

AI、特に言語処理研究について

平成28年 2月 26日

鳥澤 健太郎

国立研究開発法人 情報通信研究機構

- Web40億ページ以上の情報をもとに多様な質問文に回答。
- 語句の単純検索ではなく、**世界最大級の億単位のエントリを持つ知識ベース・辞書**を用い、テキスト間の同義、因果関係等を自動認識し、質問への回答や、**一連の世界初の技術**により仮説の推論や質問の提案まで行う。
- 百科事典や医療等の特定科学分野の知識だけではなく、**社会の潜在リスク、想定シナリオ、イノベーションのヒント**等について、**ネットで今まさに一般国民が書いている情報**も幅広く提供可能。
- **大規模クラスタ**(計算機300台)で大規模データの高速な意味的分析を行う、**日本の研究機関では前例のない大規模自然言語処理/人工知能システム**。<http://wisdom-nict.jp>にて一般公開中。

例1 「東京オリンピックで何を心配すべきか？」

質問を入力

回答を表示

検索結果【38件】

質問の回答

テキスト検索

関連する質問

質問を入力

回答を表示

資材高騰

工事費増

建設費増加

詐欺(架空の土地取引)

宿不足

物流の支障

コミケ開催

地方から関東への人材流出

関西の地盤沈下

人手不足

その他、猛暑による選手の体調不良、災害リスク、テロ行為、台風、放射能等の回答を表示

例2 ①質問：地球温暖化が進むとどうなる？

② 450件の回答

- 海水温が上がる
- 台風が巨大化する
- プラクトンが減る
- 被害総額年100兆円

③上の回答に基づき、システムが「海水温が上がるとどうなる？」という質問を提案。利用者はこの提案をクリック。

④ 450件の回答

- メタンが放出される
- サンゴの白化が進む
- 腸炎ビブリオ(大腸菌)が増える…

その後、気候変動による腸炎ビブリオ由来の食中毒の増加を専門誌が報告 Austin-Baker, C. et al., Nature Climate Change, , 3:73-77(2013)

NICT これまで：対災害技術：DISAANA & D-SUMM

- WISDOM X の技術を応用し、SNS (Twitter) 上の災害関連情報をリアルタイムに意味的に深く分析・整理して提供し、一刻を争う中での状況把握・判断の支援を行うシステム
- DISAANAでは、災害に関連する質問への回答機能(世界初)、指定されたエリア内の被災報告の自動発見機能(世界初)、デマ判定支援の機能等がリアルタイムに可能
- DISAANAはネット上に一般公開されており、<http://disaana.jp> でPCやスマホから誰でも使用可能。D-SUMMは今夏一般公開予定

対災害SNS情報分析システム *DISAANA*

平成27年9月10日、台風18号豪雨の際、質問「どこで救助を待っているか？」に対してTwitterから発見された回答

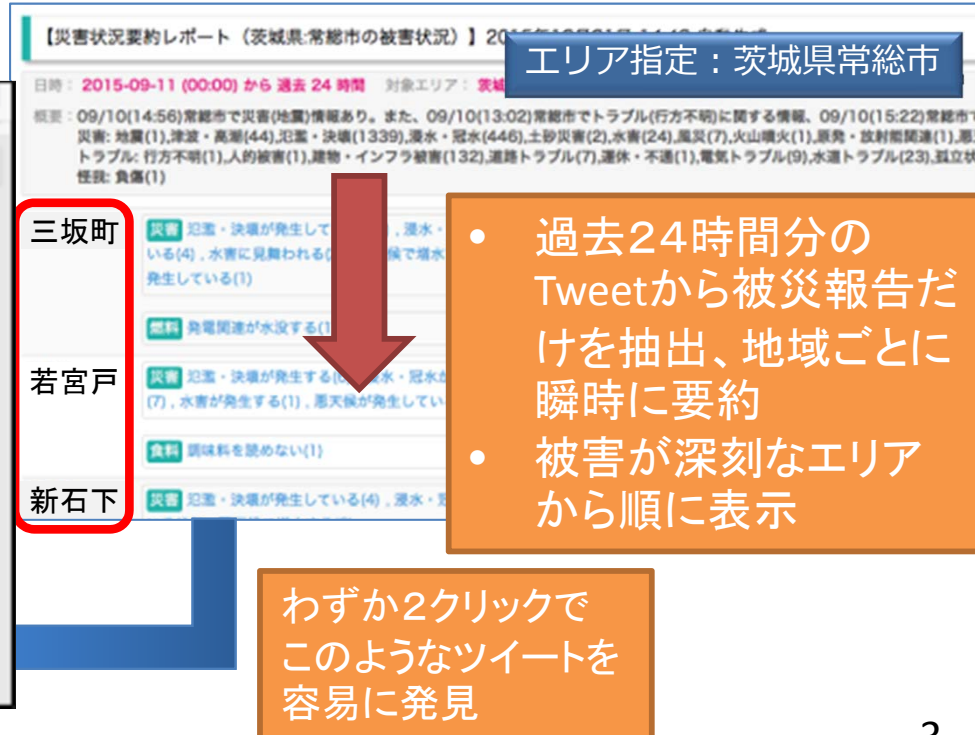
- 救助要請が出されている地点をリアルタイムに地図表示。同様の情報は、通常の検索エンジンでは1万件以上の情報を人が見て初めて取得可能
- Tweetされてから5秒後にはシステムに反映



DISAANAで発見された救助要請のTwitter情報

スマホでも利用可能

災害状況要約システム *D-SUMM*



【災害状況要約レポート (茨城県常総市の被害状況)】 2015-09-11

日時: 2015-09-11 (00:00) から 過去 24 時間 対象エリア: 茨城県

概要: 09/10(14:56)常総市で災害(地震)情報あり。また、09/10(13:02)常総市でトラブル(行方不明)に関する情報、09/10(15:22)常総市で災害: 地震(1)、津波・高潮(44)、冠水・決壊(1339)、浸水・冠水(446)、土砂災害(2)、水害(24)、風災(7)、火山噴火(1)、原発・放射能関連(1)、その他(1)、行方不明(1)、人的被害(1)、建物・インフラ被害(132)、道路トラブル(7)、運休・不通(1)、電気トラブル(9)、水道トラブル(23)、孤立状態(1)、負傷(1)

エリア指定: 茨城県常総市

三坂町 災害 冠水・決壊が発生している(4)、水害に見舞われる(1) 浸水・冠水が発生している(1)

若宮戸 災害 冠水・決壊が発生している(1)、浸水・冠水(7)、水害が発生する(1)、悪天候が発生している(1)

新石下 災害 冠水・決壊が発生している(4)、浸水・冠水(7)、水害が発生する(1)、悪天候が発生している(1)

食料 調味料を認めない(1)

過去24時間分のTweetから被災報告だけを抽出、地域ごとに瞬時に要約

被害が深刻なエリアから順に表示

わずか2クリックでこのようなツイートを容易に発見

(自律的)社会知解析技術：社会に流通している知識、すなわち**社会知**を**自律的**に分析でき、また自律的に賢くなる技術

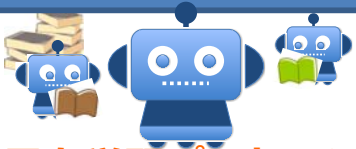
①社会における問題の自動検知技術

「少子化」は大問題！



社会問題から技術開発の課題まで様々な問題を検知

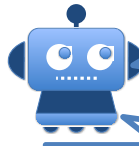
⑥自動検証結果に基づく自己学習技術



検証結果を学習プロセスにフィードバック

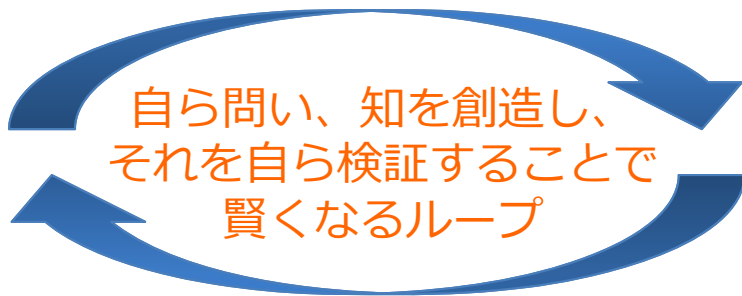
②質問自動生成技術

様々な有用な質問をシステムが自動生成



問題の解決策を問う質問：
「少子化はどうやって解決する？」

解決の具体例に関する質問：「少子化はどこで解決したか？」
「何故フランスでは少子化を解決したか？」...等々



⑤回答と仮説の自動検証技術



仮説Hは：
• 論文Aによれば...
• 白書Wによれば...
従ってHの信憑性は...

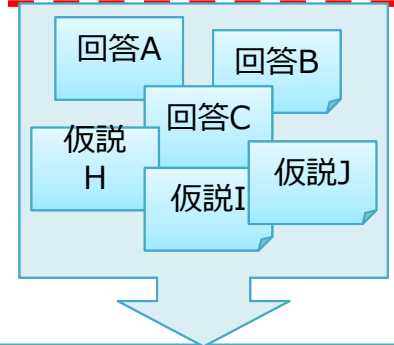
仮説やその類似物がどこかに書かれているかチェック

③WISDOM Xで回答や仮説を取得



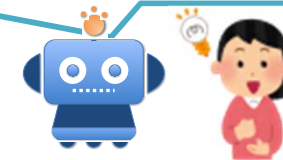
Web、論文、白書等、文脈まで考慮に入れて分析

第3期中長期計画はこの周辺の技術の一部のみカバー



④回答・仮説統合・要約技術

理解が容易な形でユーザーに提示：少子化の解決策としては、税制改革、資金援助等がある。フランスではバカンスでの子供の旅費...、税制改革では、...



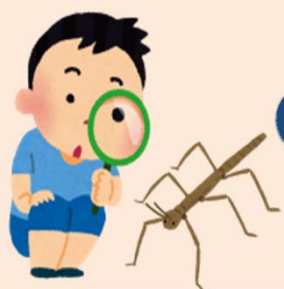
なるほど！

新開発の技術は適宜WISDOM Xに導入して一般公開

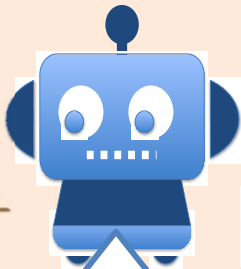
• ポイント

- 自ら問いを発するAI ← 結構、根源的
- 問いへの断片的回答・仮説だけでなく、多数の回答
 - 仮説を統合・要約。素人でも理解を容易に
 - 例：Wikipediaの記事風だけれど、特定の観点から深掘りした記事を出力(地球温暖化の経済的インパクトのみにフォーカスしたレポート等)
- 問いに対して得られた回答、仮説を検証し、自ら賢くなるAI
- もちろん、(テキスト) ビッグデータ、機械学習は出発点として必須

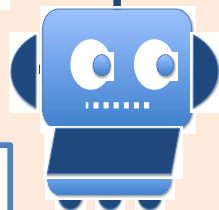
万能対話ロボット(教育、高齢者)



ナナフシってオスなしでも繁殖するよ。



車いすで楽しめるダンスがあるそうです。



シンクタンク、社会調査

少子化で耕作放棄地が急増！



それでA地方の雇用を増やせますね！



耕作放棄地で行うビジネスには、太陽光発電、魚類の養殖、植物性プランクトンの養殖。A地方に適しているには植物性プランクトンの養殖…

民間企業のイノベーション支援

南米でディーゼル油を生成する真菌(水虫の類似物)が発見される！



その作戦でいきましょう！



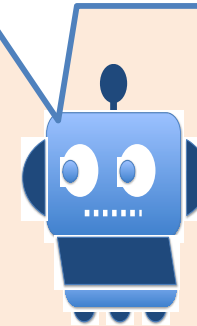
我が社のプラントによく適合しているので、プラントとセットで販売できるかも。

企業のコンプライアンス対策

排ガス試験検出のための条件分岐は…

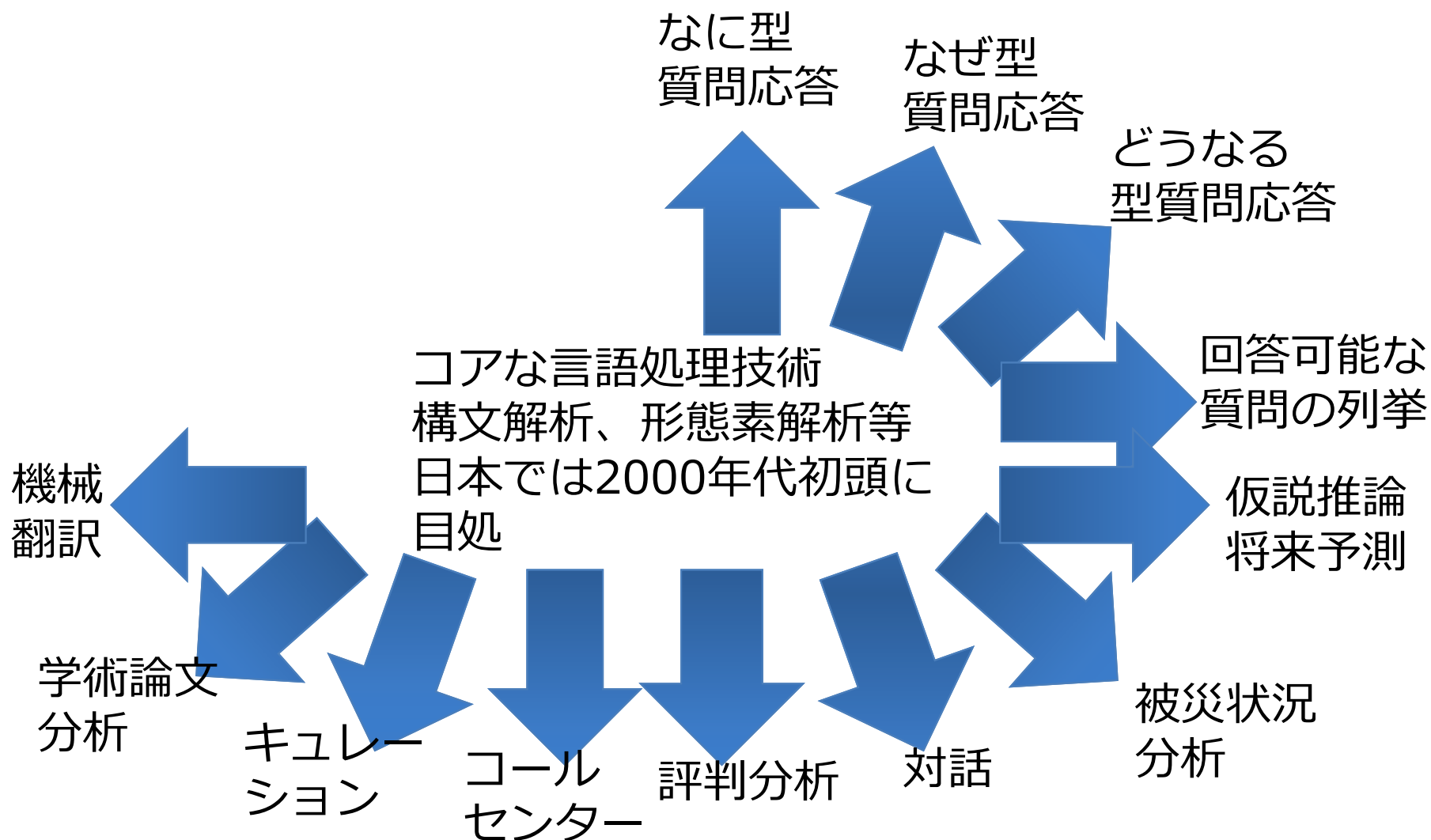


排ガス試験に関する対策を施すことは法令違反です



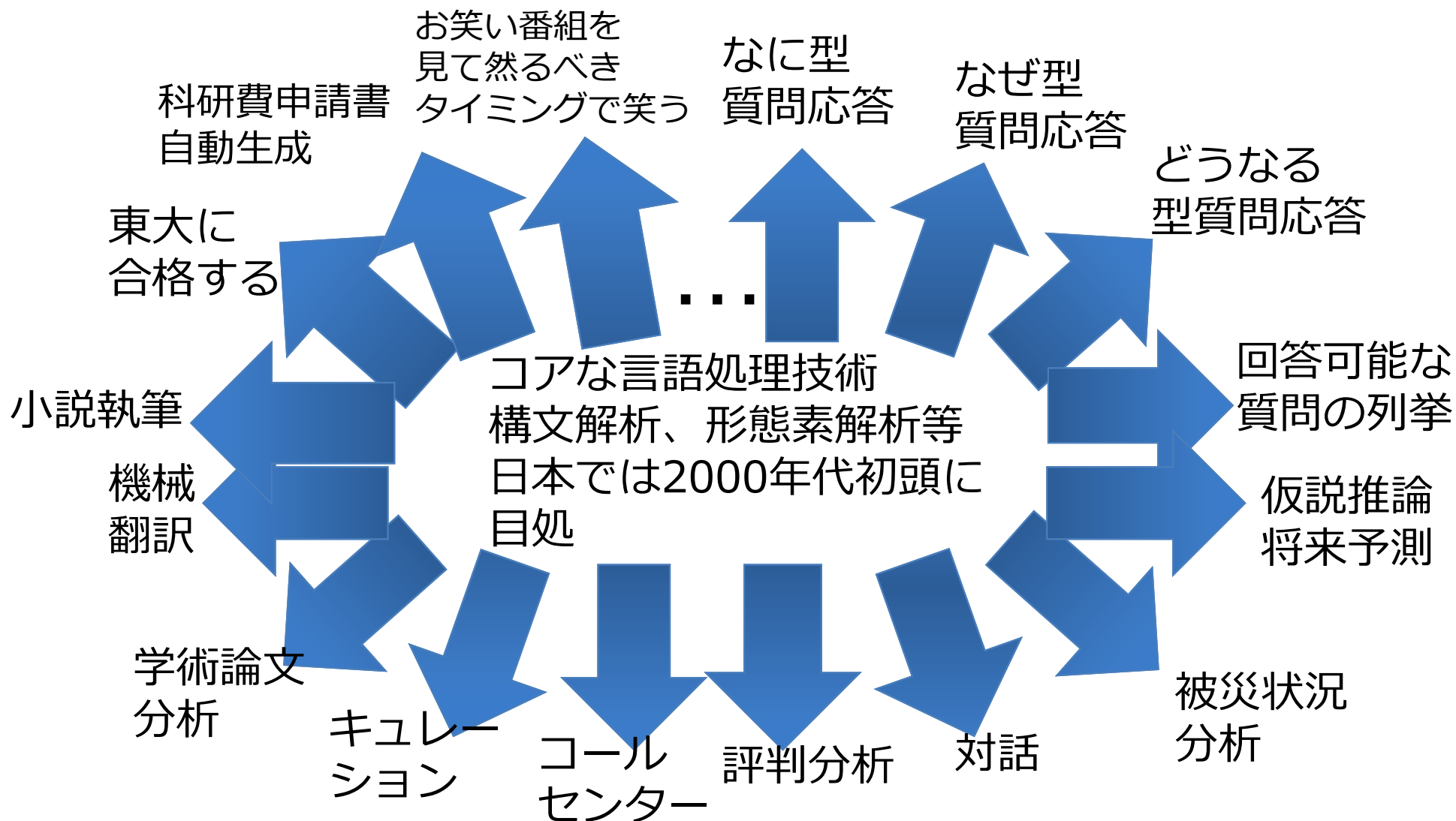
言語処理研究の現状

- 言語処理業界でカバーされているタスク・分野は人間が行えるタスク・分野のごくごく一部



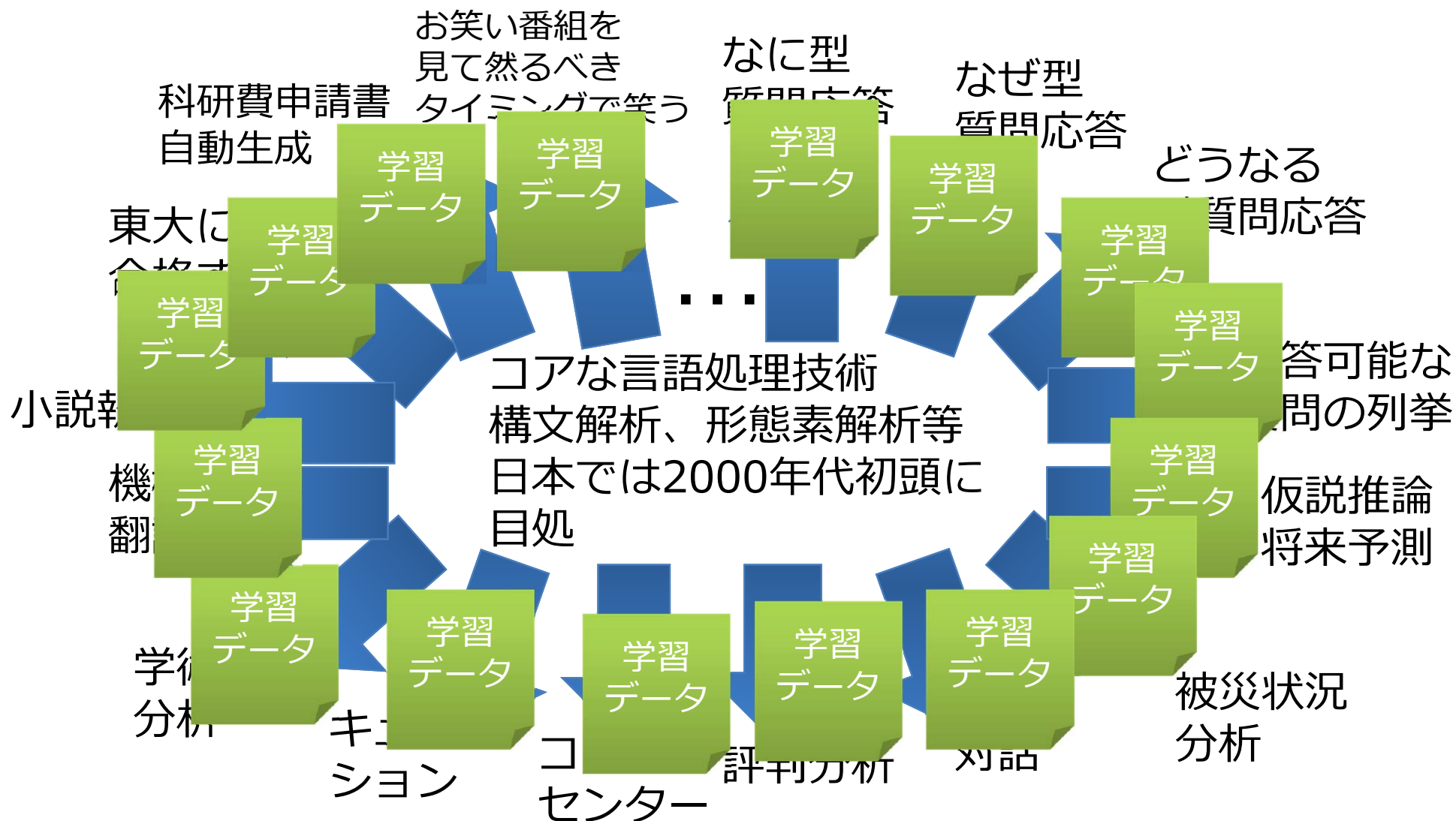
言語処理研究の現状

- 日々、新たなタスクが提案され、売り出される
- 現状は、ドラえもんには程遠いが、有望な提案も多数



言語処理研究の現状

- しかし、当分、タスク毎に学習データは必須



- しかし、当分、タスク毎に学習データは必須

お笑い番組を

今後の問題は、一昔前の「知識ボトルネック」ではなくて、「**学習データボトルネック**」か？

- 学習データ = アプリの仕様
- とはいえ、「知識ボトルネック」よりはかなり前進
- 逆手にとれば強みともなる
- 教師なし学習が有効なのは、超基礎的な話か、論文を書くときだけというのが個人的感触
- 深層学習で解決か？

分析

コミュニケーション

コ
センター

評判分析

対話

分析

学習データだけではなくて、大規模な辞書、知識ベースを機械学習と人手併用で構築

辞書、知識ベースの構築

- 数万～数億エントリ規模
- 辞書は複数のアプリで共用化
- 人手で相当量をチェック
- 最終的なアプリの精度・速度向上

同義性認識用等の辞書自動構築
(ACL 2008, 2009, 2010, 2011, EMNLP 2009:2本, NAACL 2013)

Excitation Polarities (述語の意味的極性)
(EMNLP 2012)

Sentiment Analysis
(NAACL 2010)

パターン間矛盾関係
(EMNLP 2013)

パターン間同義関係
(EMNLP 2015)

Twitter上のデマ検出

なに型質問応答

どうなる型質問応答
(ACL 2014, AAI 2015)

なぜ型質問応答
(EMNLP2012, ACL 2013, AAI 2016)

Twitterからの被災状況、救援状況抽出
(ACL 2013)

ユーザから直に見えるアプリ

- 深層学習等の新技術の導入：GPGPU搭載クラスタの導入を検討中
- 深層学習等で辞書の必要性が薄れる可能性
 - これまでの感触では、「ものによる」
- ただし、辞書なしでは必要な学習データが増える場合や、作れない場合もある
 - 例えば、ランダムサンプル中の正例が極端に少ない場合、今までは、辞書＋ヒューリスティックスで正例の濃度を高めてからラベル付与
- 研究的には機械学習は高コスト、高リスク
 - 辞書作成者の人件費<< 機械学習研究者の人件費
 - 機械学習はやってみないとわからないし、かなり時間もかかる
- 大抵の場合、人手で作った辞書の方が正確。また、辞書は使いまわせる
- 大抵の場合、辞書のlook upの方が分類器よりも速い
- ミッションクリティカルなタスクで機械学習はちょっと。。。
- 個人的にはニューラルネットだけではわかった気がしない→学問として長期的にはどうなの？



重要なのは先進的な機械学習と人手による
辞書構築の最適なバランスを見つけること

- 言語処理、特にアプリに関しては未だ無数の可能性
- 障害は「学習データボトルネック」
 - 90年代の「知識ボトルネック」に比べればだいぶ前進
- 方法論：先進的な機械学習＋人手によるデータ構築＋ビッグデータ
 - アプリのインパクト、コストパフォーマンス、アカデミックな価値等考慮しつつ、ベストなバランスを狙うべき
- 興味：脳科学とのインタラクション：脳内に実在する辞書的情報はなにか？
- アジャイルな実装力も鍵(DB、クラウド/クラスタ)
 - 実はWISDOM X開発の最大の危機は言語処理プログラムの起動コストとThread管理→自社開発のミドルウェアで解決