

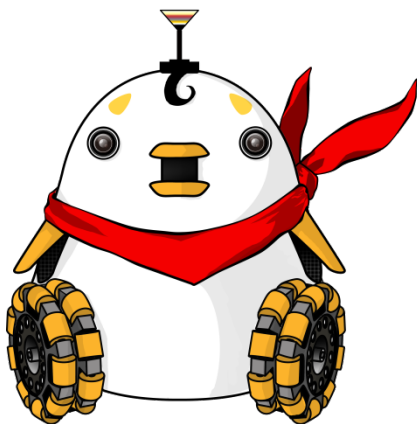
人工知能

第12回 言語と論理(1)

自然言語処理

立命館大学 情報理工学部 知能情報学科

萩原良信

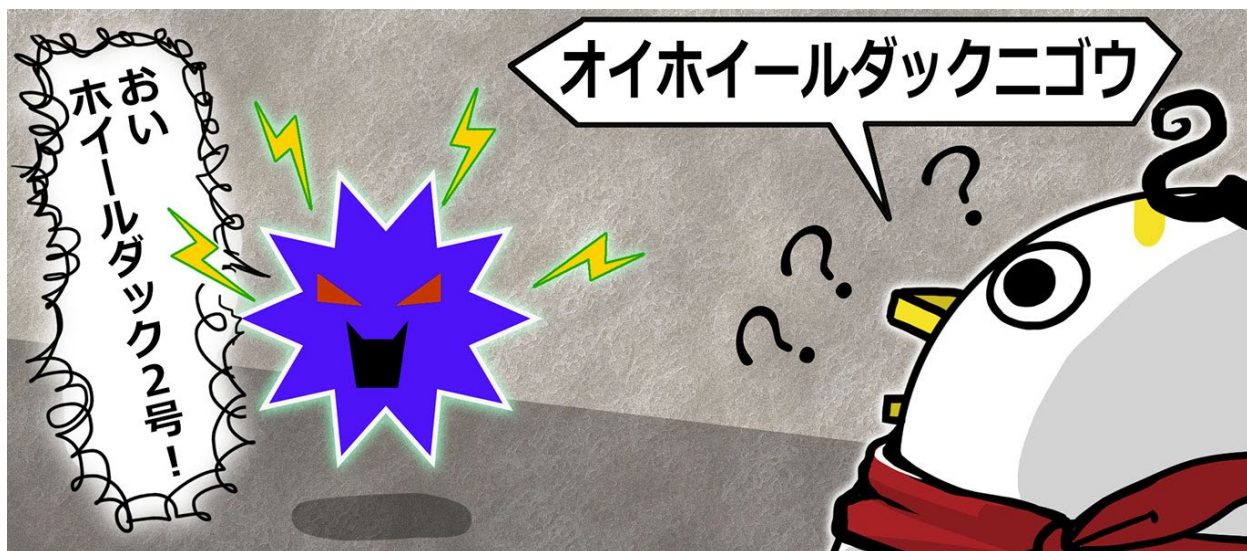


STORY 言語と論理(1)

- ホイールダック2号は迷路のゴールまで行く自信を深めた。もう、ゴールへの経路を探索するやり方だって、敵のかかし方だって覚えた。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だってできる。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。そうだ。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号に文法に関する知識, 語彙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

- 12.1 自然言語処理
- 12.2 形態素解析
- 12.3 構文解析
- 12.4 Bag-of-Words表現

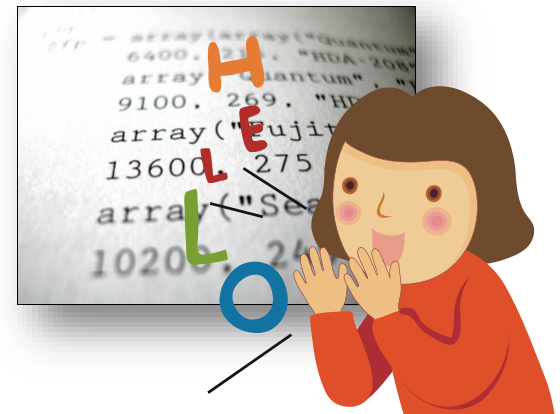
12.1.1 自然言語処理と応用分野

- 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



12.1.2 自然言語と人工言語

- 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,
 - × 小鳥のさえずり, 犬の鳴き声



12.1.3要素技術の関係

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.

12.1.3要素技術の関係

(1)形態素解析

①品詞活用の推定

名詞 助詞

動詞・活用

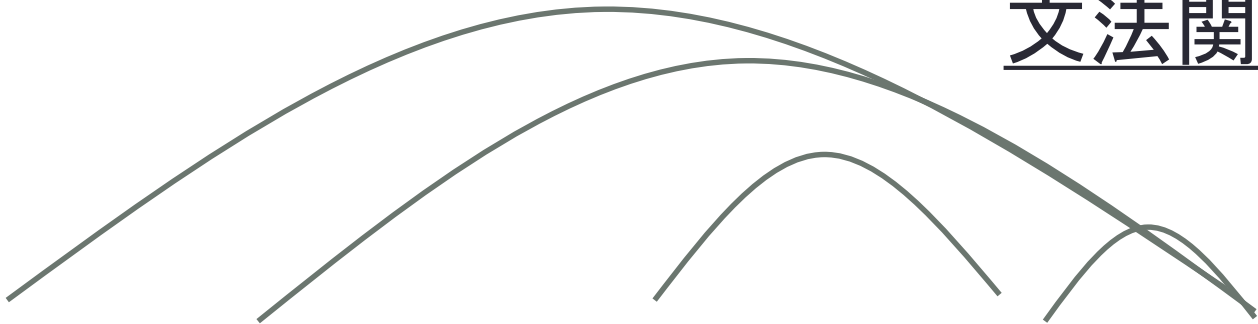
- 私|は|窓|から|降っ|て|いる|雪|を|見|た|.
- 傘|を|持っ|て|家|を|出|た|.
- それ|を|忘れ|て|き|た|.

②分かち書き

12.1.3要素技術の関係

(2)構文解析

文法関係の解析

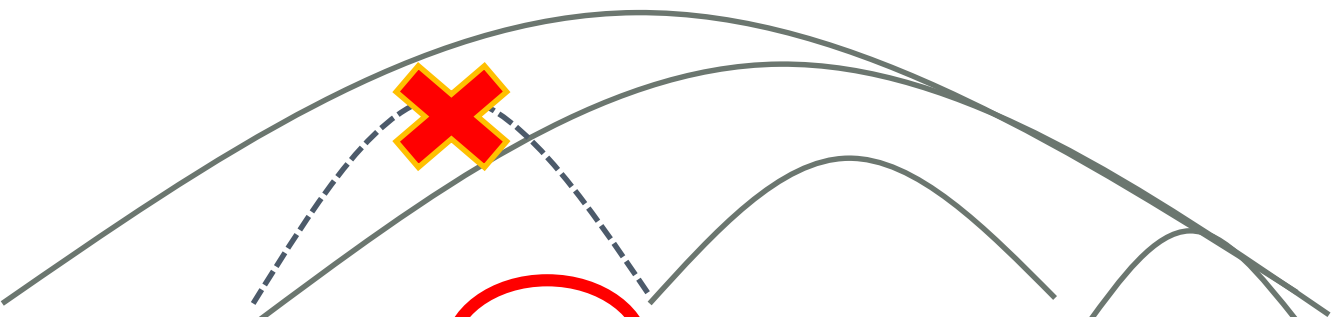


- 私|は||窓|から||降っ|て|いる||雪|を||見|た|.
- 傘|を|持っ|て|家|を|出|た|.
- それ|を|忘れ|て|き|た|.

□ 日本語では形態素を結合させた分節単位で構文解析することが多い.

12.1.3要素技術の関係

(3)意味解析

- 
- 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|.

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くさん落ちてくる。「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」

4 多く集まり寄ってくる。「一・るほど縁談がある」

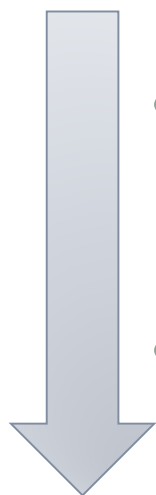
12.1.3要素技術の関係

(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

演習12-1 要素技術の関係

- 「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び。

- ① 形態素解析
- ② 構文解析
- ③ 意味解析
- ④ 文脈解析

Contents

- 12.1 自然言語処理
- 12.2 形態素解析
- 12.3 構文解析
- 12.4 Bag-of-Words表現

12.2.1 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

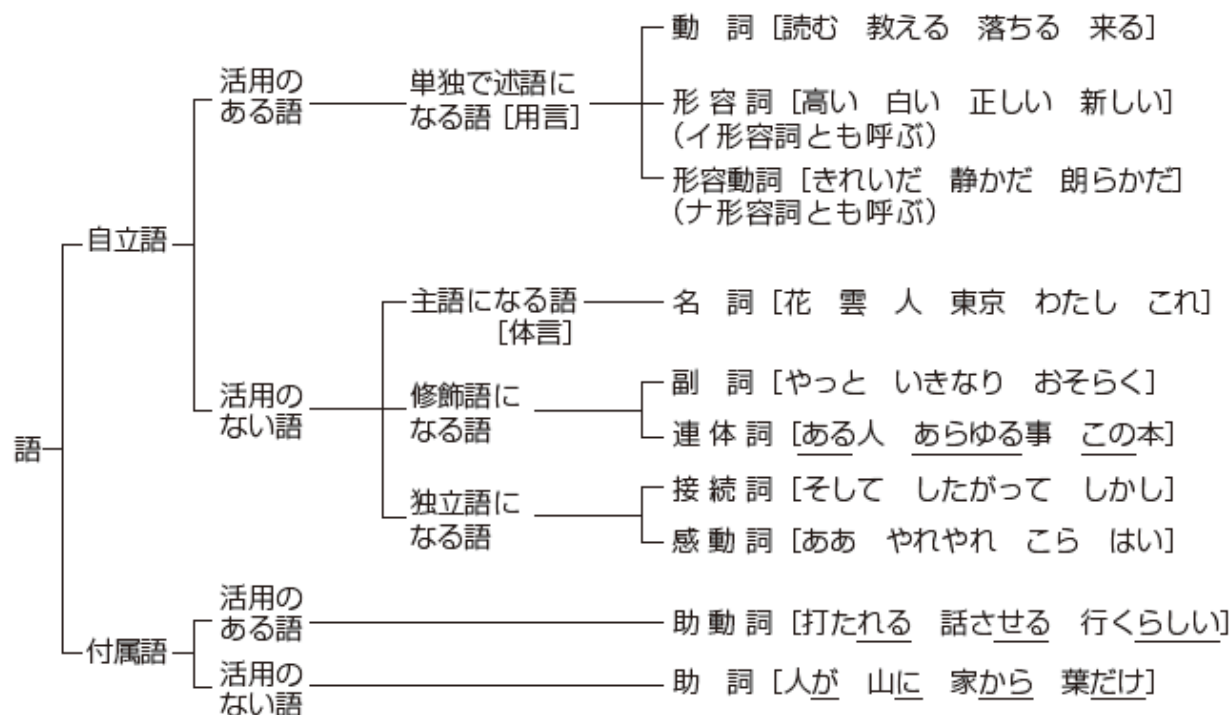


図 12.2 日本語の品詞分類

形態素解析

- 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(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
 - 行く → 行きます

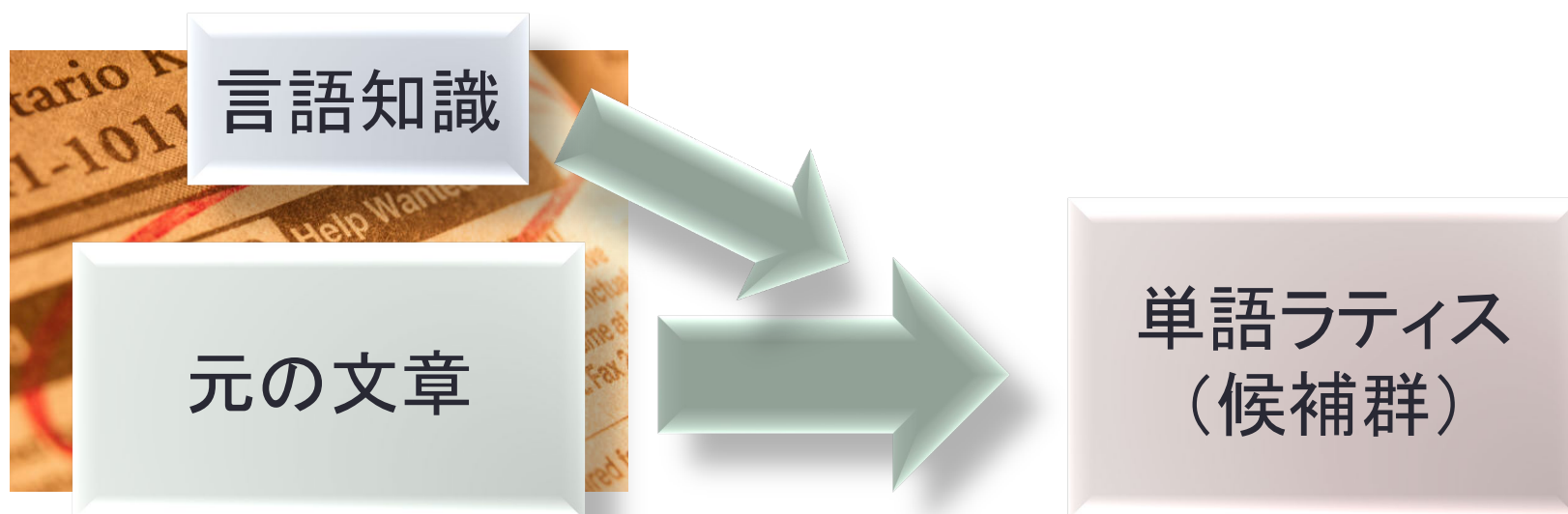
12.2.2 形態素解析に用いる情報

- 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

- 「やまだがない」

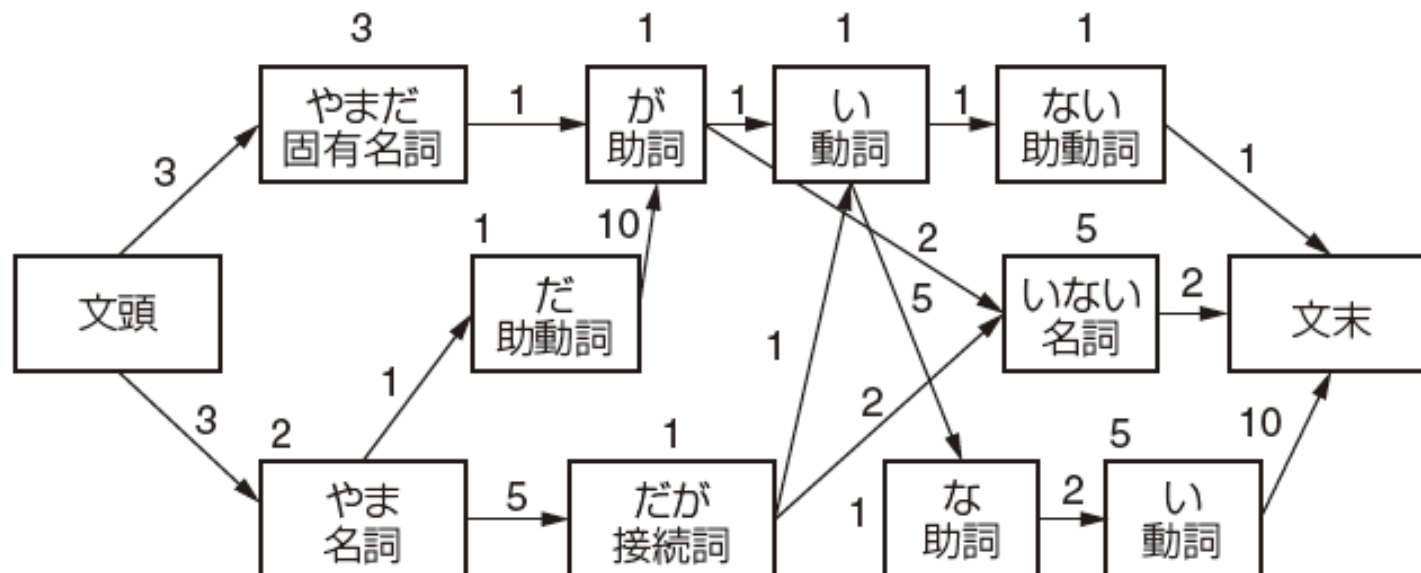


図 12.3 単語ラティスの例

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。

12.2.3 ヒューリスティックな手法

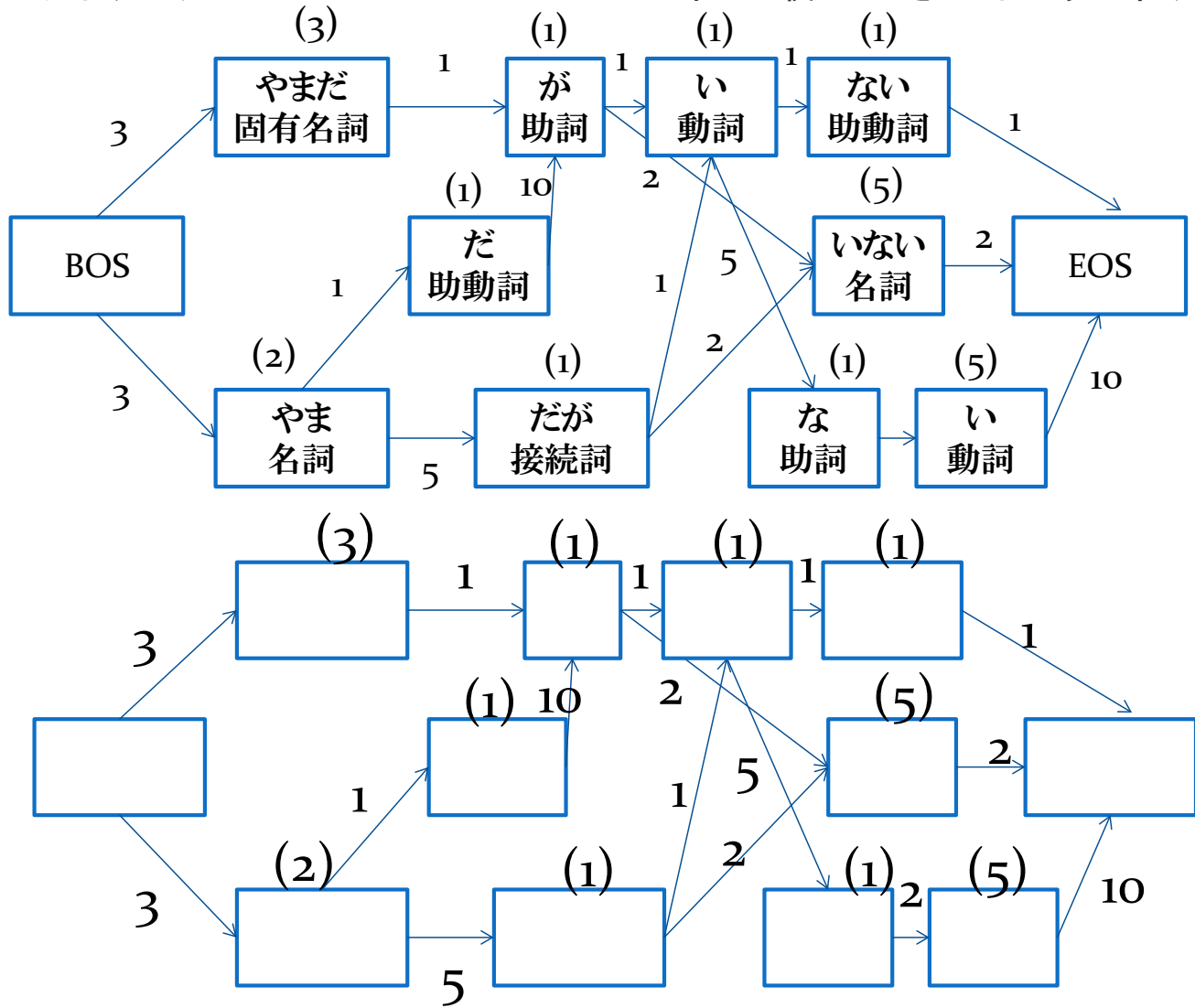
- 各形態素、文節の長さを最長にする候補を選択
- 文中の形態素、文節の数を最小にする候補を選択

表 12.1 ヒューリスティックな形態素解析の候補選択法

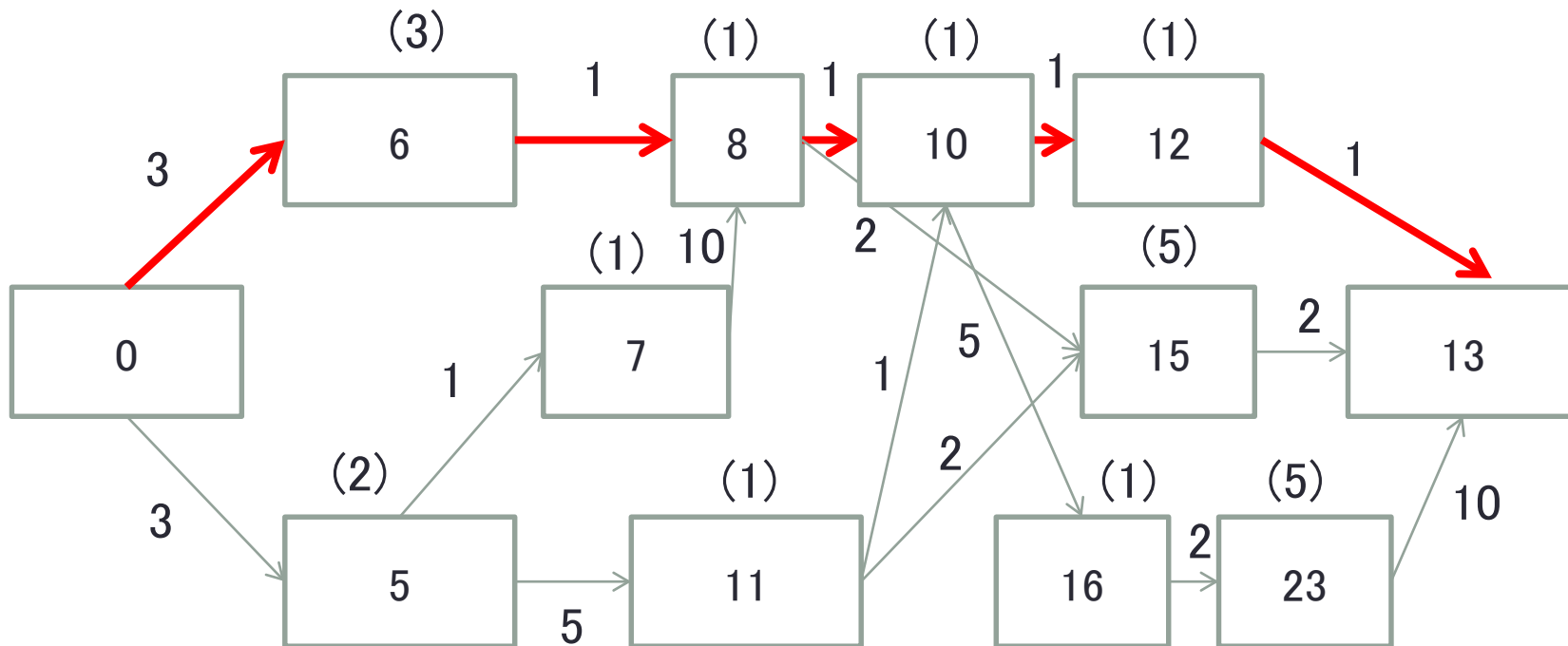
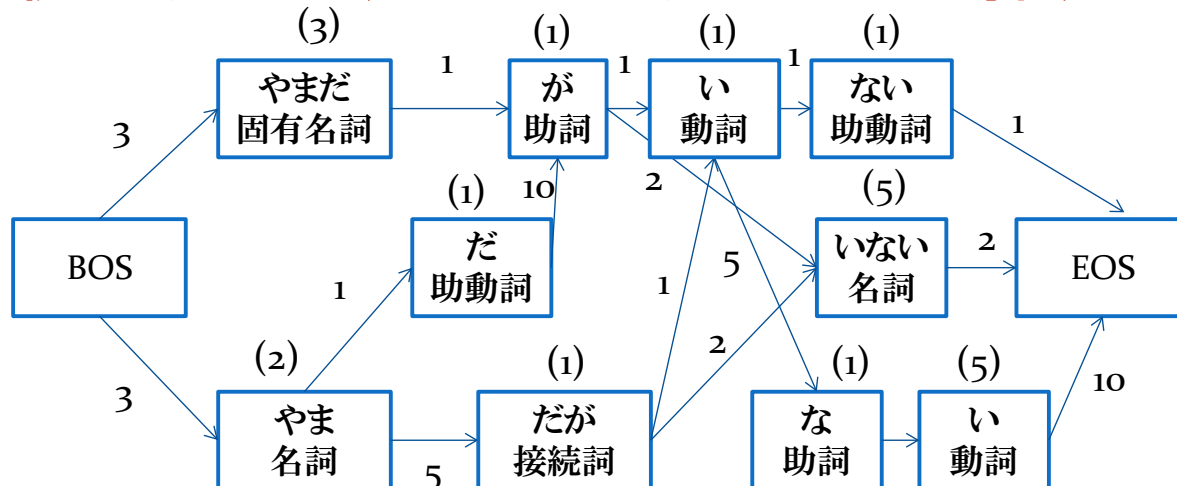
最長一致法	文頭から長い形態素の候補を優先して選択する.
二文節最長一致法	文頭から二文節ごとの長さが長い解を優先して選択する.
形態素数最小法	複数の解析結果の候補の中で形態素数が最も少ない結果を選択する.
文節数最小法	複数の解析結果の候補の中で文節数が最も少ない結果を選択する.

12.2.4 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



12.2.5 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nグラムモデルにより生成される確率を最大化する問題に置き換える



統計的自然言語処理

12.2.6 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。
 - 切れる場合と切れない場合を事前に学習しておく。

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

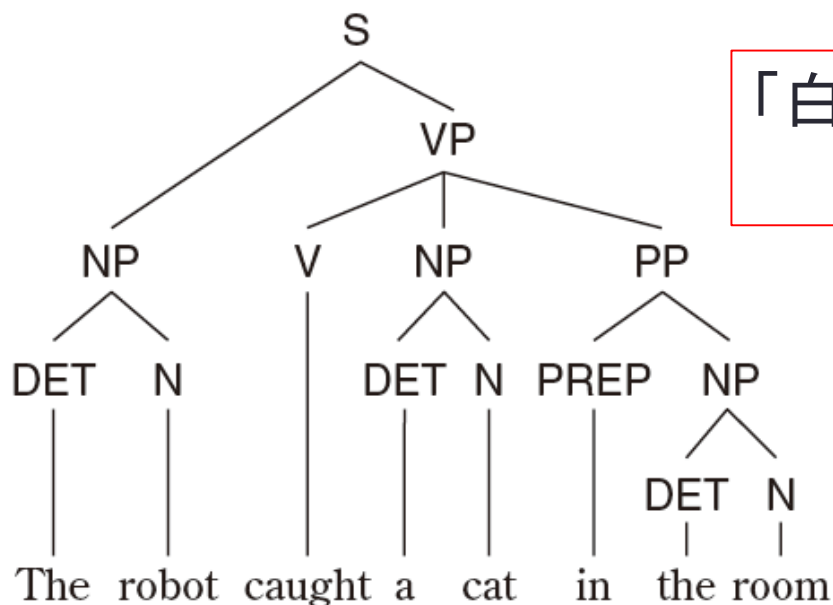
やまだ|が|いない

Contents

- 12.1 自然言語処理
- 12.2 形態素解析
- 12.3 構文解析
- 12.4 Bag-of-Words表現

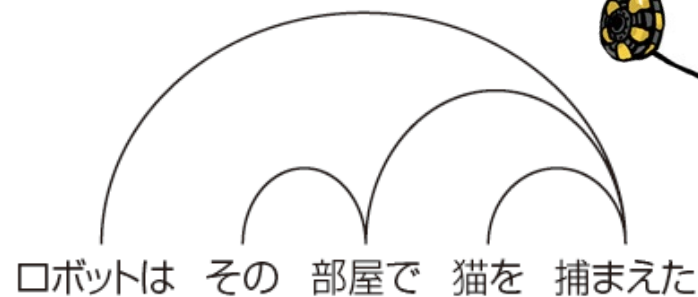
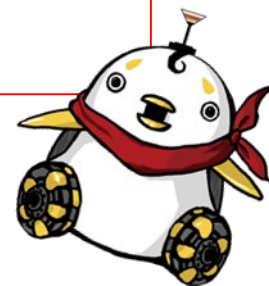
12.3.1 句構造解析と係り受け解析

- 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）



(a) 句構造解析

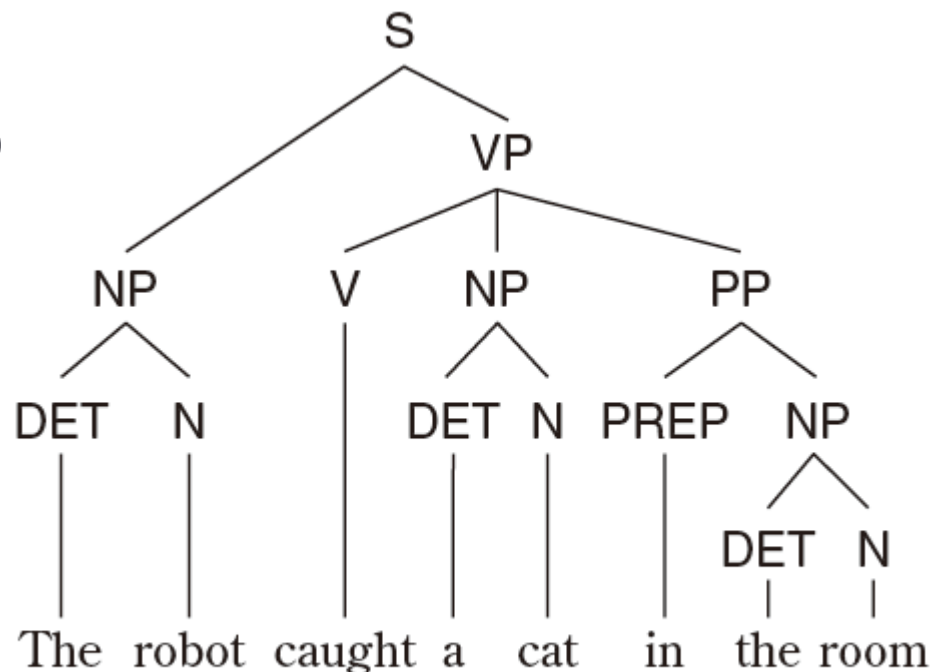
「白い机の上の箱をとってくれ」
⇒ 白いのは机？ 箱？



(b) 係り受け解析

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 12.2

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) $VP \rightarrow V NP PP$	

12.3.2 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)

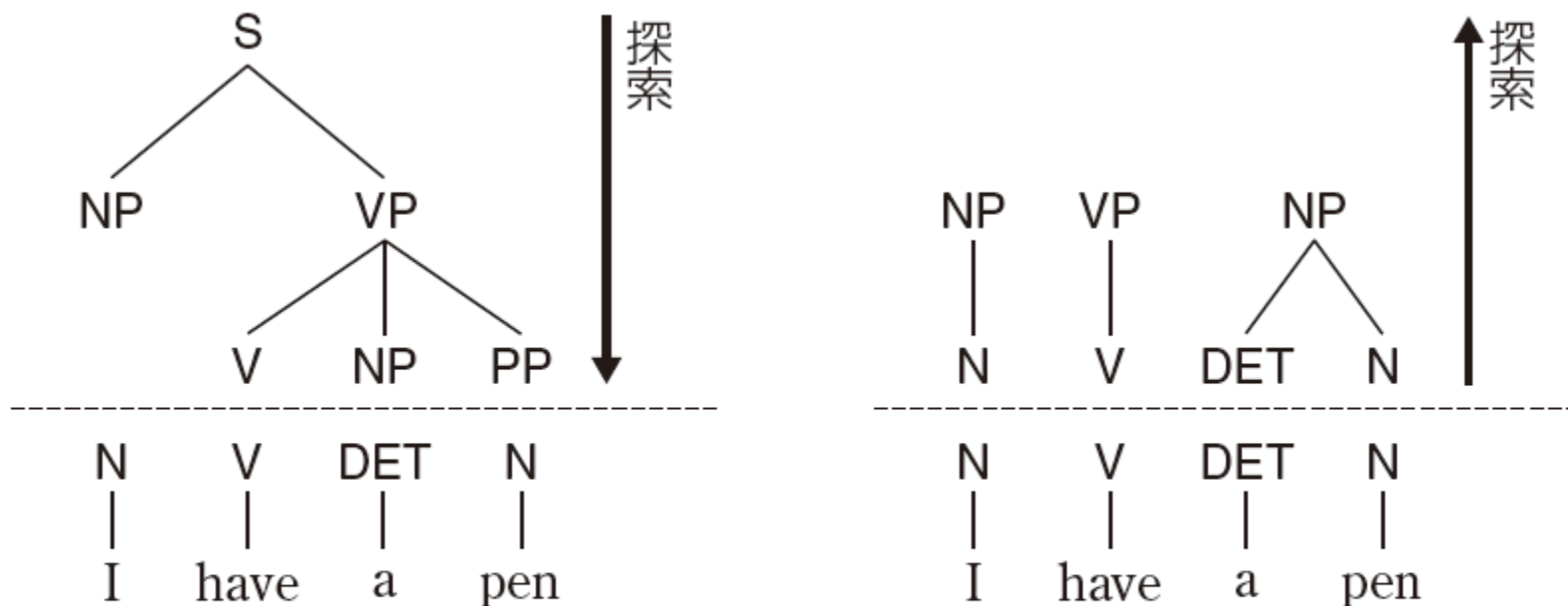


図 12.6

トップダウン法（左）とボトムアップ法（右）による構文木の探索

Contents

- 12.1 自然言語処理
- 12.2 形態素解析
- 12.3 構文解析
- 12.4 Bag-of-Words表現

12.4.1 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(Data Mining)や文書のトピック(topic)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 12.1 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	
⋮						

トピック分析, 情報推薦, 検索などに用いる

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.
- トップダウン法とボトムアップ法による構文木探索法の概略を理解した.
- 文書データの簡便な表現であるBag-of-Words 表現nについて学んだ.