S1-P10

不均衡最適輸送を用いた意味変化検出

岸野 稜[1] 山際 宏明[1] 永田 亮[2,4] 横井 祥[3,4] 下平 英寿[1,4]
[1] 京大 [2] 甲南大 [3] 東北大 [4] 理研

背景

- 単語の通時的な意味変化検出は言語学の分析に有用
- ２時代間で意味が消失/出現した事例や対応の付かない事例を特定するのは難しい

概要

- 埋め込みの集合間に不均衡最適輸送を適用し，意味が消失/出現した用例を特定
- 意味の広狭判定のための指標を提案

1．新旧事例データセット

- Diachronic Word Usage Graphs for English [1] を使用
- 47 の対象単語がある (e.g., record, tip, afternoon ...)
- 新旧コーパス (1810-1860 vs 1960–2010) の文
  - 例えば，record という単語について，

  旧:...some ***record* (記録)** may be found in his hand-writing ...

  新: The ***record* (音楽)** labels' new service ...

2．不均衡最適輸送による意味変化のモデル化

- 各対象単語について，旧コーパスの事例 $s_1, \dots, s_m$ と新コーパスの事例 $t_1, \dots, t_n$ に関する文脈付き単語ベクトル $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m, \boldsymbol{v}_1, \dots \boldsymbol{v}_n$ を計算 (XL-LEXEME [2] を利用)
- 通常の最適輸送は新旧コーパスの事例集合間の**完全な対応**をとるが，実際は時代間で意味の消失/出現がある
- 不均衡最適輸送は**対応の過不足**があってもよい：

$$\min_{\boldsymbol{T}\geq\boldsymbol{0}} \sum_{i,j} T_{ij}C_{ij} + \lambda\|\boldsymbol{T}\boldsymbol{1} - \boldsymbol{w}^s\|_2 + \lambda\|\boldsymbol{T}^\top\boldsymbol{1} - \boldsymbol{w}^t\|_2$$

消失　出現

ただし，$C_{ij} = 1 - \cos(\boldsymbol{u}_i, \boldsymbol{v}_j)$

- 事例 s_i の意味の消失率 $:= \left(w_i^s - \sum_j T_{ij}\right)/w_i^s$
- 事例 t_j の意味の出現率 $:= \left(w_j^t - \sum_i T_{ij}\right)/w_j^t$

事例を特定

3．意味の広狭を判定する

- 最適輸送距離 $\sum_{i,j} T_{ij}C_{ij}$ を広狭のスコアとする
- $\|\boldsymbol{T}\boldsymbol{1} - \boldsymbol{w}^s\| > \|\boldsymbol{T}^\top\boldsymbol{1} - \boldsymbol{w}^t\|$ のときスコアを負にする
- 実験では「エントロピー差」をスコアの真値とする
  - 対象単語には意味ラベルが存在

  record では [92, 0, 0, 1, 0] → [65, 16, 10, 0, 1] (要素は事例数)

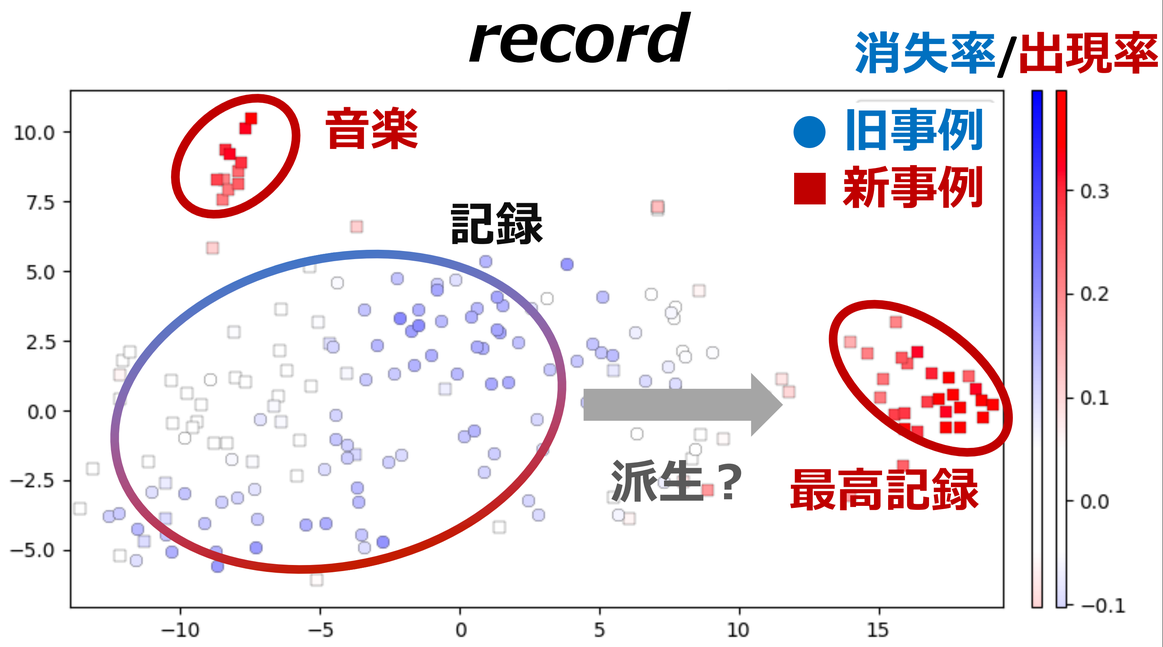
4．まとめ

- ✓ 意味の消失/出現をうまく可視化できそう
- ✓ 具体的な事例が容易に特定できそう
- ✗ 消失/出現の定量化の評価方法の定式化
- ✗ ハイパーパラメータへの依存性
- ✗ 不均衡のペナルティ項の検討：L^2 vs L^1

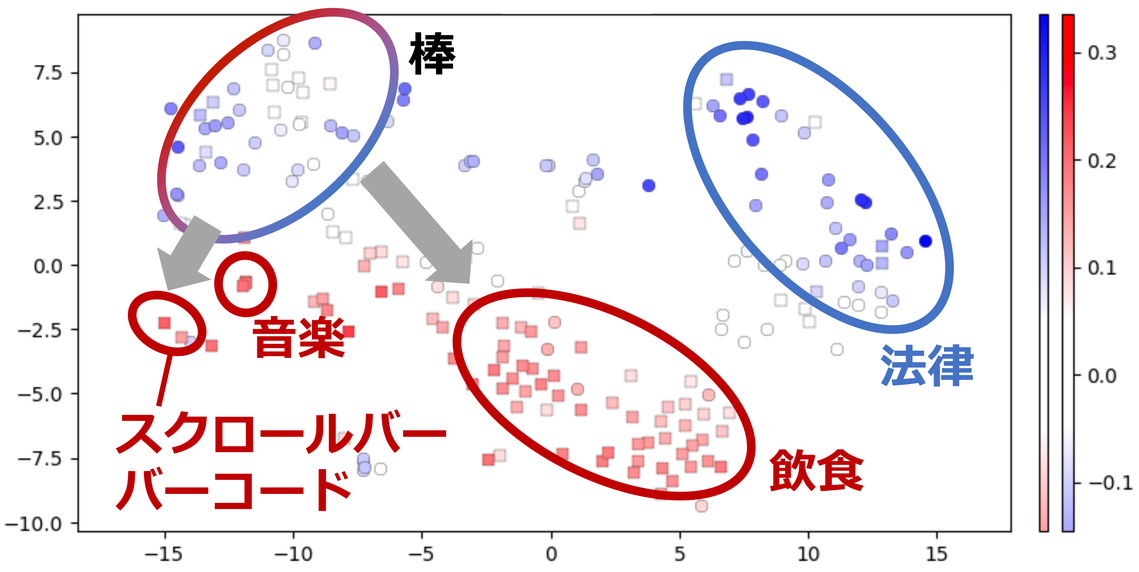
意味が消失/出現した事例の可視化

（XL-LEXEME による埋め込みに t-SNE を適用）

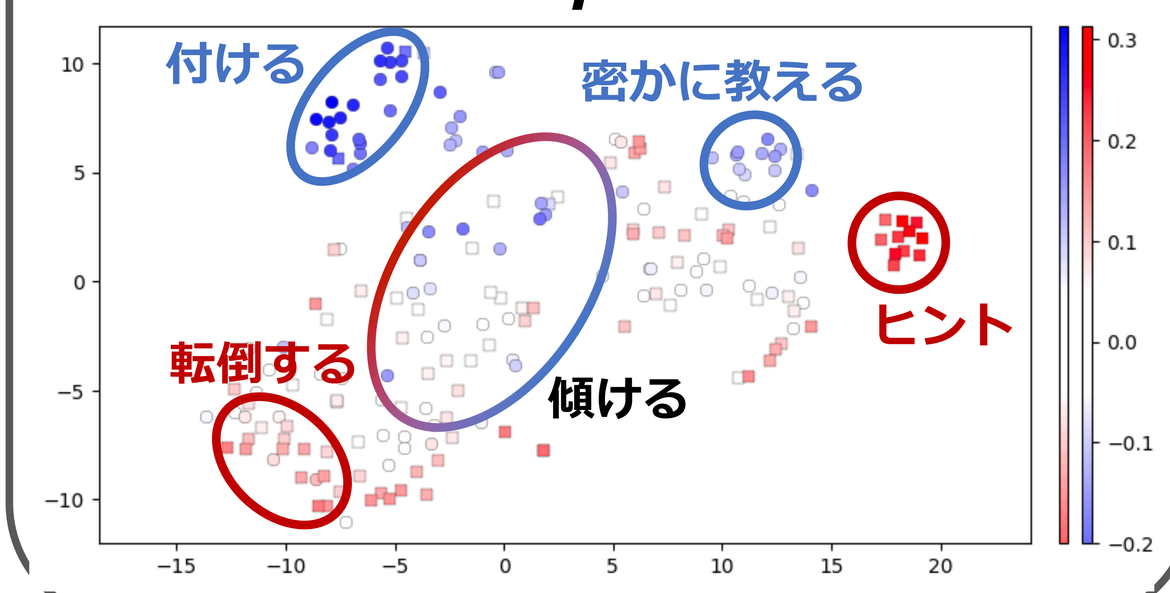
record



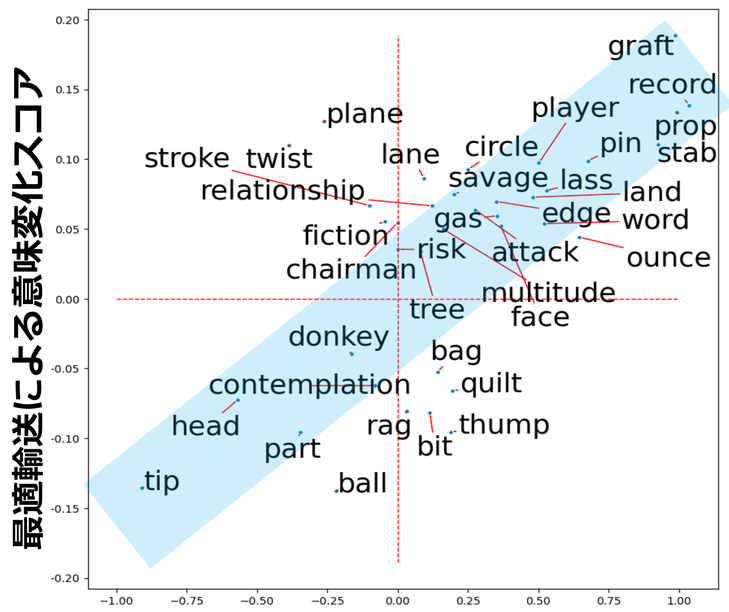
bar



tip



意味の広狭判定



提案手法の相関係数 = 0.64

既存手法 (Nagata+ [3]) の相関係数 = 0.57

[1] Schlechtweg, Dominik, et al. "DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

[2] Cassotti, Pierluigi, et al. "XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE." *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2023.

[3] Nagata, Ryo, et al. "Variance matters: Detecting semantic differences without corpus/word alignment." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023.