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」
導入と基礎「大規模言語モデルの基礎」

永田亮

甲南大学知能情報学部

理化学研究所(客員研究員)

産業技術総合研究所(客員研究員)

はじめに
少し自己紹介

自己紹介

- **永田亮(甲南大学)**
  - 専門分野:自然言語処理(NLP)
    英文誤りの検出/訂正
    文章の自動採点
    ライティング解説文の自動生成
    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.

ここから本題です

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

本日お話しする内容

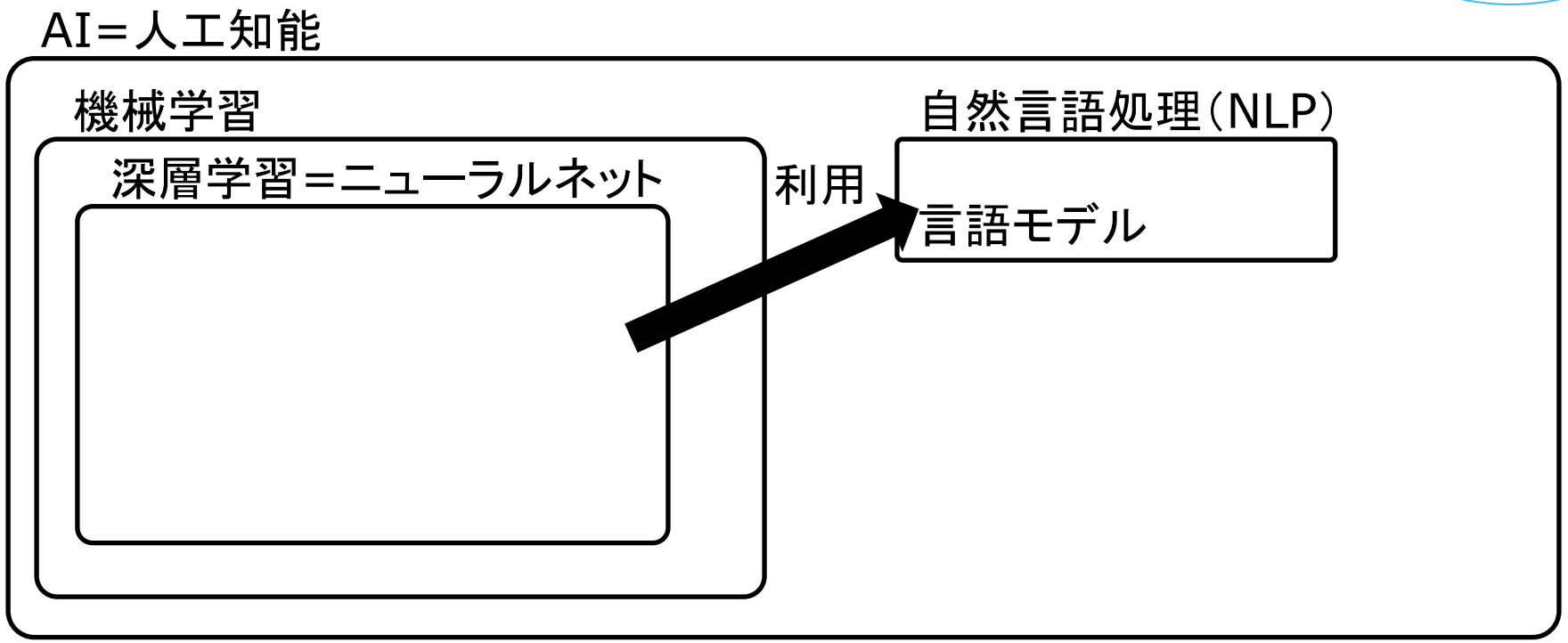
- **導入：AI，深層学習とはなにか？（10分）**
  - 入門編：ある観点から大胆に要約
- **基礎：言語モデルと言語分析（25分）**
  - 自然言語処理における言語モデル
  - 副産物（単語ベクトル）が重要
- **応用：研究事例の紹介（各1時間）**
  - 語（文内）の意味の計算（永田担当）
  - 文を超えた意味の計算（横井担当）

このセッションでお話する内容

- **導入**
  - AI: 多くの場合，深層学習を使ったプログラム
  - 深層学習：数値変換機構
- **基礎**
  - 言語モデル：単語予測器
  - 現在は深層学習に基づく
  - 副産物（単語ベクトル）が重要
  - 単語ベクトルの直感的解釈

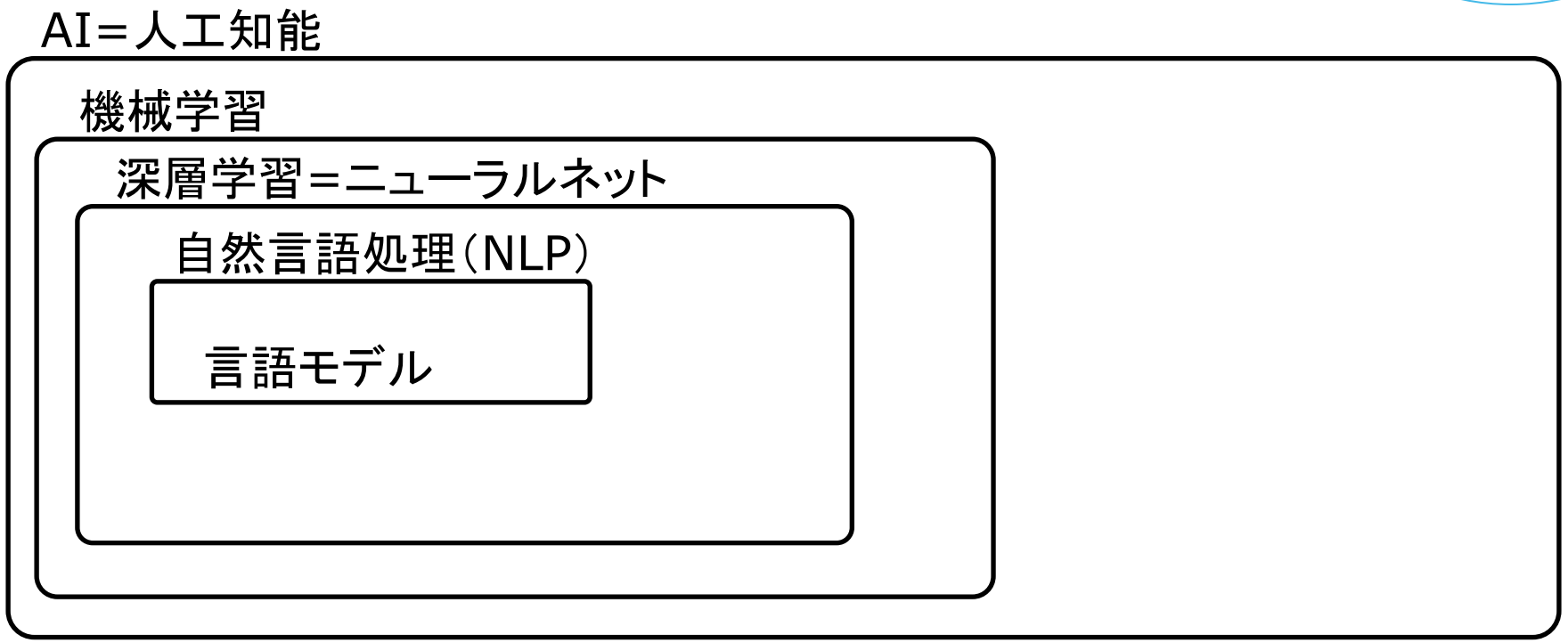
導入：
AIとは
深層学習とは

AIと深層学習



深層学習≒単なる数値（列）変換機構

こういう見方もできる？



深層学習≒単なる数値（列）変換機構

深層学習：数値列変換の世界

画像認識の場合（画素値→カテゴリID）

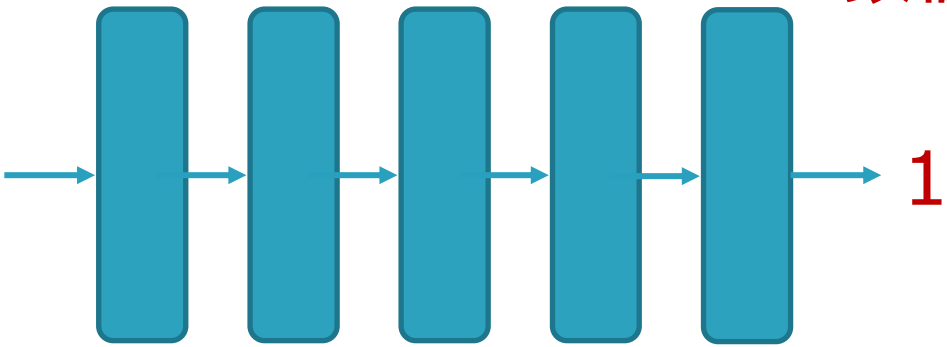
画像

カテゴリID

数値

数値列

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



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）

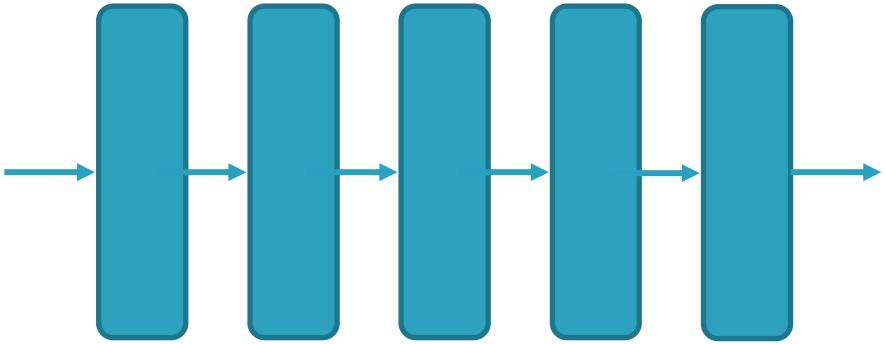
画像認識（分類）ニューラルネット

画像認識の場合（画素値→カテゴリ確率）

画像

数値列

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



カテゴリID

各カテゴリの確率数値列

$$\begin{pmatrix} 0.8 \\ 0.0 \\ \vdots \\ 0.2 \end{pmatrix}$$

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

数値列変換の中身（ひとつの例）

画像

カテゴリID

$$\begin{pmatrix} 0.7 \\ 1.7 \\ -1.6 \end{pmatrix} \rightarrow \begin{matrix} 0.7*1 + 1.7*1 - 1.6*1 \\ \vdots \\ 0.7*1/7 + 1.7*1 - 1.6*1 \end{matrix} \rightarrow \begin{pmatrix} 0.8 \\ 0 \\ \vdots \\ 0.2 \end{pmatrix} \begin{matrix} 1.\ 犬 \\ 2.\ 車 \\ \vdots \\ N.\ 猫 \end{matrix}$$

訓練（＝学習）：赤い数値の決定
パラメータ数　：赤い数値の数

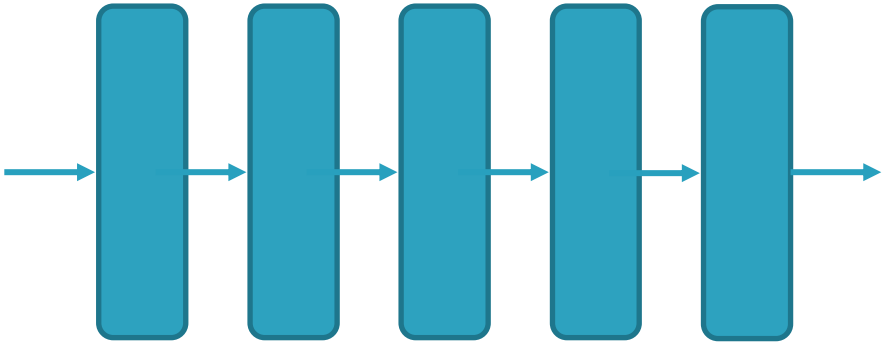
様々なものが数値変換で実現可能

白黒画像

カラー画像

数値列

$$\begin{pmatrix} 0.7 \\ 0.1 \\ 0 \end{pmatrix}$$



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

カラー画像の復元器！

逆方向にすれば白黒化

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

基礎
言語モデルとは:単語予測器

言語モデル(Language Model; LM)とは

- **自然言語処理では単語予測器**
  - 次に来る単語は？ その確率は？

    In Valencia, I enjoyed …
  - 複数回適応することで文章の生成も可！

    In Valencia, I enjoyed …

    In Valencia, I enjoyed some …

    In Valencia, I enjoyed some paella…

LMでなんでも解けてしまう！？

- **質問応答：質問文から解答文**
  - パエリアの発祥の地はどこですか？パエリアはスペインのバレンシアで生まれました.
- **翻訳：ある言語の文から別の言語の文**
  - *In Valencia, I enjoyed paella 私はバレンシアでパエリアを楽しんだ.*

LMは単語予測をしているに過ぎない！

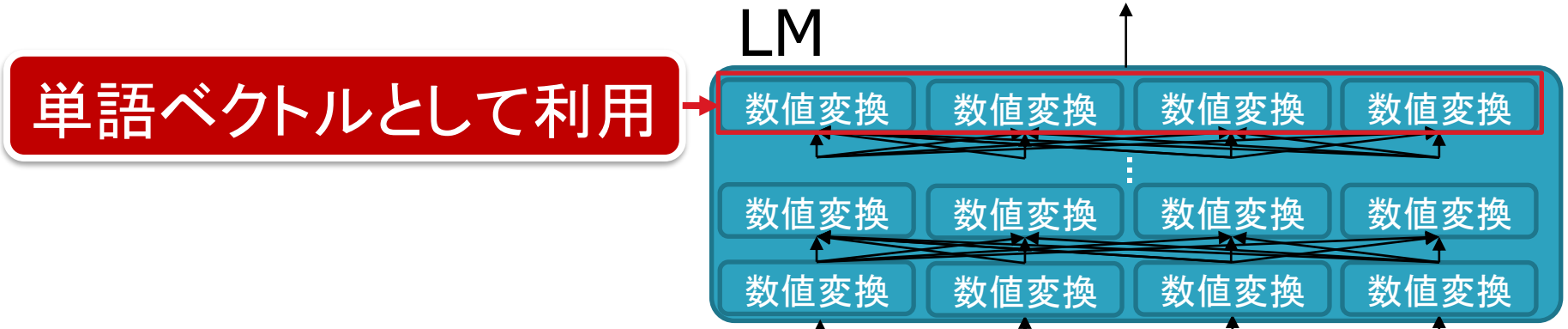
LMの素晴らしい副産物

- **処理過程で数値列（ベクトル）が得られる**
  - 単語ベクトル
  - 文ベクトル
- **記号から数値列へ**
  - 伝統的な言語学では分析対象は記号
  - 各種演算が可能（後述）
  - 単語や文の類似度が計算可能

代表的なLMの構造と処理の流れ

出力：次に来る単語とその確率

steak 0.5
stew 0.3
：　：



③ベクトル化（LMへ入力）

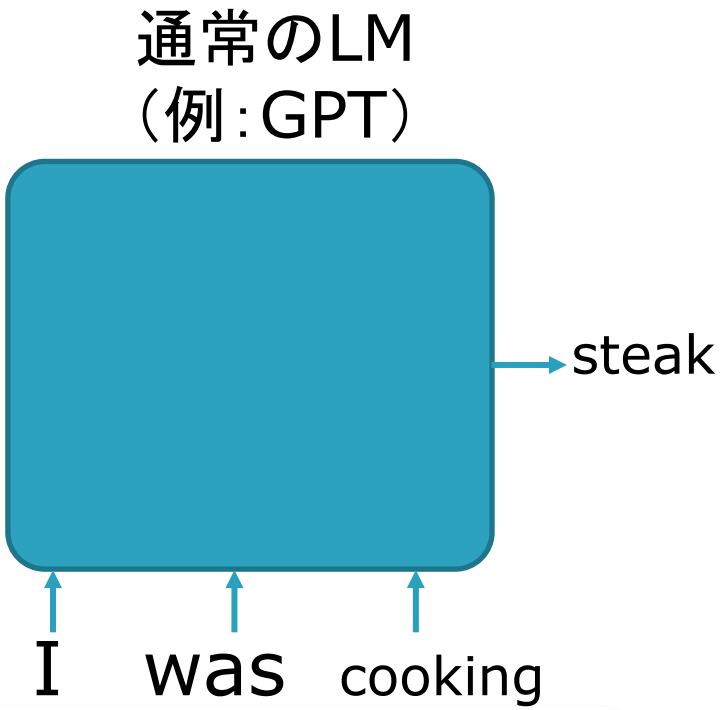
②ID化：　101,　146, 1108,　8739

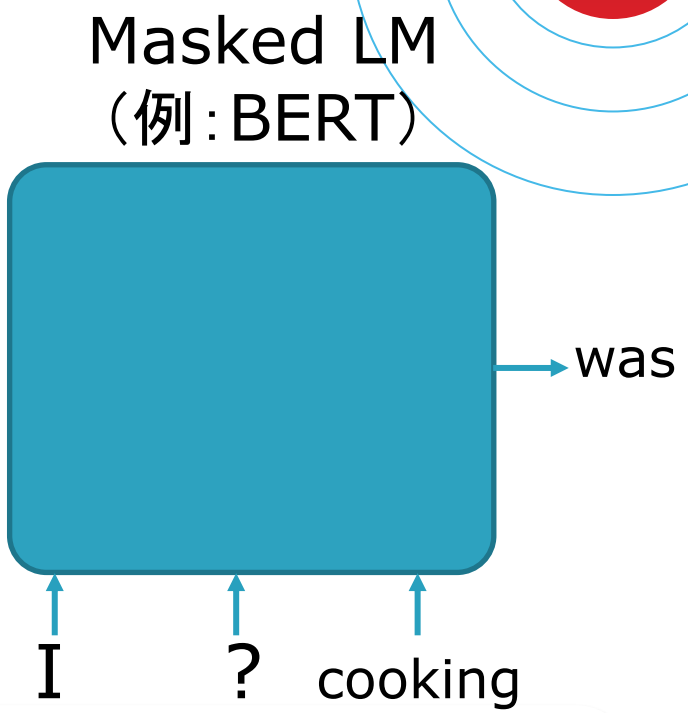
①分割（tokenize）：　文頭 | I | was | cooking

入力：I was cooking

分かり易さを重視しています．様々な構造があります

通常のLMとMasked LM





先行文脈から
次の単語を予測

文章中の任意箇
所における単語を
予測

言語分析ではMLMの方がよい場合が多い

二種類の単語ベクトル

- **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

ここまでの重要なポイントのおさらい

- **言語モデル**
  - 単語予測器
  - 現代は深層学習ベース
  - 副産物の単語ベクトル
- **単語ベクトル**
  - 2種類:トークン or タイプベース
  - 文脈(周辺単語)をソフトに反映
  - 言語の分析に有用

単語ベクトルの直感的解釈

(しばらく)トークンベースのベクトルを仮定

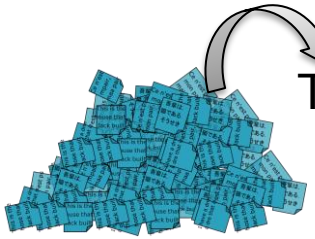
- **トークン(事例)ベース**
  - 各事例に応じた複数の単語ベクトル
  - 個別文脈を反映した単語ベクトル

  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

単語→ベクトル＝空間上へのマッピング

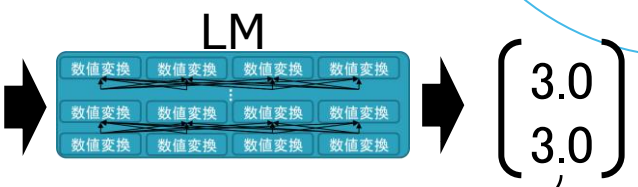
例：対象単語 *apple*

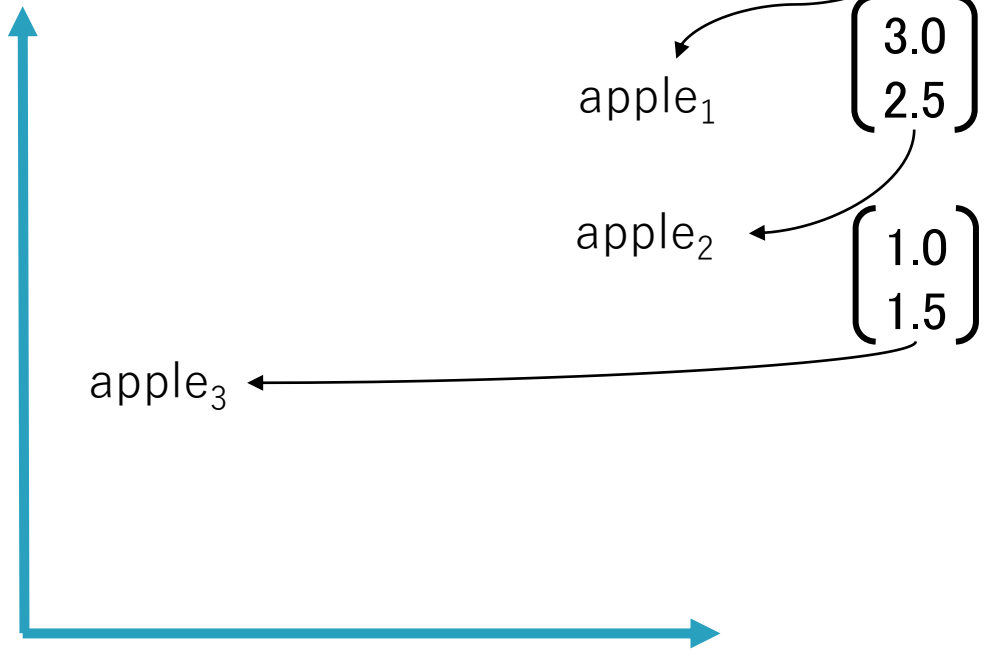


There was *apple* in the salad.

I like *apple* jam.

She uses *apple* MacBooks.

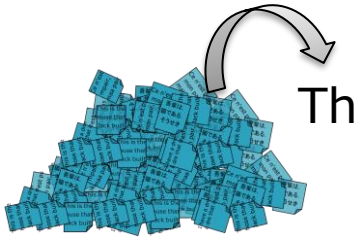




近さが測れる

向きの近さ＝類似度と考えることが多い

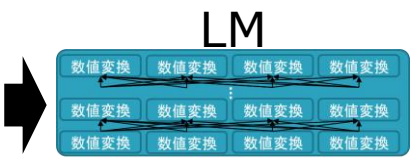
例：対象単語 apple



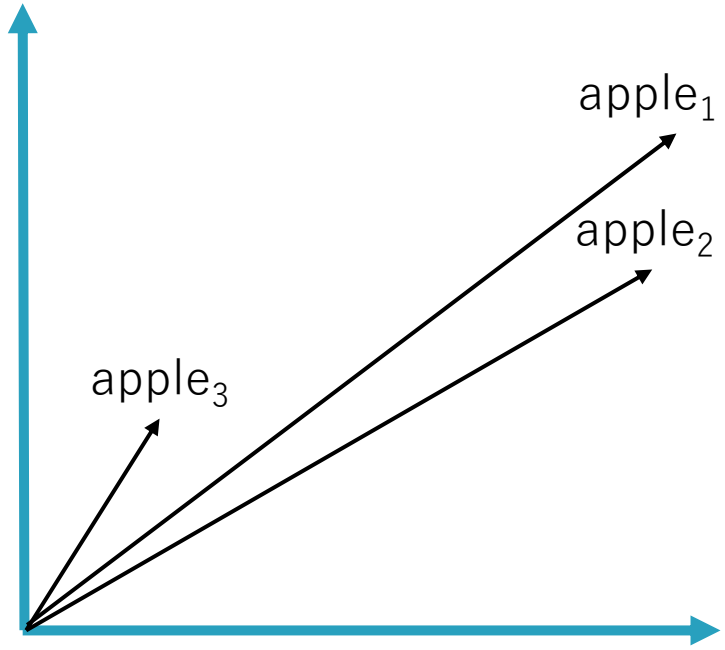
There was *apple* in the salad.

I like *apple* jam.

She uses *apple* MacBooks.



$$\begin{pmatrix} 3.0 \\ 3.0 \end{pmatrix}$$



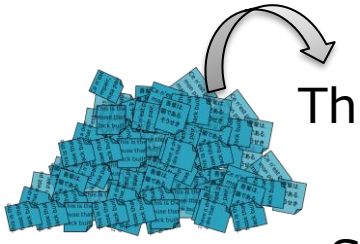
$$\begin{pmatrix} 3.0 \\ 2.5 \end{pmatrix}$$

$$\begin{pmatrix} 1.0 \\ 1.5 \end{pmatrix}$$

余弦類似度（cos類似度）

方向=類似度=円上の距離

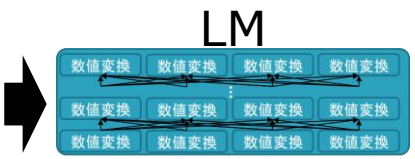
例：対象単語 apple



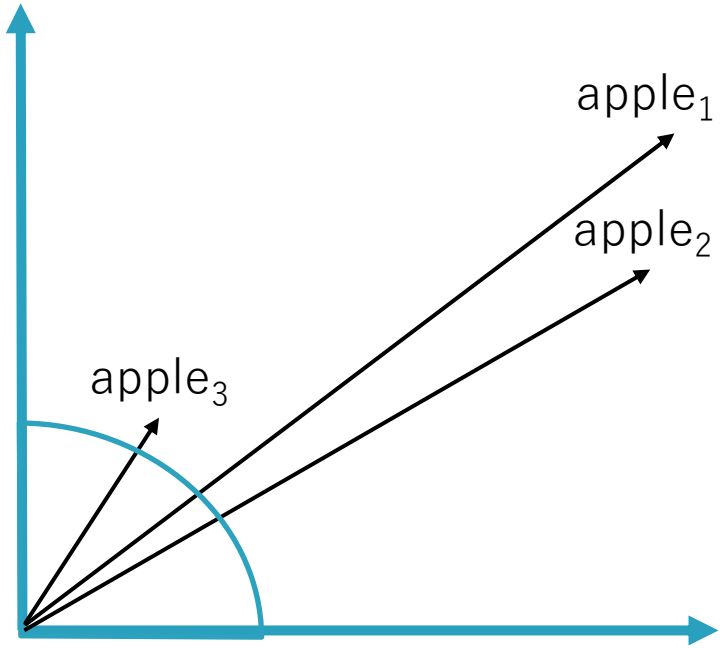
There was *apple* in the salad.

I like *apple* jam.

She uses *apple* MacBooks.



$$\begin{pmatrix} 3.0 \\ 3.0 \end{pmatrix}$$



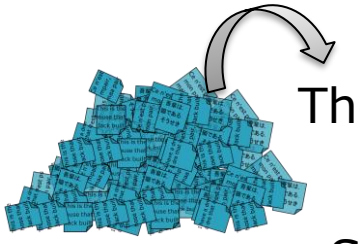
$$\begin{pmatrix} 3.0 \\ 2.5 \end{pmatrix}$$

$$\begin{pmatrix} 1.0 \\ 1.5 \end{pmatrix}$$

3次元（値が三つ）なら球面上

方向の多様性，文脈／意味の多様性

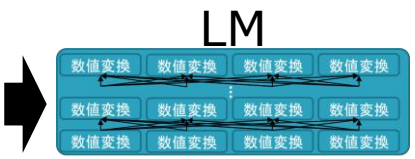
例：対象単語 apple



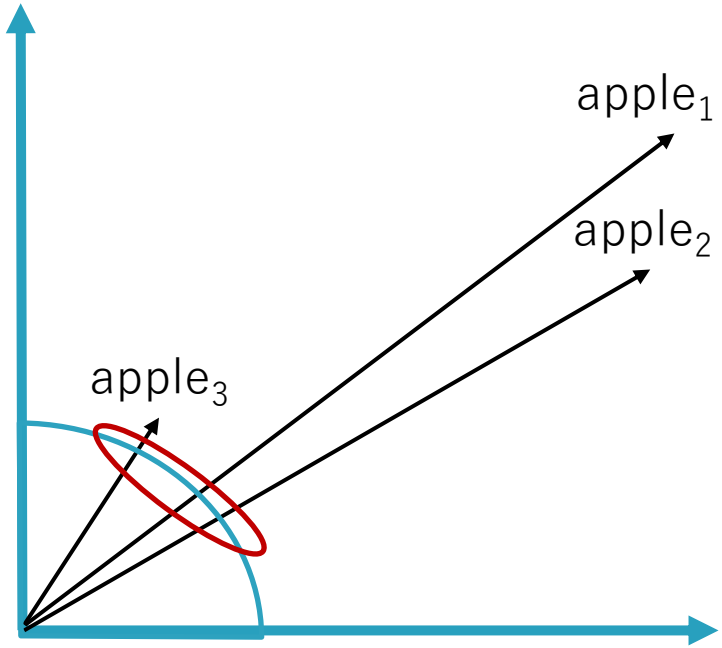
There was *apple* in the salad.

I like *apple* jam.

She uses *apple* MacBooks.



$$\begin{pmatrix} 3.0 \\ 3.0 \end{pmatrix}$$



$$\begin{pmatrix} 3.0 \\ 2.5 \end{pmatrix}$$

$$\begin{pmatrix} 1.0 \\ 1.5 \end{pmatrix}$$

様々な方向→文脈が多様→意味が多様

導入と基礎のまとめ

- **AI：最近はほとんど深層学習ベース**
- **深層学習：数値列変換機構（プログラム）**
- **言語モデル（LM）**
  - 単語予測器
  - 深層学習ベースが主流
  - 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
  - 方向性の多様性が文脈／意味の多様性に対応