

自然言語処理技術の 周辺

水谷 正大

Masahiro Mizutani

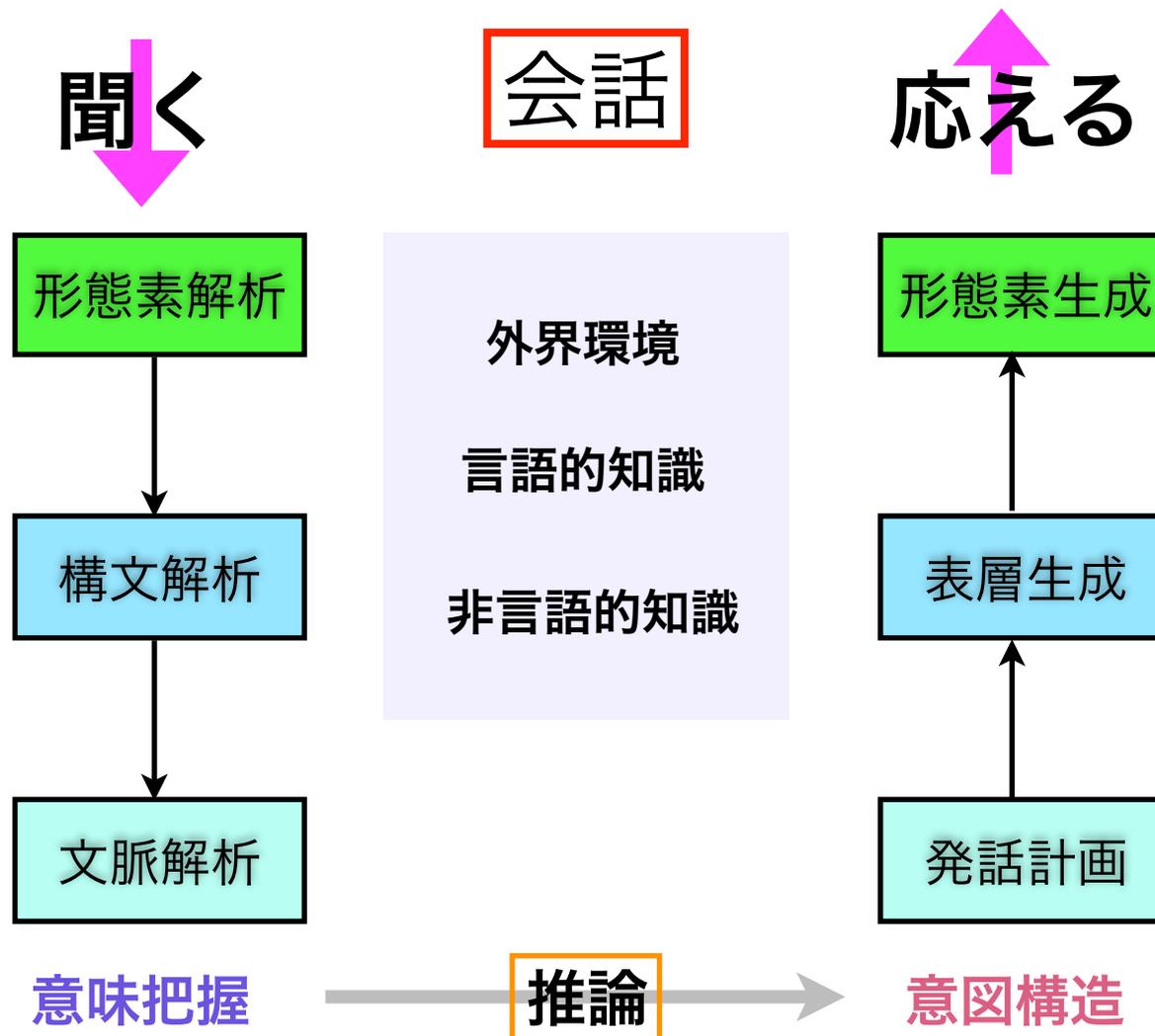
言語の情報科学的周辺

- 自然言語処理
 - 自然言語の科学的究明
 - 言語の「文法」、「意味」
 - 機械翻訳、文書情報処理
 - 形態素解析、語彙統計、内容抽出
- 数理言語理論
 - 記号論理学、数学
 - 形式言語理論とオートマトン
 - コンピュータ言語
 - 計算理論

自然言語処理技術

「聞いて」それに「応える」対話過程を処理

文字
音声認識
画像認識
パターン認識



形態素解析

- 記号列は単語や活用語尾などの**形態素**(morpheme)に分割する処理
 - 入力信号を記号列（文字、音素）に変換した後の処理
- **曖昧性**の取り扱い
 - 解消のために言語的・非言語的知識を利用
 - 「ちみたち」は「きみたち」とチとキを判断
 - 経済を話題にしていると「医者かい」⇒「社会」と聞く
 - 文節数最小・**最長一致の原則**（長いまとまりを優先）
 - にはとりがいる⇒鶏がいる
 - それでも、**非言語的知識**や**構文解析**は必要
 - ここではきものを⇒ここでは着物/ここで履物

形態素解析の応用

- 文学作品の言葉遣いは古来から研究対象
 - 作品の語彙分析
 - 作者の特徴（出身地域、年齢）
- 近年の文字情報の電子化＋情報処理技術
 - テキストマイニング
 - 自由記述文書から知りたい情報を取り出すための技術
 - 大量の文書から「傾向」「大筋」「隠された構造」を探る
 - アンケート調査等による顧客の潜在的嗜好の発見と評価
 - 国語・文学研究

タグクラウド (tag cloud)

Webサイト上で項目に**メタデータ**として付与されていたタグを集積して視覚的に表示する

単語頻度に応じて文字サイズを調整表示

<http://www.flickr.com/photos/tags/>

animals architecture **art** asia australia autumn baby band barcelona **beach** berlin bike bird
birds birthday black blackandwhite blue bw **california** canada **canon** car cat
chicago china christmas church **city** clouds color **concert** dance day de dog
england europe fall **family** fashion **festival** film florida flower flowers food
football **france** friends fun garden geotagged **germany** girl graffiti **green** halloween
hawaii holiday house india instagramapp iphone **iphoneography** island italia **italy**
japan kids la lake landscape light live **london** love macro me mexico model museum
music nature new newyork newyorkcity night **nikon** nyc ocean old **paris**
park party people photo photography photos **portrait** raw red river rock san
sanfrancisco scotland sea seattle show sky snow spain spring **square**
squareformat street **summer** sun sunset taiwan texas thailand tokyo toronto tour
travel tree trees trip uk urban **usa** vacation vintage washington **water**
wedding white winter woman yellow zoo

日本語の形態素解析

日本語や中国語では、英語などと違い
単語が空白で区切られていない

➡ 日本語文法の知識がない限り文を単語に分解できない

Q. **勉強しないことはゆかいだ** を形態素に分解せよ

名詞	動詞	助動詞	名詞	助詞	動詞	動詞	助動詞
"勉強"	"し"	"ない"	"こと"	"は"	"ゆか"	"い"	"だ"

英語はかなり簡単

Not studying is pleasant ⇒ "not" "study" "ing" "is" "pleasant"

studyingの分解には英文法の知識を使った

日本語形態素解析エンジン

- ChaSen(茶筌)
 - 奈良先端科学技術大学院大学自然言語処理学講座(松本研究室)
 - <http://chasen-legacy.sourceforge.jp/>
- MeCab(和布蕪)
 - 京都大学情報学研究科 + NTT基礎研究所
 - <http://mecab.sourceforge.net>

「走れメロス」の名詞頻度

MeCabを使い、一般名詞かつ頻度が7以上の語を列挙

名詞	頻度
わし	8
セリヌンティウス	14
メロス	73
人	20
友	18
声	9
妹	12
市	11
心	7
村	9
王	17
男	12
群衆	7
陽	8