英語コーパス学会
第49回大会ワークショップ
「深層学習時代のコーパス分析：コーパス検索では見えないこと，深層学習で見えること」

永田亮（甲南大学／理研），川崎義史（東京大学）

再配布禁止，個人利用でお願いします．

はじめに
少し自己紹介

自己紹介

- 永田亮（甲南大学/理研）
  - 専門分野：自然言語処理（NLP）
    文法誤り訂正
    文章の自動評価
    解説文の自動生成
    NLP × 言語学

NLP，深層学習の言語分析への応用に大きな可能性を感じています．

自己紹介

- 川崎義史（東京大学）
  - 専門分野：スペイン語史
    コーパス言語学
    計量文献学
    計算言語学

計算機を利用した言語変化のモデリングが最終目標

ここから本題です

本日お話しする内容

- 導入：深層学習とはなにか？（15分）
  - 入門編：ある観点から大胆に要約
- 基礎：深層学習を利用した言語分析（35分）
  - 深層学習ベースの自然言語処理
  - そこから得られる副産物が重要
- 応用：研究事例の紹介（50分；主に川崎担当）
  - 意味変化の自動分類と可視化
  - 形態変化のモデル化
  - 言語習得のモデル化

前置き

- 分かり易さ ＞ 厳密性
  - 分かり易さを重視
  - 厳密性を少し欠く部分があります
  - 詳細は関連文献などを参照してください
- 質疑
  - なんでも質問してください
  - 初歩的な質問も大歓迎

導入：What is 深層学習？

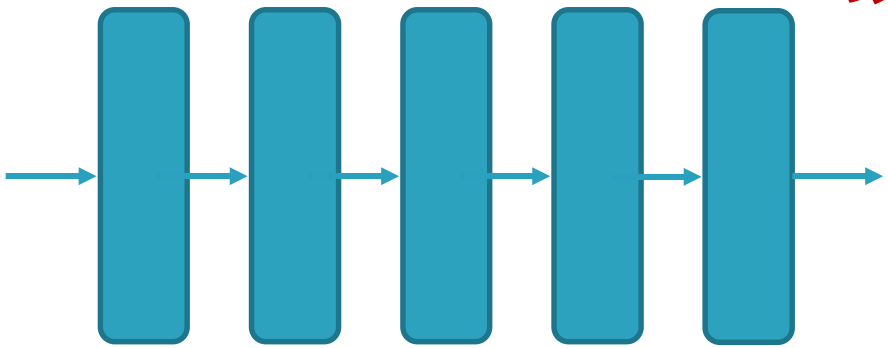
画像認識の場合（画素値→カテゴリID）

画像

カテゴリID

数値列

$$\begin{pmatrix} 0.7 \\ 1.7 \\ -1.6 \end{pmatrix}$$



数値

1

1. 犬
2. 車
:
N. 猫

深層学習（＝ニューラルネット）

（足し算，掛け算など）

画像認識（分類）ニューラルネット

画像認識の場合（画素値→カテゴリ確率）

画像

カテゴリID

各カテゴリの確率
数値列

数値列

$$\begin{pmatrix} 0.7 \\ 1.7 \\ -1.6 \end{pmatrix} \rightarrow \begin{pmatrix} 0.8 \\ 0.0 \\ \vdots \\ 0.2 \end{pmatrix}$$

1. 犬
2. 車
:
N. 猫

各カテゴリの確率も予想可能！

それでもやっぱり数値列→数値列

ここまでの重要なポイントのおさらい

- 深層学習（＝ニューラルネット）
  - 結局のところ数値列変換機構

    i.e., 数値列　→　数値列
  - 訓練：入力数値が所望の数値になるよう調整

用語の整理

- 深層学習
  - 総称的（理論，手法全体，．．．）
- ニューラルネットワーク
  - ネットワークそのもの（ネットワーク図）
  - ニューラルネット，ニューラルモデルとも
- ベクトル
  - 数値列（数値一個も含む）

基礎：深層学習の言語データへの適用

大きな壁：言語データは数値でない

- 単語，文，文書，色々あるけれど
  - 数値でない
  - カテゴリカルなデータ
- 深層学習
  - 数値列変換機構

言語データは深層学習で扱えない？

結論からいうと：単語の数値列化で対応

- 単語を数値列（=ベクトル）に変換！
  - 従来：単語は記号（例：Japan≠France）
  - 現在：単語を数値列（単語ベクトル）に

似た単語は似た数値列に

実際は数百～数千の数値の列

Tokyo Japan France Paris

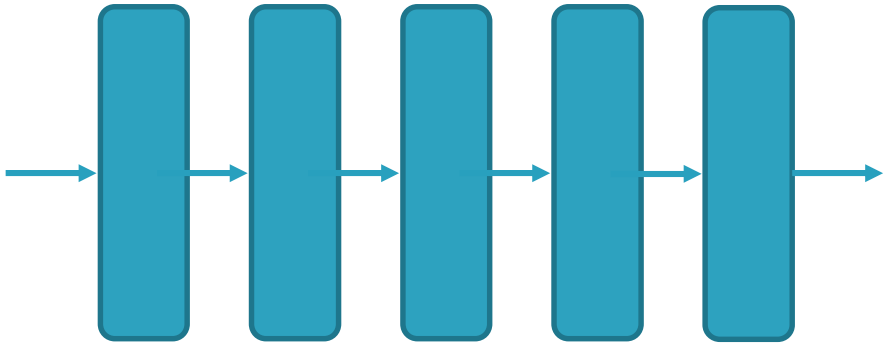
$$\begin{pmatrix} 1.0 \\ -1.2 \end{pmatrix} - \begin{pmatrix} -0.3 \\ 1.0 \end{pmatrix} + \begin{pmatrix} -0.5 \\ 0.7 \end{pmatrix} = \begin{pmatrix} 0.8 \\ -1.5 \end{pmatrix}$$

打開策：ヒントはここにあり

画像

カテゴリID

$$\begin{pmatrix} 0.7 \\ 1.7 \\ -1.6 \end{pmatrix}$$



1. 犬
2. 車
:
N. 猫

単語に数値（ID）を割り振っているのと同じ！

まずは，全単語に数値（ID）を割り振ふる！

単語（記号）→one-hotベクトル

- まずは全単語に適当に番号を振る
  - 例：アルファベット順に1, 2, ..., *N*

    a 1, *an* 2, *apple* 3, ..., *zoo* *N*
- one-hotベクトルに
  - 番号に対応する要素を1に，それ以外を0に
  - 例：単語anの場合

    ベクトルの2番目の値を1に

    それ以外を 0に

$$\left.\begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}\right\} N\text{個の } 0 \text{ or } 1$$

One-hotベクトル→単語ベクトル by 穴埋め問題

- 単語穴埋め問題

  I was cooking (?) in the kitchen.

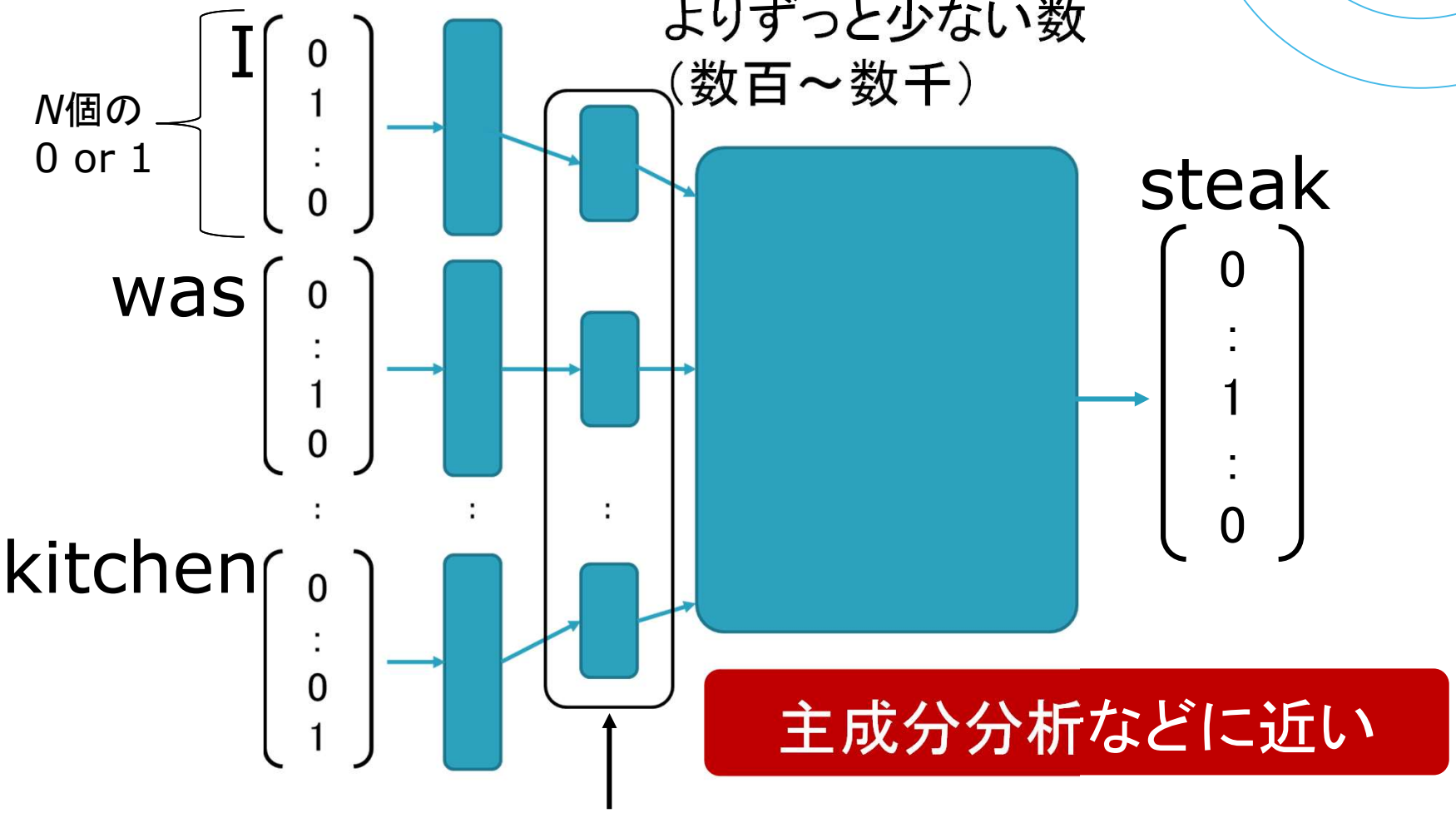
  steak

  - 入力：穴あき文章（one-hotベクトルの集合）
  - 出力：穴に入る単語（one-hotベクトル）
- 訓練データ（問題）の生成は簡単！

  1. 適当な文章を収集（例：webページ）
  2. ランダムに穴にする単語を選ぶ

穴埋め問題を解くニューラルネット

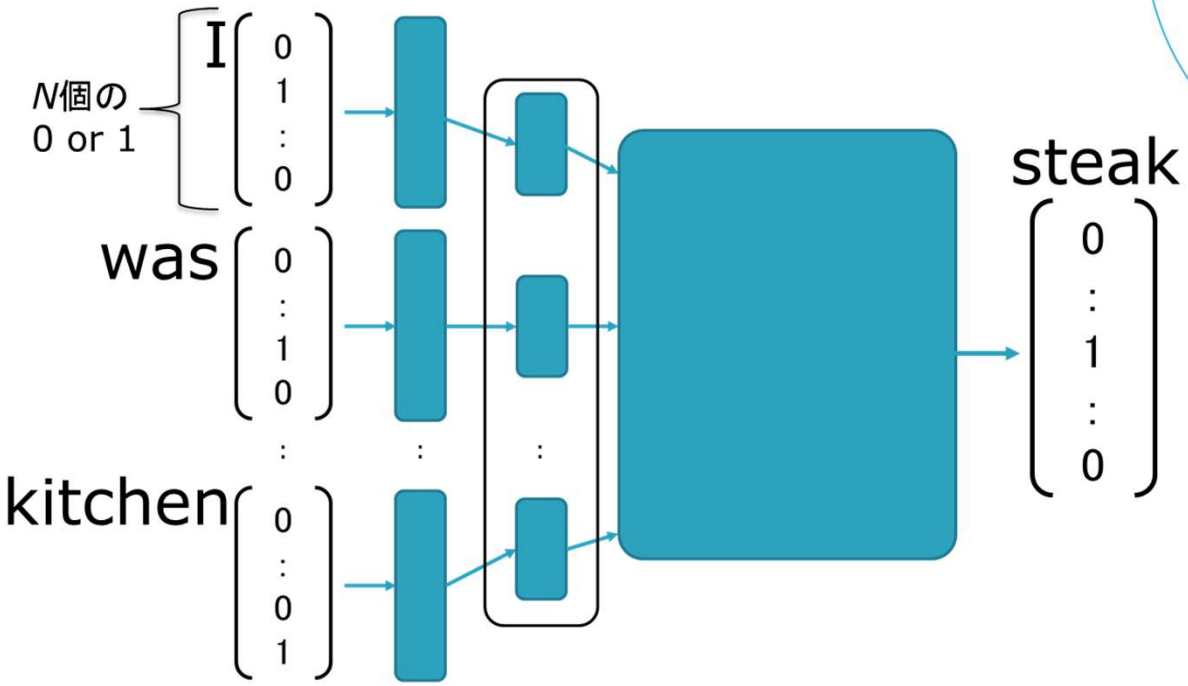
I was cooking (?) in the kitchen.



主成分分析などに近い

この部分を単語ベクトルとして使用(他の方法もあり)

実はこれが. . .



言語モデル(Language Model; LM)です！

Large LMは青い部分が大きい(計算が複雑)

言語分析で重要なのは

- 単語ベクトル
  - 言語モデルの副産物
  - (単語)分散表現とも
- 文ベクトル，文書ベクトル
  - 単語ベクトルと同様は方法で
- 言語モデルそのもの
  - 言語データ，言語現象のモデルとして
  - 人間の言語能力のモデルとして

ここまでの重要なポイントのおさらい

- 深層学習（＝ニューラルネット）
  - 結局のところ数値列変換機構
- 言語データへの適用
  - 単語も数値列に変換（e.g., by 言語モデル）
- 言語モデル
  - 穴埋め問題ソルバー，単語予測器
  - 副産物の単語／文／文書ベクトル

基礎：深層学習の言語データへの適用
強力なツール：単語ベクトル

深層学習の言語分析への応用例

- 分析ツール
  - 曖昧検索，自動グルーピング（クラスタリング）
- 言語データ／言語現象のモデル化
  - 各種現象の定量化（e.g, 意味変化の度合い）
- 人間（の言語能力）のモデルとして
  - （今回はあまり触れません）

単語や文が数値列になっているので
なんとなくイメージできる？

（文脈付き）単語ベクトル

- 単語トークンごとにベクトルを
  - コーパス中の各事例に応じたベクトル

    a river *bank* vs. a *bank* account
- 様々な呼び名があります
  - 文献を読む際には注意
- 代表的な実装例
  - BERT（から得られるベクトル）
  - GPT （から得られるベクトル）

ベクトルと類似度計算

- ベクトルは空間（平面）上の矢印
  - 類似度：二つのベクトル間の角度
    （cos類似度）

- 深層学習により単語，句，文，文書は
  - ベクトルに！

応用例：曖昧用例検索（e.g., bank）

コーパス

①対象単語の用例抽出

I have a bank account in Japan
They were sitting on the bank.
⋮

検索対象単語

He drew some bucks from the bank.

言語モデル

言語モデル

②文脈付きベクトル化

③文脈付きベクトル化

$$\begin{pmatrix} 0.5 \\ \vdots \\ 0.2 \end{pmatrix} \begin{pmatrix} 2.5 \\ \vdots \\ 1.3 \end{pmatrix} \cdots \begin{pmatrix} 0.7 \\ \vdots \\ 0.3 \end{pmatrix}$$

④類似度計算

⑤高類似事例出力

$$\begin{pmatrix} 0.6 \\ \vdots \\ 0.3 \end{pmatrix}$$

曖昧用例検索デモ

応用例：用例の自動グルーピング

- 既存のクラスタリング手法を利用可能
  - 単語，文などベクトルだから！
- 文脈付き単語ベクトル
  - 単語トークンのグルーピング

  i.e., 表層は同じだが似た意味／用法の事例群

同様なことが文，文書でも可能

自動グルーピング（対象単語：apple）

深層学習で見えるもの

歴史コーパス

①appleの事例抽出

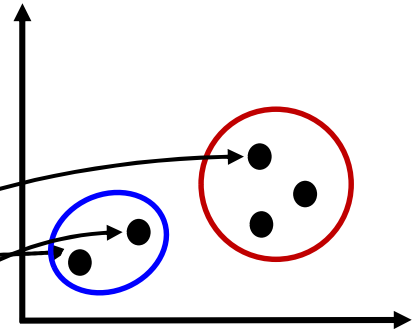
1823年: Apple trees are everywhere.
2017年: I bought an Apple PC.
⋮

言語モデル

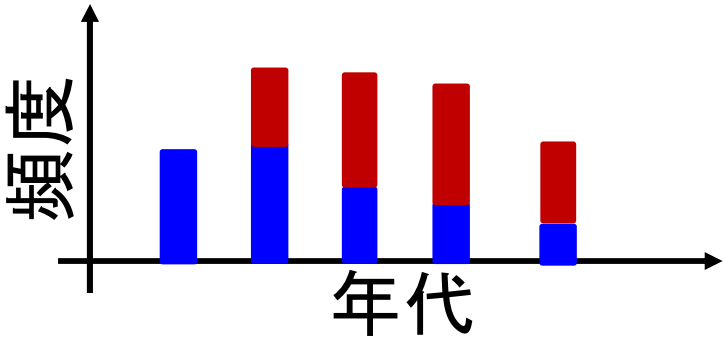
②文脈付きベクトル化

$$\begin{bmatrix} 0.5 \\ \vdots \\ 0.2 \end{bmatrix} \begin{bmatrix} 2.5 \\ \vdots \\ 1.3 \end{bmatrix} \cdots \begin{bmatrix} 0.7 \\ \vdots \\ 0.3 \end{bmatrix}$$

③自動グルーピング
（クラスタリング）



④グラフ化



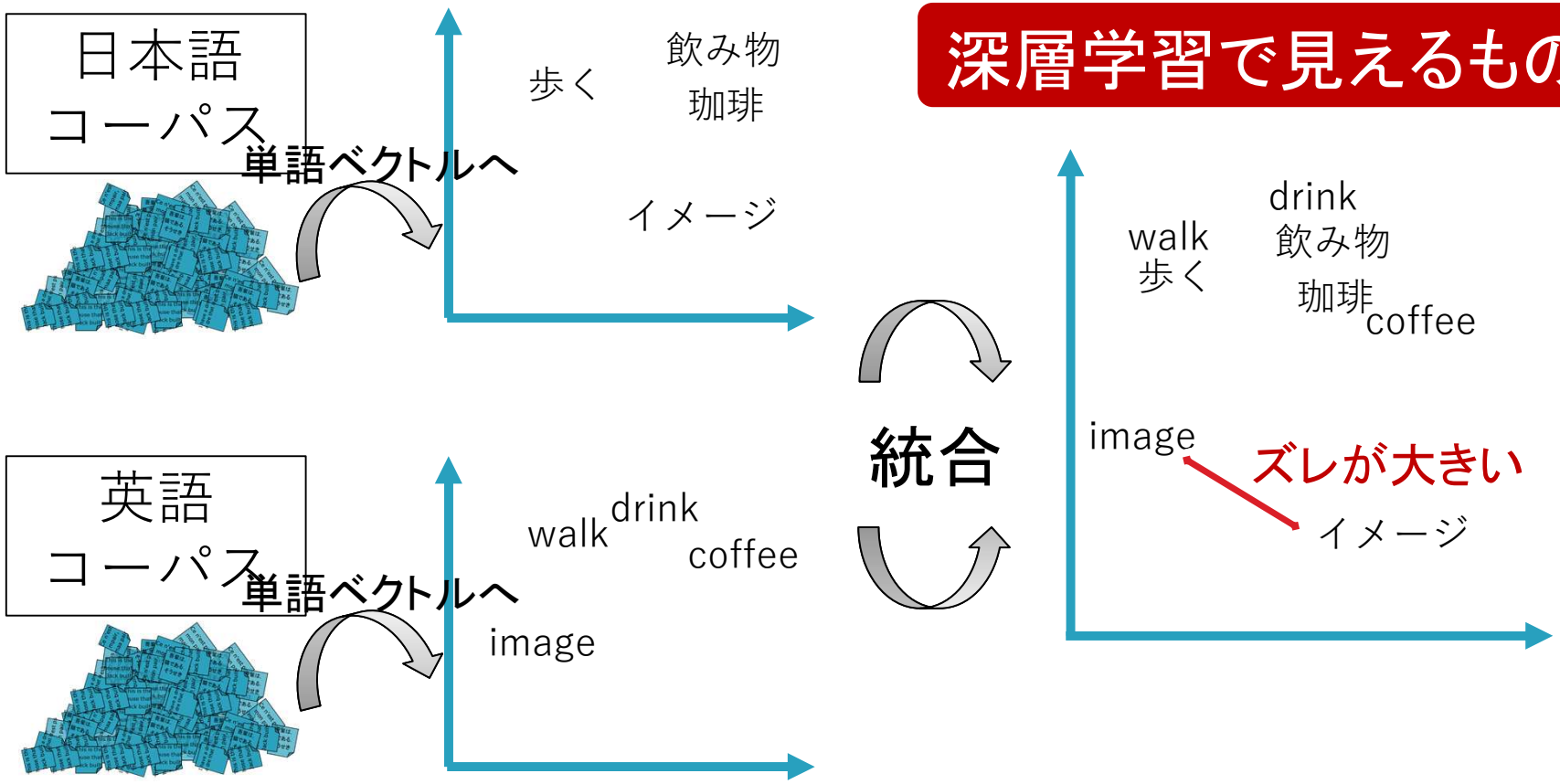
応用:
深層学習ベースの言語分析の実例

実例：意味の差異検出

- 外来語の意味変化に対する推理的分析
  - 高村，永田，川崎
  - 言語処理学会年次大会2017

- サマリ
  - 対象：カタカナ語（日本語），英語
  - 両者における意味差異の定量化
  - 単語ベクトルを利用

日本語／英単語の意味のズレの検出



自動検出したズレの大きいカタカナ語

ズレ度（0～1）	カタカナ語	英語
0.85	スプレッド	spread
0.70	プレゼント	present
0.67	コーディネート	coordinate
0.64	アッセンブリー	assembly
0.63	ソート	sort
0.61	ダイジェスト	digest
	⋮	⋮
0.52	イメージ	image
	⋮	⋮

文脈付き単語ベクトルの言語分析への応用例

意味／用法の年代変化の可視化

better off構文を題材にして

実例：意味変化の可視化による仮説の吟味

- 言語処理的アプローチによる
  better off構文の定着過程の説明
  - 永田，大谷，高村，川崎
  - 言語処理学会年次大会2022
- サマリ
  - 対象：better off構文
  - 通時的意味変化の可視化と仮説の吟味
  - 単語ベクトル＆クラスタリングを利用

better offの用法変化（大谷 2022）

Stage 1

旧用法
better off
（経済的に）より良い

He's better off than before.
以前より裕福だ.

Stage 2

適用範囲の拡張
経済状況以外にも

He's better off here.
ここにいた方がよい.

Stage 3

仮想状況と想定　増
Stage 2で適用範囲
が拡張したため

He would be better off with her.

後続句を補語と再分析

Stage 4

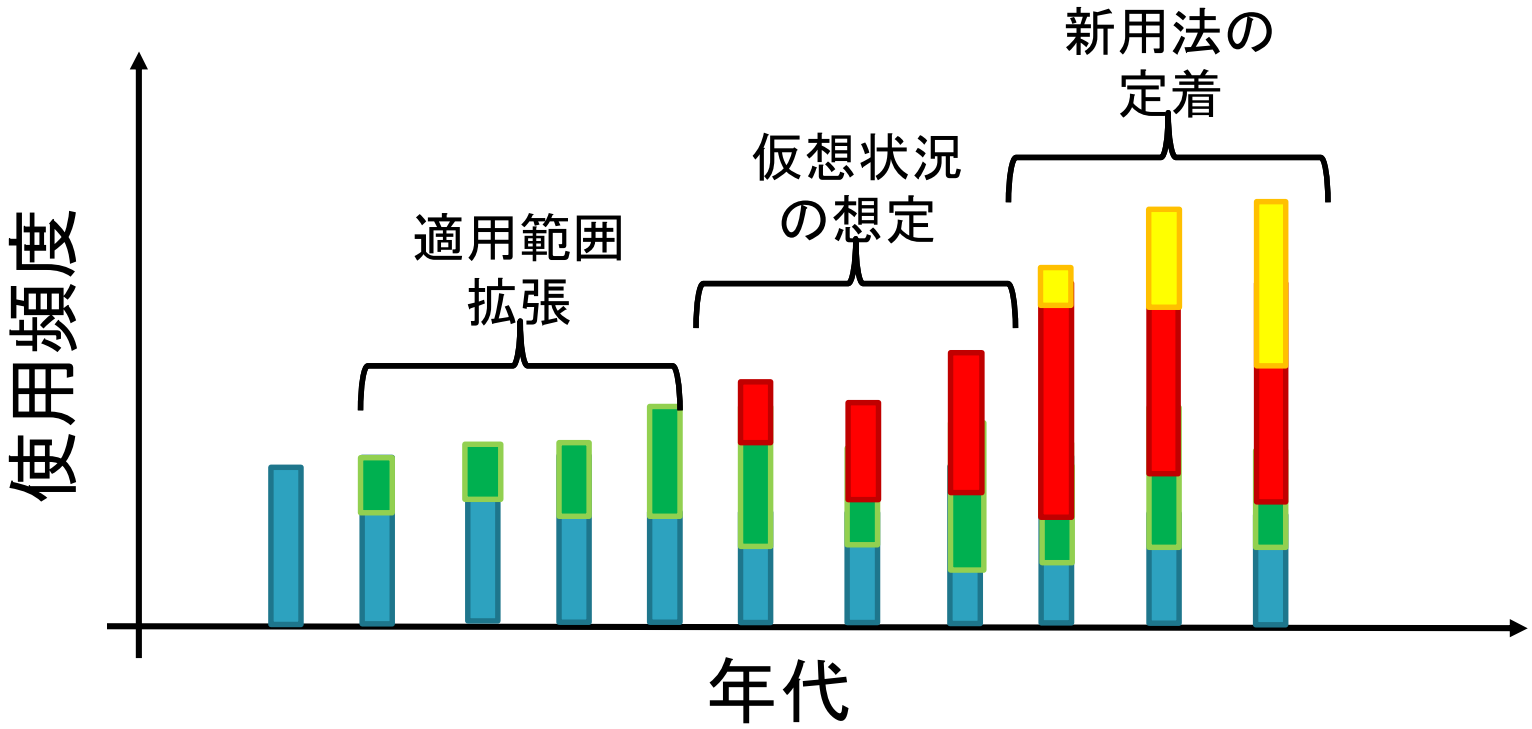
新用法
better off
～した方がましだ

He's better off dead.
死んだ方がましだ.

Stage 1 → 2 → 3 → 4と使用（数）が変化

仮説を図示すると

自然言語処理により自動的に作図



手法の流れ

歴史コーパス

①better offの事例抽出

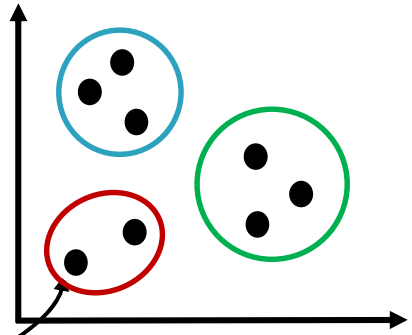
1823年: He is better off than before.
1977年: You are better off buying it.
:

言語モデル(BERT)

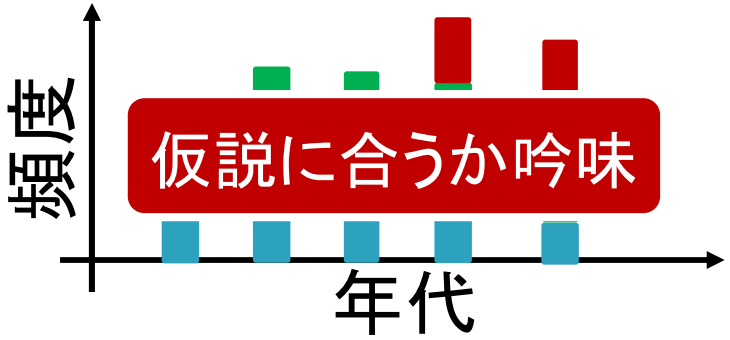
②文脈付きベクトル化

$$\begin{pmatrix} 0.8 \\ : \\ -1.5 \end{pmatrix}_{\text{better off}}$$

③自動グルーピング（クラスタリング）

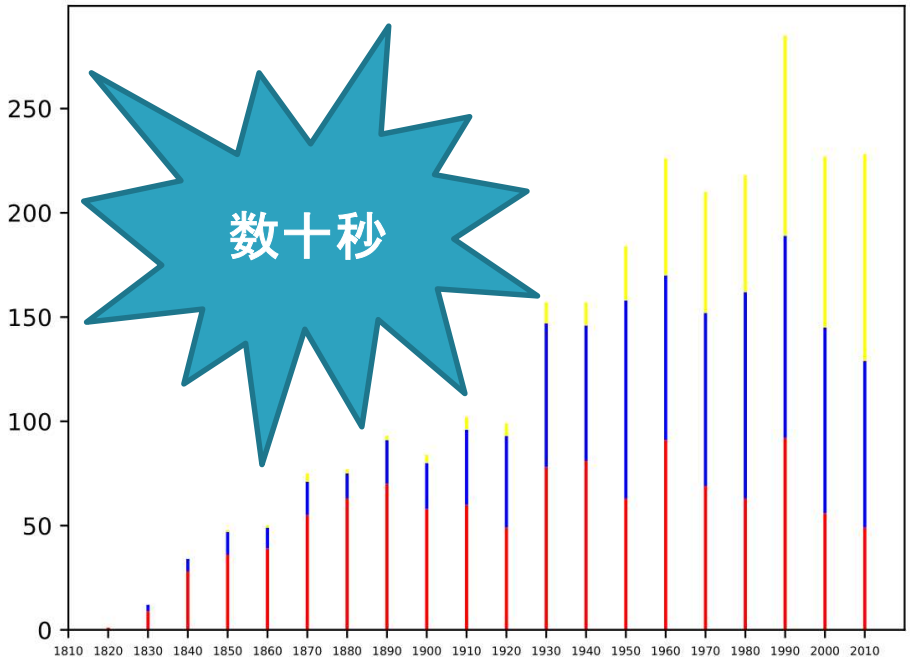


④グラフ化

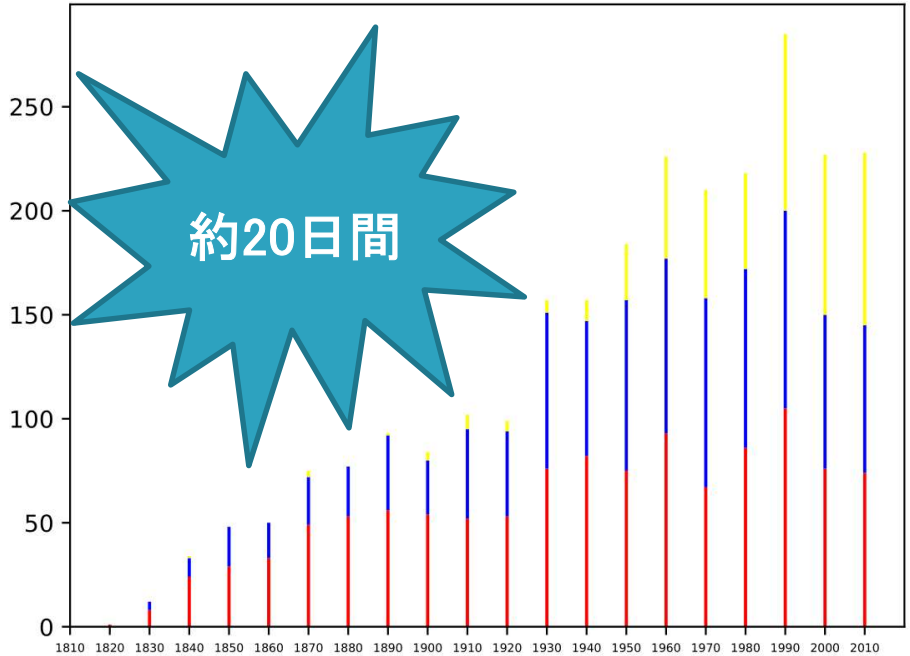


具体例:フレーズ better offの意味変化

自動分類



人手で作成
（2591事例を分類）



分類精度は人にかなわないが時間短縮は魅力的

仮説の生成，サポートになりそう

ツールとしての単語／文／文書ベクトル

- 用例の自動グルーピング
- 用例の検索
  - 意味，用法の類似性も考慮
  - キーワードを含まない類似例も検索可

    例：He ate paella *v.s* He devoured morcilla
- 類似度の定量化
  - 現在：連続値（e.g., −1 ～ 1）

この辺りが言語分析，仮説生成／確認に有益？

クロージング：関連情報案内

文献案内

- 自然言語処理の基礎
  - 岡崎，荒瀬，鈴木，鶴岡，宮尾著
  - モダンな自然言語処理を丁寧に解説
  - 深く知りたい人向け（専門的）
  - 本講演に関連する内容
    - 単語のベクトル化
    - 言語モデル

文献案内：単語の数値（ベクトル）化

- 英語教育における自動採点
  - 5章　深層学習に基づいたエッセイの自動採点（永田担当）
  - 自由記述作文の自動採点を題材に
  - 単語のベクトル化方法を文系の方向けに解説

文献案内

- 深層学習の原理に迫る
  - 今泉著
  - 深層学習がなぜうまくいくか分かり易く解説
  - 「数学の挑戦」とあるが数学の知識は不要

イベント案内

- 英語学会第41回大会シンポジウム
  - 「深層学習時代の言語研究」
  - 11月4日東大駒場キャンパス
  - 永田，川崎他4名の共同講演
  - 応用（研究事例）よりの話を
  - （今回あまり話せなかった）
    人間（言語能力）のモデルとしての深層学習

イベント案内

- 言語処理学会年次大会NLP2024
  - 2024年3月11日～15日
  - 神戸国際会議場
  - 自然言語処理の最新研究満載
  - 文系の方の参加も多数
  - NLP×言語学に関連した企画も検討中

イベント案内

- 言語処理学会言語処理セミナー
  - 30周年記念特別編
    「人文系研究者のための言語処理入門」(仮)
    (hands-onを含む講習会)
  - 永田＆川崎で実施予定
  - 2024年秋に予定
  - 詳細は現在検討中

まとめ

- 深層学習とは
  - 深層学習≒数値変換機構
- 深層学習と言語処理
  - 深層学習で扱えるように言語データを数値化
  - その結果得られる数値列（ベクトル）が重要
- 深層学習ベースの言語分析
  - 曖昧検索，自動グルーピング（クラスタリング）
  - 言語データ，言語現象，人間のモデル化

For further information

- 気軽にお問い合わせください
  - 質問，講演依頼，研究相談，etc.
  - nagata-jaecsqa ＠ ml.hyogo-u.ac.jp.

（講師二人に届きます）