言語処理学会主催（30周年記念特別編）
言語処理技術セミナー2024
「言語分析のための言語処理・深層学習」

応用1: word2vecベースの言語分析

2024年8月30日（金）

甲南大学ネットワークキャンパス東京
永田 亮（甲南大学／理研）
川崎 義史（東京大学）
内田 諭（九州大学）

アウトライン

- word2vecとは？
- word2vecの応用例
- ハンズオン

アウトライン

- word2vecとは？
- word2vecの応用例
- ハンズオン

単語（タイプ）のベクトル表現

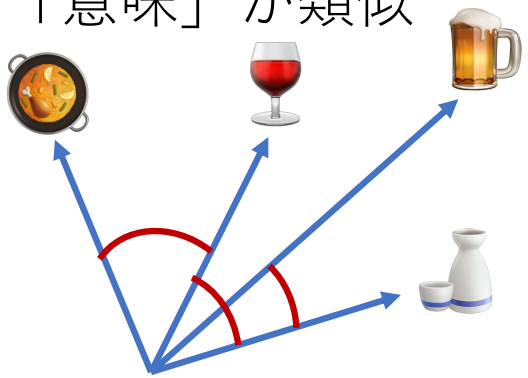
word2vec

他の呼称
分散表現（distributed representation）
単語埋め込み（word embeddings）

「単語の意味は周囲の単語によって形成される」という**分布仮説**（distributional hypothesis）に依拠（Harris 1954; Firth 1957）
→同様の文脈に出現する単語同士は「意味」が類似

バルで**ワイン**を飲む
パブで**ビール**を飲む
居酒屋で**日本酒**を飲む
レストランで**パエリア**を食べる
・・・

コーパスから学習



余弦類似度**大**（ベクトルの向きが同じ）＝「意味」が類似
余弦類似度**小**（ベクトルの向きが違う）＝「意味」が異なる

$-1 \leq$ 余弦類似度 ≤ 1

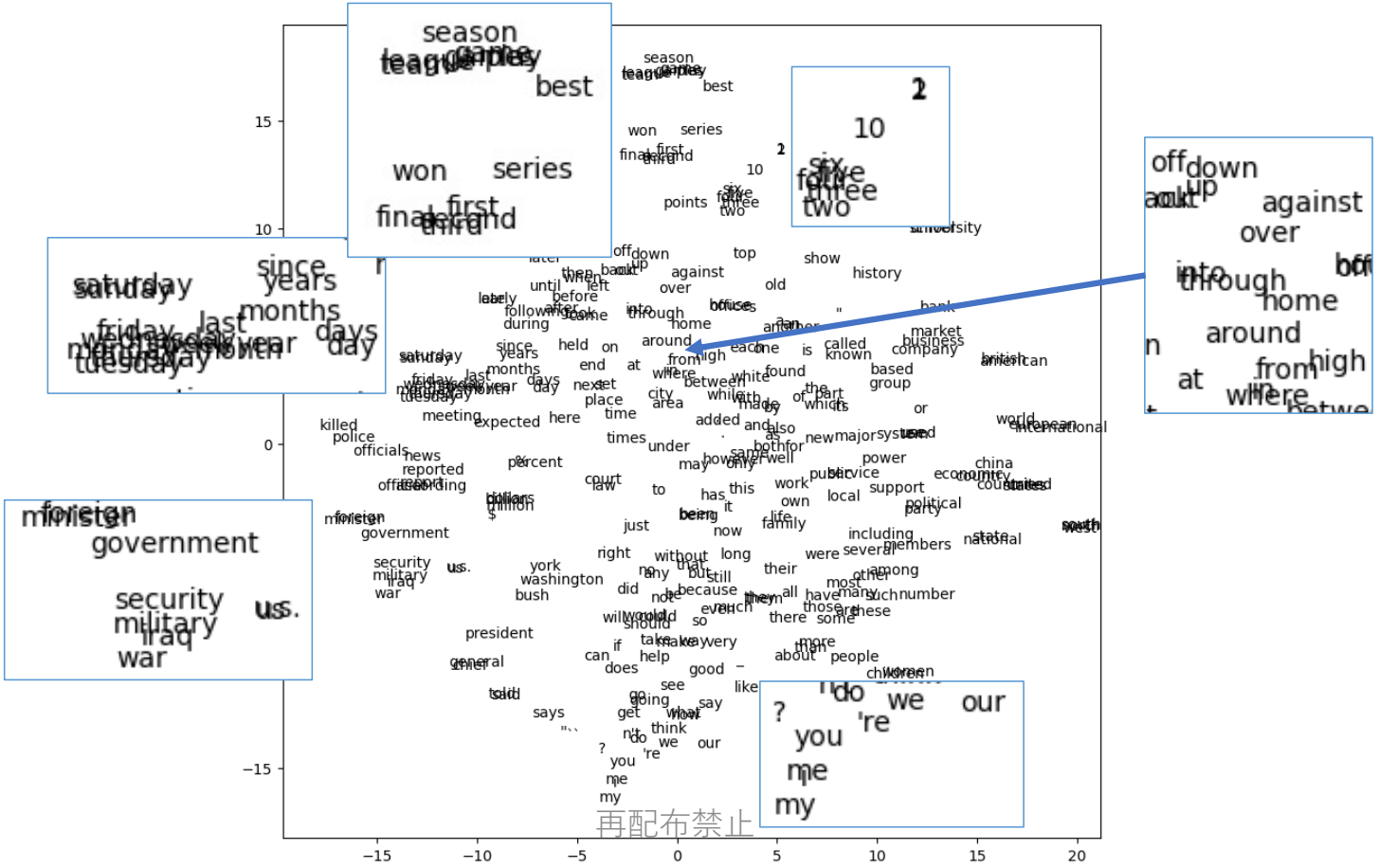
appleの単語ベクトル（50次元）

```
array([ 0.52042 , -0.8314  ,  0.49961 ,  1.2893  ,  0.1151  ,  0.057521,
       -1.3753  , -0.97313 ,  0.18346 ,  0.47672 , -0.15112 ,  0.35532 ,
        0.25912 , -0.77857 ,  0.52181 ,  0.47695 , -1.4251  ,  0.858   ,
        0.59821 , -1.0903  ,  0.33574 , -0.60891 ,  0.41742 ,  0.21569 ,
       -0.07417 , -0.5822  , -0.4502  ,  0.17253 ,  0.16448 , -0.38413 ,
        2.3283  , -0.66682 , -0.58181 ,  0.74389 ,  0.095015, -0.47865 ,
       -0.84591 ,  0.38704 ,  0.23693 , -1.5523  ,  0.64802 , -0.16521 ,
       -1.4719  , -0.16224 ,  0.79857 ,  0.97391 ,  0.40027 , -0.21912 ,
       -0.30938 ,  0.26581 ], dtype=float32)
```

- 実数値が50個並んだもの（50次元空間内の一点）
- 各次元が何を表しているかは人間には解釈不可能

単語ベクトルを50次元から2次元に圧縮

t-SNEによる単語ベクトルの可視化



類似単語検索

appleの類似単語上位10件

```
[('blackberry', 0.7543067336082458),
 ('chips', 0.7438644170761108),
 ('iphone', 0.742966413497924 8),
 ('microsoft', 0.7334205508232117),
 ('ipad', 0.7331036329269409),
 ('pc', 0.7217225432395935),
 ('ipod', 0.7199784517288208),
 ('intel', 0.7192243337631226),
 ('ibm', 0.7146540284156799),
 ('software', 0.7093585133552551)]
```

applesの類似単語上位10件

```
[('peaches', 0.8623535633087158),
 ('oranges', 0.8594476580619812),
 ('cherries', 0.8461860418319702),
 ('mangoes', 0.8264981508255005),
 ('apricots', 0.8242633938789368),
 ('strawberries', 0.8229067921638489),
 ('potatoes', 0.8179376125335693),
 ('melons', 0.7980057597160339),
 ('berries', 0.794605016708374),
 ('vegetables', 0.7920521497726 44)]
```

果物の🍎がない…
word2vecは多義語の扱いが苦手
→BERTによるトークンレベルの文脈付き単語ベクトル

類推（analogy）

単語ベクトルの足し算・引き算で意味計算が可能

意味的類推　$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} = \overrightarrow{queen}$

「支配者」ベクトル

文法的類推　$\overrightarrow{kings} - \overrightarrow{king} + \overrightarrow{queen} = \overrightarrow{queens}$

「複数形のs」ベクトル

Mikolov et al. 2013

アウトライン

- word2vecとは？
- word2vecの応用例
- ハンズオン

(Hamilton et al. 2016)

Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change

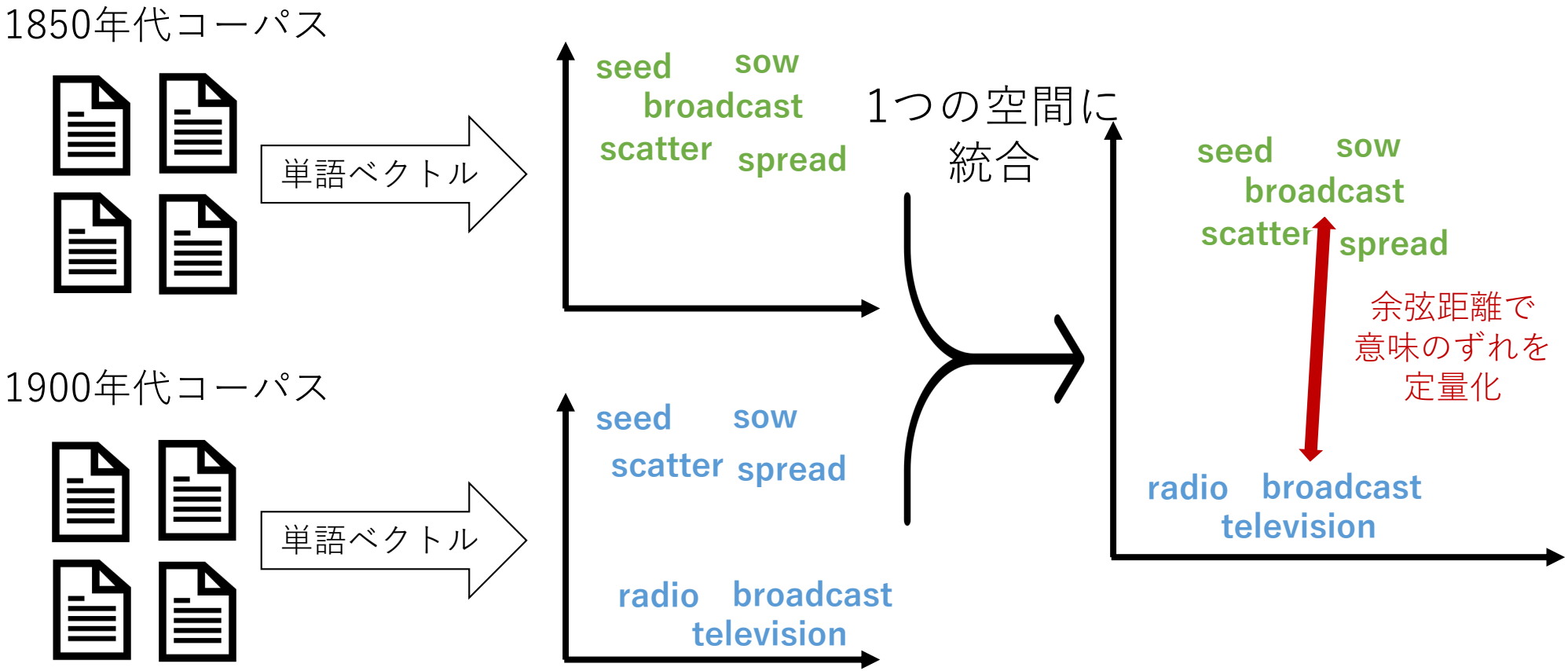
William L. Hamilton, Jure Leskovec, Dan Jurafsky
Department of Computer Science, Stanford University, Stanford CA, 94305

研究の概要

- 意味変化の統計的法則を提示
  - 高頻度語ほど意味変化の度合いが小さい
  - 多義語ほど意味変化の度合いが大きい
- **同一単語の二時点における単語ベクトル間の余弦距離（類似度）を用いて，意味変化の度合いを定量化することに成功**

  →通時的意味変化に関する計算言語学的研究が活性化！

時点間単語ベクトル



実験結果

- 対象言語：英仏独中
- 対象期間：1800－2000年
- 同一単語の二時点における単語ベクトル間の余弦距離を目的変数，頻度と多義性を説明変数として回帰分析

→全言語で，意味変化の統計的法則が成り立つことを示唆

  - 高頻度語ほど意味変化の度合いが小さい
  - 多義語ほど意味変化の度合いが大きい

（川崎ら 2024）

言語処理学会 第30回年次大会 発表論文集（2024年3月）

意味変化の統計的法則は 1000 年成り立つ

川崎義史 [1]　高村大也 [2]　永田亮 [3]
[1] 東京大学　[2] 産業技術総合研究所　[3] 甲南大学

研究の概要

ロマンス語
俗ラテン語から派生した言語の総称
🇮🇹🇫🇷🇪🇸🇵🇹🇷🇴

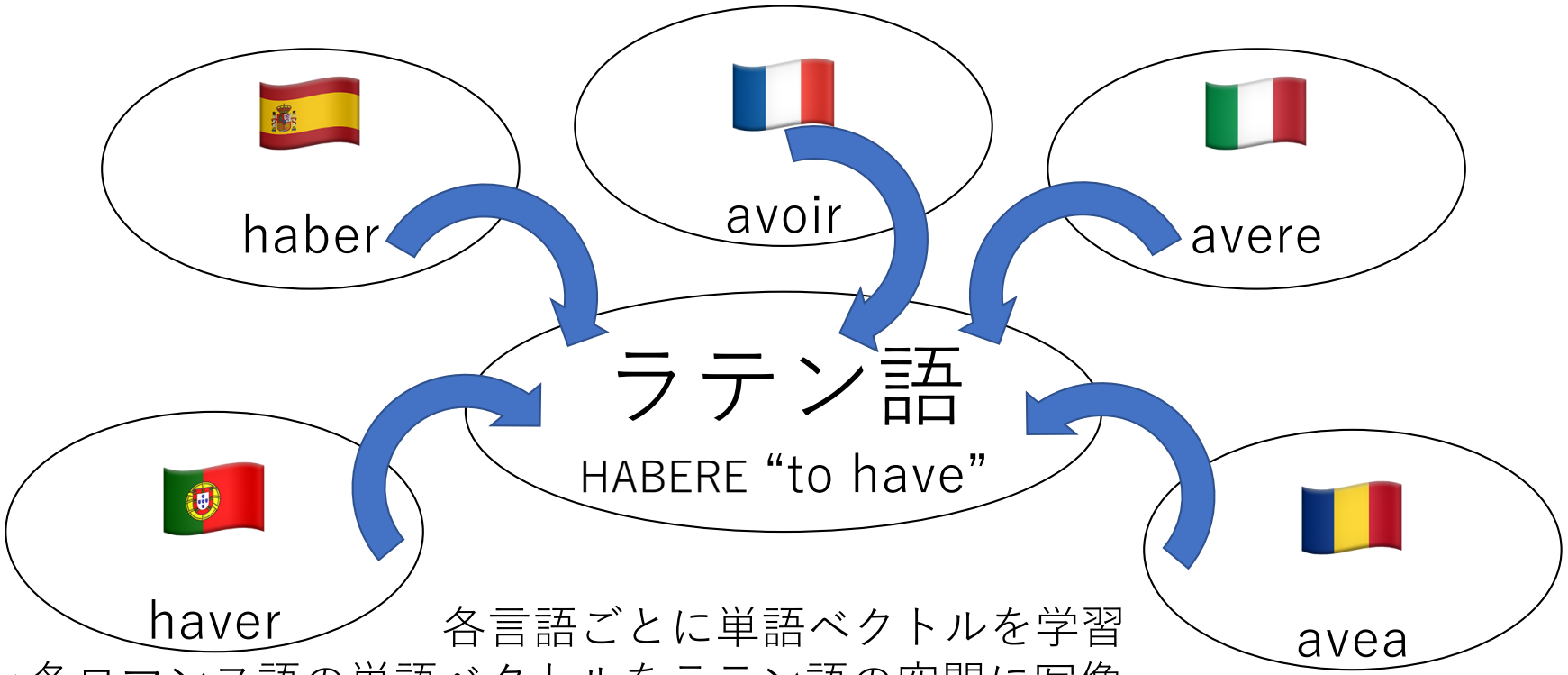
- 意味変化の統計的法則が，長期にわたり（先行研究の直近200年を超えて）成り立つか，ラテン語とロマンス語（仏伊西葡羅）の聖書を用いて統計的に分析

→長期に渡り成り立つことが示唆される

表 2　使用した聖書の翻訳年，トークン数，語彙サイズ

言語	翻訳年	トークン数	語彙サイズ
ラテン語	400 年頃	570K	13K
フランス語	1776 年	811K	13K
イタリア語	1649 年	710K	27K
スペイン語	1569 年	747K	19K
ポルトガル語	1751 年	740K	23K
ルーマニア語	1928 年	752K	20K

多言語単語ベクトル



各言語ごとに単語ベクトルを学習
→各ロマンス語の単語ベクトルをラテン語の空間に写像
→同一空間内でラテン語語源とロマンス語形の意味のずれを余弦距離で測る

実験結果

- ラテン語語源とロマンス語形の単語ベクトル間の余弦距離を目的変数，ラテン語語源の頻度と多義性を説明変数として回帰分析

→全ロマンス語で，意味変化の統計的法則が1000年以上成り立つことを示唆

表7　ラテン語語源とスペイン語形の余弦距離を従属変数とした回帰分析の結果 ($N = 147, Adj.R2 = 0.18$)

	Coef.	SE	t	$p > \|t\|$
Intercept	0.00	0.08	0.00	1.00
fr_{lat}	**-0.37**	0.08	-4.72	< 0.01
$poly_{lat}$	**0.37**	0.08	4.66	< 0.01

高頻度語→意味変化小
多義語→意味変化大

参考文献

- Firth, J. (1957). A Synopsis of Linguistic Theory 1930–1955. *Studies in Linguistic Analysis*. Oxford Philological Society. 1–32.
- Harris, Z. S. (1954). Distributional Structure. *Word*. 10: 2–3, 146–162.
- Mikolov, T., Yih, W., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. https://aclanthology.org/N13-1090/
- Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1489–1501. https://aclanthology.org/P16-1141/
- 川崎義史, 高村大也, 永田亮. (2024). 「意味変化の統計的法則は 1000 年成り立つ」. 『言語処理学会 第 30 回年次大会 発表論文集』, 1610–1615. https://www.anlp.jp/proceedings/annual_meeting/2024/pdf_dir/E6-2.pdf

アウトライン

- word2vecとは？
- word2vecの応用例
- ハンズオン

ハンズオンの流れ

- word2vecの使い方
  - データ読み込み
  - t-SNEによる次元圧縮
  - 類似単語検索
  - 類推
- word2vecの学習
- おまけ：日本語のword2vecモデル