資料2-1

DIRECT
Data-driven Intelligent System Research Center

自然言語処理のための データ整備について

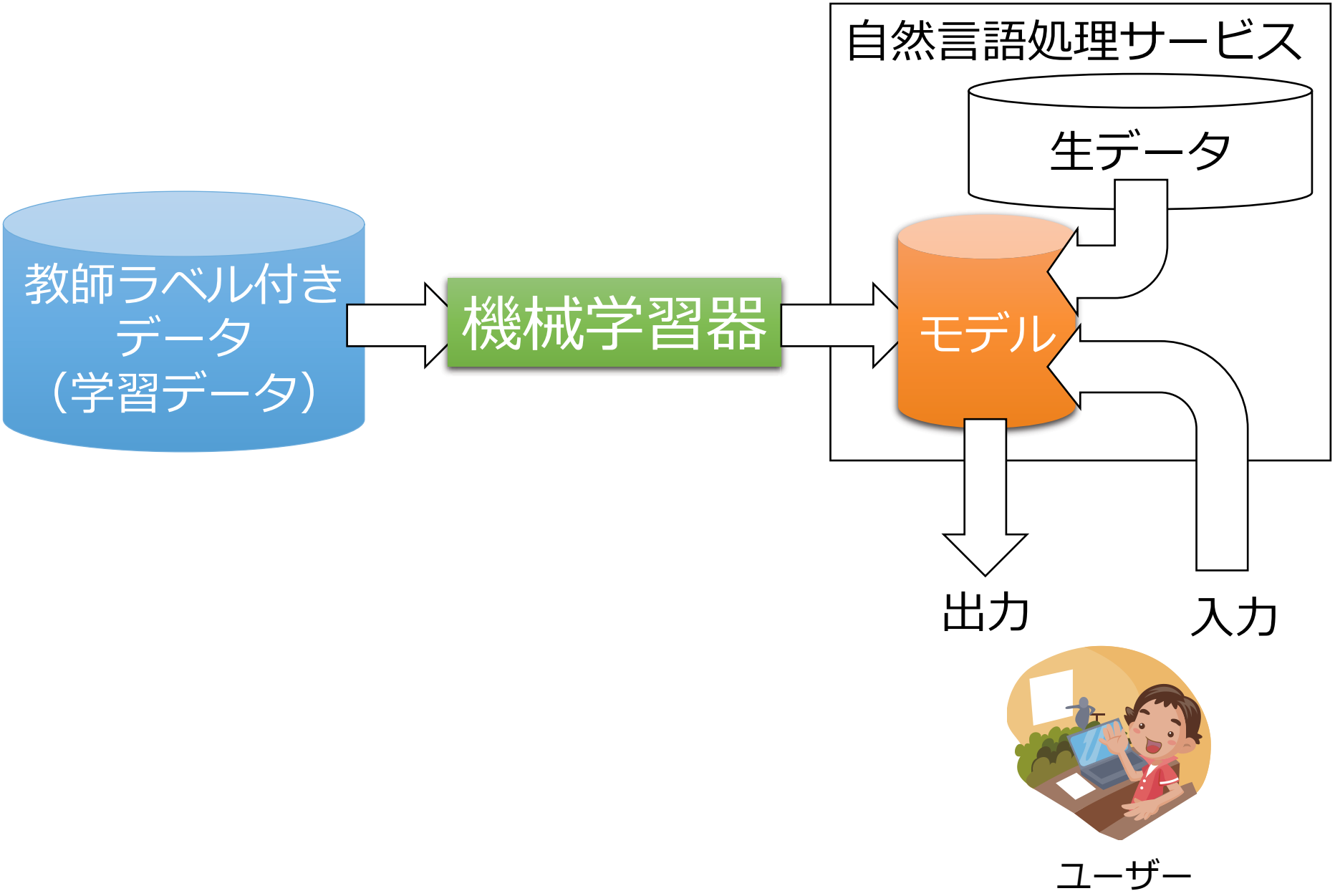
鳥澤 健太郎

NICT, DIRECT

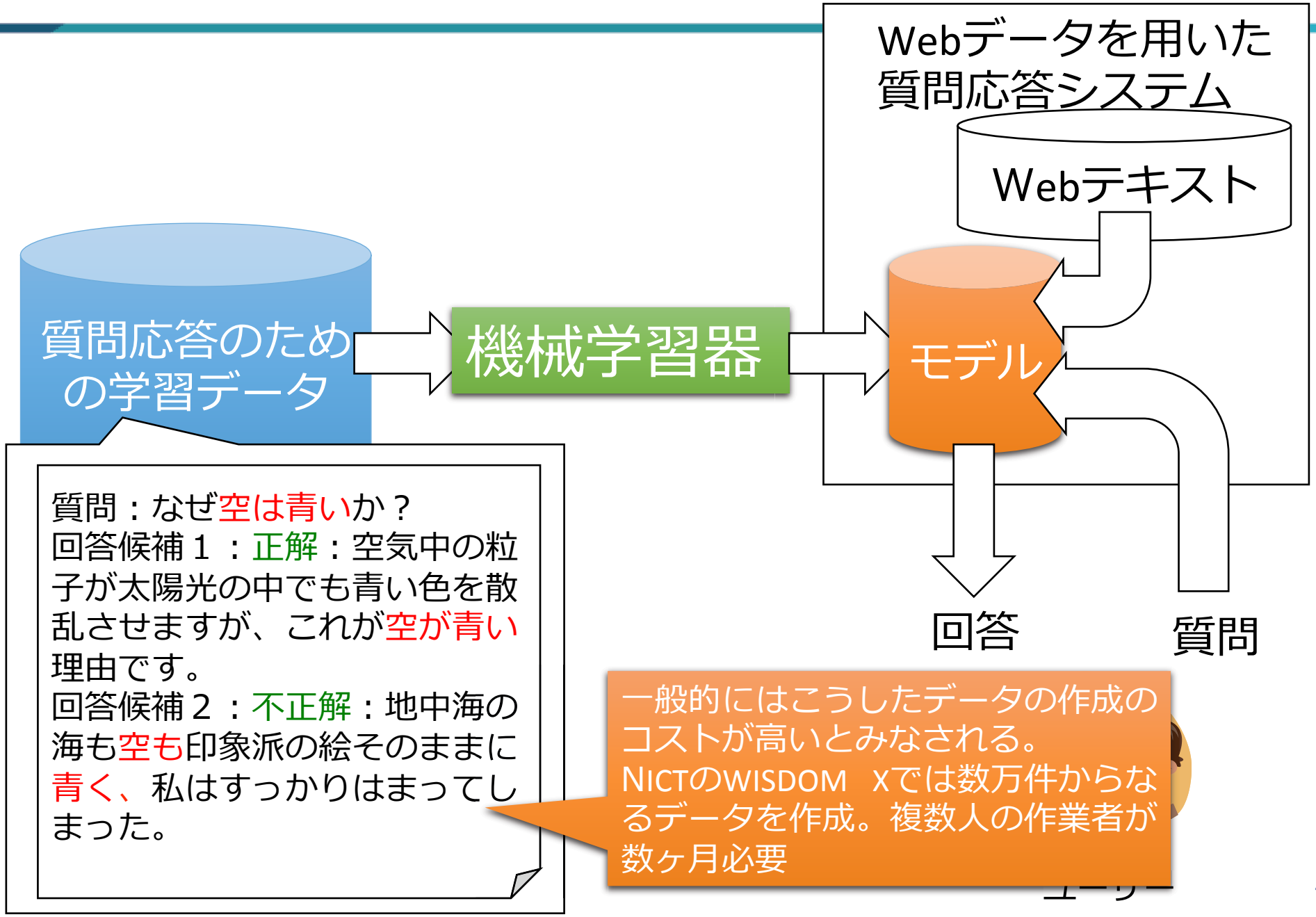
2017.2.14

NICT

自然言語処理における機械学習の利用例



WISDOM Xの一部（「なぜ型」質問応答）



データのタイプ

- 二種類のデータ
  - 生データ
    - 人工知能の研究開発での利用を意識した加工が一切なされていないもの
    - 例：生のWebテキスト、科学技術論文
    - 機械学習等でそれだけで有用な場合は「まれ」
      - 例：SNS上のユーザがタグ付けした画像情報
    - 普通は、機械学習で学習したモデルの適用先
      - 例：質問の回答をその中から探してくるテキスト（そのテキストに、学習済みのモデルを適用することで回答を見つける）
  - 教師ラベル付きデータ（いわゆる「学習データ」）
    - 生データに機械学習用のラベルが付与されたもの
    - 機械学習の学習時に学習データとして利用
    - 生データに比して作成には圧倒的にコストがかかる
      - 人がラベルを付与する作業が必要なため
    - 例：質問と、その質問の回答の候補が書かれたテキストのペア
      - 回答候補が正解か、不正解かを示すラベルがふられている

なぜ学習データが大量に必要か？

- 自然言語では同じことを言う方法が無数にある

質問：なぜ空は青いか？

回答候補１：正解：空気中の粒子が太陽光の中でも青い色を散乱させますが、これが空が青い理由です。

回答候補２：正解：空気中の粒子が青い色を散乱させるので、空が青くなります。

回答候補３：正解：空気中の粒子が青い色を散乱させ、空が青くなります。

回答候補４：不正解：地中海の海も空も印象派の絵そのままに青く、私はすっかりハマってしまいました。

８０年代、９０年代の対処法
パターン、ルールをプログラマが書く

```
If
  (これが（質問の内容）理由です
  or
   ...ので、（質問の内容）。)
then 正解
else 不正解
```

ルール中の「理由」は「原因」で置き換えてもいい。
回答候補３、４は「理由」のような手がかりがないので、抽出できない....

エンドレスな修正、悩み
⇨開発者が疲れてプロジェクト失敗

なぜ学習データが大量に必要か？

- 自然言語では同じことを言う方法が無数にある

質問：なぜ空は青いか？

回答候補１：正解：空気中の粒子が太陽光の中でも青い色を散乱させますが、これが空が青い理由です。

回答候補２：正解：空気中の粒子が青い色を散乱させるので、空が青くなります。

回答候補３：正解：空気中の粒子が青い色を散乱させ、空が青くなります。

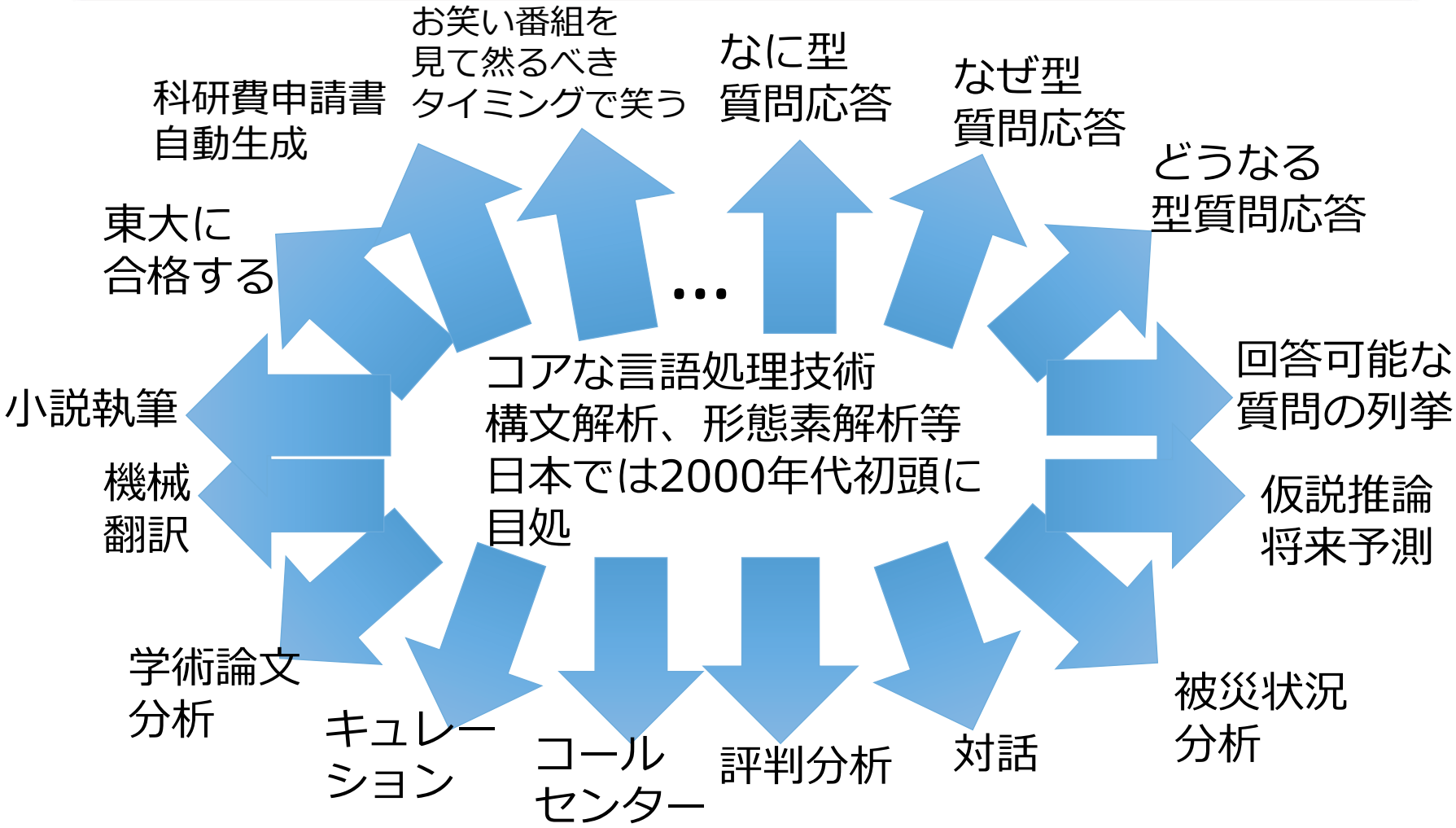
回答候補４：不正解：地中海の海も空も印象派の絵そのままに青く、私はすっかりハマってしまいました。

最近の対処法

- 疲れるプログラマではなくて、疲れない機械学習器に「パターン、ルールのようなもの」（実際には非線形の関数、これをモデルと称する）を自動的に書いてもらおう
- ただし、機械学習器には左側にあるような正解と不正解の事例（つまり、学習データ）をたくさん与える必要
- たくさん＝自然言語の無数のバリエーションをある程度カバーできる量
- パターンを書くのは熟練のプログラマでなければダメだが、正解、不正解の判定は普通の人でも可能⇨コスト的に見合う

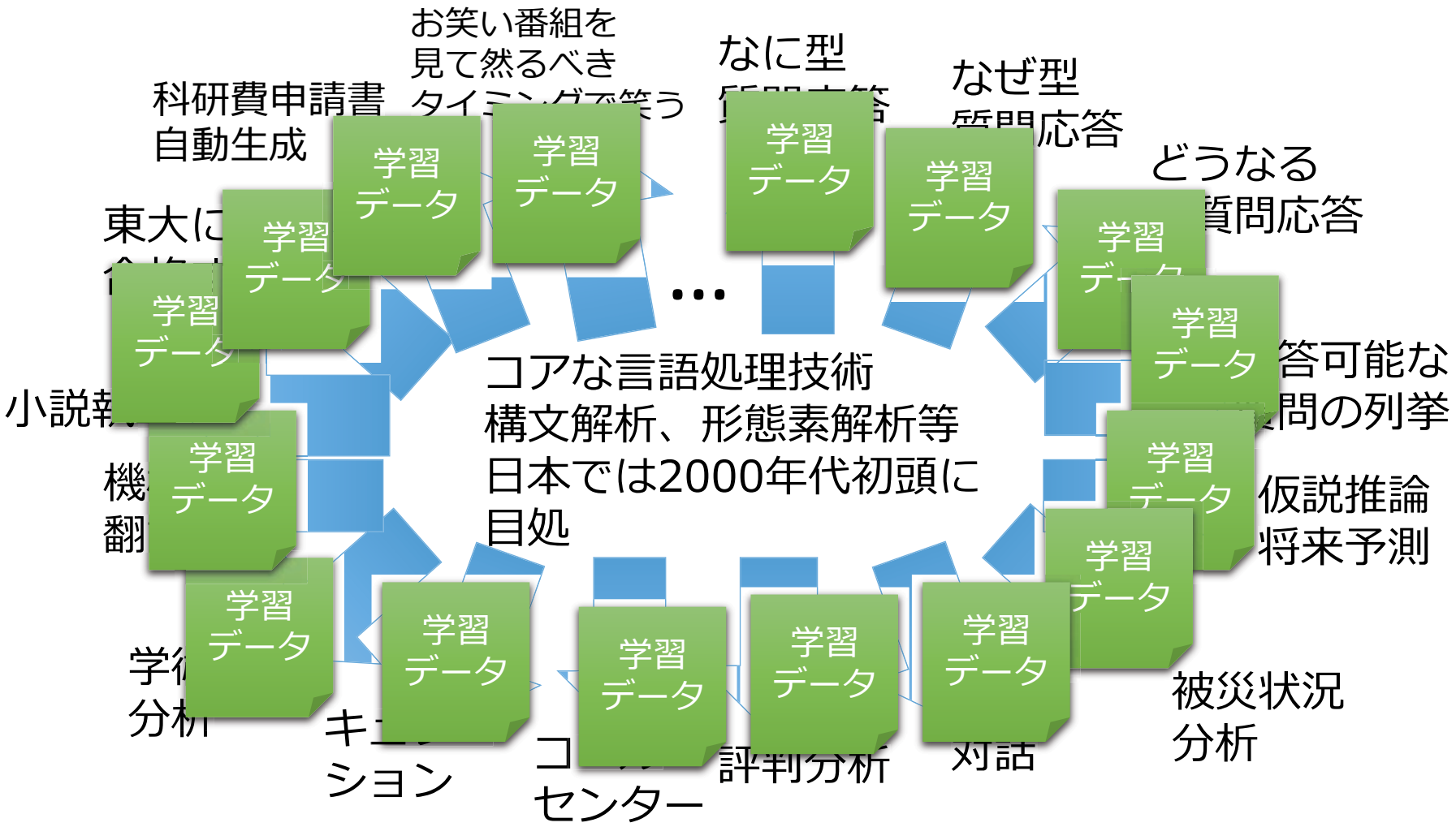
言語処理研究のタスク

- 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

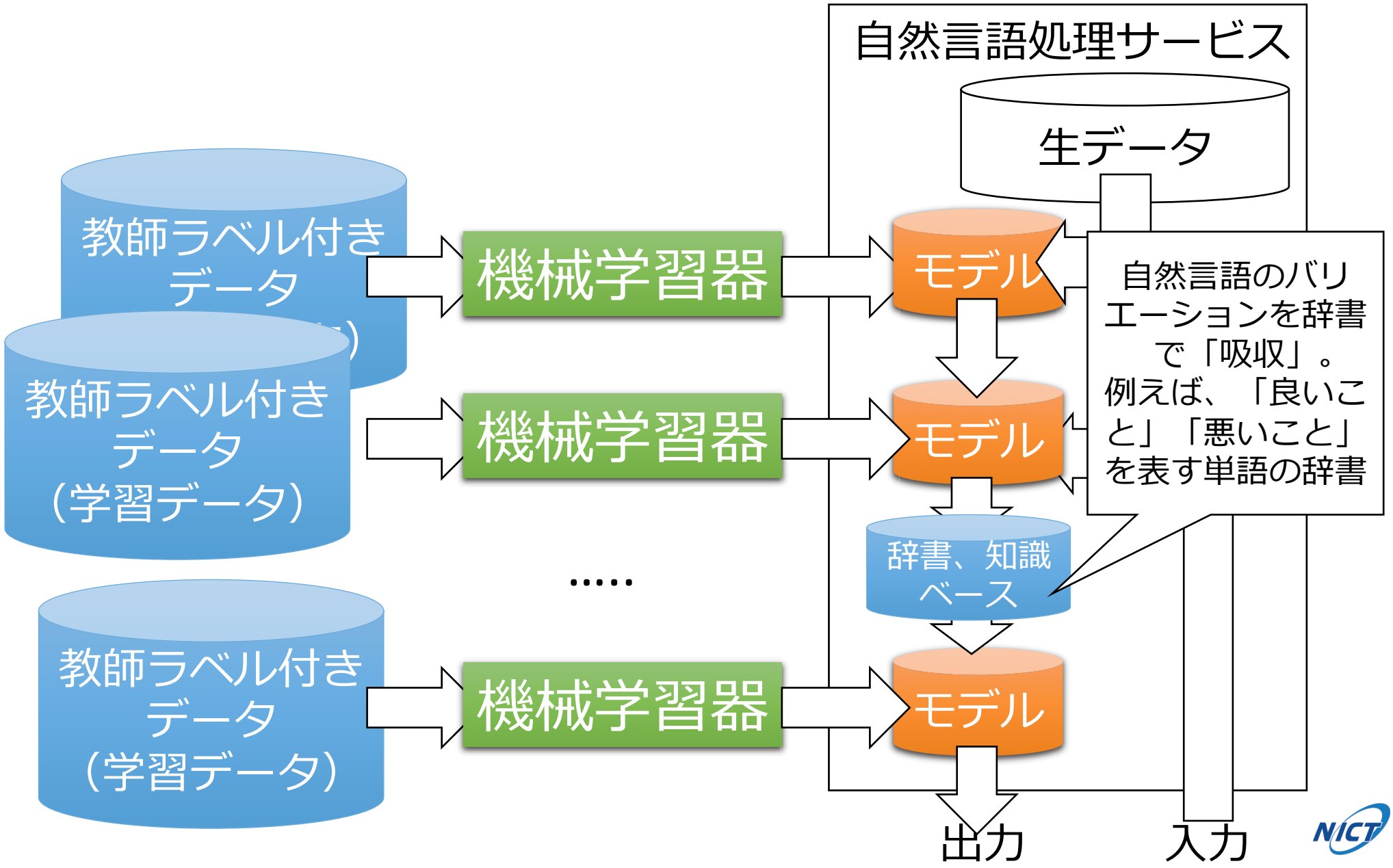


言語処理研究のタスク

- しかし、当分、タスク毎に学習データは必須



言語処理システムのアーキテクチャ



要するに何が言いたいかというと

- 学習データの設計は目標とする実サービス、アプリケーションに依存する
  - 分野、タスクの設計に依存
- 学習データの設計は目標とする実サービス、アプリケーションのアーキテクチャにも依存する
  - タスクの分割、速度、前処理等に関する制約にも依存

- この２点に十分注意を払っていなければ、データを作っても、論文が書けるだけで、実運用、社会実装には結びつかない
  - データ作成とサービスのテストはサイクルにならなければならない
- 逆に言うとデータだけがある状況というのは技術の社会実装を阻害する可能性もある
  - 実現不可能なサービスに向けた研究だけが進捗する
  - 一般公開されているデータが論文を書くのにしか使えないのは、実サービスとデータのカップリングが不十分なため
- NICTでは、サービス、アプリケーションの開発を行う研究者が学習データの構築法、設計まで考える
  - サービス、アプリケーションの実装が第一、データは重要だがその手段
  - 場合によっては実ユーザのヒアリング等も参考にしつつ学習データを設計する

要するに何が言いたいかというと

- その他に考慮しなければいけないこと
  - そもそも作業者が実行可能なタスクか？
    - 例：医療に関するデータ作成は医療従事者以外に作れるのか？
    - 例：各分野の専門家は十分に確保できるか？
  - 作業マニュアルは十分に分かりやすいか？
    - 例：ゼロ照応の同定を行ってください　=> 省略されている主語、目的語を補って、意味がとおるものを書いてください
  - 作業者間でラベル付けが一致するか？
    - 作業マニュアルの出来不出来に大きく依存

- データ作成者と自然言語処理研究者、システム開発者が一緒になって上記の点を十分に検討しなければ良いデータは作れない
- ただし、翻訳は翻訳者の基準が確立しているであろうから若干状況は違うであろう

データの提供

- NICTではWISDOM X、 DISAANA等のライセンスを開始
  - 一部データの提供は著作権法の問題から不可能
    - Web等から抽出した長い表現
  - その場合は、データではなく機械学習のモデルや抽出のプログラムを提供

- 学習データ等もサービス、アプリに依存する⇨
  ソフトウエアとデータはセットで提供しなければ意味がない

- NICTでは、今後、他組織との連携は人材交流が必須と考える
  - NICTの肩書きがなければ、著作権法等の問題によりNICT内部のデータ(特にWeb由来のもの）を触る事ができない
  - NICT内で作業を行い、成果をライセンス
  - 出向元でもデータ作成は必須。そのノウハウを伝授。
  - どういったタスク、アーキテクチャをターゲットとするかよく話し合いをする必要

本WGに期待するもの

- 上記のようなデータ作成者とサービス提供者(データ利用者）がWin-winの関係を構築し、ポジティブなスパイラルを回すにはどうしたらよいか？
  - どのような分野、タスクをアドレスすべきか？
  - サービス提供者は何が実現できると嬉しいか？
  - サービス提供者（データ利用者）はどういうコミットメントが可能か
  - データ作成側は何が実現できると嬉しいか？
  - 密なコミュニケーション、人材交流、育成？

- ボランティアベース、理想論ベースでは長期的には絶対に上手くいかない
  - データを作る側にも成果（論文？実システム？）が求められる
  - 論文の数を稼ぐだけならば、データを作るという戦略は損
  - 過去の取り組み

- NICTの立ち位置（案）
  - NICT: なるべく幅広い分野に対応するシステム＋データ（例：WISDOM X)とパイロットシステムとしての分野特化型システム(例：DISAANA)
  - 他組織：社会的ニーズと直結した分野特化型システム

NICTサイドの目標：ネット上の膨大な知識を活用する博学対話ロボット

応用例：車内対話ロボット

対話ロボット：今日の一番のニュースは英国がEUから離脱する…

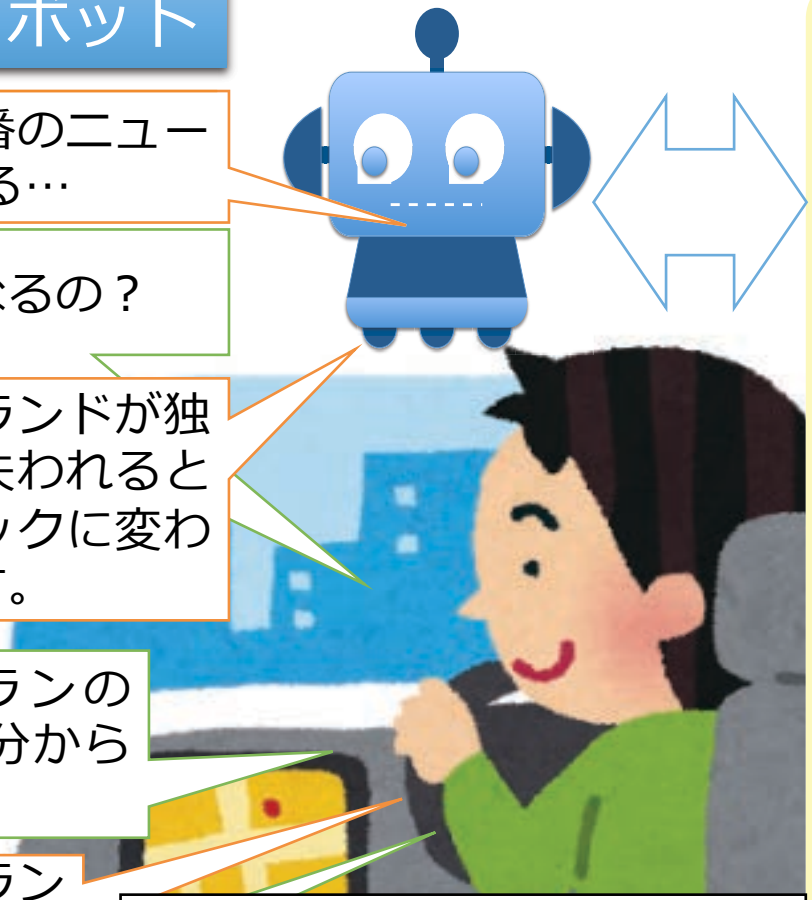
ドライバー：それでどうなるの？

対話ロボット：スコットランドが独立するようです。雇用も失われるとか。英国国教会がカトリックに変わるとかいう話まであります。

ドライバー：スコットランドの独立とか、何が起こるか分からないねえ。

対話ロボット：スコットランドが独立すると、ポンド、イギリス株に打撃です。

ドライバー：イギリス株売ります。証券会社に電話して。

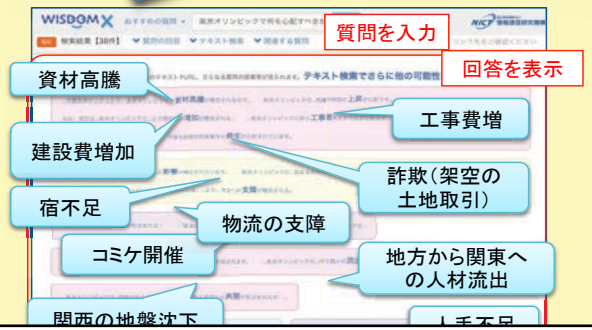


通勤中に
一人ブレインストーミング

WISDOM X

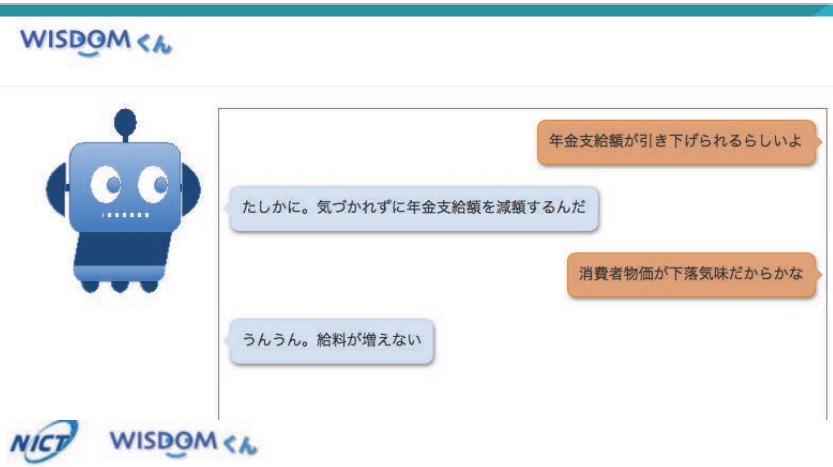
膨大なインターネット上の文字データ（40億ページ、原稿用紙220億枚相当）をもとに、質問に回答や仮説を提供するシステム
（一般公開中）
左の対話例中の情報はすべてWISDOM Xから実際に取得されたもの

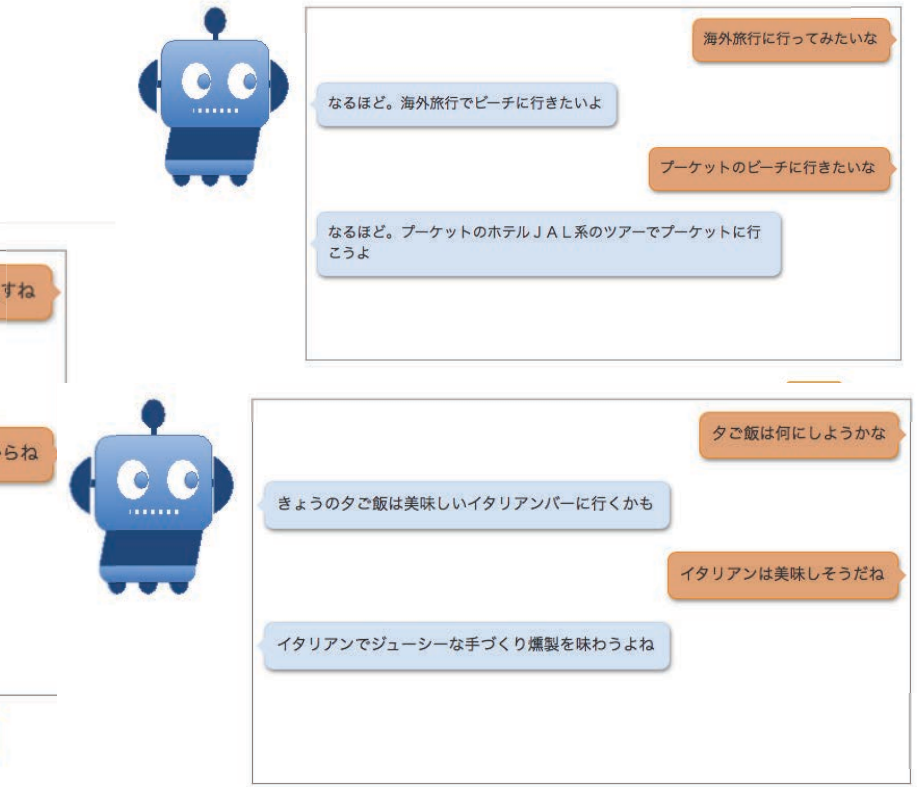
質問「東京オリンピックで何を心配すべきか？」



物理的なロボットやセンサーと対話の組み合わせはNICTの他グループが研究中。ロボカップで複数回優勝

現状：WISDOM Xを使った雑談システム





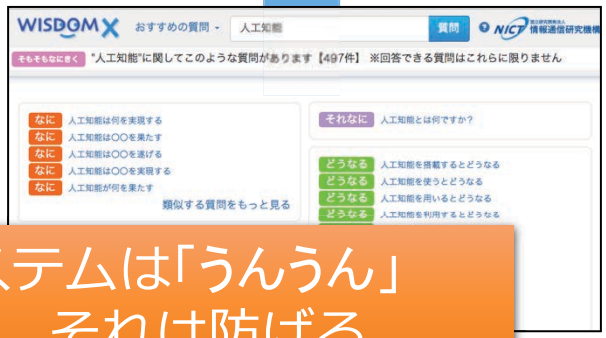
WISDOM Xを使った雑談システム(No.1)

ポイント：深層学習でシステム発話を生成するのではなくて、
WISDOM X用の質問を生成し、その回答を修正してレスポンスにする

ユーザ入力：ジョブズ,すごいよね

深層学習のモデルでユーザ入力を
WISDOM X用の質問に「翻訳」

質問（複数）
ジョブズが何を見下す?
ジョブズが何を作る?
ジョブズが何に成功する?
....



回答「Windows」
元文「ジョブズがWindowsを揶揄した」
（質問「ジョブズが何を見下す」に対して）

回答と元文からシステム出力を合成

システム出力：Ｗｉｎｄｏｗｓを揶揄するよね～

ナイーブな深層学習による対話システムは「うんうん」「あ、そう」といった相槌ばかり言うが、それは防げる

WISDOM Xを使った雑談システム(No.2)

どのような質問を生成するかで、対話システムの「人格」も決まる

例：新聞記事を題材に対話をするとして

質問：南鳥島はどこにある？

システム：南鳥島は日本最東端にあります。覚えておきましょう。
（受験生向け対話システム）

新聞記事
「南鳥島沖に球状レアメタル
海洋機構チーム」
日本経済新聞 2016/8/26

質問：誰がハイブリッド車を製造する

システム：自動車会社に影響があるかもしれません。
（ビジネスマン向け対話システム？）

質問：レアメタルは何に使う？

質問：釉薬を何に使う？

システム：新しい薩摩焼とかが現れるかもしれません。
（高齢者の趣味対応用対話システム？）

補足：NICTで開発、公開している システム群

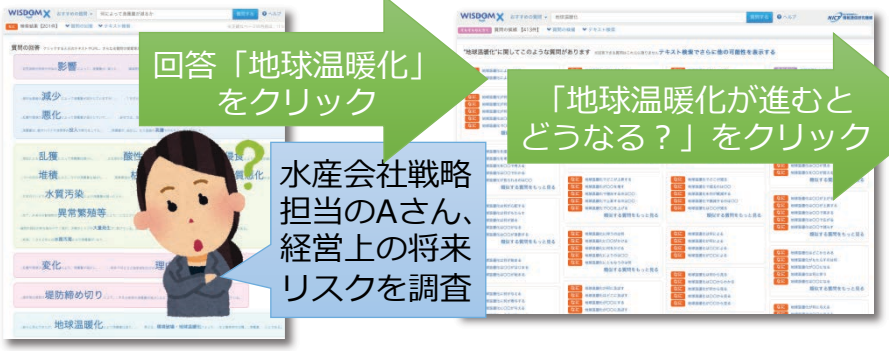
大規模Web情報分析システム WISDOM X

- 自然言語処理技術を用いてユーザーの質問に対してWeb40億ページの情報を基に様々な回答を整理して表示
- また、その回答に対して、システムがさらに質問を追加提案することができるため、 ユーザーがその質問と回答をたどることによって、新たな「仮説」を立てることも可能
- 一般公開中

ステップ１
質問「何によって漁獲量が減るか」をWISDOM Xのホームページで入力 → 回答「地球温暖化」など

ステップ２
「地球温暖化」に関して回答可能な質問が列挙される

ステップ３
約450件の回答：海水温が上がる、台風が巨大化する、プランクトンが減る…



この仮説は論文誌 Nature Climate Change で部分的に事実として報告

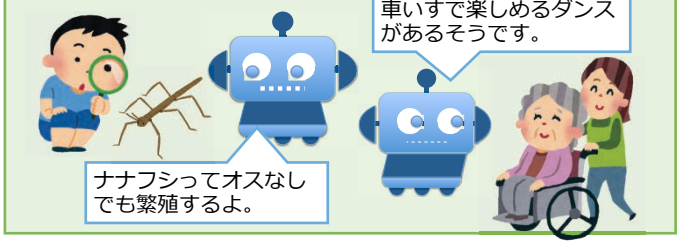
仮説「地球温暖化問題が食中毒を増加させる」
（地球温暖化 → 海水温上昇 → 大腸菌増殖 → 食中毒増加）

食中毒が起きれば水産会社ブランドイメージに打撃…

海水の大腸菌の影響を受けない「陸上養殖」のベンチャーを買収

WISDOM Xの将来像

万能対話ロボット(教育、高齢者)

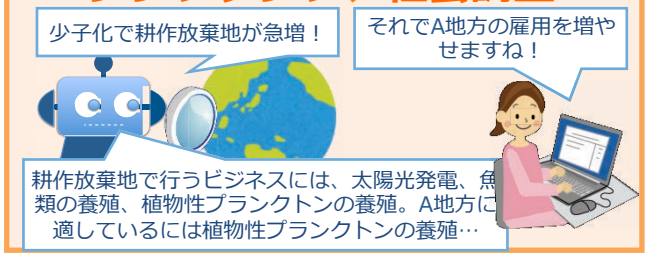


対話技術に応用することで飛躍的に対話機能が向上。ロボットの話に感化されて、ノーベル賞の受賞も夢ではないかも？

民間企業のイノベーション支援

民間企業やシンクタンクが活用することで、専門家でなくても、あらゆる技術、出来事、施策の膨大な組み合わせを、人間には実行不可能な規模でシミュレーション可能となり、この技術をきっかけにして将来有望な様々なアイディアが生まれる

シンクタンク、社会調査



対災害SNS情報分析システムDISAANA（ディサーナ）

- SNS（ツイッター）上の災害関連情報をリアルタイムに深く分析・整理して、状況把握・判断を支援し、救援、避難の支援を行う質問応答システム
- ツイートしてから5秒で分析結果を提供可能
- 熊本地震の際には、ツイッター社から人道支援として1ヶ月ほど100%のツイートを提供いただき、その分析結果を提供（平時は10%サンプル）
- WISDOM X, D-SUMMとあわせて民間企業へのライセンスも締結

住民、救援団体からの質問（例：「熊本県で何が不足していますか」）に瞬時に回答

２０１５年４月より一般公開中

熊本地震の際、首相官邸で活用
- 指定避難所以外のニーズ把握
- 日々変化する要望の把握

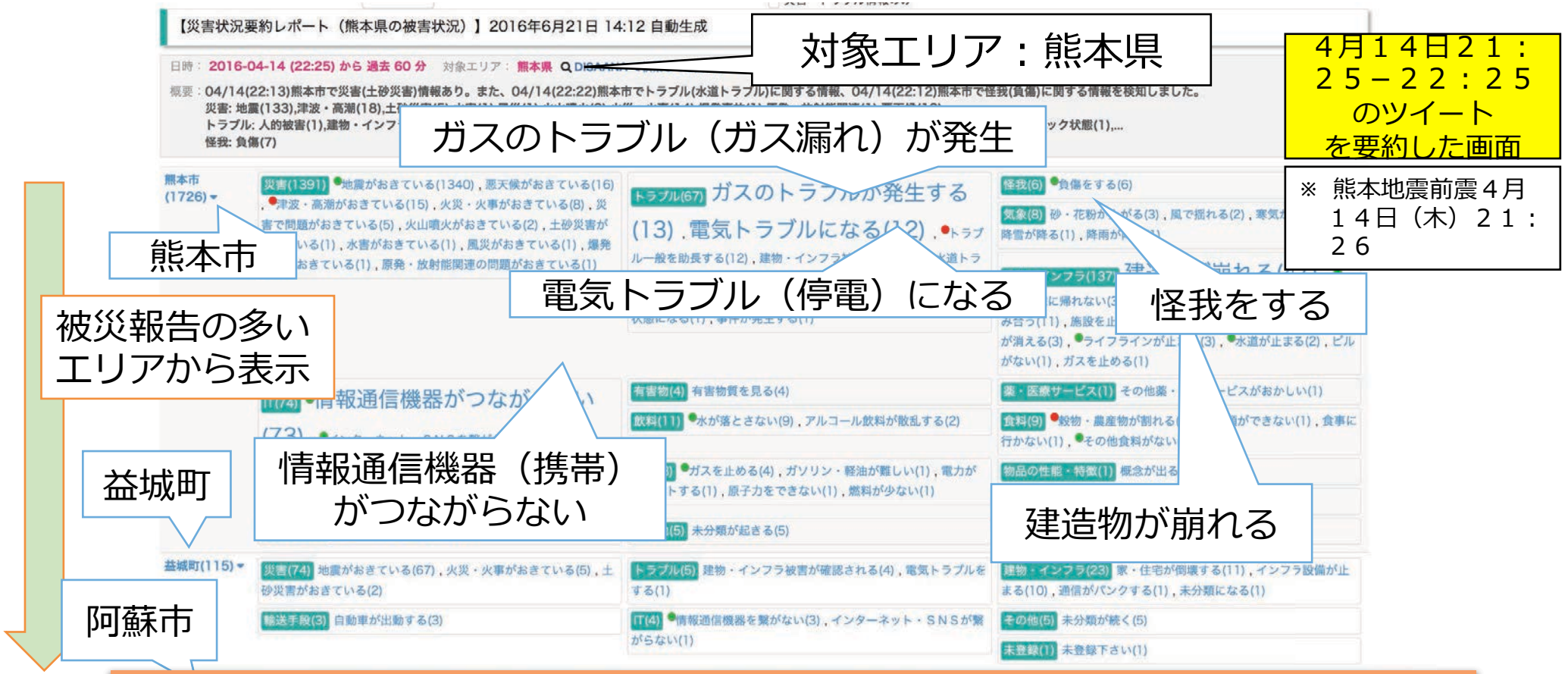
↓

熊本県へ指示

平成２８年５月１１日読売新聞夕刊一面等、報道多数

災害状況要約システムD-SUMM（ディーサム）

- 被災報告の自動抽出結果を整理して要約。リアルタイムで被災状況を把握可能。
- 熊本地震においては、発災後わずか１時間以内には、建物倒壊、負傷者発生、停電、ガス漏れ、信号故障等の被災概況が把握できていた。
- 現地機関からの報告やマスコミ報道を待つまでもなく、初動対応に活用可能。



2016年10月に一般公開開始。開発は内閣府SIPの支援で実施