資料15－6

言語資源データの重要性と戦略的整備

平成29年3月23日
国立研究開発法人 情報通信研究機構
ユニバーサルコミュニケーション研究所長
先進的音声翻訳研究開発推進センター長
木俵　豊

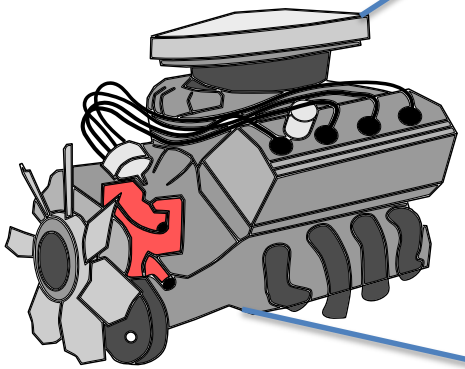
本日のご説明内容

1. NICTにおける人工知能技術の研究開発
2. 人工知能技術の核となる言語処理におけるデータ
3. 音声翻訳に必要なデータ
4. 質問応答技術・対話技術（チャットボット）に必要なデータ
5. 学習データ構築及び活用の課題
6. 人工知能技術戦略会議での議論

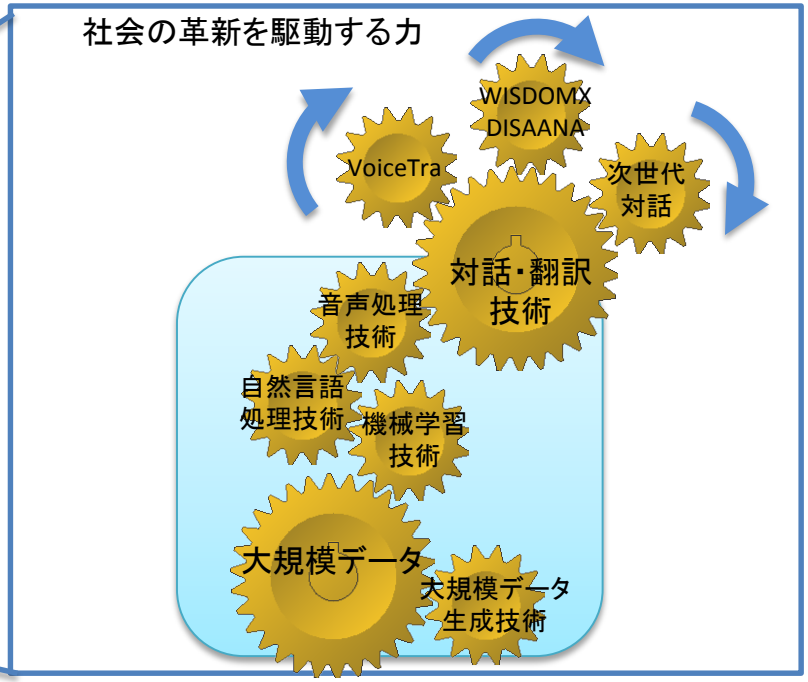
NICTにおける人工知能技術の研究開発

NICTにおける人工知能技術の研究開発

- 人工知能による第4次産業革命を実現するために
  - 実験室の技術に留まらない、社会で活用できる言語・音声処理技術の研究開発を行っている。
  - ディープラーニング技術をはじめとする機械学習技術によって、これまで実現出来なかった高度な技術の研究開発をしている。
  - しかしながら、ディープラーニング技術を活用するためには、質の良い大規模データが必要となり、さらにはそれに対応した自然言語処理・音声処理技術・計算機基盤技術が必要となる。



人工知能技術は社会を駆動する新たな知的エンジンとして期待されているが・・。



燃料（予算）をいくらつぎ込んだとしても、質のよい大規模データ（知識源）を収集し、駆動するためのデータ生成・処理技術を開発して、実用システムへと組み込むギアをかみ合わせなければ、社会を革新する技術になり得ない。

30年の研究開発を経て
翻訳技術が実験室から社会へ

パラダイム シフト

ルールベース
(If-Then-Else)
特定話者
文節発声・定型文
静かな室内
会議室予約

コーパスベース
(統計モデル＋機械学習)
不特定話者
丁寧な自発発話
展示会会場でもOK
(マイクの近くで話せば)
生活会話

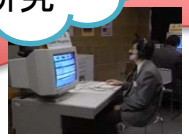
言葉の壁がない世界

フィールド

実験室

音声翻訳の研究開始

長い基礎
研究



1986

2000

2008

2009

2010

2014

2020

社会還元加速プロジェクト

全国5観光地域での大規模実証実験

世界初のN/W型音声翻訳スマホアプリVoiceTra

大量の実利用ログで精度向上

グローバルコミュニケーション計画開始

オリンピック・パラリンピック競技会

テキスト情報分析技術の研究開発

膨大なWebページやSNSの意味を多角的に分析する人工知能技術を10年以上前から着実に研究開発を進めている。

2006 2010 2011 2012 2013 2014 2015 2016

Web6億ページを対象に分析

それぞれのページがどのような意見を述べているか分析

WISDOM (2010)

意味的な関係を多角的に分析

一休 (2010)

Web40億ページやSNSを対象に分析

未来予測や様々な物事の理由等、多様な知識を発見・提示

WISDOM X (2015)

http://wisdom-nict.jp/

災害の被災状況を瞬時に提示

DISAANA (2015)

http://disaana.jp/

D-SUMM(2016)

人工知能技術の核となる言語処理におけるデータ

言語処理とデータ

- 翻訳や対話処理等の人工知能技術には、人手によって作成された大量の良質なデータが必要となる。

  機械翻訳
  =対訳データ（コーパス）＊機械翻訳向け機械学習

  質問応答・対話処理（チャットボット）
  =言語に関する学習データ＊言語処理向け機械学習

- 将来的には、これらを融合させることで、意味・文脈を理解した機械翻訳技術や多言語の意味解析技術が実現する。

人工知能技術におけるデータのタイプ

- 2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。
    - 機械学習等でそれだけで有用な場合は言語・音声処理においては殆どない。

      例：未加工のWebテキスト、科学技術論文、特許明細書、音声等
  - 機械学習の学習時に利用する「学習データ」
    - 作成には人手が必要となるため、多くのコストと時間がかかる。

      （A）SEや作業者が、生データ中のノイズの除去、フォーマットの統一等を処理する必要がある。

      例：音声と読みの対（音声コーパス）、文単位の原文と訳文の対になるデータ（対訳コーパス）等

      （B）人間のアノテータが学習のための教師データとなるラベルを付与する必要がある。

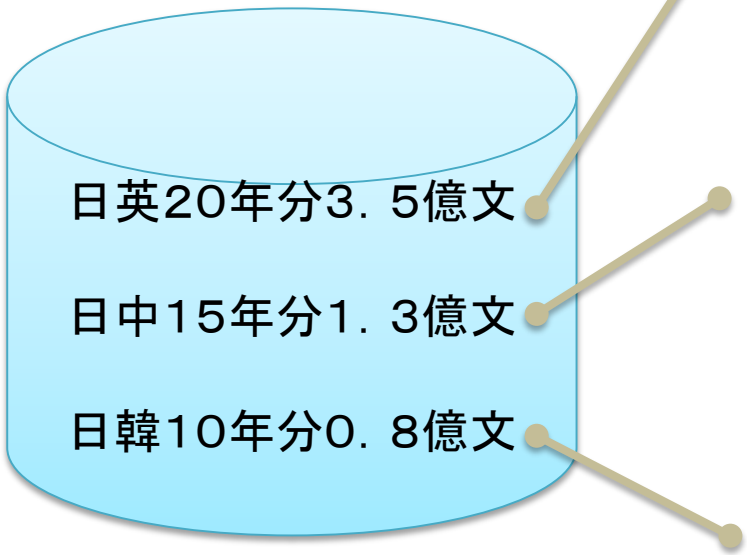
      例：質問と（正解か不正解かのラベル付き）回答候補の対になるデータ等

音声翻訳に必要なデータ

統計翻訳の仕組み

- 翻訳データから各文の確率を計算することで翻訳文を生成する。
- 翻訳データが多ければ多いほど、確率計算の精度が上がる（翻訳精度が上がる）。

対訳コーパス例　特許対訳



【日本語】受容体タンパクを濾過し、洗浄し、次いでフィルターを計数し、[3H]スピペロン特異的結合を決定した。

【英語】Receptor proteins were filtered and washed, the filters were then counted to determine [3 H]Spiperone specifically bound.

【日本語】試験は、0．5mの高さからコンクリート板上に3回落下させて、バンプ部分の断線不良の状態を調べた。

【中国語】在测试过程中，将声表面波设备从0.5米的高度三次落在坚固表面上，并检测凸缘部分的断线情况。

【日本語】従来のベストフォーカス位置を計測するためのフォーカスモニタとして、たとえばIBM社のBrunnerにより開発され、米 Benchmark Technology社より販売されている位相シフトフォーカスモニタがある。

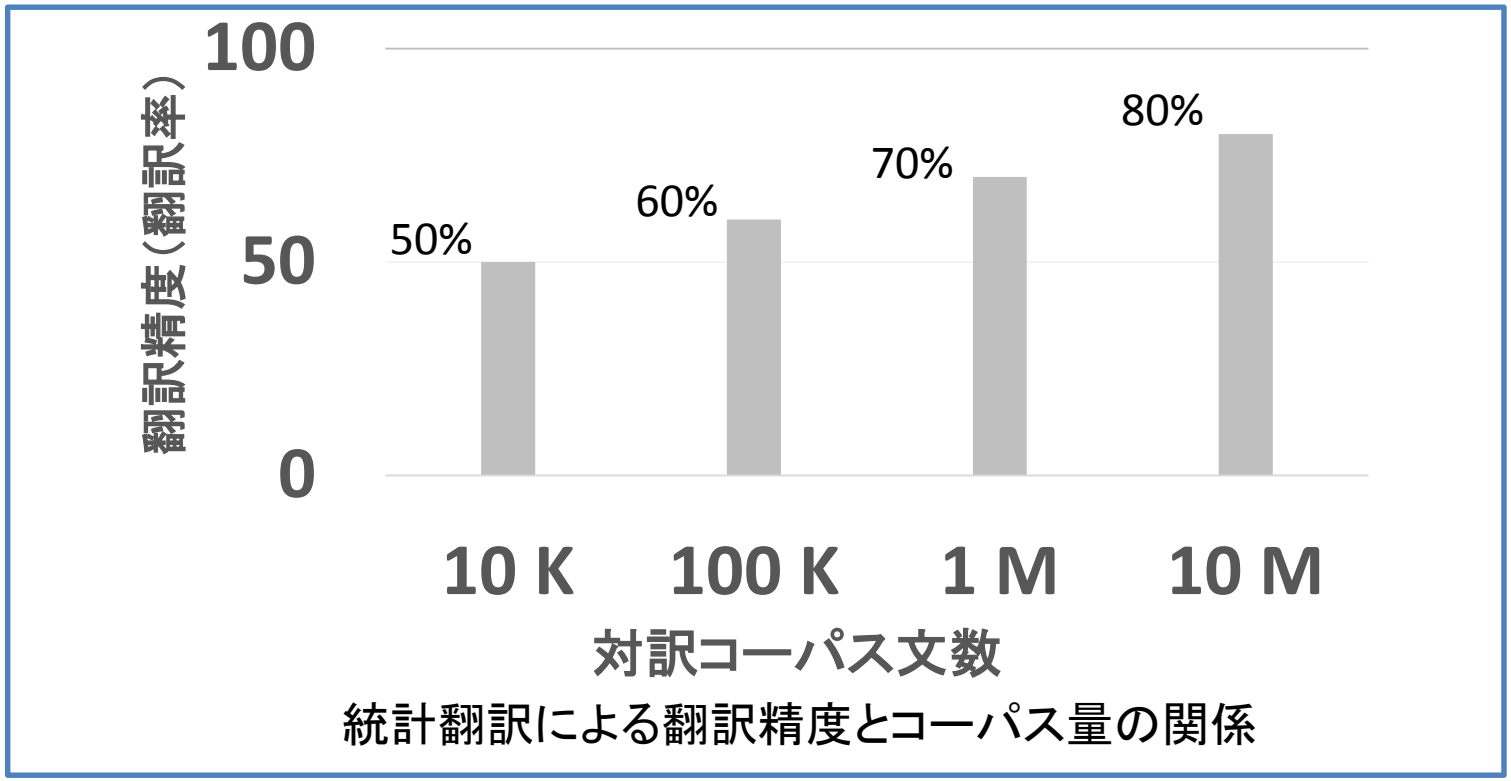
【韓国語】종래의 최상 포커스 위치를 계측하기 위한 포커스 모니터로서, 예를 들어 IBM사의 Brunner에 의해 개발되고, 미국 Benchmark Technology사로 의해 판매되고 있는 위상 시프트 포커스 모니터가 있다.

※　NICTの特許コーパスは、日本特許庁と協力し、ファミリー特許から構築

データ量は翻訳精度に直結する

- 各分野の対訳コーパスを100万文単位で収集しなければ、実用的な翻訳精度(翻訳率※ 80%以上)を達成できない（特許は専門用語が多く億文単位必要）。

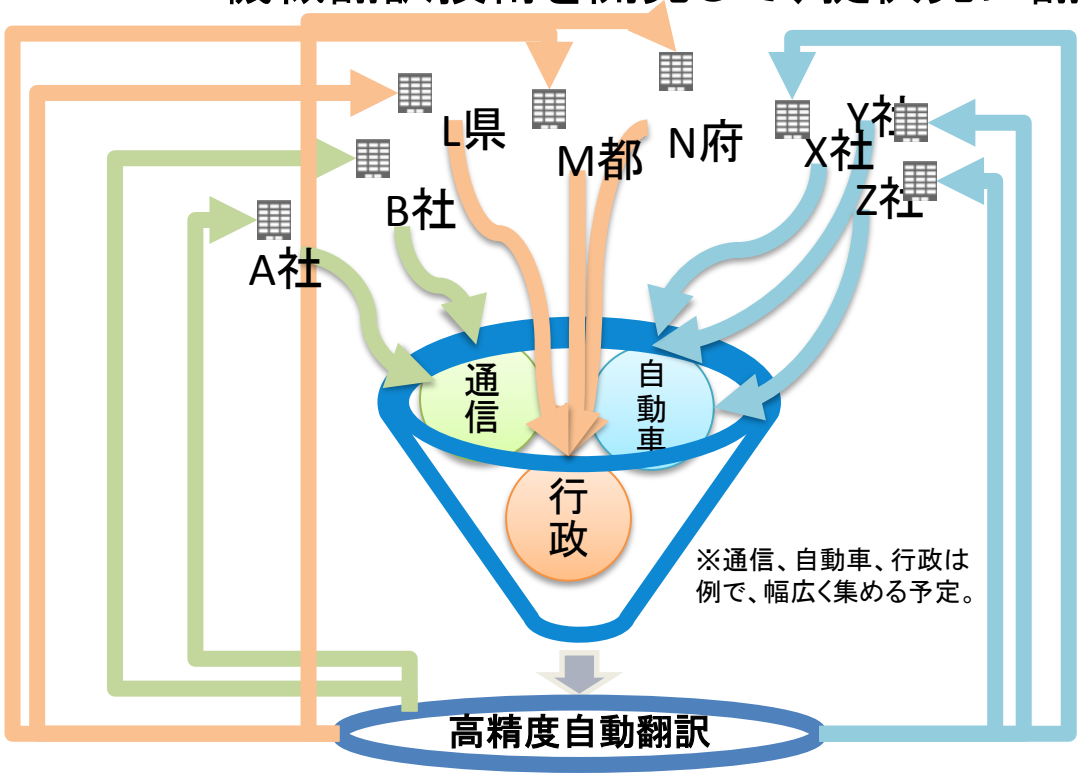
※　翻訳率とは、翻訳された結果の意味が分かる割合



統計翻訳による翻訳精度とコーパス量の関係

翻訳バンクプロジェクト

- 翻訳データ収集と高精度自動翻訳技術の提供
  - 高精度な多言語翻訳技術の開発をめざして、企業に眠る翻訳データを集積する。
  - 収集したデータを用いて、各分野の文書を高精度に翻訳できる機械翻訳技術を開発して、提供先に翻訳サービスとして提供する。



① 過去の蓄積の利活用	② 新規翻訳時に同時収集
組織内翻訳部門や翻訳会社に格納されているWORDファイルを収集	翻訳支援ツールを介して翻訳と同時に提供

質問応答技術・対話技術（チャットボット）に必要なデータ

WISDOM Xの一部(「なぜ型」質問応答)

Webデータを用いた
質問応答システム

Webテキスト

質問応答のための
の学習データ

機械学習器

モデル

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:不正解:地中海の海も空も印象派の絵そのままに青く、私はすっかりはまってしまった。

- 一般的にはこのようなラベルを付与したデータの作成のコストが高い。

- NICTのWISDOM Xでは数万件からなるデータを作成している。
- 複数人の作業者が数ヶ月必要となっている。

回答

質問

ユーザー

次世代対話技術の研究開発
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…

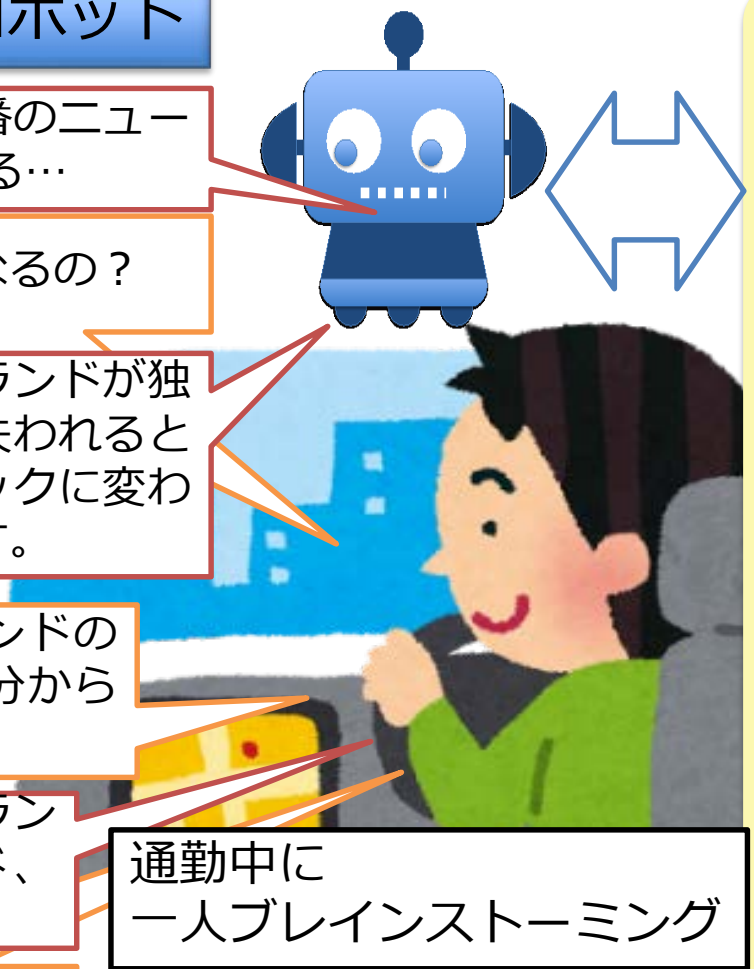
ドライバー：それでどうなるの？

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。

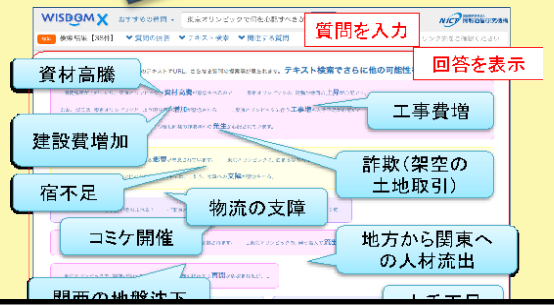


通勤中に
一人ブレインストーミング

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

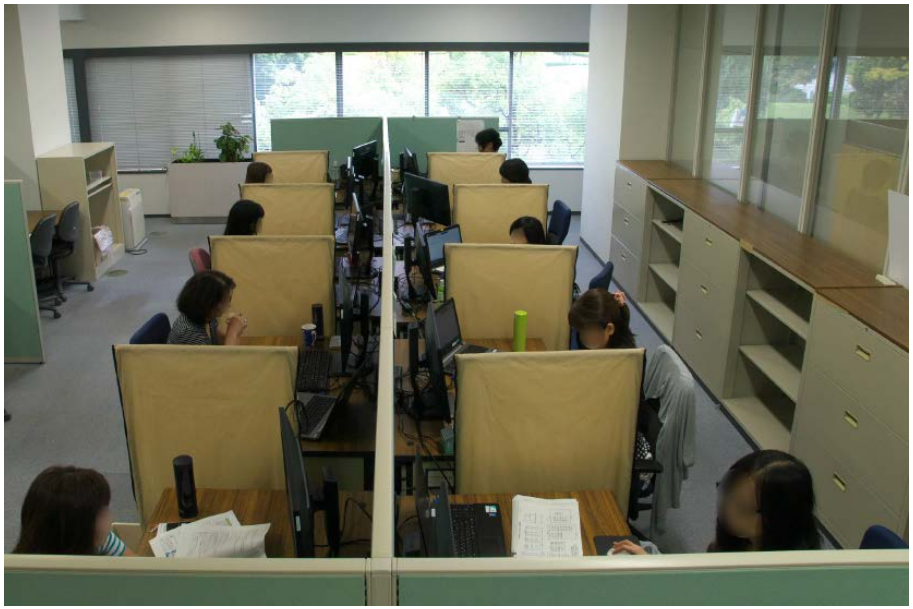
学習データ構築及び活用の課題

学習データ作成は高コスト (1/2)

- 当然ながら、データの内容が理解できる作業者が多数必要となる。
  - 例：医療に関するデータ作成は医療に関する知識が必要不可欠である。
- 作業者用マニュアルが重要である。
  - マニュアルが不備だと学習データが不整合、矛盾を含み、学習がうまくいかず目的が達成できない。
  - 1. 作業の試行、2. データの品質チェック、3. マニュアルの更新、という三段階からなるループを複数回回す必要がある。
  - 複数作業者に同一データで作業してもらい、作業者間での「揺れ」「不整合」をチェックすることも必要である。
- 学習データ作成の指導・マネージメント、学習アルゴリズムの開発の両方ができる研究者は日本では極めて稀である。

学習データ作成は高コスト (2/2)

- NICTでは学習データ作成の指導・マネージメントを専門とする研究者（言語学の学位を保有）を雇用・養成し、機械学習の開発者と共同で作業している。
  - 養成した研究者は企業等から引く手あまたな状況である。
- 作業者としては、20名から50名ほどが研究所に常駐している。
- こうした体制がないと、多数の機械学習の組み合わせが必要な実用レベルのシステムの開発は困難である。



データの有効活用の障害

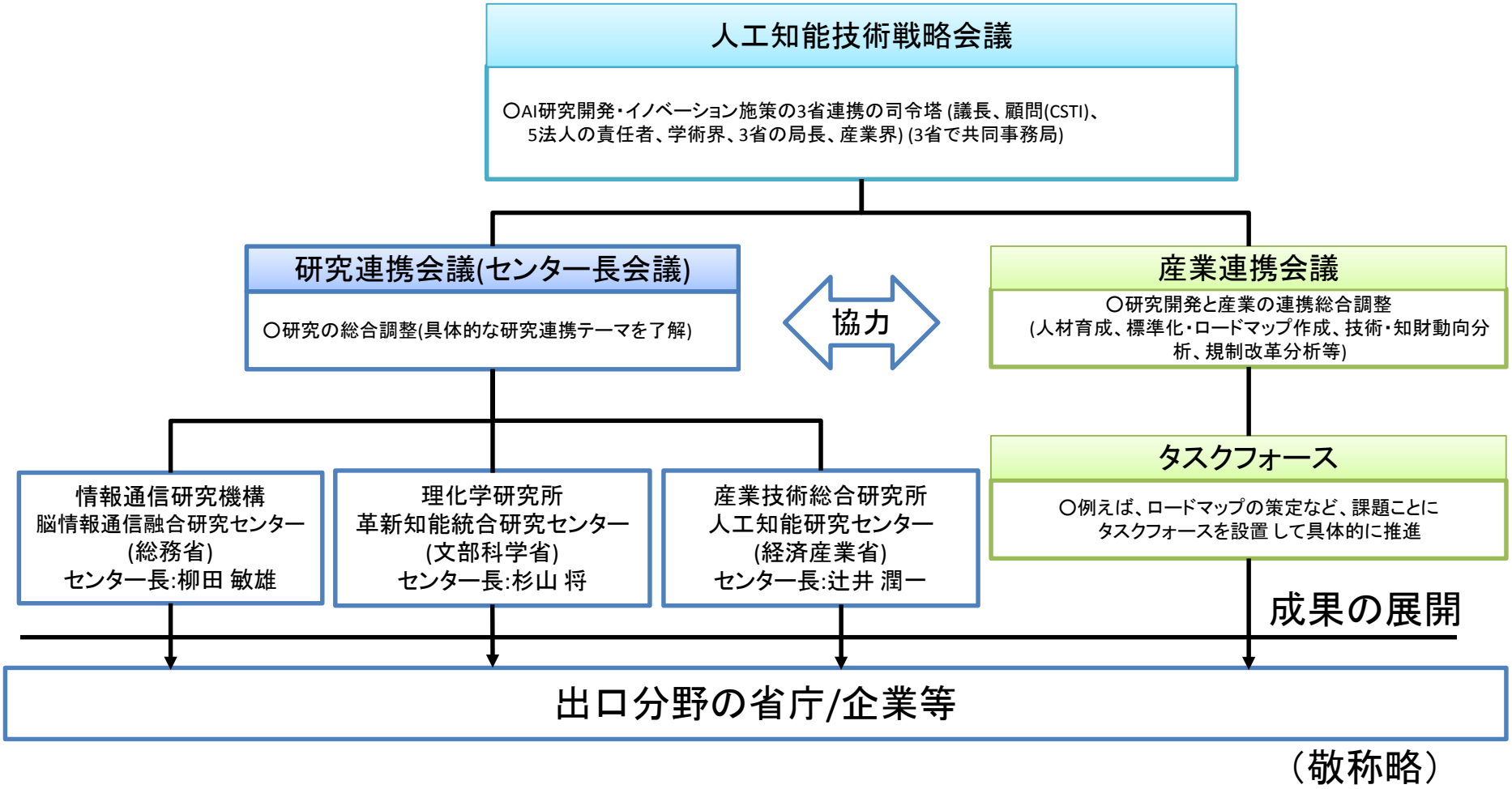
- 障害１：著作権法の規定によりWeb由来の学習データを他の研究組織等と共有することができないと認識している。
  - 共有可能とする見解も存在するが、印象として共有不可とする専門家の方が多数である。
  - 現在は学習済みのモデルをライセンスしている。
  - 欧米では、同一学習データを複数組織が共有し、それぞれが多数の学習アルゴリズムを試し、高精度なものを特定することが常識的に行われている。

- 障害２：　学習データに限らず、誰でも閲覧できるWebページを収集しても、組織をまたがっての共有は不可能である。
  - NICTでは１００億件以上のWebページを収集している。しかしながら、この収集したデータを新規サービスを計画しているベンチャー企業等に提供することはできない。
  - Webページの大規模な収集は大量の時間、資金を要し、新規ベンチャー企業等では実現が困難であり、新規サービスの実用化を阻害する要因となっている。

人工知能技術戦略会議での議論

人工知能技術戦略会議

◎ 議長　安西 祐一郎(独立行政法人日本学術振興会 理事長)



（敬称略）

産学官が有するデータ及びツール群の環境整備に関する方針について

出典：未来投資会議　構造改革徹底推進会合（平成29年2月23日）人工知能技術戦略会議提出資料

方針案	ニーズ、問題意識	期待される具体的行動のイメージ（例）
1. 重点取り組み分野のデータ整備強化	・産業化ロードマップにおける重点分野（生産性等）等については、新たなデータ取得による整備を行うべきではないか。	①府省庁連携研究におけるデータ整備の強化 ②標準画像等、AI性能評価のためのデータ整備 ③研究及び人材育成向けの学習用データセット整備
2. データ整備・提供を担う機関の強化	・研究自体よりも地味で継続的なデータ整備について、支援の強化が必要ではないか。	①データ整備提供を担う専門機関の強化等
3. データ取得やツールの検証を加速する模擬環境、実証環境の整備	・個人情報等データ取得の阻害要因がある中、データ取得できる特別の環境の確保が必要ではないか。	①工場や店舗、病院等、実物を模した模擬環境の整備 ②AI製品・サービスの実証に利用できる現場の確保
4. 産学連携によるデータ、ツールの集積の好循環	・海外AIクラウドに依存せず、国内で好循環する枠組みが必要ではないか。	①産学官における、データ解析力の提供とデータ提供を好循環させるAIクラウドの提供 ②AIクラウド提供を通じた、オープンツール開発支援
5. データセット整備を加速する技術開発、制度整備	・人海戦術となっているクレンジング、タグ付け等データセット化について、効率化が必要ではないか。	①AIで自動的に関連付けを行う技術、匿名化・秘密計算・秘密検索技術等、データ整備加速技術の開発 ②データの自動的登録（蓄積）を促す制度整備
6. 国プロから生じるデータのオープンデータ化	国が率先してオープンデータを提供すべきではないか。	①国プロで取得したデータの管理と提供 ②データ取得自体を目的とする国のプロジェクトを整備
7. データ及びツール群にかかるリソースの一覧化	・所在情報等、ユーザーが利用しやすい環境を整備すべきではないか。	①オープンデータ・オープンツール、計算機資源、データ取得環境（実証・模擬環境等）の一覧情報の提供・活用促進（3研究機関を含む）
8. 民間等保有データの共有、横断的活用等	・データ流通を巡る動きを、AI側としても積極的に対応すべき	①情報銀行、データ取引ルール等、民間主体の枠組みの活用 ②API公開等、データ連携・互換性の向上（IT本部等）

まとめ

- 現在の人工知能技術を効果的に活用するためには、大規模な学習データが必要となる。
- 特に、音声処理、言語処理技術に基づく人工知能技術の学習データは、人によるデータクレンジングやアノテーションが必要であり、高コストな作業が必要不可欠となる。
- 個人情報保護や著作権保護の観点から学習データの共有化等についても考慮が必要である。
- 我が国における人工知能技術の推進のためには、データの戦略的整備が必要不可欠である。