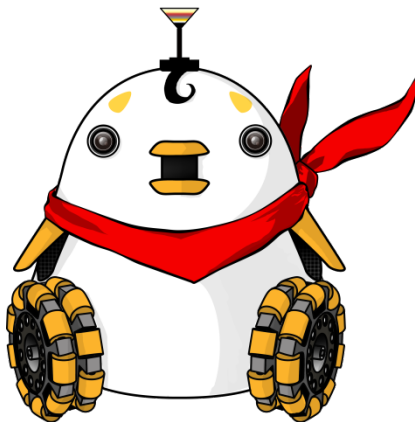


# 人工知能

## 第13回 言語と論理(1) 自然言語処理

立命館大学 情報理工学部

谷口彰

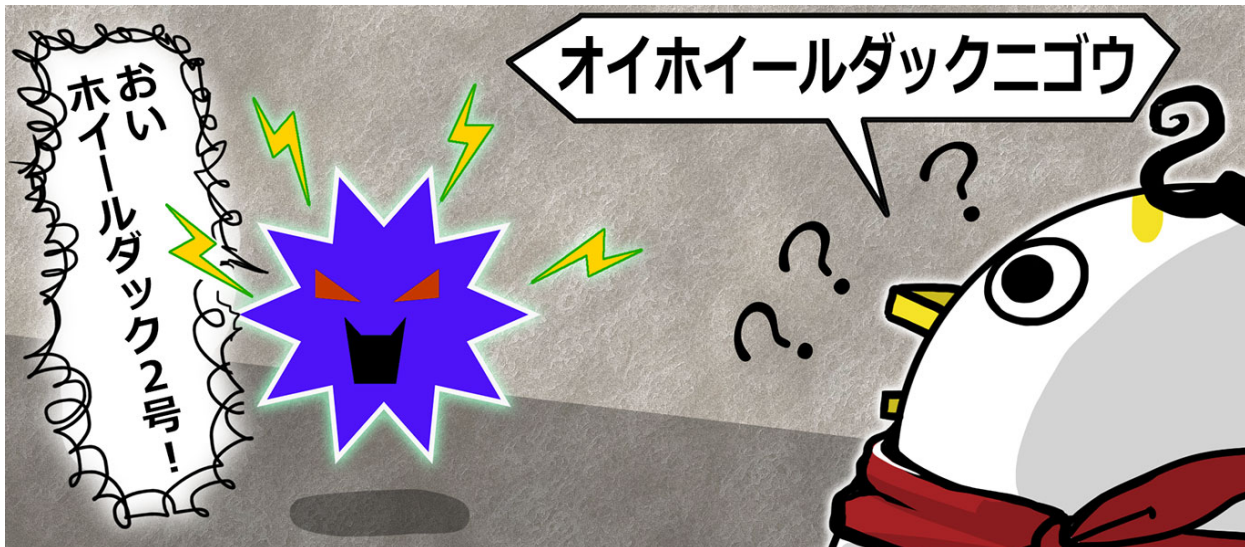


# STORY 言語と論理 (1)

- ホイールダック 2 号は、迷路のゴールまで行く自信を持った。ゴールへの経路を探索するやり方や、敵のかわし方を覚えた。
- 場所がわからなくなったときには、位置推定により自分がどこにいるかを調べることができる。また、事前に学習することで、宝箱やゴールも見分けられるようになった。これでゴールにたどり着けるだろう。
- しかし、ゴールにたどり着けば終わりではなかった。ゴールにはスフィンクスがいて、謎かけをしてくるのだ。
- 話に聞くとところによると、スフィンクスは決して難しい問題を出すわけではなく、普通に論理的に考えれば解ける程度の謎かけをしてくるらしい。
- しかし、ホイールダック 2 号には現状では大きな問題があった。ホイールダック 2 号には人間の言葉がわからないのだ。

# 仮定 言語と論理 (1)

- ホイールダック2号に文法 ( [syntax](#) ) に関する知識, 語彙 ( [lexicon](#) ) に関する知識は事前に埋め込んでよいものとする.
- ホイールダック2号は**誤りのない音声認識**が可能であるとする.



# Contents

- 14.1 自然言語処理
- 14.2 形態素解析
- 14.3 構文解析
- 14.4 意味解析\*
- 14.4 単語と文章のベクトル表現

# 14.1.1 自然言語処理と応用技術

□自然言語をコンピュータ上で処理するための研究を**自然言語処理 (natural language processing: NLP)**と呼ぶ.

□2000年代以降, WEB資源の爆発的增加によって処理可能なデータが圧倒的に増えて, 注目が増している.

□ロボットが言語理解する上でも必要.

□応用分野

□情報検索, 機械翻訳, 対話システム, 質問応答, 文書要約, など

□コンピュータ上で「言語」を扱う.

□人工言語

□プログラミング言語

□人手で作られた形式的な言語

□例) C言語, Java言語, XML, CSSなど

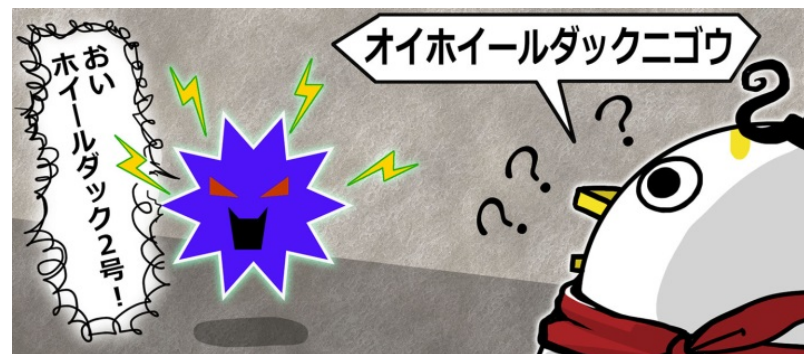
□自然言語

□人間が日常生活で用いる言語

□例) 英語, 日本語, 中国語・・・etc.etc.

□例) 大阪弁, 歌詞,

□X 小鳥のさえずり, 犬の鳴き声



## 14.1.2 自然言語処理の基礎技術

### 例文

- 私は窓から降っている雪を見た.
- 傘を持って家を出た.
- それを忘れてきた.

## 14.1.2 自然言語処理の基礎技術

### (1)形態素解析

#### ①品詞活用の推定

名詞 助詞

動詞・活用

- 私|は|窓|から|降っ|て|いる|雪|を|見|た|.
- 傘|を|持っ|て|家|を|出|た|.
- それ|を|忘れ|て|き|た.

#### ②分かち書き

## 14.1.2 自然言語処理の基礎技術

### (2)構文解析

#### 文法関係の解析

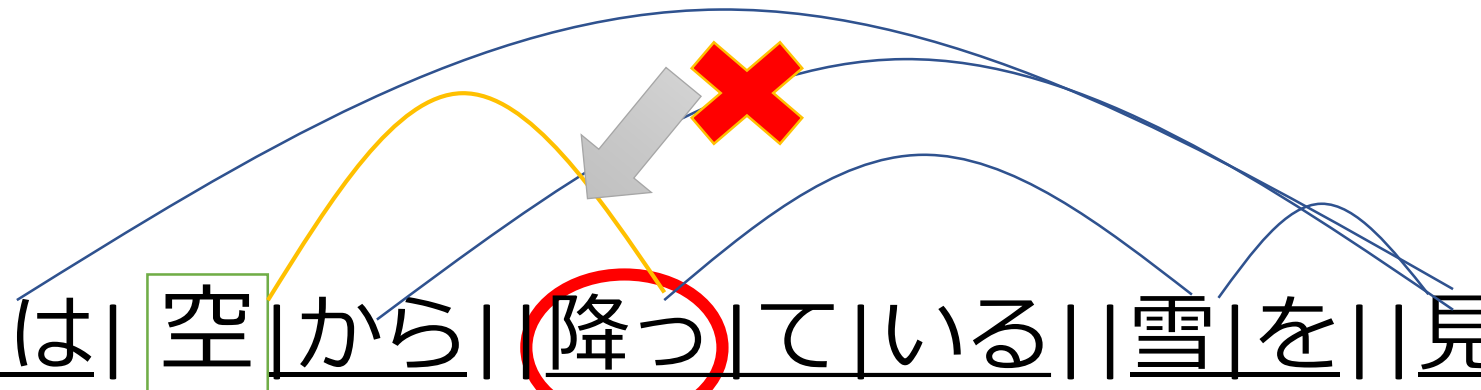
- 
- 私 | は | | 窓 | から | | 降っ | て | いる | | 雪 | を | | 見 | た | .
- 傘 | を | 持っ | て | 家 | を | 出 | た | .
- それ | を | 忘れ | て | き | た | .

- 日本語では形態素よりを結合させた分節単位で構文解析することが多い.
- 英語の場合は句構造文法, 日本語の場合は依存文法にもとづいて解析する場合が多い.



## 14.1.2 自然言語処理の基礎技術

### (3)意味解析

- 
- The diagram illustrates semantic parsing for the sentence "私|は| 空|から| 降っ|て|いる||雪|を||見|た|." (I saw snow falling from the sky). Blue arcs connect the words to their semantic roles. A yellow arc connects "空" (sky) to "降っ" (falling). A grey arrow points to the "降っ" node, which is circled in red. A red 'X' is placed above the arrow. A red arrow points from the red circle to the "意味解析" (Semantic Analysis) box.
- 私|は| 空|から| 降っ|て|いる||雪|を||見|た|.
  - 傘|を|持っ|て|家|を|出|た|.
  - それ|を|忘れ|て|き|た|.

□ 格文法(case grammar)

□ 表層格(surface case)

□ ガ格, ヲ格など

□ 深層格(deep case)

□ 動作主格, 道具格など

#### 意味解析

ふ・る【降る】

[動ラ五(四)]

1 空から雨や雪などが連続的に、広い範囲にわたって落ちてくる。また、細かい落ちてくる。「大雪が一・る」「火山灰が一・る」

2 霜がおりる。「早霜が一・る」

3 日光・月光が注ぐ。「やしの葉影に一・る月の光」

## 14.1.2 自然言語処理の基礎技術

### (4) 文脈解析

- 私 | は | | 窓 | から | | 降っ | て | いる | | 雪 | を | | 見 | た |
- 傘 | を | 持っ | て | 家 | を | 出 | た | .
- それ | を | 忘れ | て | き | た | .



文脈解析

※照応関係

# 基礎技術の関係

形態素解析でまずは区切る

構文解析

- 私は窓から降っている雪を見た。
- 傘を持って家を出た。
- それを忘れてきた。

文脈解析

意味解析

ふ・る【降る】

〔動ラ五(四)〕

1 空から雨や雪などが連続的に、広い範囲にわたって落ちてくる。また、細かいものが上方からたふさふさ落ちてくる。「大雪が一・る」「火山灰が一・る」

2 霜がおりる。「早霜が一・る」

3 日光・月光が注ぐ。「やしの葉影に一・る月」

4 多く集まり寄ってくる。「一・るほど縁談がある」

「今日」は降らないよな、  
「雪」は降るよな・・・

## 演習14-1 基礎技術の関係

- 「この道をまっすぐ行ったら交番が見えます。そこを右に曲がれば修道院ですよ」
- この文章において、「そこ」が何を指すのかを特定するために必要なのは以下のどの解析か。最も適切なものを選べ。

- ① 形態素解析
- ② 構文解析
- ③ 意味解析
- ④ 文脈解析

# Contents

□14.1 自然言語処理

□14.2 形態素解析

□14.3 構文解析

□14.4 意味解析\*

□14.4 単語と文章のベクトル表現

## 14.2.1 言語と形態素

□自然言語は音素，形態素，語，文，文章という階層構造を持つ．この中で**形態素**は言語の意味を持つ最小単位

□日本語の場合はスペースが無いので解析が必要

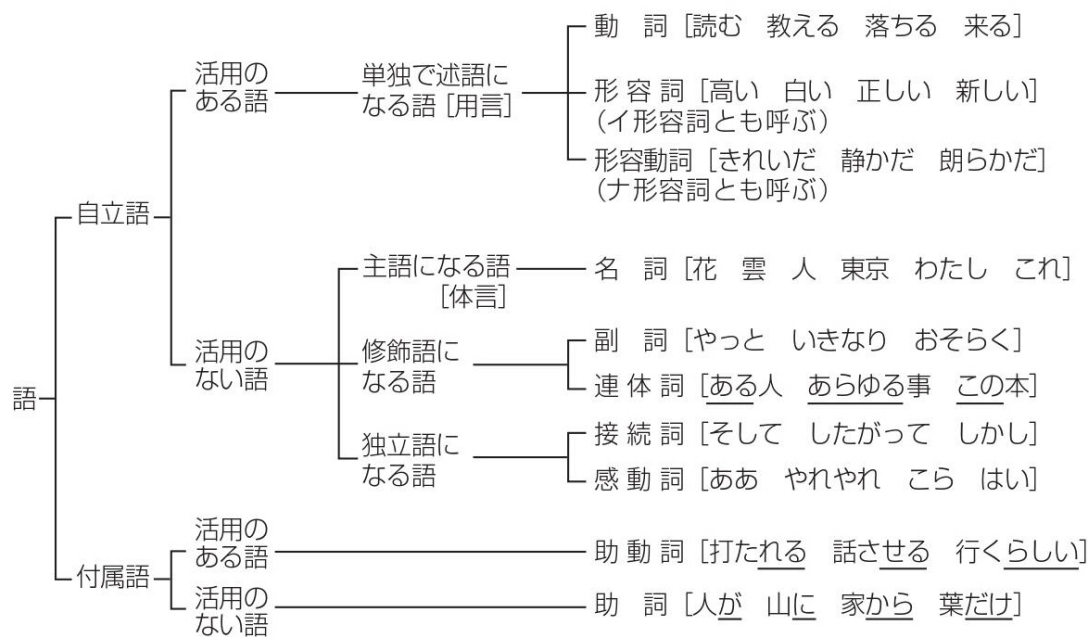


図 14.3 日本語の品詞分類

長尾真, 佐藤理史 (編) : 自然言語処理, 岩波書店, 1996.

# 形態素解析 (morphological analysis)

□形態素(morpheme)とは文字によって表記された自然言語の文において、意味を担う最小の言語単位のことを指す。(単語と同じか、より小さいまとまり)

## □形態素解析の役割

□文の形態素分割 (word segmentation; **分かち書き** 処理)

□太郎はお茶子に花をあげる.

□太郎 | は | お茶子 | に | 花 | を | あげる | .

## □形態素への品詞の付与

□太郎(**名詞**) | は (**助詞**) | お茶子 (**名詞**) | に (**助詞**) . . . .

## □形態素の語形変化の解析

□行< -> 行**き**ます

## 演習14-2 分かち書きを試みる

- 下記の文を人手により形態素解析（分かち書き）してみよ.
  - 私は人工知能概論を受講している.
  - にわにはにわうらにわにはにわにわとりがいる.



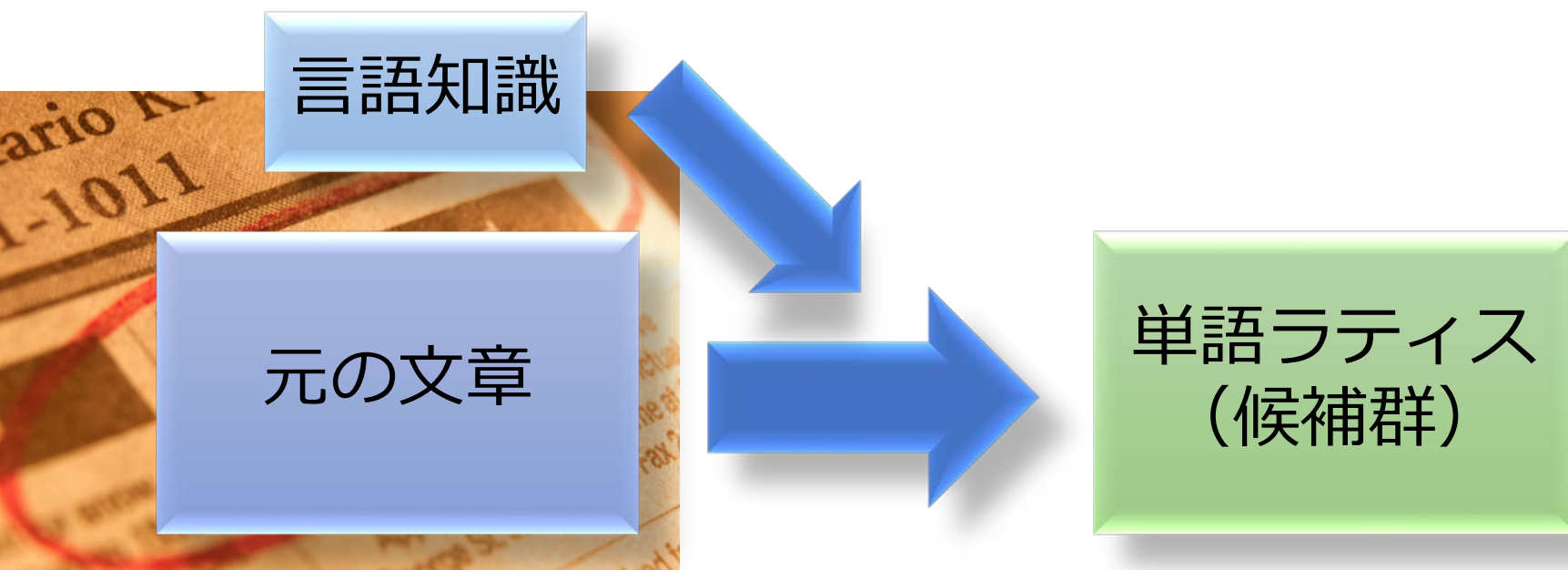
## 14.2.2 形態素解析に用いる情報

### □ 単語辞書

□ 語の品詞, 読み, 活用形などの情報を持つ.

### □ 接続辞書

□ どのような語が隣り合って並ぶことができるかについての情報を持つ.



# 単語ラティス(word lattice)

## □ 「やまだがないい」

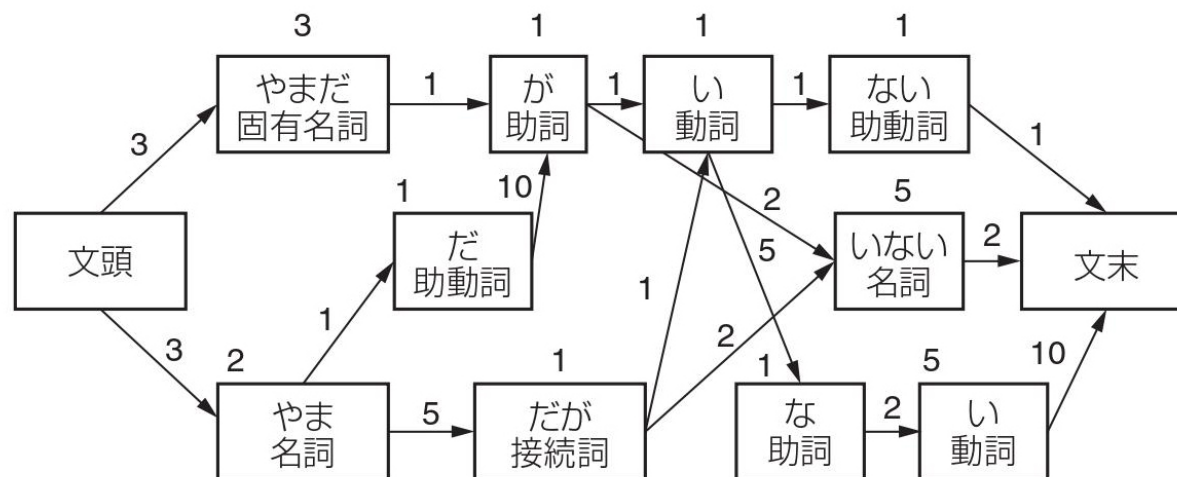
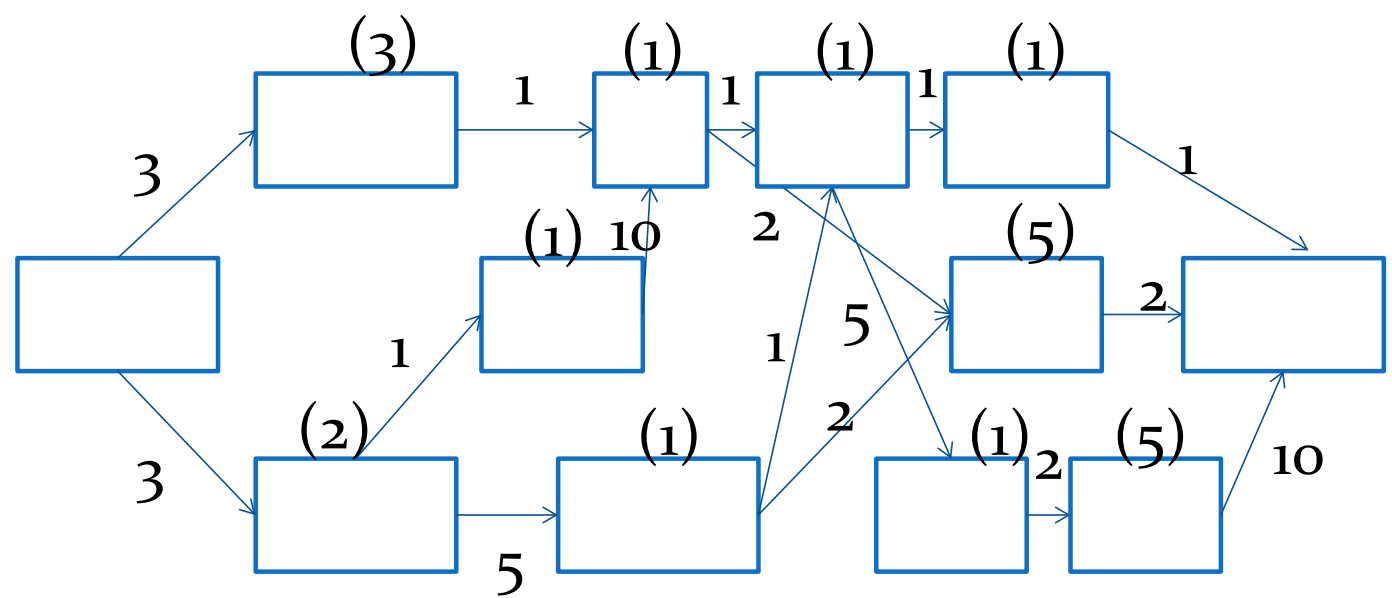
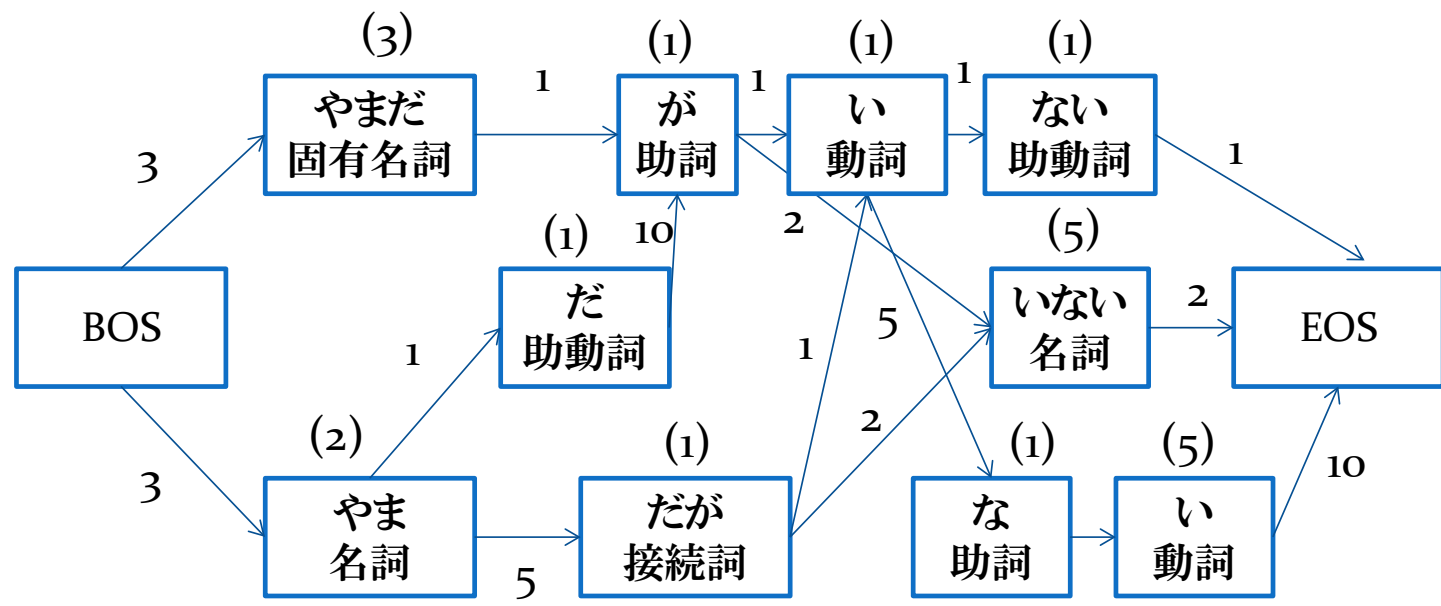


図 14.4 単語ラティスの例

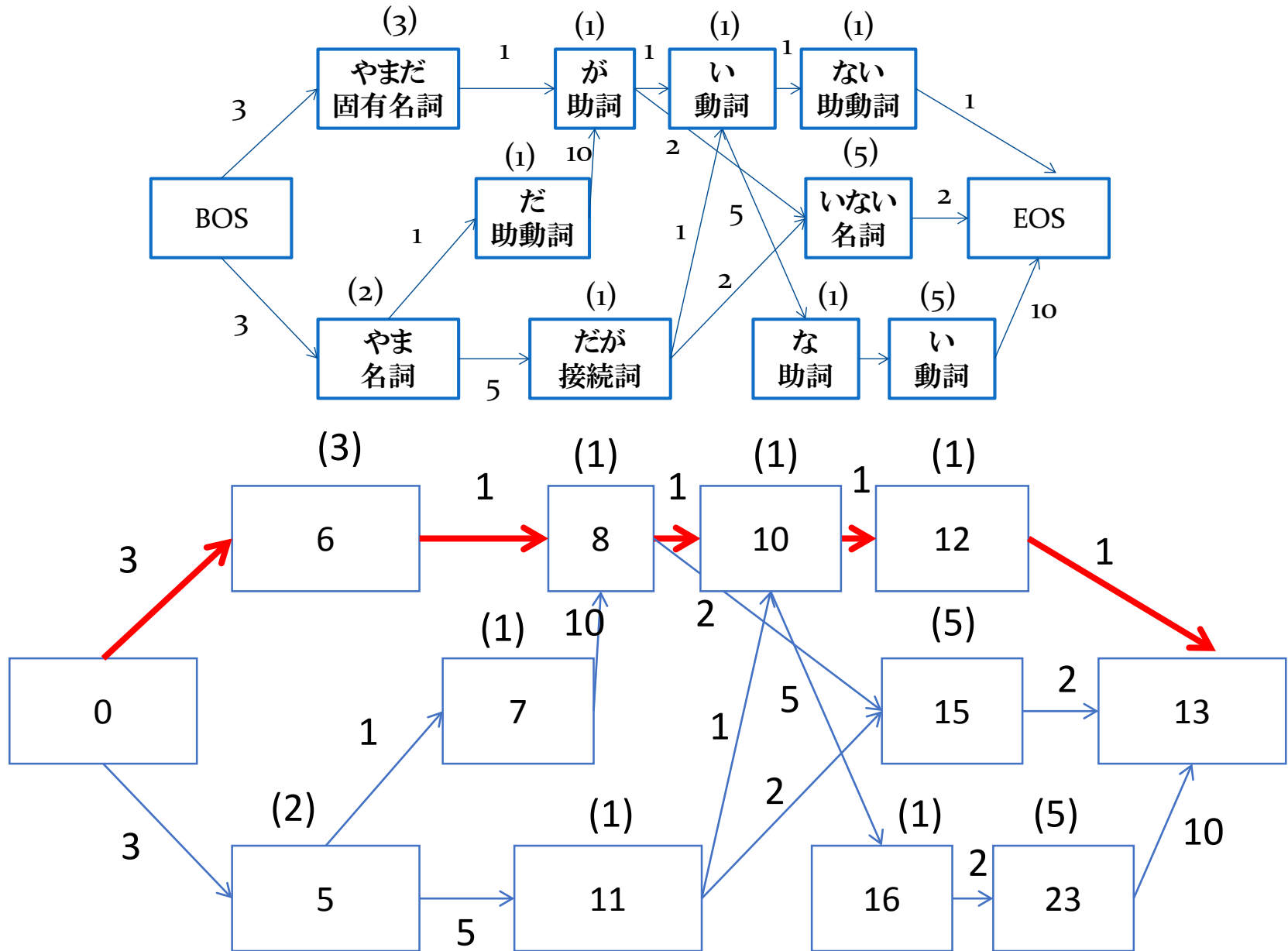
辞書に含まれている単語を形態素解析の候補としていくだけでは、形態素解析の結果は1通りには決まらない。

# 14.2.3 コスト最小法（ビタビアルゴリズム）

経路上におけるリンクのコストとノードのコストの和が最小化されるように経路探索せよ。



# コスト最小法の動的計画法による解決



## 14.2.4 統計的アプローチ

### • n-gramモデル

- 単語 $w_{t-n+1}, \dots, w_{t-1}$  が観測された後に, 単語 $w_t$  が観測される確率である n-gram 確率 $P(w_t | w_{t-1}, \dots, w_{t-n+1})$  を計算し, 情報として保持する.
- $n=1$  ユニグラム
- $n=2$  バイグラム
- $n=3$  トライグラム

### • 統計的アプローチでの形態素解析

- コスト最小化問題を単語列がバイグラムモデルにより生成される確率最大化問題に置き換える

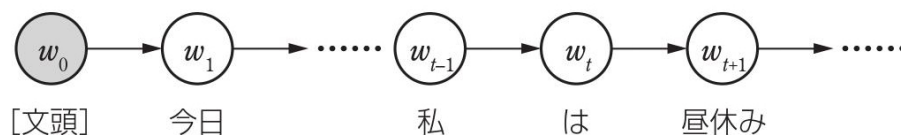


図 14.6 バイグラムによる文の生成過程を表すグラフィカルモデル



統計的自然言語処理

# 系列ラベリング

- またデータの系列に対してラベル付けを行う処理を**系列ラベリング**という
- 形態素解析における**品詞タグ**の推定はその一種である。**隠れマルコフモデル(hidden Markov model)**によりこの品詞タグ付けはある程度行うことが出来る。

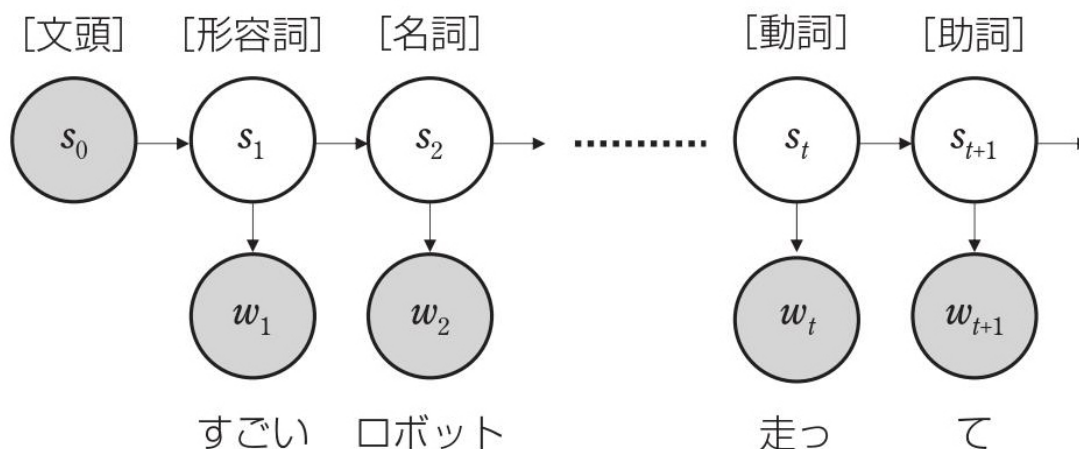


図 14.7 隠れマルコフモデルの生成過程を表すグラフィカルモデル

## 14.2.5 分類問題としてのアプローチ

- パターン認識問題としての取り扱い
  - 単語分割問題は、それぞれの文字の後に「単語が切れるか」「単語が切れないか」を判定する二値分類問題として捉えられる。

### 学習データ

- やまだ|が|たべ|た
- やまだ|も|行く|よ
- 今夜|が|やま|だ
- やまだ|が|たなか|と|あそぶ
- etc.etc.

パターン認識器

やまだがいない

# Contents

□14.1 自然言語処理

□14.2 形態素解析

□14.3 構文解析

□14.4 意味解析\*

□14.4 単語と文章のベクトル表現



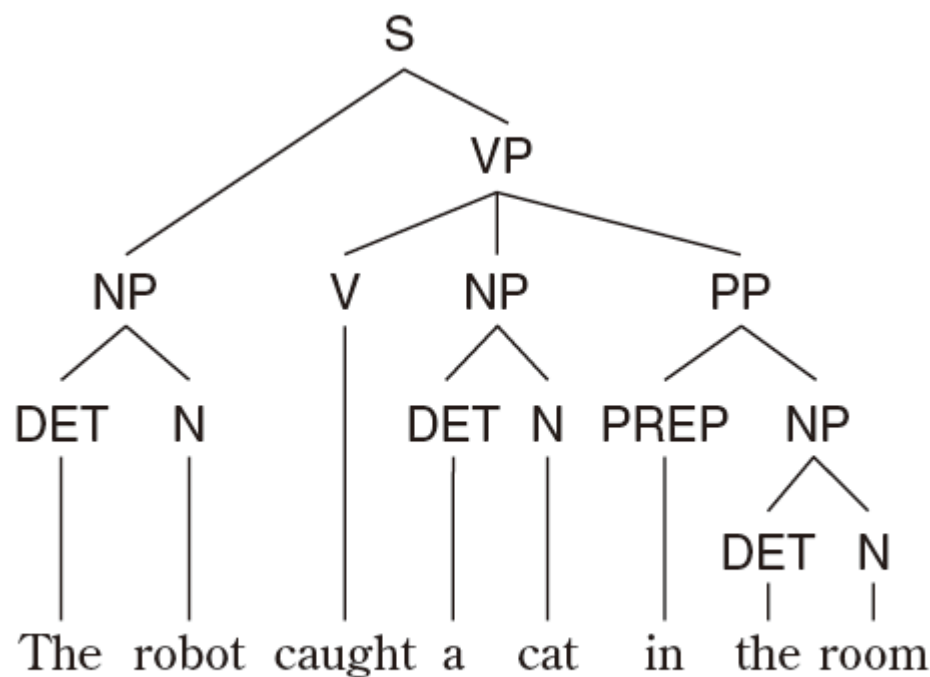
# 句構造文法

□構文木(syntactic tree)

□生成文法

(generative grammar)

□文脈自由文法(CFG)



(a) 句構造解析

表 14.1

文脈自由文法の生成規則集合の例

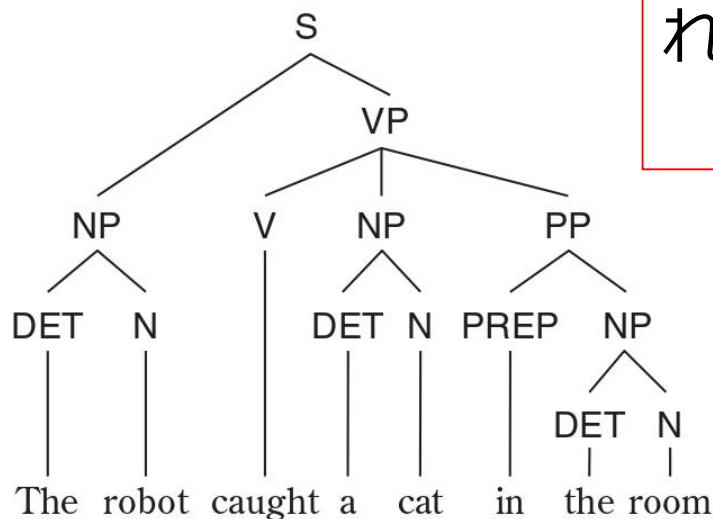
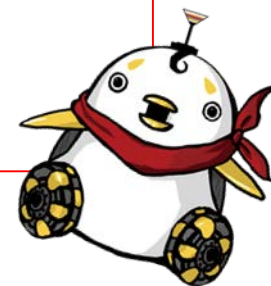
(1) $S \rightarrow NP VP$	(4) $VP \rightarrow V$	(7) $PP \rightarrow PREP NP$
(2) $NP \rightarrow N$	(5) $VP \rightarrow V NP$	
(3) $NP \rightarrow DET N$	(6) $VP \rightarrow V NP PP$	

# 14.3.1 句構造解析と係り受け解析

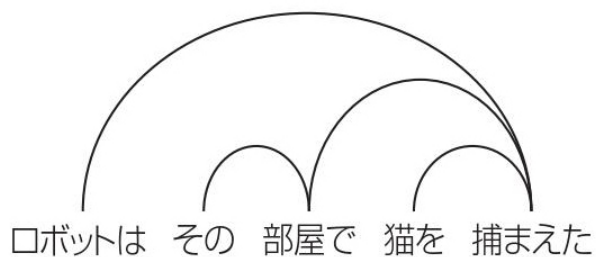
- 構文解析は与えられた言語の文法に従って、文法構造を解析することである。
  - 句構造解析・・・句構造文法に基づく（英語など）
  - 係り受け解析・・・依存文法に基づく（日本語など）

「白い机の上の箱をとってくれ」

⇒白いのは机？箱？



(a) 句構造解析



(b) 係り受け解析

## 14.3.2 構文解析のアルゴリズム

- トップダウン法(top-down method)
  - アーリー法(Earley parser)など
- ボトムアップ法(bottom-up method)
  - CKY 法(Cocke-Kasami-Younger algorithm)

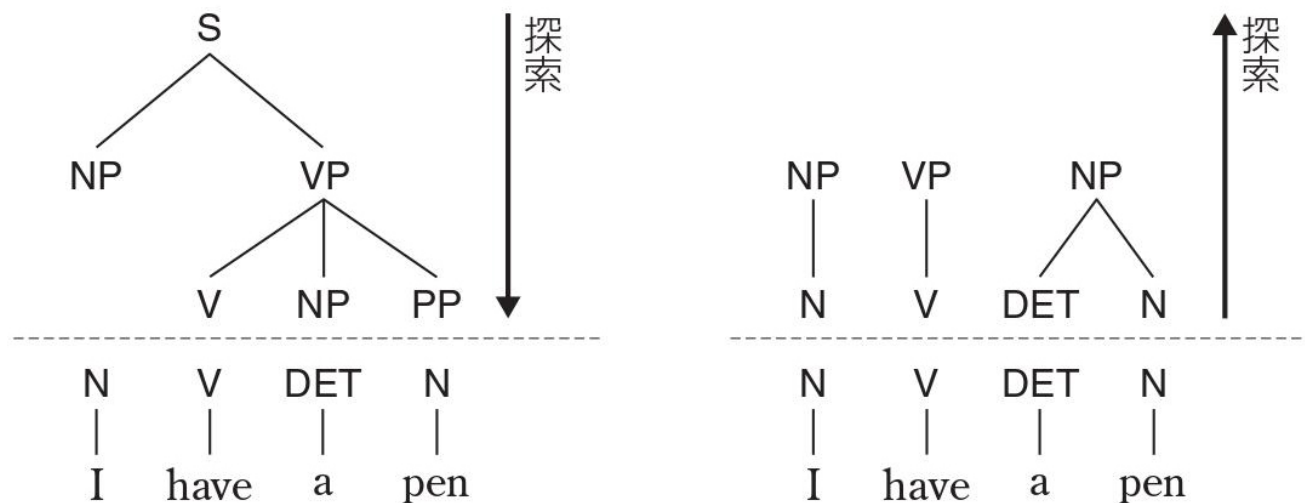


図 14.9 トップダウンな方法（左）とボトムアップな方法（右）による構文木の探索

※現在では確率モデルに基づく方法を始め様々な発展的な手法が取られる。  
※また近年では構文解析をせずに機械学習のみを用いて言語処理を行うことも多い。

# Contents

□14.1 自然言語処理

□14.2 形態素解析

□14.3 構文解析

□14.4 意味解析\*

□14.4 単語と文章のベクトル表現

## 14.4.2 意味ネットワークとフレーム理論

- 意味ネットワーク(semantic network)
  - 知識を概念 (concept) とそれらを結ぶ関係(relation)で記述することによって表現する.
- 概念はその性質を示す属性情報を持ち, 属性は具体的な値をとることにより, 概念が実体(instance)として存在する.

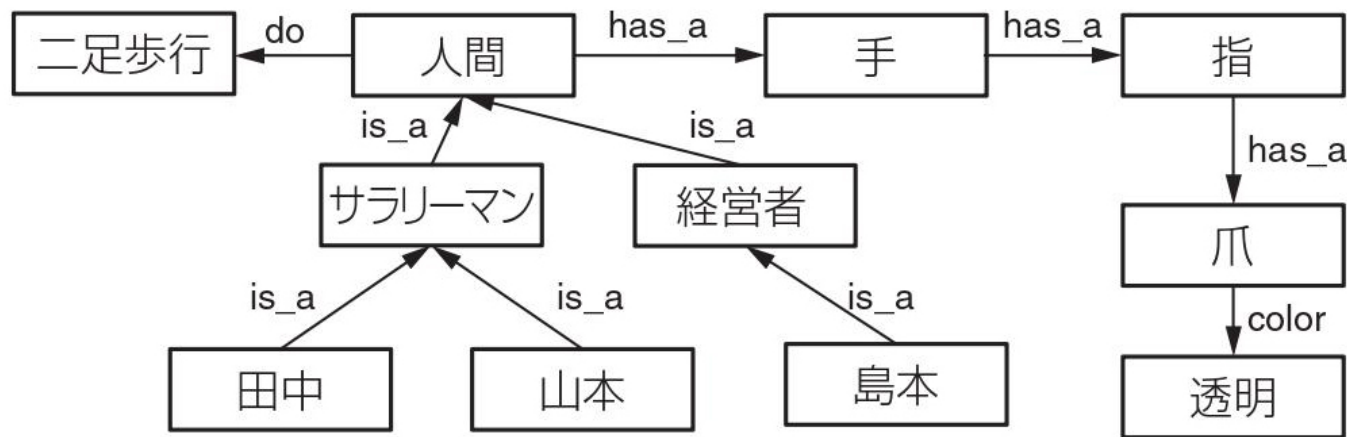


図 14.10 意味ネットワーク

# 階層構造(hierarchy)

□概念の階層構造を表現するために、以下のような関係が用いられる.

□is\_a: 上位-下位の関係を表現

□上位概念の持つ性質は基本的には下位概念にも受け継がれるという性質の継承を持つ.

□ex) 「動物」は「鳥」の上位概念

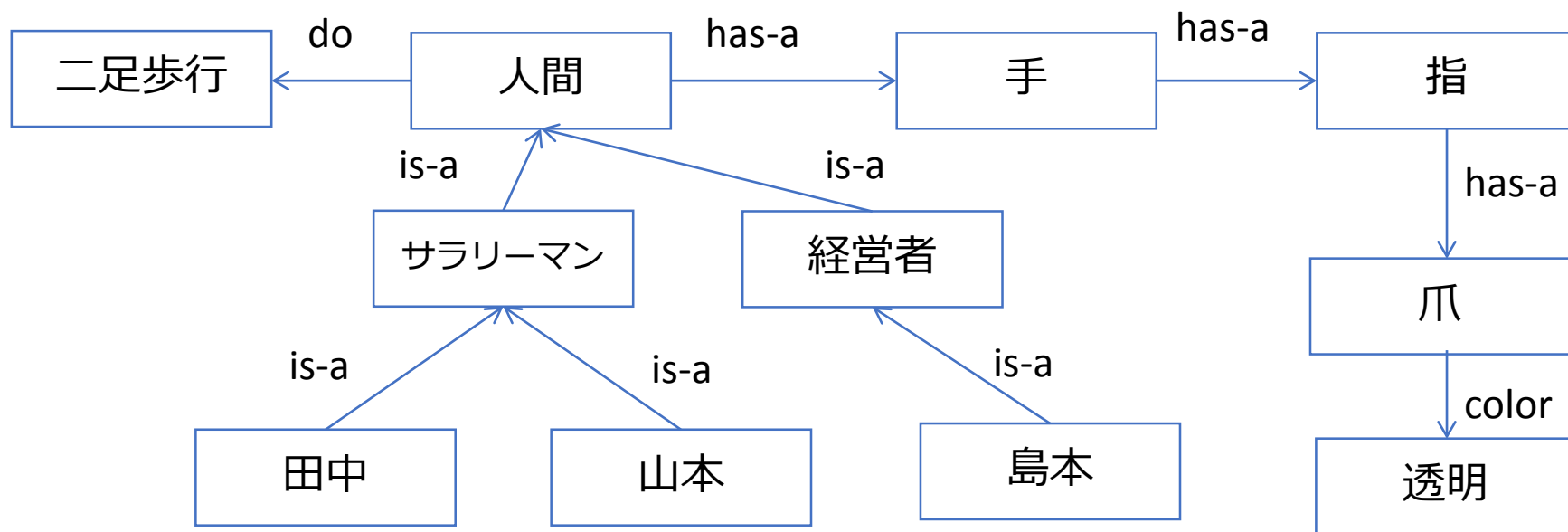
□has\_a: 部分-全体の関係を表現

□概念を構成する要素とその概念との関係を示す. 性質の継承は存在しない.

□ex) 「手」は「人」の部分である.

# 意味ネットワークにおける継承

- 田中は二足歩行する.
- 手は透明.....ではない.



# 意味ネットワークの特徴

- 利点

- 有向グラフによるネットワーク表示により視覚的に知識を表現することができるため、人が直感的に理解しやすい。
- 知識の追加・変更が比較的容易である。
- 概念の階層関係を定義することにより、概念が持つ属性の継承を階層関係において実現することができ、複雑な知識の構造化を実現できる。

- 欠点

- 推論規則を対象領域ごとに用意する必要がある。
- 知識の量が多くなると管理が難しくなる。
- 概念や関係の定義が任意であり、知識表現としての統一性が保証されない。

フレーム表現やオントロジーなどが開発されたが根本的にはよく似ている



## 演習14-3 意味ネットワーク

- 以下の知識を意味ネットワークで表現せよ.
- 関係としてはis\_a関係, has\_a関係, do関係を用いよ.
  - 「鳥は飛ぶ」
  - 「鳥は動物である」
  - 「カモメは鳥である」
  - 「ニワトリは鳥である」
  - 「人間は動物である」
  - 「人間は足を持つ」
  - 「山田は人間である」
  - 「人間は歩く」

# Contents

□14.1 自然言語処理

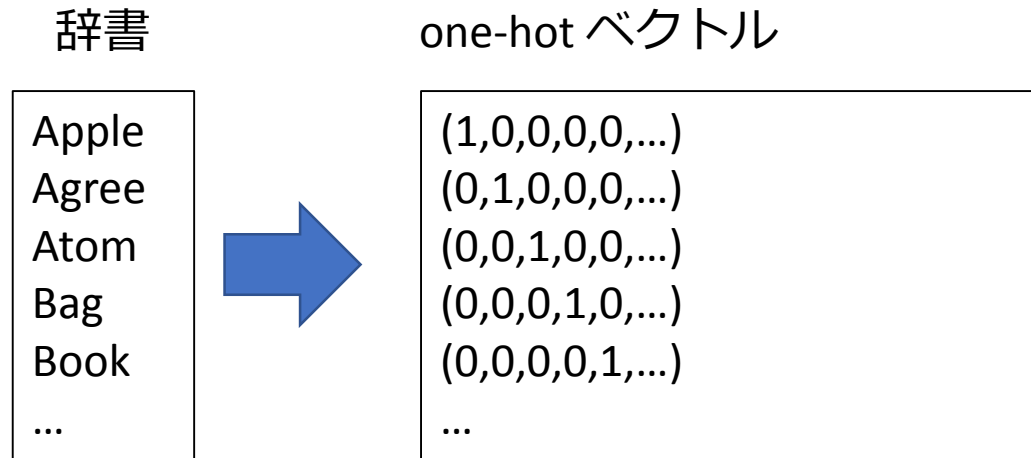
□14.2 形態素解析

□14.3 構文解析

□14.4 意味解析\*

□14.4 単語と文章のベクトル表現

## 14.5.1 one-hot ベクトル



- まず対象の文書を形態素解析により分かち書きを行う．また，必要に応じて語形変化や特殊な文字に関する処理を行い，辞書（ここでは単純に単語リストを指す）に含まれた単語だけにより構成される文書とする．
- ここで辞書に掲載されている単語には全て1から順にインデックスが振られているものとし，その最大数を $K$ とすると，それぞれの単語は1 of  $K$  表現のベクトルとして表される

## 14.5.2 文書データの簡便な表現

- Bag-of-Words(BoW表現)
  - テキストマイニングや文書のトピック分析などを行うために、簡便な表現を行う.
  - 単純に「単語」や「キーワード」がどれだけの数含まれているかをカウントする.

### Algorithm 14.1 BoW 表現

- ① 文書に含まれる各文を形態素解析にかける.
- ② 形態素解析の結果から、何らかの手法でキーワード抽出を行い、キーワード集合のリスト  $W = \{w_i\}$  を作成する.
- ③ 各文書  $d$  におけるキーワード  $w_i$  の出現回数  $c_{di}$  をカウントし文書ベクトル  $c_d = (c_{d1}, c_{d2}, \dots, c_{d\#(W)})$  を得る.

# 単語文書行列(term-document matrix)

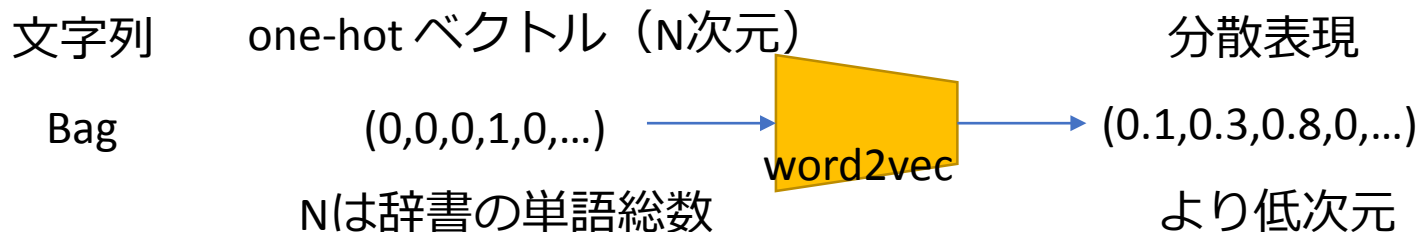
	文書1	文書2	文書3	文書4	文書5	.....
知能	3	1	2	0	0	
ロボット	1	0	4	0	0	
政府	0	0	0	0	2	
自衛隊	0	0	1	0	4	
安売り	0	0	0	5	0	
トナカイ	0	1	0	0	0	
サンタクロース	0	1	0	3	0	
⋮						

図 14.11 単語文書行列

トピック分析, 情報推薦, 検索などに用いる.

# 14.5.3 単語の分散表現

教師なし学習!



- 類似した単語は近いベクトルとして、意味的もしくは統語的な役割の異なる単語は遠いベクトルとして、より低次元なベクトルに表現する。
- 単語の表現学習を行うことに相当

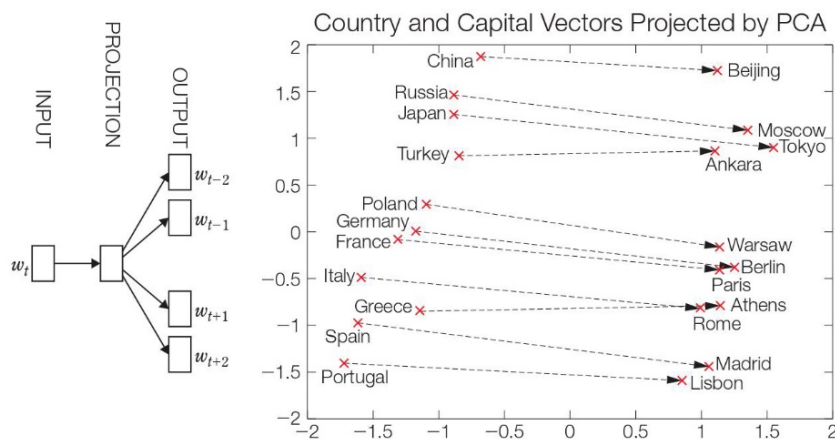


図 14.12 word2vec で用いられる skip-gram モデルと単語埋め込みの結果

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean: Distributed representations of words and phrases and their compositionality, *Proceedings of NIPS*, 2013.

word2vec : 単語の大規模なテキストコーパスから skip-gram に基づいて単語の分散表現を得ることのできるソフトウェア

## 14.5.4 文の埋め込み

教師なし学習!

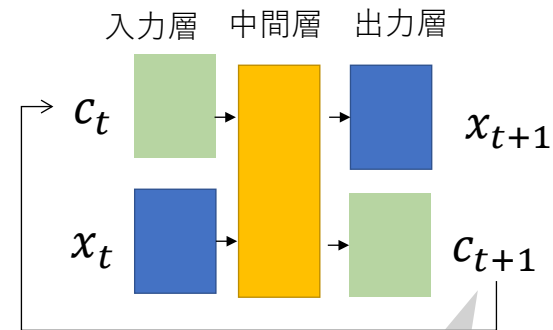
□文の意味は単語の集合によって規定されるのではなく、その並び方にも依存する。（“Tom hits the robot.”と“The robot hits Tom”では意味が違う。）

□ELMo

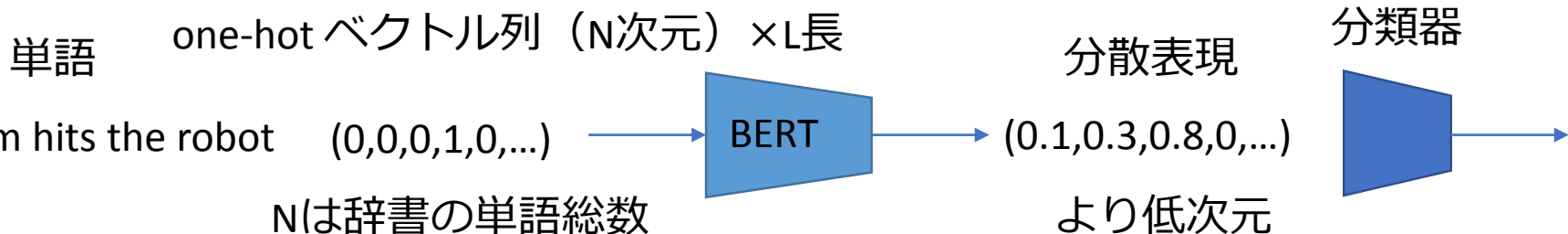
□リカレントニューラルネットワークの一種であるLSTMを双方向に用いたBi-LSTMに基づいて文の分散表現を得る。

□BERT

□深いニューラルネットワークを用いて前後の文脈を含めて入力し学習する点である。学習にはある文における一部の単語を欠損させてその単語を予測できるように、また二つの文が隣接した文かどうかを判定できるように訓練することで分散表現を学習させる。



文の情報を保持する



## 14.5.5 系列変換とエンコーダ・デコーダモデル

- 文の分散表現を介して、入力文を出力文へ変換するニューラルネットワーク。
- 後にアテンション（attention）の考え方が導入されて発展。
- それまでに比べて圧倒的に簡単に文法の知識もなく機械翻訳器が作れるようになった。

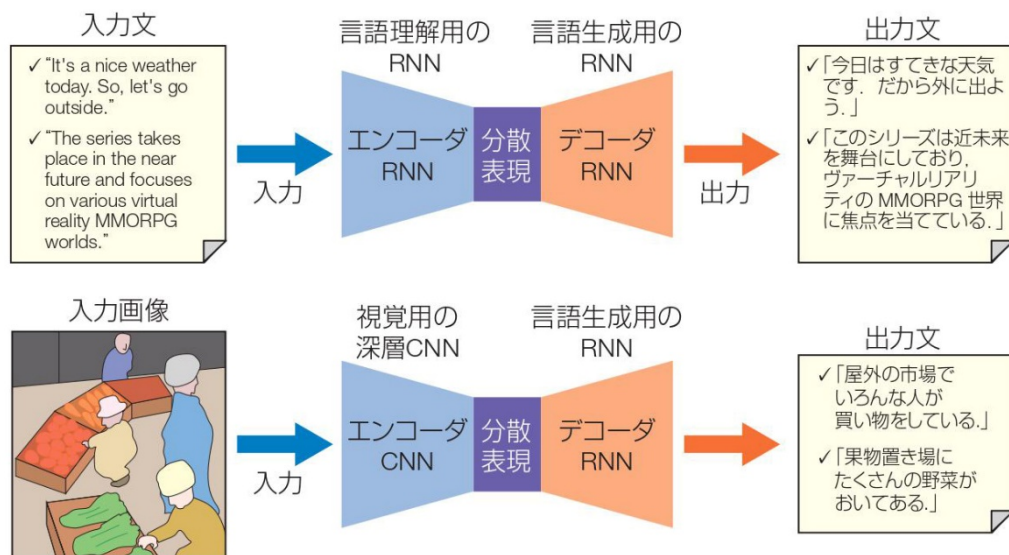


図 14.13 (上) 系列変換のモデル, (下) 画像説明文生成のモデル

O. Vinyals, A. Toshev, S. Bengio, D. Erhan: Show and tell: A neural image caption generator, *Proceedings of CVPR*, 2015.



## 演習14-4

- 以下の記述で最も不適切なものを選べ
  1. 系列変換モデルを用いることで文法に関する知識を一切人手で実装することなく機械翻訳をある程度実現出来る.
  2. 単語の埋め込み手法を用いることで似た意味を持つ単語に近いベクトルを割り当てる事ができる.
  3. 単語文書行列は構文解析の結果を保持する.
  4. 単語をベクトルとして表現するone-hotベクトルでは異なる二つの単語を表すベクトル間のユークリッド距離は単語によらず一定である.

# まとめ

- 自然言語処理の位置付けと応用分野について概観した.
- 形態素解析, 構文解析, 意味解析, 文脈解析の相互関係について例を用いて学んだ.
- 単語ラティスの最適経路を動的計画法により計算することで形態素解析を行うコスト最小化法について事例を交えながら学んだ.
- 構文解析における句構造解析と係り受け解析の区別について学んだ.
- 意味解析に関して意味ネットワークやシソーラスとは何かを学んだ.
- 単語や文の埋め込み表現に関して学んだ.