潜在表現を用いた言語変化の通時的分析

知能情報学専攻　黒橋・河原研究室　　村脇 有吾

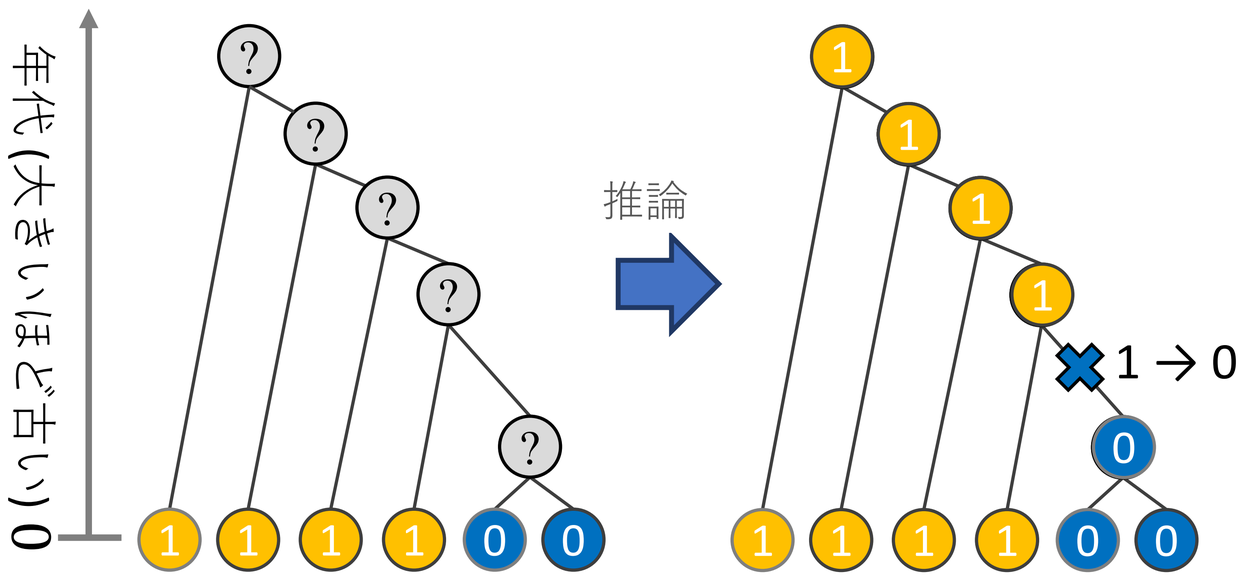
言語の歴史的変化の解明

- 言語は一般にどのように変化する傾向にあるのか?
- ある言語 (例えば日本語) は他の言語とどのような関係にあるのか?
- ある言語の過去/未来の状態はどのようなものか?

計算集約的統計手法の適用

- これまで数百年の歴史言語学の主流は**人手**による分析
- 計算集約的統計手法は人間が苦手な問題に有効
  - **数量**: 2つの言語の共通祖先 (祖語) の年代は?
  - **不確実性**: 過去を完全に復元できるだけの情報は残っていない
  - 解候補の**組合せ爆発**: 不確実性 = 解が絞り込めない

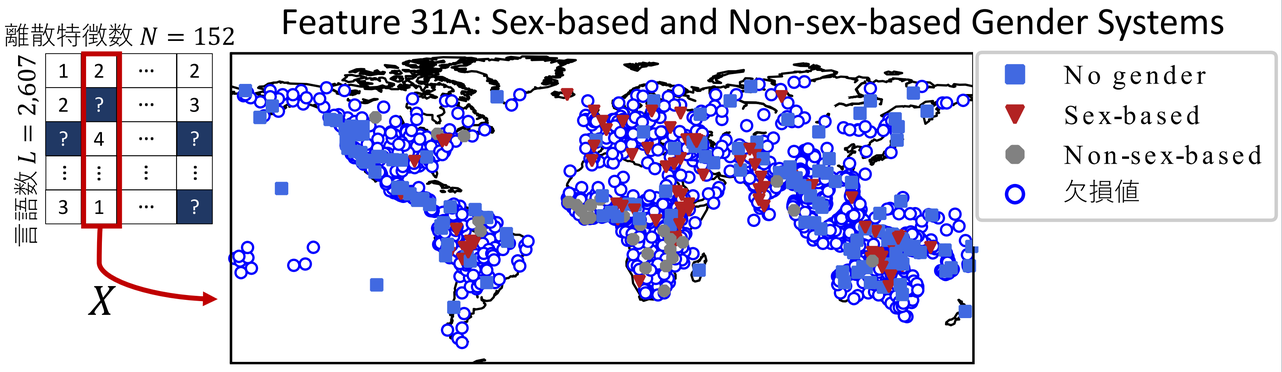
系統学的比較法



- 主に語彙的特徴を用いて得られた (年代つき) **系統樹**に他の特徴をフィッティングすると、祖語の状態や変化が起きた年代が推定できる
- 推定は不確実性が高く、統計的手法が不可欠
- 特徴の時間変化を**連続時間マルコフ連鎖**でモデル化

言語の構造的特徴 (言語類型論の特徴)

World Atlas of Language Structures (WALS)
Feature 31A: Sex-based and Non-sex-based Gender Systems



- WALS: 世界中の言語の構造的特徴 (語順、声調の有無等) を離散的にコード化したデータベース
- 歴史言語学が主な対象としてきた語彙的特徴と異なり、構造的特徴の**歴史的変化は充分に解明されていない**

問題: 特徴間の依存関係

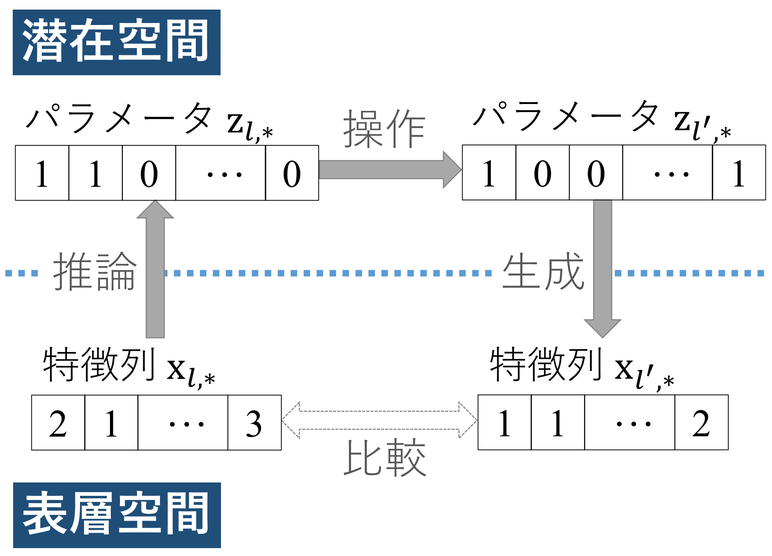
- 特徴間に依存関係があることが知られている
- 独立性を仮定して統計学的比較法を適用すると、**不自然な状態の祖語を再構**するおそれ

[Dryer, 2011]

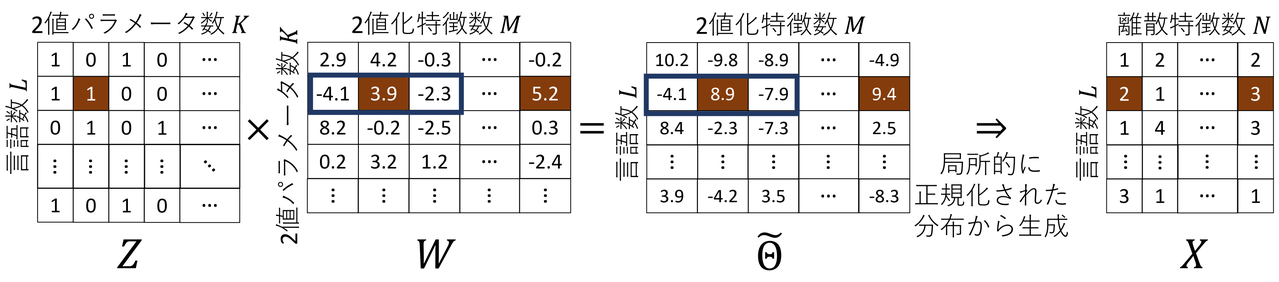
Order of object and verb \ Order of noun and relative clause	NRel	RelN
VO	✓	×
OV	✓	✓

提案手法: 潜在空間上での推論

- 複雑な依存関係のある表層特徴列を**(仮定により) 独立な**潜在パラメータ列に変換
- この潜在空間上で系統学的比較法を適用
- 得られた言語l'を表層特徴列に戻し、元の言語lと比較



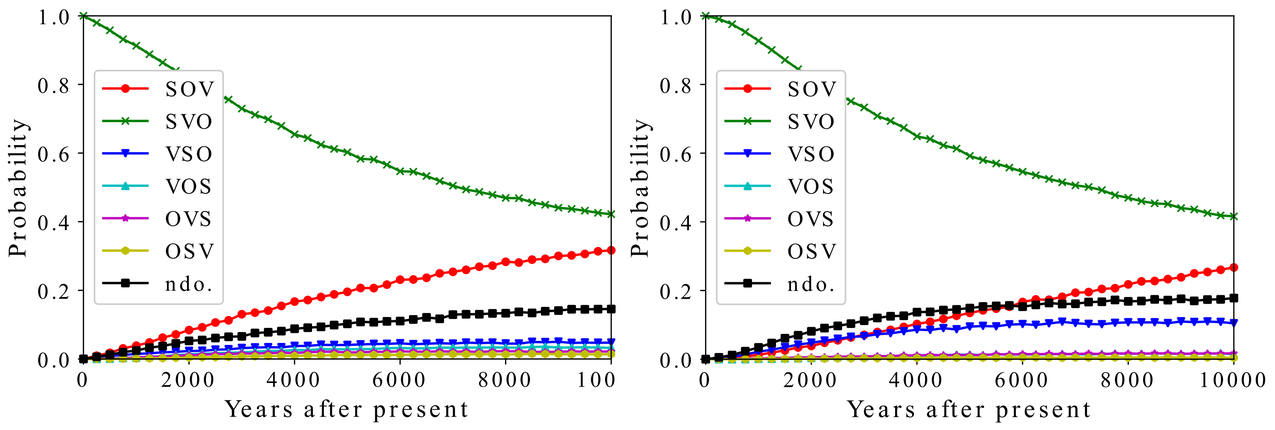
生成モデルに基づく表現学習



- 表層的特徴行列Xが潜在的2値パラメータ行列Zから確率的に生成されたと仮定
  - 重み行列Wが特徴間の依存関係を捉える
- X (欠損値あり) が与えられたとき、Z, Wを推論

推論結果

1. 世界の2,607言語の潜在パラメータ列を推論
2. 309語族 (うち154個は系統不明の孤立語) の系統樹を用いて各パラメータの確率的時間変化を推論
3. 得られたモデルを解釈するために、潜在空間上で諸言語の未来の状態をシミュレート



テトゥン語 (オーストロネシア語族)　コンゴ語 (大西洋・コンゴ諸語)

- 主語 (S)、目的語 (O)、動詞 (V) の基本語順に着目
- 同じSVO言語であっても、テトゥン語とコンゴ語では語順の安定性や変化の様態が異なる
  - 孤立語的なテトゥン語は直接SOVに変化しやすい
  - 接頭辞を多用するコンゴ語はSOVへの変化に抵抗し、支配的語順なし (ndo.) やVSOへの変化が目立つ
- 今後は他の特徴の言語学的分析も進めたい

発表文献

- Yugo Murawaki. 2019. *Bayesian Learning of Latent Representations of Language Structures*. Computational Linguistics. (to appear).
- Yugo Murawaki. 2018. *Analyzing Correlated Evolution of Multiple Features Using Latent Representations*. Proc. of EMNLP.

ソースコード: https://github.com/murawaki/lattyp