

# 計算言語学

## 単語の意味表現

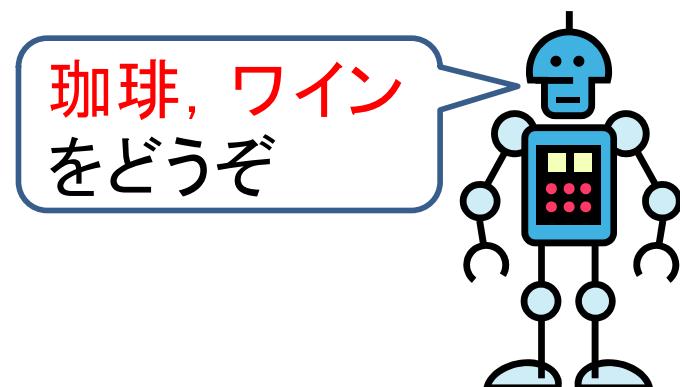
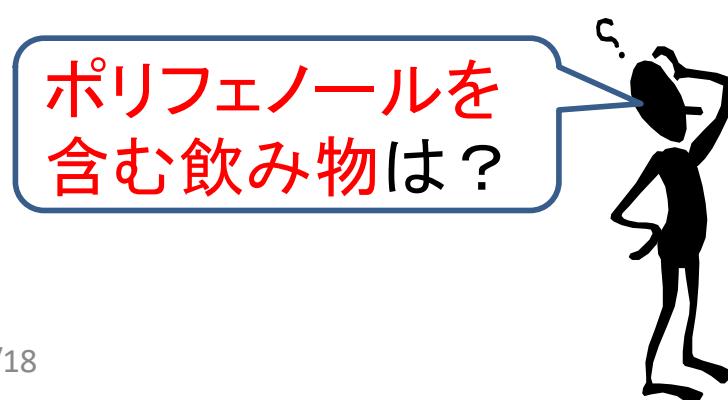
東京大学生産技術研究所

吉永 直樹

site: <http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/class/cl/>

# 自然言語の意味をどう扱うか？

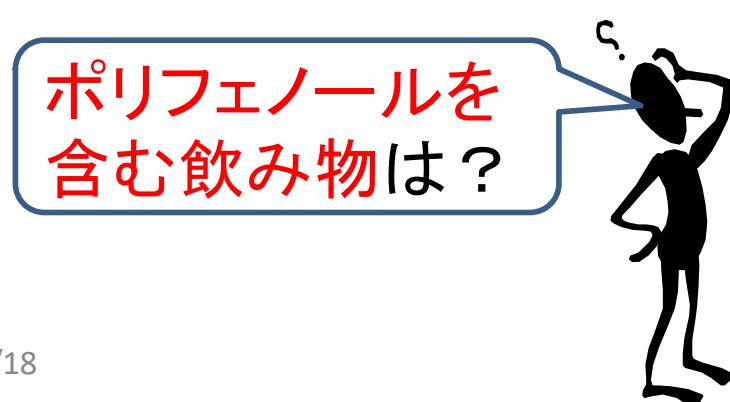
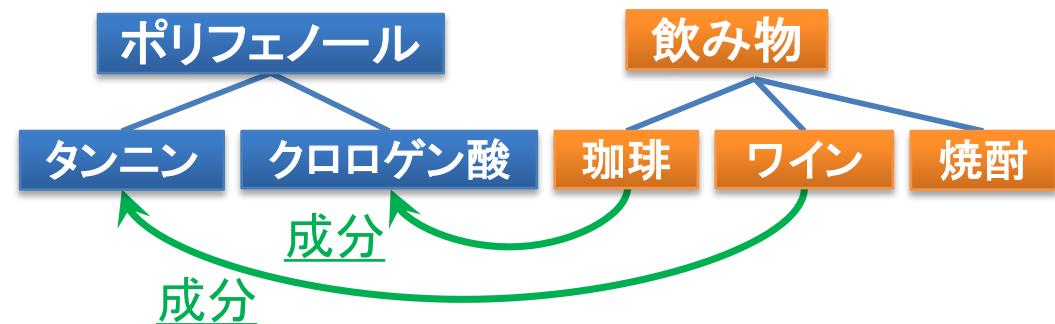
- 言語では膨大な種類の語を組み合わせて文を構成
  - Google n-gram での異なり 1-gram 数: 2,565,424
- 異なる言語表現間の意味的関係をどう捉えるか？



# 自然言語の意味をどう扱うか？

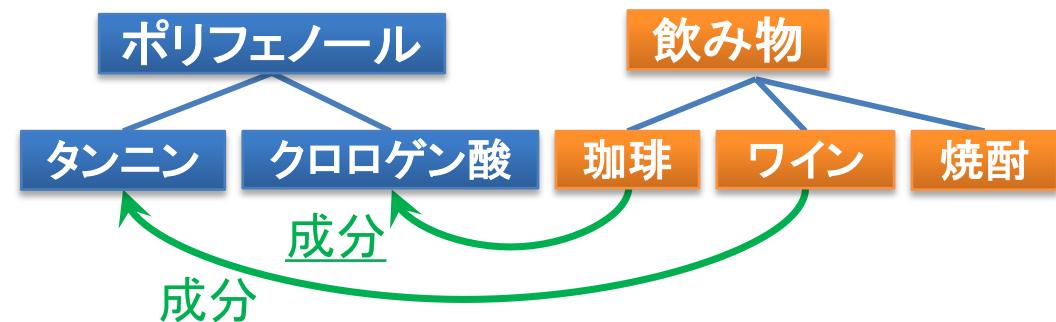
- 言語では膨大な種類の語を組み合わせて文を構成
  - Google n-gram での異なり 1-gram 数: 2,565,424
- 異なる言語表現間の意味的関係をどう捉えるか？

## 語の意味的な関係



# 自然言語の意味を扱うための基本方針

- 構成的意味論 (compositional semantics)
  - 仮定: 文の意味は部分(単語)から構成的に計算可能
  - 意味の合成方法は個別の単語の意味とは独立に規定  
例) ラムダ計算, 算術平均 など
- では単語の意味はどう表現(データ形式)するべき?
  - 異なる語の間の意味的な関係を分類・計算できる必要



# 語彙意味論 (Lexical semantics)

- 語の意味は記号(の組合せ)として離散的に記述
  - 辞書: lemma に対する語義定義

例) mouse (N)

- MOUSE<sub>1</sub>: any of numerous small rodents ...
- MOUSE<sub>2</sub>: a hand-operated device that controls cursor...

MOUSE<sub>1</sub>, MOUSE<sub>2</sub>が意味 (語義が複数 = 多義語)

- 語(義)間の関係は宣言的に分類・体系化して記述

# 多義語

- 多義語(狭義) (polysemy): 関連する(同一語源から派生した) 語義を複数持つ語

例)

mouse (, ) , newspaper (, ) , google

- 同綴異義語 (homograph): 関連しない(語源の異なる) 語義を複数持つ語

例)

bass (, ) 同綴異音異義語 (heteronym) 境界は曖昧  
bear (noun, verb), キリン (, ) 同綴同音異義語

(参考) 同音異綴語 (homophone): know/no, there/their,

# 語義間の関係 (1/6): 同義語

- 任意の文脈で文の意味を変えずに交換できる語  
文の真理条件(成立するかしないか)
  - 同義語 (狭義) (synonym): 異なる語が同一の語義を持つ  
例) car/automobile, buy/purchase, big/large  
*A lot of cars/automobiles are parking on the street.*
  - 異表記・表記ゆれ (spelling variation): 同一の語の異表記  
例)

ジョブス / Jobs	翻字 (transliteration)
重要文化財/重文	略記, 頭字語 (acronym)
バイオリン/ヴァイオリン	翻字時の音の転記揺れ
銀杏/イチョウ	字種の変換
realize/realise	

## 語義間の関係 (2/6): 反義語

- 性質, 動作, 関係が反対となる語

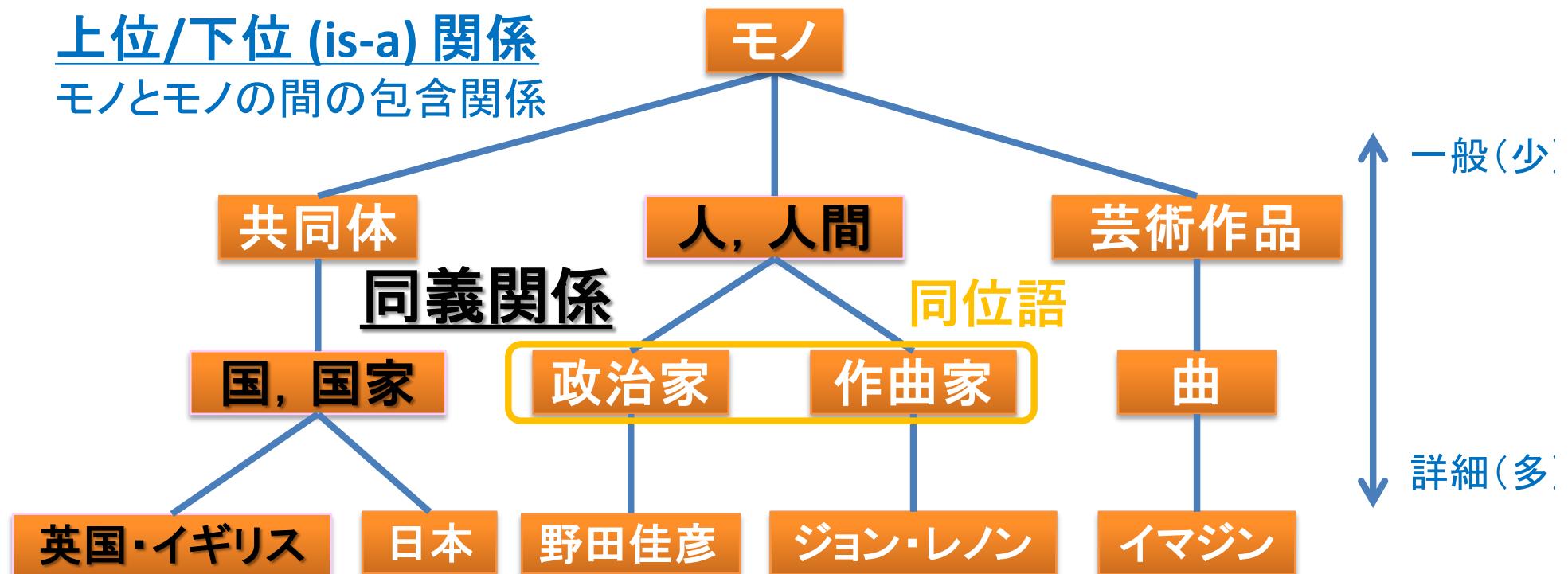
例)

long/short, big/little	性質(の程度)の反転
buy/sell, in/out	動作の反転
parent/child	関係の反転

- 反転させている性質, 動作, 関係を除き同じ意味を持つため同義語との区別が(計算機には)難しい

# 語義間の関係 (3/6): 上位語・下位語

- 語義に包含関係のある語 (is-a 関係, hyponymy)
  - 言語テスト ('A is a (kind-of) B'と言えるか) で判定可能
  - 上位下位関係は推移律が成り立つ  $\forall x A(x) \Rightarrow B(x)$
  - 下位語から上位語への置換は文の真理条件を維持する



## 語義間の関係 (4/6): 部分語・全体語

- 物理的に包含関係にある語 (part-of 関係, meronymy)
  - 言語テスト ('A is a part/member of B') により判定可能

例)

engine – car, finger – hand, leaf – tree

- 部分全体関係は推移律が成り立つ
- 部分語から全体語への置換は文の真理条件を必ずしも維持しない

# 語義間の関係 (5/6): 類義語

- 同義 or 同義とまでは言えないが似た語義を持つ語
  - 同義・反義・上位・下位語など  
例) cats - dogs, travel - go
- SimLex-999 [Hill+ 2015]
  - 複数被験者を用いて単語の類似度を定量化

SimLex-999 [Hill+ 2015]

単語対		類似度
vanish	disappear	9.8
behave	obey	7.3
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

1単語対辺り約50人の被験者が [0-6] で類似度を付与した後、正規化・平均

被験者のスコアの相関は高め

$$\rho = 0.673$$

$$\rho = 1 - \frac{6 \sum D^2}{N^3 - N}$$

## 語義間の関係 (6/6): 関連語

- 類義語ではないが関連する語義をもつ語
  - 同じ意味領域 (semantic field) で組み合わせて使われる語  
domain, topic, etc.
- 例)
  - surgeon, nurse, hospital (病院)
  - waiter, menu, food, chef (レストラン)
  - door, kitchen, bed (家)
- WordSim-353 [Finkelstein+ 2002, Agirre+ 2012]
  - 複数被験者を用いて単語の関連度を定量化
  - 類義語が混在しており被験者のスコア間の相関は低め

$$\rho = 0.611$$

# WordNet [Miller+ 1993]

- 英語の意味辞書

<https://wordnet.princeton.edu/>

- 一般的な名詞、動詞、形容詞、副詞を網羅
- 語義 = 同義語集合 (synset) とし、synset 間で意味的関係を記述

**Noun**

- S: (n) **mouse** (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)
- S: (n) **shiner, black eye, mouse** (a swollen bruise caused by a blow to the eye)
  - *direct hypernym / inherited hypernym / sister term*
- S: (n) **mouse** (person who is quiet or timid)
- S: (n) **mouse, computer mouse** (a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad) "a mouse takes much more room than a trackball"

**Verb**

- S: (v) **sneak, mouse, creep, pussyfoot** (to go stealthily or furtively) "...instead of sneaking around spying on the neighbor's house"
- S: (v) **mouse** (manipulate the mouse of a computer)

例) mouse のエントリ

固有名詞や新語(義)は、ほとんどカバーされない

発展: 語彙概念構造 (lexical conceptual structure)  
[Jackendoff 1990]

- 動詞(句)のための意味記述体系
    - 動詞を少数のタイプ(例: 「移動」に関する動詞)に分類し, 基本的意味・概念的意味を記述
    - 動詞句の意味を抽象的な意味述語を用いて記述

例) 家に花を届ける [CONTROL [BECOME [花]<sub>y</sub> BE AT [家]<sub>z</sub>]]

家に花が届く [BECOME [花]<sub>y</sub> BE AT [家]<sub>z</sub>]]

家に花がある [[花]<sub>y</sub> BE AT [家]<sub>z</sub>]

# 発展: 生成語彙論 (generative lexicon)

[Pustejovsky, 1995]

- 単語の語義は出現する文脈ごとに厳密には異なり  
生成的な側面がある
  - 窓を通って侵入した
  - ビートルズが解散する / ビートルズを聴く
  - 良いナイフ $\doteq$ 切れるナイフ, 良い人 $\doteq$ 親切な人
- 単語(特に名詞)の意味を Quolia 構造で定義し, 意味を動的に生成

例)

QUOLIA =	$\begin{bmatrix} \text{kni}fe(x) \\ \text{CONST} = \{\text{metal}, \dots\} \\ \text{FORMAL} = \text{phyobj}(x) \\ \text{TELIC} = \text{cut}(P, w, x) \\ \text{AGENTIVE} = \text{manufacture}(x) \end{bmatrix}$	<p>構成物との関係 他の概念との区別 用途 準備(生成)</p>
----------	--	---

# 語の使用者と語義の関係のモデル化

喚情的意味 (affective meaning / connotation) [Osgood+ 1957]

- 単語を語が喚起する感情に基づく3因子で数値表現
  - **Valence**: 喚起される感情の明るさ
  - **Arousal**: 喚起される感情の強さ
  - **Dominance**: 喚起される感情の(人に対する)支配度

	valence	arousal	dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24
life	6.68	5.59	5.89

単語の意味を数値ベクトルで表現



# 単語の意味をどこから計算するか？

*“the meaning of a word is its use in the language”*

[Wittgenstein, 1953]

- 分布仮説 [Harris 1954, Firth 1957]

*“difference of meaning correlates with difference of distribution”*

*“You shall know a word by the company it keeps”*

*The small dog **barks** louder.*

*His dog runs fast.*

*Eyes of the dog was very small.*

*Foxes are **barking** in the distance.*

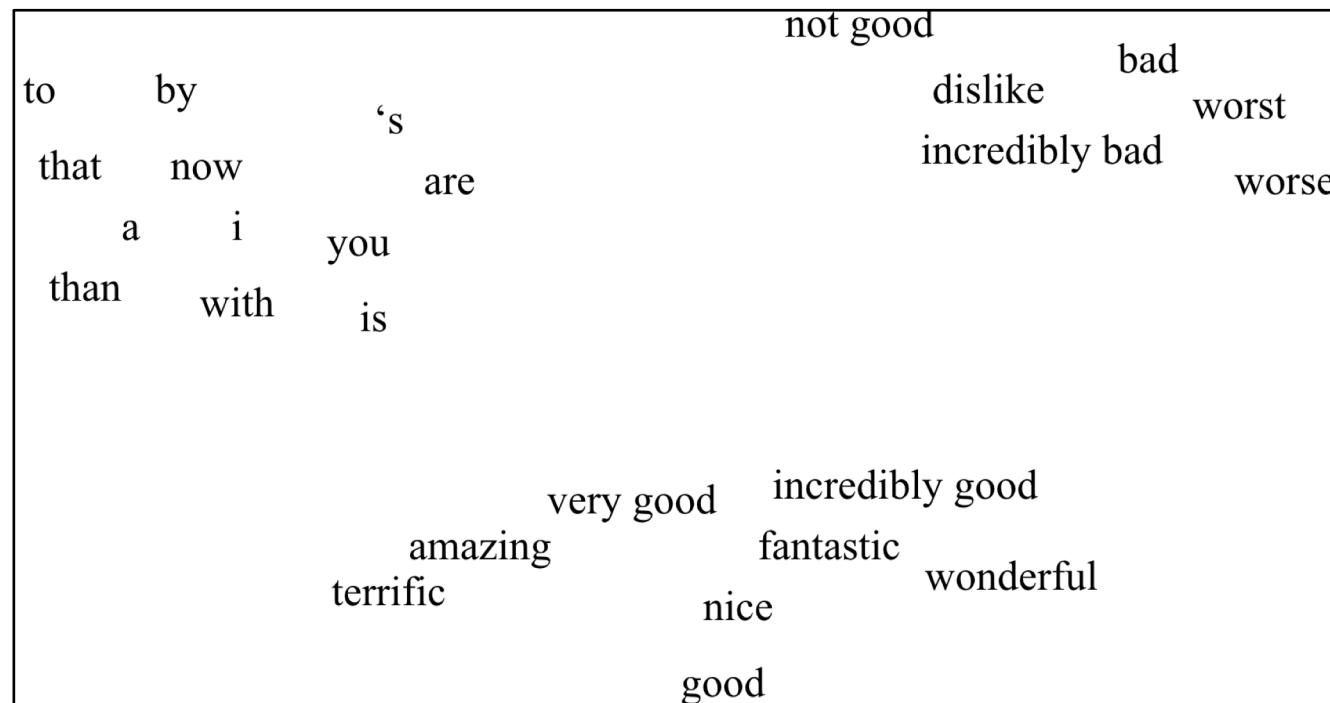
*A small fox **ran** to catch the rabbit.*

*The fox lost one of his **eyes**.*

大規模コーパスにおける単語の共起語を観察することで  
その語の意味を類推可能

# ベクトル意味論

- 語の意味を距離空間上の点で数理的に表現  
埋め込み (embeddings), 単語ベクトル
  - 単語ベクトル間の演算を用いて単語間の関係を計算



# 分布仮説に基づく単語ベクトルの計算

単語と文脈で共起する語(文脈語)の関係をモデル化  
前後  $n$  語 (window), 同一文(書), etc.

- 分布表現 (distributional representation / count-based vector)
  - ベクトルの次元が共起する単語と陽に対応
  - 各次元の値は共起語の頻度を元に計算 (例: tf-idf, PMI)
  - 高次元・疎 (非ゼロ成分が少ない)
- 分散表現 (distributed representation / predict-based vector)
  - ベクトルの次元は共起する単語と対応しない
  - 低次元・密

# 分布仮説に基づく単語ベクトルの計算

単語と文脈で共起する語(文脈語)の関係をモデル化  
前後  $n$  語 (window), 同一文(書), etc.

- 分布表現 (distributional representation / count-based vector)
  - ベクトルの次元が共起する単語と陽に対応
  - 各次元の値は共起語の頻度を元に計算 (例: tf-idf, PMI)
  - 高次元・疎 (非ゼロ成分が少ない)
- 分散表現 (distributed representation or predict-based vector)
  - ベクトルの次元は共起する単語と対応しない
  - 低次元・密

# 単語文脈行列 (term-context matrix)

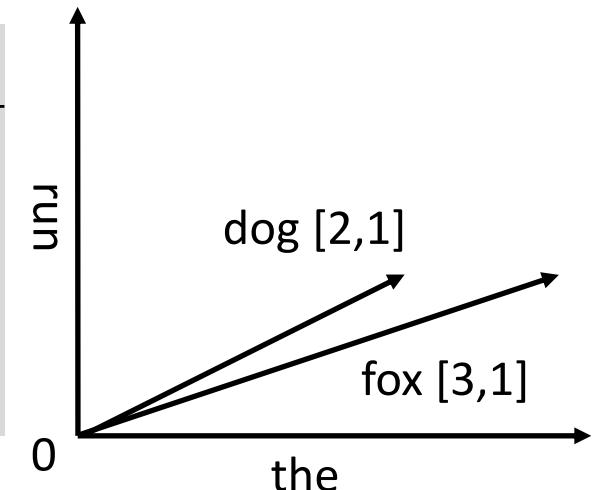
- コーパス中で、単語と同一文脈で共起する語(文脈語)の頻度を数える

前後  $n$  語, 同一文(書), etc.

*The small **dog** **barks** louder.  
His **dog** runs fast.  
**Eyes** of the **dog** was very small.*

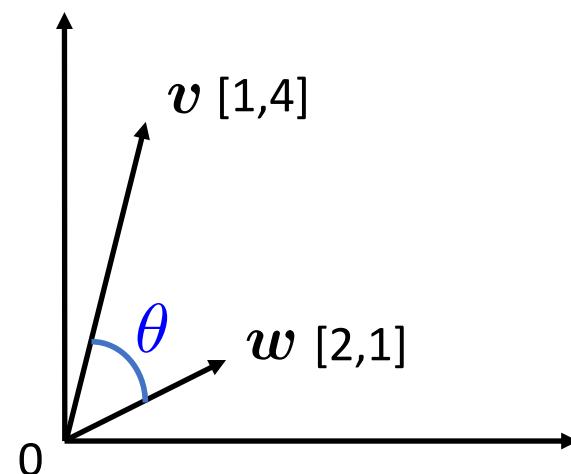
*Foxes are **barking** in the distance.  
A **small fox** **ran** to catch the rabbit.  
The **fox** lost one of his **eyes**.*

文脈語							
the	small	bark	run	eye	fast	...	
...	...	...	...	...	...	...	...
<b>dog</b>	2	1	1	1	1	1	...
<b>fox</b>	3	1	1	1	1	0	...
...	...	...	...	...	...	...	...



# コサイン類似度

- コサイン類似度(ベクトルが成す角の余弦)によって単語ベクトル間の類似度を計算
  - ベクトルの次元の値(頻度)は非負なので  $0 \leq \cos \theta \leq 1$
  - ユークリッド距離と違い单語の頻度の大小(ベクトルのノルム)に依存せず計算できるため都合が良い



$$\text{cosine}(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\| \|\mathbf{w}\|} = \frac{\sum_{i=1}^N v_i w_i}{\sqrt{\sum_{i=1}^N v_i^2} \sqrt{\sum_{i=1}^N w_i^2}}$$

$\boxed{\mathbf{v} \cdot \mathbf{w} = \|\mathbf{v}\| \|\mathbf{w}\| \cos \theta}$

$$\text{cosine}(\mathbf{v}, \mathbf{w}) = \frac{1}{4} (2, 1) = \frac{1 \cdot 2 + 4 \cdot 1}{\sqrt{1^2 + 4^2} \sqrt{2^2 + 1^2}} = \frac{6}{\sqrt{85}} = 0.6507\dots$$

# 単語ベクトルの次元の値は頻度で良いか？

- 文脈語の頻度は単語の性質を正しく表すか？
  - 高頻度語は機能語(冠詞, 前置詞など)が多く,どの語とも共起するので, 意味との関係は希薄そう  
*the, a, from, up, ...*
  - 低頻度語は詳細な内容語(固有名詞など)が多く,意味との関係は強そう  
*London, agriculture, ...*

語の性質を強く示唆する文脈語(次元)を重視したい

# tf-idf に基づく分布表現

- 単語ベクトルの各次元の値(共起語の重み)を **tf-idf** [Spark Jones 1972] を利用して計算  
(情報検索)で文書が含む語から文書ベクトルを作る際に  
単語の重み付けに使われる手法
- $$w_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$$
- 文書  $d \rightarrow$  単語  
文書が含む語  $t \rightarrow$  共起語

- tf (term frequency): 共起語の頻度

$$\text{tf}_{t,d} = \begin{cases} 1 + \log_{10} C(t, d) & \text{if } C(t, d) > 0 \\ 0 & \text{otherwise} \end{cases}$$

経験則として頻度は  
対数をとって鈍らす

- idf (inverse document frequency): 低頻度語へのボーナス

$$\text{idf}_t = \log_{10} \frac{N}{\text{df}_t}$$

单語種類数

共起語  $t$  と共に起する単語数

# PPMI (Positive Pointwise Mutual Information) に基づく分布表現

- PMI (自己相互情報量; Pointwise mutual information) により 単語と文脈語の関連の強さをモデル化

$$\text{PMI}(w, c) = \log_2 \frac{P(w, c)}{P(w)P(c)}$$

*w* と *c* が共起する確率  
と *c* の出現が独立なときに共起する確率

- 負のPMIは推定値の信頼性が低いため単語ベクトルの次元には Positive PMI (PPMI) を用いる [Bullinaria+ 2007]

$$\text{PPMI}(w, c) = \max\left(\log_2 \frac{P(w, c)}{P(w)P(c)}, 0\right)$$

# PPMIに基づく分布表現の改善 [Levy+ 2015]

- PMI 計算の工夫

- Shifted PMI: PMI の値が極端に小さい次元を無視
- スムージング: 低頻度の文脈語の過度な影響を抑える

$$\text{SPPMI}_{\alpha,k}(w, c) = \max\left(\log_2 \frac{P(w, c)}{P(w)P_\alpha(c)} - \log k, 0\right)$$

経験則:  $\alpha = 0.75, k = 5$

$$P_\alpha(c) = \frac{C(c)^\alpha}{\sum_c C(c)^\alpha}$$

- 文脈長(window幅)内の文脈語の扱いを改善

- 単語に近い文脈語を重視, 超高頻度語・低頻度語を無視

- 特異値分解 (SVD) により, 高次元・疎なベクトルを  
低次元・密なベクトルに次元圧縮

- Shifted PM + SVD ≈ Skip-gram w/ negative sampling [Levy+ 2014]

# 分布仮説に基づく単語ベクトルの計算

単語と文脈で共起する語(共起語)の関係をモデル化  
前後  $n$  語 (window), 同一文(書), etc.

- 分布表現 (distributional representation / count-based vector)
  - ベクトルの次元が共起する単語と陽に対応
  - 各次元の値は共起語の頻度を元に計算 (例: tf-idf, PMI)
  - 高次元・疎 (非ゼロ成分が少ない)
- 分散表現 (distributed representation / predict-based vector)
  - ベクトルの次元は共起する単語と対応しない
  - 低次元・密

# Skip-gram with negative sampling (SGNS) (1/2)

[Mikolov+ 2013]

- 文脈語の頻度を数える代わりに共起確率を予測することで共起をモデル化 (参照実装: word2vec)
  - ベクトルを計算する語と実際共起した語のペアを正例
  - 語彙中のそれ以外の単語とのペアを負例とみなす
  - 1, 2を用い2値分類器をロジスティック回帰で学習
  - 学習された重みベクトル = 単語ベクトル
- 文脈語を直接予測する超多値分類問題を学習する skip-gram を計算の効率のため単純化したもの

## Skip-gram with negative sampling (SGNS) (2/2) [Mikolov+ 2013]

- 与えられた単語  $t$  と window 幅  $L$  に対し文脈語候補  $c$  が window 内で共起する確率を予測

... lemon, a [tablespoon of apricot jam, a] pinch ...

$$P^+(\mathbf{t}, \mathbf{c}_{1:2L}) = \prod_{i=1}^{2L} \frac{1}{1 + e^{-\mathbf{t} \cdot \mathbf{c}_i}}$$

これらを学習

$$\log P^+(\mathbf{t}, \mathbf{c}_{1:2L}) = \sum_{i=1}^{2L} \log \frac{1}{1 + e^{-\mathbf{t} \cdot \mathbf{c}_i}}$$

# Skip-gram with negative sampling (SGNS) (2/2)

[Mikolov+ 2013]

- 与えられた単語  $t$  と window 幅  $L$  に対し文脈語候補  $c$  が window 内で共起する確率を予測

... lemon, a [tablespoon of apricot jam, a] pinch ...

- 実データから正例, 負例サンプリングにより負例を生成

正例		負例 (正例の $k$ 倍; $k=2$ )	
t	c	t	c
apricot	tablespoon	apricot	aardvark
apricot	of	apricot	puddle
apricot	jam	apricot	where
apricot	a	apricot	coaxial

ランダム  
サンプル

$$P_\alpha(c) = \frac{C(c)^\alpha}{\sum_c C(c)^\alpha}$$

## Skip-gram with negative sampling (SGNS) (2/2) [Mikolov+ 2013]

- 与えられた単語  $t$  と window 幅  $L$  に対し文脈語候補  $c$  が window 内で共起する確率を予測

... lemon, a [tablespoon of apricot jam, a] pinch ...

- 実データから正例, 負例サンプリングにより負例を生成
- 各共起語について, 以下の目的関数を最大化

$$L(\theta) = \log P^+(t, c^+) + \sum_{i=1}^k \log(1 - P^+(t, c_i^-))$$

$$= \log \frac{1}{1 + e^{-\mathbf{c}^+ \cdot \mathbf{t}}} + \sum_{i=1}^k \log \frac{1}{1 + e^{\mathbf{c}_i^- \cdot \mathbf{t}}}$$

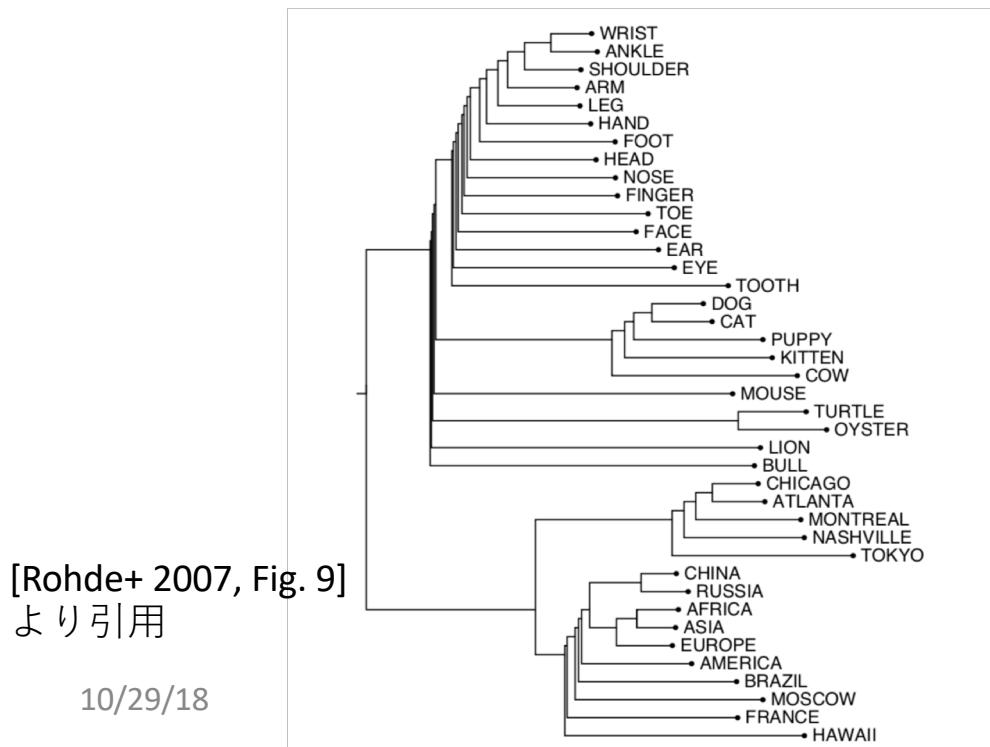
文脈語の確率  
を最大化

非文脈語の確率  
を最小化

逐次学習可能  
[Kaji+ 2017]

## 単語ベクトルの可視化

- コサイン類似度の高い単語(類似語)を計算  
例) frog → frogs, toad, litoria, leptodactylidae, rana, lizard ...
  - クラスタリングで離散化 (グルーピング・階層化) • 次元圧縮で低次元ベクトルに変換しプロット



## t-SNE [van der Maaten + 2008] による次元圧縮

			not good		
to	by	's		dislike	bad
that	now	are		incredibly bad	worst
a	i	you			worse
than	with	is			
			very good	incredibly good	
	amazing			fantastic	
	terrific				wonderful
			nice		
			good		

# 単語ベクトルの評価 (1/3)

- 言語表現(単語, 文など)の類似性判定タスク  
人による言語表現の類似度との相関で評価
  - 単語対の関連度: WordSim-353 [Finkelstein+ 2002, Agirre+ 2012]
  - 単語対の類似度: WordSim-353, SimLex-999 [Hill+ 2015]

SimLex-999 [Hill+ 2015]

単語対		類似度
vanish	disappear	9.8
behave	obey	7.3
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

1単語対辺り約50人の被験者が [0-6] で類似度を付与した後、正規化・平均

[Levy+ 2015] での比較実験

	WordSim (relatedness)	WordSim (similarity)	SimLex
PPMI	.697	.755	.393
+ SVD	.691	.793	.432
SGNS	.685	.793	.438

# 単語ベクトルの評価 (1/3)

- 言語表現(単語, 文など)の類似性判定タスク  
人による言語表現の類似度との相関で評価
  - 単語対の関連度: WordSim-353 [Finkelstein+ 2002, Agirre+ 2012]
  - 単語対の類似度: WordSim-353, SimLex-999 [Hill+ 2015]
  - 文対の類似度: Semeval-2015 task2 [Agirre+ 2015]
  - 同義語選択: TOEFL dataset

Semeval-2015 task2 [Agirre+ 2015]

文対	類似度
The bird is bathing in the sink.	5
In May 2010, the troops attempted to invade Kabul.	4
John said he is considered a witness but not a suspect.	3
They flew out of the nest in groups.	2
The woman is playing the violin.	1

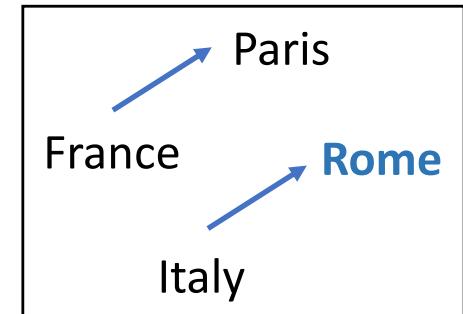
# 単語ベクトルの評価 (2/3)

- 単語のアナロジータスク [Mikolov+ 2013]  
ベクトル間の四則演算でアナロジーを求める

例) France is to Paris as Italy to \_\_\_\_\_

$$w = \operatorname{argmax}_{w'} \cos(\mathbf{v}_{w'}, \mathbf{v}_{Paris} - \mathbf{v}_{France} + \mathbf{v}_{Italy})$$

$$= \operatorname{argmax}_{w'} [\cos(\mathbf{v}_{w'}, \mathbf{v}_{Paris}) - \cos(\mathbf{v}_{w'}, \mathbf{v}_{France}) + \cos(\mathbf{v}_{w'}, \mathbf{v}_{Italy})]$$



単位ベクトル  
を仮定

[Levy+ 2015] での比較実験

- MSR's analogy dataset [Mikolov+ 2013c]  
語形変化に関するアナロジー
- Google's analogy dataset [Mikolov+ 2013a]  
語形変化 + 意味的なアナロジー

	Google	MSR
PPMI	.553	.306
+ SVD	.554	.408
SGNS	.676	.618

# 単語ベクトルの評価 (3/3)

- 単語の定義文生成タスク [Noraset+ 2017]
  - 良い単語ベクトル = 単語の意味を捉えたベクトル
  - 単語ベクトルから定義文を生成し, 辞書の定義文との類似性を評価

word2vec (CBOW) で Web 文書 1000 億語から学習した  
単語ベクトルから WordNet, GCDIE を用いて定義文生成

単語	生成された定義文
Creek	a narrow stream of water
feminine	having the nature of a woman
mathematical	of or pertaining to the science

[Noraset+ 2017] での  
S+G モデルの生成例

銀の弾丸(万能の単語ベクトル)は存在しないので注意

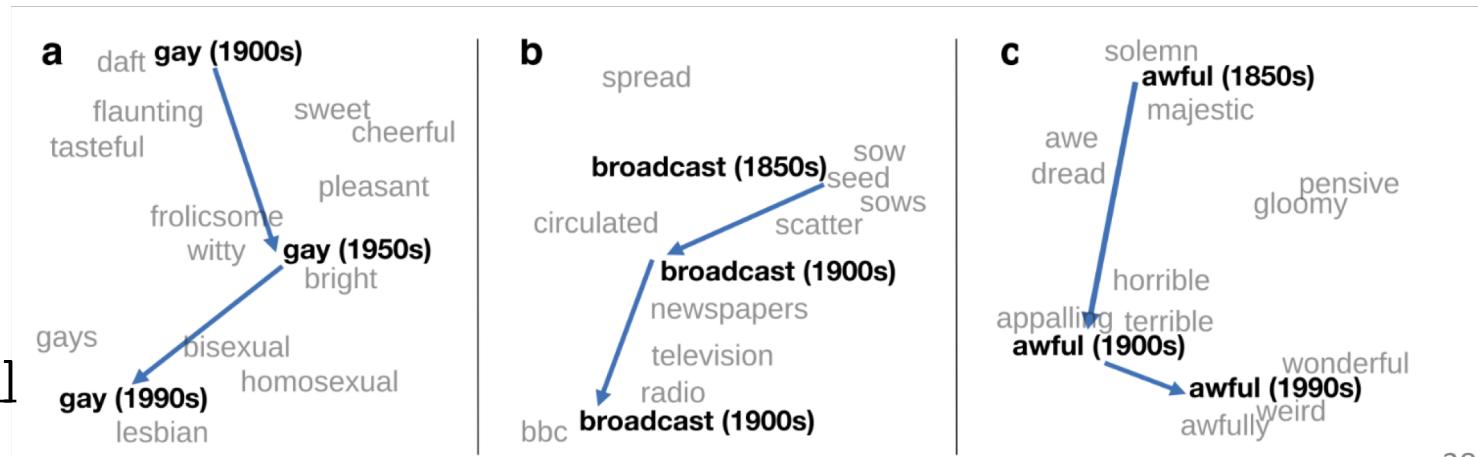
# 単語の意味をコーパスから決める難しさ： バイアスの混入

- コーパスから学習した単語ベクトルにはテキスト中に言語化されたバイアスが混入 [Bolukbasi+ 2016]
  - 応用にバイアスが波及するため、補正が必要 [Zao+ 2014]

例)

‘computer programmer’ - ‘man’ + ‘woman’ = ‘homemaker’  
‘doctor’ - ‘father’ + ‘mother’ = ‘nurse’

- 通時的・共時的なテキスト解析への応用



語の意味の変遷

[Hamilton+ (2015)]

# まとめ

- 語の記号的意味表現
  - 語義間の離散的な関係:  
同義・反義・上位・下位・同位・部分・全体
  - 語義間の連続的な関係  
類義・関連
- 語の数理的意味表現
  - 分布仮説とベクトル意味論
  - 分布表現 (高次元・疎)  
共起語を陽に次元としてモデル化したベクトル表現
  - 分散表現 (低次元・密)  
共起を陰にモデル化したベクトル表現