

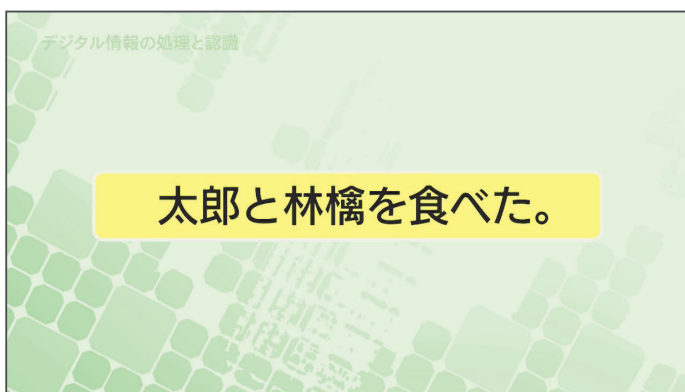
デジタル情報の処理と認識 ('22)

第2回 自然言語処理

1



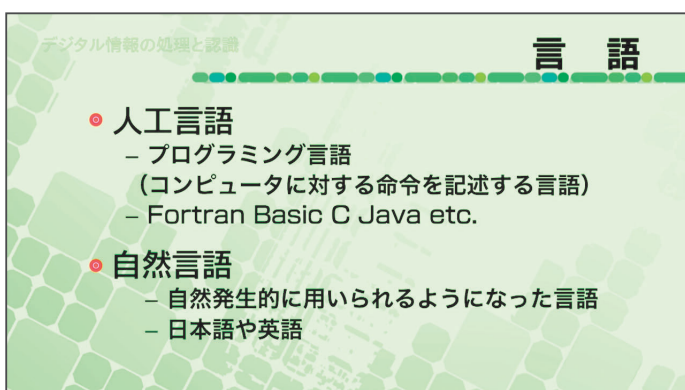
2



3



4



デジタル情報の処理と認識 ('22)

第2回 自然言語処理

5

デジタル情報の処理と認識

自然言語処理

- 日本語や英語といった自然言語の処理
- 処理手順
 - 形態素解析
 - 構文解析
 - 意味解析
 - 文脈解析

6

デジタル情報の処理と認識

形態素解析

- 形態素：triangle = tri + angle
- 形態素解析：～単語レベルの処理
- This is a pen. ↔ これはペンです。
- 日本語の形態素解析
 - 単語の分割（分かち書き）がメイン

7

デジタル情報の処理と認識

形態素解析

- 辞書を利用
- 文章と一致するように 単語を探し出す
 - 複数の候補が考えられる
- 経験則
 - 最長一致法や 分割数最小法

8

デジタル情報の処理と認識

最長一致法

- 「中国語学校」を例に単語分割
- 辞書：「中 中国 中国語 学 学校」等
- 中/国語学校 中国/語学校 **中国語** 学校
- 次に「学校」を分割
- 学/校 **学校** → 中国語 / 学校

第2回 自然言語処理

9

デジタル情報の処理と認識

分割数最小法

- 「中国語学校」を例に単語分割
- 辞書：「中 中国 中国語 学 学校」等
- 中 / 国語 / 学校 → 3分割
- 中国語 / 学校 → **2分割**
- 中国語 / 学校

10

デジタル情報の処理と認識

品詞の接続関係の利用

- 2つの単語が接続するか否かによる制限
 - 形容詞の後：名詞○ 助詞×
 - 文頭：名詞○ 助詞× など
- 最長一致法や分割数最小法と併用可能

11

デジタル情報の処理と認識

コスト最小法

- コストを設定し 最小となる分割を探す
 - つながりやすい品詞には低いコスト
 - 出現しやすい単語には低いコスト
- コストの設定
 - 人手で行う方法
 - 多量の文書データを用いて統計的に設定

12

デジタル情報の処理と認識

形態素解析結果

太郎	太郎	タロウ	タロー	名詞-固有名詞-人名-名
と	と	ト	ト	助詞-並立助詞
林檎	林檎	リンゴ	リンゴ	名詞-一般
を	を	ヲ	ヲ	助詞-格助詞-一般
食べ	食べる	タベ	タベ	動詞-自立 一段 連用形
た	た	タ	タ	助動詞 特殊・タ 基本形
。	。	。	。	記号-句点

EOS

デジタル情報の処理と認識 ('22)

第2回 自然言語処理

13

デジタル情報の処理と認識

形態素解析ソフト

- ChaSen
 - 奈良先端科学技術大学院大学で開発
- JUMAN
 - 京都大学で開発 最も古い形態素解析ソフトの1つ
- MeCab
 - 辞書に載っていない単語の処理を定義できる
- KAKASI
 - 全文検索システムNamazuなどで利用

14

デジタル情報の処理と認識

構文解析

15

デジタル情報の処理と認識

構文解析

- 形態素解析を行うことで
「太郎と林檎を食べた」という文を
「太郎/と/林檎/を/食べ/た」と単語分割
- 単語からどのように句や節が構成され また
句や節からどのように文が構成されているか
といった文の構造を解析するのが構文解析

16

デジタル情報の処理と認識

構文解析

- 文法規則を形態素解析結果に適用
- 英語「in Japan」 日本語「日本に」

文	→	後置詞句	動詞句	名詞	→	太郎
後置詞句	→	名詞	助詞	名詞	→	林檎
後置詞句	→	名詞句	助詞	助詞	→	と
動詞句	→	後置詞句	動詞句	助詞	→	を
動詞句	→	動詞	助動詞	動詞	→	食べる
名詞句	→	名詞	助詞	名詞	→	た

第2回 自然言語処理

17

デジタル情報の処理と認識

トップダウン法

- 文から文法規則を繰り返し適用
文 → 後置詞句 + 動詞句

↓

名詞 + 助詞

↓

太郎 と

18

デジタル情報の処理と認識

ボトムアップ法

- 単語列から文法規則を連続的に適用
太郎 + と + 林檎 → 名詞句

↓

- 名詞句 (太郎と林檎) + を → 後置詞句

19

デジタル情報の処理と認識

構文木1

20

デジタル情報の処理と認識

構文木2

デジタル情報の処理と認識 ('22)

第2回 自然言語処理

21

デジタル情報の処理と認識

日本語の構文解析

- 日本語は、語順の自由度が高い
- 助詞によって意味が決まる場合も多い

↓

- それぞれの文節の係り受け関係を明らかにすることを目的として構文解析が行われる場合も多い

22

デジタル情報の処理と認識

日本語の構文解析

- 文節の係り受け関係を決めるための方法としては経験則を用いる方法や統計的な性質を利用する方法などがある
- 「名詞+ガ」太郎が
 - 「動詞連用形」読み
 - 「動詞連体形」読む
 - 「名詞+ヲ」本を

23

デジタル情報の処理と認識

日本語の構文解析

- 前の文節から後ろの文節に係る
- より近い受け要素に係る

赤い 林檎と バナナ

↑

24

デジタル情報の処理と認識

日本語の構文解析

- 係り受け関係は交差しない

例外

「うどんを 香川に 食べに 行く」

デジタル情報の処理と認識 ('22)

第2回 自然言語処理

25

デジタル情報の処理と認識

単語の共起

- 「**電話**で話している 太郎を 見た」
- 「**双眼鏡**で話している 太郎を 見た」

Blue arrows indicate the relationship between the words in parentheses and the words they modify.

26

デジタル情報の処理と認識

コーパス

- 多量の例文を収集したもの
 - 単純に例文を集めたもの
 - 係り受け関係などの情報を付与したもの
- 利用方法
 - 統計的な性質を利用
 - 類似する用例を探す

27

デジタル情報の処理と認識

構文解析ソフト

- KNP
 - JUMANの形態素解析結果から構文解析を行う
- CaboCha
 - サポートベクトルマシンと呼ばれる分類アルゴリズムを用いて係り受け関係を解析

28

デジタル情報の処理と認識

構文解析結果

- KNPによる構文解析結果

太郎と<P>┐
 林檎を<P>-PARA┐
 食べた。
 EOS

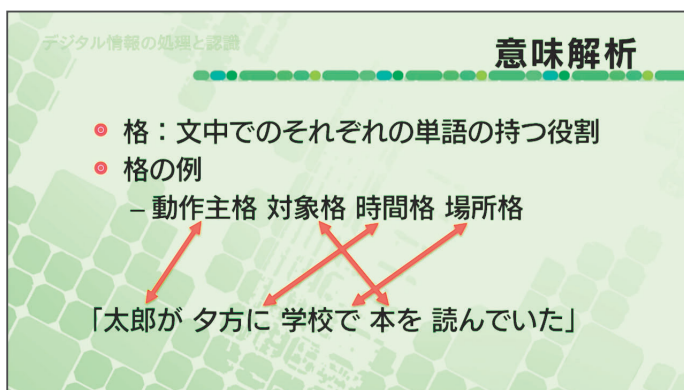
デジタル情報の処理と認識 ('22)

第2回 自然言語処理

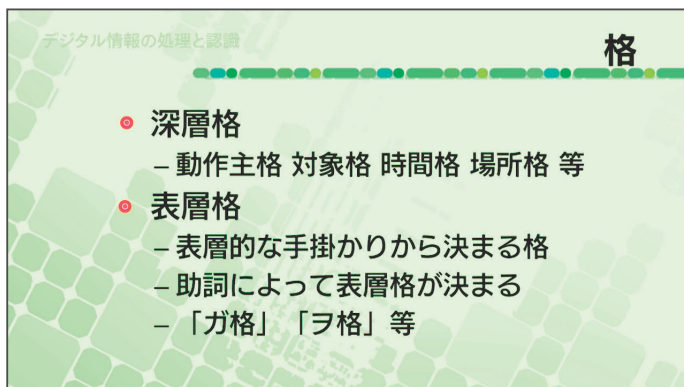
29



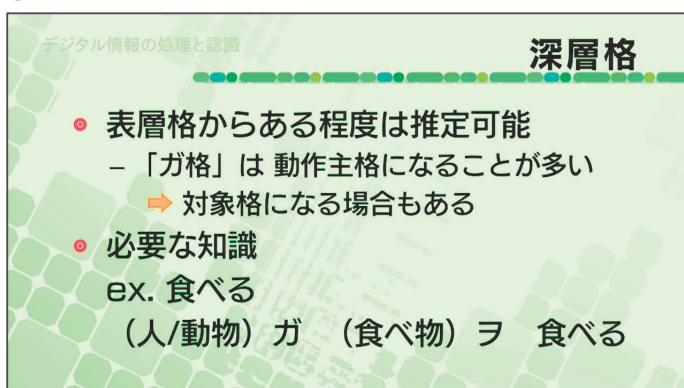
30



31



32



第2回 自然言語処理

33

デジタル情報の処理と認識

格フレーム

- 格フレーム
 - 動詞がどのような格を持つか
 - 表層格と深層格の対応などを記述
- 格フレーム辞書
 - IPALの動詞辞書
 - 京都大学格フレーム

34

デジタル情報の処理と認識

意味解析結果

35

デジタル情報の処理と認識

文脈解析

- 前後の文脈を用いることで行われる処理
 - 文と文の関係を明らかにする
 - 「それ」が何かを推定する
 - 省略された主語などを推定する

36

デジタル情報の処理と認識

2人で食事をした。

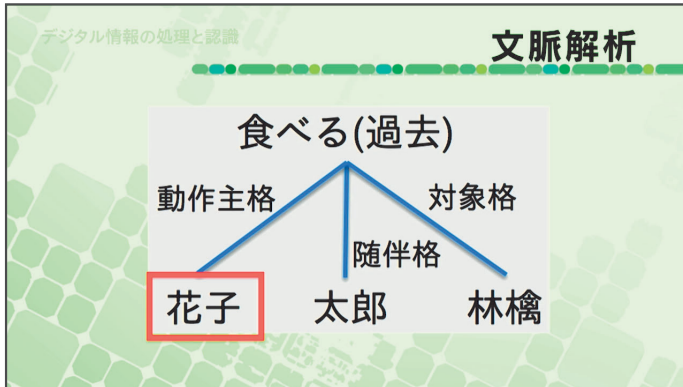
花子は林檎をむいた。

太郎と林檎を食べた。

デジタル情報の処理と認識 ('22)

第2回 自然言語処理

37



38

