験2: 実データを用いた評価 / 分析

対象単語 = ‘power’の推定結果

- 推定された語義数 = 6
- 過去支配的だった意味
  - 宗教上の力
  - 肉体上 / 精神上の自然の能力
  - 法的な力
- 新しく生まれた / 支配的になった意味
  - 強さ
  - 政治的な力 / 競争力
  - 動力 / 電力

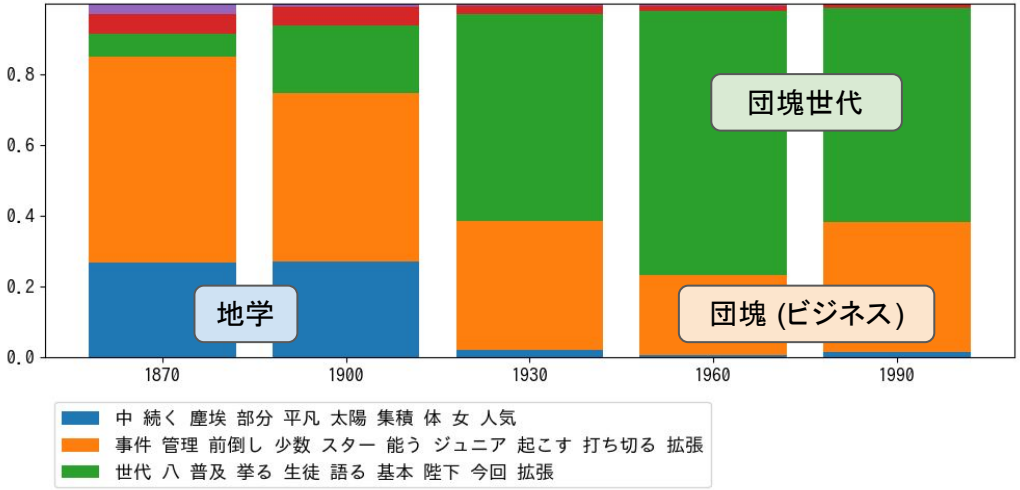


実験2: 実データを用いた評価 / 分析

対象単語 = ‘団塊’の推定結果

- 推定された語義数 = 3
- 地学関係の意味から“団塊世代”という意味に変化
  - 推定された結果だと“団塊世代”の粒度が少し細かい

#1900 - 1950あたりのデータがスパースなため平滑化が効きすぎている

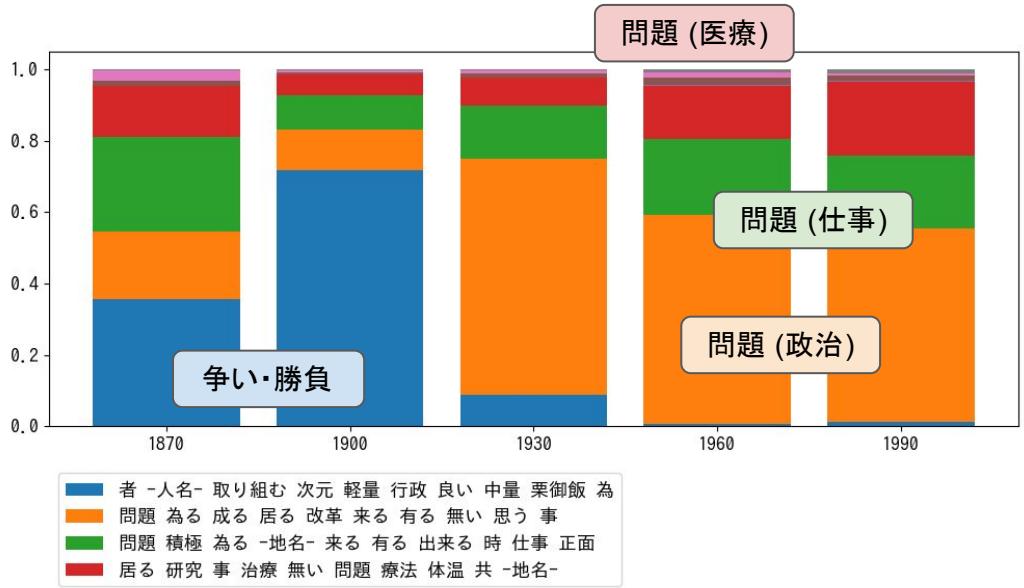


実験2: 実データを用いた評価 / 分析

対象単語 = ‘取り組む’の推定結果

- 推定された語義数 = 4
- 争いや勝負（特に相撲）における意味から問題に取り組むという意味に変化
  - “問題に取り組む”における意味の粒度が少し細かい（政治 / 仕事 / 医療）

#粒度の細かさはスニペット前処理でかなり変わる



まとめと今後の展望

- SCAN: 動的トピックモデルを用いた意味変化のモデル化
- Infinite SCAN: 単語の意味変化と語義数を同時推定する動的トピックモデル
  - ディリクレ過程をガウス確率場の上に考えることで語義数をデータから推定
  - 擬似データを用いた実験で，擬似的な **意味変化と語義数を正しく推定できる** ことを検証
  - 英語 / 日本語データを用いて単語の意味変化を推定し分析
- 今後の展望
  - 文脈単語集合（スニペット）の前処理の改善
    - 低 / 高頻度語の処理
    - 「どのような語彙だとうまく意味変化を表現できるか」の定式化
  - モデル推定の安定化
    - 時点が少ないデータに対するフィルタリング / 平滑化
  - 語義数の自動推定に対する外的な定量評価
    - 知識ベース（WordNet等）や辞書を用いて語義数が自明な単語を使用