

2025年度

英語コーパス学会春季研究会シンポジウム
「大規模言語モデルを利用したコーパスの意味分析入門」

応用1「語の意味の計算」

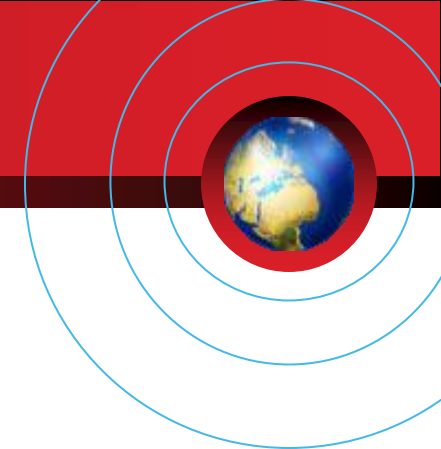


永田亮

甲南大学知能情報学部

理研AIP(客員研究員)

産業技術総合研究所(客員研究員)



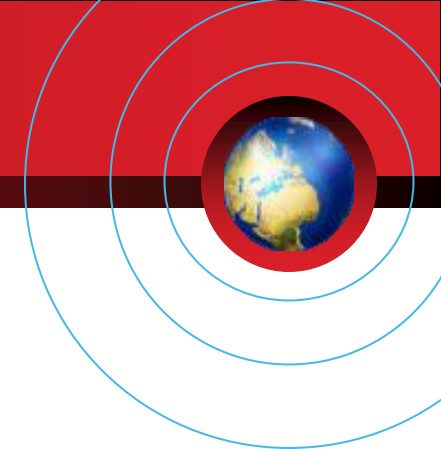
■ 言語モデル(LM)

- 単語予測器
- 現代は深層学習を利用
- 副産物の単語ベクトルが言語分析に重要

■ 単語ベクトル

- 空間上の点 or (超)球面上の点
- 類似度が測れる
- 文脈(周辺単語)をソフトに反映
- 方向性の多様性が文脈(意味)の多様性とみなせる

本セッションの内容



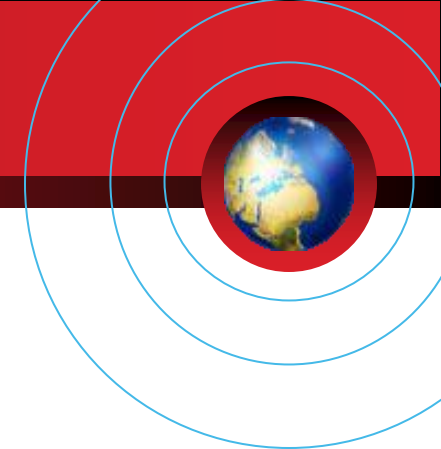
- **単語の意味の多様性の定量化**
 - 単語ベクトルの方向に基づく
- **単語の意味を取り扱った最新研究**
 - meaning-frequency lawの検証
 - 単語の意味変化検出への応用 × 2
 - 文法化の定量化

応用：初級編

単語の語義数を定量化してみる



各単語の持つ意味の数は？



■ 伝統的な方法：辞書を調べる

- 例：take

32, ジーニアス英和辞典(第 5 版)

6, プログレッシブ英和辞典(第 5 版)

■ 新しい方法：単語ベクトルを利用

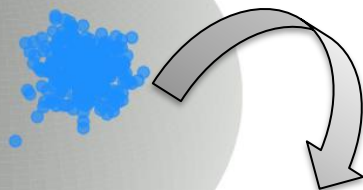
- 意味の多様性として連続値で定量化

- なにがうれしいかは後述

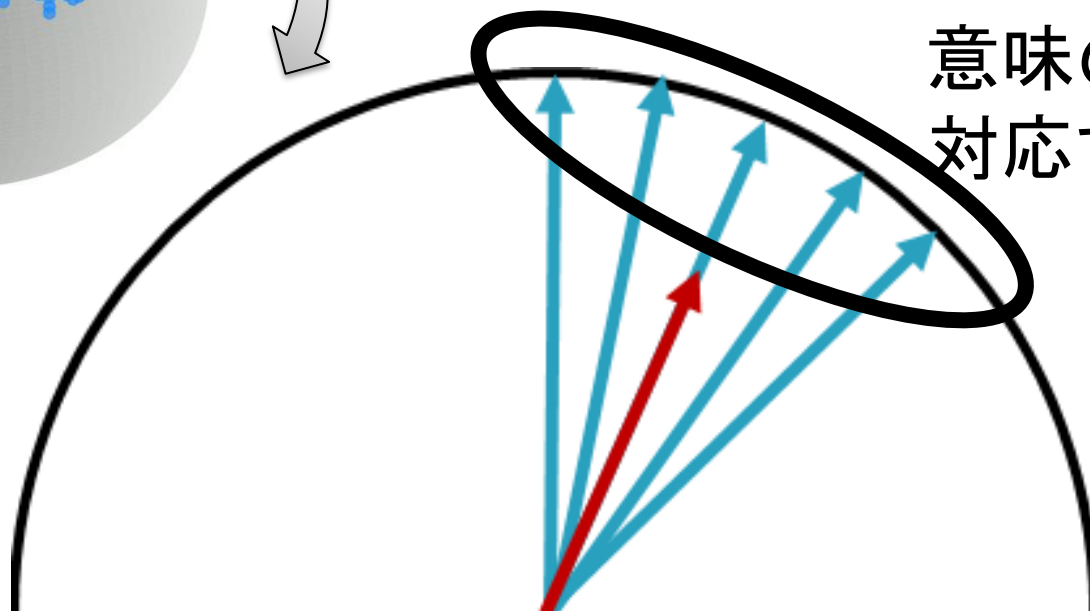
単語ベクトルのばらつきと意味の多様性



例：コーパス中のappleの全事例をベクトル化

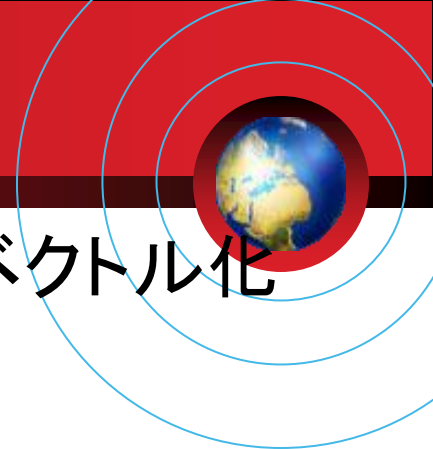


このばらつきが
意味の多様性に
対応すると考える

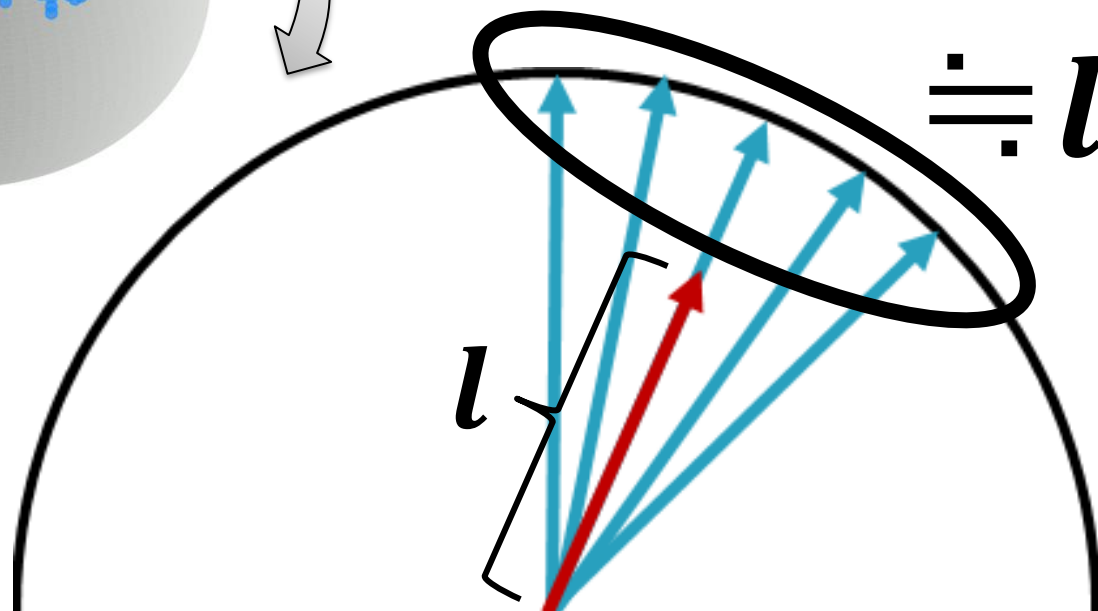
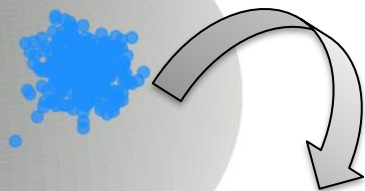


単語ベクトル 
平均単語ベクトル 

ばらつき＝平均ベクトルの長さ

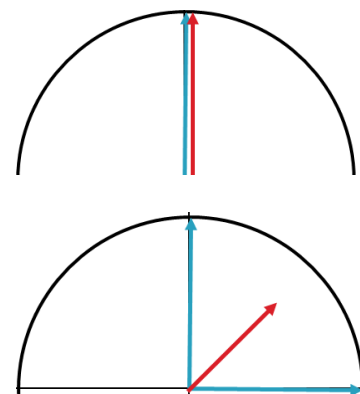


例：コーパス中のappleの全事例をベクトル化



単語ベクトル

平均単語ベクトル



平均ベクトルが短くなるほど意味が多様

意味的分析への応用



単語ベクトルに基づく
新たな meaning-frequency law の検証
(Nagata&Tanaka 2025)

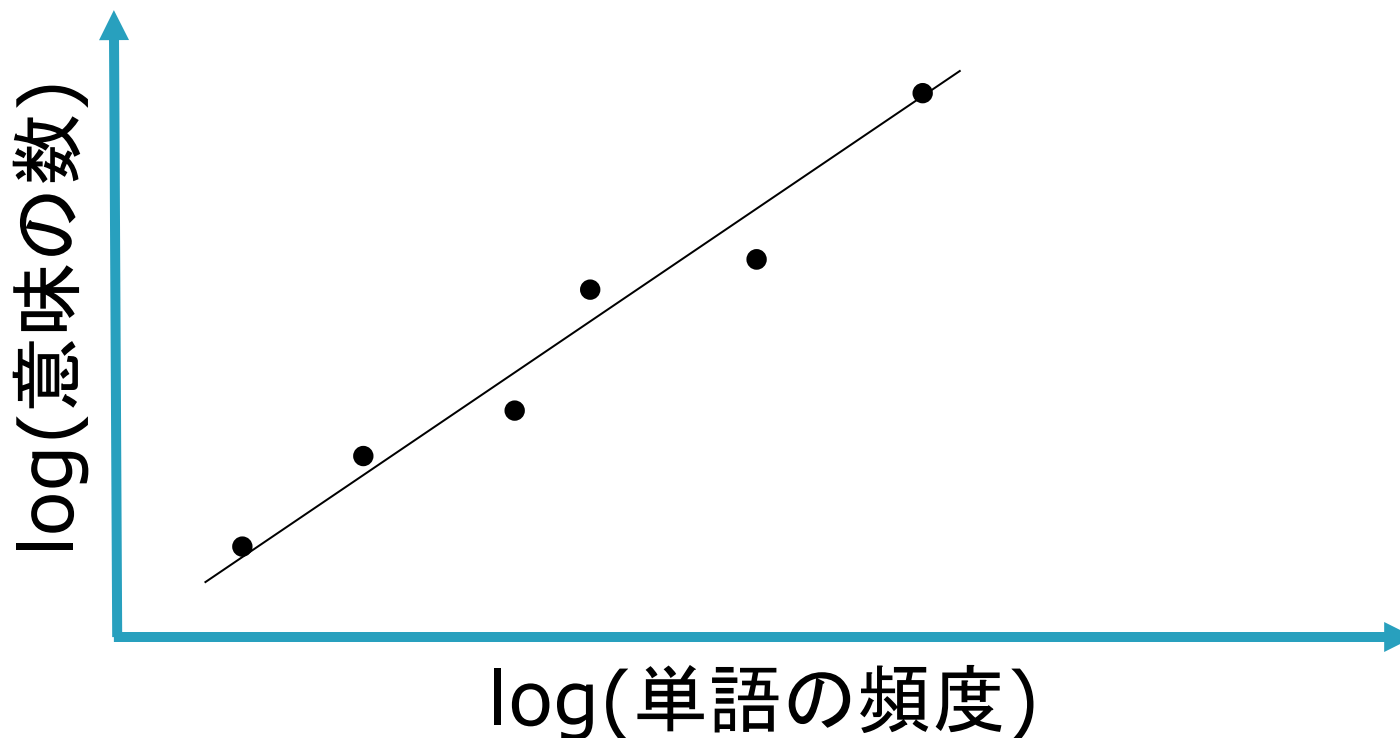


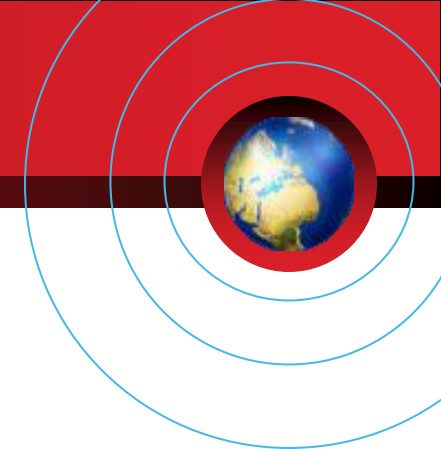
Zipfのmeaning-frequency lawとは？



■ 頻度と意味に関するZipf則

- 頻度が高い単語ほど意味の数が多い





■ 頻度と意味に関するZipf則

- 頻度が高い単語ほど意味の数が多い

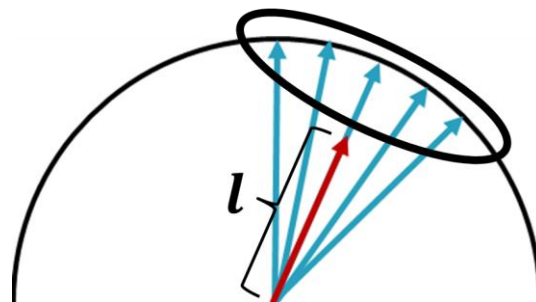
$$\log(m) = \alpha \log(f) + c$$

m : 意味の数, f : 単語の頻度, α, c : 定数

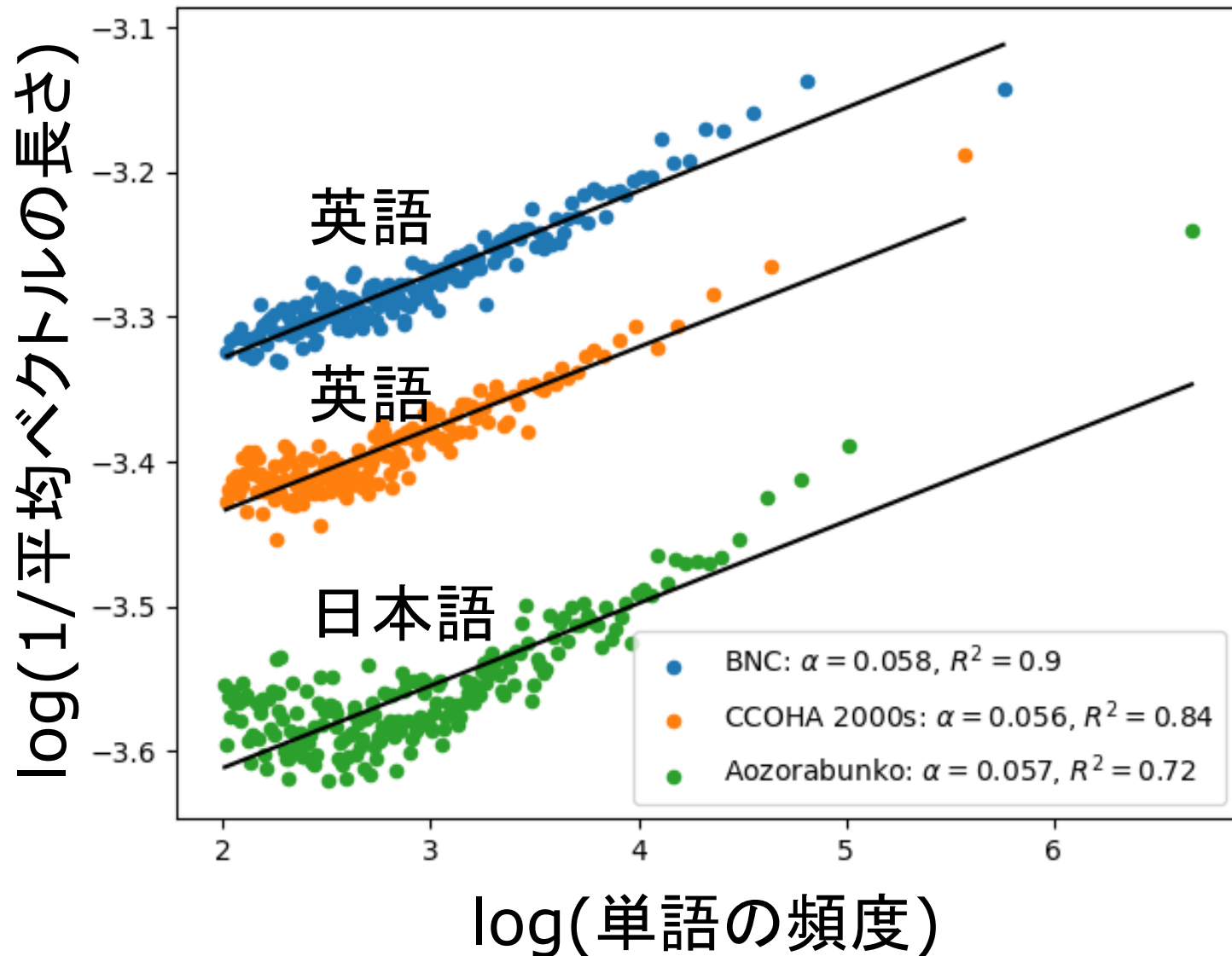
- 従来研究: 意味の数は辞書で

■ 提案手法

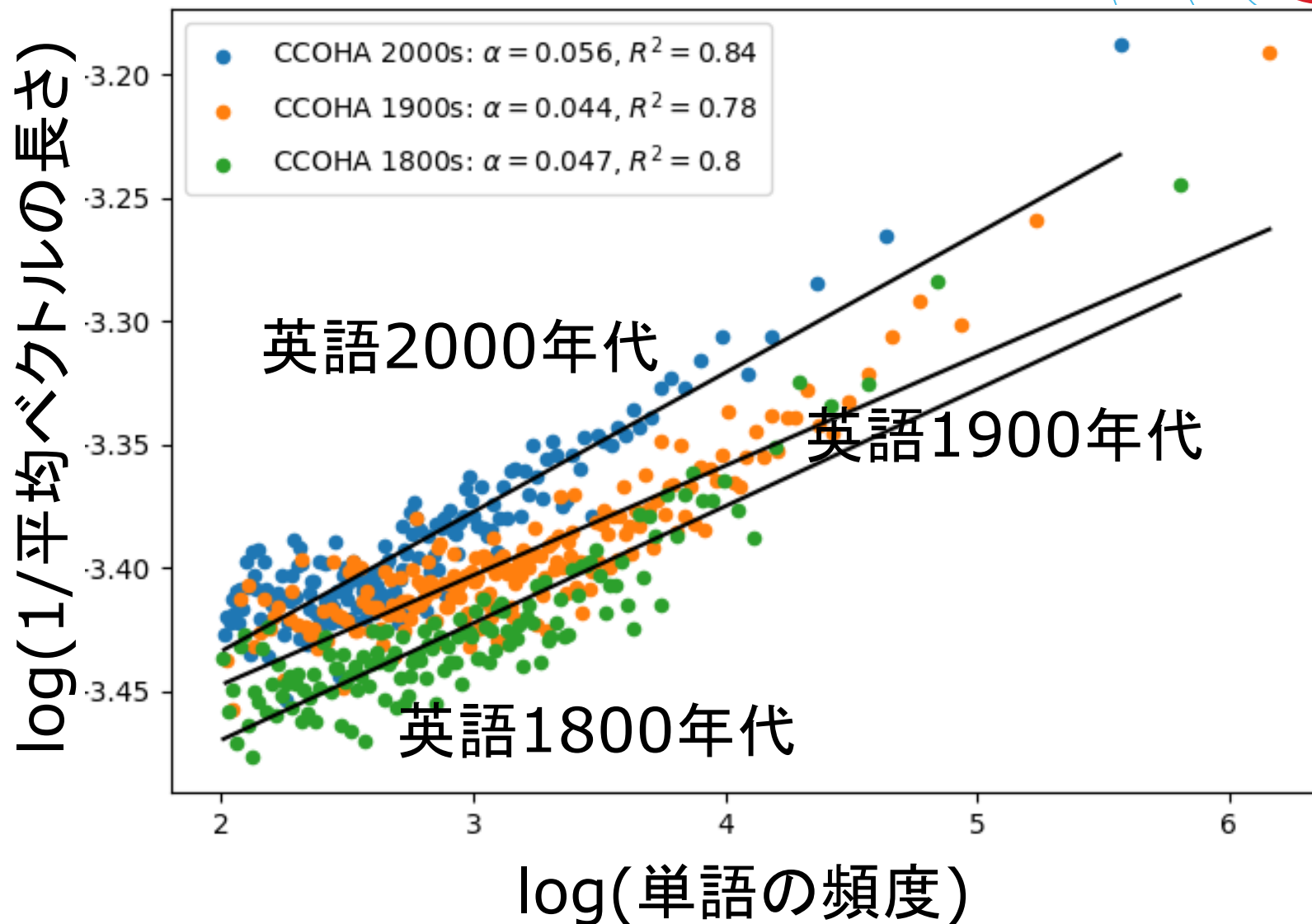
- m を多様性に置き換える(平均ベクトルの長さ)



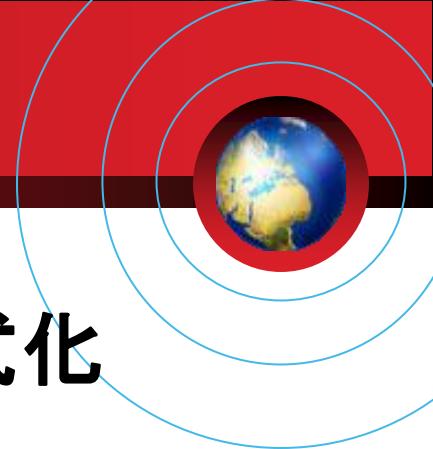
意味の多様性によるMeaning-frequency law



古い文章でも



この研究のまとめ

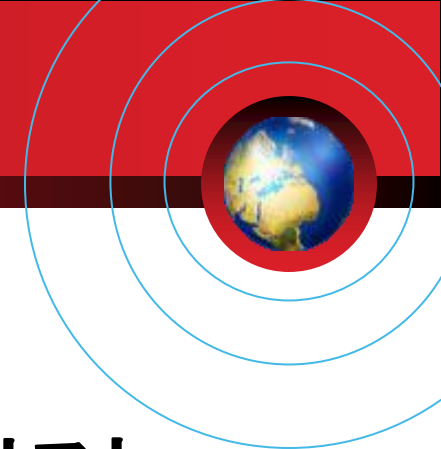


- **新しいmeaning-frequency lawの定式化**
 - 意味の多様性の定量化に基づく
 - ベクトルの方向の多様性→文脈の多様性
 - context-frequency lawとも解釈可能
- **従来を検証を拡張**
 - より幅広い単語, 言語データ

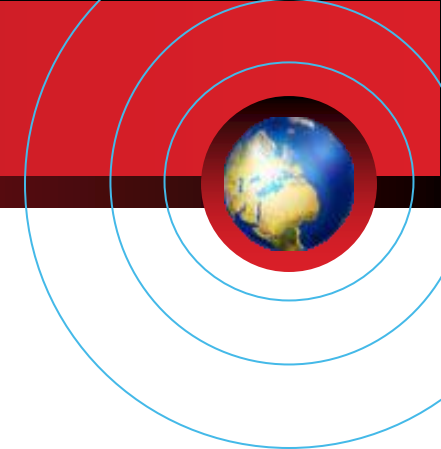
単語の意味変化検出 (Nagata+ 2023)



タスク:意味変化検出



- 入力: 二つのコーパス
- 出力: 意味の差異が大きい単語のリスト
- オプション出力:
 - 差異の程度を表すスコア
 - 意味の広／狭を表すスコア (研究例 少)



■ 歴史言語学

- 意味変化が起こった単語の発見

- 例: trunk ( →  ,  , . . . , )

■ 第二言語習得研究

- 母語話者と非母語話者の語用の違い

- 例: 副詞near (e.g., *near impossible*) 非母語話者少

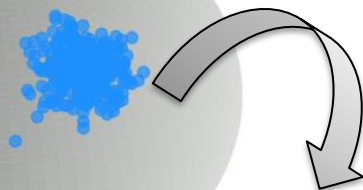
意味の多様性に基づく意味変化検出



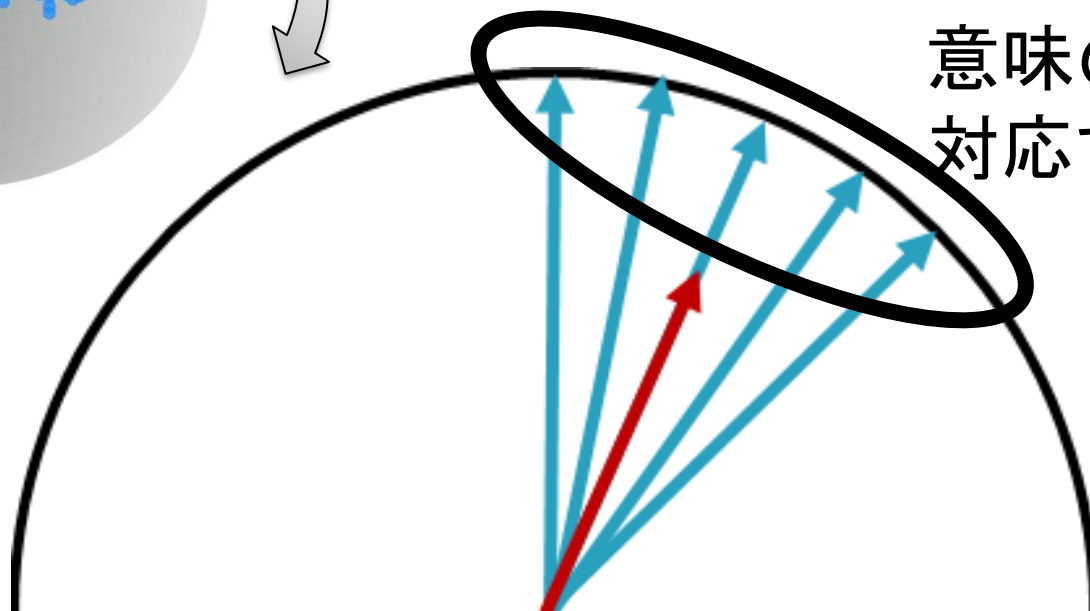
単語ベクトルのばらつきと意味の多様性



例：コーパス中のappleの全事例をベクトル化

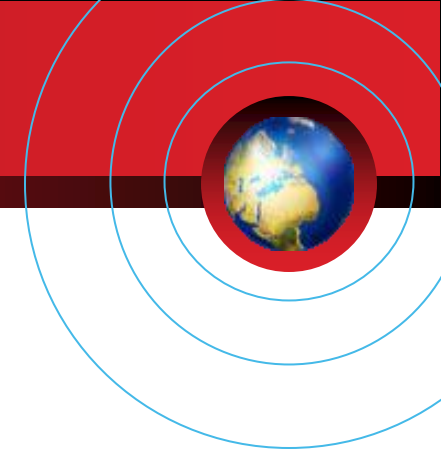


このばらつきが
意味の多様性に
対応すると考える

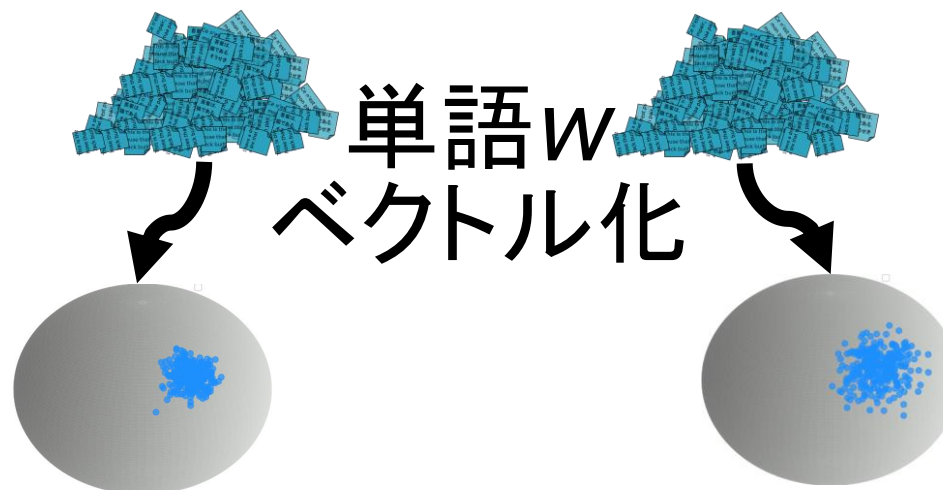


単語ベクトル 
平均単語ベクトル 

提案意味変化スコア



コーパス1 コーパス2



意味変化スコア

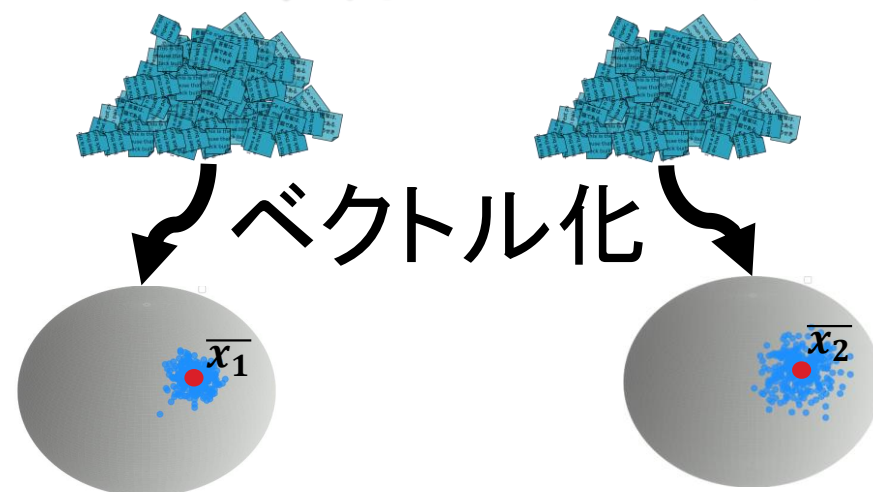
$$\frac{\text{[Diagram of vector space 2]}}{\text{[Diagram of vector space 1]}} \approx \frac{l_2}{l_1}$$

平均ベクトルと長さ l_1 , l_2 を求めるだけ！

代表事例抽出のためのスコア（非重要）



コーパス1 コーパス2



意味変化の代表性スコア

$$\frac{P(x; \text{blue cluster})}{P(x; \text{gray cluster})} \Rightarrow \left(\frac{1}{1-l_2^2} \bar{x}_2 - \frac{1}{1-l_1^2} \bar{x}_1 \right)^t x$$

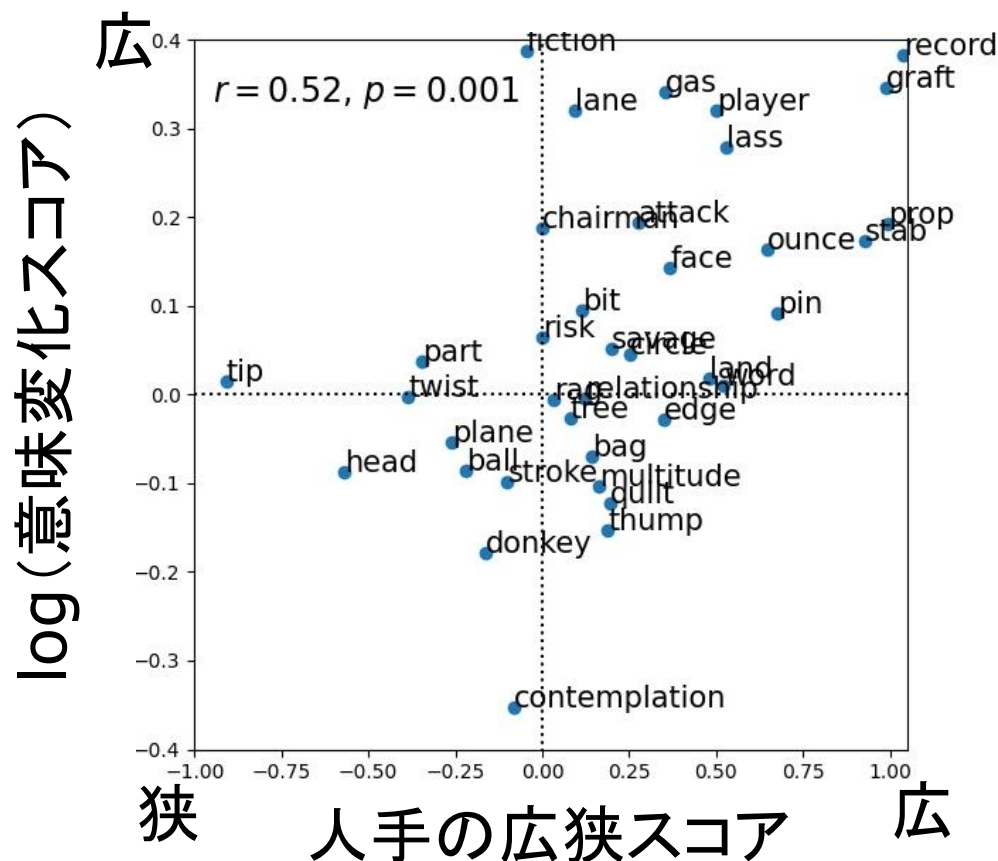
一方にのみ頻出する事例 $\Rightarrow$ スコア大

定量評価：意味の広狭判定



■ 人手で求めた意味の広狭スコアとの相関

— データ: DWUG (1810–1860 vs 1960–2010)



定性評価：母語話者 vs. 非母語話者

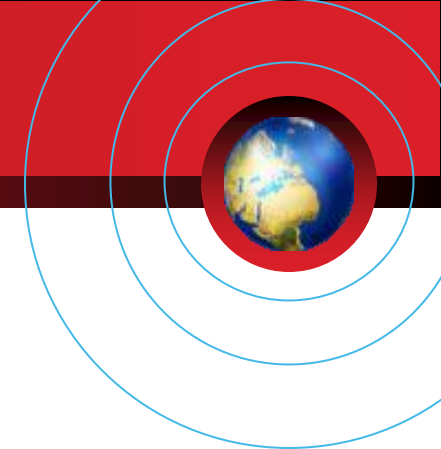


母語話者で意味が広い単語

Word		$\frac{P(x; \text{blue})}{P(x; \text{grey})}$	
near	0.54	0.35	the near global financial collapse
		0.29	it has become near impossible
		-0.63	in the near future.
		-0.83	never smoke near their children
		-0.92	smokes near me
hold	0.34	0.35	public knowledge that <i>SVO</i>
		0.29	common knowledge that <i>SVO</i>
		-0.83	knowledge that we have
		-0.92	knowledge about the world

Data: ICNALE

この研究のまとめ

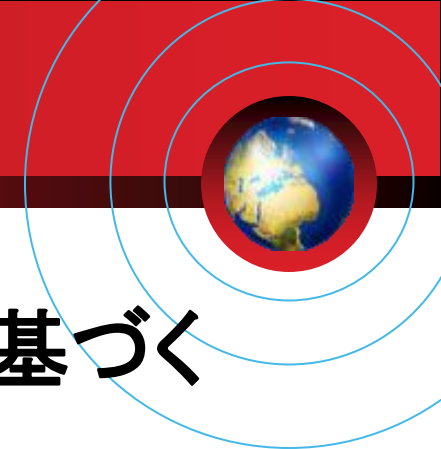


- 意味変化検出を新しいアプローチで
 - やはり意味の多様性に基づいて
- アドバンテージ
 - シンプル & 計算量 少（大規模コーパスも）
 - 制約 少 ⇒ 幅広いコーパスに適用可能
 - 意味の広狭も推定可能
- 意義
 - ベクトルの方向のばらつきが大事
(Aida+ 2023でも)

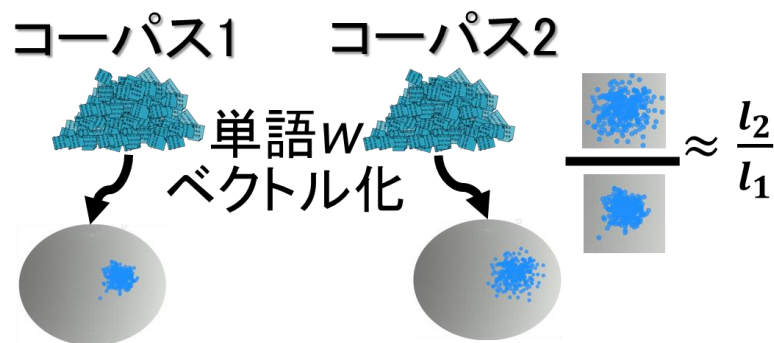
単語ベクトルの直接比較による 単語意味変化分析

(Kishino, Yamagiwa, Nagata, Yokoi,
Shimodaira, 2025)



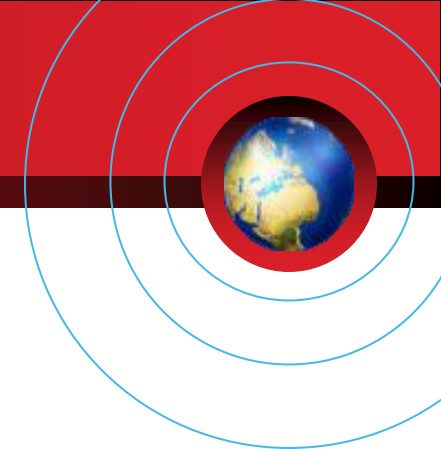


■ 単語ベクトル(の方向)の散らばりに基づく



- 個別の単語ベクトルを比較したほうが良い？
 - 片方にしかないベクトル(=事例)が見つかる？
 - 散らばりは同じだけど空間上の場所が異なる場合

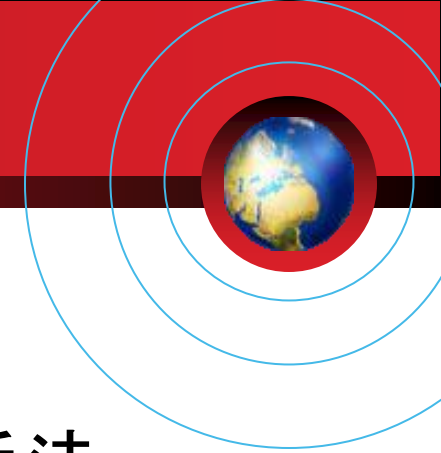
直接的に比較したい



■ 単語ベクトル同士の比較

- 最適輸送という技術を用用
- 単語ベクトル全体にわたる変化量も定量化
- 可視化もできる

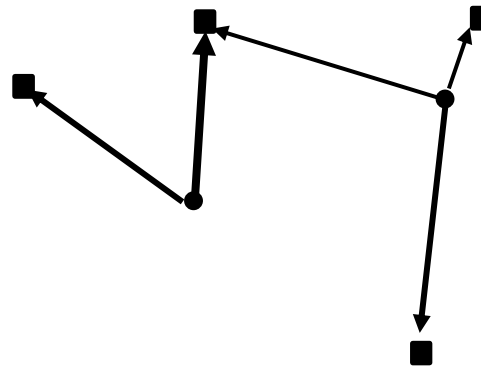
最適輸送とは



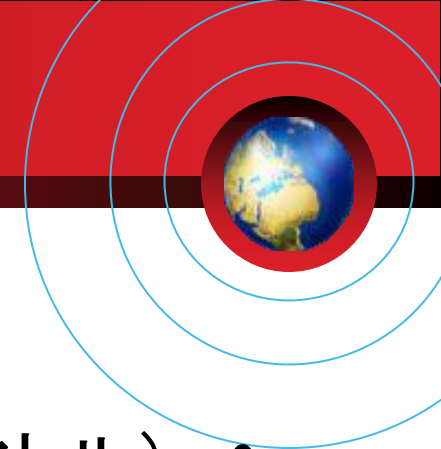
■ 最適輸送とは

- 物資を**最適に輸送**するための一群の手法
- 例：工場からデパートへの商品配送
- 最適輸送用ツールで簡単に実装可能

空間上の点の対応付けとも解釈可能



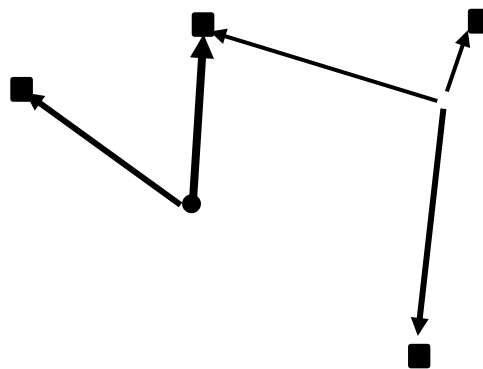
意味変化検出への応用

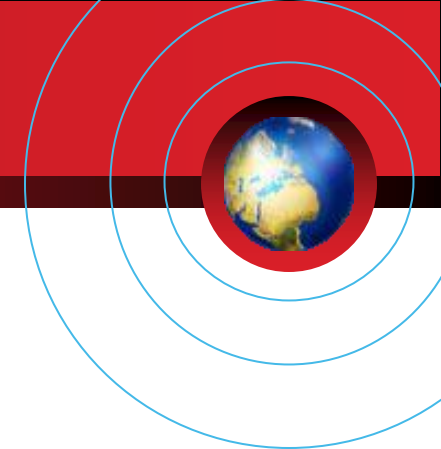


■ 次のように便宜的に認識

- 工場: 古い時代のコーパスの単語 (ベクトル) •
- デパート: 新時代のコーパスの単語 (ベクトル) ■

空間上 (球面上) の点の対応付けと等価



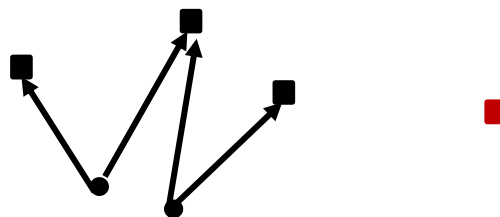


■ 最適輸送

- 輸送元(工場): 必ずどこかへ送り
- 輸送先(デパート): 必ずどこから受け取る

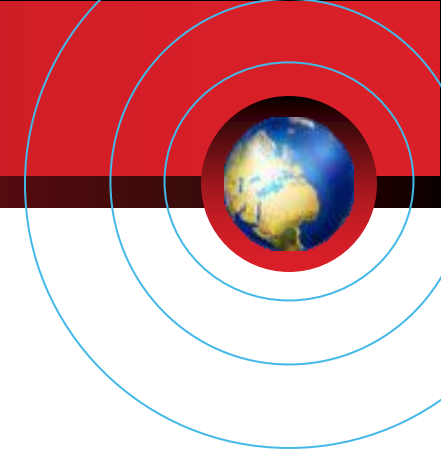
■ 単語の意味

- 輸送元(旧時代): 送らなくてもよい(意味消失)
- 輸送先(新時代): 受け取らなくてもよい(意味出現)

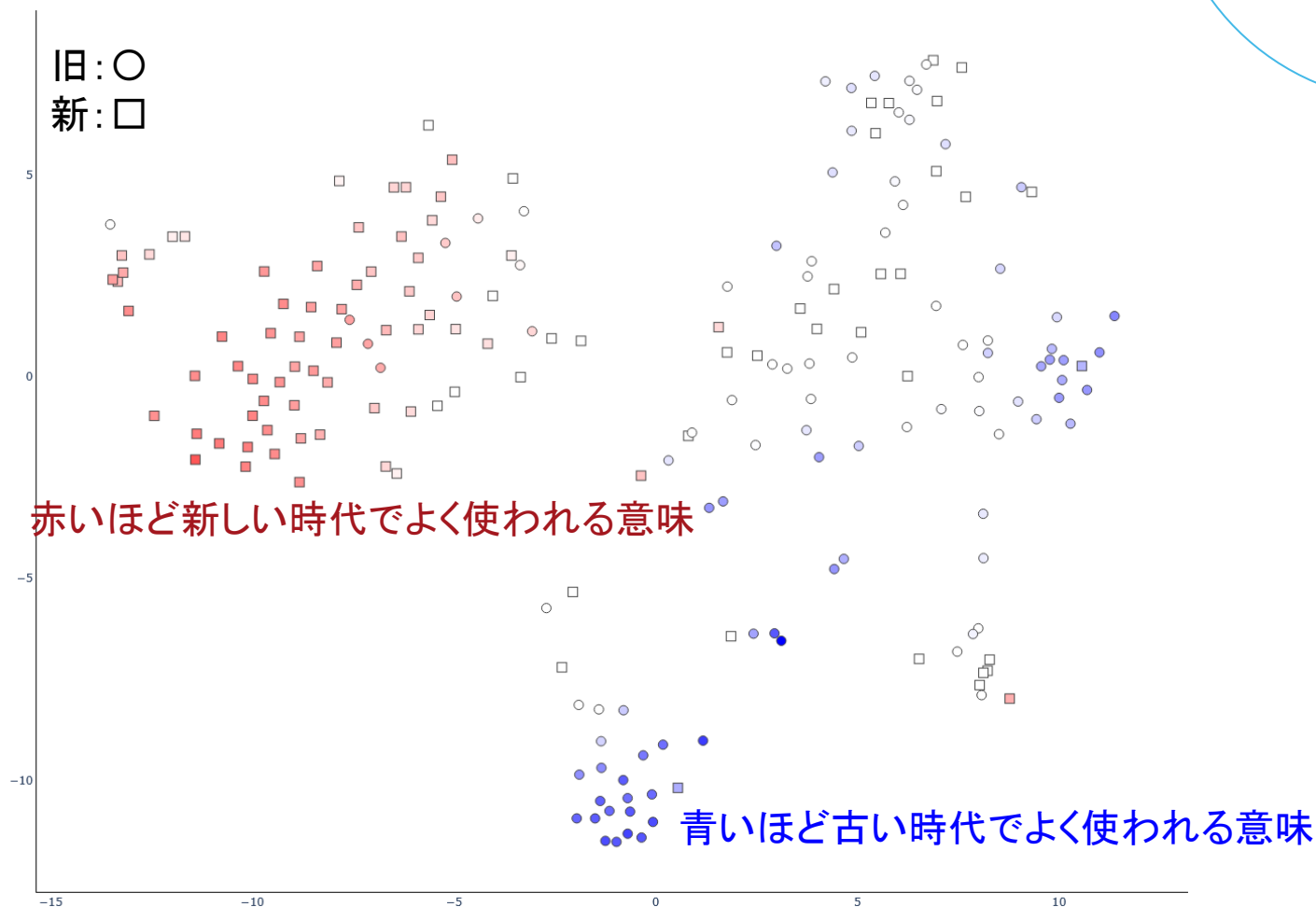


不均衡最適輸送を利用

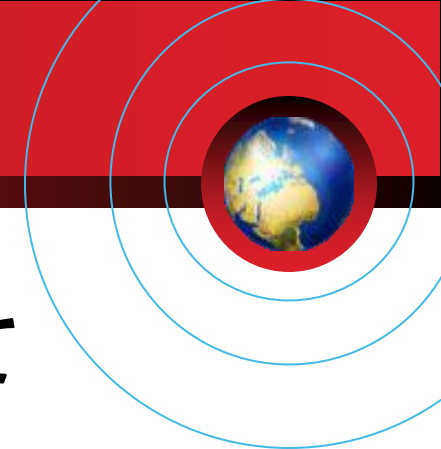
意味変化検出と可視化の実例



例: *ball* の新旧比較



この研究のまとめ

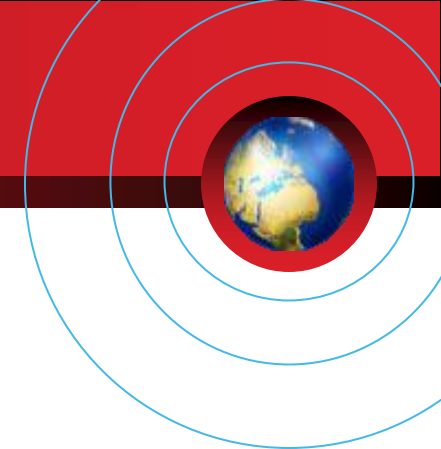


- 意味変化検出を最適輸送問題として
 - 単語（ベクトル）の直接比較
 - でもやっぱりLMから得られる単語ベクトル
- 可視化が可能
 - 意味分析に有益？
- 意義
 - 意味変化を輸送問題と捉えた
 - 不均衡最適輸送であることが重要

動詞派生前置詞の文法化の定量化 (Nagata+ 2024)

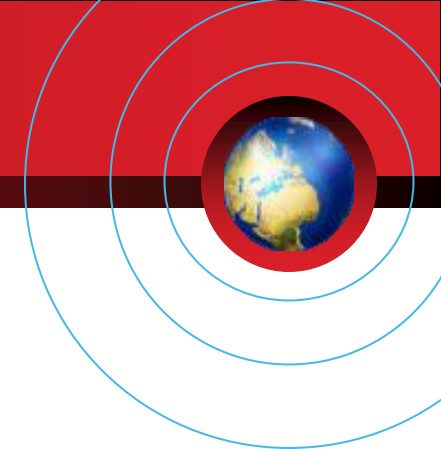


本研究の最終ゴール



■ 文法化の定量化

- 文法化の程度を定量化
 - 今回: 動詞派生前置詞
- 文法化が起こっている語の発見



■ 通時変化: 内容語 → 機能語

– 例: *can*

～14世紀 知る(動詞) → 現代 できる(助動詞)

動詞派生前置詞とは



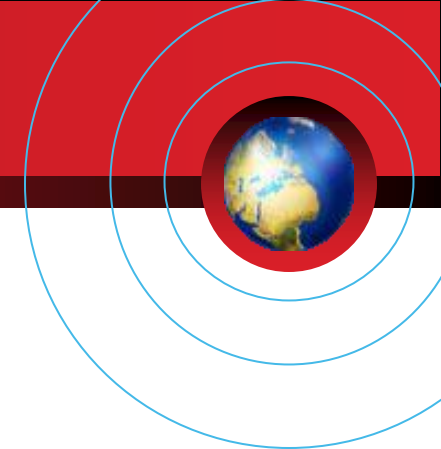
■ 動詞派生前置詞

- 動詞から派生 & 前置詞のように振る舞う

例: *during, considering, according to*

■ Deverbal prepositionsとも

- 以降, D-prepと省略



■ 基本

– 動詞 → 前置詞

■ 言語学における仮説

1. 完全には文法化していない(前置詞と動詞の間)

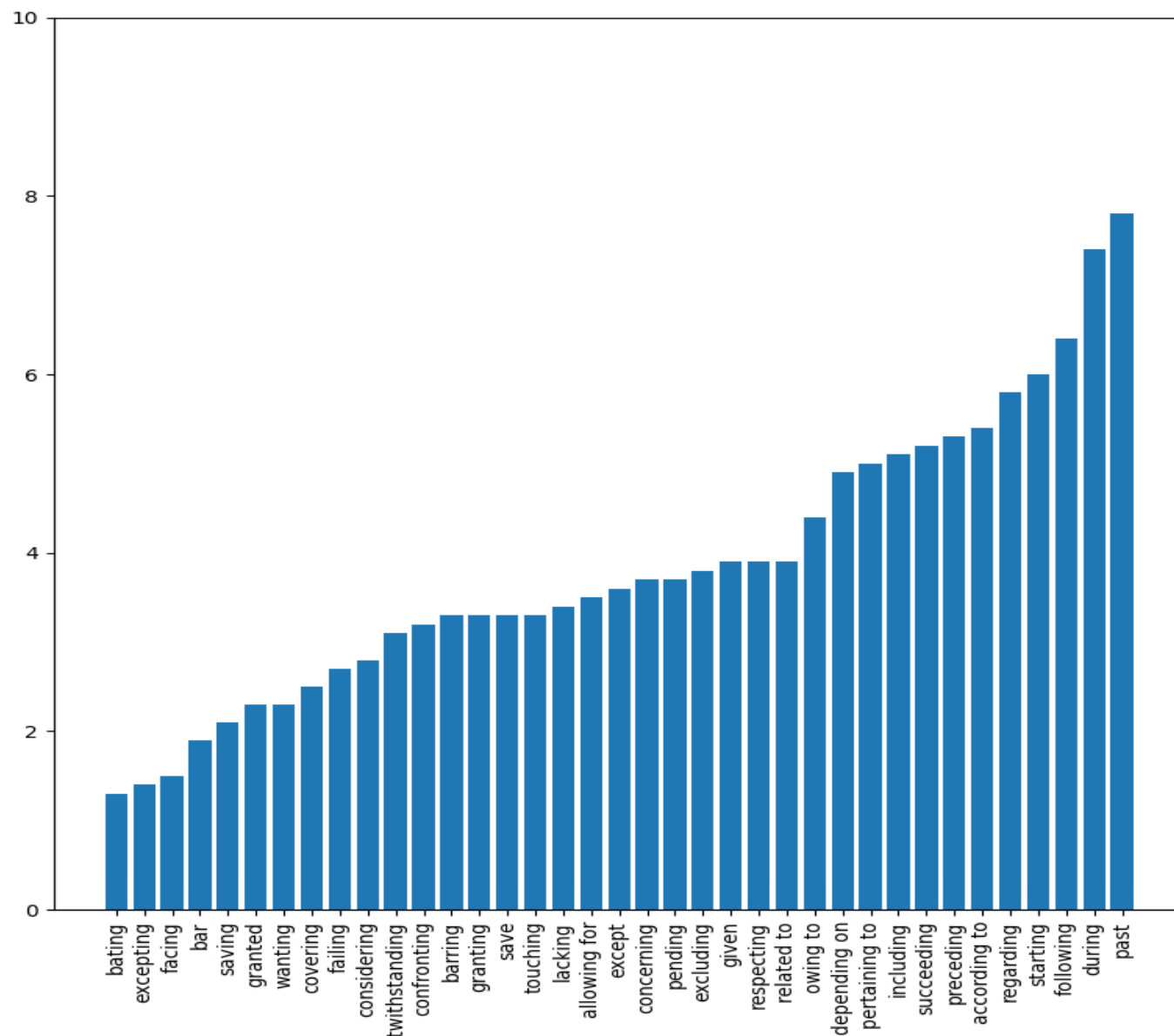
2. 文法化は連続的で, D-prepの種類による

例, 高: *past, during, except*

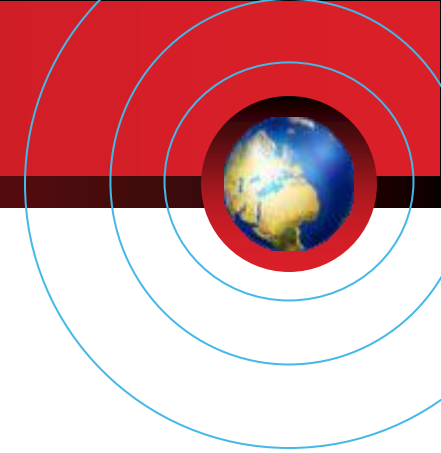
低: *succeeding, considering, allowing for*

3. 文法化の度合いは, 前置詞性と動詞性で判定可能

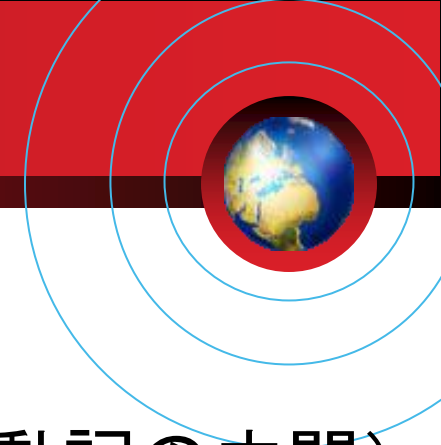
D-prepsの文法化度 (Hayashi 2015)



既存研究での仮説の検証



- 基本的に、人が介入
 - コーパス分析
 - 質問紙法
- 課題：検証規模を大きくするのが困難
 - 対象とするD-prepの種類
 - 事例数



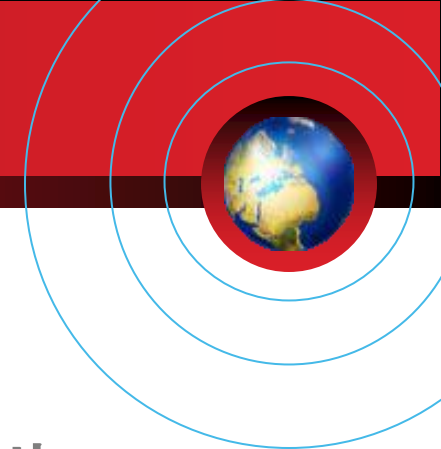
■ 三つの仮説は正しい？

1. 完全には文法化していない(前置詞と動詞の間)
2. 文法化は連続的で, D-prepの種類による
例, 高: *past, during, except*
低: *succeeding, considering, allowing for*
3. 文法化の度合いは, 前置詞性と動詞性で判定可能

本研究 D-prepの文法化度の定量化



二種類の単語ベクトル(再掲)



■ トークン(事例)ベース

- 各事例に応じた複数の単語ベクトル
- 個別文脈を反映した単語ベクトル

例: *bank clerk* と *river bank* では異なるベクトル

■ タイプベース

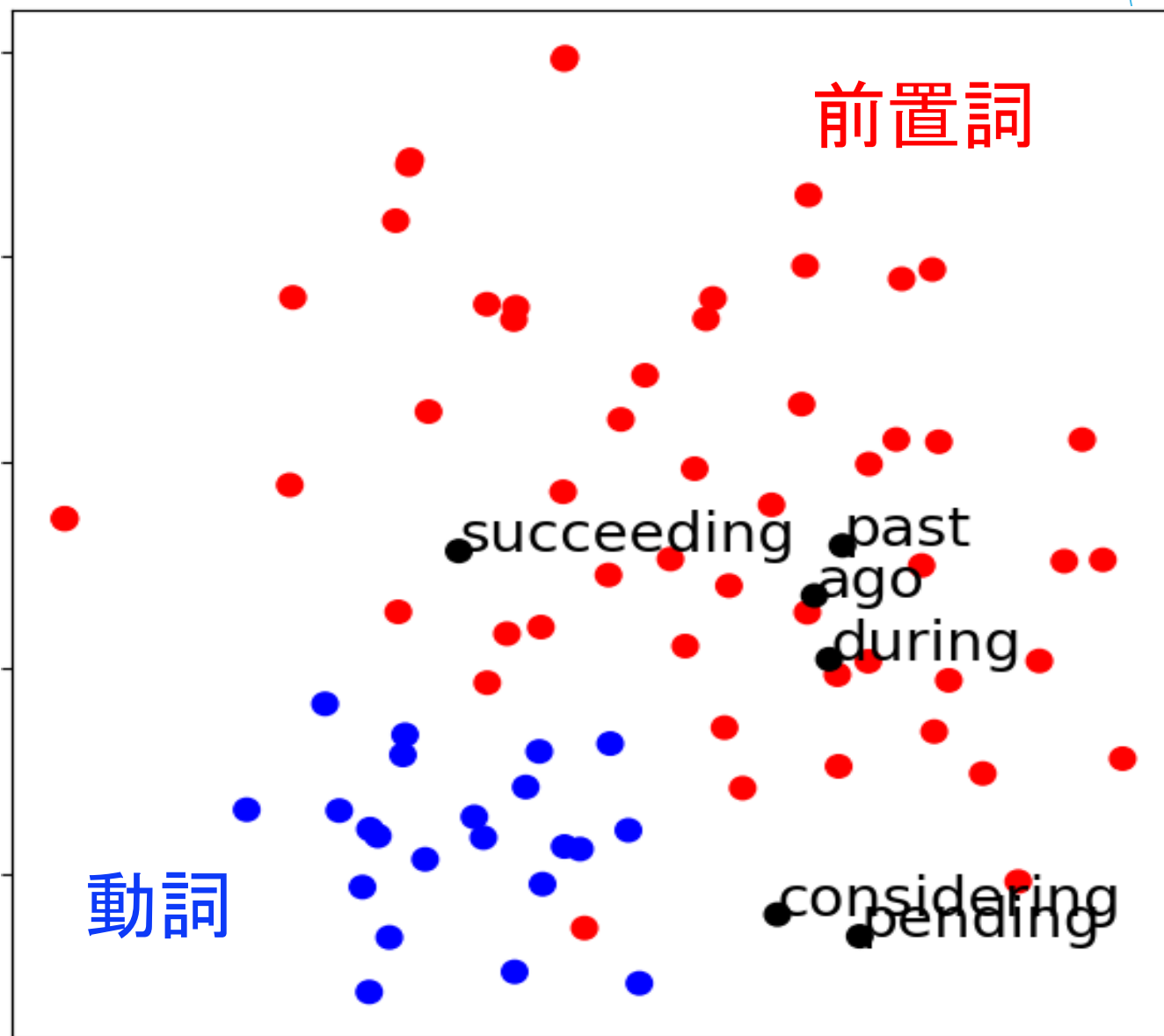
- 単語タイプに対して**一種類**の単語ベクトル
- **全文脈を反映した**単語ベクトル
- 単語のタイプとしての性質や特徴を反映

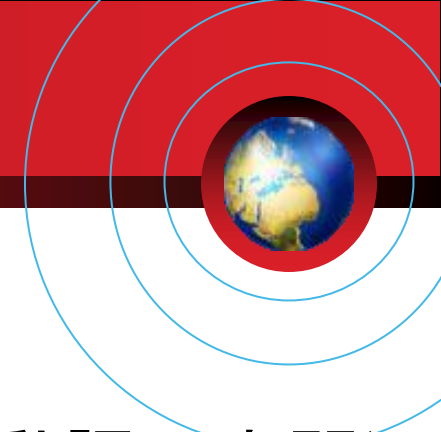
例: *bank* のベクトルは「銀行」も「土手」も反映

タイプベースベクトルを利用した 文法化度の定量化



手法の直感的な説明





■ 動詞派生前置詞:

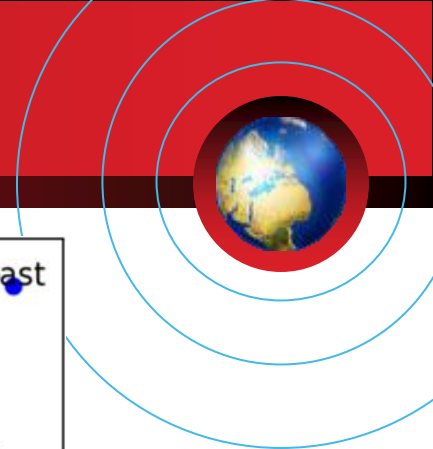
1. 完全には文法化していない(前置詞と動詞の間)
2. 文法化は連続的で, D-prepの種類による

例, 高: *past, during, except*

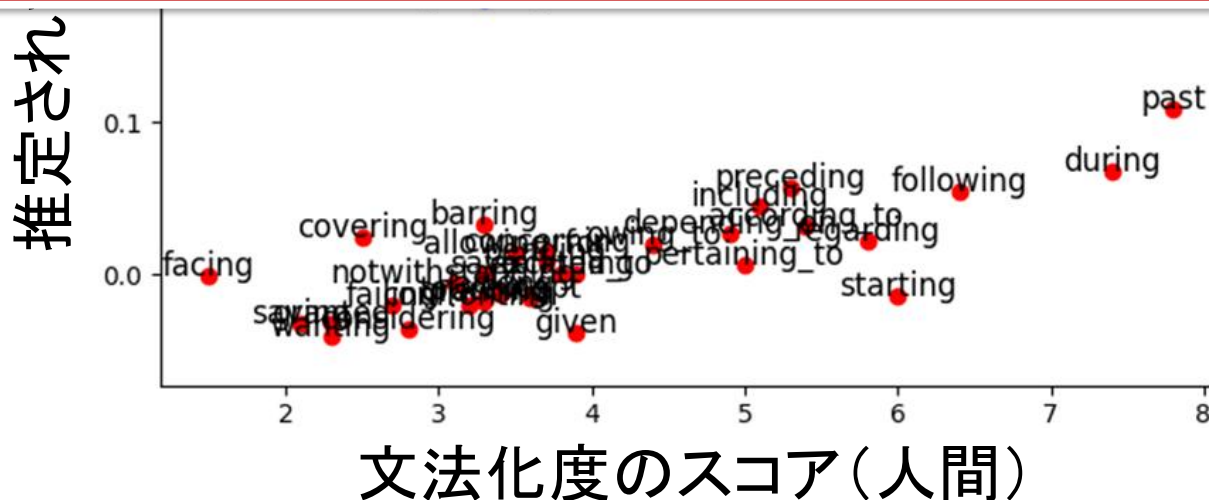
低: *succeeding, considering, allowing for*

3. 文法化の度合いは, 前置詞性と動詞性で判定可能

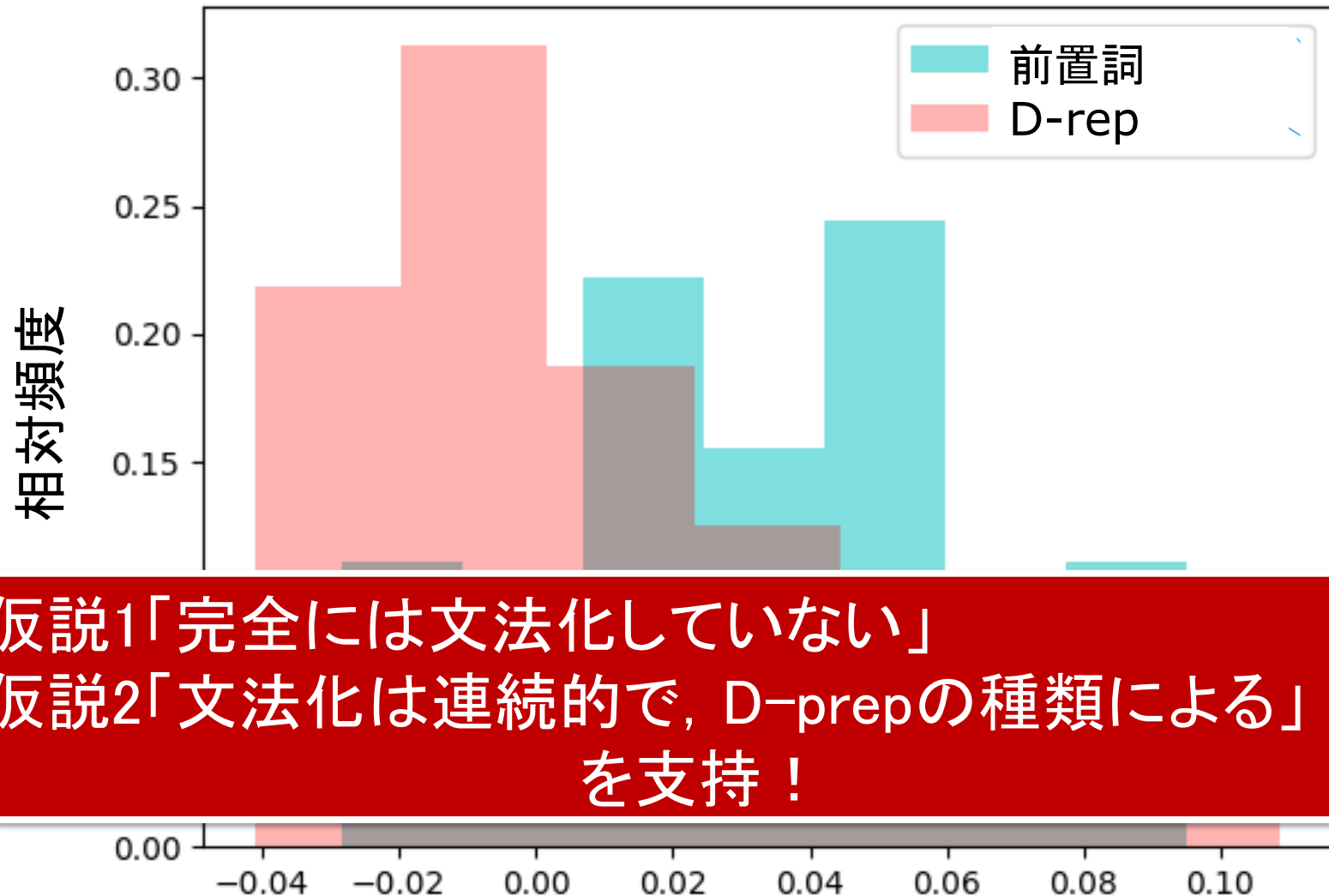
動詞の情報の効果



「仮説3: 文法化度は前置詞性と動詞性で判定可能」
を支持



前置詞, D-prepの文法化度

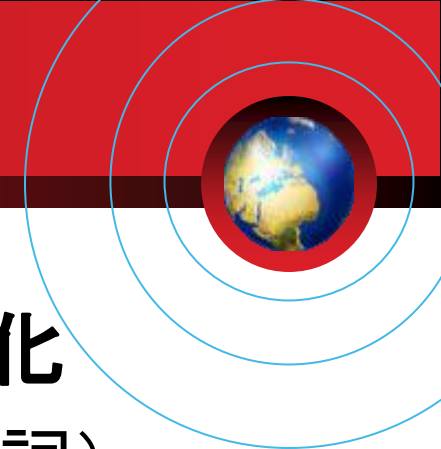


仮説1「完全には文法化していない」

仮説2「文法化は連続的で, D-prepの種類による」
を支持！

推定された文法化度のスコア(前置詞 & 動詞)

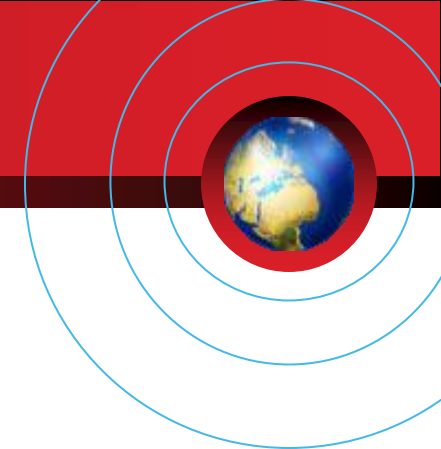
この研究のまとめ



- **動詞派生前置詞の文法化度を定量化**
 - タイプベース単語ベクトル(前置詞 & 動詞)
- **三つの仮説の検証(全て支持)**
 1. 完全には文法化していない
 2. 文法化は連続的で, D-prepの種類による
 3. 文法化度は前置詞性と動詞性で判定可能
- **意義**
 - 文法化の研究がデータドリブン & 定量的に
 - 仮説の補強に使えるという一事例

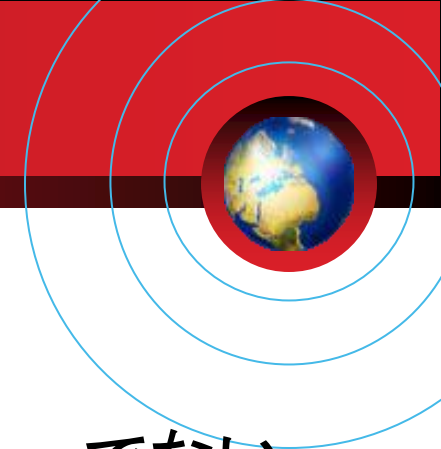
全体まとめ





- **言語モデルLMとは**
 - 単語予測器
 - 副産物の単語ベクトルが言語分析に重要
- **単語ベクトルベースの言語分析例**
 - Meaning-frequency lawの検証
 - 意味変化検出
 - 意味の広がりに基づく
 - 単語事例同士の対応付
 - 文法化の定量化
 - まだまだ応用の可能性が！

のこされた課題，問題

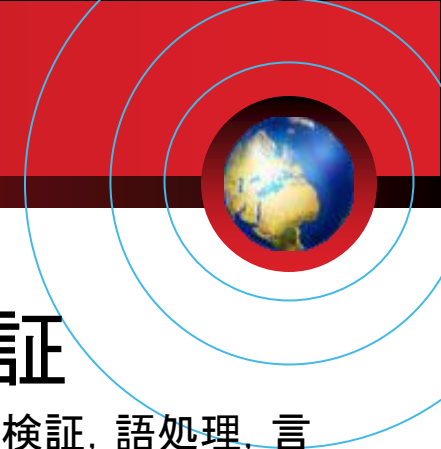


- **真にLなLM(LLM)だとどうなる**
 - 今回用いたLMは現代的な意味ではLargeでない
 - 理論的，技術的問題が
 - 現在研究中(少しずつ明らかに)
- **言語分析への適用範囲**
 - 適用できそうな言語現象はまだまだありそう
 - 得られた結果の妥当性，解釈性に改善の余地あり

関連情報案内



文献案内: 今回の内容



■ Meaning-Frequency Lawの新しい検証

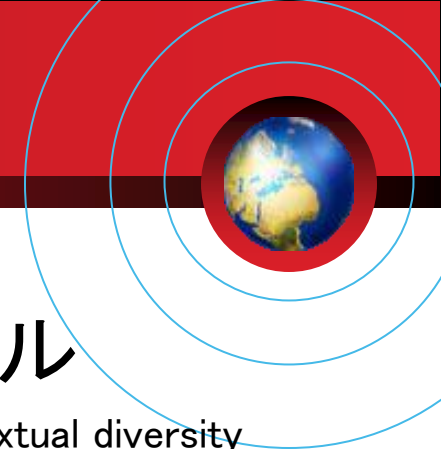
- 永田, 田中, 単語ベクトルに基づく新たな meaning-frequency law の検証, 語処理, 言語処理学会発表論文集, 2024.
- A New Formulation of Zipf's Meaning-Frequency Law through Contextual, ACL, 2025.

■ 意味変化検出手法

- Nagata, Takamura, Otani, and Kawasaki. Variance Matters: Detecting Semantic Differences without Corpus/Word Alignment, EMNLP, 2024.
- Kishino, Yamagiwa, Nagata, Yokoi, Shimodaira, Quantifying Lexical Semantic Shift via Unbalanced Optimal Transport, ACL, 2025.

■ 文法化度の定量化

- 永田, 川崎, 大谷, 高村, 動詞派生前置詞の文法化の定量化, 言語処理学会発表論文集, 2024.
- Nagata, Kawasaki, Otani, and Takamura. A Computational Approach to Quantifying Grammaticization of English Deverbal Prepositions. LREC-COLING, 2024.



■ Meaning-Frequency Law検証用ツール

- https://github.com/nagata-github/meaning_frequency_law_via_contextual_diversity

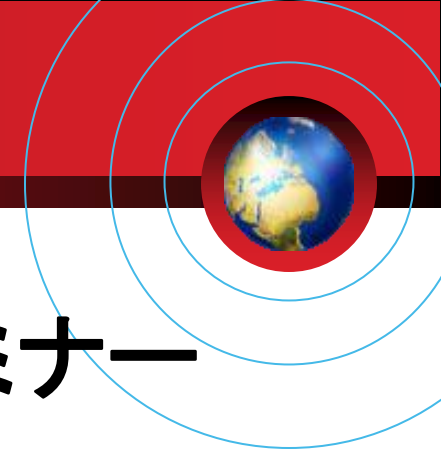
■ 意味変化検出用ツール

- 単語ベクトルのばらつきに基づく手法

https://github.com/nagata-github/vmf_meaning_change_detector

- 最適輸送に基づく手法

<https://github.com/ryo-lyo/Semantic-Shift-via-UOT>



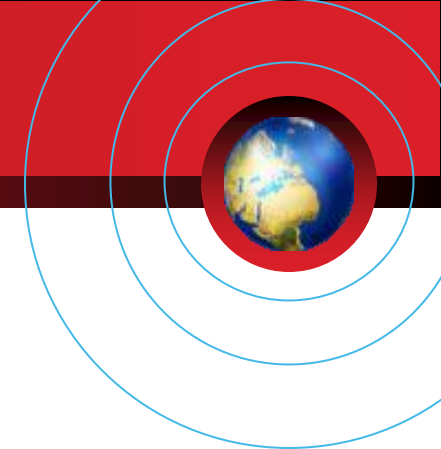
■ 言語処理学会言語処理技術セミナー

- 2025年9月13～14日(仮)
- 東大駒場キャンパス(仮)
- 言語学 × NLP

ハンズオン(プログラミング)を含む

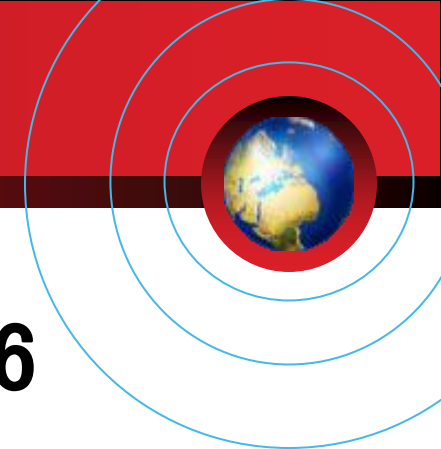
初級, 中級, 上級

今回紹介した研究の実装にもトライ



■ 言語処理学会年若手の会

- 2025年9月17–19日
- 会場:アクトシティ浜松
- 若手研究者の交流会
- 研究ポスター発表(アイデア段階のもの多数)



■ 言語処理学会年次大会NLP2026

- 2026年3月9-13
- 会場: ライトキューブ宇都宮
- 自然言語処理の最新研究満載
- 人文系に関係が深いセッションも！
(昨年)

人文学と言語処理

認知・脳と自然言語処理

人とAIの共生に向けた対話システム・言語使用の研究
言語とコミュニケーションの創発