Infinite SCAN:意味の数とその時間変化を同時に推定する統計モデル

持橋大地
情報・システム研究機構 統計数理研究所
daichi@ism.ac.jp

日本英語学会 “深層学習時代の言語研究” シンポジウム
2023-11-4 (土)

今日の話

- 自然言語処理の分野で、言語の分析のために
  開発した二つの統計モデル
- Infinite SCAN (EMNLP 2022):
  各単語の意味の数と、その時間変化を同時に推定する
  ことができる統計モデル (Inoue+ 2022)
- Holographic CCG (ACL 2023):
  潜在的なベクトル空間において、CCGの合成演算を
  二つのベクトルの再帰的な合成として表し、
  文の木構造をベクトル空間上のグラフとして表す
  研究 (Yamaki+ 2023)

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？

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、前後の文脈に同時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

スニペットの抽出

目標単語: *“coach”*

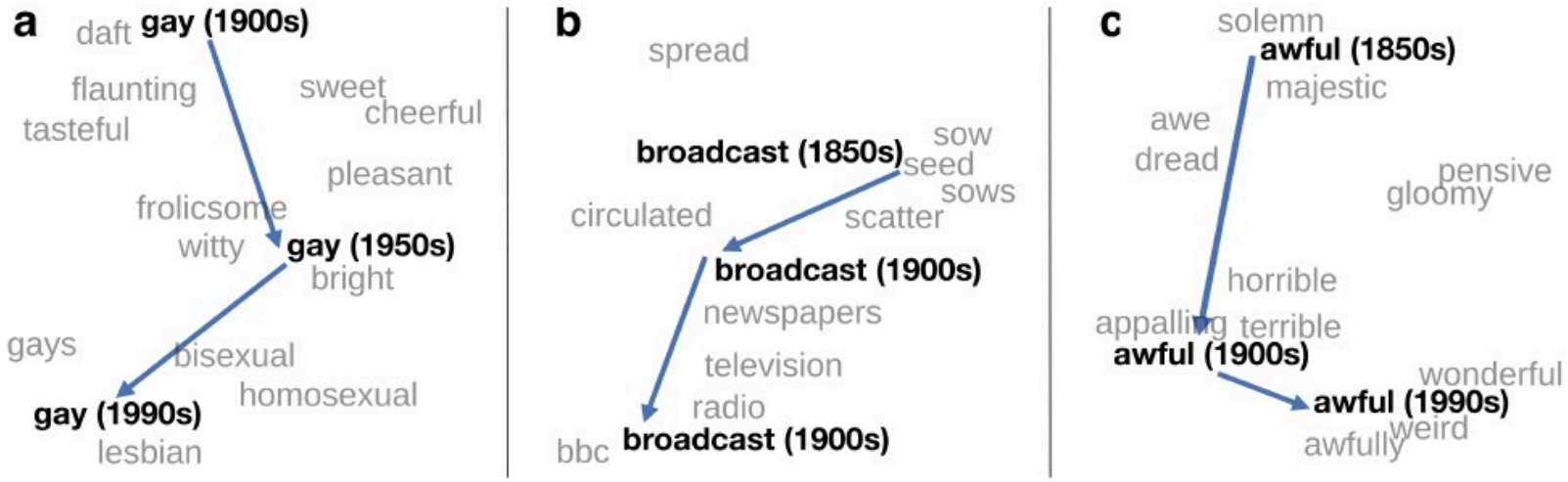
通時コーパス

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.

単語の意味変化の既存研究

(Hamilton+ 2016)

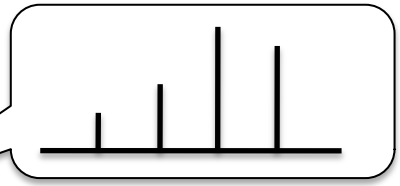


- 単語の前後の文脈(スニペット)から、単語の**埋め込みベクトル**を計算 (word2vecなど)
  →時間的な変動を観察する
- 変化したことはわかるが、何がどう変化したかはわからない！

得られた“coach”のスニペット

		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



トピックの教師なし学習

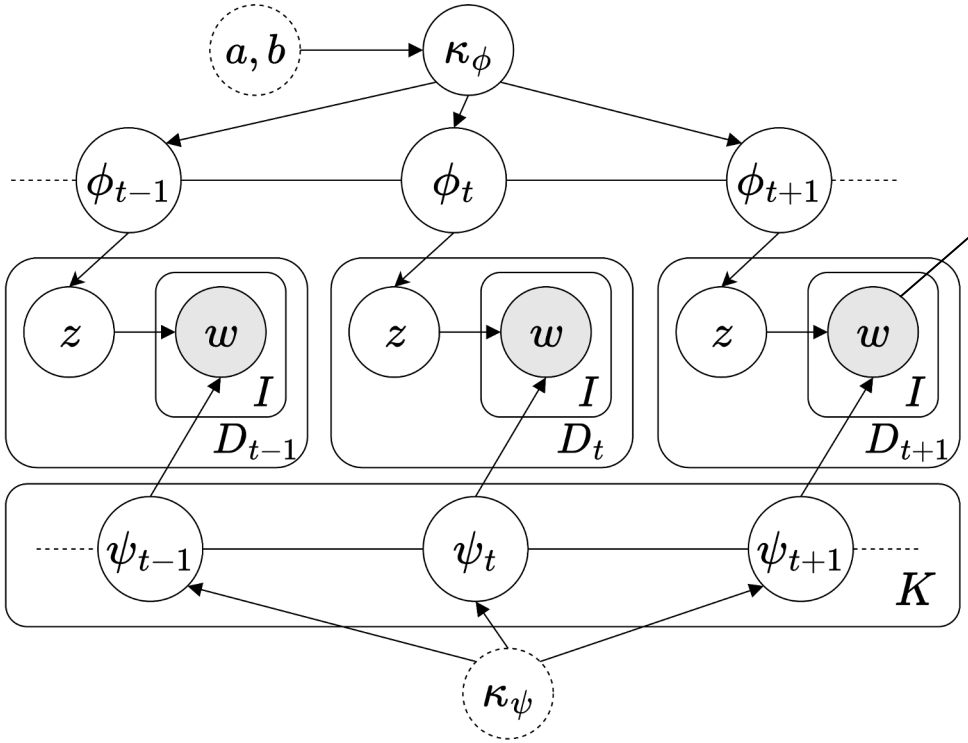
- 各スニペットの意味(トピック)は書かれていない！
  → 実は、自動的に学習できる
- アルゴリズム：
  - 各スニペットを、確率的に*K*個のトピック z のどれかに割り当てる ($\phi(z)$ がわかる)
  - 各トピック z に割り当てられたスニペットの文章全体から、$\psi(w|z)$ を更新
  - 以上を繰り返す
- スニペットのクラスタリングに相当している

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がある
- 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

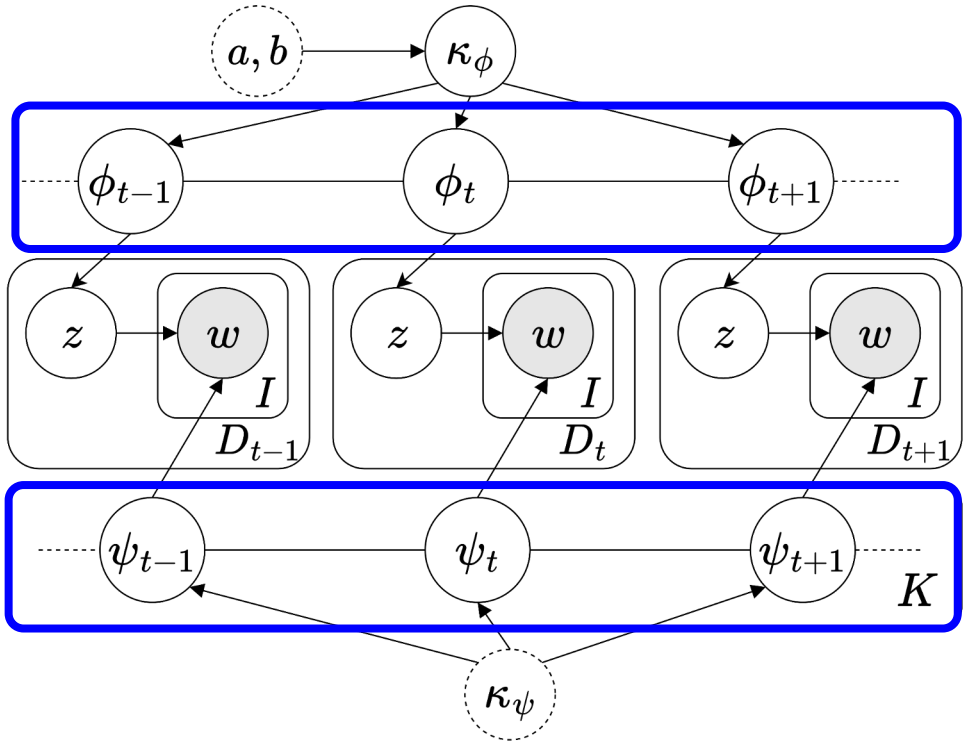


各スニペット

```
[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



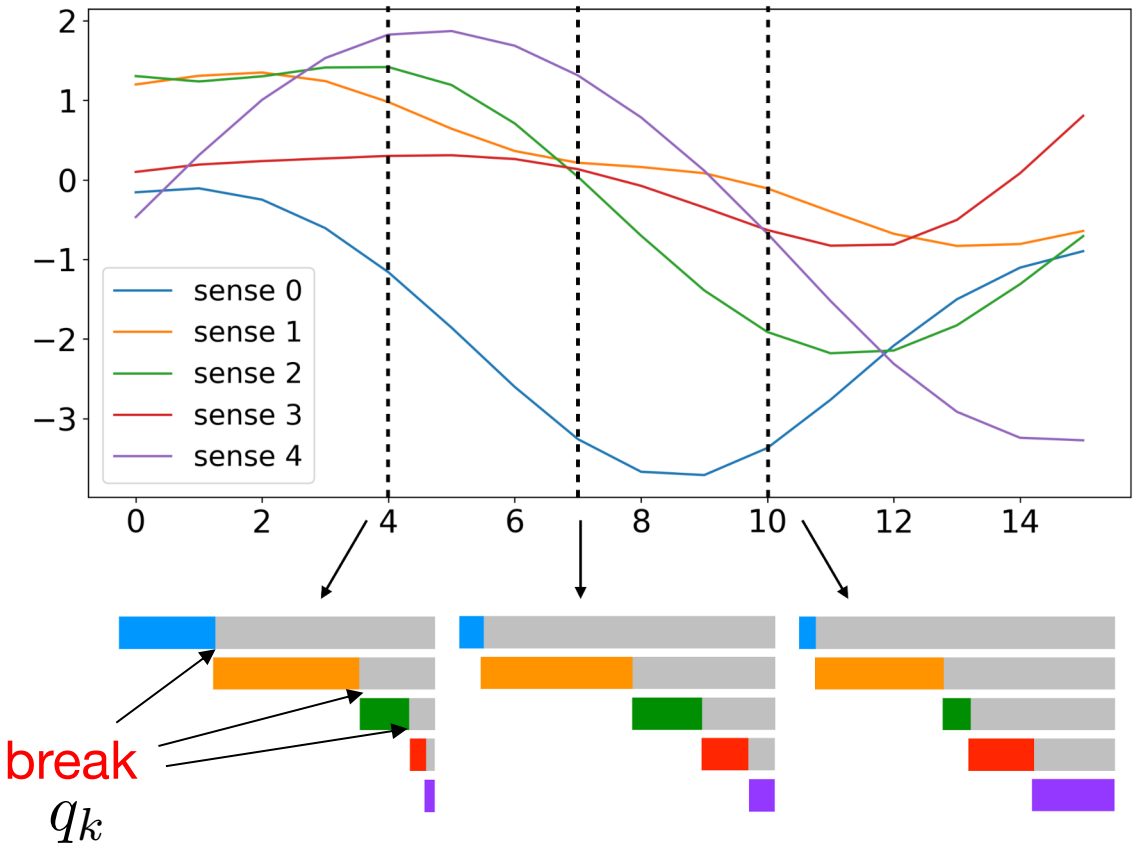
Draw $\kappa^\phi \sim Gamma(a, b)$
for time interval $t = 1..T$ **do**
  Draw sense distribution
  $\phi^t | \phi^{-t}, \kappa^\phi \sim \mathcal{N}(\frac{1}{2}(\phi^{t-1} + \phi^{t+1}), \kappa^\phi)$
  for sense $k = 1..K$ **do**
    Draw word distribution
    $\psi^{t,k} | \psi^{-t}, \kappa^\psi \sim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

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！
- 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

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

ロジスティック棒折り過程 (Ren+ 2011)



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

(シグモイド関数で確率に変換)

- 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

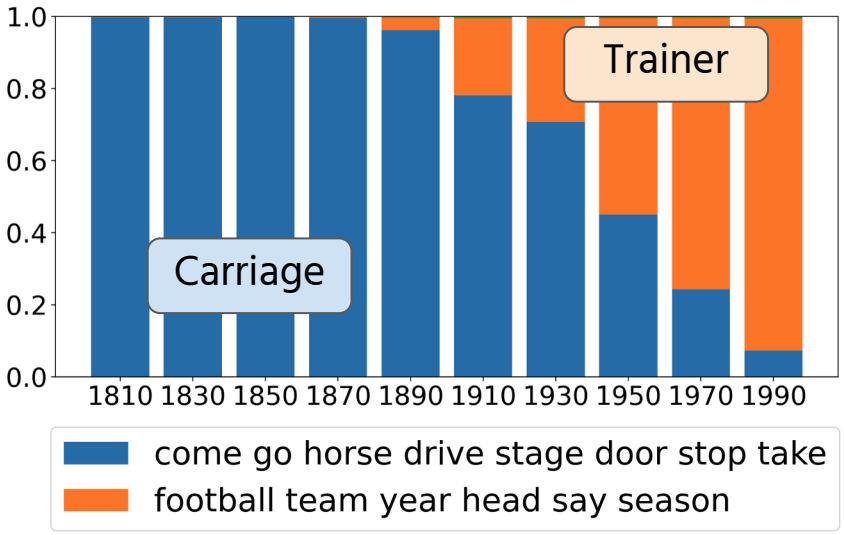
Word	Years	Samples	#Vocab
coach	1811–2009	9,758	11,962
record	1815–2009	33,992	23,886
power	1810–2009	142,527	42,932

- 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)

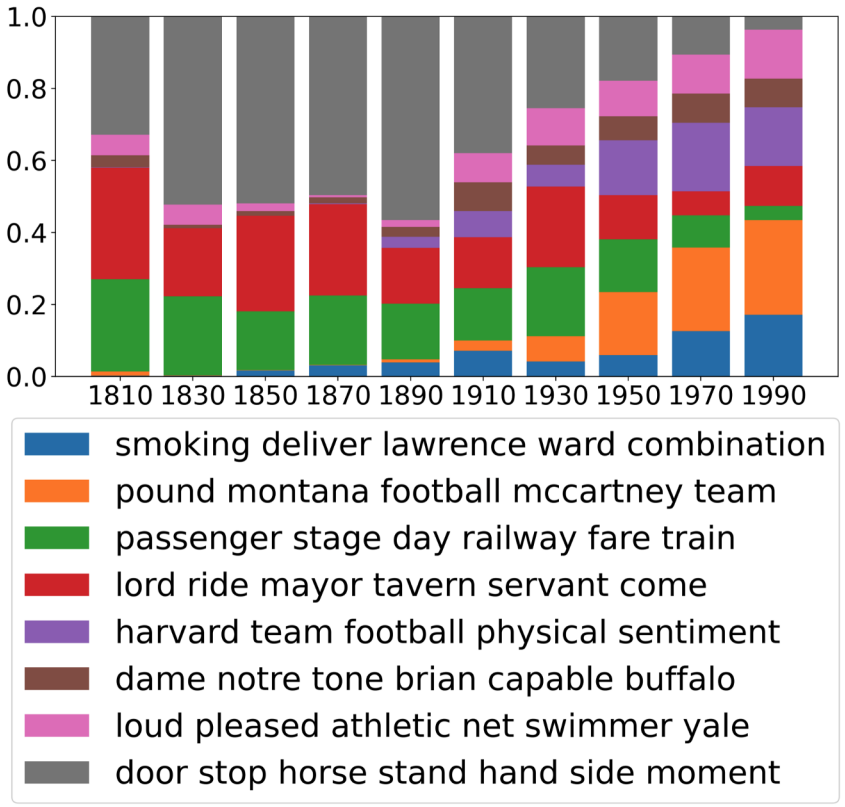
- “coach” : estimated senses=2

Infinite SCAN



Semantic shift: “carriage” → “trainer”

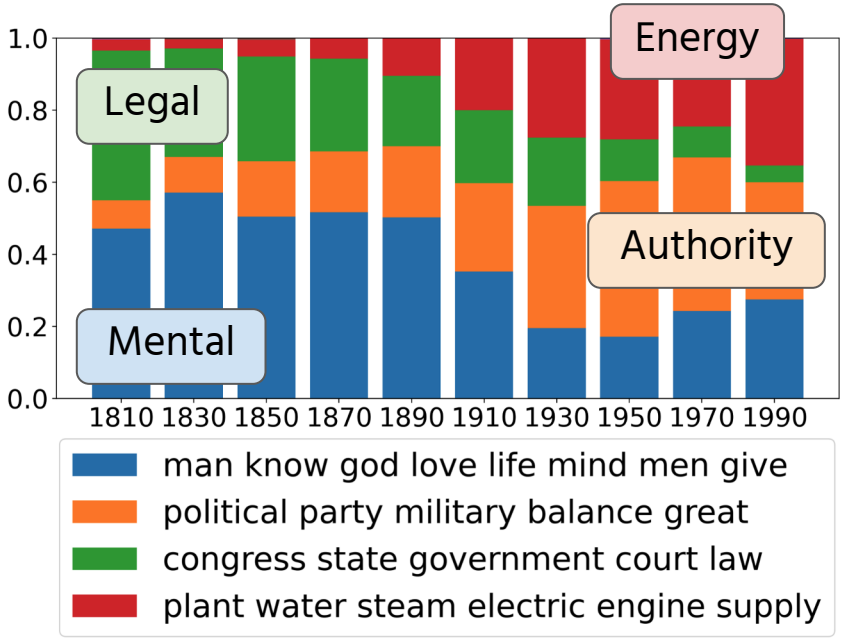
SCAN



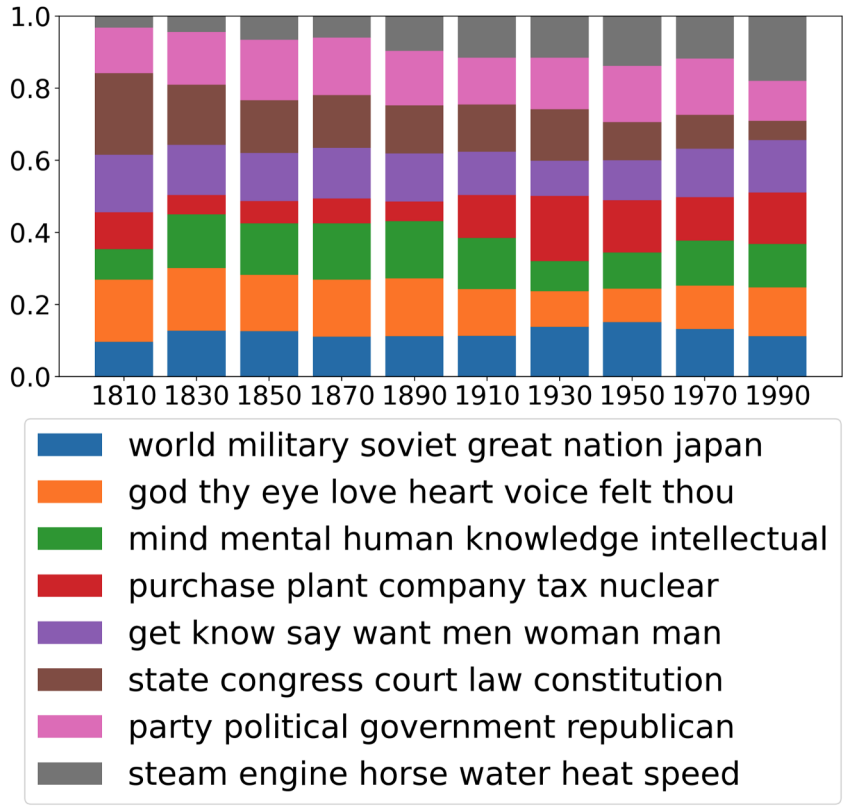
実験結果 (2)

- “power” : estimated senses=6

Infinite SCAN



SCAN



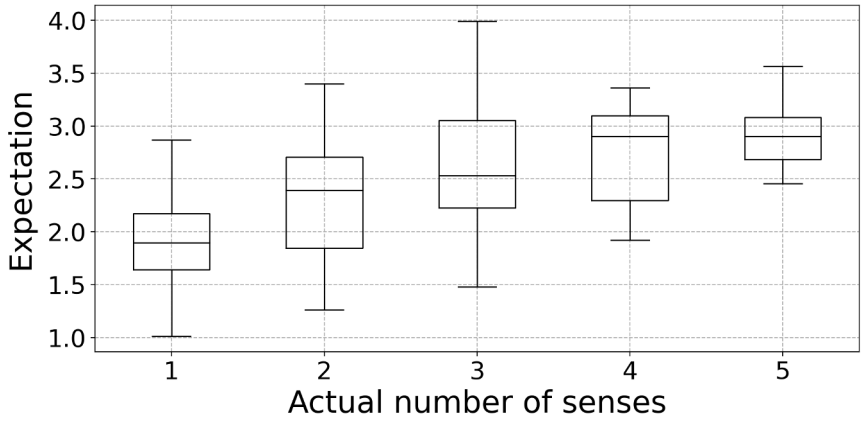
Sense birth: “energy”

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**

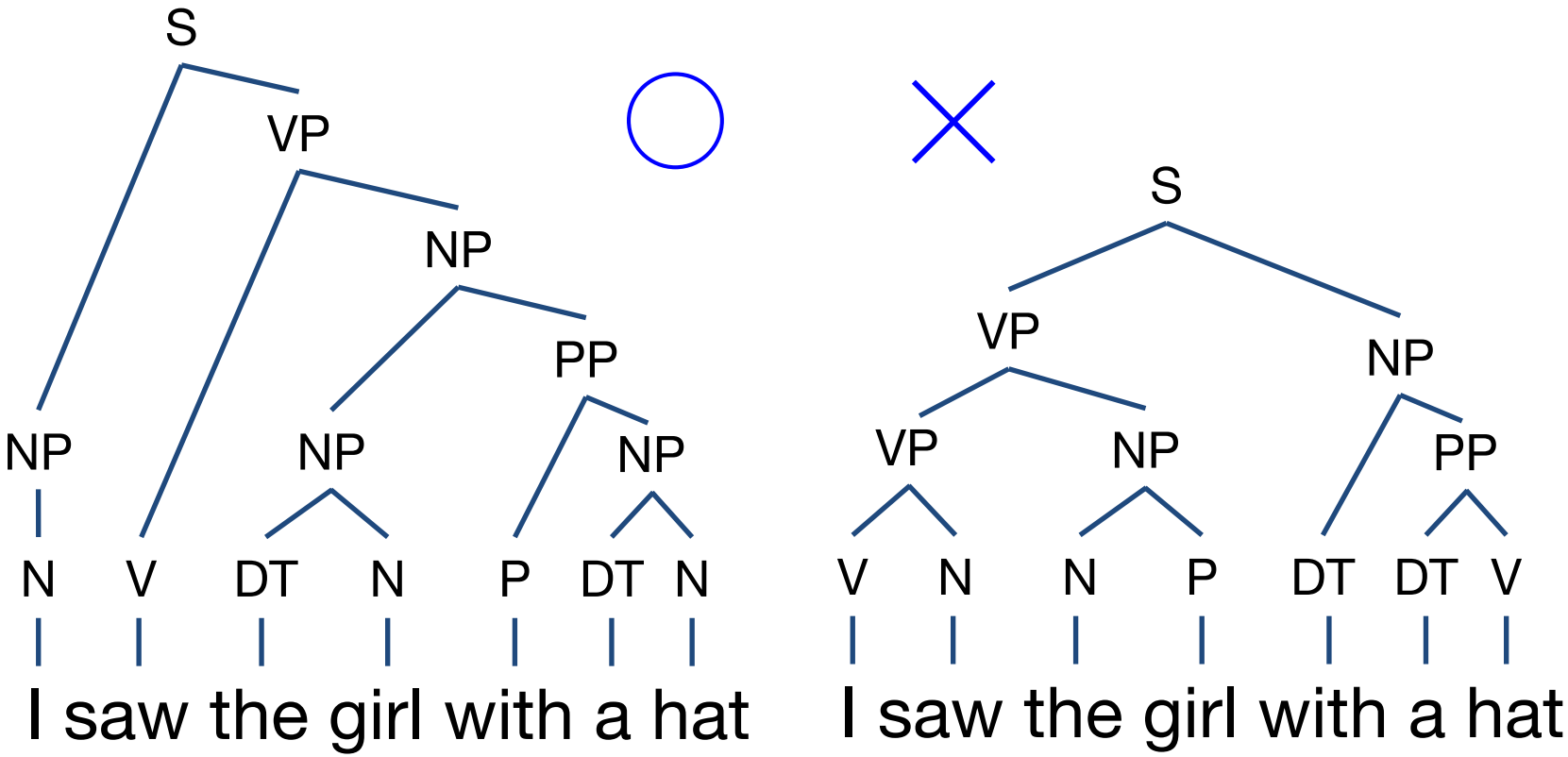
Prediction results for the number of senses.



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

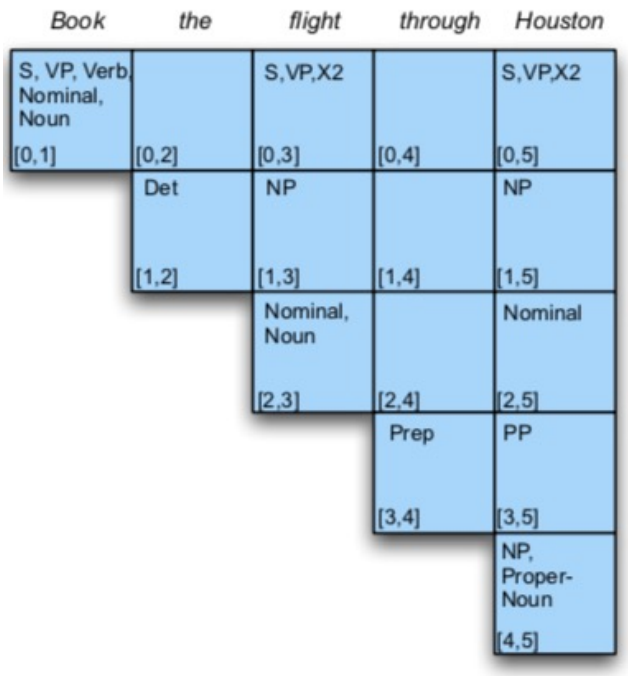
構文解析と自然言語処理

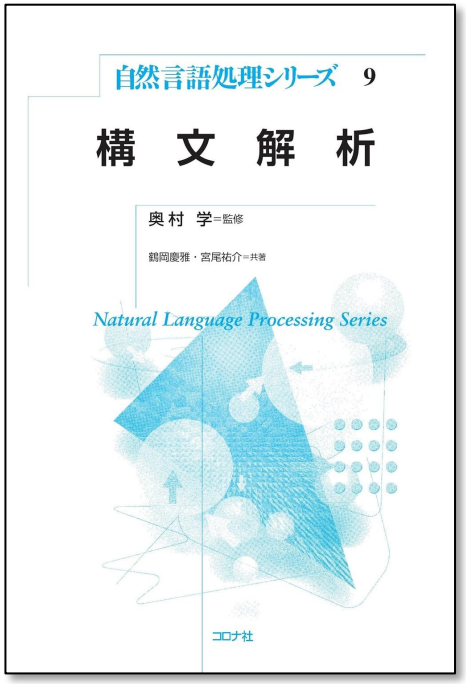
- 構文解析は、自然言語処理の中心的テーマの一つ
- 構文解析：文が与えられたとき、その木構造を
  可能な無数の組み合わせの中から自動的に求める



構文解析の難しさ

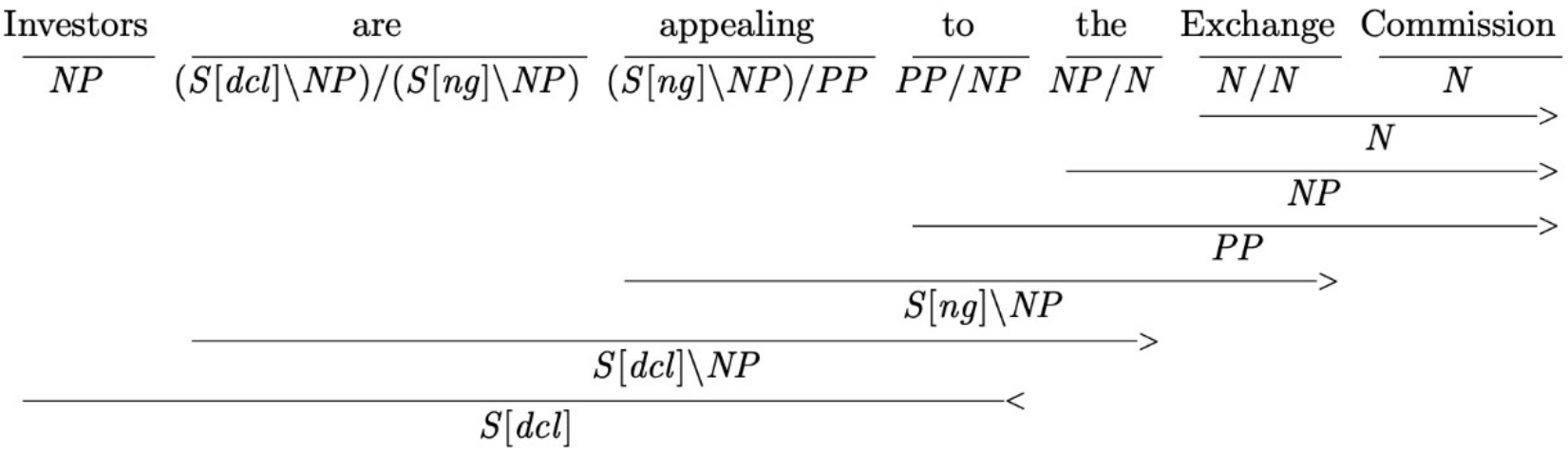
- 動的計画法を使っても、文長Nについて$O(N^3)$の計算時間 (CYKアルゴリズム)
- 与えられた文に対して可能な膨大な木構造の中から、正解の一つだけを選ぶタスク





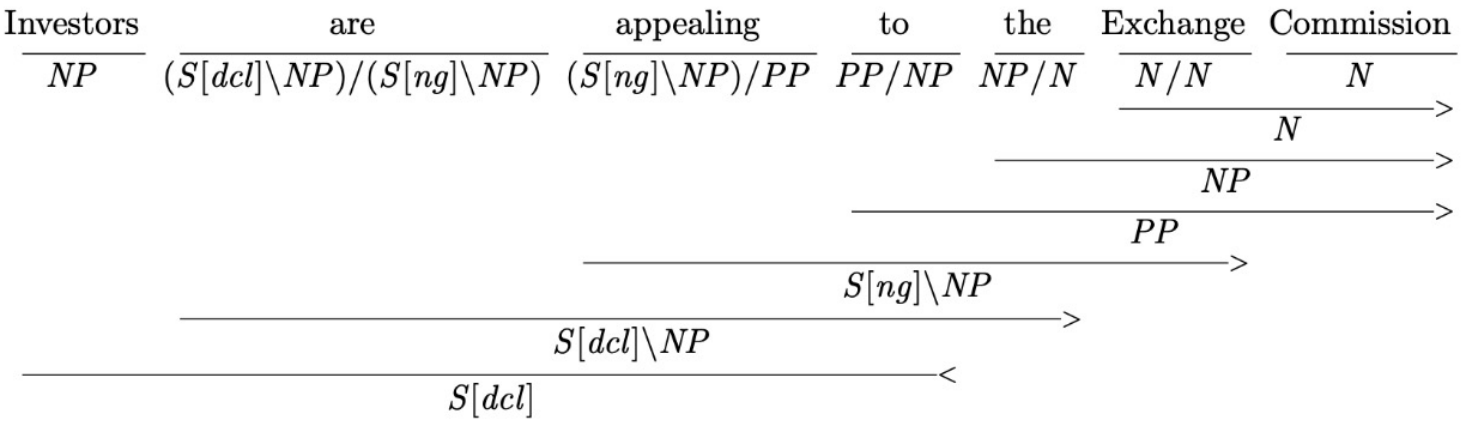
詳しくは
こうした
教科書を
参照

組み合わせ範疇文法 (CCG)



- 語を構文木に沿って組み合わせることで、句構造に対応する表現を生成
- 最終的に、Sが得られれば解析終了
- 途中の句構造に、その下にある単語の情報が反映されている

CCGの問題点



- 単語につけられた (S＼N)/Nのようなタグが離散的
  → 埋め込みベクトル化したい
  - 上位の句構造のベクトルが、合成された単語のベクトルを反映できる
- すべての単語に、あらかじめ人手でこうしたタグを付与する必要がある (教師なし学習; 今回は範囲外)

Holographic CCG

- 基本的なアイデア: CCGをベクトル合成と組み合わせられないか？
- CVGのように合成毎にパラメータがあると、厖大なパラメータが必要で過学習してしまう

- 何か、2つのベクトルを合成する見通しのよい数学的な方法はないか？
  → Holographic composition (Plate 1995)

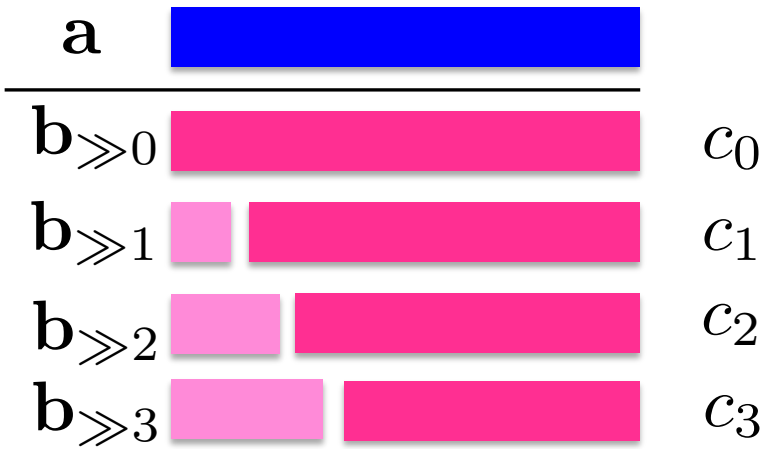
Holographic composition

- 知識グラフの埋め込み手法として、Nickel+(AAAI 2016)で提案 (最初の提案はPlate(1995))
- Circular correlation (巡回相関)

$$\mathbf{c} = \mathbf{a} \star \mathbf{b}$$

- 定義:

$$[c]_k = \sum_{i=0}^{d-1} a_i b_{(k+i) \bmod d}$$



Holographic composition (3)

- 性質:
  - Non-commutative: 方向を保存

$$\mathbf{a} \star \mathbf{b} \neq \mathbf{b} \star \mathbf{a}$$

  - Non-associative: 構造を保存

$$(\mathbf{a} \star \mathbf{b}) \star \mathbf{c} \neq \mathbf{a} \star (\mathbf{b} \star \mathbf{c})$$

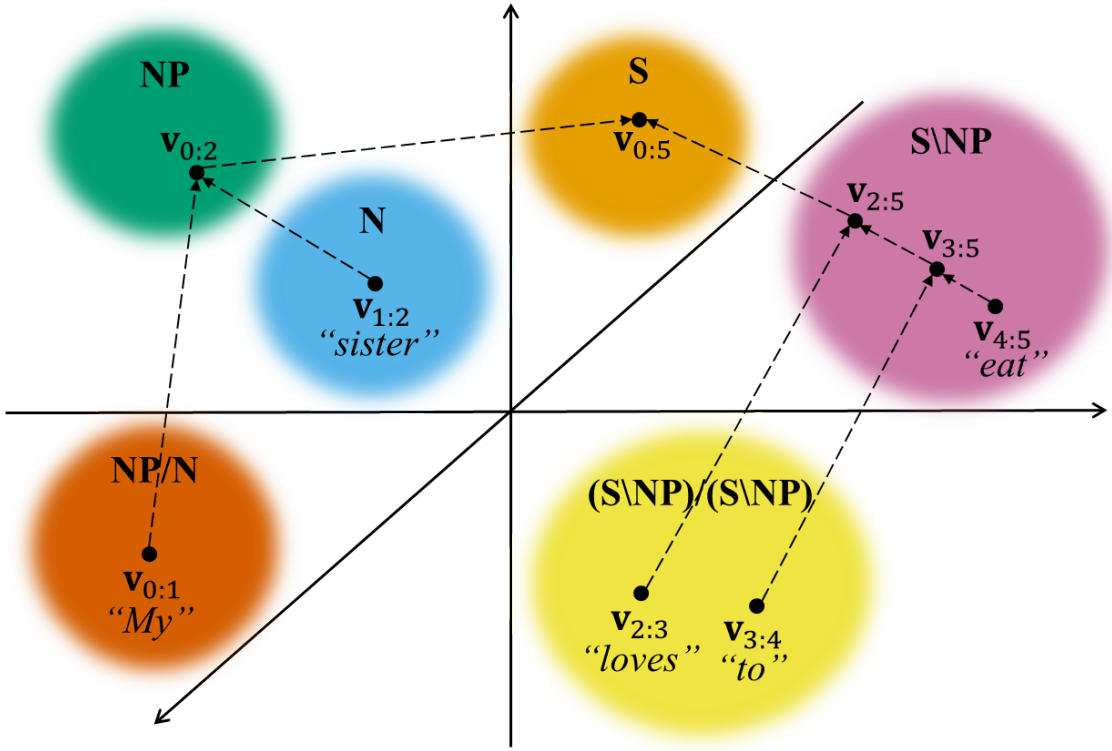
  - 構文構造を保存: たとえば “loves to eat”のベクトルは

$$\mathbf{v}_{\text{“loves to eat”}} = \mathbf{v}_{\text{loves}} \star \mathbf{v}_{\text{“to eat”}} = \mathbf{v}_{\text{loves}} \star (\mathbf{v}_{\text{to}} \star \mathbf{v}_{\text{eat}})$$

    - これは $(\mathbf{v}_{\text{loves}} \star \mathbf{v}_{\text{to}}) \star \mathbf{v}_{\text{eat}}$ とは異なる

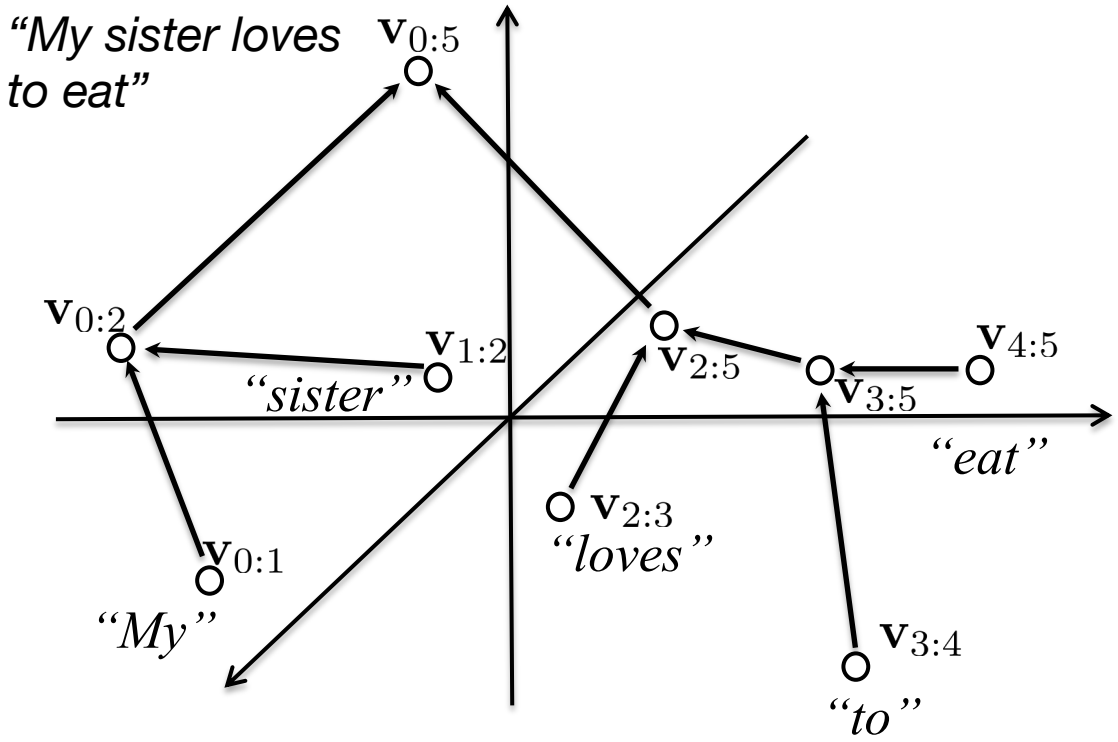
Holographic CCG

- Hol-CCGのイメージ



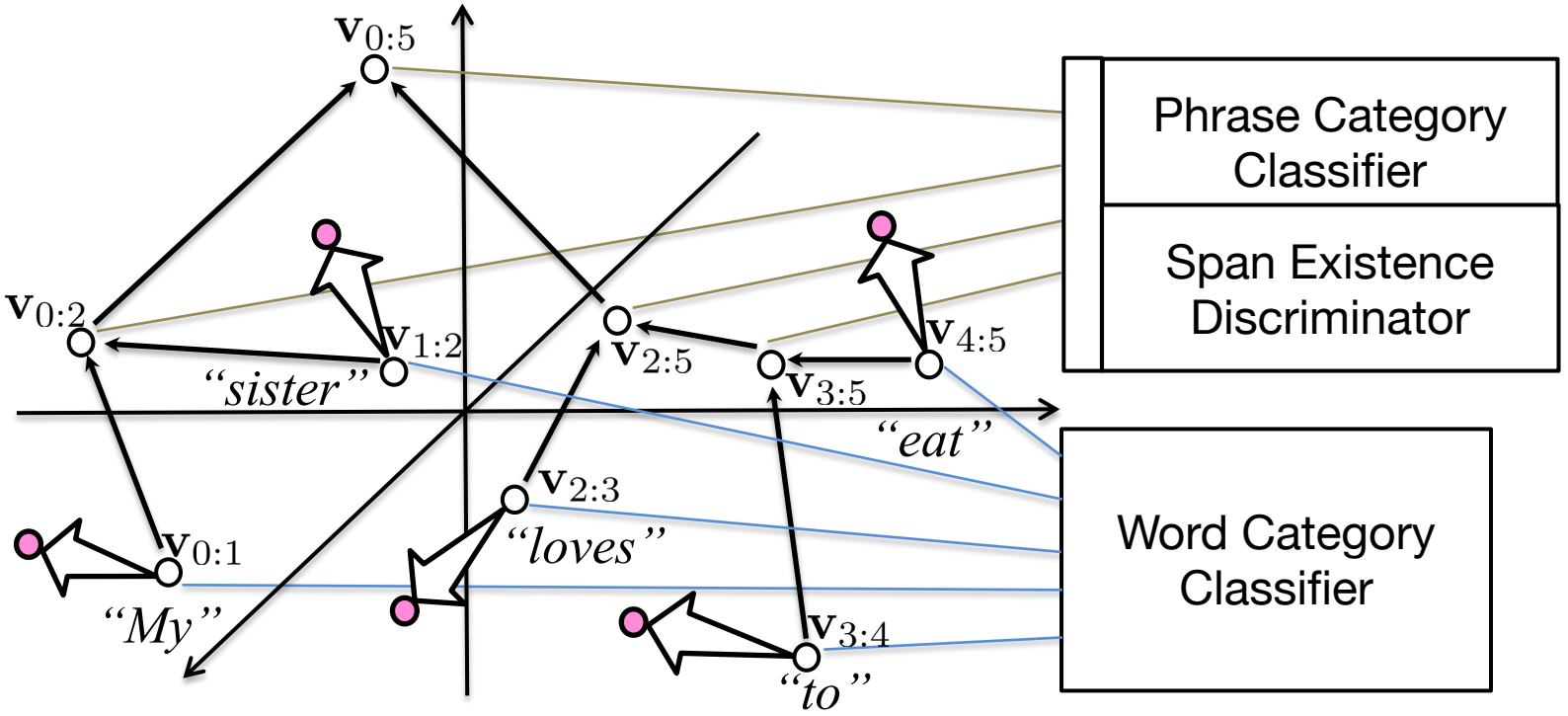
  - 単語ベクトル自体を最適化して、合成して最終的に文Sのベクトルが得られるようにする

実際の学習の様子 “My sister loves to eat”



- 単語のベクトルを構文構造に従って合成して、句構造のベクトルを再帰的に計算する
- 最終的に、文全体を表すベクトルが得られる

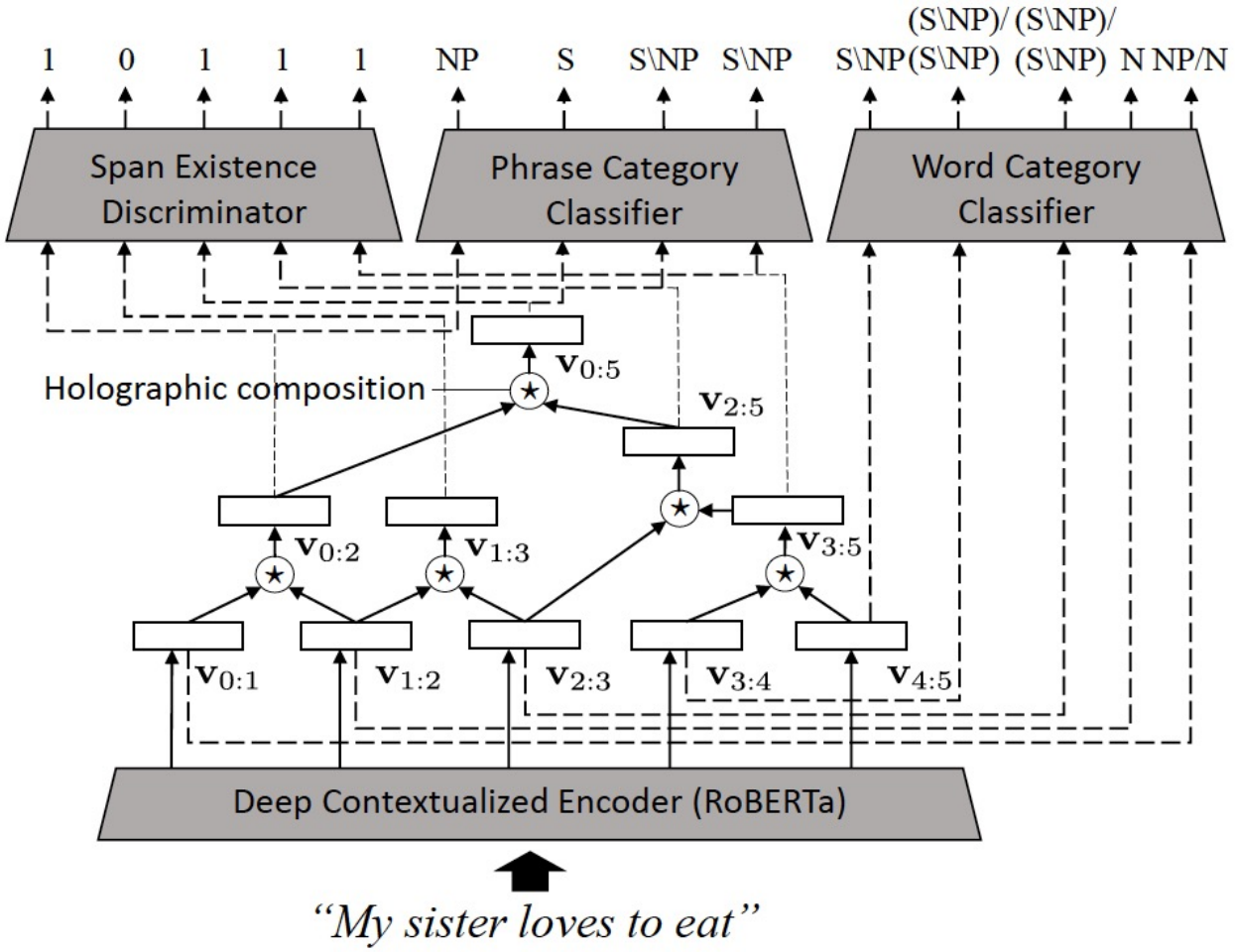
実際の学習の様子 “My sister loves to eat”



- 合成した結果ベクトルがそれぞれの句に正しく分類されるよう、逆伝播で最初の単語ベクトルを最適化
- 最終的には、文Sとなればよい

Transformer(深層学習)から始める場合

- 各単語の入力埋め込みとしてRoBERTaを使用



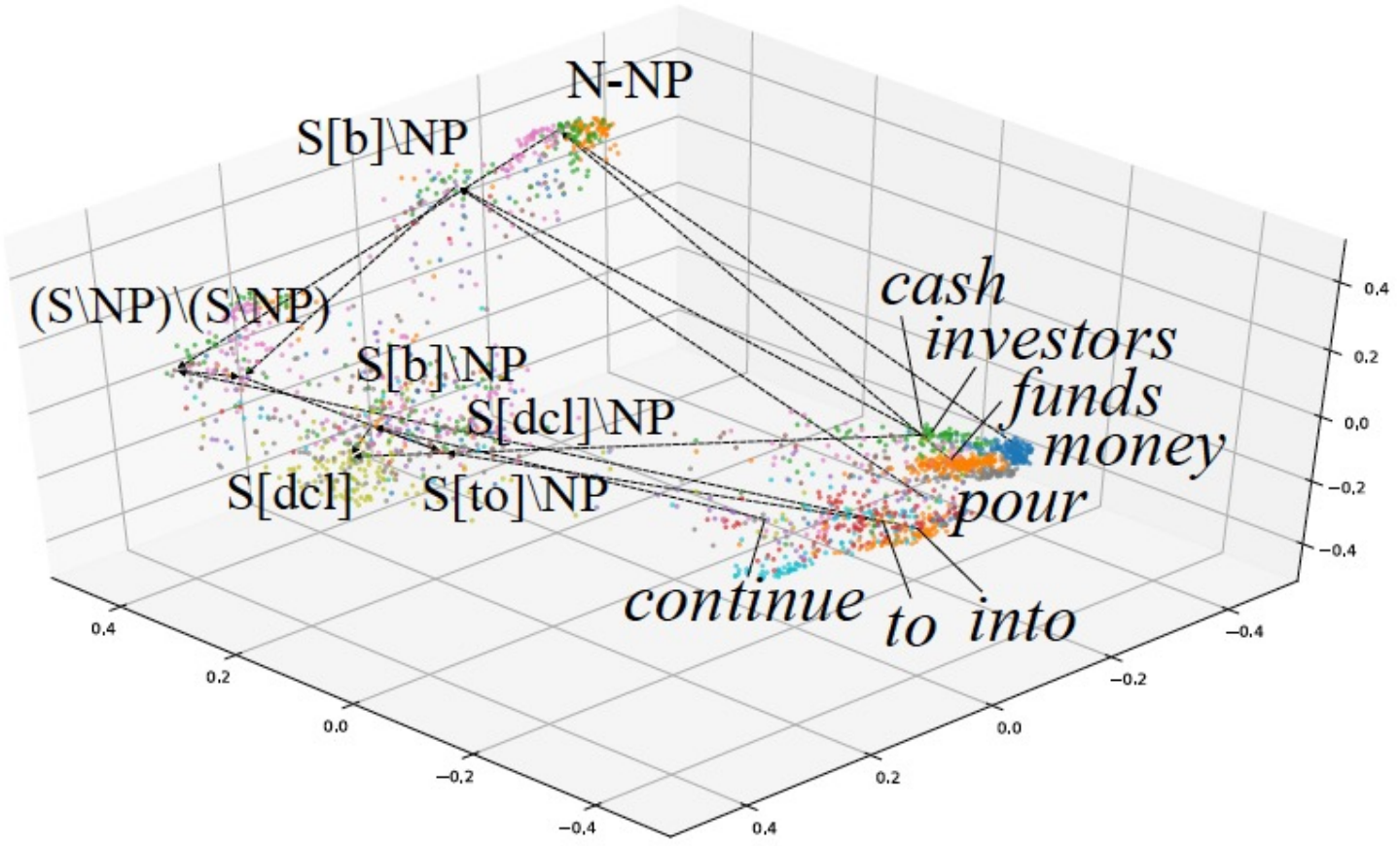
実験結果

- CCGBankというWall Street Journalのデータで評価
- 現在、ヒューリスティックな方法を抜いて世界最高性能

Model	Super-Tagger	Parser	Acc	LF
Lewis et al. (2016)	LSTM	A*	94.7	88.1
Vaswani et al. (2016)	LSTM	C&C	94.5	88.32
Yoshikawa et al. (2017)	LSTM	A* (LSTM)	–	88.8
Stanojević and Steedman (2020)	LSTM	Shift-Reduce (LSTM)	–	90.6
Tian et al. (2020)	Attentive-GCNN	EasyCCG	96.25	90.58
Bhargava and Penn (2020)	LSTM decoder	C&C	96.00	90.9
Liu et al. (2021)	Category Generator	C&C	96.05	90.87
Prange et al. (2021)	Tree-Structured decoder	C&C	96.22	90.91
Kogkalidis and Moortgat (2022)	Heterogeneous Dynamic Convolutions	–	96.29	–
Clark (2021)	Tian et al. (2020)	C&C	–	91.9
		Span-based	–	**92.9**
Ours ($\mathcal{L}_w + \mathcal{L}_p + \mathcal{L}_s$, Real)	Holographic	C&C	**96.60**	**92.12**
		Span-based	–	92.67

Table 2: Comparison of the proposed model and existing methods; best results are shown in bold.

実際のベクトルの合成の様子



- 768次元のベクトル空間をPCAで3次元に可視化

文法に即した穴埋め

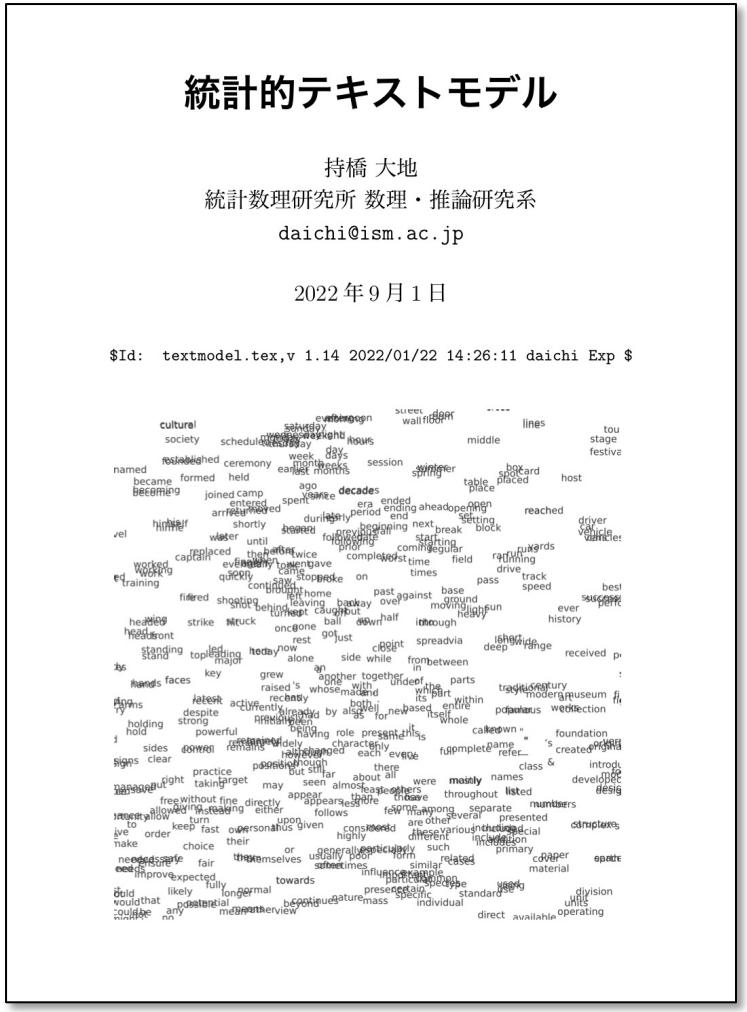
- 「句」のベクトルから、ランダムに単語に戻すことが可能 → 文法に準拠した穴埋めが可能に

ID	Sentence	Replacement Candidate	Sim.	NPMI
1	*Mr. Vinken is chairman of Elsevier N.V. , the Dutch publishing group .*	*Mr. Baris*	1.00	0.19
		Dr. Novello	1.00	0.10
		Ms. Ensrud	1.00	0.11
2	*When Scoring High first came out in 1979 , it was a publication of Random House .*	*turned up*	0.94	0.27
		sold out	0.91	0.29
		sells out	0.90	0.24
3	*In early trading in Hong Kong Thursday , gold was quoted at $ 374.19 an ounce .*	*for $ 25.50 a share*	0.94	0.33
		for $ 60 a bottle	0.94	0.29
		at $ 51.25 a share	0.93	0.34
4	*Judges are not getting what they deserve .*	*what she did*	0.96	0.28
		what they do	0.96	0.36
		what we do	0.89	0.35
5	*Despite recent declines in yields , investors continue to pour cash into money funds .*	*Despite the flap over transplants*	0.89	0.22
		In a victory for environmentalists	0.86	0.22
		On the issue of abortion	0.82	0.27
6	*Despite recent declines in yields , investors continue to pour cash into money funds .*	*to provide maintenance for other manufacturers*	0.83	0.27
		to share data via the telephone	0.79	0.21
		to cut costs throughout the organization	0.77	0.26

まとめ

- 言語の意味と文法に関する、講演者の最新の研究を一つずつ紹介
- 適切な統計モデルを立てることで、従来は自動的にはできないと思われていた分析が可能になる
- 計算機でモデル化することで、網羅的なモデル化が可能
  → どんな言葉や文に、どんな特徴があるのか
  をデータから抽出することができる
- 言語学と自然言語処理の境目は、どんどん縮まりつつある

統計的テキストモデル (教科書)



- 岩波書店より、2024年頃に発売予定
- 言語の統計モデル化の基礎について、丁寧に説明
- 類書と異なり、パッケージに頼らず、数学的背景もその場で説明しています
- http://chasen.org/~daiti-m/textmodel/ で原稿の一部を公開中