

計算言語学

東京大学生産技術研究所
吉永 直樹

site: <http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/class/cl/>

最初に質問

- Q1: 計算言語学という学問領域を(この講義以前に)聞いたことがある人はいますか？
- Q2: 昨年度開講された鶴岡先生の「自然言語処理」の講義を履修した人はいますか？
- Q3: 機械学習、特に深層学習を使って研究している人はいますか？

参考書

- D. Jurafsky and J. H. Martin, **Speech and Language Processing**, Pearson Education
 - <https://www.cs.colorado.edu/~martin/slp.html>
 - Draft of 3rd edition available at
<https://web.stanford.edu/~jurafsky/slp3/>
- 適宜、計算言語学的な内容や最新の研究の話題を補完します

自然言語処理と計算言語学

- 自然言語処理 (natural language processing)
 - 人間の行う様々な言語を用いた行為の計算機による代替・支援を目標とした**工学的研究**
 - 情報検索、自動要約、機械翻訳、評判分析のような言語処理応用の実現を目的とし、言語をモデル化
- 計算言語学 (computational linguistics)
 - 言語という自然現象の理解・説明を計算論的な観点で行うことを見目標とした**理学的研究**
 - 言語獲得、言語の複雑性、人間の言語処理過程などの理解を目的とし、言語をモデル化

本講義のゴール

- 言語に関する問題を設定し、計算機を用いて解き、客観的に性能を検証するための方法論の修得
- そのために以下を学ぶ
 - 様々な言語現象をモデル化する方法
 - 単語、文、文章の意味・構文構造
 - モデル化に用いる技術
 - 機械学習、形式文法理論、組合せ最適化
 - 言語に関する具体的な応用課題
 - 翻訳、要約、対話、評判分析
- ただし、本講義は今年度新たに開講された講義のため、内容については適宜微調整を行います

成績評価

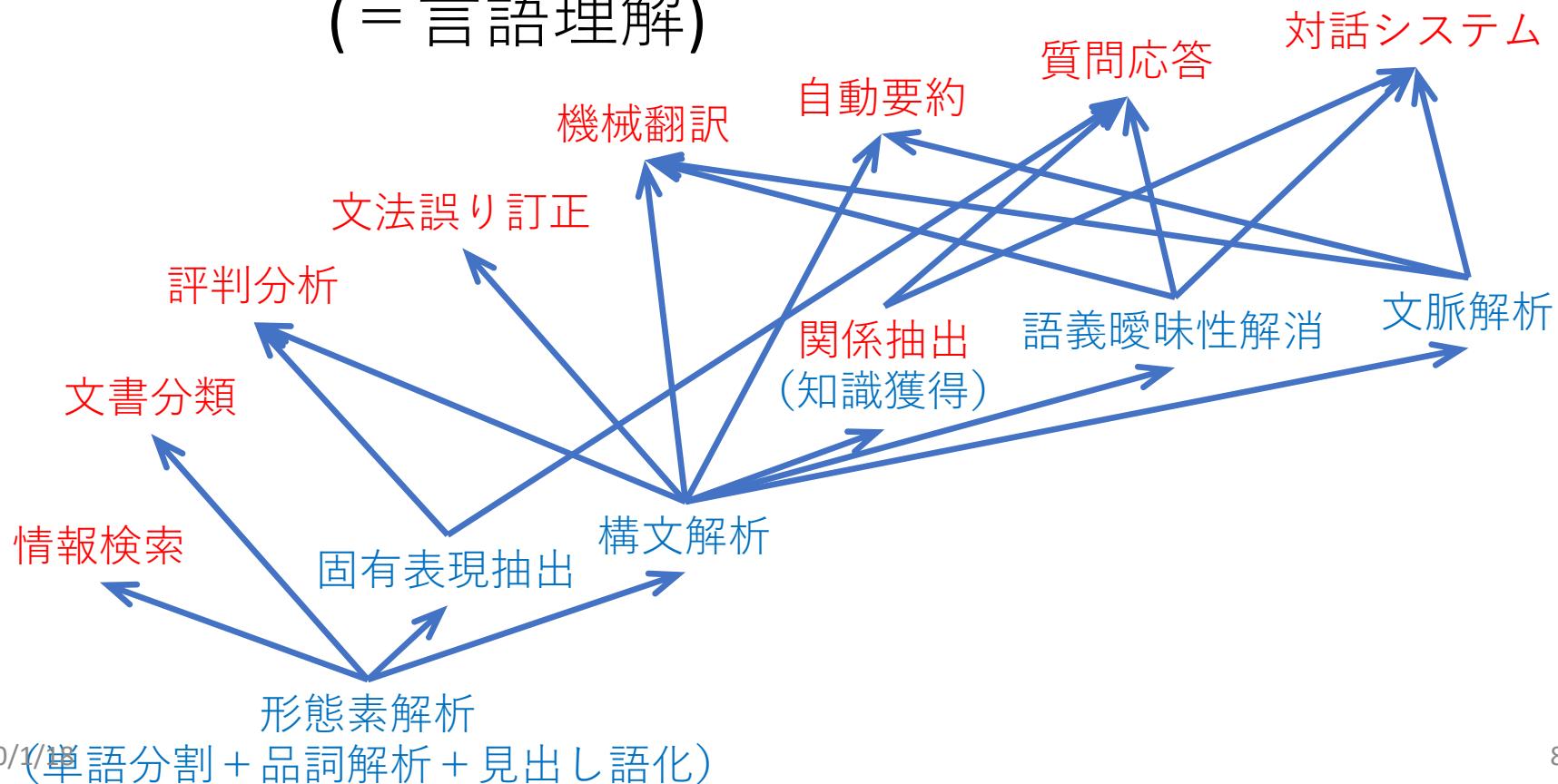
- レポートを予定。内容・回数は未定ですが、
 - 計算言語学/自然言語処理の問題を新規に設定して解く
 - 既存の計算言語学/自然言語処理の問題を新規な手法で解く
 - 論文数本を読んで紹介する
- などを検討しています

スケジュール(予定)

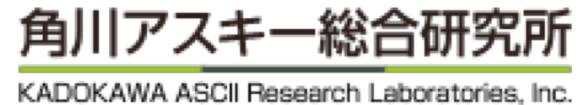
- 10/1, 8(祝日), 15, 22, 29
- 11/5, 12, 19, 26
- 12/3, 10, 17, 24(祝日)
- 1/7, 14(祝日), 21

自然言語処理/計算言語学の応用・要素技術

- 応用技術: 言語を入出力とするアプリケーション
- 要素技術: 非構造化データである言語を構造化
(= 言語理解)



言語情報を利活用するサービス例



SNS 口コミ解析

Twitter トレンド解析



感染症流行分析



対災害SNS情報分析
大規模Web情報分析



天気リポート



レシピ



商品レビュー解析

言語を入出力とする情報デバイス例



<https://www.flickr.com/photos/cinz/41175173715>



<https://www.flickr.com/photos/iphonedigital/26954179704>



<https://www.instagram.com/p/BbkKuHynxTT/>

高度な自然言語処理技術が日常的に利用されるように

研究状況・社会状況の変化

最近20年の動向を中心に

自然言語処理/計算言語学の歴史: 黎明期

- 1947 Weaver が暗号解読とのアナロジーに基づく機械翻訳の実現可能性を示唆
- 1956 Chomsky が文脈自由文法を定義
- 1958 Luhn が技術文書の自動要約システムを開発
- 1962 AMTCL (Association for Machine translation and Computational Linguistics, 後の ACL) 発足
- 1964 最初の電子化されたコーパスであるBrownコーパス
- 1966 ALPACレポートにより翻訳研究が以後10年停滞
- 1966 Weizenbaum が対話システム Eliza を開発
- 1967 Woods が質問応答システム LUNAR を開発
- 1968 Fillmore が格文法を提唱

自然言語処理/計算言語学の歴史: 合理主義 vs. 経験主義の時代

- 1978 東芝が初の日本語ワードプロセッサを販売
- 1984 人手で一般常識のDB化を行う Cyc プロジェクト開始
- 1986 Steedman が組合せ範疇文法 (CCG) を提唱
- 1988 Fred Jelinek ``*Anytime a linguist leaves the group the recognition rate goes up*''
- 1992 言語資源の配布を行う Linguistic Data Consortium 発足
- 1992 Penn Treebank コーパスがリリースされ、以後構文解析において統計的手法とコーパスに基づく評価が主流に
- 1993 Brown が統計的機械翻訳モデル(IBM) モデルを発表
- 1993 EMNLP の前身の Workshop on Very Large Corpora 開催
(2011 IBM Watson がクイズ番組で人間と対戦し勝利)

コーパスに基づく統計的手法と評価の普及 (1990s～)

- 各言語処理タスクについて、注釈付きコーパス
(タスクの正解データ)と自動評価指標の整備が進む
例) Penn TreeBank (Wall Street Journal + 品詞・構文情報)
 - クラウドソーシングの発達により、さらに加速
- Pros:
 - 機械学習に基づく統計的手法の活用
 - 精度等による客観的評価が可能となり研究が加速
(既存手法を再実装することなく性能比較が可能に)
- Cons:
 - データバイアス (陳腐化した標準コーパス)
 - 再実装が必要となる評価尺度の軽視 (解析速度など)

ウェブの発達による言語情報の大規模・多様化 (1993~)

- ソーシャルメディアとスマートフォンの普及により誰もがいつでもどこでも言語で情報発信・共有



大規模テキストを利活用する自然言語処理への期待

深層学習の衝撃 (2013～)

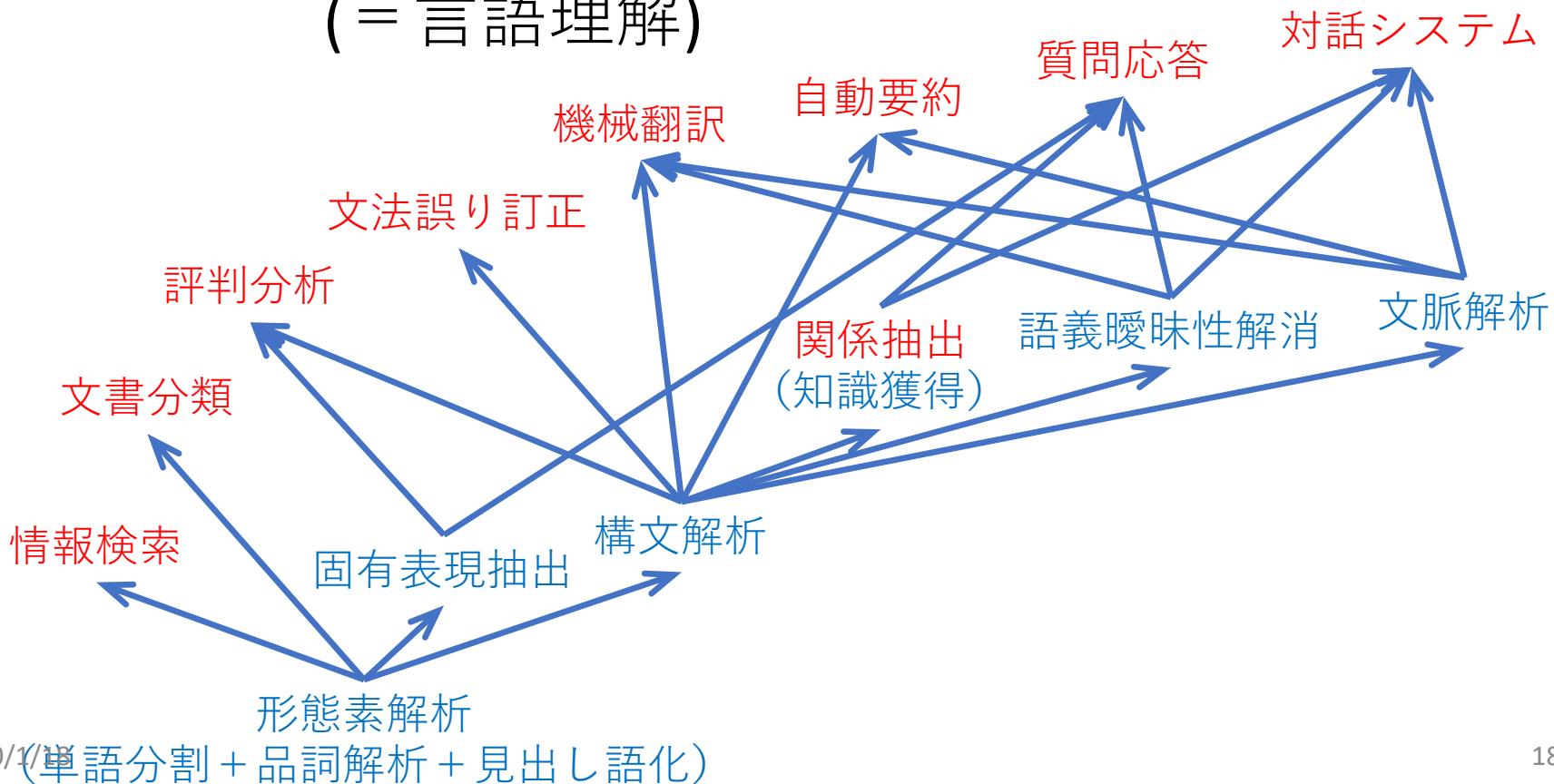
- 深層ニューラルネットワークの適用
 - Recurrent Neural Network による可変長入出力処理
- Pros
 - 意味の連続表現(单語埋め込み)・特徴量の自動学習
 - 非言語データ (画像等) との横断・融合処理が現実的に
 - End-to-end learning により、言語生成を含む複雑な応用タスクで要素技術を全く利用しないモデルが実現可能に例) 機械翻訳・要約・対話・キャプション生成
- Cons:
 - Resource hungry (学習データ、計算機資源)
 - モデルの振る舞いに対する解釈性

應用技術

典型的な問題設定

自然言語処理/計算言語学の応用・要素技術 (再掲)

- 応用技術: 言語を入出力とするアプリケーション
- 要素技術: 非構造化データである言語を構造化
(= 言語理解)



機械翻訳

- ある言語で書かれたテキストを同じ意味を持つ他の言語のテキストに変換

The screenshot shows the Google Translate website. At the top, there's a navigation bar with the Google logo, a grid icon, and a 'ログイン' (Login) button. Below the navigation, the word '翻訳' (Translate) is displayed in red. There are two language selection dropdowns: one for the source language ('英語') and one for the target language ('日本語'). Between them is a double-headed arrow icon. To the right of the target language dropdown is a blue '翻訳' (Translate) button. The input text on the left reads: 'Toto, I've a feeling we're not in Kansas any more. We must be over the rainbow!' The output text on the right is the Japanese translation: 'トト、私はもうカンザスにいないと感じています。私たちは虹の上にいなければなりません！'. Below the input text, there are icons for microphone, keyboard, and a dropdown menu, along with a character count '79/5000'. Below the output text, there are icons for star rating, square, microphone, and a pencil.

<https://translate.google.com/>

自動要約

- 与えられたテキスト(集合)からユーザの要求に応じて重要な内容を抽出・簡略化して出力

入力テキスト :

東京都のJR東日本管内の線路などで相次いでいる不審火で、警視庁は9月15日に、8月23日のJR東日本品川変電所の不審火に関与した疑いで、東京都内に住む42歳の男を威力業務妨害の疑いで逮捕したと発表した。威力業務妨害の疑いで逮捕されたのは、東京都武蔵野市吉祥寺に住む自称ミュージシャンの野田伊佐也容疑者・42歳で、「やったことはやったが、業務を妨害したとは思っていない」と、警視庁に対して話している。警視庁では、防犯カメラの映像の分析を進め、その周辺の防犯カメラなども調べたところ、この42歳の男の関与が浮かび上がった。捜査関係者の話では、この42歳の男は8月23日20時頃に、品川区のJR東日本品川変電所の敷地内において、ペットボトルを燃やして、JR東日本の業務を妨害した疑いが持たれている。かつて野田容疑者が住んでいたアパートの近くに住む50代男性はサンケイスポーツの取材に「夜遅くや朝早く、部屋からロックが聞こえてきた」と話しているが「人付き合いもなさうで目立つ感じではなかった」とも話している。東京都内のJR東日本管内においては、この8月以降、線路脇のケーブルや変電所において、あわせて7件の不審火が発生した。また、品川区にあるJR東日本品川変電所を含む、4か所の現場から、焼けていたペットボトルと針金が発見されている為、警視庁では一連の事件との関連についても捜査を進めている。

要約文字数 : 200



東京都のJR東日本管内の線路などで相次いでいる不審火で、警視庁は9月15日に、8月23日のJR東日本品川変電所の不審火に関与した疑いで、東京都内に住む42歳の男を威力業務妨害の疑いで逮捕したと発表した。捜査関係者の話では、この42歳の男は8月23日20時頃に、品川区のJR東日本品川変電所の敷地内において、ペットボトルを燃やして、JR東日本の業務を妨害した疑いが持たれている。

要約文字数 : 188

https://hitoshin.github.io/tiny_summarizer/

評判分析

- 与えられた評価文書に対し、書き手の評価極性を推定

肯定的評判と否定的評判の書き込み数を集計して表示

肯定的な書き込みと否定的な書き込みを分類して提示

評判指数	49.4
ポジティブな表現	300件 (60%)
ネガティブな表現	79件 (16%)
そのほかの表現	124件 (25%)
合計	503件

VAIOを安定してきたし、大事な会も無事終わった。久しぶりにまたりした
気分で迎える週末。
yu-chika.air-nifty.com/yuyuzanmai/2009/03/post-00b3.html - 詳細情報
由々三昧

VAIO typeC 欲しい!!!!!!! でも通信費超高けえ!!!!!!! 携
88skynet.com/brands/2009/03/pspthe_idolmster_sp_82.html - 詳細情報
ブランド商品研究所

VAIOです 中右は今はやりのASUSのミニPCです ちょっと使いにくいです 下
は奥
blogs.yahoo.co.jp/boya_49574/58707046.html - 詳細情報
★ボーヤのきまぐれフライパン★

質問応答

- 与えられた質問文に対し、知識ベース(KB)や(Web)文書を知識源として回答を提示

The screenshot shows the WolframAlpha mobile website. At the top is the WolframAlpha logo with a red star icon. Below it is a search bar containing the query "How tall is Mt. Everest?" with a red rectangular border around it. To the right of the search bar is an orange equals sign button. Below the search bar is a section titled "Input interpretation" with the text "Mount Everest elevation". Underneath that is a "Result" section with the text "8848 meters". There is also a "Show non-metric" button.

<https://m.wolframalpha.com/>
10/1/18

The screenshot shows the Wisdom-NICT search results page. The search query is "日本はなぜスマホの普及率が低いのか". The results are listed in three numbered sections:

- 中国.普及率.低いの...インフラ整備..問題.....ひとつ..背景..して考え..の....固定回線..存在..日本で..早い時期....NTT.....国策..して光ファイバー..整備..実施して....このため.. 固定回線..インフラ..整って...モバイル..使わ..なら..利用者...数..少な
い可能性..ある。この予測には米国の数字は入っていないが、米国は人口の増加が続いており、日本のような高齢化は起きない可能性が高い。このため、オーストラリア...同様、スマート
普及率..高くなる..考え....問題....基本的にITビジネス..ルール..米国中心..決まって....いう現実....
http://blogos.com/article/90747/?ignore_lite
- ..いうことで..なぜガラホ..作る必要..あった.....いうこと..説明して....思う.....スマートホンが登場して....形態..スマートホンに変化し
て..ガラケー..見かける事....なくなつて....しかし..現状..スマートホンの普及率..落ちて....ガラケー..普及率....高くなつて....そ
の為..スマートホンの生産..少なくなつて....そしてガラケー..生産....増えて....いう状況....その原因として....お年寄り..スマートホンを
使うの..難しくて使い..ため....いうこと....日本はどんどん高齢化してきているため、その影響もあり
ガラケーの普及率が国の中でダントツに高いのである。..お年寄り..ニーズ....考えて..し....編
み出さ..の....ガラホ....ガラホ..メリット..デメリット
<http://kasimasi4309.net/post-853/>
-値....スマートフォン..普及率..見て..良い....スペイン..6.6%、イタリア..5.8%、イギリス..5.5%..目立つところ。....
一対..挙げ..一般携帯電話..各國..数%..留まって..の..対し..日本は2.3%..いう高い値..示して....これは先行する別
途記事でも解説している通り、日本は事実上マルチメディアフォンとして使われる、
高性能な一般携帯電話（非スマートホン）が普及したため、スマートフォンの普及が遅れて
いる。..影響....値....表れて....2.3%....マルチメディアフォン..見て良い。タブレット機..2.0%..外..利用率
<http://www.garbagenews.net/archives/2134305.html>

<https://wisdom-nict.jp/>

対話システム（応答生成）

- 与えられた発話に対し、適切な応答を生成



文法誤り訂正

- 入力に含まれる綴りや文法誤りを認識し、訂正

Grammar Check by Grammarly

Get a free grammar check and correct mistakes in your text
with Grammarly's online grammar checker.



Grammatical Errors



Spelling Errors



1



Incorrect Punctuation



Misused Words

Are you ready to move beyond standard grammar checking tools that misses even basic grammar and spelling errors?

Check your text

<https://www.grammarly.com/grammar-check>

デモ: 話題追跡

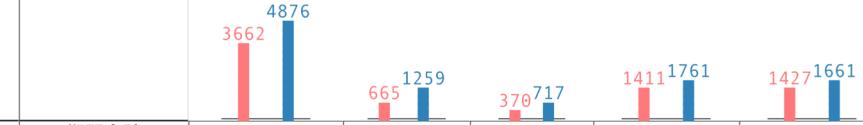
オンラインブログ解析

リアルタイムツイート解析

Blog Feeds (monthly)

2009/02	2009/03	2009/04	2009/05	2009/06	2009/07
4797: 花粉	6832: イチロー	6304: GW	8538: 新型インフルエンザ	3605: 紫陽花	3891: 短冊
3131: 花粉症	9124: WBC	3238: ゴールデンウィーク	8143: 感染	1986: ホタル	2549: 七夕
940: 黄砂	2323: お花見	7447: お花見	14334: マスク	2235: マイケル	12773: 夏休み
1396: 豆まさき	3271: ギューパ	3776: 入学式	12081: GW	1681: アジサイ	2294: 花火大会
3334: バレンタインデー	16911: 桜	6242: 花見	1169: エイブルフルール	1237: あじさい	1561: 梅雨明け
14600: チョコ	1363: 検察	1174: 新緑	1393: 交流戦	1597: 蛍	2718: マニフェス
1673: お雛様	2024: 花見	2208: 連休中	749: 梅雨入り	5669: 花火	
5364: バレンタイン	1708: ダルビッシュ	998: 鯉のぼり	1506: 田植え	1844: ドラクエ	
5948: 豆	5774: 卒業式	38623: 桜	1258: カーネーション	1140: 扇風機	1321: 夏祭り
2758: 節分	2074: 東京マラソン	5928: 花びら	1477: 運動会	1471: プロレス	3767: 浴衣

新型インフルエンザ (nRANK: 3856)



期間合計

	2009/05	2009/06	2009/07	2009/08	2009/09
流行する (1423)	流行る (325)	発生する (89)	流行する (251)	流行る (216)	
流行る (1383)	流行る (64)	流行る (35)	流行る (184)	流行る (190)	
発生する (638)	確認される (63)	流行する (29)	ふるう (69)	ふるう (61)	
ふるう (395)	広がる (279)	流行る (41)	出る (29)	振るう (64)	
振るう (351)	広がる (271)	上陸する (40)	出る (28)	振るう (59)	
蔓延する (332)	確認される (285)	発症する (20)	出る (47)	振るう (44)	
蔓延する (332)	蔓延する (178)	広がる (29)	入る (46)	蔓延する (43)	
出る (304)	上陸する (175)	やってくる (8)	発生する (43)	発生する (31)	
亡くなる (146)	種ぐ (44)	死亡する (11)	死ぬ (6)	亡くなる (36)	
出る (144)	大騒ぎする (31)	学校問題する (8)	受け取る (5)	亡くなる (44)	
死亡する (132)	出る (31)	控えられる (7)	死ぬ (6)	死亡する (23)	
死ぬ (25)	死ぬ (17)	休む (7)	死ぬ (3)	死ぬ (17)	
種ぐ (56)	死ぬ (6)	休む (6)	流れれる (3)	亡くなる (11)	
大騒ぎする (46)	受け取る (4)	発生する (4)	減る (3)	種ぐ (8)	
入院する (42)	亡くなる (13)	休む (4)	確認される (7)	亡くなる (5)	

か

で

10/1/18
10/1/18

Real-Time Parsing / tweet+retweet grep: follow: 0 2011/03/11 14:30 → 2011/03/12 14:30 , per 10 posts, sampling 1000 posts, 20 words per 10m (freq > 0) 全て 名前 scroll 2 burst 3 guess が で を に から まで より 無 無 sent debug polar en parse!

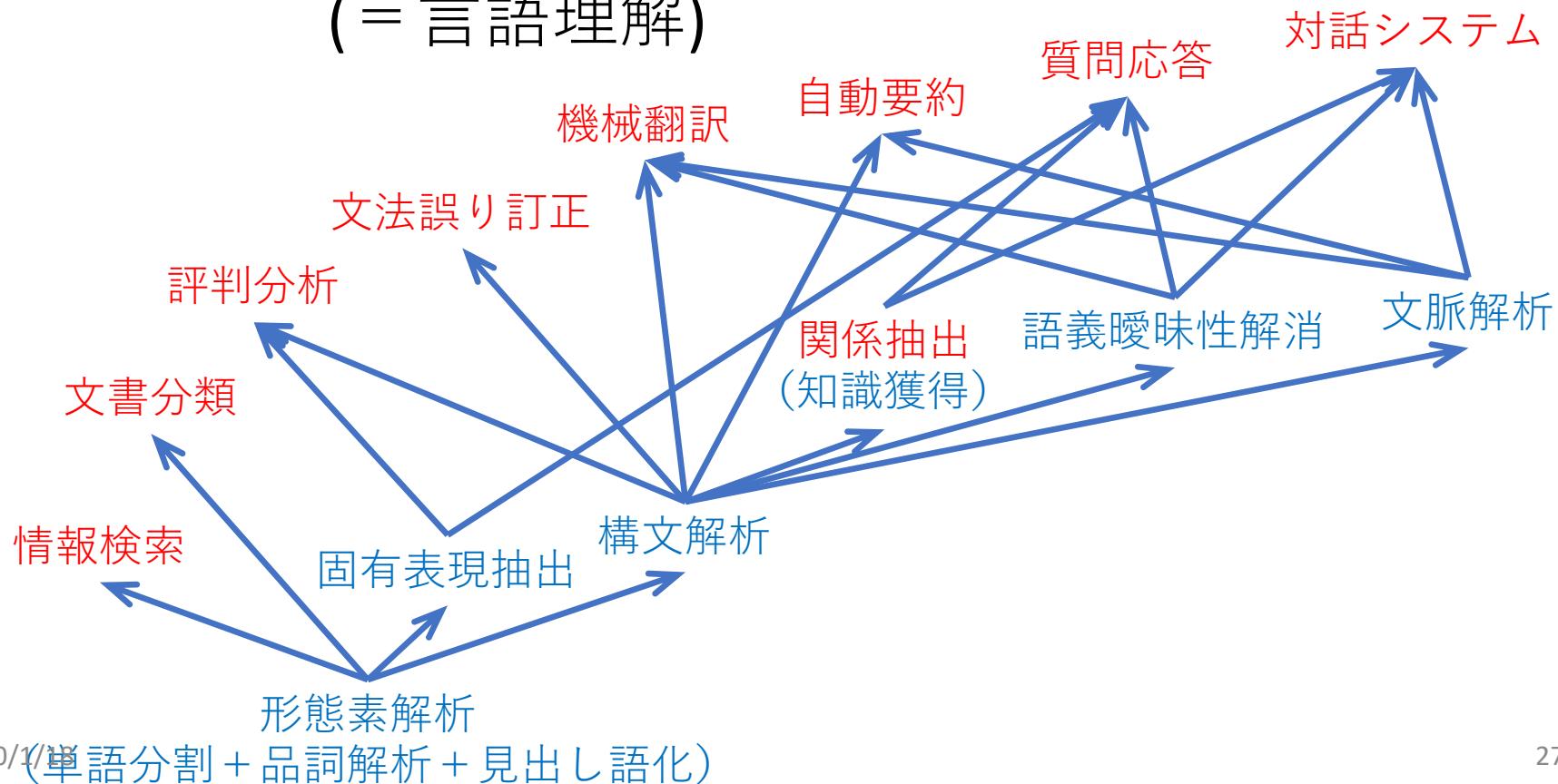
*	ツイート数
2011/03/11 14:30 - 14:40	12820
30 日本	23 出馬
20 石原都知事	19 石原
17 都知事	16 苏真
23 やばいやばい!	14 石原
448 地震↑	34 日本
64 じん!	24 東京
390 地震情報↑	346 紛糾場↑
389 宮城↑	346 紛糾場↑
354 宮城県南部↑	351 岩手県内陸北部↑
351 岩手県内陸南部↑	351 岩手県内陸北部↑
351 岩手県内陸南部↑	351 福島県中通り↑
1873 地震↑	346 紛糾場↑
1213 地震	491 余震
596 電話↑	346 紛糾場↑
1126 地震	987 余震
1038 電話	338 家
617 家	262 お台場
690 余震	253 地域↑
617 家	248 東京
582 水	371 地域
793 水	524 お風呂
636 水	510 窓
762 水	548 建設
636 水	484 大津波警報
508 水	430 情報
508 水	358 安否
640 波浪	615 家
885 地震	580 重話
517 希望	428 抵抗希望
517 希望	404 情報
545 抵抗希望	483 電話
499 備蓄	414 大津波警報
499 備蓄	402 水
444 電話	377 余震
441 水	358 水
438 抵抗希望	367 電話
397 電話	358 水
385 水	347 余震
443 水	366 情報
337 電話	311 高台
398 抵抗希望	373 情報
331 連絡	311 高台
376 情報	361 抵抗希望
288 連絡	271 水
347 水	341 抵抗希望
327 情報	295 阪神淡路大震災
282 水	286 水
233 緊急地震速報?	232 水
393 情報	347 余震
397 電話	289 避難所
322 波浪	284 水
427 抵抗希望	267 抵抗
318 情報	313 女性
310 場所	281 被害
344 1号館	278 状況
338 1号館	278 場所
306 水	256 余震
304 余震	225 女性

要素技術

応用技術実現のために解決されるべき基礎タスク

自然言語処理/計算言語学の応用・要素技術 (再掲)

- 応用技術: 言語を入出力とするアプリケーション
- 要素技術: 非構造化データである言語を構造化
(= 言語理解)



(計算)言語学が対象とする言語の領域

- Morphology: 語形成に関する言語現象
- Syntax: 文中の単語と単語の関係
- Semantics: 語や句、文の意味
- Discourse: 文を越える言語現象
- Pragmatics: 言語表現と使用者・状況の関係

Morphology (1/2)

- 対象: 語形成に関する言語現象
- 要素技術

形態素解析

東京都にあったカフェ

She loves you

単語分割

東京都 | に | あった | カフェ

品詞タグ付与

名詞 助詞 動詞 名詞

PRP VBZ PRP

見出し語化

東京都 に ある カフェ

she love you

Morphology (1/2): 形態素解析の工学的利点

- 単語の適切な照合・抽出処理に必須(情報検索等)

東京都にあるムーンファクトリーコーヒー

外国人参政権

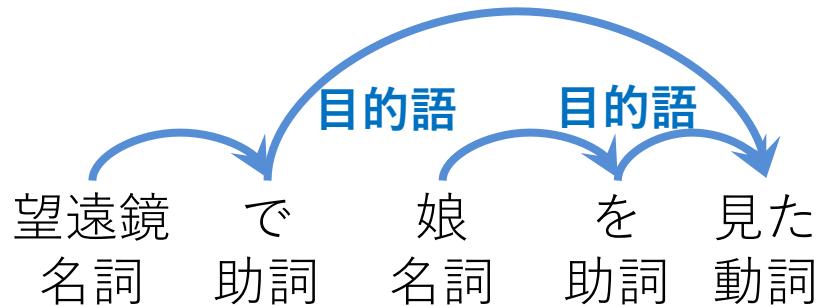
人工知能ブームの今、この先生生きのこるには

Syntax (1/2)

- 対象: 文内における単語間の統語的関係
- 要素技術

依存構造解析 (dependency parsing)

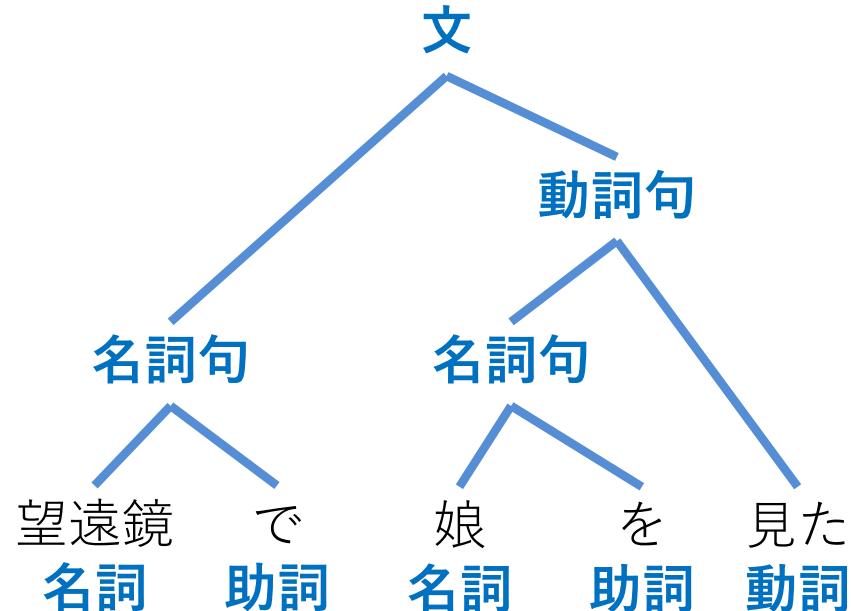
文 → 依存構造木



10/1/18

句構造解析 (constituent parsing)

文 → 句構造木 (構文木)



31

Syntax (2/2): 構文解析の工学的利点

- 構文的な曖昧性の解消
 - 実家の可愛い犬 / 尻尾の可愛い犬
 - 背泳ぎで泳ぐ娘を見る / 望遠鏡で泳ぐ娘を見る
 - 京都の話の好きなおじさん
- 構文的な表現の揺れの吸收
 - 僕はカレーを彼女に作った
 - 僕は彼女にカレーを作った
 - 彼女は僕にカレーを作った
 - • •
- 「共起語」の分類・選別
 - 彼女にカレーを作って自分はパスタを食べた

Semantics (1/2)

- 対象: 単語や文の意味の計算、同義・包含関係
- 基盤技術

語義曖昧性解消 (Word Sense Disambiguation)



単語 -> 語義

キリンの子供が生まれたと聞き、
キリンの一番搾りで乾杯した。



ORGANIZATION ARTIFACT

固有名詞抽出 (Named Entity Recognition)

固有名詞 → カテゴリ

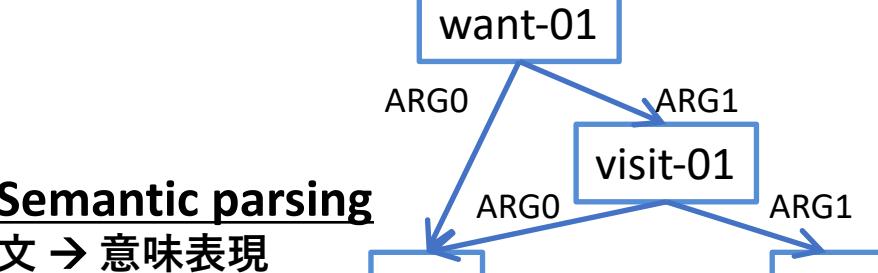
意味役割付与 (Semantic Role Labeling)
述語に対する項の意味役割同定

The boy wants to visit London

ARG0

ARG1

述語 (visit-01)



Semantic parsing
文 → 意味表現

Semantics (2/2): 意味解析の工学的利点

- 意味的な曖昧性の解消 (情報検索・機械翻訳等)
 - キリンの子供 / キリンの一番搾り
- 意味的な表現の揺れの吸収 (質問応答/関係抽出等)
 - 東京都は北海道に支援物資を送った
 - 東京から北海道に送られた支援物資
 - 北海道は東京から支援物資を送られた
 - 札幌市は東京から支援物資を送られた

Discourse

- 対象: 文を超える文脈に跨る言語現象
- 基盤技術

共参照解析 (coreference resolution)

同じ実態を指す名詞句をグルーピング

太郎は Apple の iPhone SE を買った。

次の日、それを次郎に見せびらかした。

次郎もそのスマートフォンを買った。

照応解析 (anaphora resolution)

照応詞 → (照応詞が指す) 名詞句

Entity Linking

固有名詞 → 知識ベースのエントリ



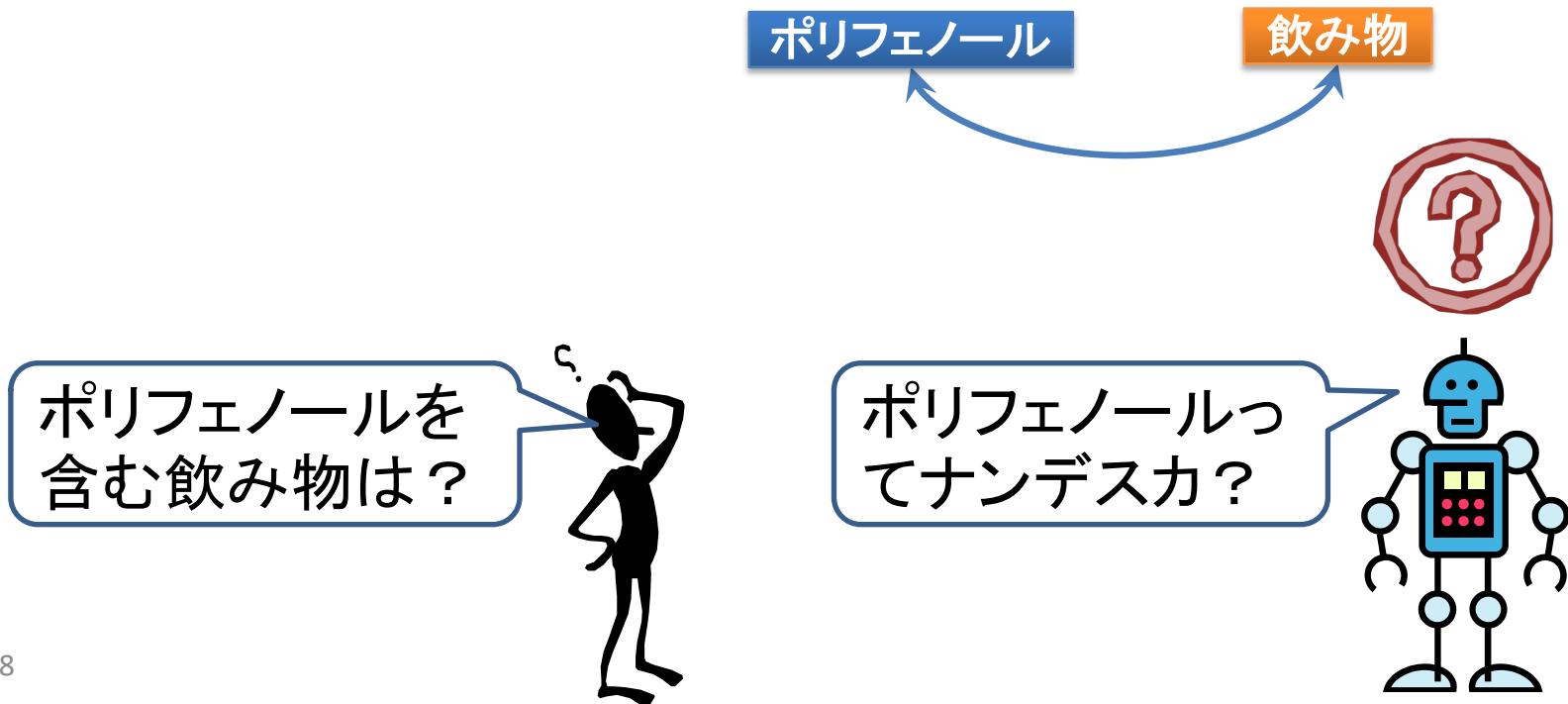
談話構造解析

Cause

關係抽出 (知識獲得)

関係抽出（知識獲得）

- 非構造化データであるテキストから構造化された知識（関係知識）を抽出・整理
 - 得られた知識はユーザまたは計算機が利用

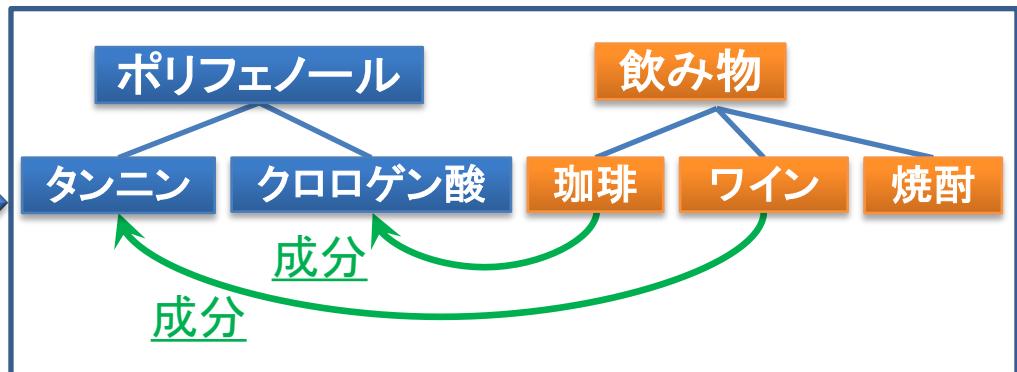


関係抽出（知識獲得）

- 非構造化データであるテキストから構造化された知識（関係知識）を抽出・整理
 - 得られた知識はユーザまたは計算機が利用

実世界の記述

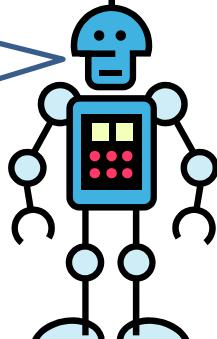
体系化された世界知識



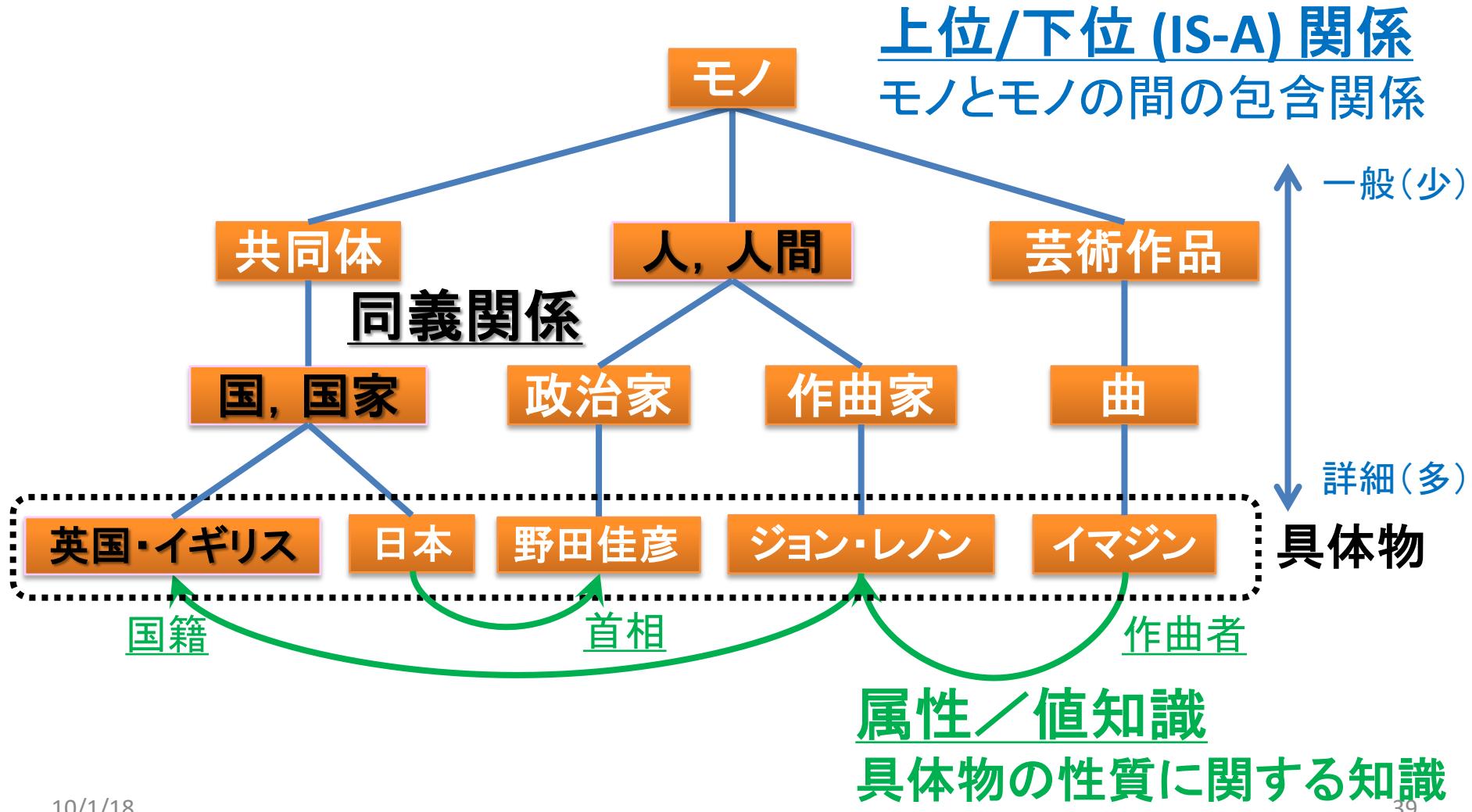
ポリフェノールを
含む飲み物は？



珈琲、ワイン
をどうぞ



モノに関する知識



コトに関する知識

- 固有表現間の関係知識
 - 受賞する（山中伸弥, ノーベル生理学・医学賞）
- 推論規則
 - 飲食店に行く → メニューを見る → 料理を注文する
 - 雨が降る → 洗濯物を取り込む（因果関係）
- 含意関係
 - 彼に事実を告げる → 彼が事実を知る
 - 外貨を売買する ↔ 通貨を交換する（言い換え）