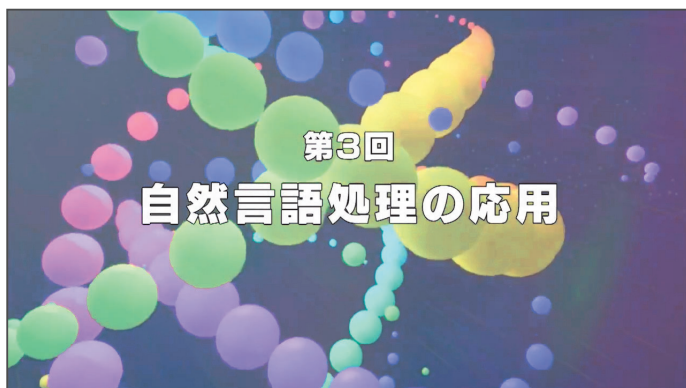


第3回 自然言語処理の応用

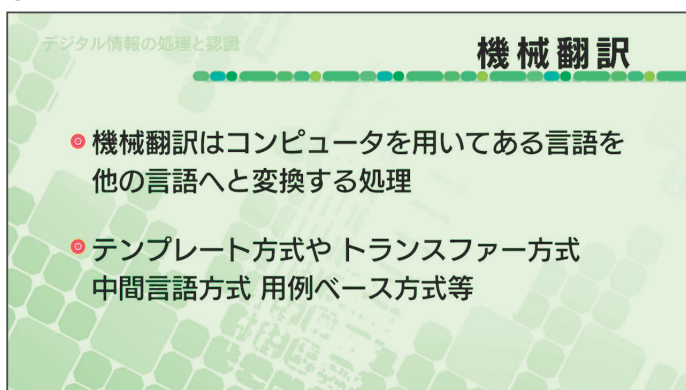
1



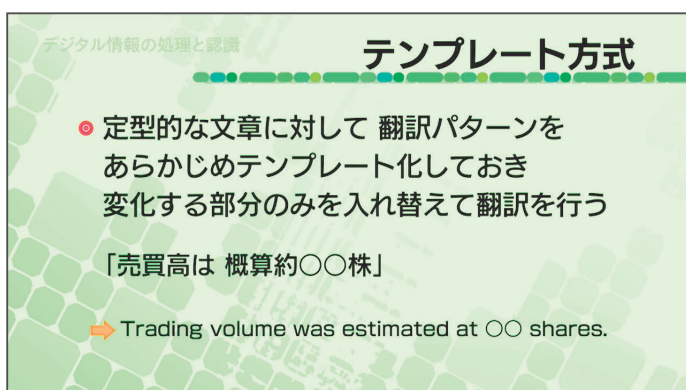
2



3



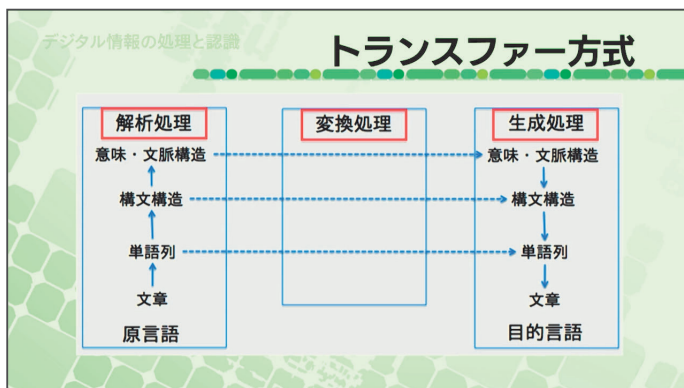
4



デジタル情報の処理と認識 ('22)

第3回 自然言語処理の応用

5



6

デジタル情報の処理と認識

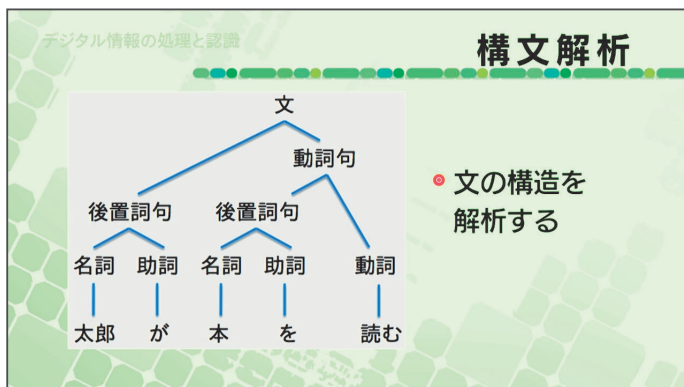
解析処理

- 「太郎が本を読む」を翻訳する

形態素解析

太郎	太郎	名詞-固有名詞-人名-名
が	が	助詞-格助詞-一般
本	本	名詞-一般
を	を	助詞-格助詞-一般
読む	読む	動詞-自立 五段・マ行 基本形
。	。	記号-句点

7



8

デジタル情報の処理と認識

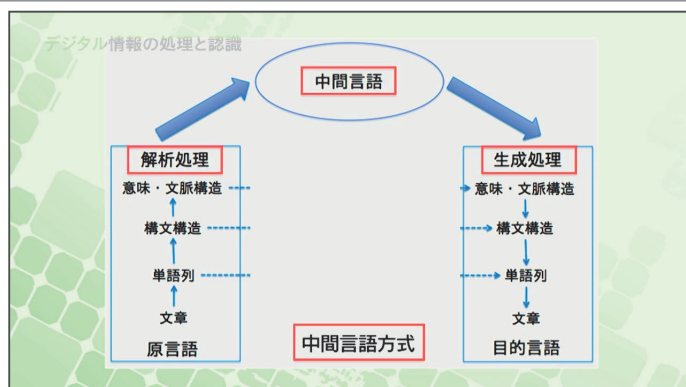
変換処理/生成処理

- 日本語の「名詞句 が 動詞句」
 - 英語の「名詞句 動詞句」
- 「太郎が本を読む」
 - (太郎) (読む) (本)
 - Taro reads a book.

デジタル情報の処理と認識 ('22)

第3回 自然言語処理の応用

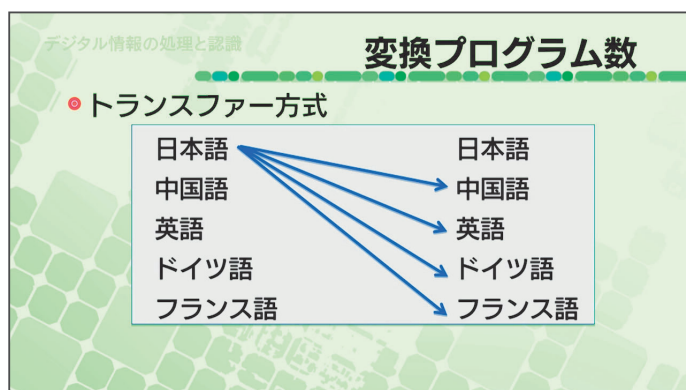
9



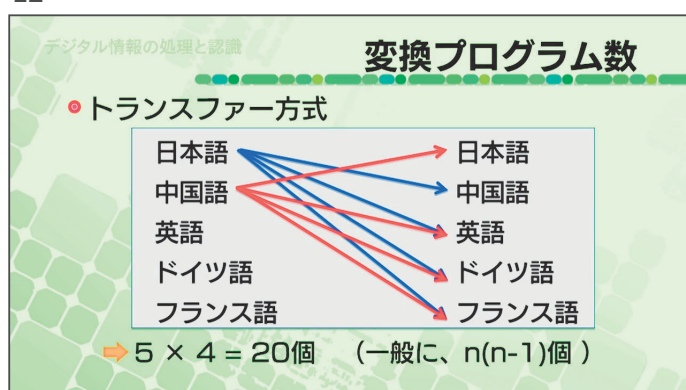
10



11



12



デジタル情報の処理と認識 ('22)

第3回 自然言語処理の応用

13

デジタル情報の処理と認識

変換プログラム数

- 中間言語方式

日本語
中国語
英語
ドイツ語
フランス語

中間言語

日本語
中国語
英語
ドイツ語
フランス語

⇒ $5 \times 2 = 10$ 個 (一般に、 $2n$ 個)

14

デジタル情報の処理と認識

機械翻訳

- ルールベース方式
 - 翻訳のための変換ルールを予め用意しておく
 - それに基づき翻訳を行う方式
 - トランスファー方式 中間言語方式
- 「Time flies like an arrow.」
 - ⇒ 「時間蠅は矢を好む」??

15

デジタル情報の処理と認識

機械翻訳

- 用例ベース方式
 - 「太郎が本を読む」 ↔ Taro reads a book.
 - 「太郎が新聞を読む」
 - ⇒ Taro reads a newspaper.
- 機械が理解しやすいように もとの文を書き換えてから翻訳を行うといったことも必要

16

デジタル情報の処理と認識

全文検索

デジタル情報の処理と認識 ('22)

第3回 自然言語処理の応用

17

デジタル情報の処理と認識

全文検索

- ◎ 総当たりによる方法
 - 検索キーワードと一致する文字列を文書の最初から最後まで順次照合しながら探し出す方法
 - あらかじめ前処理を行う必要が無く任意の文書を対象に検索を行うことができる
 - 文書量に比例して検索時間が増える

18

デジタル情報の処理と認識

全文検索

- ◎ 索引情報（インデックス）を作成
 - 転置インデックス

「多量に蓄積された文書の中から目的の検索キーワードが含まれる文書やその位置を探し出す処理が全文検索です」

19

デジタル情報の処理と認識

多量	多量	名詞-形容動詞語幹
に	に	助詞-副詞化
蓄積	蓄積	名詞-サ変接続
さ	する	動詞-自立 サ変・スル 未然レル接続
れ	れる	動詞-接尾 一段 連用形
た	た	助動詞 特殊・タ 基本形
文書	文書	名詞-一般
の	の	助詞-連体化

➡ 「多量、蓄積、文書、中、目的、検索、キーワード、位置、処理、全文」

20

デジタル情報の処理と認識

転置インデックス

単語	文書ID
物理	2, 17
情報	1, 17
音楽	5, 9, 14
数学	4, 7, 9
体育	13, 15

デジタル情報の処理と認識 ('22)

第3回 自然言語処理の応用

21

デジタル情報の処理と認識

転置インデックス

単語	文書ID
物理	2, 17
情報	1, 17
音楽	5, 9, 14
数学	4, 7, 9
体育	13, 15

22

デジタル情報の処理と認識

転置インデックス

単語	文書ID
物理	2, 17
情報	1, 17
音楽	5, 9, 14
数学	4, 7, 9
体育	13, 15

23

デジタル情報の処理と認識

検索の高速化

- 不要語をインデックスから除外
 - 「to」や「be」を除外
 - 「To be or not to be. (生きるべきか死ぬべきか)」を検索できない

24

デジタル情報の処理と認識

キーワードの重要度

- TF_{ni} (Term Frequency)
 - 文書 n における単語 i の出現頻度

「多量に蓄積された文書の中から
目的の検索キーワードが含まれる文書や
その位置を探し出す処理が全文検索です」

➡ この文書番号が1 単語「検索」の番号が1であれば、 $TF_{11}=2$

デジタル情報の処理と認識 ('22)

第3回 自然言語処理の応用

25

デジタル情報の処理と認識

IDF

- ◉ IDF_i (Inverse Document Frequency)
 - 全文書数(=N)のうち 単語 i を含む文書数(=DF _{i})の割合を求め 逆数をとったもの
 - 通常は 更にその対数をとる
 - 特定の文書にのみに存在する単語は重要
- ◉ 「検索」を含む文書数10 全文書数1000
 - ⇒ $IDF_1 = \log(1000/10) = 2$

26

デジタル情報の処理と認識

TF・IDF

- ◉ TFとIDFの積
 - その文書に出現する頻度が高い [TF大]
 - 他の文書にはあまり出現しない [IDF大]
- ⇒ 両方大きければ重要度が高い
- ◉ 重要なキーワードの抽出や
検索の際の順位付けなどに利用

27

デジタル情報の処理と認識

ベクトル空間法

- ◉ 複数のキーワードを用いた類似度を定義
- ◉ 例えば N個のキーワードについて TF・IDFを
並べたものを文書の特徴として利用

文書1 ⇒ (1.2, 1.7, 3.4, ...)

↑ ↑

キーワード1のTF・IDF キーワード2のTF・IDF

⇒ 文書をN次元のベクトルとして表現

28

デジタル情報の処理と認識

ベクトル空間法

文書1

検索キーワード

θ_1

θ_2

文書2

デジタル情報の処理と認識 ('22)

第3回 自然言語処理の応用

29



30

デジタル情報の処理と認識

データマイニング

- 多量に蓄積されたデータの中から知識を発見する処理
 - あるスーパーの売り上げについて調べ
 - アサリを買った人の多くが白ワインも買っている
 - アサリのワイン蒸しを作る人が多い?
 - アサリの売り場の脇に 白ワインを置くことで店の売り上げが上がる?

31

デジタル情報の処理と認識

テキストマイニング

- テキストデータを対象としたデータマイニング
 - Webページを解析して評判を分析
 - 自由記述アンケートを分析
- 共起分析
 - ある単語と同時に出現する単語を分析

32

デジタル情報の処理と認識

対応分析

- 類似する文書が近くなるように配置することで内容によって文書を分類することができる
- 例として 講義で使用されたスライド中のテキスト情報によってスライドを分類

デジタル情報の処理と認識 ('22)

第3回 自然言語処理の応用

33

デジタル情報の処理と認識

対応分析

- スライドからテキスト情報を抽出
- 形態素解析
- キーワードの出現頻度のベクトルを作成

→ 「権利,情報,技術,倫理,論文」などを抽出
これらのキーワードの出現頻度のベクトルを
それぞれの文書の特徴量として利用

34

デジタル情報の処理と認識

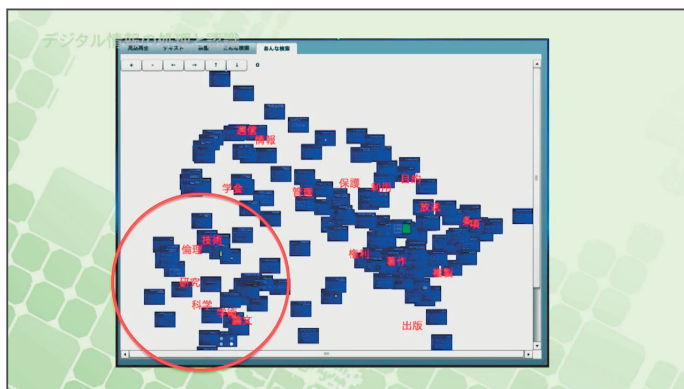
対応分析

		J列			
		kw1	kw2	kw3	...
I行	スライドA	65	55	90	
	スライドB	50	66	50	
	スライドC	68	63	63	
	...				

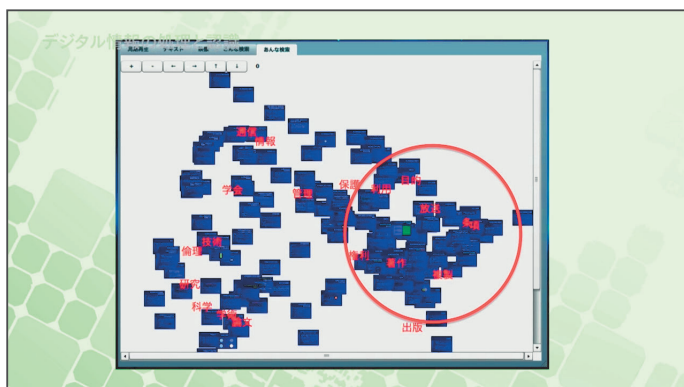
比率が行のプロファイル

比率が列のプロファイル

35



36



第3回 自然言語処理の応用

37

デジタル情報の処理と認識

まとめ

- 機械翻訳
 - テンプレート方式 トランスファー方式
 - 中間言語方式 用例ベース方式
- 全文検索
 - 転置インデックス ベクトル空間法
- テキストマイニング
 - 共起分析 対応分析

5

10

15

20

25

30

35