ビッグデータ、人工知能と知識の有効な活用

Big Data, Artificial Intelligence and Exploitation of Knowledge

東京大学名誉教授
国立研究開発法人産業技術総合研究所　人工知能研究センター　研究センター長
辻井　潤一

人工知能研究センター　研究センター長、英国マンチェスター大学客員教授、国際計算言語委員会（ICCL）委員長、AAMT／Japio 特許翻訳研究会委員長

j-tsujii@aist.go.jp

1. はじめに

本年５月に国立研究開発法人産業技術総合研究所の中に人工知能研究センター（AIRC − Artificial Intelligence Research Center）が設立され、Microsoft 研究所から移籍して、このセンターのセンター長を務めています。

英国の BBC がこの９月に１週間にわたって人工知能の特別番組を組むなど、世界的にも、人工知能はブームとなっています。特に、日本では、経済産業省の傘下である我々のセンター以外にも、文部科学省や総務省も、人工知能研究を重点的な研究分野として、センター設立や研究推進を宣言するなど、過熱気味でもあります。

本稿では、AIRC の設立理念や研究方向を紹介し、現在の人工知能ブームを単なる一過性のものではない、息の長い研究や開発に結びつけていくための方策を議論したいと思います。その中で、特許情報をはじめとする科学技術テキストの処理やそこに埋め込まれた知識を有効活用するための研究についても議論したいと思います。

2. 人工知能の２つの流れ

前世紀の中葉、人工知能の研究は、まず、知能とは何かを定義することから始まりました。人工知能という言葉が一般化する以前には、機械で実現するという知能そのものをまず定義する必要がありました。しかし、知能を定義することはそれほど簡単なことではありません。分析的な定義はできず、結局は、人間を知能的な存在のモデルとすることになりました。人間が行う様々なことの中でも、我々が知的と感じることを実行できる人工物ができれば、それを人工知能のひとつの実現形と考える、という間接的な定義をとったわけです。

人間と言葉による会話ができて、その会話の流れが人間と区別できないほどこなれたものになれば、人工知能が実現できたと考えようというチューリングテストは、その典型的なものです。自然な会話を行うためには幅の広い知的な能力が求められると思われますが、もっと特殊なタスクだけに注目し、人間が行うと知的であると考えられるような特定のタスクが実行できるプログラムを人工知能の実現形としようとする考え方もあります。チェスや将棋をするプログラム、数学の定理を証明するプログラムといったものが、その典型でしょう。

徐々にタスクの幅を広く取るようになりますが、この考え方は、現在でも強く、たとえば、東京大学の入学試験に合格するプログラムを作ろうという東ロボのプロジェクトも、この分類に入れていいでしょう（図 1）。

人間を知能のモデルとして、人間に迫ろうという人工知能に対して、ビッグデータの技術の延長に、もう一つの、比較的新しい人工知能の流れをみることができます。膨大なデータを収集し、その背後にある規則性をとらえることは、人間は得意ではありません。商品の購買データや気象データから気象状況と特定商品の購買との相互関係を把握したり、大量のたんぱく質の発現データから、たんぱく質間の相互関係のネットワークを構築したりといったことは、本来、人間にとってはむつかしいタスクです。このような作業は、むしろ、大規模な記憶容量と高速な計算機能を持つコンピュータが得意とするところです。

ビッグデータ解析という研究分野は、大量のデータを収集し、それを統計処理する技術、あるいは、その処理結果を人間にうまく提示する視覚化技術を発展させてきました。このビッグデータの分野では、データサイエンティストと呼ばれる人間が、大量データをデータ分析器（Data　Analytics、統計処理プログラム）を使って分析し、それを視覚化してみることで、対象理解を行うこと前提としています（図2）。

図 1　人工知能

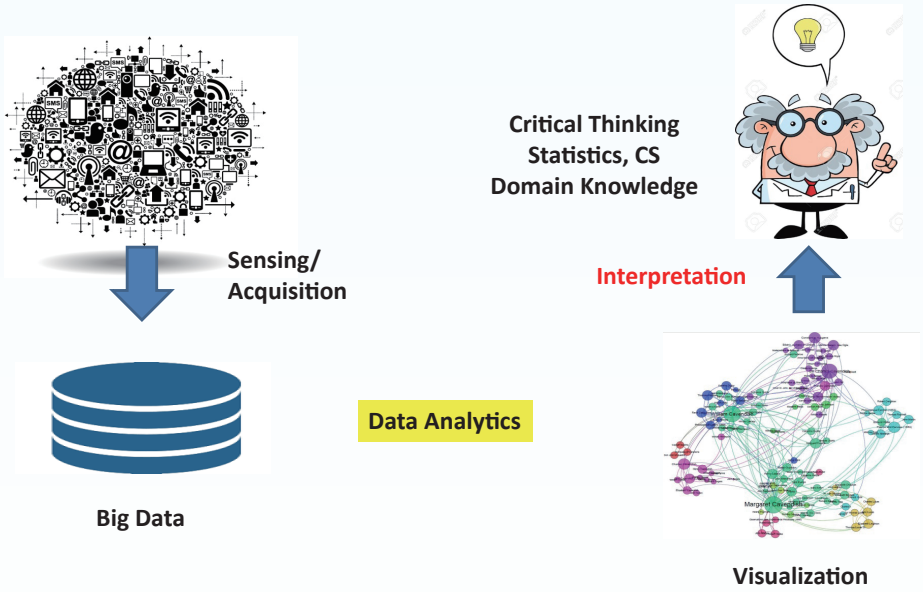

図 2　データサイエンス

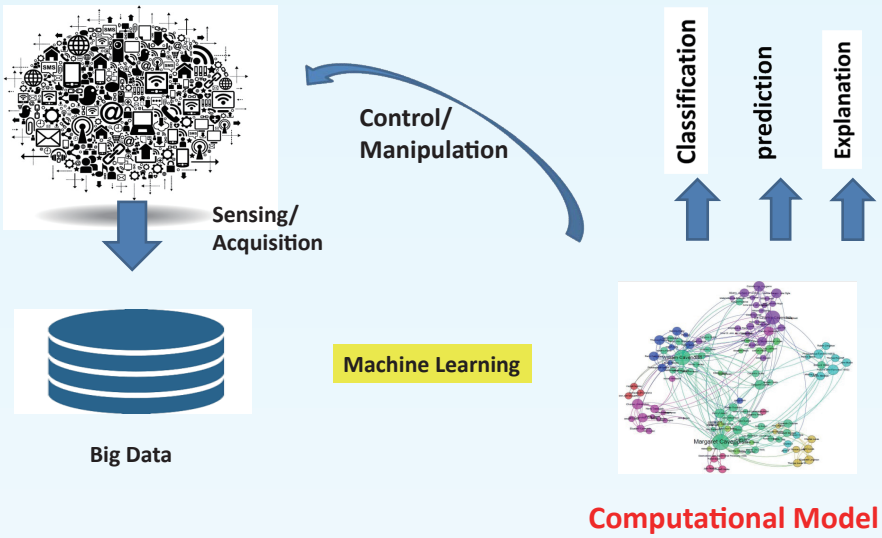

図３　機械学習による人工知能

これに対して、現在ひろく使われるようになってきた機械学習の技術は、大量のデータを使って分類や予測を行う計算モデルを自動構築します。この構築された計算モデルは、新たなデータを分類したり、新たなデータから次に起こることを予測したりすることができます。すなわち、データサイエンティストの仕事であったデータの解釈や理解の作業を計算機に置き換えようというわけで、これが新たな人工知能研究の流れを形成しています。

大勢の患者さんの大量データから、病疾患に関する計算モデルを構築し、新たな患者さんのデータを分類したり（すなわち、患者さんの病疾患の診断をしたり）、その患者さんがたどるであろう病疾患の過程を予測したりするプログラムが、機械学習で作られます。このようなプログラムは、医者が行う Diagnosis や Prognosis を行うわけですから、医者の行う知的なタスクを実行できる人工知能プログラムです。また、どのような薬剤を投与すべきかといった対象（患者）の操作、治療行為を行うことも可能になります（図３）。

知的な行為者を、「外界のモデルをもち、外界の変化を予測したり、外界を自分の目的に合うように操作することができる」行為者であると定義すると、大量データから対象の計算モデルを構築し、これにより対象の分類、予測、操作ができる機械学習のプログラムは、人工知能プログラムと呼んでよいでしょう。

このような機械学習に基づく人工知能は、大量データからの計算モデルの構築という、人間が苦手とすることを行うという点で、また、人間とはかなり違ったやり方で分類、予測、操作を行うという点で、人間を超える人工知能、人間知能とは異質な人工知能といってもよいものでしょう。

現在の人工知能ブームは、この２つの流れが統合され始めたことで起こったものだと考えられます。

3. 人間に寄り添う人工知能

図４は、AIRC のロゴです。このロゴは、「人に寄り添うしなやか人工知能」を目指すセンターの基本的な立場を象徴するものです。

現在、人工知能への期待が高まると同時に、人工知能への不信感、それがもたらす社会変動への恐れも顕在化しつつあります。ステファン・ホーキンス博士、ビル・ゲー

図４　AIRC のロゴマーク

ツなど、多くの著名人が人工知能の危険性に言及し、その極端な場合には、人間知能を超える超知能が出現するという警告もあります。

私自身は、このような議論の根拠を深く理解しているものではありませんが、人工知能への過剰な期待をあおる人と同様に、人工知能への極端な警戒を主張する人も、人工知能の能力や可能性を買いかぶり過ぎていると思っています。

人工知能ブームの立役者の一人、深層学習の研究者でフェースブックの人工知能研究所所長であるルカン博士の「現在もっともすぐれている人工知能でも、実際はとんでもない馬鹿だ」（Right now, even the best AI systems are dumb）というのが、私の実感に近いものです。

「人工知能は人間を超える」という議論では、人工知能が人間と同じような種を形成していて、それが人間を超えるという議論をしているかに聞こえます。しかし、現在の人工知能は、特定の知的タスクを実行するために作られた、まったくバラバラなプログラムの集合体です。プロ棋士と対等に戦う将棋のプログラムは、ただそれだけであり、たとえば、東ロボや IBM のワトソンといったプログラムとは全く関係がないものです。人工知能という、動物種に相当するような種が存在するわけではありません。いろいろな全く違ったプログラムを便宜上で人工知能プログラムと総称してしまうために、人工知能という、共通の能力基盤をもった種があり、それが人間という種と比較できる存在であるかのように錯覚しているだけの議論に思われます。

大量データから対象の計算モデルを学習する人工知能が、人間が理解できない、自己完結的なブラックボックスの知能になっていることは事実でしょう。このような自己完結的な知能は、他の知能体（すなわち、人間）に、何らの説明もなく分類や予測の結果だけを示し、その受け入れを強制するもので、他の知能体（すなわち、人間）と協働して挑戦的なタスクを実行していくことはできません。

我々の「人に寄り添うしなやかな」人工知能とは、大量データに基づく「人間を超える」人工知能と、人間の知能をモデルとする「人間に迫る」人工知能の技術を融合させることにより、人間と協働できる、人間に理解でき、人間が協働できる人工知能を実現していこうというものです。

より具体的には、（1）データで考える人工知能を知識で考える人に近づけるデータ知識融合型 AI、（2）しなやかな人の知能を支えている脳の働きをうまく反映した脳型 AI、という 2 つの柱で研究、開発を進めていこうとしています。

4. 日本に向いた組織づくり

現在の人工知能研究は、米国、特にその巨大 IT 企業がリードする形で進展してきました。大量データに基づく現在の機械学習技術が、大量データを集積する巨大 IT 企業を中心にして発展してきたのは当然なことだろうと思います。

インターネットから大量データを収集し、それらに付加価値をつけることをビジネスとする巨大な IT 企業は、ビッグデータ処理の技術や機械学習の技術を発展させる核となっていました。大量データ、大勢の研究者と開発者、大量データの価値化という明確なニーズという 3 つの要素が、単一の企業内に集積していたことが、人工知能技術を急速に発展させる原動力になってきました。

しかしながら、日本やヨーロッパなど、米国外の地域では、この 3 つの要素を集積する巨大 IT 企業は存在しません。人材の流動性が低い日本では、技術的なシーズをもつ研究者も、多くの大学や研究機関に散在し、小規模な研究グループを作っているのが現状です（図5）。

このような散在した研究者は、大量データへのアクセスも不可能です。自らが持つ技術的シーズを現実の問題に適用し、その結果として、次に解くべき技術的課題を定義していく契機を欠き、結果として、都合のよい仮想的なタスクを設定して、研究のための研究に終始しているのが現状です。

この現状を打破して、データ、シーズ、ニーズという 3 つの要素を集中させるためのハブ・センターとして、AIRC が設立されたわけです。

これからの人工知能研究にとって、3 つの要素を内部に取り込む巨大 IT 企業のモデルが最適なモデルになっているわけではない、と我々は考えています。実際、インターネット中のデータだけが大量データではないとい

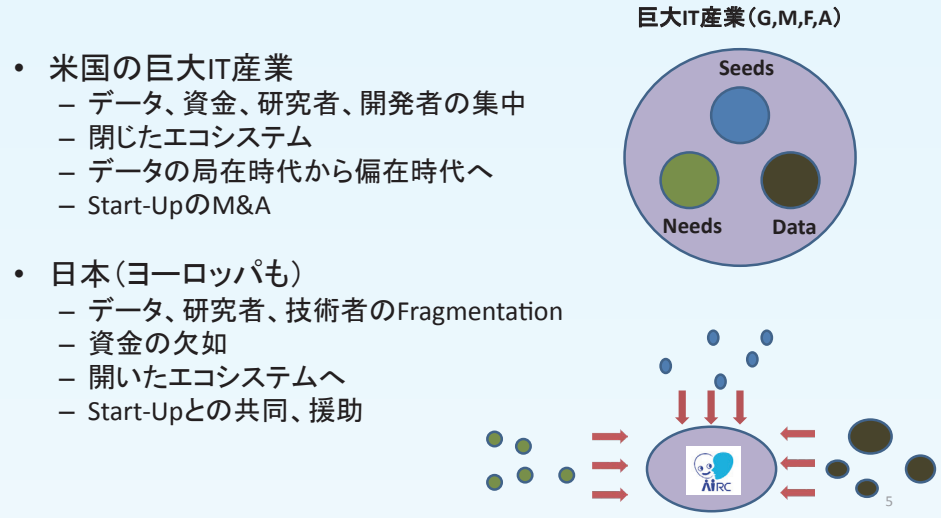

図５　人工知能の技術開発：現状

う時代を迎え、巨大 IT 企業も、例えば、医療データや自動運転のためのデータなど、彼らの本来のビジネスからは逸脱した彼らの内部には集積していないデータへの人工知能技術の適用を試みる時代に入りつつあります。この傾向は、IoT や CPS など、現在のインターネットとは形態の異なるデータ収集が活発化することで、さらに強まっていくでしょう。また、医療データ、健康データ、生命科学や物質科学などのビッグサイエンスからのデータなど、巨大 IT 企業の外にある、別の組織がデータ収集の主体となっている分野でも、人工知能技術の適用が不可欠になってくるでしょう。

今後は、巨大 IT 企業の中にデータ、技術シーズ、応用ニーズが集中するという閉じた研究開発のモデルから、これらの３つ要素がバラバラにあるという前提で、この３つを集中させる研究組織を構築していく必要があろうと考えています。

5. 研究テーマの例

「データと知識とを融合する」といっても、データとか知識をきちんと定義するのはむつかしい。

例えば、「テーブルの上にリンゴがある」写真を考えてみましょう。写真は、計算機の内部では、写真中の各点の色情報として記録されます。色の情報は、赤・青・黄色の３原色それぞれの強度であらわされ、各点にこの強度の情報が記憶されます。この連続量の情報を「テーブル」、「リンゴ」、「A の上に B がある」という言語表現に置き換える操作は、データと知識とを結びつける処理の第一歩でしょう。このためには、同じような色が広がる領域を認識して、それが「テーブル」とか「リンゴ」と呼ばれる、すこし大げさな言い方になるが、「概念」の具体例であると認識できる必要があります。また、この２つの具体例（特定のリンゴとテーブル）が、「上にある」という関係の具体例となっていることがわからなければ、言語表現はできません。

さらに、「テーブルの上にリンゴがある」という言語表現（すなわち、文）は、計算機にとっては、テ、ー、ブ、ルといった文字が並んでいるだけです。この文字列という非構造化情報を、（ON　APPLE　TABLE）といったように、計算機で操作可能な構造化情報に置き換えができれば、このテーブルとリンゴの関係を使った推論も可能になるでしょう。

このように、視覚系からは入ってきた連続量を構造化された概念間の関係としてとらえ、それを使うことで、外界に対する行動（例えば、リンゴをつかんで食べる）を起こすことができるようになります。人は、このような連続量や非構造データを記号的な構造表現に置き換え、その上で思考し、さらにその結果を再び連続量の世

界である外界での行動に移すことができます。視覚から思考、思考から行動への移行が極めてスムーズに、しなやかに行われます。このしなやかな移行が、現在の人工知能プログラムには非常にむつかしいことです。

たとえば、写真に写っているものを猫、テーブル、車といった分類するという課題は、一般画像認識といわれますが、この課題を解くだけでも、100 万枚以上の正解が付いた写真を使ってトレーニングする必要があります。深層学習を使うことで、人間よりも優れた性能を示すシステムが構築されたと喧伝されていますが、そのためには 100 万枚のトレーニング用の正解付きデータを用意する必要があったわけです。しかも、写真に２つ以上のものが移っている場合の認識や、複数のものの間の関係（「上にある」といった関係概念の認識）などを認識する研究は、まだ手についたばかりの研究です。

このように、視覚系、概念系、行動系間の情報の流れをしなやかに行うことは、次世代の人工知能の基盤技術として、AIRC の主要な研究課題となっています。

現在の自動運転技術は、視覚系から直接に行動系に結び付ける、いわば、条件反射に基づく自動運転になっています。これに対して、視覚系の情報から「子供が赤信号に気付かず交差点を渡ろうとしている」といった、世界での出来事を明示的に認識し、その結果を適切な運転という行動に結びつけていくことは、次の段階の基盤的な研究として、AIRC と九州工大、情報学研究所、早稲田大学などとの共同研究のテーマとなっています（図６）。

この視覚系、概念系、行動系のしなやかな連携を、人間の脳は、学習によって比較的速やかに獲得していくように見えます。脳型 AI の研究では、人間の脳が、視覚系と概念系、概念系と行動系とのしなやかな連携を学習で獲得していく過程を計算論的に明らかにし、これを次世代 AI 技術の基盤技術に育てていくことを目指しています。

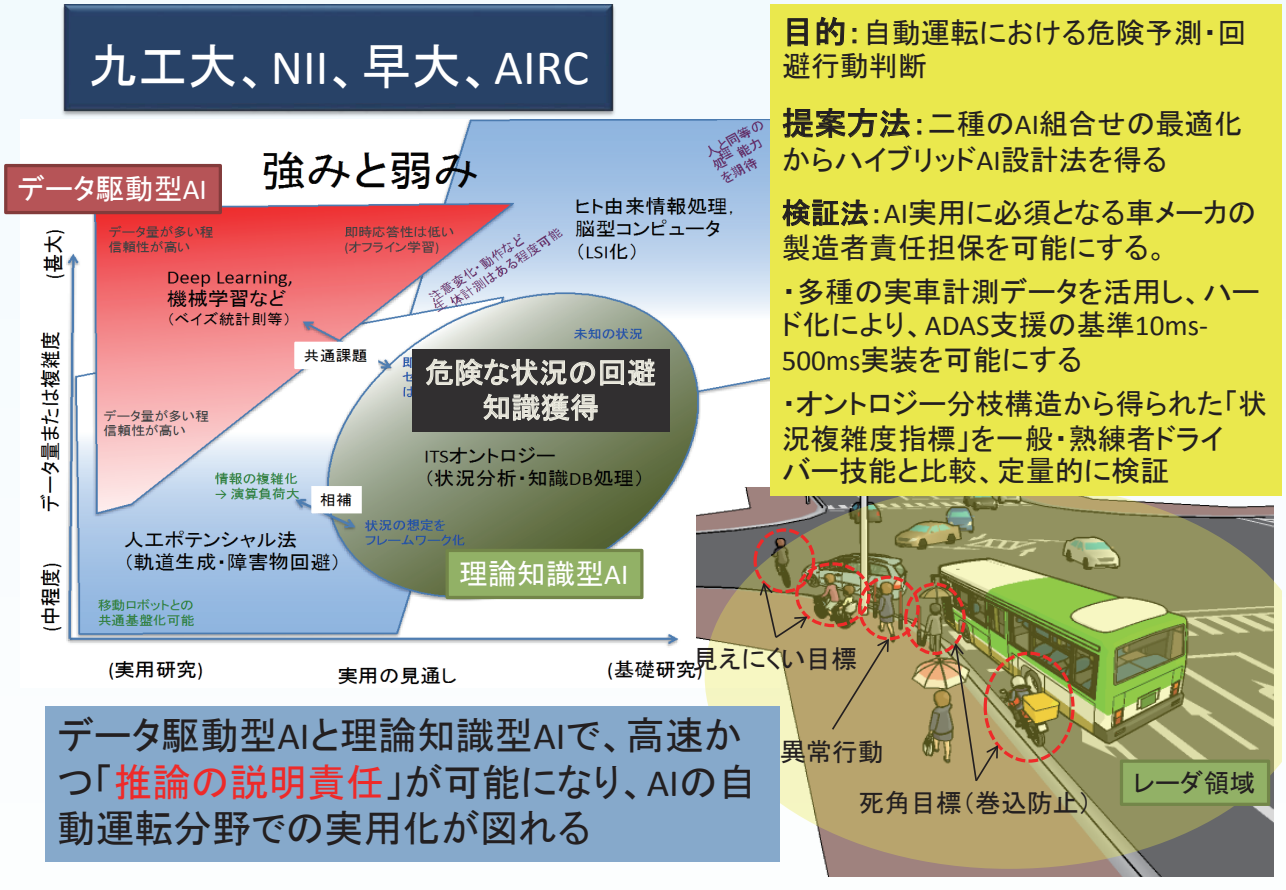

図６　状況理解と自動運転

6. ロボット・サイエンティスト

人工知能研究の大規模科学（Big Sciences）への適用に、ロボット・サイエンティストがあります。このアイデアは、かつての私の同僚である英国マンチェスター大学のロス・キング教授が最初に提唱してもので、その後、ケンブリッジ大学などでも取り上げられています。米国 DARPA の Big　Mechanism でも取り上げられ、このプロジェクトには、私もシカゴ大学、マンチェスター大学のチームに所属して、参加しています。Big Mechanism プロジェクトの主目的は、膨大に出版される生命科学の文献の中の情報を、すでに蓄積されている構造化されたパスウェイの知識に結び付けること、また、マイクロアレーデータによるたんぱく質の発現データをパスウェイに結び付けて解釈することです。非構造化情報である文献、連続量で非構造化データであるマイクロアレーデータという２つの情報を、構造化データであるパスウェイに結び付けることが目的です（図７）。

プロジェクトは、「センサーなどで観察、獲得できるのは、実際に起こっていることのごく一部であり、観察データを理解するためには、その背後にあるメカニズムを理解する必要がある」という前提で、データを背後のメカニズムに結びつけようというものです。生命科学では、この背後にあるメカニズムが、パスウエイという構造化された知識として表現されます。実際には、この背後にあるメカニズムを究明することが生命科学の究極の目的ですから、データに解釈を与えるという作業は、構造化されている未完結のパスウエイをもとにして背後にあるメカニズムに関する仮説を作っていく過程になります。さらには、文献に断片的に発表されているたんぱく質の相互関係をより大きな相互関係のネットワーク（すなわち、パスウエイ）にまとめ上げること自体が、このメカニズムを究明する重要なプロセスとなります。生命科学者によってパスウエイとして構造化され、データベース化されているのは、既発表の文献に現れているたんぱく質相互関係の１０%以下でしかないと推察されています。

マンチェスター大学のテキストマイニングセンターでは、発表論文からのたんぱく質相互作用の情報を取り出して構造化すること、これを使って既存の構造化された

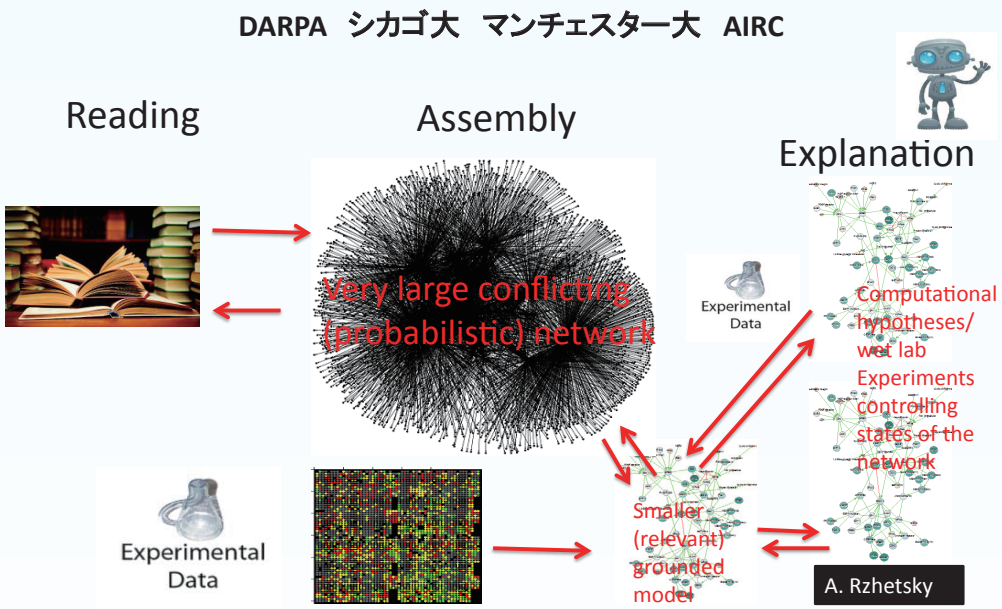

図７　Big Mechanism: ロボット・サイエンティスト

パスウエイを豊富化する研究を推進しています。テキストという非構造化情報からたんぱく質相互作用の情報を明示的に取り出し、それを既存の知識と対応付けるテキスト理解のプロセスの研究です。

これに対して、シカゴ大学のバイオインフォマティクスのグループは、観測データを説明するパスウェイの仮説を既存の大規模パスウエイから切り取った、説明のための小規模なパスウェイを構築すること、また、この仮説を検証する実験手順の自動構築を行うことに重点を当てた研究を行っています。

プロジェクトは、（1）観察データを説明する背後のメカニズムに関する仮説を自動構築すること、（2）仮説を構築するための要素的な知識を既発表の論文から見つけ出すこと、（3）仮説を検証する実験をアレンジすること、という生命科学者が行っている知的作業を人工知能に置き換えることを目指しているといってよいでしょう。

実際には、この３つの過程を人工知能プログラムと人間とが共同して行うことで、科学の進展を加速するというのが現実的なシナリオであり、人に寄り添う人工知能、すなわち、科学者のアシスタントとしての人工知能として、AIRC としても積極的に取り組んでいくテーマとなっています。

7. 特許と技術動向調査

ロボット・サイエンティストにおける論文からのたんぱく質相互作用の情報抽出に見られるように、大量の文献データの内容を処理し、それを既存の知識に結び付ける技術は、かなりの程度成熟してきています。

このことは、現在の研究が、文献に分類コードを振るという文献を単位とする処理から、文単位やパラグラフ単位という、粒度の細かな処理へと向かっていることの一例です。一方では、出現した単語や専門用語でテキストを特徴づけ検索するフルテキスト処理という用語単位の最も細かな粒度の処理に、意味処理を連動させることで、同義語や関連語群に広げる方向の研究も盛んになりつつあります。

用語単位の処理と文献全体を単位とする処理の中間に、文やパラグラフを単位にその意味を構造化する方向が顕著になってきています。このことは、私がこれまで強調してきた特許文献と用語オントロジーとの関連付け、特許文献からの情報抽出による技術動向調査のツール作成といった研究方向と一致しています。AIRC でも、特許文献は、非構造化情報の典型であるテキストを構造化された知識に結び付ける研究を行う最適な分野の一つとして、研究を進めていこうと考えています。

8. おわりに

今回の人工知能ブームは、３次のブームになると言われています。これまでの２回のブームが人間と同じような知的なタスクを行うという、理想主義で抽象的なブームであったのに対して、現在のブームには、ビッグデータを価値化したいという、現実的なニーズとビッグデータ処理で成熟してきた技術の基盤に基づいたブームになっています。70 年の歴史をもつ人工知能の研究が、ようやく現実の問題を解いていくための技術になってきたということでしょう。同様に、私が関わってきたテキストの意味理解の研究も、大量のテキストと大規模な構造化された知識とが使えるようになり、大きな飛躍を遂げる時期を迎えたと考えています。