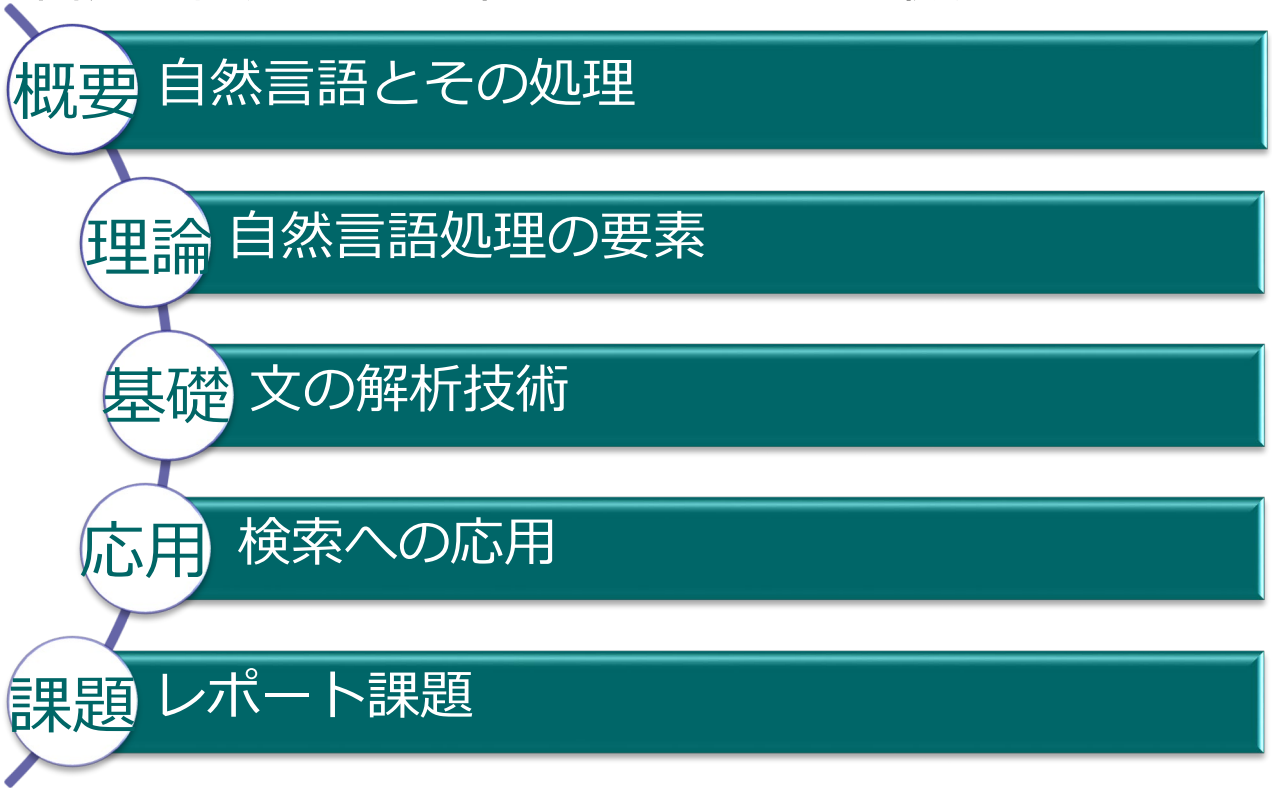
自然言語処理

松原 茂樹 （情報学部コンピュータ科学科）

Astronomers saw stars with ears.

内容

- **自然言語処理 Natural Language Processing**
  - 自然言語処理の基本的なトピックを取り上げます



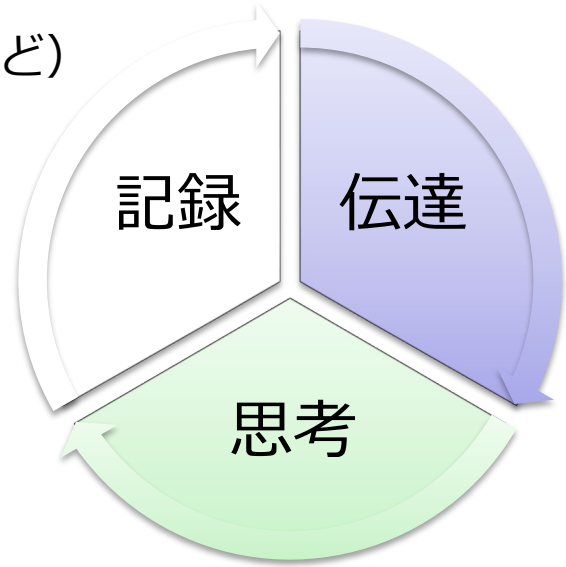
文献
- Foundations of Statistical Natural Language Processing, C. D. Manning (1999).
- Speech and Language Processing, 2nd Edition, D. Jurafsky (2008).
- 自然言語処理の基礎、奥村学、コロナ社 (2010).
- 自然言語処理概論、黒橋禎夫、サイエンス社 (2016).

自然言語とその処理

- 自然言語とは
- 自然言語処理の歴史
- 自然言語処理の普及

自然言語とは

- 自然言語　※人工言語（プログラミング言語など）
  - 日本語、英語、フランス語、…



- 言語の働き
  - モノやコトに名前を付け、
    それらの関係を記すこと

名前	• 名前の付け方は恣意的
用法	• 社会の慣習次第
語彙	• 時代によって変化
内容	• モノ・コトの関係を１次元で記述
意味	• 多義性、曖昧性、同義性

自然言語処理の難しさ

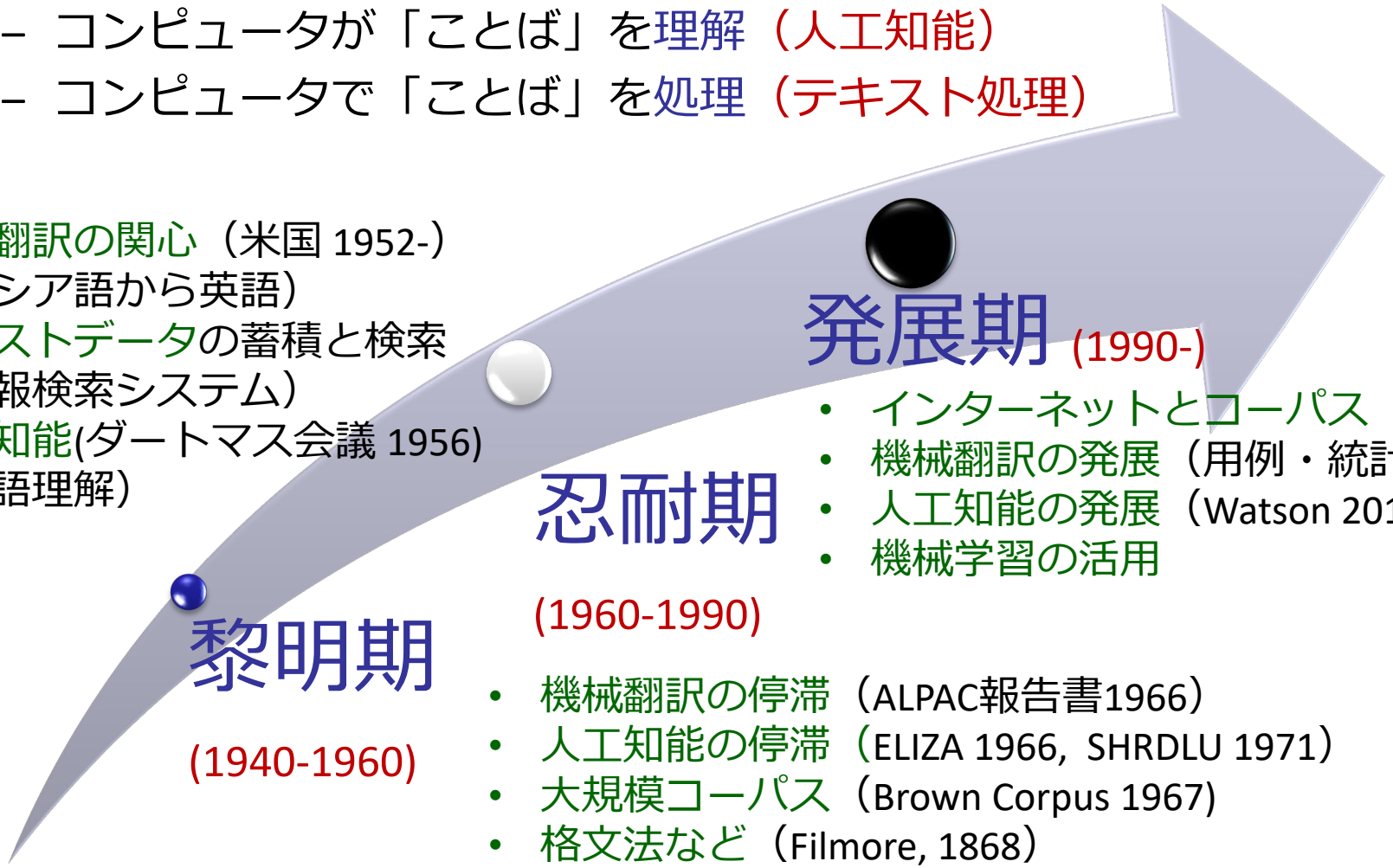
自然言語処理の難しさ

名前	• 名前の付け方は恣意的
用法	• 社会の慣習次第
語彙	• 時代によって変化
内容	• モノ・コトの関係を１次元で記述
意味	• 多義性、曖昧性、同義性

- 言語の知識
  - 新語、専門用語、同義・類義の関係
  - 語の用法、語と語のつながり
- 言語の曖昧性
  - 曖昧性の解消

自然言語処理とその歴史

- 自然言語処理
  - コンピュータが「ことば」を理解（人工知能）
  - コンピュータで「ことば」を処理（テキスト処理）



- 機械翻訳の関心（米国 1952-）（ロシア語から英語）
- テキストデータの蓄積と検索（情報検索システム）
- 人工知能(ダートマス会議 1956)（言語理解）

黎明期

(1940-1960)

忍耐期

(1960-1990)

- 機械翻訳の停滞（ALPAC報告書1966）
- 人工知能の停滞（ELIZA 1966, SHRDLU 1971）
- 大規模コーパス（Brown Corpus 1967)
- 格文法など（Filmore, 1868）
- 医学文献サービス（MEDLINE 1971）

発展期 (1990-)

- インターネットとコーパス
- 機械翻訳の発展（用例・統計）
- 人工知能の発展（Watson 2011）
- 機械学習の活用

自然言語処理の応用

- 実用化された自然言語処理
  - Web検索エンジン
  - 仮名漢字変換
  - 文法チェッカ
  - 会話エージェント
  - 機械翻訳

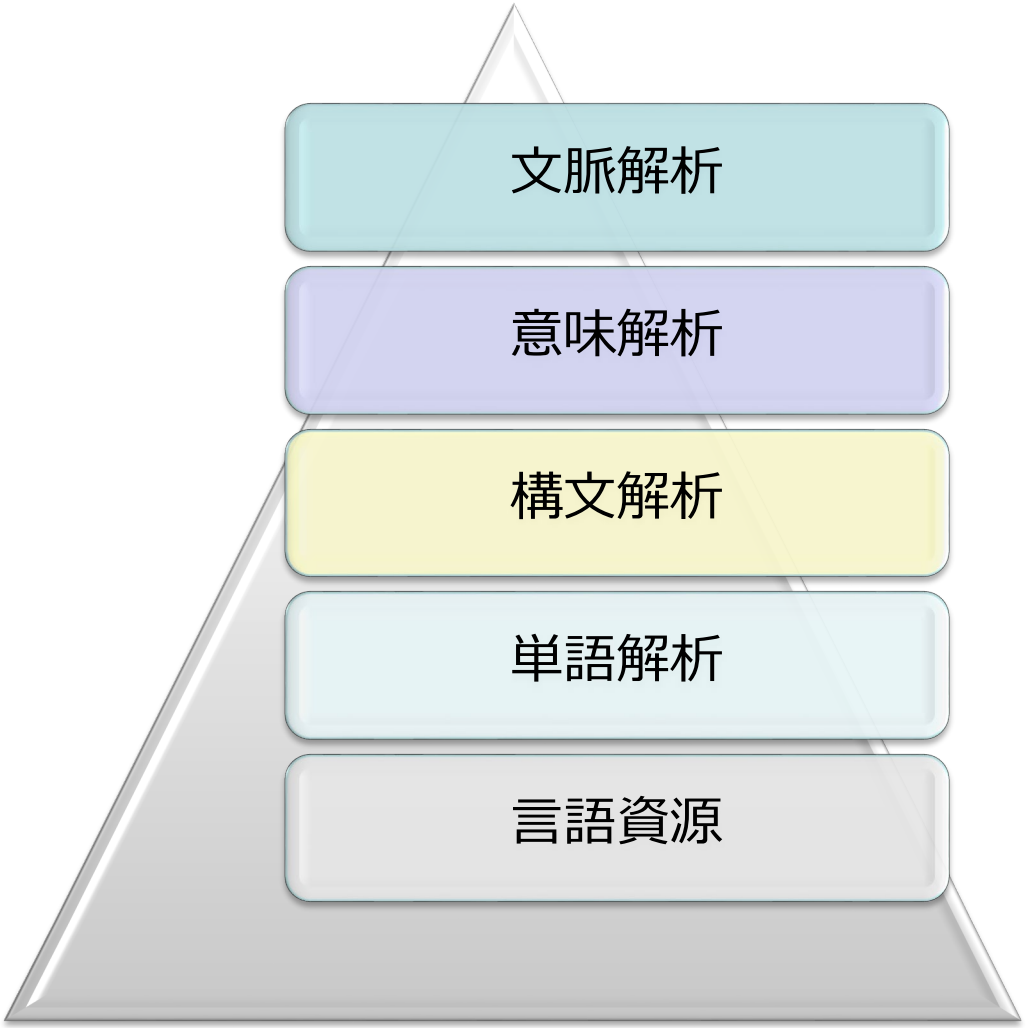
- 研究が進んでいる応用例
  - 情報検索、情報抽出、知識獲得、文書分類など
  - 文書要約、言い換え、機械翻訳、文生成など
  - 質問応答、言語インタフェース、音声対話など

自然言語処理の要素

- 系列解析
- 構文解析
- 意味解析

自然言語処理の要素

- 言語の基本単位
  - 語（文字のならび）
  - 句（語のならび）
  - 文（句のならび）
    - 何がどうした
  - 文章（文のならび）
    - 因果関係など



文の解析（語の系列）

- 系列解析（＝形態素解析）
  - 語に分割＋語形の解析＋固有表現認識

    - **花子はフランス語も話せる**

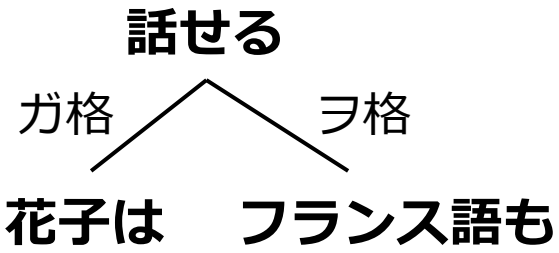
花子　　名詞 人名
は　　助詞 副助詞
フランス 名詞 地名
語　　名詞 普通名詞
も　　助詞 副助詞
話せる　動詞 母音動詞 基本形
EOS

文の解析（構文）

- 構文解析
  - 文内の語句間の（修飾・被修飾）関係の解析
    - **花子はフランス語も話せる**

文の解析（意味）

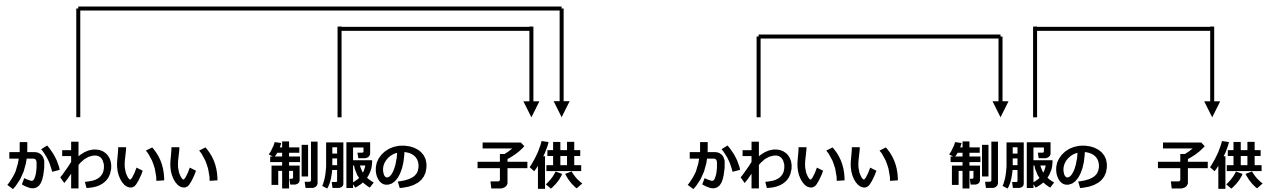
- 意味解析
  - 文内の述語と項の関係の解析
    - **花子はフランス語も話せる**



  - 構文と意味： 非文?
    - 東京で行く。
    - 愛する花子を太郎は。
    - 海は本を切る。

自然言語の曖昧性と解消

- 自然言語文の曖昧性
  - A) 語境界の曖昧性
  - B) 品詞の曖昧性
  - C) 語義の曖昧性
  - D) 文構造の曖昧性
    - **かわいい制服の子供**



文章の解析（文脈）

- 文脈解析
  - 花子はフランス語も話せる
  - フランスに留学していたからだ

– 照応・省略解析

花子はフランス語も話せる

フランスに（省略）留学していたからだ

– 談話構造解析

（理由）
- 花子はフランス語も話せる
- フランスに留学していたからだ

言語資源（辞書）

- 単語辞書
  - 形態情報：読み、品詞、活用型など

見出し	読み	品詞	活用型	活用形	基本形
は	は	助詞			
歯	は	名詞			
橋	はし	名詞			
弾く	はじく	動詞	カ行五段	終止形	弾く

  - 意味情報：語義を格フレームで表現
    - 「はじく」（接触してきたものをはね返す）
      - [レインコート、下敷き、ガラス、羽] **ガ**
      - [水、雨、インク] **ヲ**

        （出典：IPA日本語基本動詞辞書）

言語資源（コーパス）

- コーパス
  - 電子化された言語データ
    - 生コーパス （収集したまま）
    - タグ付きコーパス（情報を付与）

（例）**構文情報付きコーパス**

```
( (S
    (NP-SBJ
      (NP (NNP Pierre) (NNP Vinken) )
      (, ,)
      (ADJP
        (NP (CD 61) (NNS years) )
        (JJ old) )
      (, ,) )
    (VP (MD will)
      (VP (VB join)
        (NP (DT the) (NN board) )
        (PP-CLR (IN as)
          (NP (DT a) (JJ nonexecutive) (NN director) ))
        (NP-TMP (NNP Nov.) (CD 29) )))
    (. .) ))
```

（出典： Penn Treebank コーパス）

```
# S-ID:950101003-001 KNP:96/10/27 MOD:2005/03/08
* 26D
村山 むらやま 村山 名詞 6 人名 5 * 0 * 0
富市 とみいち 富市 名詞 6 人名 5 * 0 * 0
首相 しゅしょう 首相 名詞 6 普通名詞 1 * 0 * 0
は は は 助詞 9 副助詞 2 * 0 * 0
* 2D
年頭 ねんとう 年頭 名詞 6 普通名詞 1 * 0 * 0
に に に 助詞 9 格助詞 1 * 0 * 0
* 6D
あたり あたり あたる 動詞 2 * 0 子音動詞ラ行 10 基本連用形 8
* 6D
首相 しゅしょう 首相 名詞 6 普通名詞 1 * 0 * 0
官邸 かんてい 官邸 名詞 6 普通名詞 1 * 0 * 0
で で で 助詞 9 格助詞 1 * 0 * 0
* 6D
内閣 ないかく 内閣 名詞 6 普通名詞 1 * 0 * 0
記者 きしゃ 記者 名詞 6 普通名詞 1 * 0 * 0
会 かい 会 名詞 6 普通名詞 1 * 0 * 0
と と と 助詞 9 格助詞 1 * 0 * 0
* 6D
二十八 にじゅうはち 二十八 名詞 6 数詞 7 * 0 * 0
日 にち 日 接尾辞 14 名詞性名詞助数辞 3 * 0 * 0
* 26D
会見 かいけん 会見 名詞 6 サ変名詞 2 * 0 * 0
し し する 動詞 2 * 0 サ変動詞 16 基本連用形 8
、 、 、 特殊 1 読点 2 * 0 * 0
```

（出典：京都大学テキストコーパス）

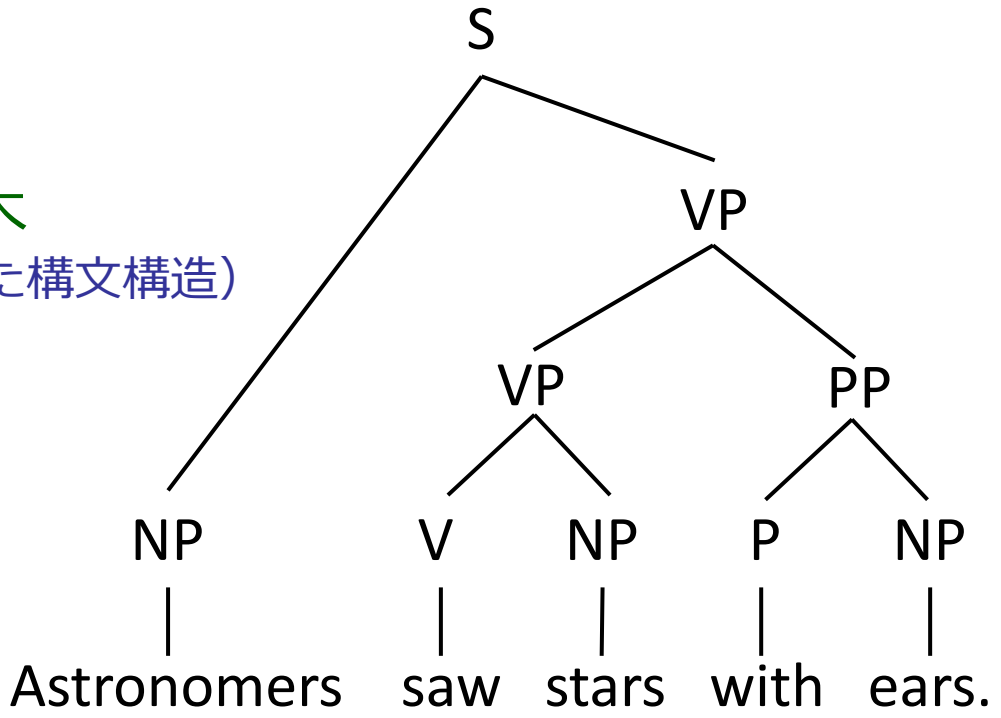
文の解析技術

- 構文解析とは
- 文法
- 構文解析の方法

構文解析とは

- 自然言語の文は単語の並び
  - 構文解析
    - 単語の並びは文法的かを検査する！
    - 文の構造（＝構文構造）を明らかにする！
      - Astronomers saw stars with ears.

構文木
（木構造で表された構文構造）



文法

- 文法：言語の規則の集合
  - 例）

S （＝文）
NP（＝名詞句）
VP（＝動詞句）
PP（＝前置詞句）
V （＝動詞）
P （＝前置詞）

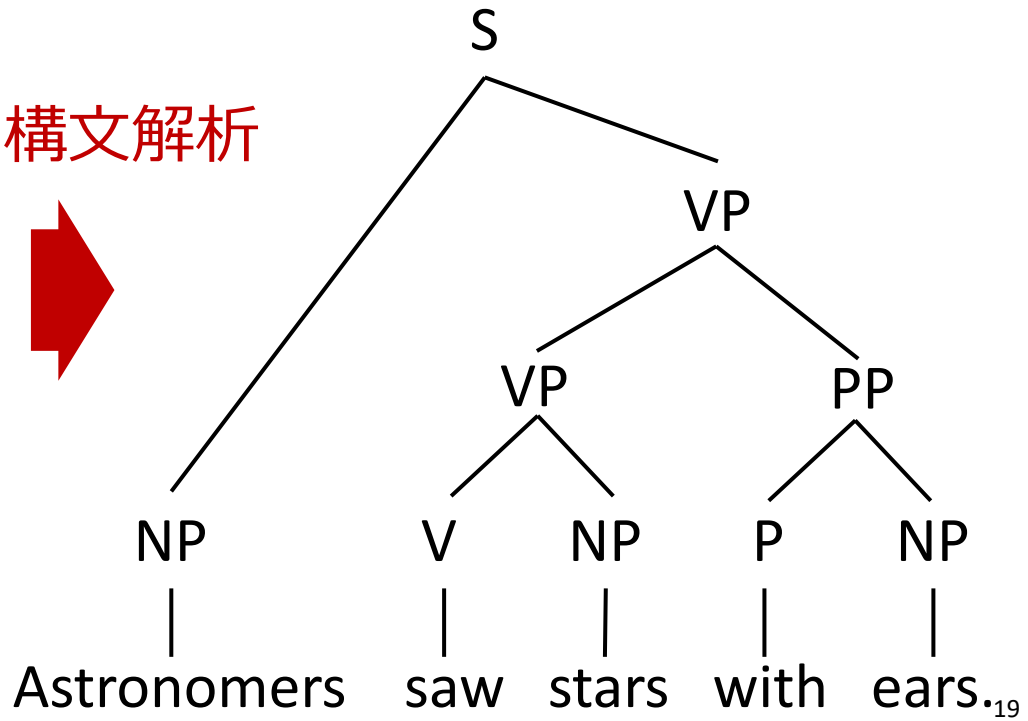
文法規則

(1)	S → NP VP
(2)	VP → V NP
(3)	VP → VP PP
(4)	NP → NP PP
(5)	PP → P NP

辞書規則

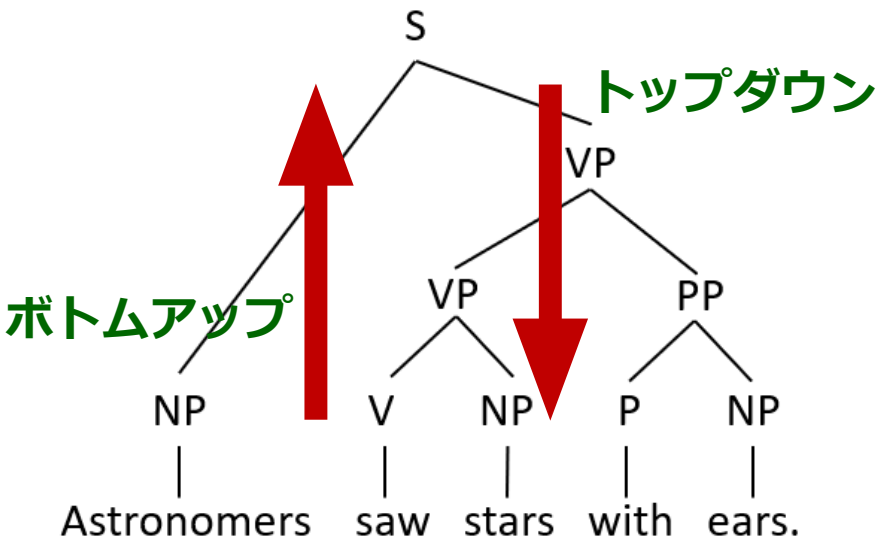
(6)	NP → astronomers
(7)	NP → stars
(8)	NP → ears
(9)	V → saw
(10)	P → with

構文解析



構文解析

- 構文解析の方法
  - トップダウン解析
    - 文(= S) から始めて単語列に向けて解析木を成長
  - ボトムアップ解析
    - 単語列から文(＝S)に向けて解析木を成長



  - 縦型解析
    - 複数の規則が適用できる場合、１つを選択して解析を続け、行き詰まると後戻りする
  - 横型解析
    - 複数の規則が適用できる場合、すべての結果を保持しながら並行して処理を進める。

文の導出

- 縦型・トップダウン構文解析により
  文を導出する。
  - 文の導出の方法（ここでの決まり）
    - **文(＝S) から始め**て単語列が得られるまで
      「文法規則」と「辞書規則」を使って書き換える。
    - **最左導出**（＝書き換える可能な記号のうち、
      最も左の記号を書き換える）を採用する。
    - １回の書き換えには、**１つの「文法規則」または「辞書規則」のみ**を使用する。
    - ある記号に対して**複数の「文法規則」または「辞書規則」**が適用する場合、どれを選んでもよい。
    - 途中で書き換えできなくなったら、直前の
      選択可能な場面に**後戻り**する。

文の導出例

- 縦型・トップダウン構文解析により
  Astronomers saw stars with ears
  を導出する。

S ⇒ NP VP (1)
⇒ astronomers VP (6)
⇒ astronomers VP PP (3)
⇒ astronomers V NP PP (2)
⇒ astronomers saw NP PP (9)
⇒ astronomers saw stars PP (7)
⇒ astronomers saw stars P NP (5)
⇒ astronomers saw stars with NP (10)
⇒ astronomers saw stars with ears (8)

文法規則

(1)	S → NP VP
(2)	VP → V NP
(3)	VP → VP PP
(4)	NP → NP PP
(5)	PP → P NP

辞書規則

(6)	NP → astronomers
(7)	NP → stars
(8)	NP → ears
(9)	V → saw
(10)	P → with

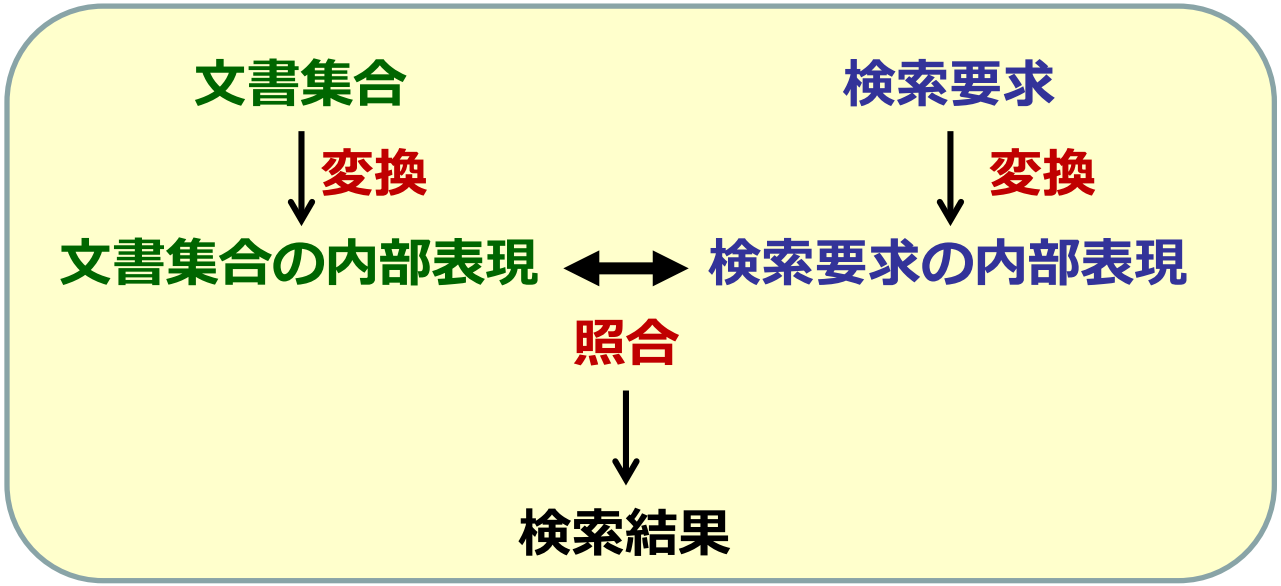
検索への応用

- 情報検索とは
- 情報検索モデル
- 索引付け

情報検索とは

- 情報検索
  - 大量の情報から必要な情報を探し出すこと

– 情報検索の仕組み



- 内部表現：**索引語（＝インデックス）**の集合
- 内部表現への変換：**索引付け（＝インデキシング）**

情報検索のモデル（１）

- ブーリアンモデル

文書集合

↓ **変換**

文書集合の内部表現

（“1” は出現する、 “0” は出現しない）

	索引語１	索引語２	索引語３	索引語４
文書１	1	1	1	0
文書２	0	1	1	1
文書３	1	0	1	1
文書４	0	0	1	1

検索要求

索引語１, 索引語２を含む

↓ **変換**

検索要求の内部表現

索引語１∧索引語２

↔ **照合**

↓

検索結果

{文書１, 文書３} ∩ {文書１, 文書２}
= {**文書１**}

- 索引語の重みの違いは？
- 検索結果のランキングは？

情報検索のモデル（２-１）

- ベクトル空間モデル

文書集合

↓ **変換**

文書集合の内部表現

（数字は文書における索引語の重み）

	索引語１	索引語２	索引語３	索引語４
文書１	0.2	0.5	0.6	0
文書２	0	0.3	0.1	0.8
文書３	0.5	0	0.5	0.2
文書４	0	0	0.3	0.3

検索要求

索引語２, 索引語３を含む

↓ **変換**

検索要求の内部表現

(0, 1, 1, 0)

↔ **照合**

↓

検索結果

情報検索のモデル（２-２）

- ベクトル空間モデル（続き）
  - （例）照合に内積を用いる場合

照合

文書	文書 内部表現	検索要求 内部表現	内積	順位
文書１	(0.2, 0.5, 0.6, 0)	(0, 1, 1, 0)	1.1	1
文書２	(0, 0.3, 0.1, 0.8)	(0, 1, 1, 0)	0.4	3
文書３	(0.5, 0, 0.5, 0.2)	(0, 1, 1, 0)	0.5	2
文書４	(0, 0, 0.3, 0.3)	(0, 1, 1, 0)	0.3	4

検索結果
１位：文書１
２位：文書３
３位：文書２
４位：文書４

  - 照合の方法（＝ベクトル間の類似度計算）
    - コサイン類似度、など様々

索引付け

- 索引付けの処理
  1. 単語の同定
     - 単語解析による単語の分割
  2. 索引語の選択
     - 不要語リスト（不要語＝ストップワード）で除去
       - 付属語（助動詞、助詞など）
       - 自立語のうち、どの文書でも高頻度で出現する語
  3. 索引語の重み付け

（例）

	索引語１	索引語２	索引語３	索引語４
文書１	0.2	0.5	0.6	0
文書２	0	0.3	0.1	0.8
文書３	0.5	0	0.5	0.2
文書４	0	0	0.3	0.3

索引語の重み付け（出現頻度：TF）

- 文書における索引語の重みを計算する
  - **仮定１）**文書に出現する頻度が多い索引語ほど、文書をより特徴付ける
  - 文書ごとの索引語の出現頻度（転置していることに注意！）

	文書１	文書２	文書３	文書４	文書５
索引語１	1	0	5	2	3
索引語２	0	3	3	2	0
索引語３	3	2	0	4	0
索引語４	6	6	8	7	5
索引語５	4	1	4	0	0
索引語６	0	5	0	3	2

  - 文書（document: d）における索引語（term: t）の出現頻度（term frequency: tf）を重み（weight: w) とする

$$w_{t,d} = tf(t,d)$$

索引語の重み付け（文書頻度：IDF）

- 文書における索引語の重みを計算する
  - 出現頻度を重みとすることの問題

	文書1	文書2	文書3	文書4	文書5
索引語1	1	0	5	2	3
索引語2	0	3	3	2	0
索引語3	3	2	0	4	0
索引語4	6	6	8	7	5
索引語5	4	1	4	0	0
索引語6	0	5	0	3	2

特徴付けている？ →

  - **仮定２）**出現する文書数が少ない索引語ほど、出現する文書をより特徴付ける

  - 索引語（term: t）の、総文書（number: N）あたりの文書頻度（document frequency: idf）の逆数を重み（weight: w) とする

$$w_{t,d} = idf(t) = \log \frac{N}{df(t)} + 1$$

	idf
索引語1	1.10
索引語2	1.22
索引語3	1.22
索引語4	1.00
索引語5	1.22
索引語6	1.22

索引語の重み付け（出現頻度×文書頻度：TF・IDF）

- 文書における索引語の重みを計算する
  - 出現頻度（TF）と文書頻度（IDF）を考慮

$$w_{t,d} = tf(t,d) \times idf(t)$$

出現頻度（TF）

	文書1	文書2	文書3	文書4	文書5
索引語1	1	0	5	2	3
索引語2	0	3	3	2	0
索引語3	3	2	0	4	0
索引語4	6	6	8	7	5
索引語5	4	1	4	0	0
索引語6	0	5	0	3	2

×

文書頻度（IDF）

	idf
索引語1	1.10
索引語2	1.22
索引語3	1.22
索引語4	1.00
索引語5	1.22
索引語6	1.22

=

	文書1	文書2	文書3	文書4	文書5
索引語1	1.10	0	5.50	2.20	3.30
索引語2	0	3.60	3.66	2.44	0
索引語3	3.66	2.44	0	4.88	0
索引語4	6	6	8	7	5
索引語5	4.88	1.22	4.88	0	0
索引語6	0	6.10	0	3.66	2.44

【応用】文解析による検索

- 学生が英語で論文を書く
  - ◯日本語論文、英語の知識、分野の専門用語
  - ×（大量の）英文用例

- 英文用例を検索できる環境
  - 「3節で要約手法について述べる」：述べる？

検索クエリ section -v method　※ -v : 動詞

◯ Section 3 describes our method.

◯ Section 4 describes the translation method based on TSC.

◯ This section briefly describes the evaluation methods we employed for automatic synonym acquisition.

× Finally, section 6 describes how the method was generalized to cover other semantic features .

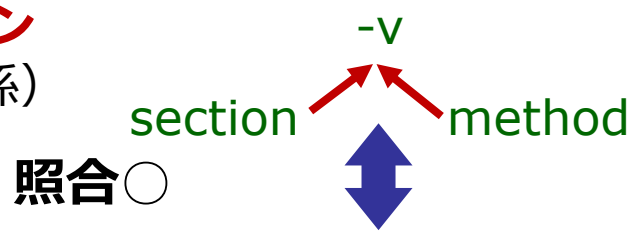
  - ユーザの意図に合致した用例文を提示

【応用】文解析による検索

- 検索クエリと用例文との照合

検索クエリ section -v method

構文パターン
(単語間の関係)



※ →: 依存関係
(単語間の修飾関係)

照合○

Section 4 describes the translation method based on TSC.

- 英文検索システム **ESCORT**

http://escort.slplab.org/

  - キーワードを入力するだけ
  - 品詞入力も可
  - キーワードの補完
  - 構文パターンで結果を分類提示

国際会議
論文
など