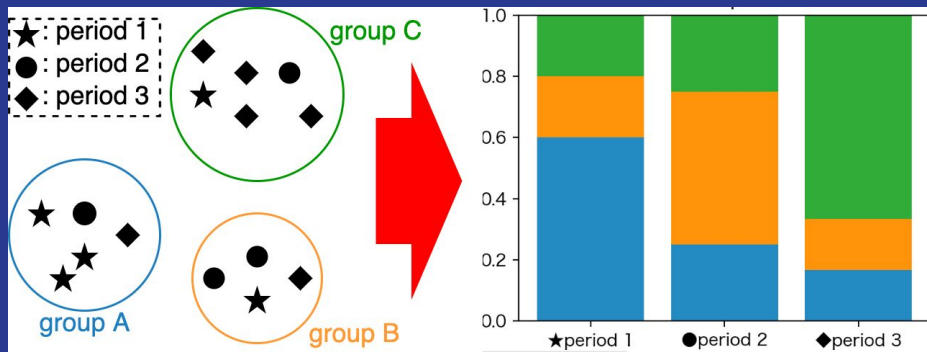


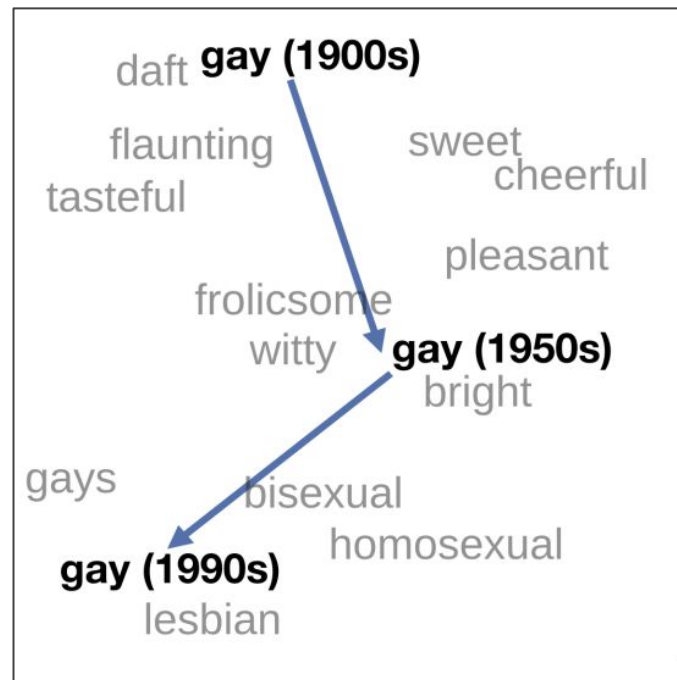
BERT を使用した 日本語の単語の 通時的な意味変化の分析

小林 千真、相田 太一、小町 守
東京都立大学



語義変化

- 時間経過によって、意味が変化する単語がある
- **代表的な検出方法:**
異なる時期間の単語分散表現を比較



Hamilton et al. (2016) Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1489–1501

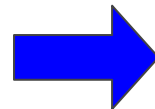
単語分散表現を使うことで意味変化の検出が可能

異なる時期の単語タイプを比較

- 意味変化を検出するには十分
- **意味変化の特徴の解釈・分析が困難**

1900s

1. The city's **gay** and lesbian people.
 2. We had a **gay** old time down at the dance hall.
- ...

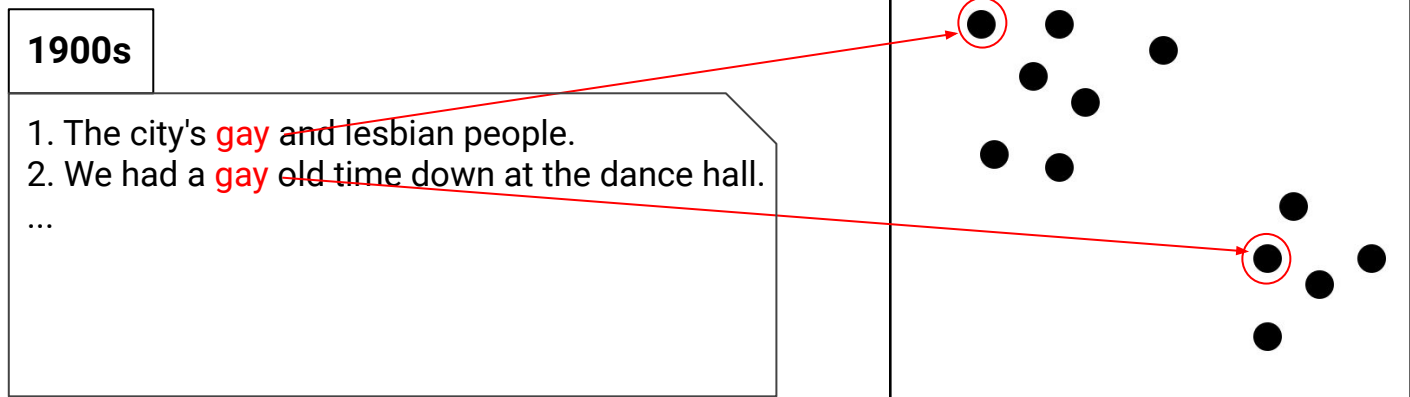


集約！

gay = [0.94, 0.52, ...]

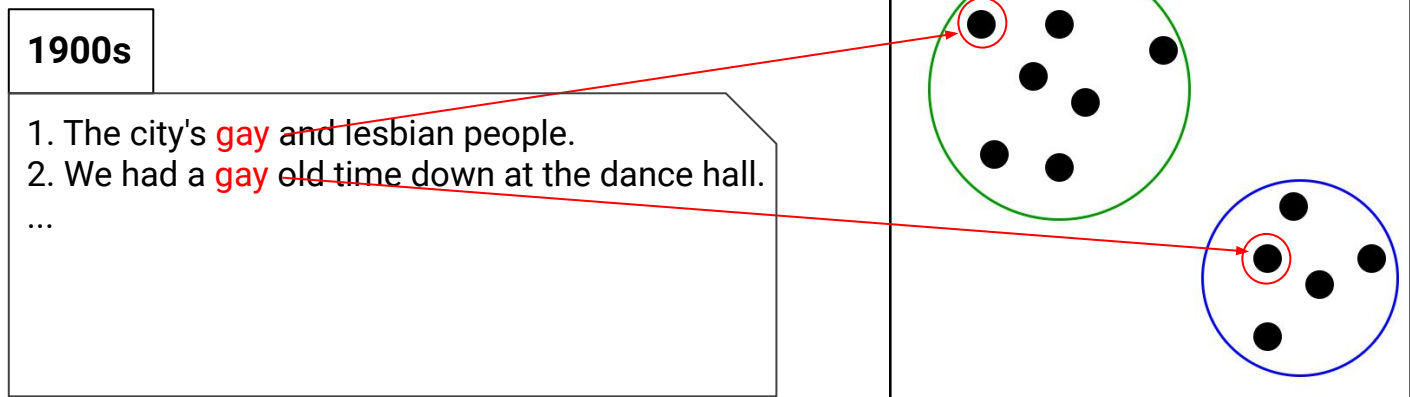
BERT によって解釈可能に！

- 単語トークンごとに単語分散表現を獲得する手法 (e.g. **BERT**, ELMo)
 - **語義変化の特徴を解釈・分析できる**



BERT によって解釈可能に！

- 単語トークンごとに単語分散表現を獲得する手法 (e.g. **BERT**, ELMo)
 - **語義変化の特徴を解釈・分析できる**

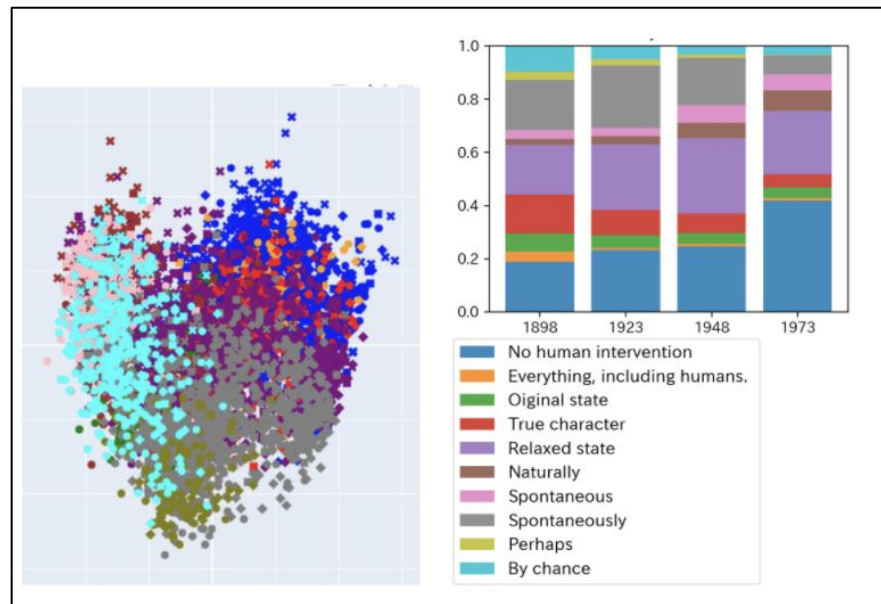


本研究の貢献

2つの BERT ベースの手法を比較

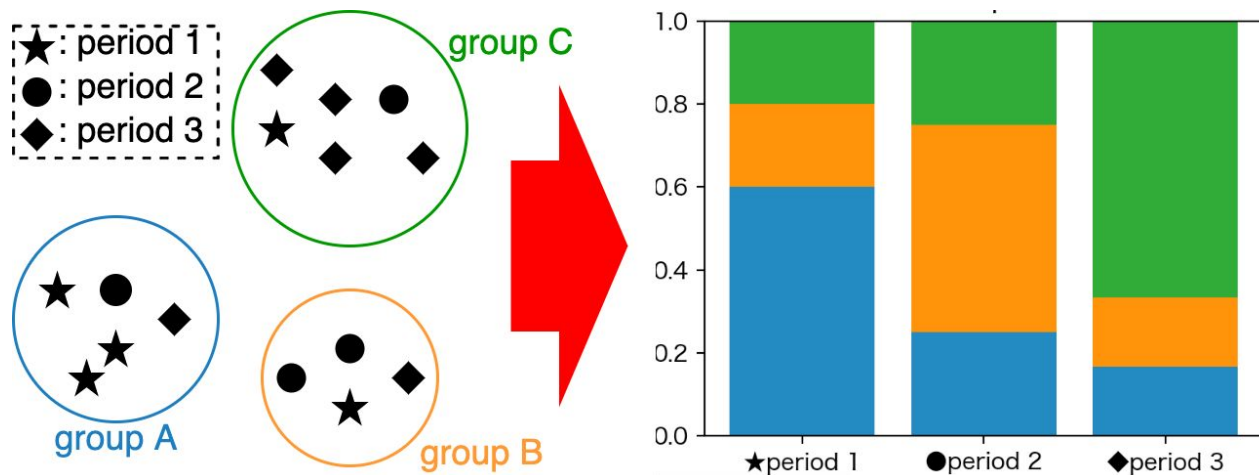
		クラスタリングベースの手法	
		Good	Bad
辞書ベースの手法	Good	16 単語 (障害、柔軟、要領、...)	5 単語 (教養、貴族、要領、....)
	Bad	13 単語 (遊撃、風俗、了解、...)	8 単語 (自然、女性、モデル、...)

語義変化を分析



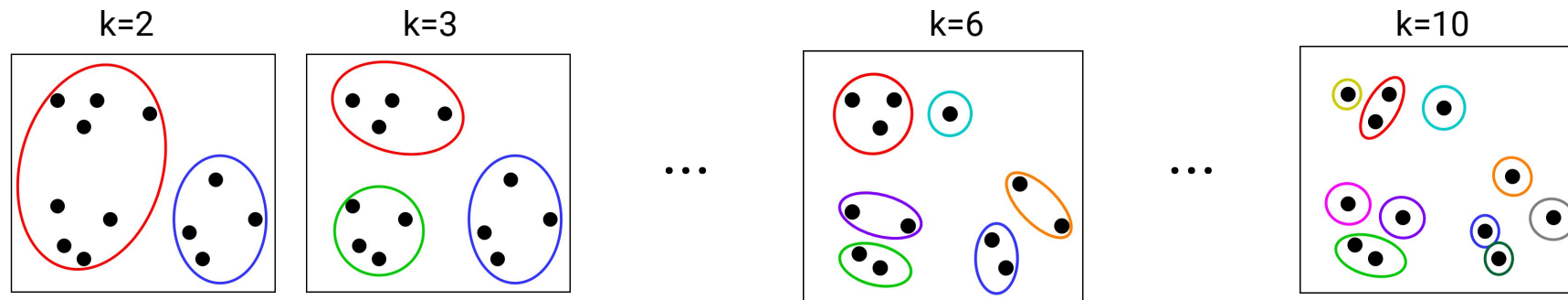
手法

手法



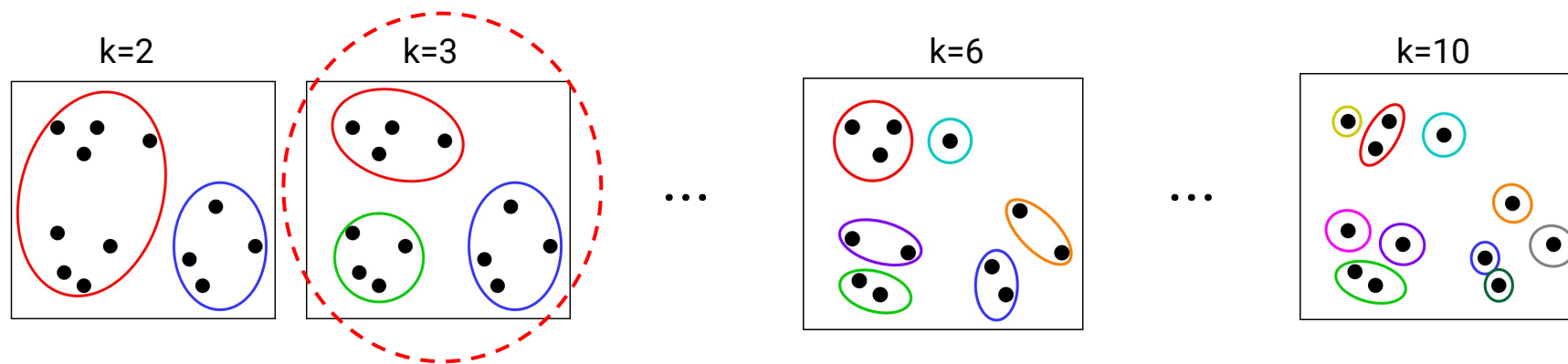
1. 意味ごとに単語ベクトルをグルーピング
(単語ベクトルは BERT により獲得)
2. 時期ごとに各語義の使用比率を計算

クラスタリングベースの手法



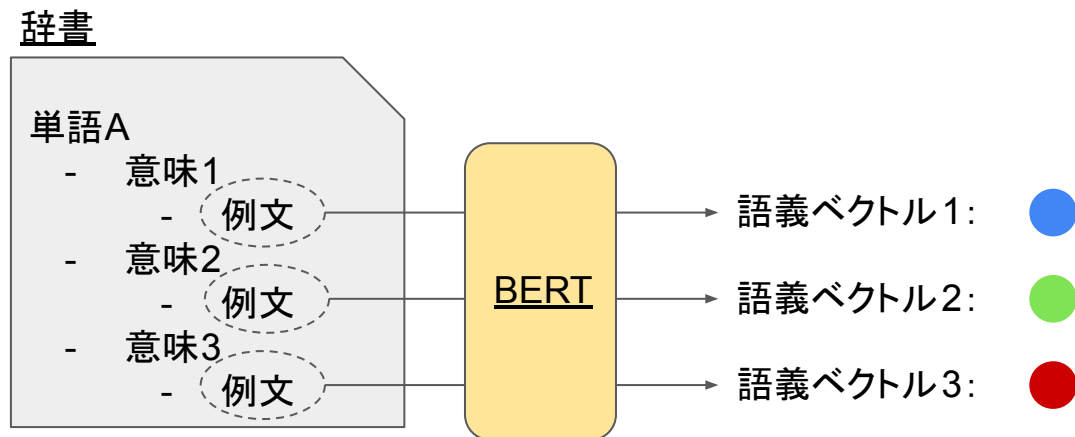
- クラスタリング : k-means 法
- クラスタ数 k : 2~10
- クラスタ数の決定指標: シルエットスコア

クラスタリングベースの手法

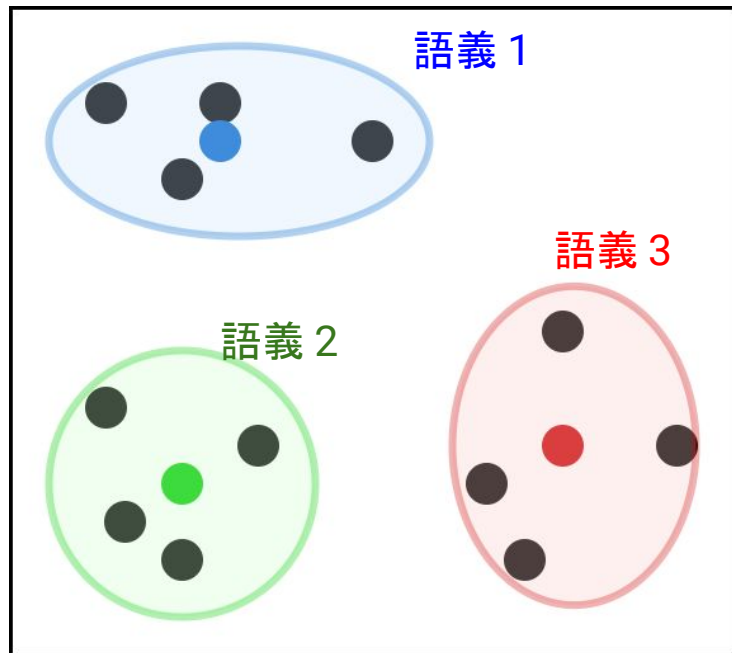


- クラスタリング : k-means 法
- クラスタ数 k : 2~10
- クラスタ数の決定指標: シルエットスコア

辞書ベースの手法



辞書ベースの手法



最近傍の語義に割り当てることで
グルーピング

語義ベクトル1: ●

語義ベクトル2: ●

語義ベクトル3: ●

実験設定

- コーパス

- 近代雑誌コーパス
- 期間: 1898~1997の100年間

(<https://ccd.ninjal.ac.jp/chj/index.html>
<https://ccd.ninjal.ac.jp/cmj/woman-mag/index.html>)

- 辞書

- 日本国語大辞典(電子版)

(<https://japanknowledge.com/en/contents/nikkoku/>)

- 対象単語

- 間淵ら(2021) が作成したリストから語義変化されていることが広く知られている単語を42単語使用

- 事前学習済み BERT モデル

- 日本語 Wikipedia で学習されている東北大 BERT を使用

(<https://github.com/cl-tohoku/bert-japanese>)

- 評価

- 著者による定性評価

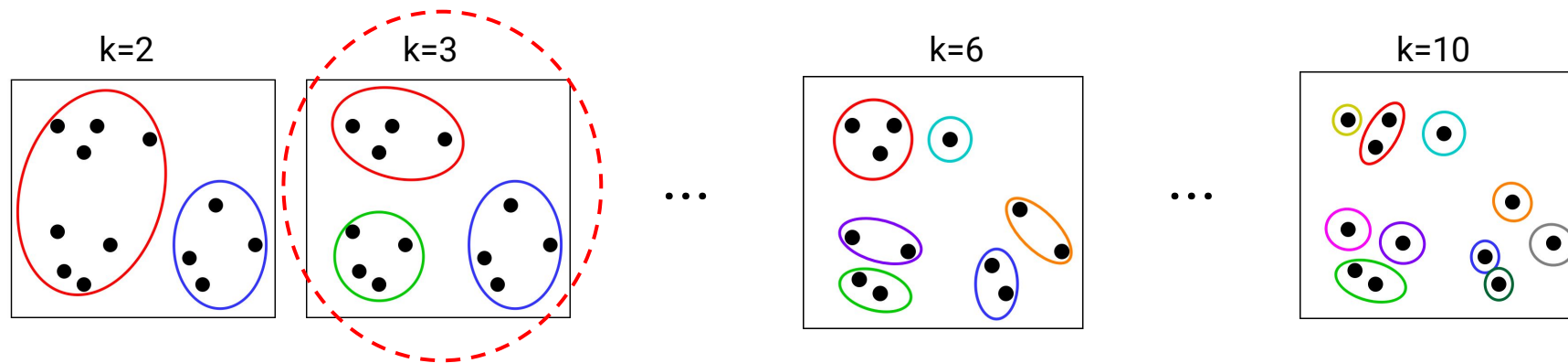
実験結果

実験結果

		クラスタリングベースの手法	
		Good	Bad
密書ベースの手法	Good	16 単語 (普段、障害、柔軟、結構、要領、ケース、免許、優勝、明細、非常、全然、精々、渋滞、ポイント、管制、故障)	5 単語 (教養、貴族、要領、教授、心持ち)
	Bad	13 単語 (遊撃、風俗、カフェ、団塊、普通、尋常、愛人、ボタン、広告、了解、端末、住居、スーパー)	8 単語 (自然、女性、モデル、設備、婦人、情報、主婦、こだわり)

クラスタリングベースの手法

再掲



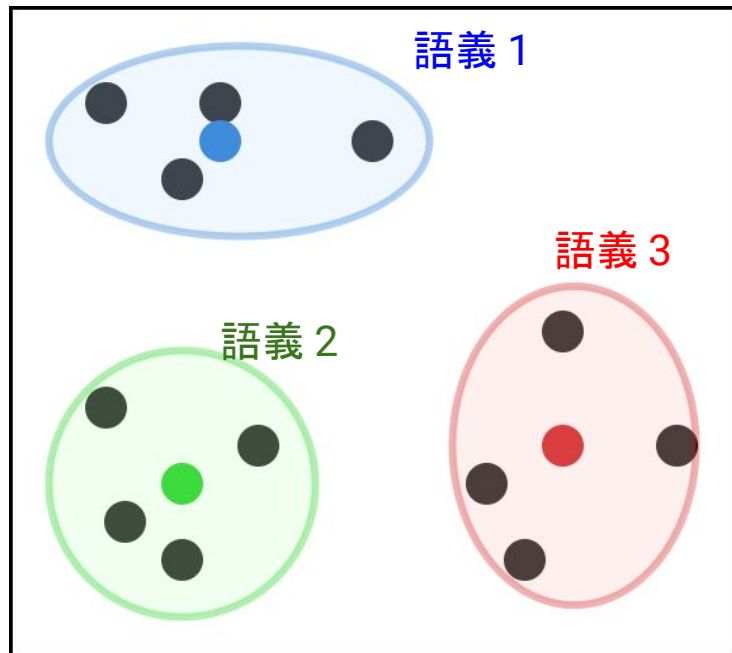
- クラスタリング : k-means 法
- クラスタ数 k : 2~10
- クラスタ数の決定指標: シルエットスコア

実験結果

		クラスタリングベースの手法	
		Good	Bad
密書ベースの手法	Good	16 単語 (普段、障害、柔軟、結構、要領、ケース、免許、優勝、明細、非常、全然、精々、渋滞、ポイント、管制、故障)	5 単語 (教養、貴族、要領、教授、心持ち)
	Bad	13 単語 (遊撃、風俗、カフェ、団塊、普通、尋常、愛人、ボタン、広告、了解、端末、住居、スーパー)	8 単語 (自然、女性、モデル、設備、婦人、情報、主婦、こだわり)

辞書ベースの手法

再掲



最近傍の語義に割り当てることで
グルーピング

語義ベクトル1: ●

語義ベクトル2: ●

語義ベクトル3: ●

実験結果

		クラスタリングベースの手法	
		Good	Bad
密書ベースの手法	Good	16 単語 (普段、障害、柔軟、結構、要領、ケース、免許、優勝、明細、非常、全然、精々、渋滞、ポイント、管制、故障)	5 単語 (教養、貴族、要領、教授、心持ち)
	Bad	13 単語 (遊撃、風俗、カフェ、団塊、普通、尋常、愛人、ボタン、広告、了解、端末、住居、スーパー)	8 単語 (自然、女性、モデル、設備、婦人、情報、主婦、こだわり)

クラスタリングベースで失敗したケース

		クラスタリングベースの手法	
		Good	Bad
辞書ベースの手法	Good	16 単語 (普段、障害、柔軟、結構、要領、ケース、免許、優勝、明細、非常、全然、精々、渋滞、ポイント、管制、故障)	5 単語 (教養、貴族、要領、教授、心持ち)
	Bad	13 単語 (遊撃、風俗、カフェ、団塊、普通、尋常、愛人、ボタン、広告、了解、端末、住居、スーパー)	8 単語 (自然、女性、モデル、設備、婦人、情報、主婦、こだわり)

変化前と変化後で
各語義の使用頻度に
大きな違いある単語

辞書ベースで失敗したケース

		クラスタリングベースの手法	
		Good	Bad
辞書ベースの手法	Good	16 単語 (普段、障害、柔軟、結構、要領、ケース、免許、優勝、明細、非常、全然、精々、渋滞、ポイント、管制、故障)	5 単語 (教養、貴族、要領、教授、心持ち)
	Bad	13 単語 (遊撃、風俗、カフェ、団塊、普通、尋常、愛人、ボタン、広告、了解、端末、住居、スーパー)	8 単語 (自然、女性、モデル、設備、婦人、情報、主婦、こだわり)

辞書の内容が
不十分だった単語
(語義の分割や例文の
充実度合い)

どちらの手法でもうまくいったケース

		クラスタリングベースの手法	
		Good	Bad
辞書ベースの手法	Good	16 単語 (普段、障害、柔軟、結構、要領、ケース、免許、優勝、明細、非常、全然、精々、渋滞、ポイント、管制、故障)	5 単語 (教養、貴族、要領、教授、心持ち)
	Bad	13 単語 (遊撃、風俗、カフェ、団塊、普通、尋常、愛人、ボタン、広告、了解、端末、住居、スーパー)	8 単語 (自然、女性、モデル、設備、婦人、情報、主婦、こだわり)

前に述べたどちらの問題もなかった単語

どちらの手法でも失敗したケース

		クラスタリングベースの手法	
		Good	Bad
辞書ベースの手法	Good	16 単語 (普段、障害、柔軟、結構、要領、ケース、免許、優勝、明細、非常、全然、精々、渋滞、ポイント、管制、故障)	5 単語 (教養、貴族、要領、教授、心持ち)
	Bad	13 単語 (遊撃、風俗、カフェ、団塊、普通、尋常、愛人、ボタン、広告、了解、端末、住居、スーパー)	8 単語 (自然、女性、モデル、設備、婦人、情報、主婦、こだわり)

前に述べた両方の問題があった単語

どちらの手法でも失敗したケース

		クラスタリングベースの手法	
		Good	Bad
く 雑	Bad	13 単語 (遊撃、風俗、 カフェ、団塊、普通、尋 常、愛人、ボタン、広 告、了解、端末、 住居、スーパー)	8 単語 (自然、女性、 モデル、設備、 婦人、情報、 主婦、こだわり)
		対象期間内のコーパスで 出現頻度が少なかった 単語	単語 (教養、貴族、 要領、教授、 お持ち)

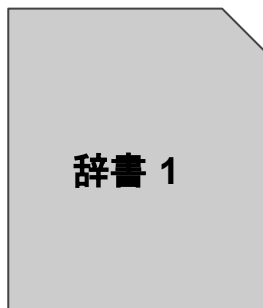
対象期間内のコーパスで
出現頻度が少なかった
単語

前に述べた両方の
問題があった単語

議論

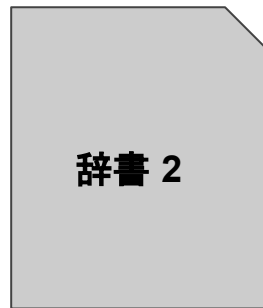
辞書ベースの手法は辞書の違いに敏感

異なる辞書を使用すると大きく結果が変わる



日本国語大辞典
(original)

VS.

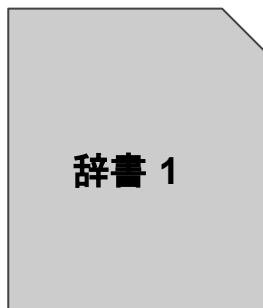


デジタル大辞泉 (new)

		日本国語大辞典	
		Good	Bad
デジタル大辞泉	Good	14 単語 要領、優勝、免許、明細、非常、適當、精々、障害、柔軟、故障、結構、教養、教授、心持ち	2 単語 自然、スーパー
	Bad	6 単語 普段、全然、ケース、ポイント、渋滞、管制、貴族	13 単語 風俗、尋常、愛人、モデル、普通、遊撃、ボタン、カフェ、団塊、普通、広告、情報、こだわり

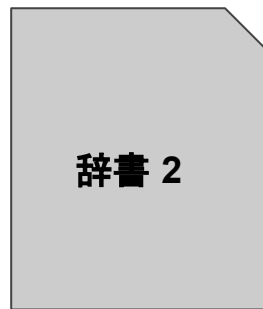
辞書ベースの手法は辞書の違いに敏感

異なる辞書を使用すると大きく結果が変わる



日本国語大辞典
(original)

VS.



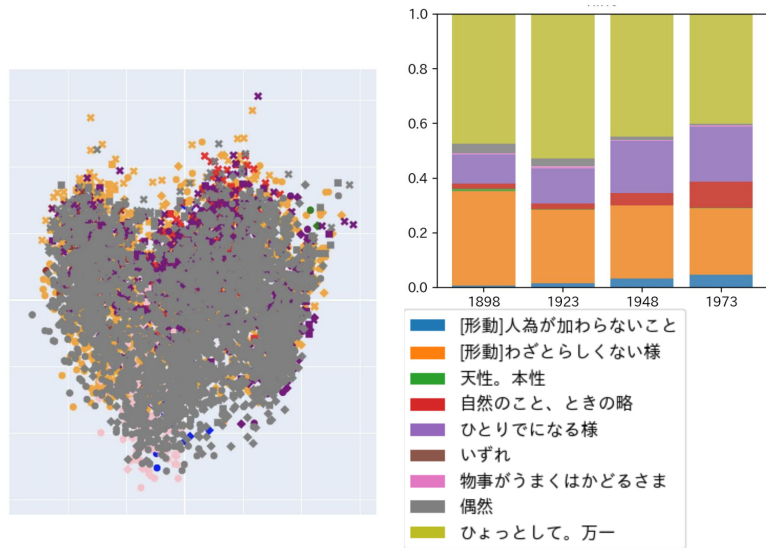
デジタル大辞泉 (new)

		日本国語大辞典	
		Good	Bad
デジタル大辞泉	Good	14 単語 要領、優勝、免許、明細、非常、適當、精々、障害、柔軟、故障、結構、教養、教授、心持ち	2 単語 自然、スーパー
	Bad	6 単語 普段、全然、ケース、ポイント、渋滞、管制、貴族	13 単語 風俗、尋常、愛人、モデル、普通、遊撃、ボタン、カフェ、団塊、普通、広告、情報、こだわり

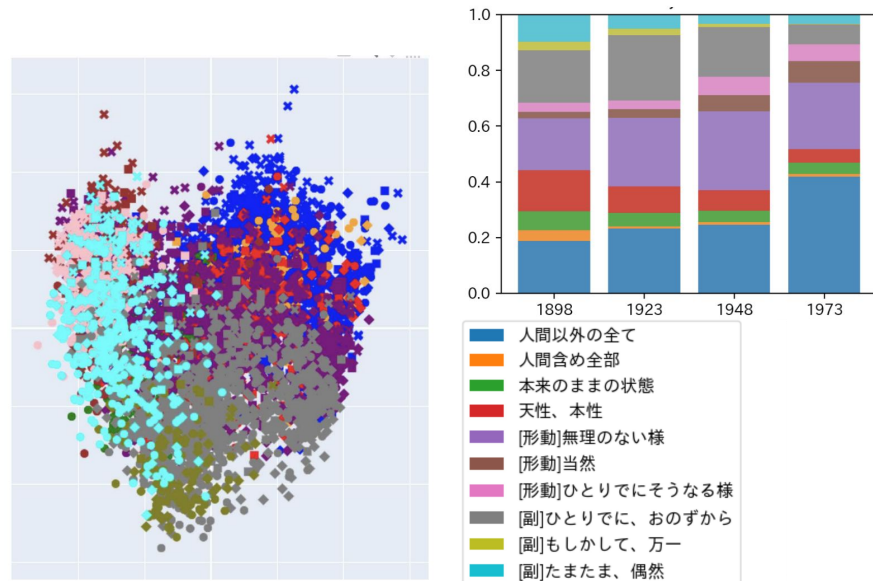
辞書の違いによる結果の違いの例(単語: 自然)

大辞泉では捉えているが、日本国語大辞典では捉えられていない

日本国語大辞典(original)

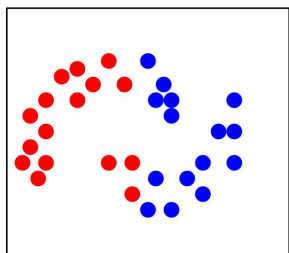


デジタル大辞泉(new)

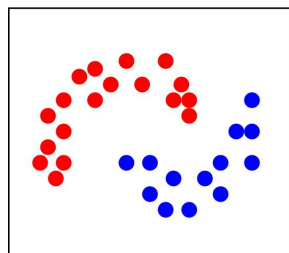


本タスクにおいて k-means 法はシンプルかつ有効

k-means(original) vs. DBSCAN(new)



密度ベースでない



密度ベース

		k-means	
		Good	Bad
DBSCAN	Good	11単語 スーパー、住居、渋滞、障害、非常、普通、免許、優勝、要領、了解、全然	0単語
	Bad	18単語 カフェ、愛人、端末、ケース、管制、故障、ポイント、ボタン、結構、柔軟、団塊、普段、風俗、明細、遊撃、尋常、精々	13単語 こだわり、教授、モデル、貴族、教養、自然、主婦、女性、設備、適当、婦人、情報、心持ち

クラスタリングベースで失敗したケース

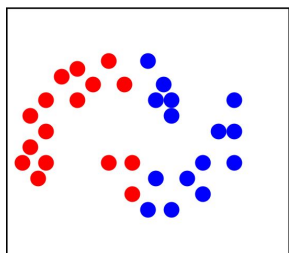
再掲

		クラスタリングベースの手法	
		Good	Bad
辞書ベースの手法	Good	16 単語 (普段、障害、柔軟、結構、要領、ケース、免許、優勝、明細、非常、全然、精々、渋滞、ポイント、管制、故障)	5 単語 (教養、貴族、要領、教授、心持ち)
	Bad	13 単語 (遊撃、風俗、カフェ、団塊、普通、尋常、愛人、ボタン、広告、了解、端末、住居、スーパー)	8 単語 (自然、女性、モデル、設備、婦人、情報、主婦、こだわり)

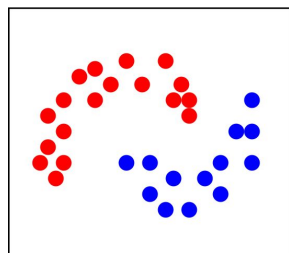
変化前と変化後で
各語義の使用頻度に
大きな違いある単語

本タスクにおいて k-means 法はシンプルかつ有効

k-means(original) vs. DBSCAN(new)



密度ベースでない

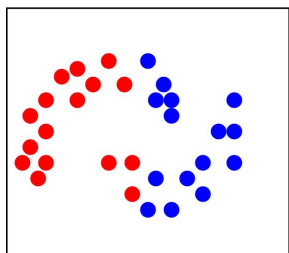


密度ベース

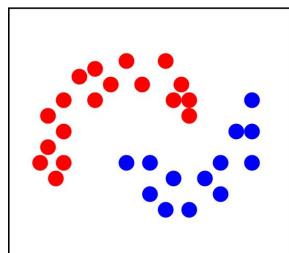
		k-means	
		Good	Bad
DBSCAN	Good	11単語 スーパー、住居、渋滞、障害、非常、普通、免許、優勝、要領、了解、全然	0単語
	Bad	18単語 カフェ、愛人、端末、ケース、管制、故障、ポイント、ボタン、結構、柔軟、団塊、普段、風俗、明細、遊撃、尋常、精々	13単語 こだわり、教授、モデル、貴族、教養、自然、主婦、女性、設備、適当、婦人、情報、心持ち

本タスクにおいて k-means 法はシンプルかつ有効

k-means(original) vs. DBSCAN(new)



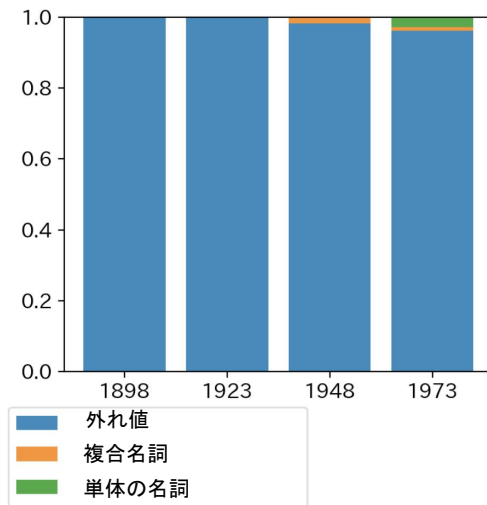
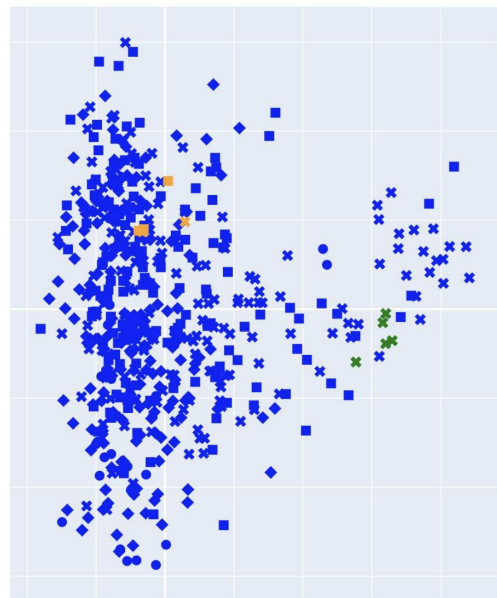
密度ベースでない



密度ベース

		k-means	
		Good	Bad
DBSCAN	Good	11単語 スーパー、住居、渋滞、障害、非常、普通、免許、優勝、要領、了解、全然	0単語
	Bad	18単語 カフェ、愛人、端末、ケース、管制、故障、ポイント、ボタン、結構、柔軟、団塊、普段、風俗、明細、遊撃、尋常、精々	13単語 こだわり、教授、モデル、貴族、教養、自然、主婦、女性、設備、適当、婦人、情報、心持ち

DBSCAN を使用した時の失敗例(単語:教養)

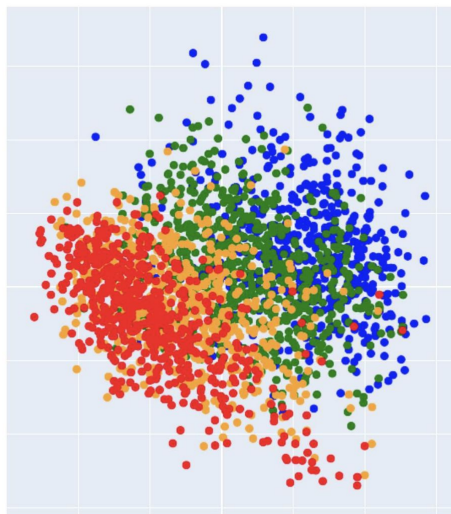


青い点は全て外れ値。。。

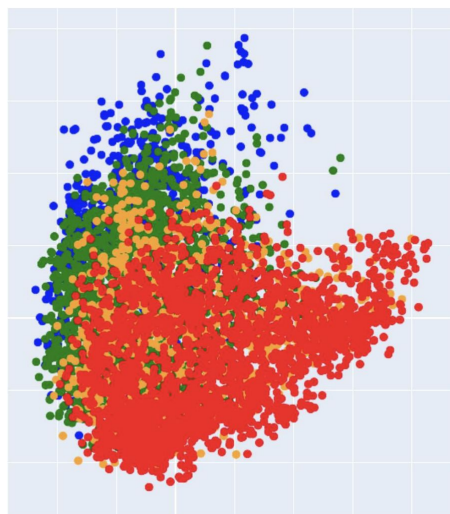
図.「教養」に対して DBSCAN を用いた

BERT から獲得した単語ベクトルの分布の遷移

時間経過による単語ベクトルの分布の遷移



「全然」



「教授」

青:1898~1922



緑:1923~1947



黄:1948~1972



赤:1973~1997

まとめ

- BERT ベースの2つの手法を日本語単語を対象として比較した
- 本タスクにおいて、クラスタリングベース、特に k-means を用いた手法は、辞書を用いた手法よりもシンプルかつ有効だった
- 辞書ベースの手法は辞書の特徴の違いに敏感だった
- クラスタリングベースの手法では DBSCAN よりも k-means を用いた手法が有効だった
- BERT による単語ベクトルの分布が時間経過により遷移する単語が見受けられた

今後の展望

- 単語ベクトルの改善
 - BERT はファインチューニングを前提としたモデル
 - 年代の影響を考慮
 - 単語ベクトル自体の性能の向上
- 対象単語の拡張
 - 活用がある単語
 - 語義変化しない単語
- 評価
 - 定量的な評価

補足資料

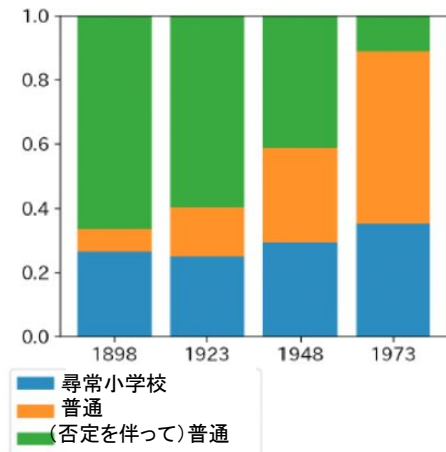
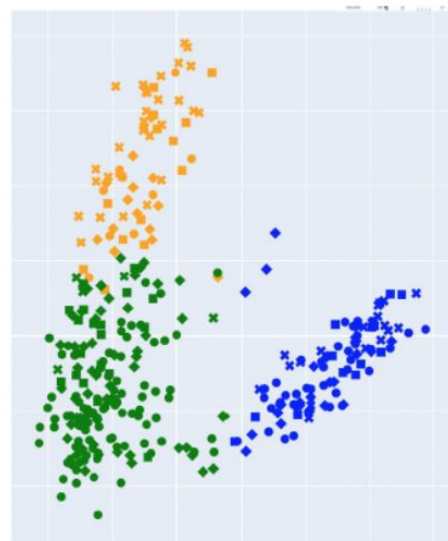
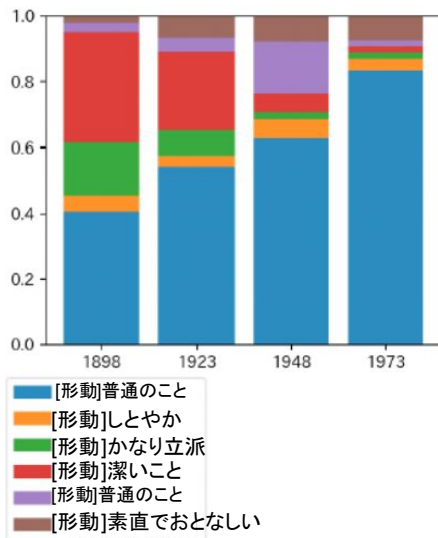
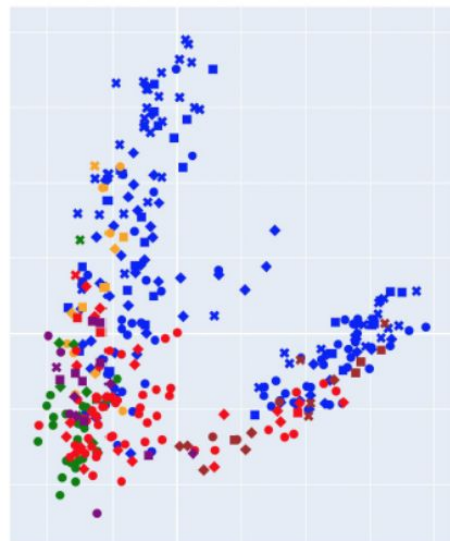
評価

- 著者による定性評価
- 2つのポイント
 - 語義変化の割合を示す棒グラフの変化
 - 実例をみてクラスと意味が対応しているか

一つ目の評価ポイント

クラスタリングの結果が正しいかどうか

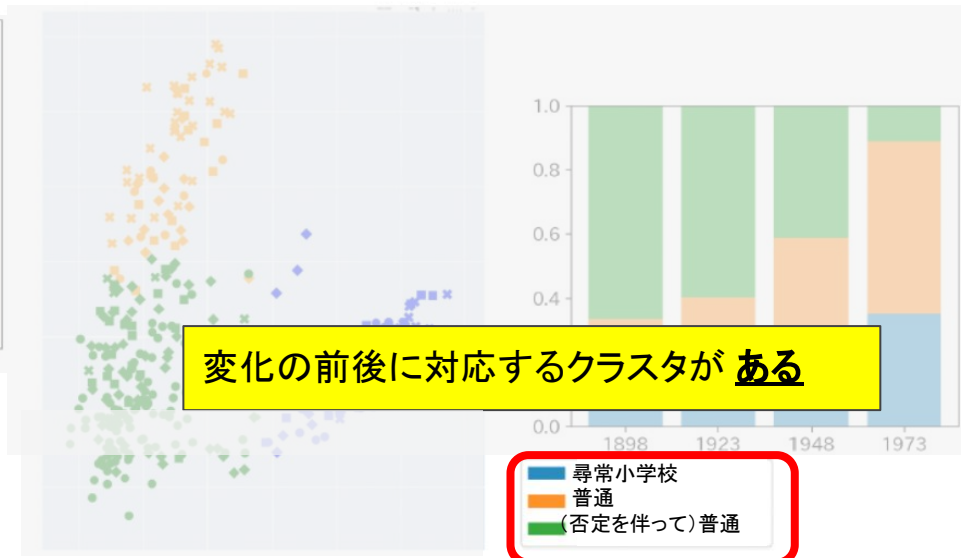
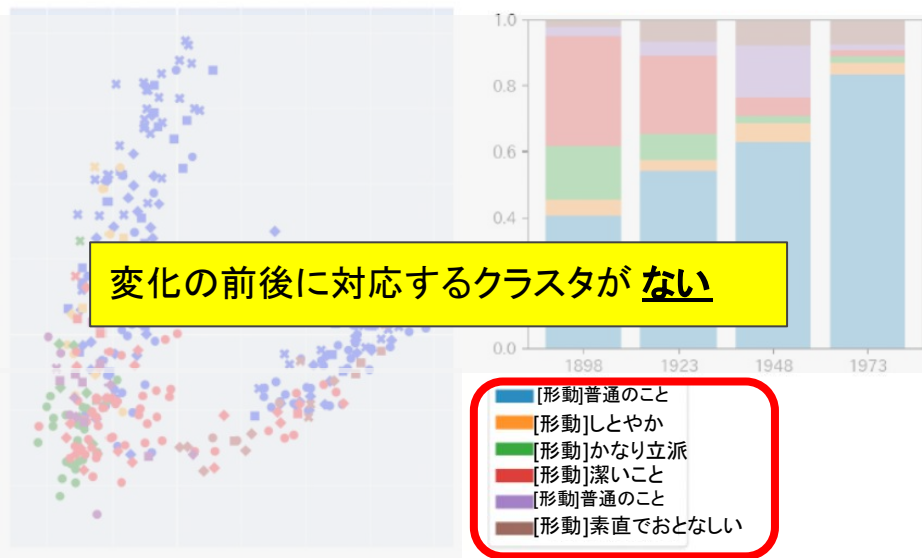
- e.g. 尋常 : 「普通」→「(否定を伴って)普通」



一つ目の評価ポイント

クラスタリングの結果が正しいかどうか

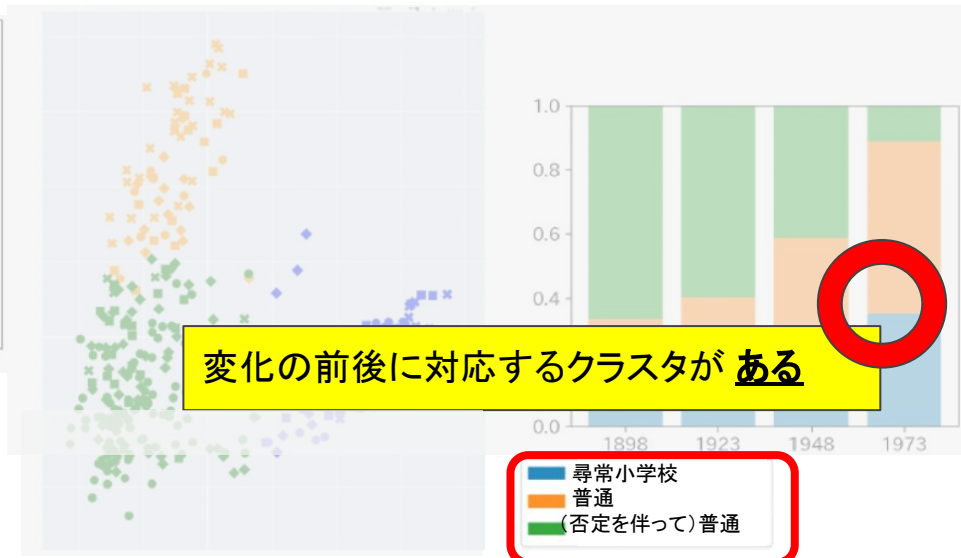
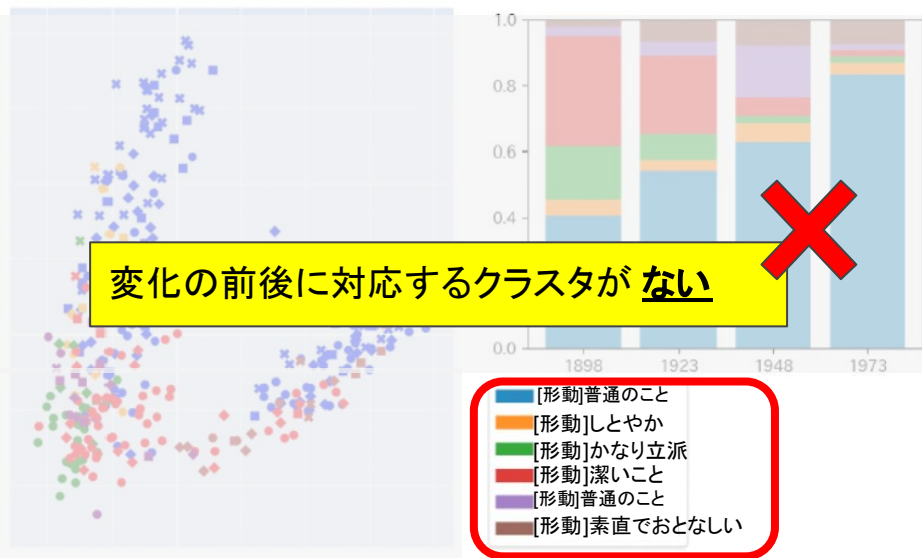
- e.g. 尋常 : 「普通」→「(否定を伴って)普通」



一つ目の評価ポイント

クラスタリングの結果が正しいかどうか

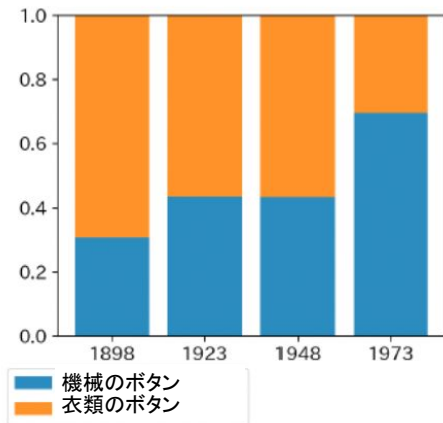
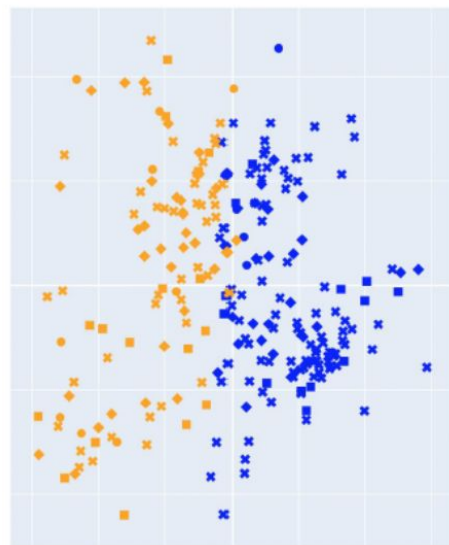
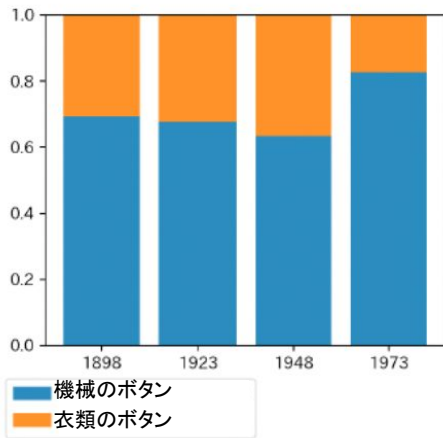
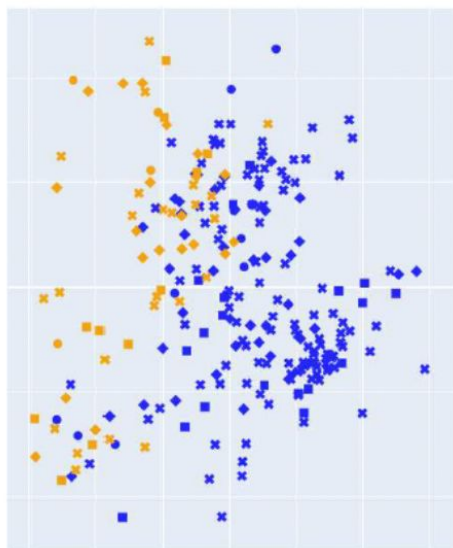
- e.g. 尋常 : 「普通」→「(否定を伴って)普通」



2つ目の評価

棒グラフの増減が正しく対応しているか

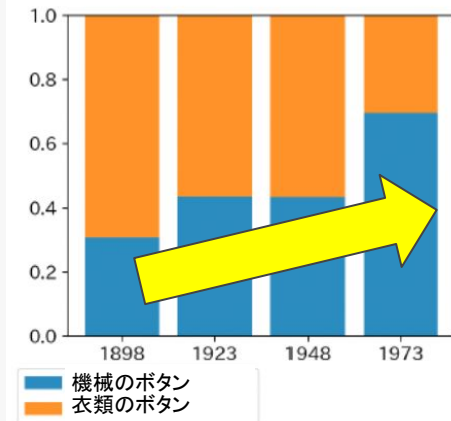
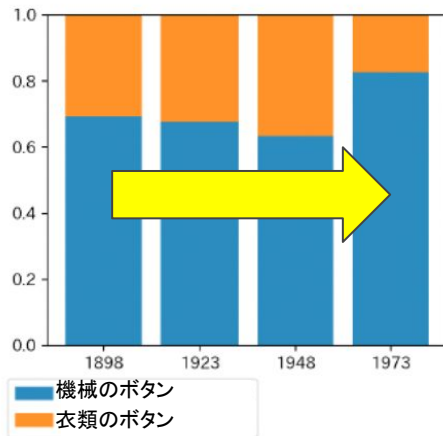
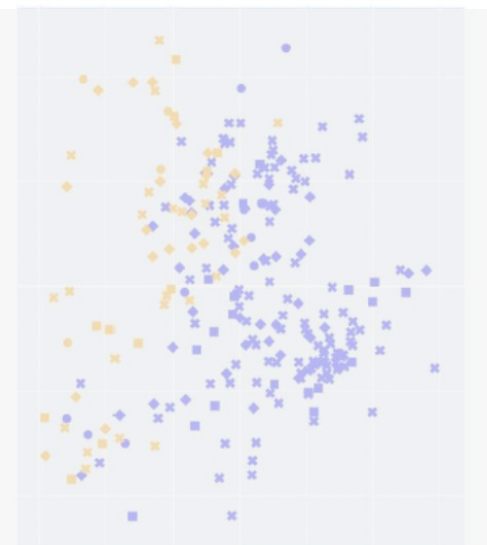
- e.g. ボタン:「衣類のボタン」→「機械のボタン」



2つ目の評価

棒グラフの増減が正しく対応しているか

- e.g. ボタン:「衣類のボタン」→「機械のボタン」



2つ目の評価

棒グラフの増減が正しく対応しているか

- e.g. ボタン:「衣類のボタン」→「機械のボタン」

