CRDS-FY2019-WR-08

科学技術未来戦略ワークショップ報告書

深層学習と知識・記号推論の融合による AI基盤技術の発展

国立研究開発法人科学技術振興機構 研究開発戦略センター
Center for Research and Development Strategy, Japan Science and Technology Agency

エグゼクティブサマリー

国立研究開発法人 科学技術振興機構（JST）研究開発戦略センター（CRDS）は、国の科学技術イノベーション政策に関する調査・分析・提案を中立的な立場に立って行う組織として、今後わが国が取り組むべき重要な研究領域に関する提言を行ってきている。また、その提言に向けた前段階として、広く技術分野の状況を俯瞰し、注目すべき研究動向の抽出を行ってきている。これら俯瞰や提言は、文部科学省・内閣府をはじめとする政府の各種政策立案のためのインプットとなる。

この一環として、2020 年 1 月 30 日に科学技術未来戦略ワークショップ「深層学習と知識・記号推論の融合による AI 基盤技術の発展」を開催した。本報告書は、その内容をまとめたものである。以下、注目する研究動向やワークショップの狙い、および、ワークショップでの発表・議論の概要を述べる。

深層学習（Deep Learning）が第 3 次 AI ブームを牽引している。画像認識、文書読解、囲碁・ポーカー等、さまざまな特定タスクで人間を上回る結果を出して話題になり、分類・予測・異常検知等の用途で、さまざまな産業応用も広がっている。その反面、大量の教師データや計算リソースが必要であること、想定外の状況での振る舞いが分からないこと、理由の説明が難しいこと等、現在の深層学習の限界も指摘されるようになった。

このような限界を克服する次世代 AI の方向性として、深層学習に記号推論を融合し、認識・行動のレベルから言語・推論のレベルまでを扱える仕組みが検討されつつある。ここでは、深層学習と知識・記号処理の単純な組み合わせ・併用ではなく、より本質的な融合が必要になる。

そこで、特に 2 つの論点について掘り下げるために、本ワークショップを開催した。第 1 の論点は「次世代 AI の中核的な技術チャレンジは何か？」であり、第 2 の論点は「それがもたらす社会・産業インパクトは何か？」である。5 件の発表と総合討議を通して、これら 2 点について検討した。

1 件目は麻生英樹氏（産業技術総合研究所人工知能研究センター 副研究センター長）が「自然知能と人工知能の共進化ー「知の創生」に向けてー」と題して発表した。これまでの AI 研究の 2 つの流れ（データから学習する即応的な知能と、記号的知識を使って推論・試行する熟考的知能）や課題を概観しつつ、2 つの融合に向けて、注目する取り組みや考え方が幅広く紹介された。

2 件目は尾形哲也氏（早稲田大学基幹理工学部 教授、産業技術総合研究所人工知能研究センター 特定フェロー）が「深層学習モデルのロボットへの応用」と題して発表した。ロボットへの深層学習の適用を通して知能を考えるというスタンスから、模倣予測学習や連続動作学習等の具体的な応用事例を交えつつ、ロボットにおける記号接地問題や動作の分節化の現状と課題が示された。

3 件目は乾健太郎氏（東北大学大学院情報科学研究科 教授、理化学研究所革新知能統合研究センター自然言語理解チームリーダー）が「自然言語理解への道のり」と題して発表した。現在の自然言語処理技術は、文章読解のベンチマークで一見高いスコアが得られていても、文脈理解・常識推論を含め、真の理解はいまだできておらず、今後、生のテキストの処理と知識ベースによる高次の処理を共通構造上に融合させる方向への発展が必要なことが示唆された。

4 件目は橋田浩一氏（東京大学大学院情報理工学系研究科教授）が「知能の基本構成素としてのサイクル＝価値＝意味」と題して発表した。仮説検証サイクルが、意味（価値）の維持・向上・創造に重要な役割を果たし、このサイクルを組み込んだ循環ニューラルネットワークや、サイクルをベースとしたモジュラリティーが今後向かう方向になるとの示唆がなされた。

5 件目は長井隆行氏（大阪大学大学院基礎工学研究科 教授、電気通信大学大学院情報理工学研究科 教授）が「マルチモーダル確率的生成モデルと汎用知能」と題して発表した。記号世界の正解ラベルに実世界のものを結びつけるという記号接地問題の捉え方は適切でなく、実世界とのインタラクションや社会的な共有の中で記号が創発されるという記号創発ロボティクスの考え方、汎用的なロボット・知能へのアプローチ、そこでキーとなるマルチモーダル確率的生成モデルが示された。

続く総合討論では、まず、深層学習と知識・記号推論の融合、あるいは、知覚・運動系の即応的知能（System 1）から言語・論理系の熟考的知能（System 2）までを統一的に扱う仕組みをどう捉えるかを論じた。次に、この方向で研究開発を推進するうえで、新たにエクスペリエンス（経験）のデータ化が重要になることが指摘された。さらに、国際競争力という面に関して、まだ皆がスタートラインに立った段階であり、日本にもチャンスがあること、および、ロボット、自然言語、教育といった切り口から産業応用が広がる可能性が示唆された。

目　次

1．趣旨説明

福島俊一（科学技術振興機構 研究開発戦略センター）

1．1　AI 研究発展の方向性

人工知能（AI）分野の研究開発に関する戦略プロポーザルの第 3 弾の発行に向けて、今回のワークショップを開催した。既に発行済みの 2 件の戦略プロポーザル[1]では、AI 分野の俯瞰に基づき、AI と社会との関係が大きくなってきた中で新たに生まれた研究課題に着目した。具体的には、AI 応用システムの安全性・信頼性の確保のための研究開発、および、フェイク問題への対策も含めた人間の主体的意思決定の支援に関わる研究開発について提言した。

今回の第 3 弾では、AI 基盤技術、AI の枠組み自体の発展に着目する。つまり、第 2 次ブームのときはルールベースが主流で、第 3 次ブームの現在は深層学習を中心とした機械学習ベースが主流になった。では、その次はどのような形へ発展するか、ということである。図 1-1 では「第 4 世代 AI」としている。この今回着目する AI 基盤技術の発展方向と、第 1 弾・第 2 弾で取り上げた社会や人間との関係は、実は非常に関係が深く、両輪として進めていく必要があるが、本日のワークショップでは、AI 基盤技術の発展にフォーカスして深く議論したい。

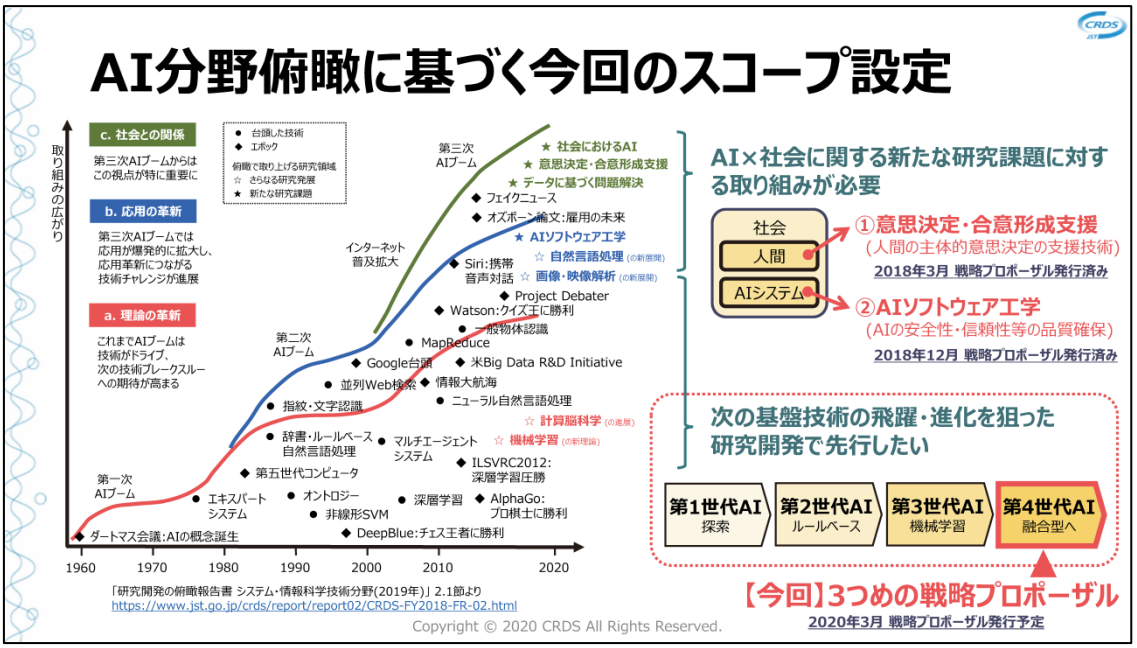

図 1-1　AI 分野俯瞰に基づく今回のスコープ設定

よく知られているように、AI 研究には 2 つの流れがある。ここでは、コネクショニズム的なアプローチとシンボリズム的なアプローチと呼ぶ（図 1-2）。第 2 次ブームのときにはシンボリズム側で、第 3 次ブームではコネクショニズム側と、主流が入れ替わった。そして、次はこれらが融合する方向に向かっている。以前からこのような融合の方向は考えられていたが、深層学習が発

[1] 戦略プロポーザル（1）「複雑社会における意思決定・合意形成を支える情報科学技術」2018 年 3 月
https://www.jst.go.jp/crds/report/report01/CRDS-FY2017-SP-03.html
戦略プロポーザル（2）「AI 応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立」2018 年 12 月
https://www.jst.go.jp/crds/report/report01/CRDS-FY2018-SP-03.html

展して、コネクショニズム側がシンボルまで扱えそうなところまで近づいてきたことで、漠然とした方向性として語られるレベルから、実現できそうなレベルに変わってきた。

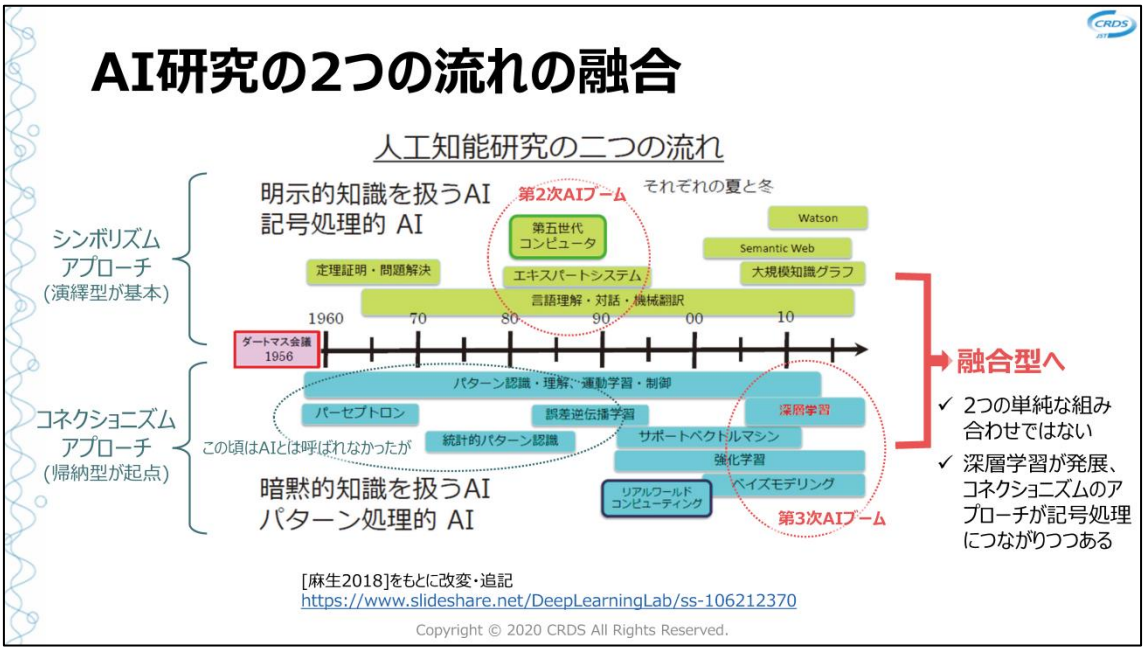

図 1-2　AI 研究の 2 つの流れ

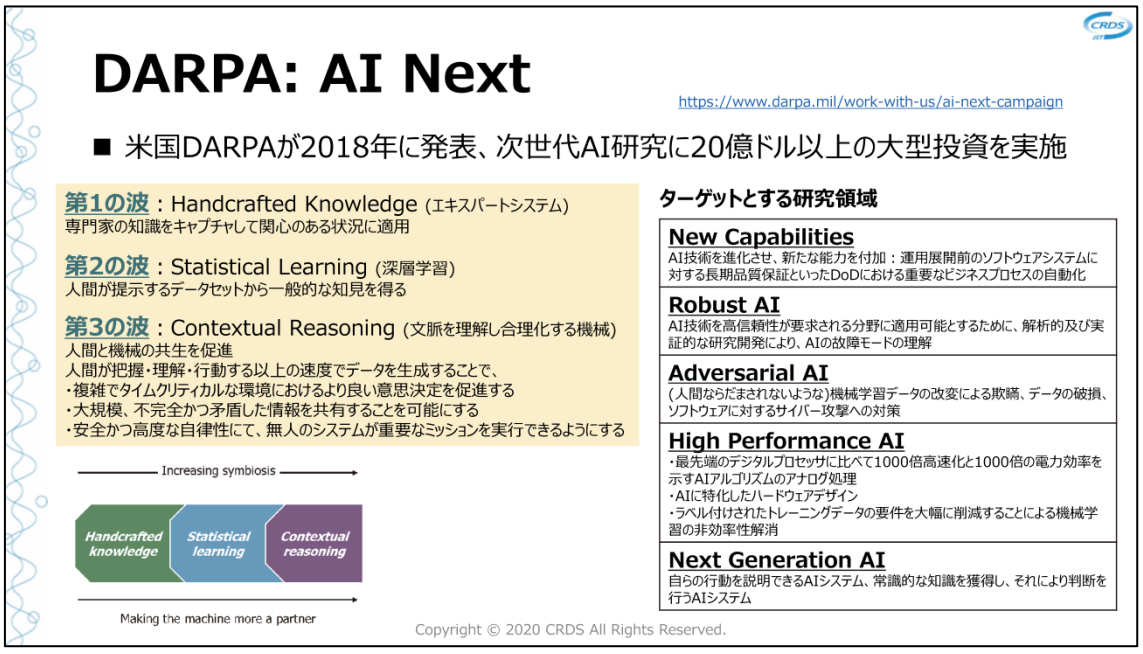

図 1-3　DARPA: AI Next

このような方向性に関わる最近のトピックをいくつか挙げる。

まずよく知られているのが米国 DARPA の AI Next である（図 1-3）。第 2 次 AI ブームのときが「第 1 の波」で Handcrafted Knowledge、すなわちエキスパートシステムのような人手で記述

したルールベースの AI である。第 3 次 AI ブームの今は「第 2 の波」で Statistical Learning、深層学習のような機械学習ベースの AI である。そして、DARPA のいう「第 3 の波」は Contextual Reasoning として、文脈を理解するとか、人間と機械の共生を促進するといった姿が語られているが、先ほど述べた 2 つの流れの融合という形をとるのではないか。

昨年 12 月の NeurIPS 2019 での Yoshua Bengio 先生による基調講演は、そのような方向性を取り上げたものだった[2]。つまり、今の深層学習は System 1 に相当するもので、将来の方向は System 2 の深層学習だと。ここでいう System 1、System 2 という考えは、行動経済学で有名で、ノーベル経済学賞も受賞した Daniel Kahneman 先生が「ファスト＆スロー」という著書で示してよく知られるようになった。人間は意思決定の際に、直感的・無意識的・非言語的で高速な System 1 と、論理的・意識的・言語的で低速な System 2 という 2 タイプのシステムを用いているという話である（表 1-1）。深層学習は現状 System 1 に相当するが、System 2 に拡張するには、アテンションや意識が重要な役割を果たすだろうとのことである。

表 1-1　System 1 と System 2

System 1	System 2
● 高速	● 低速
● 直感的	● 論理的
● 無意識的	● 意識的
● 非言語的	● 言語的
● 習慣的	● 計画・推論

また、NeurIPS と名前が変わる前の NIPS 2016 の基調講演では、Yann Le Cun 先生が、次の大きな挑戦は教師なし学習・予測学習だと、ケーキに例えて話をされた[3]。「知能をケーキに例えるならば、教師なし学習はケーキ本体であり、教師あり学習はケーキの飾り、強化学習はケーキ上のサクランボぐらいである。私達はケーキの飾りやサクランボの作り方は分かってきたがケーキ本体の作り方は分かっていない」と。

こういった動きは国内で早く、2 年前、中島秀之先生らが中心となって、東大で「深層学習の先にあるものー記号推論との融合を目指してー」という公開シンポジウムを開催した[4]。その第 2 回は昨年開催され、本日ここにおられる麻生先生や尾形先生も発表されているが、計 2 回の発表者を見ると、日本の AI 分野のリーダー的な方々が、次はこういう方向が重要だと、国内においても先んじて発信していることは注目に値する。

[2] https://youtu.be/T3sxeTgT4qc で講演の動画と資料が見られる。Yoshua Bengio 先生（カナダのモントリオール大学教授）は、2018 年のチューリング賞の受賞者（深層学習の父たち）の一人である。
[3] https://drive.google.com/file/d/0BxKBnD5y2M8NREZod0tVdW5FLTQ/view で講演資料が公開されている。Yann Le Cun 先生（Facebook のチーフ AI サイエンティスト、ニューヨーク大学クーラント数学科学研究所のシルバー教授）も、2018 年のチューリング賞の受賞者（深層学習の父たち）の一人である。
[4] 第 1 回は 2018 年 1 月 22 日に開催された。
http://park.itc.u-tokyo.ac.jp/FAIRE/event/「深層学習の先にあるもの-記号推論との融合を/
第 2 回は 2019 年 3 月 5 日に開催された。
http://park.itc.u-tokyo.ac.jp/FAIRE/event/「深層学習の先にあるもの-記号推論との融合を-2/
第 2 回の講演動画は https://todai.tv/contents-list/2018FY/beyond_deep_learning で公開されている。

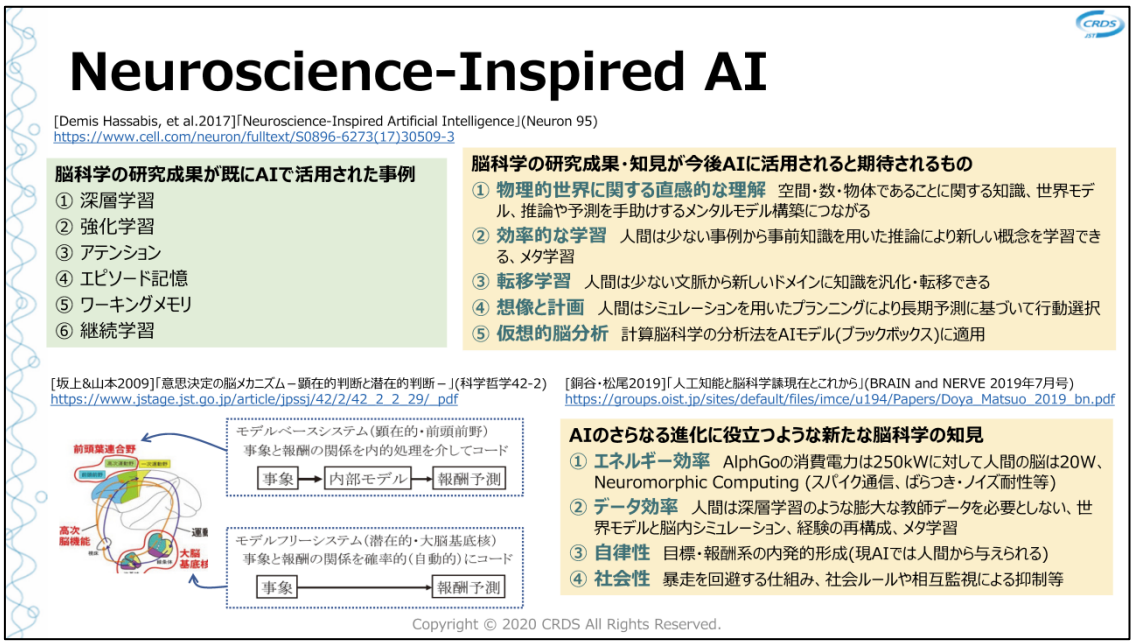

図 1-4　Neuroscience-Inspired AI

関連して、Neuroscience-Inspired AI にも注目したい（図 1-4）。AlphaGo で有名な DeepMind 社が取り組んでいることがよく知られている。図 1-4 の上半分はその DeepMind の論文から要点を抜粋したもので、深層学習・強化学習・アテンション等、脳科学の研究成果・知見が既に AI で活用されている一方、今後 AI に取り込まれることが期待される知見として、大量の教師データを必要としない学習、学習結果の異なるドメインへの応用、世界モデルや想像等が挙げられている。また、図 1-4 の左下に示すように、脳科学においても、モデルフリーとモデルベースという 2 タイプの情報処理が、脳内の異なる部位で行われているという見方がある。今の深層学習・強化学習はモデルフリーに相当し、ここでも、もう一つのタイプが必要なことが示唆されている。人間の脳から学ぶ、ヒントを得るということが有効と思われ、計算脳科学や認知発達ロボティクスのような研究分野の知見が参考になると期待される。

1．2　論点

このような動向を踏まえて、今回のワークショップでは、現在を第 3 世代と呼ぶとき、その次の第 4 世代 AI の方向性として、深層学習と知識・記号推論の融合を考える（図 1-5）。帰納型 AI と演繹型 AI の融合という見方もある。教師なしでも知識獲得・成長していく AI という面もある。先ほど述べたように、DARPA は「第 3 の波」と言っているが、フェーズとしては同じでも、その内容は微妙に異なるので、同じ言い方をすることはやめた。「第 4 世代 AI」という言い方は広く認められているわけではないので、違和感を持つ方もいるかもしれないが、とりあえず仮置きで進めさせていただきたい。

図 1-5 に「第 4 世代 AI へ向けて」ということで開催趣旨を示した。この中で「融合」というのは「単純な組み合わせ・併用ではなく、より本質的な融合が必要」としている。それはどのようなものが要求されるのか、次の世代へ飛躍すると言えるようなブレークスルーとなる融合とは

どのようなものか、といった点が今日の議論での技術的な面からの論点である。

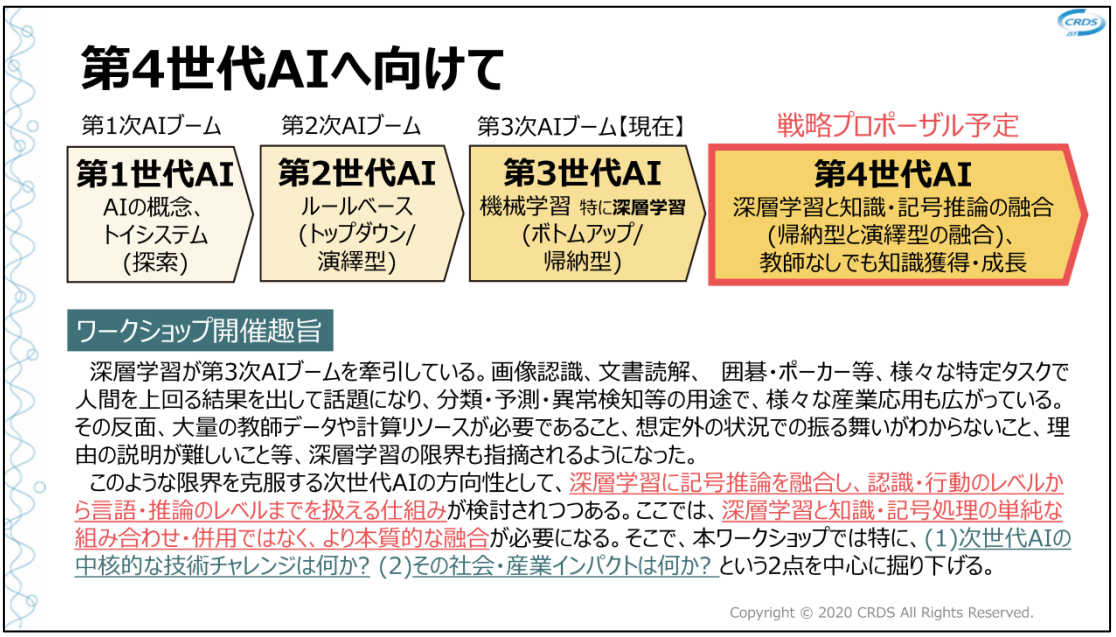

図 1-5　ワークショップ開催趣旨：第 4 世代 AI へ向けて

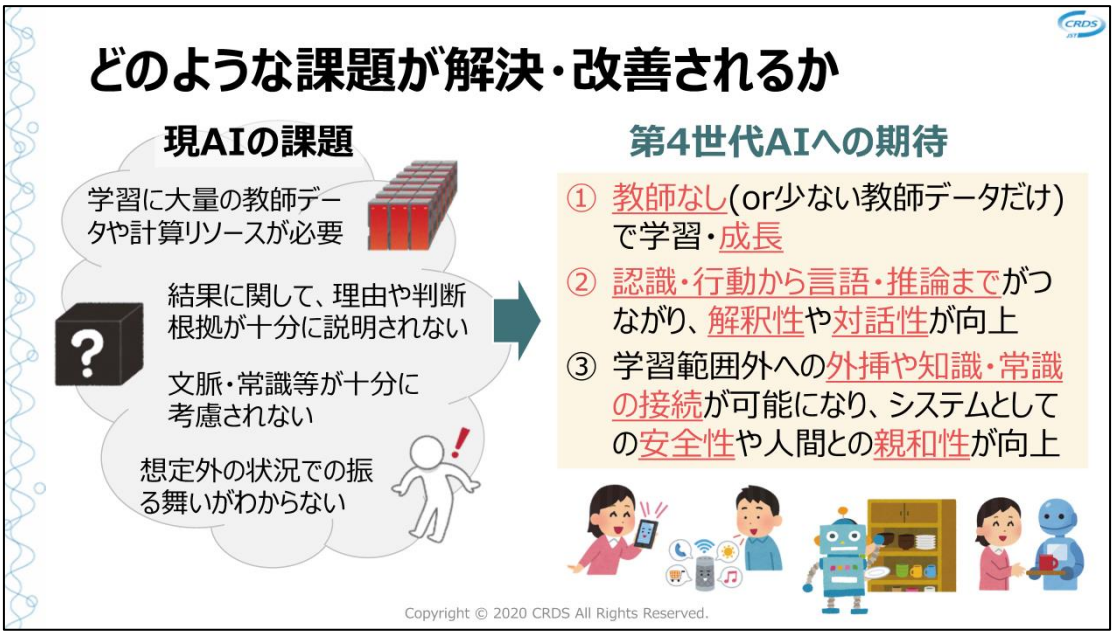

図 1-6　第 4 世代 AI でどのような課題が解決・改善されるか

ところで、どのような融合かというのは AI の構成論であって、その結果として、どのようなありがたみがあるのか、実際のどのようなシーンでどのような効果が出るのか、それは例えば産業面ではどのような分野の発展が期待できるのか、といった面についても、もう少し明確化したいというのが、もう一つの論点である。

この後の各発表で具体的に示されるものがあると思うが、図 1-6 に、現 AI の課題解決という面から 3 つの期待を挙げた。1 つめは、大量の教師データがなくても学習・成長できるというこ

と。2 つめは、認識・行動から言語・推論までがつながって、解釈性や対話性が向上するということ。3 つめは、機械学習が学習した範囲内だけでうまくいくという問題に対して、学習範囲外への外挿や知識・常識といったものも扱えるようになり、システムとしての安全性や人間との親和性が向上するということである。これらが適切なものかについても議論したい。

1．3　融合の形態

先ほど述べたように、単純な組み合わせや併用ではなく、より本質的な融合はどういうものかが論点になるが、帰納型 AI と演繹型 AI を組み合わせる試みは既にいくつか行われている。それをここで簡単に振り返ってみたい。こんなレベルの組み合わせではだめだとか、今後のどういうものが重要になっていくかとか、といった議論の素材として 4 つほどの形態に分けて紹介する[5]。

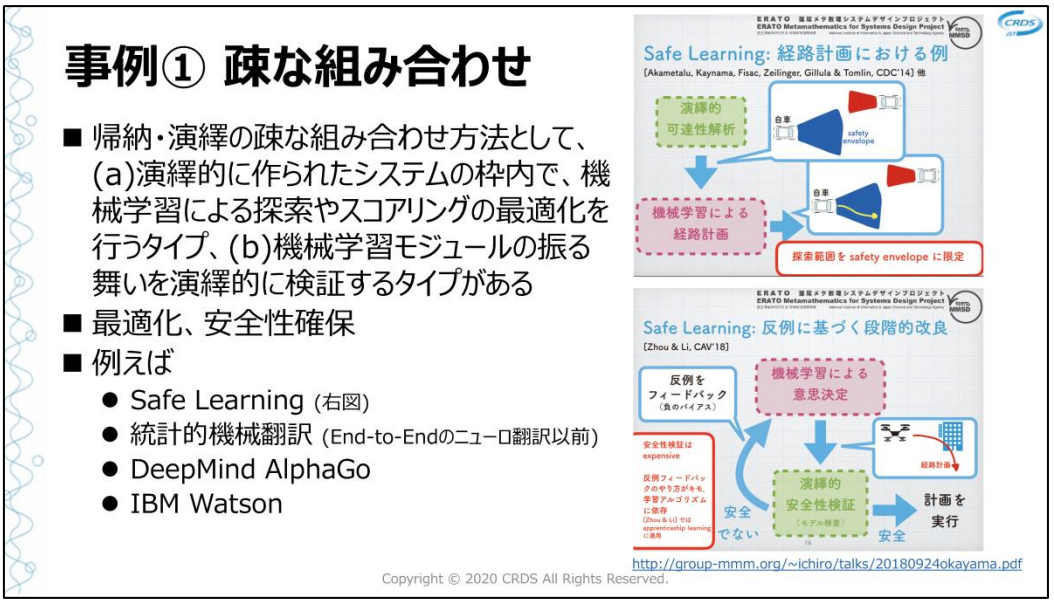

図 1-7　融合形態①：疎な組み合わせ

（1）疎な組み合わせ

1 つめは疎な組み合わせだが、図 1-7 ではこれには 2 通りがあることを示した。その 1 つは、従来型の演繹的なプログラミングで作られたシステムの枠内で、探索を効率化するとか、評価関数やスコアリングを最適化するために機械学習を使っているタイプである。DeepMind の AlphaGo や IBM の Watson では、このような方法が使われている。統計的機械翻訳と言われていた時代の機械翻訳システムも該当する。

もう 1 つは、機械学習モジュールの振る舞いを外側から検証して、問題あるケースは回避するという形態で、Safe Learning と言われる、安全性を確保するという目的等で使われている。図 1-7 の右上は、まず演繹的に安全な範囲を決めて、その中で機械学習を動かすという組み合わせ方で、右下は、機械学習で何か結果を出した後で演繹的に検証するという組み合わせ方である。

[5] この発表ではパターンを 4 つ紹介したが、2020 年 3 月公開予定の戦略プロポーザル「第 4 世代 AI の研究開発 ―深層学習と知識・記号推論の融合―」では、「シミュレーションと帰納的手法の組み合わせ」「記号創発ロボティクス」という 2 つを増やして 6 パターンとした。後者については、本ワークショップにおいて、長井先生の発表で取り上げられている。

このような疎な組み合わせはこれまでも多数行われてきている。たぶん、これではまだ足りなくて、より深い組み合わせ・融合に進みつつあるのだと思う。

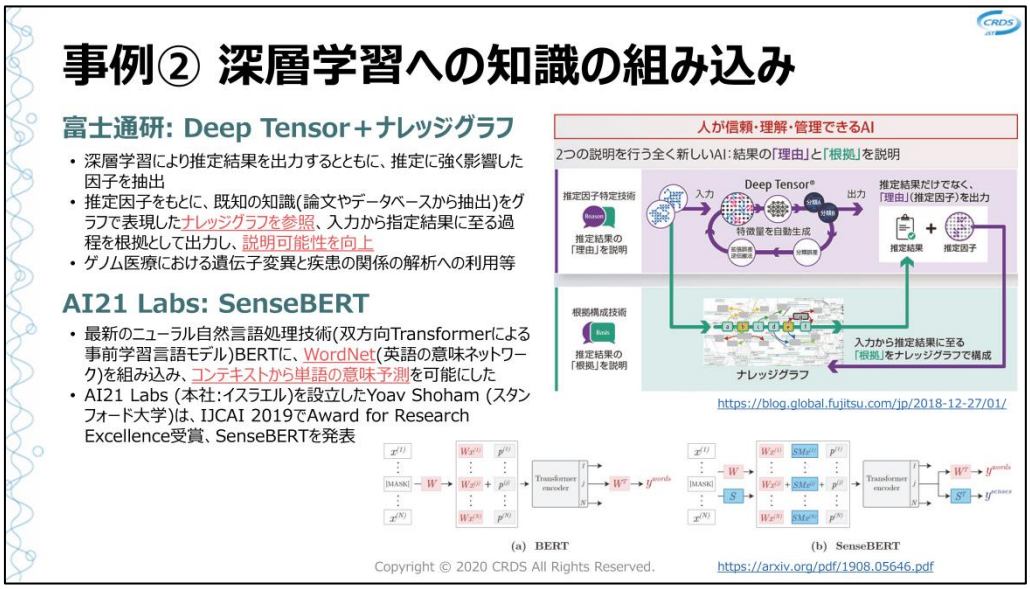

図 1-8　融合形態②：深層学習への知識の組み込み

（2）深層学習への知識の組み込み

2 つめは、深層学習に知識を組み込むというもので、これもいくつかの事例がある（図 1-8）。富士通研では、深層学習の Deep Tensor にナレッジグラフを組み合わせて、その間の対応をうまくとれるようにすることで解釈性を高めている。イスラエルに本社のある AI21 Labs が発表した SenseBERT は、BERT の中に意味ネットワーク WordNet を組み込んで、コンテキストを考慮した単語の意味予測をできるようにしようというものである。

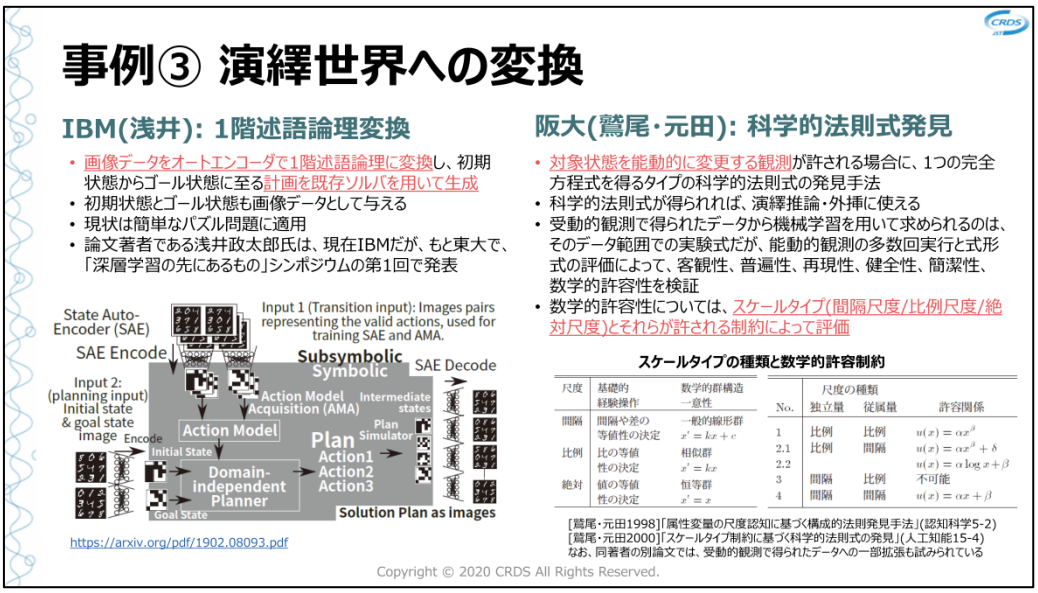

図 1-9　融合形態③：演繹世界への変換

（3）演繹世界への変換

今紹介した形態は深層学習の側に組み込むものだったが、逆に演繹の世界に持ってくるタイプのものもある（図 1-9）。最近発表された IBM の事例では、画像をオートエンコーダーで一階述語論理に変換して、既存のソルバを使ってプランニングするという方法をとっている。また、大阪大学の鷲尾先生らによる科学的法則式の発見の試みもこのタイプに該当する。機械学習によって大量データから帰納的にルールを見つけることができるが、それは実験で測定された範囲だけをカバーする経験則でしかない。それを演繹に使えるような法則式に格上げするため、能動的な観測を通して、客観性や再現性等を高めるという方法が示されている。

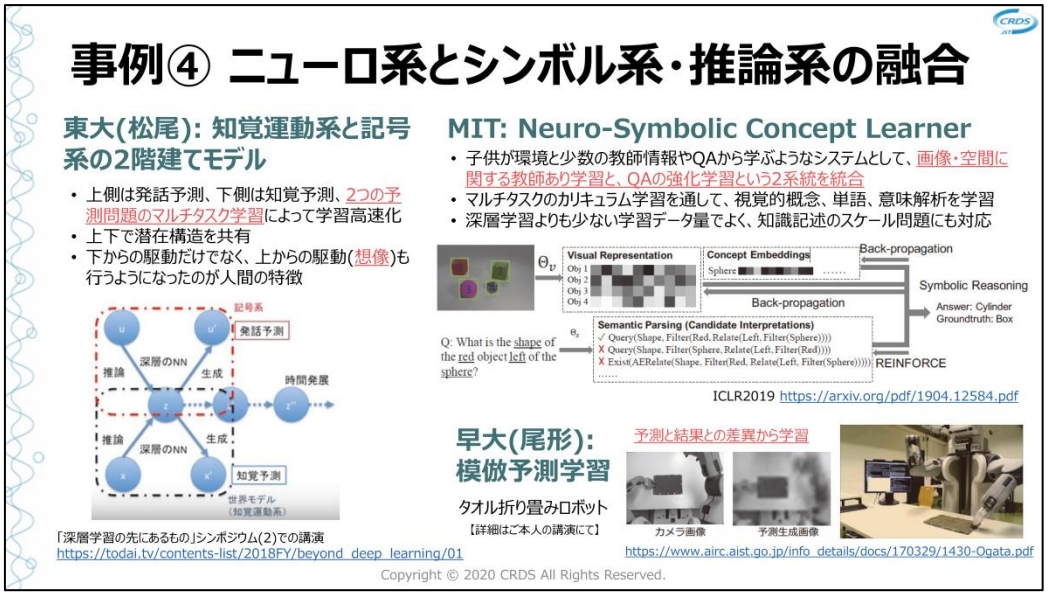

図 1-10　融合形態④：ニューロ系とシンボル系・推論系の融合

（4）ニューロ系とシンボル系・推論系の融合

4 つめの形態は、ニューロ系とシンボル系あるいは推論系を融合するというもので、かなり密な融合になる（図 1-10）。先ほど触れた東大での公開シンポジウムにおいて、松尾豊先生が発表された知覚運動系と記号系の 2 階建てモデルがこのタイプである。上側の記号系の方が発話予測をして、下側の知覚運動系が知覚予測をする。そして、これらがマルチタスク学習を通して連動する。このような枠組みで「想像」も可能になると。

また、MIT の Neuro-Symbolic Concept Learner、IBM と共同の研究チームだったと思うが、このシステムは、画像や環境・空間の認識と質問応答（QA）という 2 系統を持ち、画像・空間系は教師あり学習、QA の方は強化学習のような形で、2 系統から学習していく。これによって、大量の学習データを与えなくても、より効率よく学習できるというものである。まだ積み木の世界で試されているレベルだが、2 系統の連動・融合が試みられている。

さらに、模倣予測学習という方法もこのタイプに相当する。本日は尾形先生から説明されるはずなので、私から詳しい説明は省くが、帰納的に得られたモデルで予測するというレベルもあるかもしれないが、物理モデルを用いたり、演繹的な知識を組み合わせたりして、演繹的な意味合いの強い予測になると、帰納・演繹の融合と言ってもよい。

以上、4 つのタイプに整理して、現状の関連する事例を紹介したが、後半の方が、より融合の度合いが高い感がある。このような形になってくると、いわゆる「記号接地問題」にアプローチすることになる。図 1-11 は、中島秀之先生による「記号接地問題」（シンボルグラウンディング問題）の用語解説を引用したもので、深層学習によって一部の記号接地は学習可能になりつつあると書かれている。記号接地問題という定義自体が適切なのかという議論もある[6]。実世界の対象物に対応する記号と、抽象的な概念や行為に相当する記号とではレベルが異なるとも思われる。記号接地問題をより適切に再定義して、問題を段階的に定義し直すとか、分解するとかしたうえで、取り組むことが必要になっているようにも思う。このあたりも、第 4 世代 AI に移行するうえでのポイントになってくるのではないか。

シンボルグラウンディング問題
人工知能（AI）の知識表現において、そこで使われる記号を実世界の実体がもつ意味に結び付けられるかという問題。「記号接地問題」ともいう。（中略）最近ではディープラーニング（深層学習）によって記号と画像の関係を学ばせ、記号表現を画像に落とすことができるようになってきているので、一部の記号接地は学習可能になりつつある。
［中島秀之］2019 年 8 月 20 日　©Shogakukan Inc

図 1-11　記号接地問題（シンボルグラウンディング問題）
［引用元］日本大百科全書 https://japanknowledge.com/contents/nipponica/sample_koumoku.html?entryid=1164

これで、私からの趣旨説明は終わりとし、この後、麻生先生、尾形先生、乾先生、橋田先生、長井先生の順に発表をお願いする。改めて本日の主な論点を示すと、先ほど述べたように、この 2 点である（図 1-12）。

論点（1）次世代 AI の中核的な技術チャレンジは何か？
深層学習＋記号推論 ～ 記号接地問題へのアプローチ等
論点（2）それがもたらす社会・産業インパクトは何か？
ロボットおよび自然言語処理が有望と見込まれる等

図 1-12　本日のワークショップでの主な論点

1 点目は技術的な面で、現在の AI から次の世代の AI と言えるような飛躍あるいは進化をもたらす、中核的な技術チャレンジは何か。深層学習と知識・記号推論の融合、帰納型 AI と演繹型 AI の融合という方向で、単純な組み合わせというよりは、記号接地問題的なところにも踏み込んだ深い融合というものがどのようなものか。このあたり技術的にディープな議論になるかもしれない。

2 点目は、そういったことができた次の世代の AI は、社会的・産業的インパクトとして、どのような分野でどのような効果を生み出せるか。私の方でこれまでいろいろな方々にお話を聞いてきた中では、ロボットや自然言語、対話といったところが有望という話は出ていたが、本日のメンバーはどうお考えか。

では、順に発表をお願いしたい。

[6] 例えば、「記号創発問題 ―記号創発ロボティクスによる記号接地問題の本質的解決に向けて―」（谷口忠大、人工知能学会誌 31 巻 1 号、2016 年）

2．発表

2．1　自然知能と人工知能の共進化ー「知の創生」に向けてー

麻生英樹（産業技術総合研究所）

人工知能の２つの流れについては、最近 Bengio らも「System 1 深層学習から System 2 深層学習へ」と言っているが、このような話は昔からされていた。私も、2015 年に人工知能研究センターができて以来いろいろなところで講演をするときに、図 2-1-1 のような図を使っている。「人工知能」と「統計的機械学習」はもともと異なる discipline であり、私はもともとニューラルネットをやっていたのでどちらかと言えば「統計的機械学習」がベースである。ブレークスルーは主に機械学習の側で起こっていて、パターン情報の認識と生成の性能が大幅に向上した。その要因の一つは、潜在特徴表現獲得の性能向上であろう。その一方、解けていない問題もあり、データ量・計算量・消費電力の壁や性能保証、説明可能性などは福島さんの説明でも言及されている。

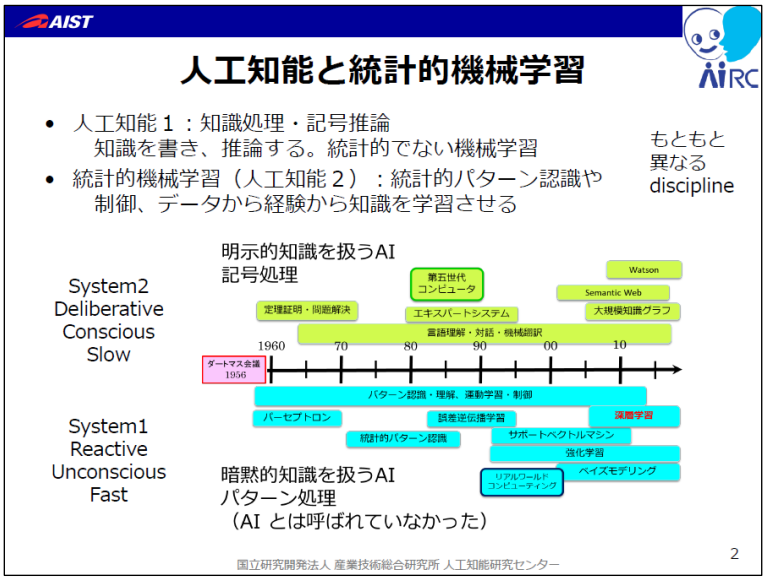

図 2-1-1　人工知能と統計的機械学習

いろいろな課題があるが、根源的な課題は、データから学習する即応的な知能と、記号的知識を使って推論・試行する熟考的知能の融合ではないかと思う。これが解けていないから、いろいろな問題が解けてないと考えている。この問題もすごく古い話ではある。私は 1988 年に書いたニューラルネットの本で「柔らかな記号」という話をしていたし、長尾真先生や安西祐一郎先生の岩波新書の書籍でもニューラルネットのような非シンボル的扱いと記号処理的人工知能という２つのモデルを総合したシステムを検討することが重要とされている。Harnad が 1990 年に Symbol Grounding Problem という題名の論文を書いたことから「記号接地問題」と呼ばれることも多いが、もっと広い文脈で考えることができる。

この話を遡ると、例えば、1982 年に出版された David Marr の「ビジョン」が挙げられる。これはコンピュータビジョンの研究についてまとめた書籍だが、Marr 自身は元々脳の情報処理の研究をしており、1970 年代には脳の情報処理に関する３つの大作の論文（大脳皮質の理論、小脳

の理論、古皮質の理論）を書いている。そうした研究も踏まえて「ビジョン」の中で、脳のような複雑な情報処理システムを理解するためには少なくとも３つのレベルを考える必要があり、一番下は物理的実装、２番目は情報表現とアルゴリズム、３番目は計算理論（Computational Theory）であると言った（図 2-1-2）。そして、視覚情報処理で用いられる情報表現として、Primal Sketch、2 1/2 D Sketch、3D Model という抽象度や目的の異なる表現を提案している。ここで言われているような「構造をもった階層的な情報表現」をどうやって観測データから構築するか、が最も重要な問題ではないかと考えている。

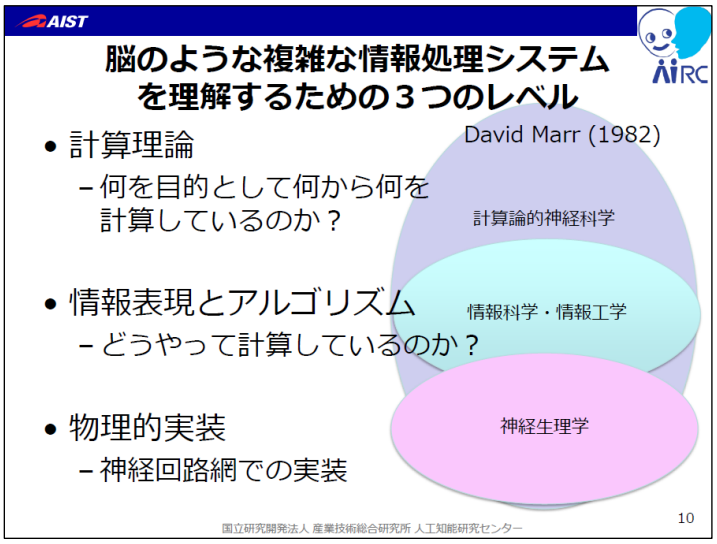

図 2-1-2　脳のような複雑な情報処理システムを理解するための３つのレベル

脳のような複雑なシステムを理解するためには複数の情報処理のレベルと、そこで使われる情報表現が必要だということは、他にもさまざまなコンテキストで言われている。例えば、Rodney Brooks のサブサンプション・アーキテクチャも System 1、System 2 のような話である。一番下の反射的に行動する部分から、もう少し場合ごとに対応、抽象的に試行、理論的に考えるという階層があり、それらが相互作用するアーキテクチャである。それまでの AI に対する批判として提唱され、その後サブサンプション・アーキテクチャのシステムはたくさん作られた。他にも、川人光男さんの「視覚と運動の計算理論」や尾形さんが使っている「Predictive Cording」、Friston らの「自由エネルギー原理と Active Inference」など、ボトムアップとトップダウンの処理を融合させようという考え方も、抽象的には同類と言えるのではないかと思う。

最近の深層学習の分野では、松尾豊さんも注目している「World Model の学習」がある。ここで World=世界というのは、図 2-1-3 に示したように、観測すると何か観測結果が得られる関数で、ある種の質問応答する対象であると考えられる。World Model とはそのような質問応答ができるようなモデルである。１個の World Model ではなく、例えばこの部屋の World Model、違う部屋の World Model などと、多数の World Model がある。そのさまざまな World Model を学習するときに、1 個ずつ全てゼロから学習するのではなく、もう少しメタな学習が必要だろうというあたりがポイントであると思う。

World Model の論文では、３次元の空間に積み木がいくつか置いてあり、それをいくつかの視点から観察した画像データだけからこの世界のモデルを学習し、任意の視点で観測したときにど

ういう画像が得られるかを推測する問題を扱っている。これは、ある立体をさまざまな角度から見たときに、どう見えるかを学習する問題だが、松尾さんたちの研究では、この問題を、複数の学習課題の一つ一つを事例とする学習である「メタ学習」ととらえて、ベースラインとなる方法よりも早く収束する方法を提案している。ポイントは、1 個の対象を学習するのではなく、たくさんのいろいろな対象を学習するときに早く学習できるというところである。こうした研究は、昔から Learning to Learn や Lifelong Learning などと呼ばれているが、学習対象、環境、世界の組み合わせ的変化に対応して高速に学習するための事前知識（Prior）の獲得である。

メタ学習の分野では、最近は、Abbeel らの MAML（Model Agnostic Meta Learning）が有名だが、これは、いくつかの学習課題があるときに、課題全部に適した初期値を Prior としてメタ学習するものである。Bengio の Consciousness Prior では、抽象概念の間の Sparse Factor Graph で表現される結合分布を Prior として使うことを提案している。この方法は、そもそも我々の世界の変化は局所的で独立だということを仮定していることに対応する。そうでないと、そもそもいろいろな世界に対応して学習するなどはできない。これはとても一般的な仮定で、こういう Prior を入れておくと、我々の世界の多くの問題に関しては有効なことが多く、しかもこの性質は、因果的知識が我々の世界で有効であることと密接に関係していると主張している。これは、McCarthy と Hayes が指摘した「フレーム問題」とも密接に関わる。例えば、あるエージェントがあるオブジェクトを操作するといったときに、操作した結果の影響が局所的に留まるという仮定がないと、未来の予測や行動ができない。そういうことを Prior として入れてゆこうというのが Consciousness Prior の基本的な考え方だと理解している。

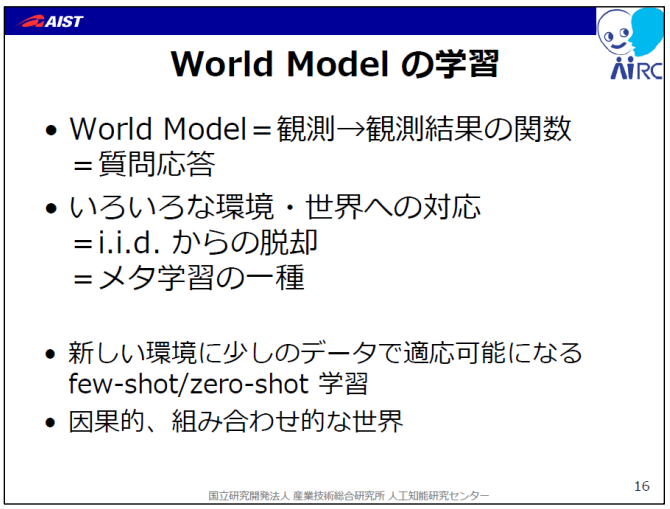

図 2-1-3　World Model の学習

最近話題になっている Self-Supervised Learning（自己教師あり学習）も情報表現の学習に関するものだ。Self-Supervised という言葉自体は、Hinton らも含め昔から使われているが、一番基盤にある情報表現を Unsupervised や Self-Supervised な学習でしっかり学習しておいて、そこで得られた表現を使うと、教師あり学習や強化学習が簡単になる、という話である。Predictive Coding も典型的な Self-Supervised Learning の一種で、未来の予測は、時間がたてば必ず正解が得られるので、教師となるラベルをつけるが必要ない。この意味で、予測は知能の源泉だと言

われるが、Self-Supervised Learning は、それを、未来に限らない形に拡張している。過去の知っているところをわざと隠して学習したり、画像の一部をわざと隠して学習したりなど、教師ラベルを新規に作らなくてもよい学習課題というのはさまざまに設定できる。そうした新しくラベルを付けなくてもよい問題設定の中で、内部表現学習をしっかりやればよいという、とても理にかなった考え方である。

自然言語理解の分野でこの考え方を展開しているのが BERT（Bidirectional Encoder Repre-sentations from Transformers）である。この研究では、自己教師あり学習で単語や文章の埋め込み表現を学習するために、モデルとして Attention（注意）メカニズムをベースにした Transformer を使っている。これによって LSTM のような回帰結合型のネットワークは自然言語処理の世界では要らなくなった、と称している。Attention は、状況に応じて必要な情報にフォーカスする仕組みである。例えば、畳み込みニューラルネットワークで使われている畳み込み計算は、入力画像の中のある限定された部分から情報を集めて、畳み込みフィルターをかける計算だが、Attention はそうではなくて、入力画像の全体を対象にして、問題や状況に合わせて必要なところだけから情報を取ってこられるように、重みのつけ方自体も学習する。従って、畳み込みフィルターを使わずに、Attention メカニズムだけで画像認識をするとことも可能である。

また Generative Adversarial Learning（生成的敵対的学習）もとても有力なアイデアである。これは複数のネットワークを敵対的に学習させることで、高次元データの生成分布の学習の損失関数をうまくサロゲートできるというとてもよくできたアイデアである。現在は、データの生成のほうに注目が集まっているが、これと、Consciousness Prior や World Model などの内部表現に関する仮定をうまく組み合わせて、内部表現学習のために利用すると、いろいろ面白いことができるのではないかと思う。

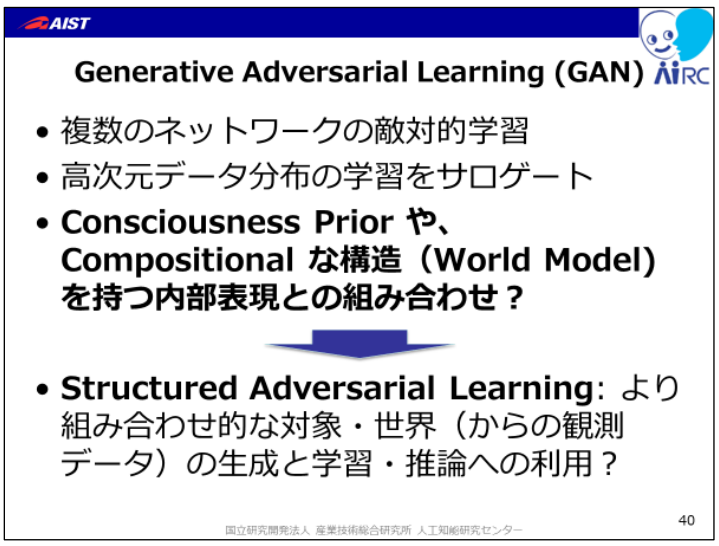

図 2-1-4　Generative Adversarial Learning

以上のように、二つの情報処理の融合という根源的な問題に関して、昔から、多くの研究者がいろいろなアイデアを積み重ねてきているが、まだ何が足りないものがあるのだろうか？ トップダウンとボトムアップをつなぐ仕組みや、「良い」構造のある特徴表現を学習する仕組み、その表現の操作系、データや計算機が足りないということも考えられるが、個人的には「何のために統

合するのか？」というところが最も足りていないと考えている。融合を必要とするタスクとして、Q&A（質問応答）などが提案はされているが、ImageNet のような、研究を進めるのにちょうど良いタスクがまだない。そして、さらに、融合がどのような社会的・産業的なインパクトを生むのかという点もあまり検討されていない。以降では、そうした、２つの情報処理の融合のタスクとユースケースについて少し考えてみたい。

囲碁は盤面評価が System 1 で、先読みは System 2 なので、AlphaGo で既に融合できているとも考えられる。実際、先読みのコントロールに盤面評価を使ったり、評価結果を先読みで検証したりできている。もう少し深い融合を考えるとしたら、例えば GAN で囲碁の盤面を生成し、それをイメージとして入力して評価したり、イメージの中で先読みしたりといったことが可能かもしれない。しかし、囲碁の場合には、盤面を非常にコンパクトに記述可能で、それ以上の情報は要らないため、わざわざ画像にするメリットはない。コンピュータ・グラフィクスを使ったゲームであれば、ゲームのある状況を記号に完全に落とし込むのは難しいかもしれないなので、何か、そうした生成パターンを使う意味があるかもしれない。

もう一つ、私たちが、1990 年から 2000 年頃に取り組んでいた「事情通ロボット」の研究を挙げよう。この研究では、自律的に環境から情報収集して学習するオフィスロボットを作ることを目指した。実際には、とても単純なことしかできていなかったが、やろうとしていたことは、人と対話することによって学習するということである。対話システムは、基本的に人にサービスするものとして作られるが、事情通ロボットの場合はそうではなく、自分の利益のために対話する。ロボット自身が人間に質問する文、人間からの回答文などを、センサーの情報と組み合わせて、何か構造を持った潜在的な意味構造のようなものができ、これを通していろいろな情報がつながるということを考えた。この研究の後、対話は非常に難しいことが分かったので、しばらく対話の研究はするまいと思っていたが、最近になって、例えば、Embodied Q&A という、実世界で移動するロボットと質問応答をする研究が、コンピュータビジョンのトップ国際会議である CVPR で発表され、ベンチマークデータも作られているようである。深層学習の発展のおかげで、事情通ロボットのような話も、実現に近づいてきているのかもしれない。

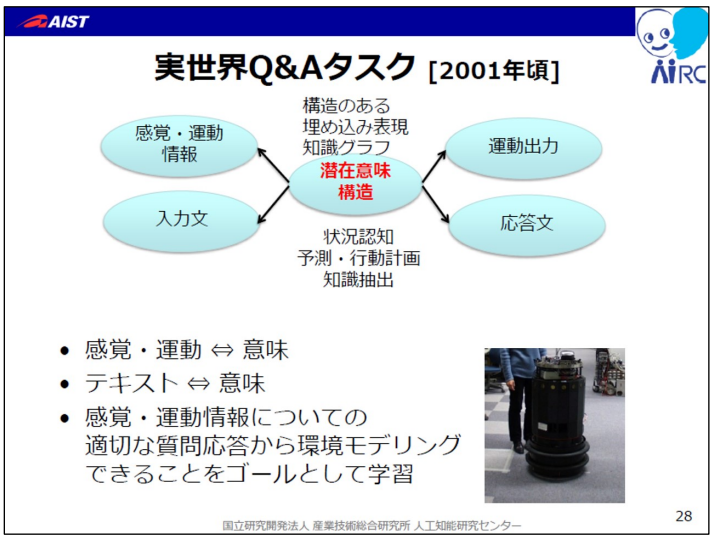

図 2-1-5　実世界 Q&A タスク

また、人工知能研究センターで行っている研究の中には、イベント会場などで、会場内に設置された固定センサーや移動ロボットで観測を行い、人の流れや３次元の地図などの情報を集約してゆくというものがある。リアルタイムでどんどんデータが集まってくるので、それをプラットフォームに集積して、いろいろなことに役立てることを狙っている。これはいわゆる Cyber-Physical System（CPS）の一例だが、事情通ロボットのようなものをこの CPS の中に入れて質問応答できるようにすると、面白いことができるのではないかと思っている。

また、AI センターの別の研究として、工場での組み立て作業などの複数ステップの作業をロボットに学習させるというものがある。工場では変種変量生産で作る物がいろいろと変わるが、その度にロボットのプログラムを作り替えるのは大変である。そこで、ロボットがかなり自律的に学習し、自分で計画して、対象の変化などに合わせられるようにしたい。この問題に対して、原田研介さん（阪大／産総研）は、人間の作業を観測したデータやシミュレーションを使って計画を立てロボットが実行するという、どちらかというと、古典的な AI を発展させてゆくアプローチをとっている。一方、尾形さん（早大／産総研）は、全部 End-to-End でやるアプローチで、人間が最初にいくつかの例を動かして見せると、その経験から深層学習する。タオルを巻くというタスクで、複数の動作の動作が深層ニューラルネットワークに埋め込まれて学習されて、状況にあわせてその切り替えができるようになる。松原崇充さん（奈良先端大）は、ハンカチをひっくり返し、目的の形に折りたたむという課題を、全部強化学習でやるアプローチを取った。

次のステップとして、これらのアプローチがうまく統合されると、面白いことができるのではないかと個人的には思っている。強化学習だけで全てをやるのは大変なので、人間の教示を使うことや、古典的なプランニング型の AI に深層学習をもう少し取り入れること、などは考えられ、今後上手に統合できると良いと思われる。さらに、こうした研究が図 2-1-6 に示すような熟練者の技を AI が取り込むというニーズにつながっていき、その結果として、熟練者を支援するだけでなく、熟練者の育成の社会的コストと時間を劇的に下げるようなことができると、こういう研究も社会的な意味が大きくなると思う。

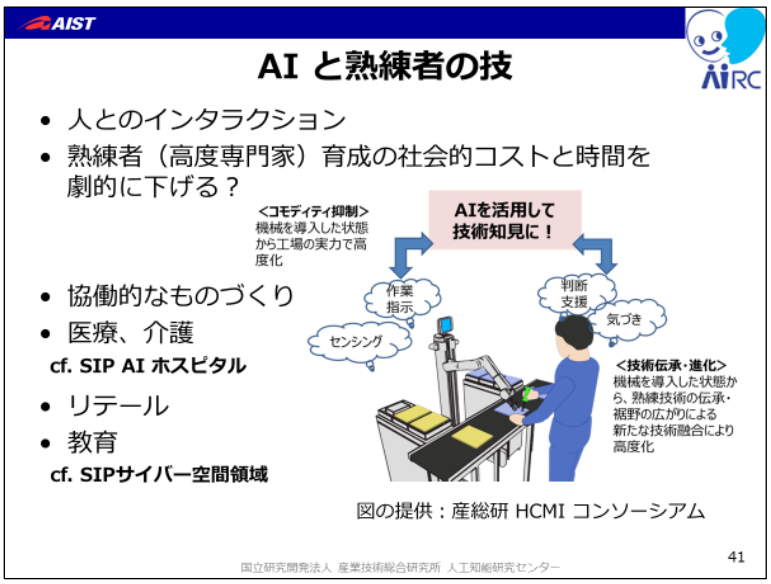

図 2-1-6　AI と熟練者の技

もう一つの社会的に意味のあることは、AI を科学技術研究に使うということだろう。科学技術研究こそ蓄積された人間の知識の宝庫であるので、そういうものと、AI によって大量のデータから学習されるものをつながないといけない。深層学習と記号推論の融合の社会的な意義の一つだろう（図 2-1-7）。最近、人工知能研究センターの辻井センター長が「共進化 AI」という言葉を使っている。人間は System 1 と System 2 の両方をもつが、今の機械学習 AI は System 1 だけで、異質な知能である。今、人間の専門家は AI に対して、ラベルをつけて教えるだけしかしていないが、この異質な知能が、もっと豊かなインタラクションをして、共に育っていくようなことを目指すべきだと言っている。この話は当然、「知識社会」につながっていくと思う。領域アドバイザーとして関わった JST さきがけ「知の創生と情報社会」（2008～2014 年）で議論されていたことを、社会的なインパクトとして考えてはどうかと思う。

今の AI は、機械学習がコアにあり、その本質はデータの利活用である。データを利活用して社会をどれだけ賢くできるかというのが社会の競争力を決めるわけなので、２つのレベルの統合のようなことも、そこに貢献してゆくことが必要だろうし、それによって科学を拡張することが考えられると思う。こうしたことについて、何かうまい言い方がないかと模索する中で思いついたのは「データが主役の科学」である。これまでの科学は、間違っているかもしれないが、データは理論に奉仕する、つまり、理論屋さんのほうが偉く、実験屋さんはデータをとって、それをサポートする、というような印象があったと思う。そうではなくてデータのほうが主役で、抽象化された理論のほうは、例えばデータの検索性を上げたり整理したりして、使いやすくするためにある、というような「科学」も可能ではないかと思うが、どうだろうか？

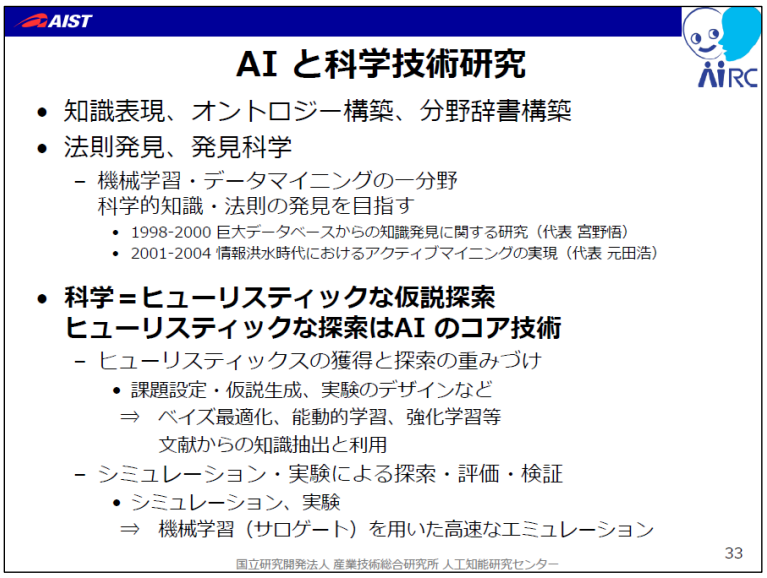

図 2-1-7　AI と科学技術研究

質問応答

Q：トップダウンとボトムアップというのは、何か違和感がある。どちらもトップダウンでもボトムアップでもないような気がするので、何か別の言い方はないか？

A：そもそもこの２つのシステムの融合という話は、上手く表現しにくい。我々も「データと知識の融合」などと言っているが、機械学習で得たものも知識であるので、厳密に言うと

変である。何か別の言い方があったほうが良いというのはそのとおり。

Q：トップダウンとボトムアップは明らかにまずいと思う。

A：同意する。「記号接地問題」というのも問題を矮小化するので、あまり良くないと思っている。接地さえすればいいのであれば、それこそ、もうできているのでは、という気がする。

Q：ロボットの研究で３つのアプローチが融合しなかったのは、タオルを畳むタスクがシンプルだったからではないか。ロボットのタスクとして大変なのは分かるが、そこをマルチタスクなど実世界に持ってくると、かなり融合しないといけなくなると思う。

A：その意味では、組み立てタスクは良いのではないかと思っている。まさに Compositionality や Independence など Consciousness Prior が働く世界で、しかもさまざまなアプローチがとれる。

Q：例えば Embodied Q&A みたいな、実世界の話と言語とつながってきて、対話ができるようになると面白いという話は昔からたくさんあると思うが、言語処理の視点からは、そこで可能になる対話は割と単純なものでしかないという印象がある。事情通ロボットを CPS に入れたときの対話として、どのようなことが可能なのか？

A：まだシステムとして動いたりしているわけではない。対話はこれからである。情報を集めて CPS 化する部分は動いているが、それを使って対話するという部分は動いていない。

Q：どういう対話がここでできると面白いと考えているか？

A：事情通ロボットで考えていたオフィスに関する Q&A は、いろいろなことが聞けてしまうので、タスクとしてはまだ少し難し過ぎるかもしれない。例えば、MIT のグループの最近の研究（"The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision"）では、例えば、「赤い箱の右にあって、青い小さなボールと同じ素材の物はいくつある？」といった、記号的、抽象的な推論ができないと答えられない少し複雑な質問応答が取り上げられている。そのような質問応答や対話ができると面白いと思う。

２．２　深層学習モデルのロボットへの応用

尾形哲也（早稲田大学）

本日は、深層学習（Deep Learning）を通して見えてくる知能をロボットの視点で考え直すという趣旨で、深層学習モデルのロボットへの応用に関して発表する。

我々は、深層学習を利用した「物体折り畳みタスク」に関する研究を産業技術総合研究所（産総研）人工知能研究センターと共同で行っている。タオルを折りたたむロボットの動画は、産総研 YouTube 広報チャンネル[7]で見ることができる（図 2-2-1）。認知科学系もしくはロボットの知能に関して、予測、順モデル、逆モデルといったところも含めて、World Model[8]を考えるというのは必要不可欠であるという大前提に立ち、それを深層学習に拡張したらどうなるのか？という

[7] https://www.youtube.com/watch?v=YH1TrL1q6Po
[8] 米 Google Brain の David Ha 博士とスイス IDSIA の Jürgen Schmidhuber 博士らが提案。

観点から研究を行っている。

図 2-2-1　深層学習を利用した「物体折り畳みタスク」

見ている映像から未来の映像を予測する画像予測を一種のサブタスクとして与えている。本来は関節の動かし方を学習すればいいところを、敢えて少し先の画像も予測せよということも学習タスクに加えることによって、学習に使える教師誤差を増やして、与えなければならないサンプル数を減らしている。予測は、深層学習にマルチシーケンスを埋め込む学習により実現される。我々が使っているものはリカレントネットと言われるニューラルネットワークだが、中にアトラクター（システムの安定状態の軌道を、そのシステムの状態を表す状態変数から成る状態空間で表現したもの）の形でいろいろなシーケンスを埋め込むことができる。その埋め込んだシーケンスを状況に合わせて切り替えることで、設計者の意図通りに出力が得られるような形で学習をさせており、各モーションの構造が設計者から容易に理解できるようになっている。ここで柔らかい物のモデリングは非常に難しいタスクである。しかしながら、柔らかい物体の画像から直接行動を生成するという End-to-End の仕組みがつくれてしまえば、柔らかい物体の部分的な表現が潜在的に深層学習の中に獲得される。

図 2-2-2　ドアを開けて通過するロボット

図 2-2-3　模倣予測学習

「マルチモーダル AI による粉体秤量ロボットに関する研究」（エクサウィザーズとデンソーウェーブとの共同研究）、「マルチモーダル AI による液体秤量ロボットに関する研究」（エクサウィザーズと大成建設との共同研究）等、さまざまな企業と共同研究を行っているが、紙面の都合上、ここでは「ドアを開けて通過するロボット」に関する日立製作所との共同研究の成果を紹介する（図 2-2-2）。開発したロボットは、人間の動きを模倣させる「模倣学習」の技術（図 2-2-3）を取り入れ、ドアを開けて自律的に通過できる。このタスクでは、人間の操作だけをコピー（いわゆる動作軌道だけをコピー）するよりは、その見え方からモーションがどのように生成されているかということに関する、感覚と運動を統合のさせ方（スキル）の学習を実現している。まずドアに近づくというモーションと、開けるというモーションと、通り抜けるというモーションをばらばらに教えている。これはそれぞれの動作を予測学習する複数の深層学習が並列に走っていて、それぞれ少し先の画像を予測している。そして、その画像予測の精度に合わせて、どの深層学習が今ロボットのコントロールをすべきかを切り替えている。ほぼ深層学習だけで実現しており、ドアのモデル、ロボットのモデルはもちろん台車のモデルも必要ない。

多くのロボット研究者は現実の世界にいて、制御対象の運動を微分方程式で解いている。その世界では、同じ状態というのは二度と起こらない。常にダイナミックに変化していて、簡単にカオスが起こる世界である。例えば、２リンクのフリージョイントはカオスになる。このようなカオスを起こさない、もしくは制御する、ように難しいシステムを組まなければならない。一方、情報工学の研究者の考えは、主に記号の世界にある。一つの事象が１か０で表現され、その曖昧性を確率で扱う。この二つの世界をつなぐときに悩む問題が２種類ある。一つは、連続世界の分節化、そしてもう一つが「多対多（many to many）」の問題である。言語の意味は使われる場面で変わる。また、時々刻々と意味が変わっていく場面もある。したがって、立てられたモデルは次に使えるかどうかは保証できない。

多対多の問題の例を挙げる。ロボットがオレンジにタッチするモーションがあったとき、このモーションを説明する言葉は無限にあり得る。touch the orange でも、touch the fruit でも、あるいは hand on that fruit でもいい。つまり、どの言語が与えられたとしても、この運動をするようにしなければならない。逆に、当然だが、一つの言語がいろいろな運動になり得る。touch the orange と言っても、無限にタッチの仕方がある。もちろん無限対無限ではなく、制約はある。ロボットが人間から「あれ取って」とか言われるのは困るし、「これ取って」と言われても困る。な

ぜなら、コンテキストによって「これ」の意味は変わるからである。

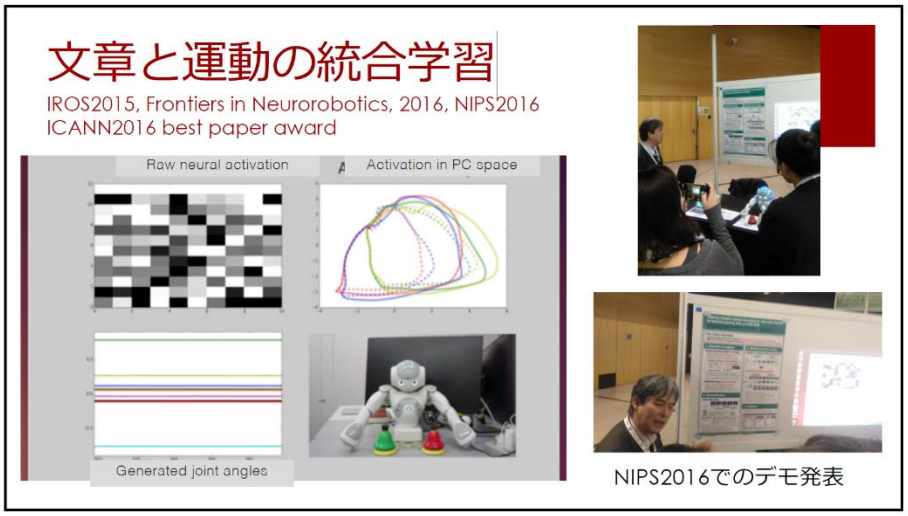

図 2-2-4　文章と運動の統合学習

実際に、ロボットが人間とインタラクションするとき、扱える文章、対話はすごくプリミティブであるにも関わらず、書き切れない。そもそも何が状況に応じた変数なのかも事前には設定しきれない。我々は、そのような問題に挑戦しようと思って、言葉による命令（ベルを指す、ベルを叩く）に従うロボットの学習実験を行った。この命令自体はとても素朴で、「赤を叩け」と言っているだけである。ただ、赤を叩けといっても、その赤が右にあるか左にあるかで使う手を代えなければならないし、両方に赤があったら両方叩かなければならない。赤がなかったら無視しなければならない。赤というものが変数だということを最初に与えていればよいが、もともと何が変数なのか分からない状況にある。このような問題を運動シーケンスと言語シーケンスだけから学習ができないだろうか？と考えて研究したのである。つまり、言語シーケンスと運動シーケンスが、その指し表すものが同じだという例を複数与えて、今まで学習したことがない言語と運動の組み合わせに対しても対応ができるようにならないだろうか？ということを期待した。そして、その成果として、30th Annual Conference on Neural Information Processing Systems（NIPS 2016, Barcelona, Spain）において、「入力された命令の多義性を理解して条件によってベルの叩き方を変えることができるロボット」のデモンストレーション発表を行った。

図 2-2-5 の例では、あるシーケンスに伴う運動やビジョンの変化を再現できるように学習する Action Encoder RNN と、同じ言語を表現する Description Encoder RNN を考えて、この間で潜在変数をシェアして（つまり、お互いの内部表現を近づけるような損失関数を用いて）学習している。このアイデアは、マルチ言語でやられていることからヒントを得て、運動にも適用してみたのである。

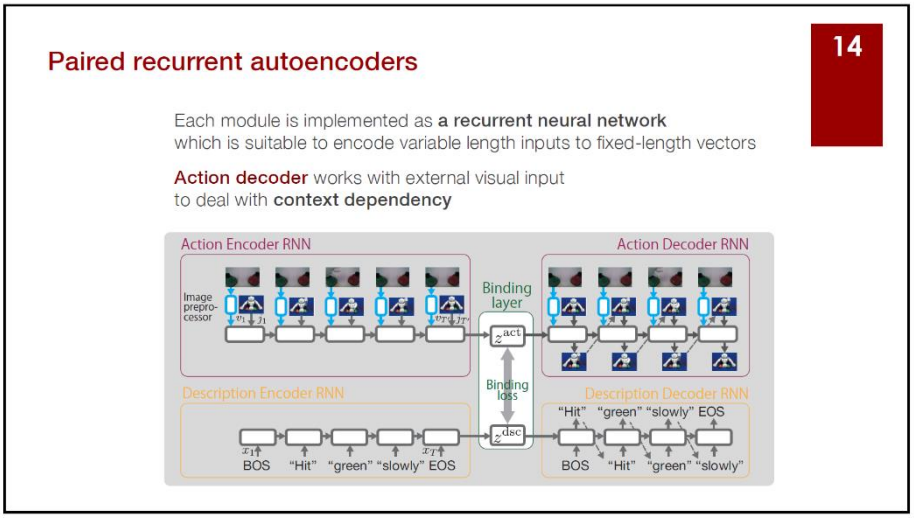

図 2-2-5　対となったリカレントオートエンコーダー

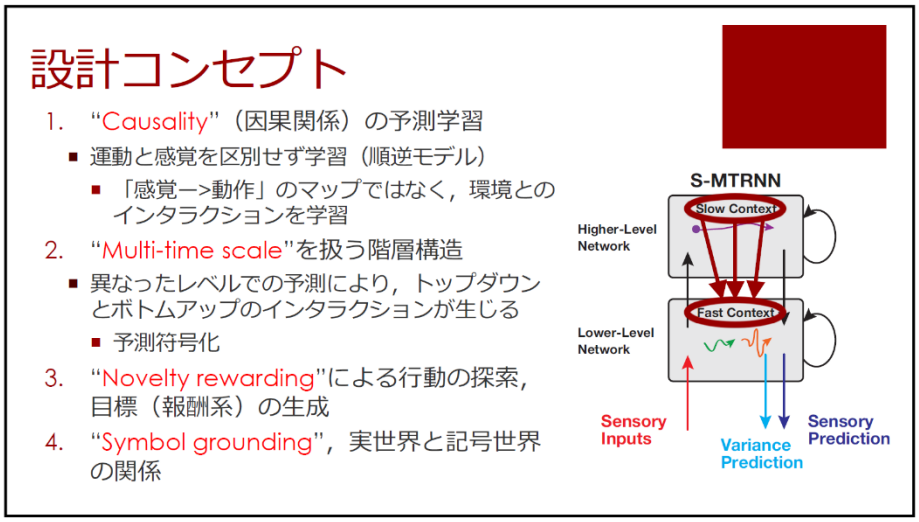

図 2-2-6　モデルの設計コンセプト

「Push the right」とか、「Push the red」とか、すごくシンプルな言葉（短いセンテンス）を与えるとロボットは動き出す。一つの運動が複数の言語で表現できるようになっている。「多対多」を意識し、曖昧性が残るようなシンプルな言葉を与えて学習させている。潜在変数を学習し終わったときの言語側の変数を見ると、一つの文章がさまざまな運動を指し示す、つまり、一つの運動がさまざまな文章に対応することを示唆している。また、運動側に関しても同様で、一つの文章がさまざまな運動を表す。興味深いことに、損失関数のシェアを行わないと、例えば運動だけで学習させてしまうと視覚が無視されてしまい、運動の類似差だけで統合されてしまう。つまり、言語が入ることで、その構造化が視覚に向くのである。運動だけではどうも上手くいかない。このようなことが、言語側でも起こらないかと、今考えている。つまり、言語側の構造化も、運動と統合したことで構造が少し変わるということが起こってくれると面白いと考える。もちろん、「多対多」の性質は、数多ある記号接地問題[9]のほんの一部だと思っている。まさにコンテキスト

[9] 記号接地問題（シンボルグラウンディング問題）については図 1-11 を参照。

やシチュエーションに依存する言葉の意味が、実世界で変わるということを探求してみたい。すぐに実応用にはならない気がするが、そのような基礎研究をやってみたい。

図 2-2-6 に示すように、今行っているモデルの設計コンセプトは 4 つあり、①Causality（因果関係）の予測学習、②Multi-time scale を扱う階層構造、③Novelty rewarding による行動の探索および目標（報酬系）の生成、④シンボルグランディング（実世界と記号世界の関係）である。

システムにとって予測学習しているということは、自分の行為と世界の関係（つまり、こうすればこうなる、もしくはこういうイベントが起こればこうなるということ）を学習しているということである。実際にはどうかは分からないが、システムにとってはそういうことを主観的に思い込むように学習していく、という感覚が重要である。時間（もしくは次元数）のスケールが異なったレベルでの予測を、我々はリカレントネットで学習している。異なったレベルでの予測により、トップダウンとボトムアップのインタラクションが生じるものと期待している。また、Novelty rewarding による行動の探索および目標（報酬系）の生成は重要だと思っている。行動の学習が報酬系だけでできるとは限らない。むしろ逆に行動の探索から報酬系が創発されるプロセスに我々は興味がある。なぜ赤ちゃんは寝返りを打つのか、まさに、Novelty rewarding の概念をロボットにどこまで取り入れられるか、非常に興味を持っている。そして最後に、記号接地問題は、設計コンセプトにおいて極めて重要だと思っている。上記したように「多対多」の問題は、記号接地問題のほんの一部である。まさにコンテキストやシチュエーションに依存する言葉の意味が実世界で変わるということを探求してみたい。

最後に、AI の信頼性に関して考えているところを述べたい。深層学習は 100%は信頼できない。データにはノイズが含まれており、完璧な学習ができない。そのようなことを考えると、シンボル系と統合するやり方としては、いくつか段階があると思う。

図 2-2-7　AI の信頼性について

図 2-2-7 に示すように、第 1 レベルとしては従来の AI による抑制がある。例えば、我々のタオルを畳むロボットの場合には、バーチャルバリアを導入している。つまり、タオルを畳むタスクは信用するが、少しでもおかしなことをやり始めたら止める。AI が明らかにおかしいことをやり始めた場合、深層学習に対して比較的シンプルな AI が介入する仕組みは当然つくれる。第 2 レベルとしては、深層学習が獲得する表現を理解することである。深層学習モデルの中に獲得される潜在表現は非常に不連続で複雑な構造になっており、それが深層学習の能力を高めているとも

言えるが、その仕組みを理解するための研究は当然ながら非常に重要である。最近では、今泉允聡氏（統計数理研究所）をはじめ、いろいろな研究者により行われている。そして、第３のレベルとして、学習の方法に身体知の視点を取り込み、人間に近い学習のさせ方をさせることで信頼性を確保する方法があり得る。例えば、認識の問題において目の前のパソコンを「コンピューター」とラベル付けすると、それ以外の意味をとることができない。しかし実際は、コンピューターになるのか、パワーポイントになるのか、それともキーボードになるのかは簡単には決まらない。これはもちろん確率で決まるわけでもない。「身体を持った主体」がどういうコンテキストで、どういう行動目的で見るかによって意味が立ち現れてくる。環境と身体を持ってインタラクションできるシステム（ロボット）を考え、あるコンテキストで物はどう見えるのか？を探求することが必要であると考える。学習すべきデータをトップダウン的にラベル付けする、もしくは確率を与えるのではなく、システムが行動に必要な範囲で経験（Experience base）として獲得されるべきものだと私は考えている。そういう意味で、改めて「身体知」を考えなければならない。

質問応答

Q：「ドアを開けて通過するロボット」のタスクにおいて、少し先の画像を予測するという話があったが、予測するものは画像がいいのか？何か別の World Model とか、何かの表現（Representation）を予測してもいいような気もするが…。

A：はい、その通り。

Q：画像を予測することの意味をどのように考えたらよいか？

A：取得できる感覚を全て予測するのが理想だと思う。そういう意味では、主体にとって世界を知覚できるものを予測していく。今の場合だと、ターゲットが画像から運動を生成するということがターゲットなので、運動の他には画像が分かりやすい学習対象になる。それを予測対象にすると誤差逆伝播が楽になる。

Q：なるほど、それが常にできるわけか？

A：できるだけ一番大きく情報を得ることが大事だ。力覚や触覚を入れてもすごく効果がある。画像のみである必要はない。また予測する画像の解像度は、それほど高くなくても学習としては十分に働く。

Q：言葉による命令（ベルを指す、ベルを叩く）をロボットが命令通りに実行し動作することは、ある種、翻訳であると思う。翻訳のアーキテクチャは、いろいろ考えられるが、基本的には言語シーケンスと運動シーケンスの、違うモダリティーのシーケンス間の翻訳であると言える。もし世の中に結構ある画像や言語のデータを全部放り込んだらどういうことが起こりそうか？

A：まずは、ロボットが必要な言語を学習できればいいと考えている。つまり、このロボットは物体に対して指さすことと叩くことしかできないという前提の中で、そのバラエティーを表現できることを徐々に増やしていこうという発想である。そうすると全く新しい単語が、似たような意味の違う単語として入ってきても解釈できる可能性が出てくる。動画を事前学習させたものと言語を事前学習させたものを準備して、それら両方を統合するやり方はあり得ると考えており、認知発達ロボティクスの観点から探求したい。ただ、すぐ応用せよと言われると、それはちょっと現実的には…。

Q：ドアを開けて通るロボットのときに、動作に分解して、それぞれにネットワークを学習さ

せていたが、そのようなレベルから一歩進んで、わざわざ分解しないでも動作ごとに学習できるような汎化は大変と思う…。理想的にはどのような形が向かうべき方向なのか？何か一つのネットワークでいろいろ動作を学習するようになるのか？それとも、ある程度抽象化された動作が有限個にカテゴライズされるのか？

A： 大変難しい質問だ。人間は構造の分節化そのものが（切れているか切れていないかも含めて）曖昧である（これも記号接地の重要課題)。連続的な運動に対して、言語のどこが再使用可能か？どこが組み合わせ可能か？どこが変形するだけなのか？そのようなところまで探求しようとすると、まさにメタラーニング（メタ学習）の研究領域に入ってくると思っている。これは簡単ではないというのが正直な気持ちで、いろいろなアイデアを持って私の研究室の学生も挑戦しているが、基礎研究になってしまうのが現状である。ばらばらなニューラルネットワークでやるのか、１個のニューラルネットワークで埋め込むのかというのも、昔からの局所表現・分散表現論争の議論がある。局所表現でやるとすごく安定する。学習パターンが、別々のニューラルネットに学習されるので分かりやすいしデザインもしやすい。そのかわり、汎化という点では、今の組み合わせ以上の汎化は出てこないことになる。１つのニューラルネットワークにいろいろなパターンを埋め込む分散表現だと、創造性のある行動が出てくる可能性がある。制御する側からすると邪魔かもしれないが、非常に面白い表現が出てくると思う。

Q： 言語は単語のシーケンスであるので、大量に言語データのシーケンスを学習していくと、それなりに何か塊のようなもの（要するに、ある種のチャンク）ができて、それから意味的なチャンクになって、結果的に表現のようなものが学習できるのではないか？

A： それは、今まさに研究している最中である。

Q： 赤ちゃんは言葉が分かっていなくても、何かにスベスベするような動作をやっているように、アクションだけ、あるいは動画だけで、ある種の分節みたいなものが生データだけで学習できて、それで良い表現をある程度つくるようなことは考えられるのか？

A： やはり、そこでも「多対多」のようなことが起こると思っている。今、動画の分節化の研究（コンテスト）がたくさんあるが、答え（教師信号）がそもそも安定していないという問題点がある。ラベルをつける人間によって、名前がまず変わるし、どこで切れるかも変わったりする。

Q： 言語は、別にラベル付けは全然しなくても……。

A： そこは大きい。予測、教師なしでずっとやっていると構造化できると考えている。

Q： 教師なしでも、ひたすらやる？

A： ある種、表現は、すごく刹那的なものだと思っている。ある世界では使えるが、別の世界だと全く別の意味になってしまうようなことが起る方が、何か本当のような気がしている。例えばモジュールという言葉も、あまり固定して考えてしまうと面白くない。むしろ、いろいろな形に変容できる元の表現があって、それがコンテキストに依存して意味が変わる(内部構造のレベルでも変わる)。今使っているパソコンをコンピューターと認識するか、パワーポイントと認識するか、キーボードと認識するか、どういう行動目的で見るかによって、その意味構造が変わるようなことが起こるかもしれない…。

Q： 触覚を実装したロボットは、触覚と画像の両方を予測学習しているのか？

A： はい。現在のロボット業界では、むしろ触覚・力覚の方がマジョリティーになっている。

なぜなら、画像と音声に関して（そして自然言語も）研究されてしまっているからである。カナダ・モントリオールで 2019 年 5 月 20 日-24 日に開催されたロボットのトップカンファレンスである ICRA（The International Conference on Robotics and Automation）2019 において、ベストペーパーアワードを受賞したのは Michelle A. Lee ら（Stanford 大学）による ”Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks”という論文題目の、力覚に関する発表であった。そして、彼女たちに惜しくも敗れてベストファイナリストになったのが、”Deep Visuo-Tactile Learning: Estimation of Tactile Properties from Images”と題して触覚に関する発表をした Kuniyuki Takahashi & Jethro Tan（Preferred Networks）であった。触覚・力覚というのは、まさに身体経験と結びついたものであり、それを深層学習でやることによってロボットの運動がよくなる。学会では、そのようなことにだんだん興味が移ってきつつある。視覚をさまざまなモダリティーときちんと統合できることがすごく大事であると私は思っている。

Q：彼らも、本日の発表にあった予測学習と同じ枠組みで学習している？

A：彼らは予測学習も含んでいるが、強化学習などとも組み合わせている。ただ、身体を持つということについて考えるときに、深層学習に予測学習という考え方を入れるのは大事であると考えている。

2．3　自然言語理解への道のり

乾健太郎（東北大学）

東北大学で自然言語処理の研究をしている。本日は、言語の中に比較的閉じて言語処理がどのようになっているかという話をする。自然言語理解は非常に長い歴史を持つ研究分野だが、深層学習の躍進も含めて大きな進展がいろいろ出てきている。

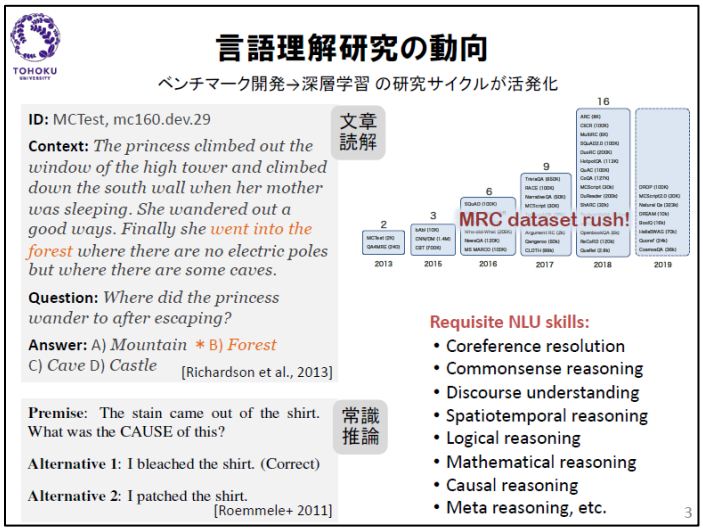

図 2-3-1　言語理解研究の動向

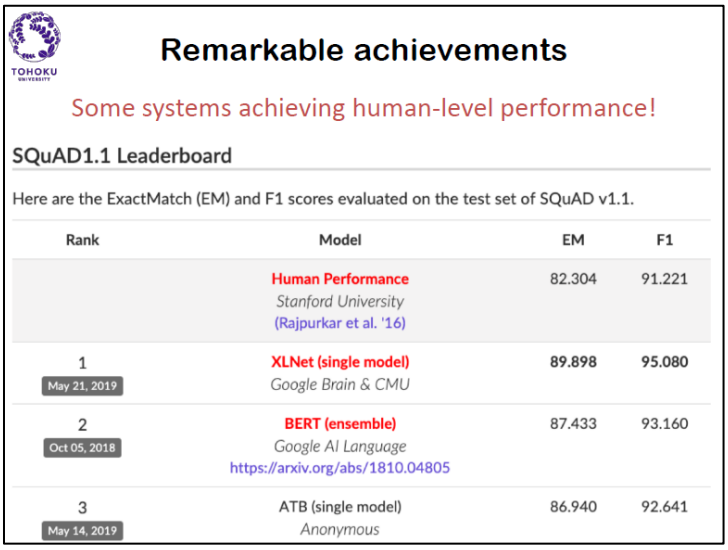

図 2-3-2　言語理解の成果

一つのシンボリックなものは機械翻訳。もう一つのシンボリックなものは、言語理解のベンチマーク。文章を与え、それに対する質問に答えさせるというもの。質問に答えられたら、国語や英語の問題が解けることと同様に言語が大体分かっているだろうということ。このための大量のデータが作られている。今は、クラウドソーシングがあるため、世界中の人にこういう問題を作ってもらい、それを解いてもらっている。人間が確認した問題だけを解かせており、これによりさまざまな言語のデータ、テスト、ベンチマークのデータができてきている。

すばらしい結果も出始めている。例えば、図 2-3-1 は文章とそれに対する質問が入っているデータセットで、人の精度に対して、場合によってはシステムがそのパフォーマンスを超えている（図 2-3-2）。これをもって、「もう読解問題は終わりました。人間を超えています」というような記事が本当に出回ったりする。しかし、全くそういうところまで到達していない。まだまだできていないことのほうがはるかに多いと思っている。

これは本当に分かって当たっているわけではない。例えば、図 2-3-3 は実際の Machine Reading Comprehension のものだが、本当に Language Understanding のスキルを測っているかというと、そうではないと言う人が少しずつ出てきている。我々は、この中では一番 comprehensive に実際のデータセットを見て、そうではないということを強く言っている。

図 2-3-3 では、あるシステムが質問に対する答え（John Elway）をうまく当てられている。しかし、そこに何の関係もない文（青字）を加えると全然違う答え（Jeff Dean）をしている。加えられた文章が全然関係ないということも分かっていないし、こういう Destructor が出てくると容易に間違ってしまう。つまり、本当に分かって回答しているのではないことが簡単に分かる。

こうした背景があり、言語理解はかなりできているという話もあるが、実はまだまだである。本質的にできていないのは、本当に言語を理解するということ。それはすごく単純に言うと図 2-3-4 に示す古典的な例でシンボリックに表現されている。何か文があるとその間は意味的な関係でつながっていて、単にばらばらな文が並んでいるわけではない。

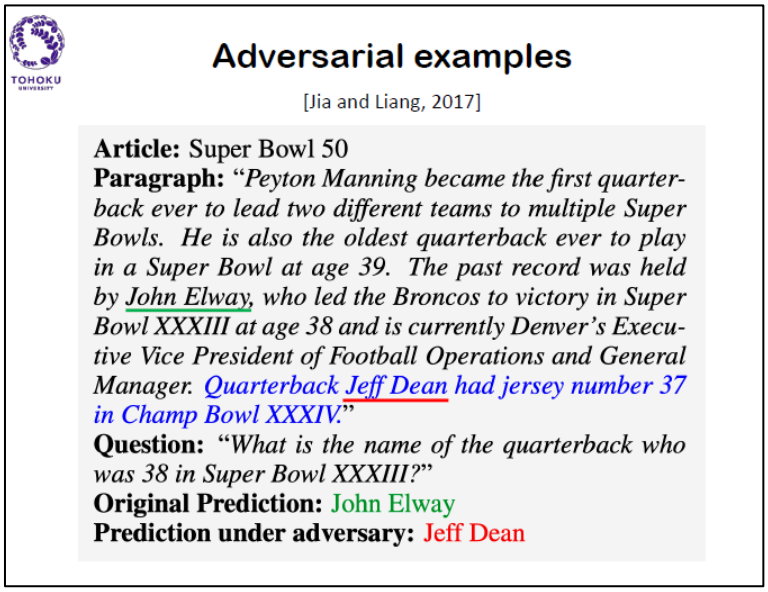

図 2-3-3　Adversarial Examples

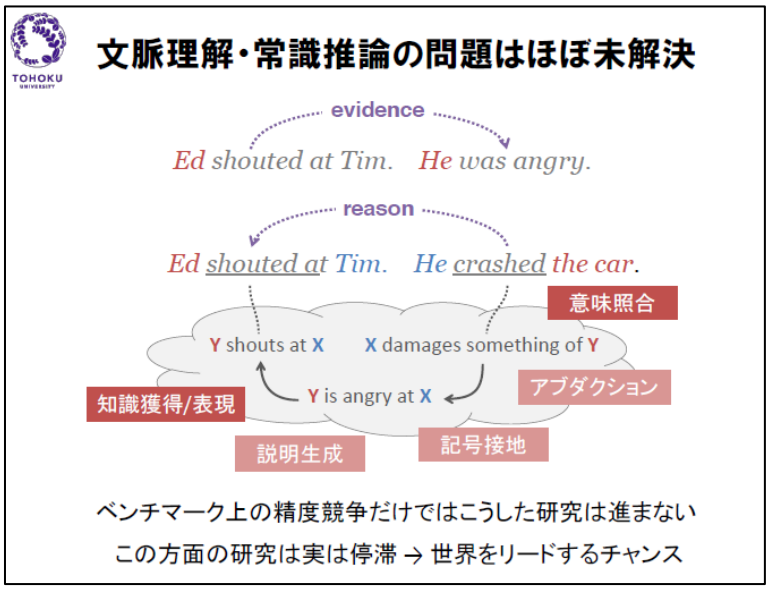

図 2-3-4 文脈理解・常識推論の問題

図 2-3-4 に示す 2 つのイベントの間に我々は関係を見いだす。1 文目の He は Ed であり、2 文目の He は Tim であろう。2 文目の He は、車を壊したために怒られたほうの Tim であり、この車の持ち主は Ed だろうと、我々は推測しながら読める。こうした推測ができるのは、頭の中でどう表現されているかは分からないが、何らかの知識を持っているからだろう。例えば、誰かのものに損害を与えてしまえば､与えられた人は怒るだろうし､怒ったらどなるだろうということ。

こうしたある種の世界に関する知識を、恐らく我々はいろいろ持っていて、それを必要に応じて想起しながら、組み合わせて、何らかの推論をして、理解した気になっているのだろう。その副産物として、これがこれの理由を言っているとか、これは Ed ではなく Tim を指しているということが分かることが、言語を理解するということの一つの形である。これは 1980 年代、1990 年代からずっと言われていることだが、実際に先ほどの言語文書読解の問題で、こういうことは

全くできていない。これをできるようにするためには知識を集めないといけないし、知識がある種シンボルで書かれているとすると、シンボル間の非常にセマンティックなマッチングもしないといけないだろう。必要なものをうまく retrieve して、あるいは組み合わせて、場合によっては abductive に推論しなくてはならないが、まだまだできていない。

ベンチマーク上の単純な精度競争だけでは、こうした研究は進まない。実際にはこうした研究は 1980 年代から言われているが、深層学習の時代になっても、なかなか進んでいかない。

1990 年に Interpretation as Abduction ということを言った Jerry Hobbs という計算言語学の方など、いろいろな話があるが、我々のチームもこういうことをしっかり取り組まないといけないと思っている。知識を獲得するとか Reasoning をやるということを、それぞれのコンポーネントはいろいろ研究している。もちろん、これは我々だけでなく、さまざまな言語処理の研究チームが知識獲得やある種の推論の重要性に言及している。それから、先ほどからずっと出てきている深層学習に基づく Representation Learning、あるいは深層学習全体でこういうものを組み合わせていくということが重要だという個別の研究が今盛んになされている。

個別の研究はなされているが、知識獲得や Reasoning をどのように進めていけばいいのかが、大きな問題である。なかなかそれを進めるドライビングフォースがない。

まずは、先ほど麻生先生の発表でもタスクが重要という話があったが、言語処理でもこれを考えるのに、いいタスクをうまく考える必要がある。いいタスクがあればその研究が進むし、場合によっては、いいアプリケーションが開けていくだろう。そこにはできれば、先ほどのようなさまざまなスキルが必要なタスクであって、かつ Interpretability、Explainability が求められるようなタスクがいいだろう。

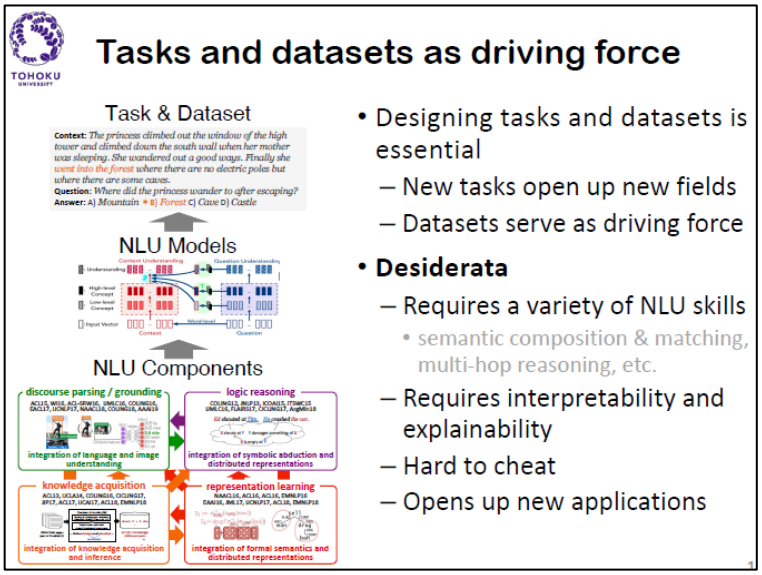

図 2-3-5　Tasks and datasets as driving force

そう考えたときに、いろいろなタスクが考えられるが、我々自身、一つは Machine Reading Comprehension がやはり非常に重要なタスクと考えている。もう一つは、あまり言っている人はいないが、アセスメントが非常にいいタスクと考えている（図 2-3-6）。前者の Machine Reading Comprehension は、先ほど述べたようにデータはたくさんあるが、現在うまくドライブできてお

らず、どうすればいいか考えなければいけない。後者のアセスメントでは、これはある種の赤ペン先生の AI 版のようなものをつくることを考えている。これは言語の理解を本当にしないといけない、かつ全く無理ということでもなさそうな、いいタスクと思っている。あまり振り向いてもらえていないので、いろいろなところで少しずつ宣伝をしていきたいと思っている。

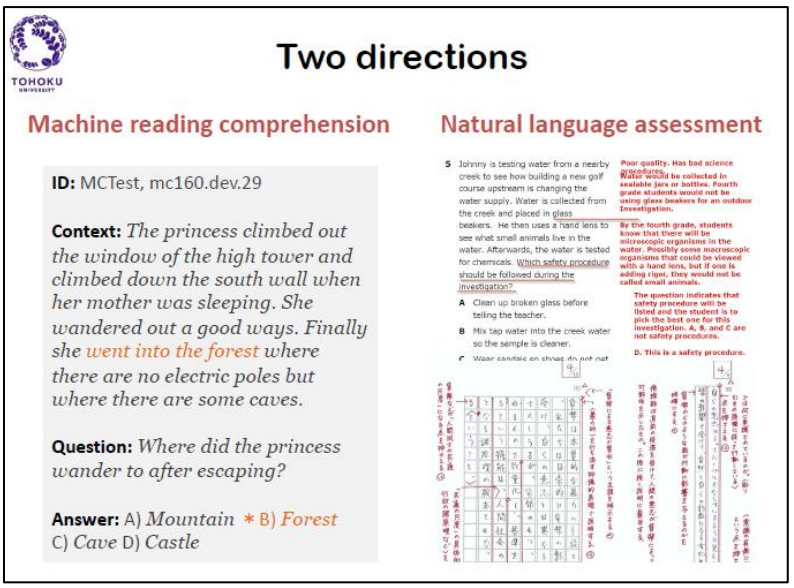

図 2-3-6　2 つの方向性

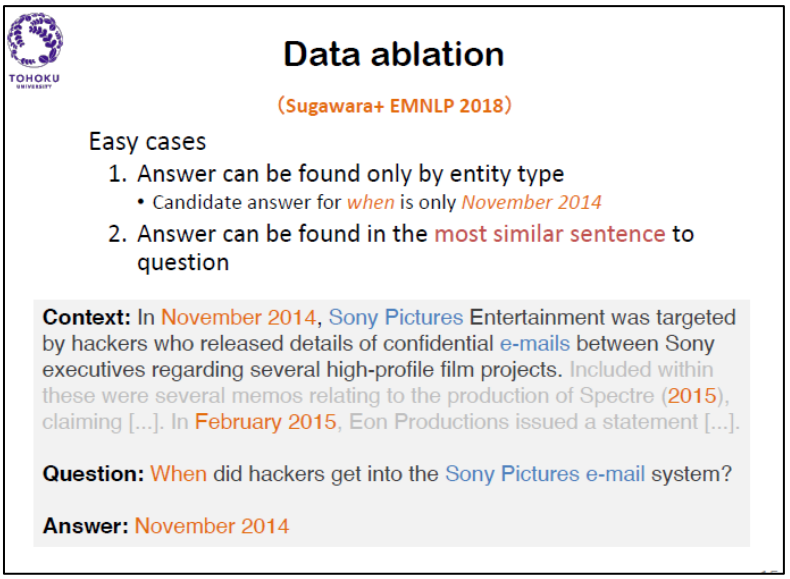

図 2-3-7　Data ablation

まず、Machine Reading Comprehension のほうは、たくさんのデータがあるので、それをどううまく再利用するかが重要と考えていて、いくつかのことに取り組んでいる。1 つは Ablation。モデルの Ablation はよくされているが、データの Ablation はまだ極めて少ない。しかし、これが非常に重要である。問題の大部分を ablate して、例えば When だけで問題が解けるのであれば、それはこの問題は簡単過ぎるとか、あるいは、ワードの重なりだけを見て一番似ている文の

中に、When に相当するものは 1 個しかなかったら簡単に解けるだろうなど、意地悪しても勝手に解けてしまうものは、いい問題ではないということで、フィルタリングアウトするようなことをする。

実際に今まであるデータセットを横断的に見ているが、今のようなことで解けてしまう問題は簡単な問題で、そうでない問題はいい問題である可能性が高い。今のデータセットは、簡単な問題の割合がかなり大きく、何か変なトリックを使って、先ほどの人間と同じようなパフォーマンスが出せているということが分かってくる。

あるいは、さらに意地悪をして、例えば文章と質問を全部アノニマイズしてしまう。無茶苦茶だが、reputedly や appeared to のような単語の意味を全く教えずに、言語理解と関係ない手がかりだけで解かさせても、かなり解けてしまうことが分かっている。例えば図 2-3-8 のようなデータセットでも 60%解けてしまう。このため、こうした問題をちゃんとフィルタリングして、いい問題にフォーカスして、実際に研究を進めていくことが必要だろう。

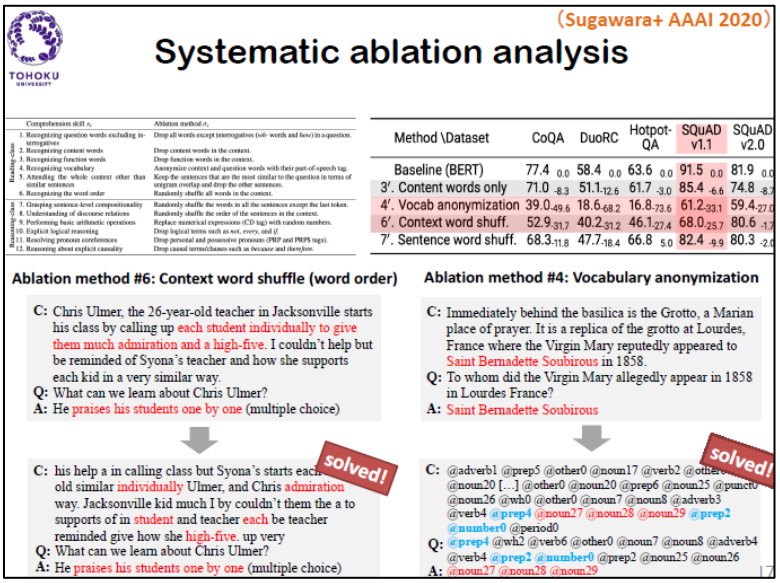

図 2-3-8　Systematic ablation analysis

あるいは、こうした問題に対して関連する記事から答えを見つけてきなさいというような問題が過去のデータセットにもある。どういう推論をしてこの答えに至るのか、この答えでないとおかしいというために必要十分な推論は何なのかということを明示的に人間に作らせて、これを当てさせる。このように「推論する問題」を設計していく必要があるだろう。クラウドソーシングを使い、うまく設計すればいい問題ができて、こちらのほうに少しでもドライブすることができるのではないか。

もう一つは、アセスメント。本当に赤ペン先生のようなものを作れると、社会的には大きなインパクトがある。人間が書いたものをアセスメントするということは、例えば How to say や、What to say のレベルでアセスメントするということ。最終的にはディベートしているものを見て、アセスメントして、それに対して何か論評するとか、あるいはディベートの相手をするとか、そうしたことができるような技術を作っていきたいと思っている。

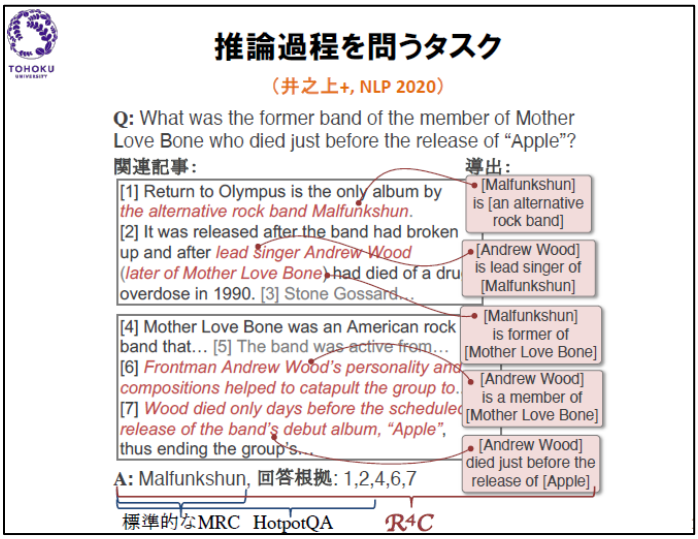

図 2-3-9　推論過程を問うタスク

これは教育に使えるので、社会的なインパクトがあることはほとんど自明と思うが、これができるためには、そもそも Interpretability、Explainability が本質的に必要。アセスメントした結果を説明できないと教育的には使えない。また、生徒が書くものであり、不完全な文章であるため、それに対する頑健性も必要。さらに、生徒が書く文章に対して、この人は何が言いたいのか、どういうことを言おうとしているかを不完全な文章からも推論できるようにしたい。それにはある種の世界知識を持ち、それを使って推論し、この人は何が言いたいのかということについて、Tim と Ed の例のような推論ができないといけない。

これをできるようにするタスクは、非常に面白いと思っており、我々東北大のチーム、あるいは理研の AIP の自然言語理解チームではメインに展開しようとしている。これは非常に面白いタスクだが、データがないことが大問題である。データがないので、研究者がみんなやりたがらず、なかなか研究が進んでいない。逆に言えば、データを用意すれば、手に入れれば、世界をリードできる可能性があると思っている。

こういうタスクも重要だが、そこに共通して必要なのは、先ほどの議論でも出てきている Interpretability、Explainability である。これを本当にやっていく必要があると思っている。言語でも、当然、解釈性、説明性というのは非常に重要。例えば翻訳する際、本当に正しいかどうかという確証が欲しいとか、対話システムが変なことを言い始め時に、どうしたらいいのかとかいうようなことが必要である。画像の場合は、この辺を見て猫だと見ているといった、ある種のアテンションのようなもので、説明された気になる。しかし、言語の場合は、この辺がポジティブだから全体がポジティブだと判断したと言われても、当たり前過ぎるような気がする。画像の認識タスクで必要な説明と、言語でもう少し理解に踏み込んだところで必要とされる説明というのは、恐らくレイヤーが違い、求められるものも違うのだろう。

恐らく重要なのは、言語レイヤーで納得感のある説明をするには、福島さんの話にもあったが、AI の情報処理を人間の知識に紐付けて説明できることだろう。このため、人間が持つ常識や専門知識を深層学習にどうやって教え込むかが一つの大きな課題であると考える。

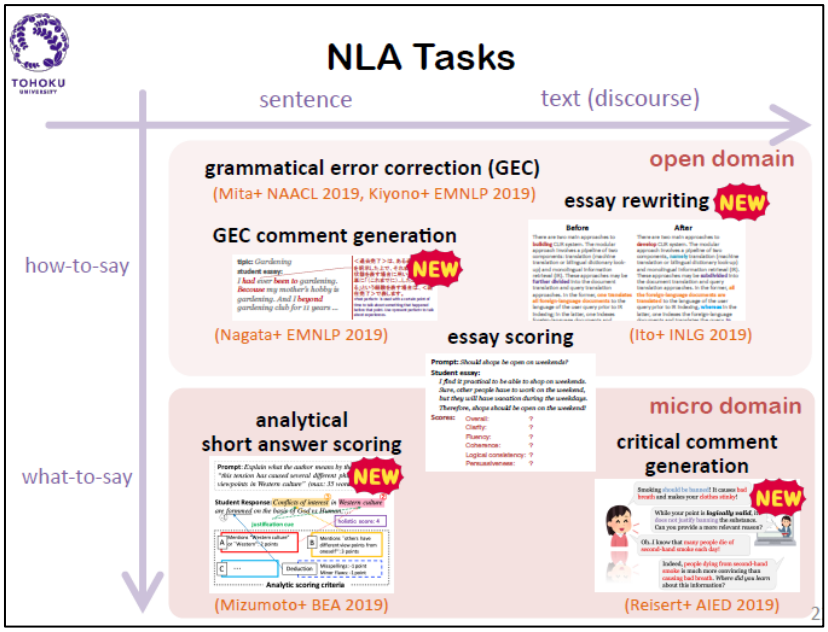

図 2-3-10　自然言語アセスメントタスク

方法は本当にいろいろあると思うが、今、1 つ我々でできかかっていることを話す。まず言語は、例えば科学技術論文でも文書が山のようにある。それから、場合によっては知識ベースがある程度整備されている分野もある。知識ベースは人間がある種一般化した知識である。一方の言語データはインスタンスが本当に山のようにある。このインスタンスはラベルがついていないかもしれないが、BERT などの言語モデルによって教師なしでどういう単語がどんな意味を持っていそうかを学習することができる。

実際には、テキストと知識ベースを統合したようなグラフとして考える。テキストのほうがはるかにたくさんの情報があるが、その中に一部、ナレッジグラフのような知識ベースからの情報も入っている。これ自身を全てニューラルネットの世界、一つの共通の空間に embed する。これはいろいろな方法があり、我々は 3 つぐらい考えているが、とにかく共通の空間に持っていく。そうすると、例えばアルツハイマー病に対して、それに効きそうな薬をこの空間からとってくること（即応的な推論）ができるようになってくる。これは先ほどからの議論で考えると、System 1 の推論である。

シンボルの世界に推論の結果をもう一回戻すことで、結果の説明ができる。そういうことで、知識ベースの情報の上にテキストの情報を大量に重ねて、一緒に推論をすると、知識ベース上の推論だけではなかなかできなかった推論が、テキストの情報が手に入り、精度的にもブーストするということがある。

図 2-3-11 でアルツハイマー病にアポモルフィンが効きそうだということが、図の右下のニューラルネットの世界で直感的に分かったときに、その判断の理由として、どういう説明がよさそうかということ自身も、図の左下のニューラルネットで生成させる。テキストとナレッジグラフの情報を使いながら説明を生成するような、そういうモデルをつくることもできる。

そうすると、ニューラルネットで即応的な推論をしながら、熟考的な推論をすることで、シンボリックな説明ができる。この 2 つのニューラルネットを同時に学習することもできるので、即応的な推論の精度をずっと上げながら、説明もすることができるということが実際に可能になり

つつある。

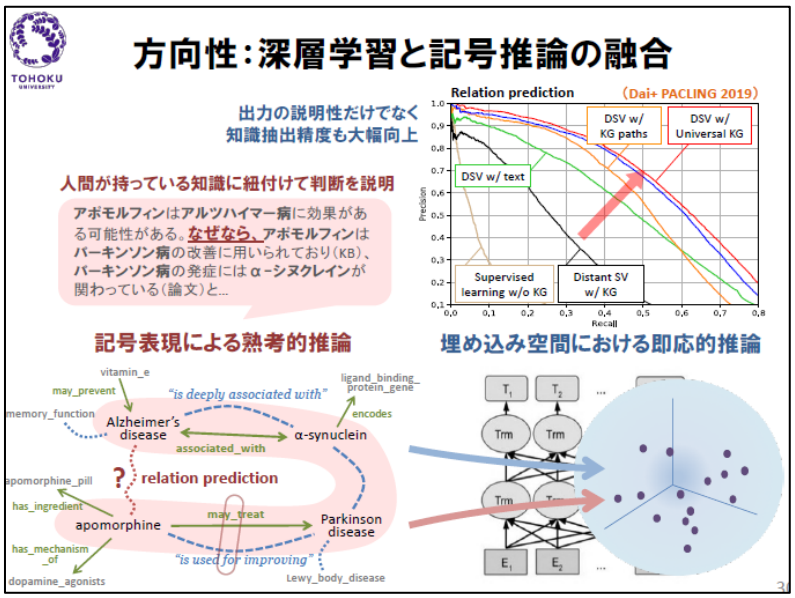

図 2-3-11　深層学習と記号推論の融合

これが、言語におけるニューラルネットの世界とシンボルの世界をどのようにインテグレートして、精度を上げながら、かつ説明ができるようにするかということの一つの解ではないかと考えているが、もちろんここから先のいろいろな発展もあると思っている。

こうした発展を実際の文章に対して、関連する言説をたくさん集めてきて、それに関連するある種の知識ベースも集めてきて、それを共通の空間に入れて、その間で推論させながら、どの言説とどの言説が同じようなことを言っている、あるいは対立することを言っているということを判断するような、そういう世界をつくることができるのではないかと思っている。

そのため、最初に福島さんの話もあったが、こうしたシンボルの世界と深層学習の世界を、まさに System 1 の世界と System 2 の世界を統合するようなことが、言説空間の集約とか、あるいはファクトチェッキングとか、そうしたところに実際に応用として出せていけると面白いと考えている。

質問応答

Q：冒頭の読解問題は、人間の読解力が低いという事か？

A：人間の読解力はそんなに悪くないと思うが、ケアレスミスもあるし、間違える。機械ももちろん間違えるが、例えば図 2-3-1 の 4 択問題では、チャンスレートは 25%だが、正答率はこれを大きく上回っている。

Q：たくさんある自然言語から、まさに知識ベースのようなものができるという方向性もあると思うが、そういった研究はどのくらいあるのか。

A：ここでやっていることはまさにそれに相当すると思う。テキストの中の一部は知識ベースに既にあることが書かれている。それらを対応づけることによって、この種の関係はどういう書き方がテキストでなされるのかということを学習することができる。

そうすると、他にもいろいろ似たような言い方で、他の知識のインスタンスがテキストには山のように書かれているので、それをこの中に取り入れると、全体を一つの空間に embed するということは、知識ベースに書かれていない知識をその空間の中に取り込むということであり、それはナレッジグラフの中にあまたの知識を吸収するということでもある。

それをシンボリックに、これはこの関係ですかといった問題をエクスプリシットに解こうとすると、どうしてもノイズがのってしまう。それから全体最適みたいなことがなかなか難しくなる。それを一つの空間に全部放り込むと。そこには単純でない、いろいろな技術的なテクニックが必要で、これを BERT でやろうと思っても単純にはできない。BERT のようなものにいくつかの技術的テクニックを入れないとこういうことはできない。けれども、そのヒントのようなものは今、手に入りつつあり実際先ほどのようなことができつつある。

Q：今まで人間が持っていたナレッジベースと矛盾するような構造があるということはないか。それでコンフリクトが起き、一緒に入れようと思っても入らないとか、知識ベースそのものを書き換えたほうがよいとかいうこともあり得るのではないか。

A：実際には矛盾することもたくさん書かれている。そのため、先ほど最後に申したような言説を集め、それで矛盾するようなことも認識しながら、知識ベースとつなげていくことをやっていかなければいけない。

今、矛盾する知識を同じ embedding の空間にどう入れればいいかというのは、エクスプリシットには何もできていない。そこをしっかり表現する仕組みが必要だと思っている。今、このリンクが非常にやわらかいので、少々矛盾するようなものが入っていても、システムとして破綻はしない。しかし、恐らくそこでは変なねじれが起きていて、このままだと本当にはうまくいっていないというところがいくつも出てきているのではないかと思う。そのため、そういうことを解消して、最後に申したファクトチェックや言説の集合を集めてつなげていけると面白いと考えている。

Q：ドメイン依存はどれくらいあるか。同じ言語でもこのドメイン、このデータベースで見る解釈がいろいろありそうだがどうか。

A：言語も知識ベースも、今のこの中で、人間がラベリングをするようなものは何も入っていない。知識ベースがあって、言語の生のデータがあれば、基本的にはこういうことができるということで、ドメインごとにつくってもいいし、あるいはどこかでマルチタスクでやって、共通に何かリプリゼンテーションを学習させるようなこともできるかもしれない。

Q：発表の中で、怒っているから叫んでいるだろうという推論と、図 2-3-11 に書かれていたような話で、ギャップがあるように見える。

A：2 つの間にはまだギャップがある。できていることをもう少し精緻に確認できるような、Reading Comprehension なり、Reasoning Dataset を作る必要がある。しかし、既存のデータセットは山のようにあるので、それをうまく再利用することで、そちらの世界にドライブできるようなデータセットになるのではないかと考えている。

先ほどの推論ともちろん違うが、図 2-3-11 は、アポモルフィンとパーキンソン病とα-シヌクレインの間にはある種の関係があり、一部はテキストで書かれていて、一部は知識ベースになっている。このある種 causal chain を、ここで推論をしながら、これとこれの間

に何か関係があるのではないかということを推論している。それは先ほどの例では、怒ったらどなるといった、そういう推論チェーンの一歩にはなっているのではないかと思う。それがテキストの生文とうまく対応させることができると、スケーラビリティーも出てくるし、パラフレーズのようなものにも対応できるようになってくる。その上で即応的な推論と熟考的な説明ができるようになってくると、何かそちらのほうに一歩進めるように思う。

Q：矛盾に関しては、確率化すればいいと思っている。データがたくさんあり、それを高速にサーチできれば、ハイヤーオーダーは要らないのではないかとも思うがどうか。

A：ハイヤーオーダーは要ると思っている。しかし、それをロジックで言うところのハイヤーオーダーのような話がどこまで要るのか分からない。こういう世界でそれをうまく embed することは難しい。ただ、知識ベースの上で、ある種のイベントの上位下位のような話はある。そのため、ある種の一般化が知識ベースの中でもあって、テキストの中でも一般的な記述と具体的な記述はある。そのぐらいであれば対応はできるのではないかと思う。しかし、もう一つは、これは全部命題論理の話なので、変数が出てくると大問題だと思う。

Q：テキストの世界だけだと、どうしても中国語の部屋的な印象が出てくる。画像や動画など、それこそリアルワールドのデータとグラウンディングするということはどう考えているか。

A：実世界とのグラウンディングは、私自身も大変関心がある。生文にはありとあらゆることが書かれている。その生文をこういう割とかっちりした世界とつなげていきたい。このやり方はスケールする。また、生文と画像やほかのモダリティーとの対応関係は、それなりにデータがあればつく可能性も出てきている。そのため、生文をいい媒介にして、かっちりした世界と実世界をつなげていくというのは、悪くないのではないかなと思っている。

Q：人間を考えると、即応的に考えて、後で理由をゆっくり考えるといった行き来をする。先ほどの質問では、個別の事象から新たなナレッジグラフができるのではないかと言われたと思うが、逆に、個別の事象でこうだと思って、よくよく考えたら違うぞというので、即応的な方の修正を人間はすると思う。そういった可能性はどうか。

A：いろいろ考えられると思うが、今はできていない。今はむしろ逆で、深層学習で結構精度が上がるので、その精度と、シンボリックな方で説明しようとすることが乖離してしまうことが大きな問題であることの方が多い。つまり、精度を十分に保ったまま、しっかりシンボルも使うということが、なかなか両立できないことがまずは問題だと思う。そのため、今はむしろ逆で、シンボリックに推論することと、それから深層学習の世界だけで推論することを、なるべく共通の結果が出るように両方学習をするので、これが謀反を起こすようなことはむしろしないようにする。しかし、そこはまだまだいろいろ考えられると思う。

Q：Message Understanding の間違いの比較があったが、間違った内容は、機械も人も同じようなところで間違うのか、それとも間違いのパターンは機械とは違うのか。

A：それを我々は調べていないし、まだあまり調べていないと思うがおそらく違うと思う。機械はそもそも分かっていなくて、全然人間と違う手がかりを使って解いている。入試で問題文を見ずに、答えの部分だけ見て当てられることと同じようなことを機械が今やっていると思う。

2．4　知能の基本構成素としてのサイクル＝価値＝意味

橋田浩一（東京大学）

前半でそもそも意味とは何かについて話し、後半でメインの話をしたい（図 2-4-1）。

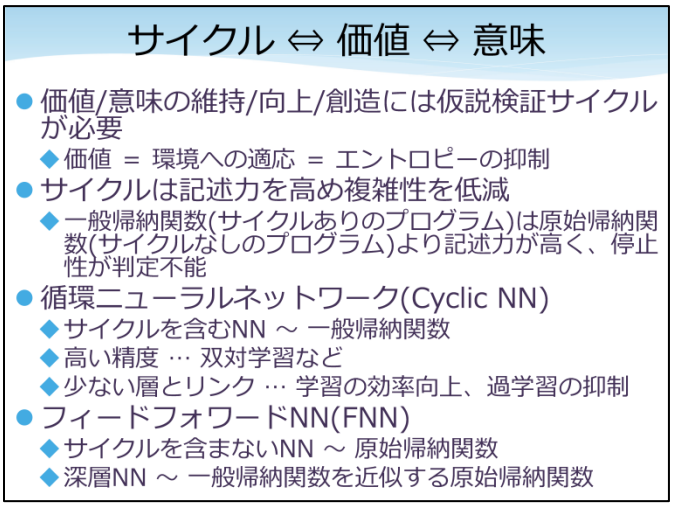

図 2-4-1 サイクル ⇔ 価値 ⇔ 意味

生物というのは熱力学的なエントロピーを抑制するものであり、知能というのは恐らく情報論的なエントロピーを抑制するものである。エントロピーの抑制というのは、環境への適応で、すなわち価値であり、それが意味でもある。価値・意味を実装するのは仮説検証サイクル、フィードバックループだというのが出発点である。

もう一つは、システムがサイクルを含むと記述力が本質的に高まる。サイクルなしのプログラムは原始帰納関数と呼ばれ、サイクルを含む一般的なプログラムは一般帰納関数と呼ばれる。一般帰納関数は原始帰納関数よりも記述力が高く、一般帰納関数が停止するかどうかを判定するアルゴリズムは存在しない。これは、サイクルがあると非常に複雑なことができるということである。

サイクルを含むニューラルネットはさほど多くないが、それらは一般帰納関数のようなものである。双対学習などでは、サイクルを含むことによって精度が上がっている。また、サイクルによって記述力が高まれば、層の数やリンクの本数を減らすことができ、学習の効率の向上やオーバーラーニングの抑制が期待される。世の中の主流はフィードフォワードニューラルネットワーク（FNN）だが、これはサイクルを含まないので、原始帰納関数に相当する。

ディープニューラルネットワークは、本来であればサイクルを含むはずのニューラルネットワーク、すなわち一般帰納関数でサイクルを不定回実行することを、サイクルを回す回数を何十回とかいう割と大きな回数に固定することによって、つまり一般帰納関数を原始帰納関数によって近似しているようなものと捉えられる。

したがって、サイクルをいかにうまく取り仕切るか、飼いならすかが重要だと思う。サイクルにはさまざまな種類がある。実行のレイヤーが一番基本的で、これは環境に対する適応度を直接改善するような、リアルタイムの環境とのインタラクションである。また、その実行の方法を改

善するためのサイクルが学習である。さらにメタ学習というのもあり得る。これはひょっとしたら進化のようなものかもしれないが、さまざまなレイヤーがあって、意味の階層を構成する（図 2-4-2）。

学習や進化は、サイクルになっているのは当然であるが、まず押さえておくべきことは実行のサイクルである。さまざまなエビデンスがあり、神経回路網の中には多くのサイクルが含まれている。また、生体の中では、ホメオスタシスなどいろいろな種類の振動が起こっているという話もある。

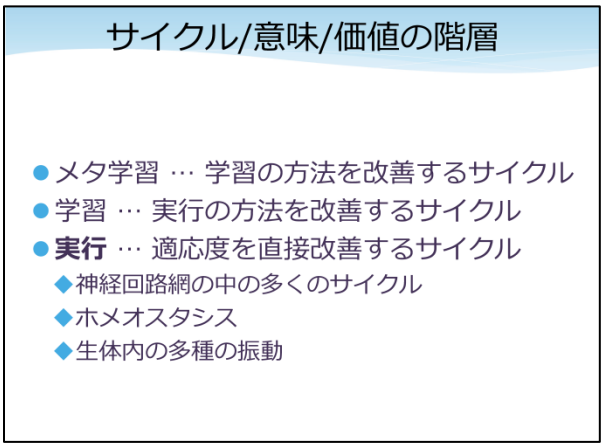

図 2-4-2　サイクル/意味/価値の階層

この絵（脚注[10]参照）を被験者に見せ、「女性を見てください」、「小屋を見てください」という指示を行い、その際の被験者の目の焦点距離を測定する。すると、遠方に描かれた小屋を見ているときは焦点距離が長く、手前に描かれた女性を見ているときは焦点距離が短い。

焦点距離といっても、絵は平面なので、被験者の目から小屋までの距離と女性までの距離とは物理的には同じである。しかし、絵のセマンティクスを解釈することによって、被験者は近景を近景として見ており、遠景を遠景として見ている。つまり、焦点距離の調整のようなことにおいて、単にはっきり見えるようにフォーカスを合わせるというローレベルのプロセスと、絵の意味を理解するというハイレベルのプロセスがインタラクションしている、図 2-4-3 のようなサイクルをなすと推測される。

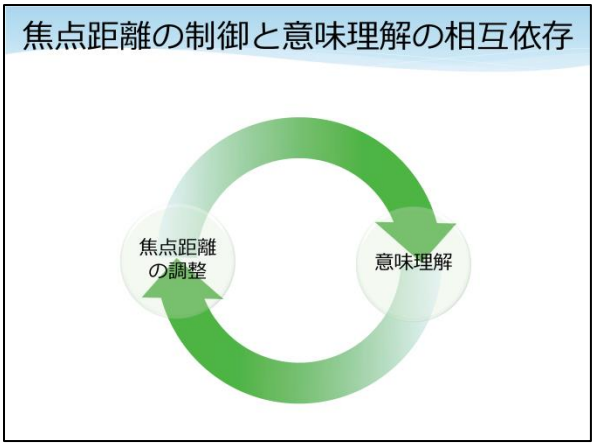

図 2-4-3 焦点距離の制御と意味理解の相互作用

[10] Andrew Wyeth の絵画作品”Christina’s World”　https://www.moma.org/collection/works/78455
手前に女性、遠方の地平線に小屋がある風景が描かれ、手前の女性は遠方の小屋を見上げているように見える絵である。

また、音韻論や音声学で昔から言われている、音声知覚の運動理論がある（図 2-4-4）。音声言語を発話するときに、話者の頭の中で起こっているプロセスと、それを聞いたときに聴者の頭の中で起こっているプロセスは同じだという話である。

先ほどの麻生さんの資料にあった、順逆変換ネットワークも同様の話である。

言語に関して、ブローカ失語症という障害がある。ブローカ失語症の患者さんに、「この 4 コマ漫画を見て何が起こっているかを言ってください」と聞くと、回答の文を発話することがほぼできない。

一方で、例えば「ドアを開けてください」とか「窓を閉めてください」と言うと、ドアを開けてくれるし、窓を閉めてくれる。つまり、一見、この人たちは文をつくり出すことはできないけど理解することはできる、すなわち、前述の音声知覚の運動理論とは違って、言語の産出と理解とが異なる能力のように見える。しかし、ブローカ失語症の患者が文を本当に理解しているかどうかを、例えば「The car is pushed by the truck.」という文を聞かせるか読ませるかし、自家用車がトラックを押している絵とトラックが自家用車を押している絵を提示して文意に合うものを選ばせると、正解率は 50%となる。

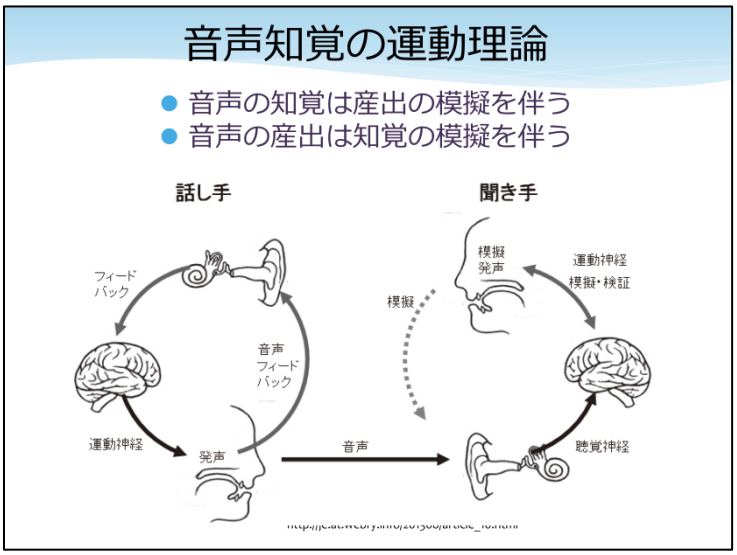

図 2-4-4　音声知覚の運動理論

要するに、ブローカ失語症の患者は、少しひねった文を理解することができない。つまりブローカ失語症では、発話の能力が失われているのではなく、文法へのアクセスが失われていると考えられる。言い換えると、言語表現の産出と理解のプロセスは、図 2-4-3 や図 2-4-4 と同様にサイクルとして一体になっていて、分離できない。

我々認知主体がさまざまな状況に置かれたときに、その状況で多様な情報が利用可能になるわけだが、利用可能な情報の分布が文脈によって大きく変わる。例えば、何か物理的な対象が部分的に隠されていて見えないとか、ノイズが入って人の発話が部分的に聞こえない、ということがあるし、話をしているときに、ある種の話題に関して自分は知識がないので分からない、など、関連する情報・知識の中で、どの部分が利用可能でどの部分が利用可能でないかという、その分布が目まぐるしく変わるというのがごく普通である。

したがって、ニューラルネットの場合にも、値が分かっているユニット（固定ユニット）はごく一部で、どのユニットの値が分かっているかという与えられた情報の分布は、やはり文脈によって変わると考えるべきであろう。例えばリコメンデーションは、いくつかの既知の商材に関するユーザーの評価の情報が分かっているときに、そのユーザーが未知の商材をどう評価するかを推測するわけだから､利用可能な情報の種類が文脈によって異なるという問題設定である｡また、文の認識でも、文の一部しか聞こえない、見えないときに、文全体を推測するといった問題もある。

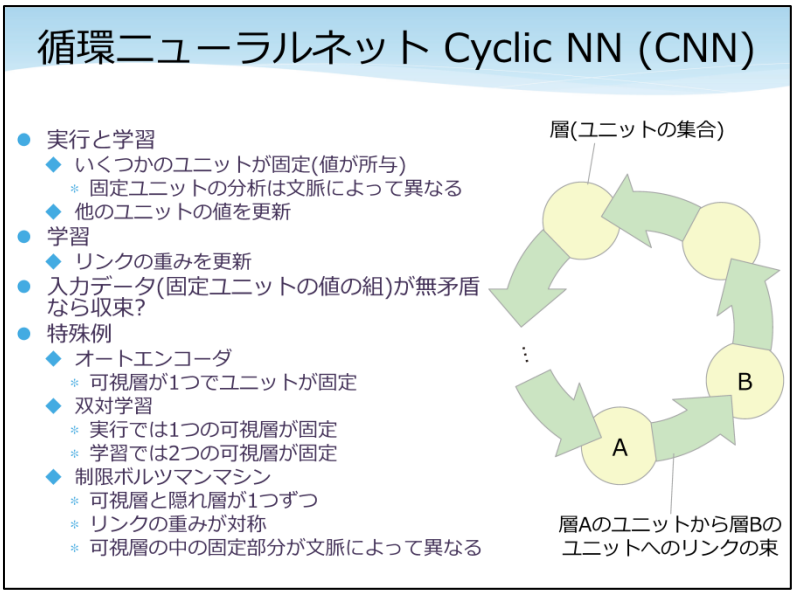

図 2-4-5　循環ニューラルネット Cyclic NN（CNN）

サイクルをなすメカニズムで、いろいろな文脈に対処できるものとして、図 2-4-5 の循環ニュートラルネットワークのようなものが良いのではないかと思う。実行時にも学習時にも、この可視層の中のいくつかのユニットの値が既知である。ただし、どのユニットの値が既知であるかは文脈によって変わる。また、他の既知でないユニットの値を更新するということも起こる。学習時にはリンクの重みを更新するということも起こる。そういう動きをするネットワークである。このようなサイクルが収束するかどうかは自明でないが、おそらく入力データ、つまり固定ユニットの値とその分布が現実的なものであれば、つまり世の中の実情と矛盾しないものであれば収束するのではないかと期待している。

オートエンコーダー、双対学習、制限ボルツマンマシン等は、この循環ニューラルネットワークの特殊例である。各循環ニューラルネットワークをモジュールと考えたときに、複数の循環ニューラルネットワークを比較的簡単に結びつける、組み合わせることができる。複数のモジュールの間でユニットを共有すればよい。

こうして、複合した循環ニューラルネットワーク全体の機能や意味を、各モジュールが持つ機能や意味の組み合わせとして設計する、あるいは説明することができる。かつ、スケーラブルな設計ができる。つまり、いくらでも多くのモジュールを組み合わせて大きなネットワークを作ることができるので、比較的自由にさまざまなモジュール構造を持った、大きなシステムができるのではないかと考える。

応用としては、言語を使用する場合には音声学、音韻論、辞書などさまざまな知識のモジュールがあり、それらの間のインタラクションをどうやって実現するかという話がある。ロボットの場合には、力学的な制約や、視覚、言語、常識などいろいろなモジュールを組み合わせる必要がある。

最初の例題としてはリコメンデーションがよいと考えている。翻訳のような例題もある。「太郎が花子を家に招待した。」のような文を、機械翻訳するとうまくできない場合がある。また「招待した」を「帰した」にしてもできないことがある。つまり、「AがBをCに招待するというときには、CはAの持ち物であるとかAの息のかかったものでなければいけない」、とか、「AがBをCに帰すというときには、CはBが普段いるところでないといけない」、といった推論規則のような知識のモジュールが必要である。

これをフィードフォワードネットワークのモジュールの並列接続で行うと、組み合わせの仕方そのものを学習する必要があり大変だと思う。つまりモジュラリティーが低い。循環ニューラルネットワークの上のような組み合わせのほうが良いと考える。

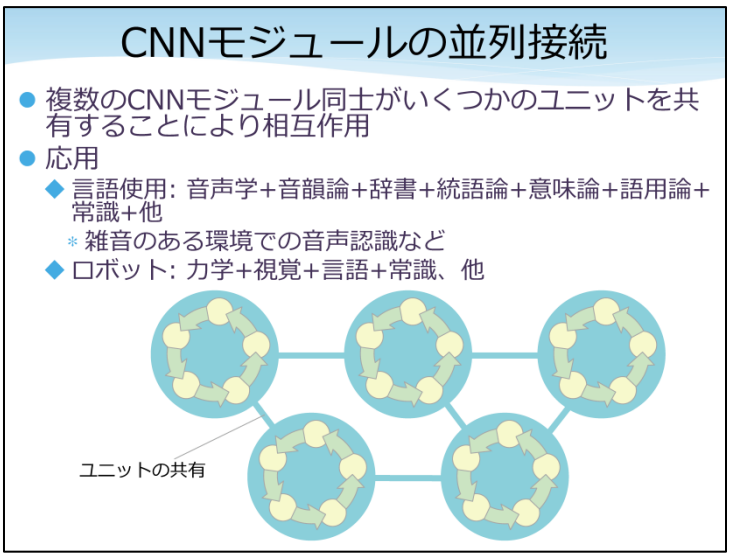

図 2-4-6　CNN モジュールの並列接続

最後に、さらに難しい問題はいろいろあるが特に２つある。まず、先ほどのようなモジュールをどうやって創発させるのか。モジュール構造が初めから分かっていれば、あまり苦労しないと思うし、そういう問題も多いと思う。しかし、一般にはそうではない。モジュールのないところからどうやってモジュールをつくっていくのかはすごく難しそうである。もし、モジュールごとに学習せずに、多数のモジュールの学習を、モジュールの間の区別をせずにやろうと思うと、学習に必要なデータは、モジュールの個数の指数関数になり、非現実的である。したがって、学習によってモジュール構造をつくる方法を確立しないと先に進めない。

もう一つは記号推論。先ほどの翻訳の例では「AがBをCに招待するというときには、CはAの持ち物である」のような推論規則が働いていると想定される。しかし、その想定そのものが不確かであり、もし正しいとしても、推論規則を一般に運用しようと思うと、各推論規則の複数のインスタンスを作る必要がある、など数々の問題がある。

質問応答

Q： ボルツマンマシンとは違うのか。

A： ボルツマンマシンは極端な例である。リンクとサイクルが多過ぎて学習時間がかかる等、扱いにくい。それを工夫しようとして、深層学習ができたとも言えるが、それは工夫のし過ぎで、適度にサイクリックなネットワークが必要だろうということである。

Q： 学習して、何かアクションして、フィードバックが返ってきて、また学習してという意味では、すべてある種のサイクルではないか。またリカレントなネットワークとの違いは。

A： ここで言うサイクルは、主には学習のサイクルではなくて実行時のサイクルである。また、リカレントネットワークは、仮説検証ではなくシークエンスを展開するためにサイクルを回している。使い方の問題である。

Q： 人が話を聞いていて、途中で違うかな、と思うと予測が変わってくる。オンラインでコンテキストを変えていき、予測するにしてもバックグラウンドのコンテキストが動的に変わるイメージで、一回フィックスで初期値を決めて、全部をデコードするのではなく、オンラインでどんどんコンテキストが変わっていくということかと思う。そうであれば、プレディクションコーディングのような感じがする。自分が今見ているバックグラウンドが、解釈のずれに合わせて変わっていく系である。プレディクションコーディングの実装は、きちんとやった人はいないと思っているが、さらに少しシンボリックなところまでレベルが上がっているような印象である。

A： ある文の前半を聞くと後半が予測できるというのはそういうことである。人間は、そういった予測を自然に行っている。かつ、聞こえないはずの音を聞いているし、見えないはずの色や形を見ている。

Q： 音声知覚の運動理論は、以前、声道モデルで行ったことがある。運動情報を入れるか入れないかで大きく結果が変わる。音だけでは、アとオは、よく似ているが、声道モデルを入れて学習させると、きれいにクラスターできる。やはり運動を意識できる、聞いているところからそのバックにあるコンテキストをイメージするという感覚が、聞いている音をよりクリアにしていっているというのが、まさに、その場でサイクルが回っているということかと思う。

A： それは以前からシステム理論で言われていることであるが、そういうことを深層学習のようなもので実現できそうになってきたということであろう。まず先ほど説明したような循環ニューラルネットワークをつくりたいが、既存のニューラルネットワークのライブラリーで簡単にカスタマイズできるものがあれば教えていただきたい。

Q： 制約ボルツマンマシンでやればよいのではないか。

A： 制約ボルツマンマシンには層が２つしかない。制約ボルツマンマシンを積み重ねたものでもない。制約ボルツマンマシンの積み重ねだと、隣り合った層の間で双方向の結合があるが、先ほど説明した循環ニューラルネットワークの場合は片方向であり、フィードフォワードニューラルネットワークの場合に似た学習アルゴリズムが使えそうである。

Q： オンライン中に動いていくという感覚がちょっと独特である。ボルツマンマシンは何ステップか後に収束する。これは、収束を待たずに次に行く感覚が印象に残っている。

A： 実行時には、理想的には収束するのだが、実際には収束を待たずに次に行く。

Q：具体的にやろうとされているタスクとか、具体的にどうやって使うかというのを教えていただきたい。

A：まずはリコメンデーションに使いたいと考えている。制約ボルツマンマシンで協調フィルタリングを行う話はあるが、循環ニューラルネットワークのほうが精度が良いのではないかと思うのでまず試してみたい。

Q：リコメンデーションについて、もう少し具体的にはどのようになるか。

A：例えば、可視層のうちの一部のユニットの値が分かっているというのは、各ユニットが一つの商材で、その商材に関するユーザーの評価が分かっていることに相当する。しかし、同じ可視層の中の別のユニットの値が分かってないということは、それらのユニットに対応する商材はユーザーにとって未知であり、その評価を知りたいということである。

Q：ここでのサイクルと、記号創発、記号接地、あるいは概念の分節化、そういうものとの関係というのはどうなるか。

A：このサイクルというのは頭の中で起こっているサイクルでもあり、認知主体と外界を結ぶようなサイクルでもあり得る。後者の場合は、いわゆる知覚行動カップリングである。また、サイクルが全部頭の中に入っていると、何らかの記号推論のような話になると思う。

Q：そのときのサイクルの単位と、記号や概念との対応は１対１であるか。

A：それは全く分からない。研究課題である。一つのサイクル、つまりデータ・情報の流れが回っているのは、何らかの記号のインスタンスだろうと思う。

Q：このサイクルは、外界とのインタラクションの中で変わっていくというか徐々にできていくのか。

A：徐々にできていくやり方を考えるのが、最後に述べたモジュールの創発という非常に難しい問題で、まだ全く答えの見当がつかない。一つ一つのモジュールを何とか手なずけ、それらを組み合わせて、もう少し複雑な問題を解く、というのが、当面の話だと思う。

Q：理解イメージを追加すると、認知をする際、本当は落ちる解はいろいろあり得るが、サイクルが回るにつれて、このコンテキストならこの落ち、このコンテキストならこの落ちということが、意味理解が入ると見方の意味が変わるということであろう。したがって、１対１でマップしてはだめで、再解釈して、落とすべき場所を変えなければならないとうことかと思う。いろいろな可能性がちゃんと準備されていて、それでちゃんとどこに行くかという話になる。

A：そのためにさまざまなモジュールの間のインタラクションが必要であり、そのインタラクションの実装法として、フィードフォワードニューラルネットワークより循環ニューラルネットワークのほうが良いと思う。

２．５　マルチモーダル確率的生成モデルと汎用知能

長井隆行（大阪大学）

本日は、確率的生成モデルをキーワードとして、汎用的な知能をどのように実現していくかについてお話ししたい。私自身は、知能は脳の計算と、身体と、社会的な関係性が非常に重要であ

ると思っているので、人間の知能とは何かということを少しでも理解できるようになったらよいと思っている。

人間の知能活動の実現を目指す概念として、記号創発ロボティクスがある（図 2-5-1）。記号創発ロボティクスという言葉自体は、2010～2011 年頃に創発したもので、身体を基盤とした主体が、知能をソフトウェアとしてとらえずに、環境と相互作用することで概念や言語を獲得していく過程を、機械学習やロボティクスの技術を用いて実現することを一つの目的とした研究領域である。我々のグループは、記号創発ロボティクスの視点から、言語を含むロボットによる実世界の理解の検討を行ってきた。

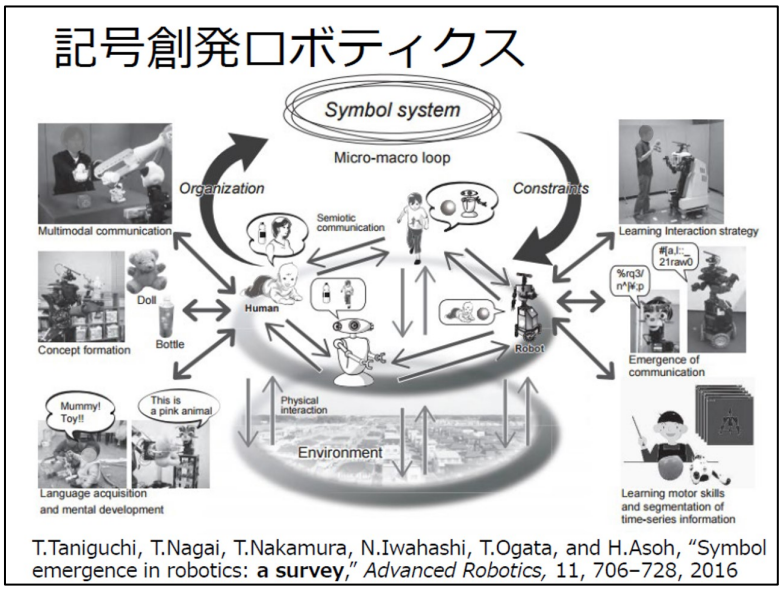

図 2-5-1　記号創発ロボティックスの概念

そこで、本日の論点として、

（1）次世代 AI の中核的な技術チャレンジとして、汎用的なロボットを作れないか。

（2）その汎用的なロボットがもたらす社会・産業へのインパクトは何か。

について考えてみたい。

図 2-5-2　ロボットによる実世界理解

そもそも、ロボットによる実世界での言葉の意味や理解とは何であろうか。ここで重要なことは、実装可能なレベルの定義を行うことである。

我々は、「言葉の意味を理解するということは、その言葉を聞いた際に、その場の状況や過去の経験を通して、関連する情報を予測することであり、予測された内容がその言葉の意味である」と定義している（図 2-5-2）。

また、予測の精度を高めるには、経験によって得たマルチモーダル情報を分類することでカテゴリを形成し、そのカテゴリを通して現状を把握し、予測すべきである。このマルチモーダル情報を分類したカテゴリをマルチモーダル概念と呼んでおり、我々は、確率的生成モデルを開発してきた（図 2-5-3）。

例えば、人が「ぬいぐるみ」と言ったときに、ロボットが「ぬいぐるみ」という言葉の意味を理解するとはどういうことかというと、その言葉、音韻列から、視覚、触感が予測できることで、これをもって「ぬいぐるみ」という言葉をロボットが理解するということである。これは、当然、過去のぬいぐるみの視覚、触感の経験を用いるわけである（図 2-5-2）。

ロボットは、自らが経験して得るマルチモーダルなセンサー情報等を範疇化することによって概念構造を獲得し、その概念構造を通して未観測または隠れた情報を推論することで意味を理解できることになる。ここで、意味のある概念構造を、ロボットが経験に基づきボトムアップ的に獲得する必要があるということが重要である。

ロボットが自らの身体によって得るマルチモーダル情報を構造化し、それによって実現される予測が本質な意味を持つのである。我々は、この仕組みをマルチモーダルカテゴリゼーションと呼び、階層ベイズの枠組みである MLDA （Multimodal Latent Dirichlet Allocation）でモデル化できることを示した（図 2-5-4）。

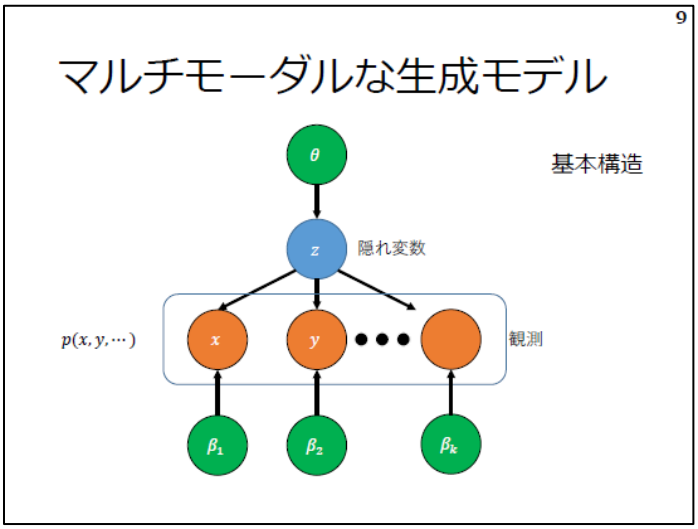

図 2-5-3 マルチモーダルな生成モデル

マルチモーダルカテゴリゼーションでは、身体性の利用により、人の感覚に近いカテゴリを軽形成することができ、人と自然なコミュニケーションにとって重要な概念となる。

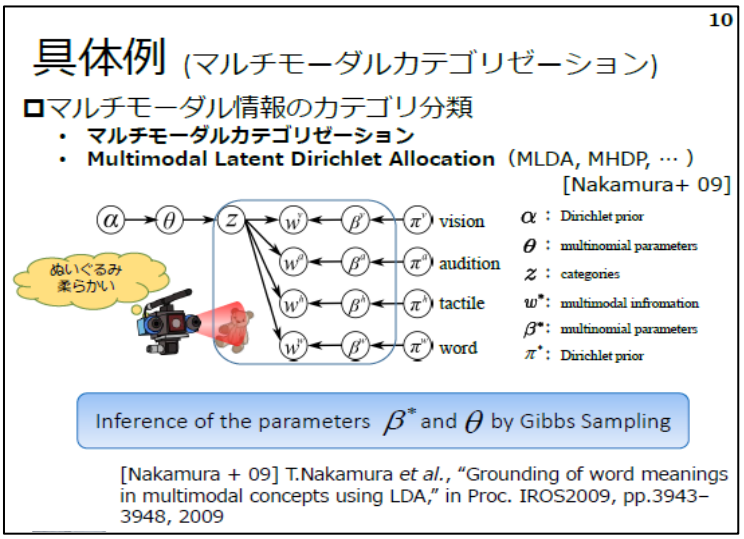

図 2-5-4 マルチモーダルカテゴリゼーション

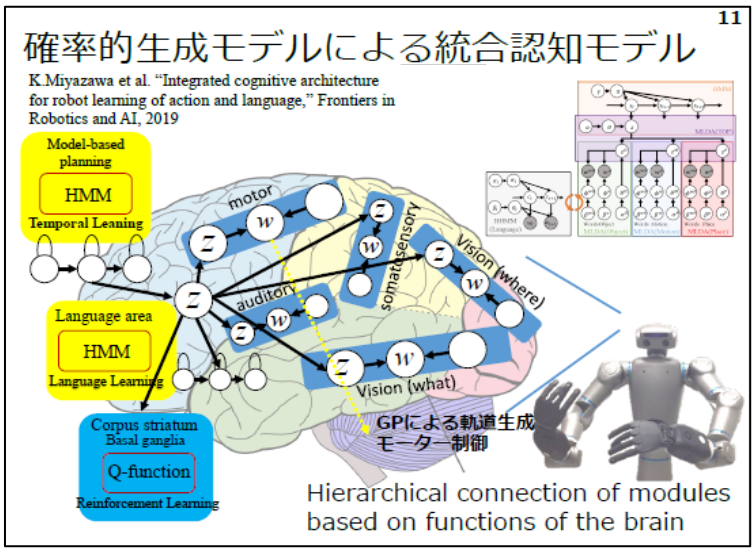

図 2-5-5　確率的生成モデルによる統合認知モデル

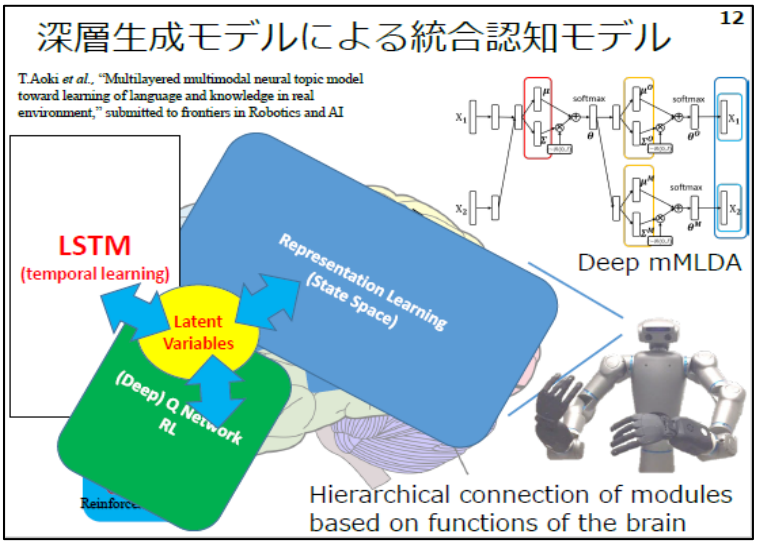

図 2-5-6　深層生成モデルによる統合認知モデル

一方、統合認知モデルは、多層マルチモーダル LDA（mMLA）による概念形成を中心に複数のモジュールを統合したものである。形成した概念を各モジュールが共有することで、複数の認知機能を統合することができる。各モジュールは、言語を扱う HMM（Hidden Markov Model）、行動の時系列をモデル化し長期的行動計画を行う HMM、即時的な強化学習（Q-function）である（図 2-5-5）。

また、近年の深層学習の進展により、確率生成モデルを深層学習によって推論する深層学習モデルによる統合認知モデルも提案されている（図 2-5-6）。

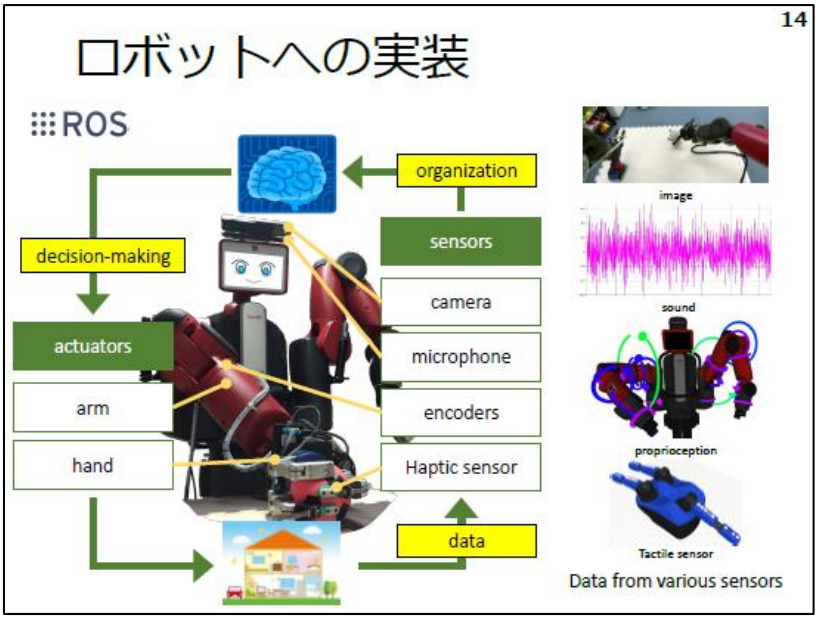

図 2-5-7　マルチモーダル情報のロボットへの実装

マルチモーダル情報を実際に使用した例として、ロボットが長期間にわたって概念・語彙の獲得を行った実験を紹介する。この実験では、マルチモーダル概念と音声を認識するための言語モデルを相互学習し、ロボットが音響モデルのみの状態から、概念と語彙を獲得できることを示した。この実験では、図 2-5-7 に示すロボットを使用して、人と物体を介したインタラクションを 1 日 3～5 時間、それを約 1 ケ月間（100 時間以上）行った。この期間内にロボットは、499 個の物体（81 カテゴリ）を逐次的に学習した。学習物体には、ペットボトル、カン、ビン等の食料品、ぬいぐるみ、ボール等のおもちゃなど多様なものが含まれている。ロボットは、頭部に取り付けた RGB センサーにより視覚情報を、指先に取り付けた触覚センサーにより物体を掴んだときの触覚情報を、手先に取り付けたマイクロフォンにより物体を振ったときの聴覚情報を取得した。さらに、物体を観察している間、被験者の発話をヘッドセットにより取得した。実際、このような 1 ケ月にわたるロボットと人とのインタラクションを経て、語彙と概念の相互学習のモデルを搭載したロボットは、全学習物体の約 6 割を正しく分類することができた。この実験の詳細な分析を行うと、学習が進むにつれて、ロボットがマルチモーダルなカテゴリを形成していく様子を観察することができる。

こうしたロボットによる、長期的な学習は、乳幼児の言語獲得プロセスとの比較を可能にするとともに、これらの知能モデルを結合することで、身体制御、言語理解、言語生成までの発達プロセスを獲得するロボットの実現が期待される。

次の議論として、我々が研究を行っている統合認知モデルが、この分野でどのような位置づけにあり、今後の課題としてどのようなものがあるか、社会的なインパクトはどのようなものかをお話ししたいと思う。そのために、さまざまな視点で統合認知モデルの解釈を行っていく。

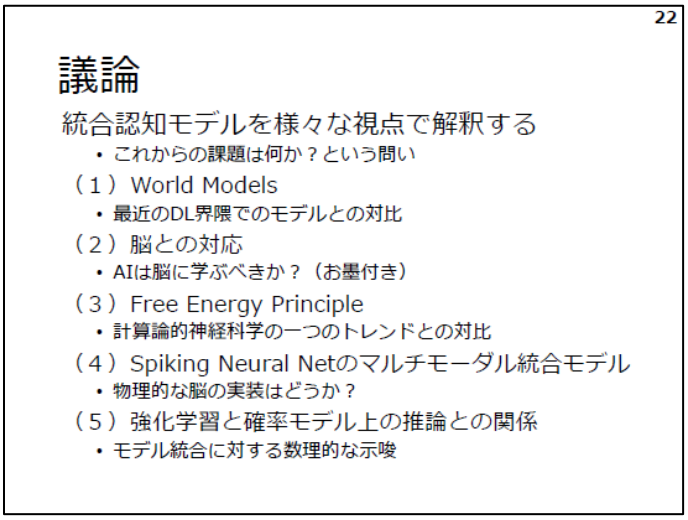

図 2-5-8　統合認知モデルの解釈の視点

2018 年頃から、World Model（世界モデル）と呼ばれる、自己教師あり学習によって、自己をとりまく環境をモデル化しようという研究が盛んになってきたが、この考えは、我々の研究の発想とかなり近い。人間はあらゆるものを知覚できるわけではなく、脳にインプットされる情報が限定的である。したがって、脳の内部では、インプットされた情報から世界をモデルとした内部モデルを作っている。つまり、内部モデルとは、外界からの刺激を基に、外界世界をシミュレートするモデルということができる。

World Model は、内部モデルに近く、以下の 3 つの要素から構成されている。

- Vision Model（V）：高次元の観測データを、VAE（変分自己符号化器）を使用して低次元のコードに圧縮
- Memory RNN（M）：RNN を用いて、過去のコードから未来の状態を予測
- Controller（C）：V と M から良い行動を選択

我々がやっている研究も、基本的には World Model と同じで、時間的な情報を見ながら、その時にどのような行動をとるべきか、その行動をとると次にどのような状態に遷移するかという意味で、World Model を確率的な生成モデルで記述しているということになる。

次に脳との関係についてお話ししたい。

脳の生理学的な結線を考えていくと、ある種の制約条件みたいなことが分かるのではないかと言われており、また、脳がどのように学習しているかを知ることは、人間の知能を考えるうえで重要である。したがって、今後は、我々のモデルをどのように組み合わせれば、人間の脳の全体として表現できるかを探求していきたい。そして、仮にそれが実現できたとしたら、ロボットに実装することも重要となる。この一環として、我々は、複数のモジュールを結合して相互学習するフレームワークとして、SERKET（Symbol Emergence in Robotics tool KIT）を開発している。SERKET は、小規模なモデルをモジュール化し、階層的に接続することにより大規模なモデルの構築と、そのパラメータ推定を容易に行うことができるフレームワークである。これにより、脳機能は生成モデルを使用して描き切ることができる（図 2-5-9）。

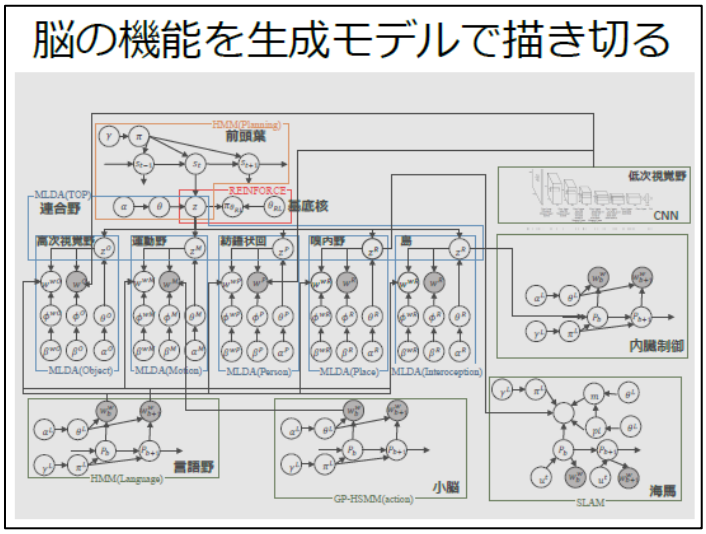

図 2-5-9　生成モデルによる脳機能の表現
（ただしこの図はイメージを伝えるための図であり正確なものではない）

自由エネルギー原理（Free Energy Principle：FEP）は、Karl Friston が提唱した、脳における情報処理機能全般を対象とした知覚と行動と学習の統一原理である。

「いかなる自己組織化されたシステムでも、環境内で平衡状態であり続けるためには、そのシステムの（情報的）自由エネルギーを最小化する必要がある。」というものである。これは、適応的なシステムが無秩序へ向かう自然的な傾向に抗して持続的に存在し続けるために必要な条件である。自由エネルギーは、変分推論の ELBO（Evidence Lower bound）と符号が異なるだけであることから、統合認知モデルは、自由エネルギー原理の一つの実装であると言える。

Sergey Levine は、強化学習の Value Iteration 系の話はグラフィカルモデルで記述できて、確率的推論と数理的には同じであると言う論文を発表している。我々のモデルも表現学習を行っていて、その状態空間を使用して強化学習を行ったり、その状態変化を時系列のモデルで捉えようとしたりしている。ただ、プランニングをするということと、強化学習をするということは、ほぼ数理的には等価なので、2 つある意味はあまりないと考えられる。では、この意味は何なのかというと、反実仮想のモデルを積極的に使用してデータを生成し、強化学習を行うということである。これは、ある意味、モデルベースの強化学習とモデルフリーの強化学習と捉えてもいいかもしれない。

HMM の文脈で考えると、過去から未来を予測していく中で、行動しながら、状態をアップデートしながら進んでいくという形で、これができると、仮説推論のような推論ができることになる。そうすると、ロボットにある種の説明性が実現できることになる。また、ロボットは、実際に行動してみて、因果関係を持っているかどうか確認することができ、Judea Pearl が言っているように、do inference の do ができるわけで、そう考えると、因果推論、反実仮想、仮説推論系が全部実現できるのではないかと考えられる。

図 2-5-10　生成モデルに足りない機能

以上、我々の生成モデルをさまざまな視点から解釈してきたが、我々のモデルには、まだ足りない機能もたくさんある（図 2-5-10）。

この中で、情動・感情については、先ほどの外界に対する自由エネルギーの話を、自分の身体に対する自由エネルギーの話に置き換えればできそうだと議論されており、同じような枠組みに取り込むことができると思われる。また、高次な認知機能については、トリレス線図（現在の内部状態と出力を表す方法）の中のパスの推論で実現できると考えられる。

汎用的なロボットは実現できれば、いろいろな所に有用でインパクトが大きいということは、おそらく、皆さんの共通認識であると思う。

質問応答

Q：ロボットの実験で、長時間の学習を行ったという紹介があったが、どのくらい汎化できて、どのくらいスケールするのか。また、語彙獲得は人間みたいにエクスポーネンシャル的に増加するのか。

A：ロボットのインタラクションモデルでは主に 3 つの理由によりスケールはしない。それが問題だとの指摘もある。1 つ目の理由は確率的生成モデルを使用しているからである。深層生成系のアルゴリズムを使用すれば回避できそうなところもある。2 つ目の理由は、ロボットの実験では学習すべきものが限定されているからである。3 つ目の理由は、現状のモデルでは自他分離が取り込めていないからである。短い範囲内では、エクスポーネンシャルに近いものもあったが、全体的には飽和すると思われる。それは、スケールしないという理由とほぼ同じである。

Q：ロボットの実験で、ずっと言葉だけ教えているのはかわいそうである。

A：今回紹介した実験では、本当に言葉だけだったが、それは、音の情報をきちんと区切って、それと実質的な感覚や資格の情報が相互に結びつくということを示したかったためである。その後の実験では、運動学習も入れて実験しているので、対象範囲はもう少し広くなる。

Q： 意味は、予測も入るかもしれないが、それよりは行為選択の可能性だと思っている。例えば、ぬいぐるみは、普通は可愛がるというコンテキストで使用するが、何かたたく道具に使用する人もいるかもしれない。つまり、一つのものがいろいろな言葉に解釈可能であるため、ロボットがどういう行動の可能性で扱えるかというところまでいって、初めて意味の議論になるのではないか。インタラクションのコンテキストの中で、言語が出てくると、やはり記号接地問題を考えてしまう。

A： 初期のロボットの研究では、情報を得て、それをどのように構造化するかということに意味を見いだしたかったので、行為の可能性についてはあまり考えていなかった。ただ、実際には、確率生成系のモデルには、行為の可能性というのは入っている。だから、自分がこういう行動を行うから、それがこういう言葉で表現されるという部分はモデルとしては表現できていると思う。

Q： 例えば、ある物体があって、これを PC と呼んでいたが、今はメールをやっているのでメールと呼ぼうというような行為によって追解釈されるようなことをやろうとすると確率によるモデルでは難しくなると思う。やっぱり力学的なことを議論したいと思っている。

A： それは誤解であると思う。

Q： 図 2-5-3 の基本構造のところで矢印がいくつかあるが、この情報の流れは双方向か。

A： 双方向である。

Q： 具体的にはどういうネットワークで表現されるのか。

A： ベイジアンネットである。深層生成モデルになると、推論するときに、オートエンコーダーになっているので、ある方向から推論するネットワークと、逆方向から推論するネットワークが形成される。

Q： この研究は、どこを落としどころにするのか。

A： 構成論的なアプローチをとったときに、人間の赤ちゃんは、プレトレーニングされて言語形成ができるわけではないので、ロボットにプレトレーニングの言語モデルを持たせるのはよくないと思っている。ただ、人間も、モダリティーと切り離された世界の情報も得て学習するという側面もあるので、それを知識としてどのように学習していくのかは研究してみたいと思っている。それが、もしかしたら、エクスポーネンシャルに仕上がっていくのかもしれない。

Q： 人間の胎児の研究では、言語形成は、おなかの中のプレトレーニングがきいているという知見もあるが、生まれた後の言語形成がモジュール化していくプロセスはよく分かっていない。そういう意味で、最後のモジュール化を海馬で強くやり過ぎない方がいいというイメージを持っている。

A： 構造自体を決めていくという問題は重要であるとは思うが、かなり難しい。

Q： どこまでサイエンスでやり、どこまで実用にするかの問題だと思うので、研究の方向性としては、その境目をうまく狙っていくことになるのではないか。

A： 進化と発達がどう関係しているかという問題もある。

3．総合討論

融合の方向性

福島　これまでの議論をプロポーザルに反映していくにあたり、今回取り上げたい論点を冒頭で 2 つ挙げた。その 1 つめは「次世代 AI の中核的な技術チャレンジは何か」である。今、深層学習でいろいろ成果が出ているが、その先を見据えて、きちんと中長期的に研究投資していきたい。現在の AI から次の飛躍のためにどういう取り組みをしていくべきかというところを明確にしていきたい。

これからの方向性の仮置きとして、何らかの融合に向かうと考えている。「深層学習と知識・記号推論の融合」、「帰納型と演繹型の融合」、「コネクショニズムとシンボリズムの融合」などいくつかの言い方をしているが、実は共通する部分はあるものの、微妙に違う。厳密な技術的定義という面と、一般的な分かりやすさという面があるが、どういう融合を目指すのが AI の進化として本質的か。

そこでは記号接地問題のようなものも絡んでくるはず。この問題も、実世界のものとラベル（記号）を対応づければよいというものではなく、文脈や文化に応じてラベル付け・接地が変わってくるとか、ラベル同士の関係といったメタレベルの概念にラベルを付けるとか、さまざまなレベルの問題を含んでいる。記号接地問題は再定義されるべきと思う。

また、大量の教師データがなくても、実世界とのインタラクションを通して学習、成長、あるいは発達していく仕組みも非常に重要なポイントではないか。

橋田　スケーラビリティーが重要ではないか。いろいろなモジュールの組み合わせにより、より大きな問題を解くということを一般にどうすればいいのか、よく分かっていない。でもそれができないと、いろいろなリアルワードのアプリケーションがつくりにくいと考えると、サイエンスとエンジニアリングの両方をにらんだテーマ設定になり得る。

スケーラビリティーはモジュラリティーとも関係が深い。モジュラリティーがアカウンタビリティーを高める等、いろいろなメリットがある。スケーラビリティーを前面に出すと、いろいろな問題がすんなりとつながるのではないか。

福島　今の深層学習で、スケーラビリティーが出ていないことの原因はどこにあるのか。

橋田　複数のディープニューラルネットのモジュールを並列につなごうとすると、つなぎ方自体を学習する必要があり、あまりスケールする感じがしない。モジュラリティーが低い。それをどうするか。

尾形　Yoshua Bengio は ICRA2019 の基調講演で「因果性」、「マルチタイムスケール」といったキーワードを使っていた。要するに、自分の行為の可能性に関係する World Model を使って、物の見方を考えられるのかということ。また、事前モデルによる予測とのずれを利用することで常に行為の仕方、認識の仕方を変えていくことも必要になってくるのだろうと思う。それが複数の解釈が出てくることにもつながっているのではないか。

また、強化学習における探索とは何なのかという議論がある。つまり、本来はまず行為探索によってでき上がってくる報酬系があり、これこそがシンボルになる。そういう意味で、記号推論に行けたらよい。

記号推論が中途半端に入るとパフォーマンスが落ちてしまうと言われているが（部分的には確かにそうだが）、本当は違うのではと思っている。人間の場合でも、本を読むとか辞書

を見るとかによってうまく問題を解決することもある。スケールが違うというところに自然に記号が入ってきて、再解釈をしているうちに、こうやればいいんだ、データベースを用いたほうがすっきり説明しているじゃないかということが、人間の場合はよくあるだろうと思う。

だから、まさに自分が行為をするために物をボトムアップ的に見直しているところに、記号のレベルがトップダウンで見てくれているような、そのようなイメージができたらいいなと思っている。どうやるか分からないが、今日の乾先生のモデルは一つの可能性を見せてくれている。

今のところ我々は、簡単なセンテンスが行為に多義性を含めてどうつながるかというところをやっているレベルだが、乾先生のモデルではいろいろな多義性やノイズが乗っているデータが、シンボリックな構造に結びつくものになっていると思う。我々のやっている運動のような下のレベルから、シンボリックな構造まで行くような話になったらよいと思う。

福島 そのあたりは、融合形態④（図 1-10）に挙げた 2 階建てモデルが、イメージとして近いか。

尾形 そのアプローチは、上の強いシンボルはあまり意識せず、全部深層学習で考える立場ではないかと思っている。「創発」に近い。私は、もちろん記号創発は重要だが、人間の利用しているシンボルそのものは非常に強いものだと思っている。「1 足す 1 はどんなときでも 2」、「交通ルールは常に守れ」といった、コンテキストに依存しない普遍な表現のようなものだ。実際には、人間は臨機応変に適当に実世界に記号なしでも対応しているのだけど、そこに普遍的なもの（記号）があると気づける能力をボトムアップ（創発）でつくろうとすると（できないわけではないはずだが）相当難しいと考えている。ちゃんと記号体系があるのであれば、それをうまく運用することも同時に行うべきだと思う。

ルールを入れると、却ってそれに引っ張られてだめになってしまったり、臨機応変にできなかったりということが起こるのかもしれないが、それでもルールを知っていたほうが良いと直感的には思う。どうやったらルール（記号）が深層学習と自然にくっつくのだろうか。

乾 長井先生がやっているようなサイエンス系の話と、AI でできることを増やしたいというテクノロジー系の話では、大分違う。研究としては両輪であり、両方やることになるのだろうが、説明の仕方によって、どちらの指向を持っている人が来るのか大分違ってきそうな印象を持った。

テクノロジー系でいうと、人間とともに何かするとなると、人間には記号でしか解ってもらえないというレイヤーがあると思う。例えば、人間は言語という記号をみんな既に持っていて、それを共有している。それをオントロジーなどよって、AI ともある種の共有ができればいいということを自然言語処理は今やろうとしているが、記号としての言語と、ニューラルネットのような連続的な世界との間がうまく対応がとれるようになってきた。埋め込んでいるというのはまさにそういうことだと思うが、それをもう少しエクスプリシットにやろうというのが、私が今日お話をしたこと。

一方、今日の尾形先生や長井先生のお話によると、人間にも記号でない言語化されない世界があり、我々が言語のレベルで見ていたことと少し違う。そうすると、ディープニューラルネットワークのようなものと人間との間には、シンボリックなインターフェースとそれ以外の何かがあるということなのだろうか。

尾形　言語も本来、綺麗には記号化されてない生々しいものの一つで、加えて感覚運動などいろいろな生々しいチャンネルがあるということなのだと思う。

乾　それ以外のいろいろなチャンネルがあるということか。今の深層学習プラス記号というのは、上の記号化された層だと分かりやすいけれども、下の層には、生々しく、マルチモーダルで、記号化されていないものをいろいろ持っているということか。そのときに、記号と深層学習がどう対立して、何を融合しようとしているのか、分かりにくくなる印象を持った。

福島　世の中に伝わりやすいから深層学習という言葉を使っているが、知覚運動系や認識行動系という言い方もしている。知覚運動系のようなものと言語推論系をうまくつなぐときに共通的な表現とか構造のようなものがあって、そこにモジュラリティーも関わってくるというイメージを持っている。

乾　対立には 2 軸あるということか。言語・ルールといった記号とニューラルネットワークとの対立があり、言語とそれ以外のマルチモーダルな世界との対立。二重の意味を持たせている感じがする。これら 2 つの融合というのは、問うているものが大分違う。

福島　総合討論の冒頭でも述べたように、融合にいろいろな側面があり、一番の本質は何だと言うべきか悩む。

尾形　深層学習はどちらにも利用できるものなので、シンボリックなデジタルの世界とアナログな現実世界について、「上と下」という言い方をした。深層学習を使うと、ナレッジまで有機的につながる可能性がある。これによって、人間との言語によるコミュニケーション、人間と一緒に物を運ぶロボット、実世界での自動運転といったものができるということではないか。すなわち、ただ入力と出力をつなげる、マップをするというのではなく、深層学習が回帰構造をちゃんと持って一つのシステムになるということが必要で、深層学習の再解釈につながると思う。

乾　それは長井先生の統合認知モデルの「統合」と良く似ている。

福島　Daniel Kahneman が論じている System 1、System 2 とも同じことなのか。

尾形　System 1、System 2 についてだが、System 1 とは下の 1 層だけで考えてしまっているように思う。System 1 はフォワードだけでやってしまう感じだが、本当は、運動レベル、マルチモーダルのレベル、言語レベルといった異なる階層があり、インタラクションがある。それらによって、純粋なシンボルまで行くのは難しいかも知れないが、シンボルの入り口ぐらいまでは行ける可能性があるのではないかと思う。

長井　最初に名前について議論していたが、やはり言葉はすごく難しいし、キャッチーである必要もある。そういう意味では、谷口忠大先生が論じている「創発」が重要なキーワードではないかと思う。ただし「創発」とは何かということから考える必要があるため結構難しい。ちゃんとした知識を持った上で、創発という言葉を使うのは良いと思うが、それを一般に伝えるのは難しい。

麻生　本質的なアイデアはいろいろ出ており、何かやればいいという印象を持ったが、一方、長井先生の話を聞くと、まだいろいろ難しいとも思える。長井先生がやられていることは基本的に、ベイジアンで双方向の推論をすること。ただし、ベイジアンでは学習が大変だから深層学習を導入したり、モジュールの創発をノンパラメトリック・モデルでしたりなど、いろいろ手だてはあるでしょうと言うことだと思うが、やはり、それでどこまで行けるか

一回やってみていただきたいというのはある。

エクスペリエンスデータの収集

麻生 そのときに問題となるのは、データがないことではないか。つまり、ビデオやテキストといったデータはあるが、エクスペリエンス（経験）をデータ化したものはない。Google DeepMind は、StarCraft、Minecraft などのゲームの世界でのエクスペリエンスのデータを大量につくって学習させているが、そこでのエクスペリエンスでどこまで行けるのか。

尾形 まさにエクスペリエンスデータを収集するという趣旨で産総研の「AI テストベッド事業」を実施しているのではないか。

麻生 確かに、サイバー・フィジカルシステム研究棟では、例えば工場の組み立て作業やコンビニでの陳列での作業をエクスペリエンスの対象としている。

尾形 あそこでリアルに実世界で活動するエージェント（ロボット）をあれだけ準備できているというのは、一つの方向性だと思う。

橋田 そのエクスペリエンスデータとはどのようなものか。

麻生 人間が実世界で活動したときの、いろいろな五感の同期データのようなもの。

尾形 ロボットのモーター、つまり行為も含めたというイメージ。まさに何が何の基点なのかというときに、自分の行為が因果性には関係するはずなので。それは別にロボットでなくても何でも良いのだが。

麻生 ただし、一般的なコンテキストをやるのは難しいので、組み立てやコンビニのような環境にタスクをある程度絞って、そこでデータをとっていきましょうということ。そういうところで、できるだけ、構成性とか因果性が重要であり、Consciousness Prior が有効で、しかも単一タスクではなくマルチタスクであるという状況を設定して、頑張ってエクスペリエンスデータを収集して、それで長井先生のアプローチでどこまで行くのか試してみたい。

橋田 そこでは、私が先に述べたのとは別の意味のスケーラビリティーを考慮する必要があるのではないか。つまり、研究開発のためだけに特別な仕組みをつくって作成・収集するデータではなくて、日常の生活や業務の一環として広く行われているサービスにおいて、自然に､大量の､しかもなるべく構造化されたデータを収集できるようにしたいということか。そうしないと、サステナビリティーもスケーラビリティーもない。

麻生 そのとおり。それもあわせて、エクスペリエンスのデータ化というのが一番難しいのではないか。CPS 棟は、模擬環境であり、それだけだと、研究用のデータというところに留まってしまう面はある。CPS 棟でデータの取り方を研究して、その先は、利用していただく企業などと連携して、継続的にデータが取れるようにする必要がある。

長井 ウェアラブルセンサーを用いてライフログを取得することが一時期流行ったが、我々の研究室では、ウェアラブルセンサーを使って学生の日常生活からデータを取得し、そのデータを使ってロボットに私がお話したものと同じような枠組みで学習、行動させると、それなりの何かができる。人間と同じようなタスクをロボットにやらせると、身体性が全く違うので、つくられる概念はやはり違う。そこの差を見るのは面白い。
しかしながら、データを収集するのはやはり大変である。同期をきちんととらなければい

けないし、データもものすごい量になってくる。プライバシーの問題も重要だ。橋田先生がおっしゃったように、きちんとサービスとしてデータがうまくとれるような仕組みがあると、やりようはかなりある。アルゴリズムとしてはいろいろあるが、やっぱりその中身であるデータをどうするかというのは、なかなか難しい。

尾形　「雇用の未来」を書いた Michael Osborne 教授は、e -lancing というものを紹介していた。これは、家にいながら遠隔でいろいろなタスクをするようになるということ。実際に、フィジカルに体を動かせない方でも、ロボットを使えば仕事ができるといった事例が出てきている。この遠隔ロボットのデータが集められたら、間違いなくロボットの自律化ができる。彼も、人間がネットを通じて行ういろいろな行為、仕事のデータが自動で集められ、それが学習の対象になっていくだろうと予測している。

国際ロボット展 2020 においてエクサウィザーズとデンソーウェーブは、本当に大事な学習システムや知識のデータベースをベンダー側において、現場でロボットを動かしたデータを全部ベンダー側に吸い上げて学習させ、学習結果をまたデリバーするというサービスモデルを見せていた。今後、そういうふうに操作履歴などのデータを取得するやりかたが増えるだろう思っている。今はまだ、工場現場の産業ロボットに IoT が全然入っていないが、ロボットの手前側でネットを通じて操作したデータが取得できるので、そういうことは起こるのではないかと思う。

橋田　遠隔医療からも操作データがとれるのではないか。

尾形　倫理上の問題など、対象によってはやってはいけないことがあるだろうが、遠隔医療でもいろいろできるだろう。もともと遠隔医療はいろいろな方に医療を受けられる機会を与えるという趣旨でつくられているシステムなので、データを集めればいいのだと簡単には言えないが、サービスとしては進んでしまうだろう。

橋田　我々はプライバシー情報の保護技術について取り組んでおり、これを使えば、プライバシー等を守りながら、個人情報や企業秘密を含むデータを集めて分析することが容易になる。

尾形　そのようなデータが使えるようになると、深層学習が入力と出力のマップを継続的に学習できるようになっていくので、すごく使えるようになると思う。

乾　それぞれのロボットには、アームの関節の数など、いろいろ異なる点があるが、あるロボットで学習した経験のデータを別のロボットに移植することは可能だろうか。あるいは、経験データの標準化の様なことが必要だろうか。

尾形　転移学習の応用はあるが、現状ではちゃんとした研究はほとんどない。関節を 1 個 2 個増ふやして上手くいくといった議論はあるが、コンフィギュレーションが全く違うと全然使えないというのが本当のところだ。そういう意味では身体にすごく依存している。我々のやり方（身体知の考え方）だと、むしろ身体に依存して知識や構造ができるということが重要になっている。先ほどの異なる階層というのはまさにそういうところ。身体にほとんど依存しないような世界の理解というものが入ってこないといけない。記号接地とは、そのような話なのかもしれない。また、橋田先生が論じたモジュラリティーも、そういうところではないか。

橋田　共有ができるためには、各文脈とか各個人とかに余り特化しているとだめで、抽象化する必要がある。したがって、腕の長さが違う人や、関節の数などが違うロボットの間で、動作やスキルのようなものを共有しようとするならば、例えばある運動するときに、このパ

ラメータが最小化されているといった、あるレベルで抽象化したものを共有することになるのだろう。

乾　一方で、映画を見たり、人がやっていることを見たりして、自分が経験した気になるし、そこから学んでいることはたくさんある。「あのようにやると落下するんだ」というように分かる。一体一体のロボットが経験のデータを積むということは難しいので、共有とはすごく重要なことであり、次にやらなくてはならない大きなエンジニアリング的なテーマであり、サイエンス的にも面白いテーマではないかと思う。それがある種の抽象化なりシンボルにしないと、経験として共有できない。

尾形　自分と他者のコンテキストがずれていても大丈夫なところを見つけるという行為が、コミュニケーションをするとか、言語にするというときのきっかけだったと思う。それが一体どこから来たのか、相当不思議なことだと思う。しかも、お互いにシェアできていると確信を持っているが、これも相当不思議なことだと思う。先ほど挙げた「1 足す 1 はどんなときでも 2」というのは普遍的に正しい究極の例だが、一般的にはシェアできていると確信を持っていることでも、実は間違っている可能性もある。長井先生が論じていたように、コミュニケーションしないとどうしようもないというモチベーションがないと、そもそもコミュニケーションはできない。今のロボットでは、人間のスキルをトランスファーするだけでも結構大変で、どのパラメータをどう最適化したらいいかが分からないので、仕方がなく深層学習でやっているというところがある。それによって、今までできなかったことができるようになったということだが、それだけでもインパクトがあるため、皆が深層学習の上の記号まではまだ議論できてない。ただ、乾先生が言うとおり「本当にほかのロボットに使えるのか」という汎用性の問題が起こってくる。その表現形態を考えるうえでも、記号というキーワードがもう一回出てくると思う。

長井　CREST「記号創発ロボティクスによる人間機械コラボレーション基盤創成」での一つの課題は、いろいろな家庭で学習したことを共有して、ある程度の背景知識や、ある種の事前知識、Prior のようなものを持っている状態で働き始めることができ、そこから適応していくというもの。そのため、ある種のシンボルシステムをいかにちゃんと作るかということが本題であり、そのときに、身体まで違ってしまうと相当難しいが、環境が違うくらいであれば、ある程度実現できる。

さらにそこから、身体が違うときはどうかということになるが、言語、記号化がすごく重要であり、何か言葉で説明されると、その言葉を軸にして自分の行為を予測できるので、その方向でやってみようと考えている。試行錯誤をする際、記号的な上位の知識を使うことにより、探索するスペースをすごく限ることができ、その中でうまくできていくようなプロセスが考えられる。

麻生　先ほど言及された 5G 通信を介して遠隔から工場で働く、といったサービスからデータを取得するというのもあると思う。また、3D-CG の映画にはモデルがあり、それぞれのエージェント・登場人物が、どういう環境で、何を見て、何を聞いて、どういう姿勢をとっているかということをつくり出している。そうしたものもデータを取得の候補になると思う。知り合いのソウル大学の研究者が、ビデオの内容についての Q&A をして、それが人間と区別できるかどうかを判定する「ビデオチューリングテスト」を提唱している。これをトイ・ストーリーのような 3D-CG の映画ですれば、「ロボットの対話はすごくシンプルなも

のばかりだ」という批判も少しは避けられるのではないか。

橋田 そういうものは、ゲームの方が使いやすいのではないか。

麻生 ゲームの方が良いかもしれない。だから Google DeepMind はゲームをやっている。

尾形 ネットゲームから物理的世界を介した言語インタラクションのデータを集めたいとスクエア・エニックスの研究者と相談したことがある。

国際競争力および応用分野

福島 論点を変えて、国際競争力についてもお聞きしたい。日本がこのような取り組みにおいて、どれくらい強みがあるか、強みを出すポイントはどこか。現状では、データに関しては強みがまだ出ていないが、今のうちに取りかかれば先行できるということもあると思う。

尾形 あくまで私見だが、ロボットに関しては、米国の基本はアカデミックな研究であり、多くはシミュレーターを中心にして研究が行われている。一方、日本はロボットの生産台数の 6 割を占めており、市場を大きく押さえてというところが強みとして挙げられる。しかも、日本の企業は深層学習の導入について、比較的積極的に取り組んでおり、既にサービスを提供するというところも出てきている。ただし、あまり統一感はない。まだ投入し始めているところであり、ロボットと AI の両方をやれる人材が必要となるだろう。また、研究の中心になるようなセンターや研究組織が出てきて、いろいろ企業を取り込んでいくような取り組みをもっと積極的にやるべきだと思う。

中国はこの 2～3 年でロボットの生産台数を 10 倍程度増やしている。彼らは機械学習についても強みがあり、今はそれほど目立った成果は出していないが、今後、間違いなく主要なプレイヤーになってくる。

ただし、先ほどの私の発表で示した粉体秤量ロボットのように、人が現在もせざるを得ない作業をロボット化するところに実際に取り組んでいるのは日本の企業であり、ロボットに関しては、AI を取り込むことで十分にアドバンテージというのは出せると考えている。

乾 言語については、まだできていないことのほうが多い。特に文脈を捉えて、あるいは常識、論理付けといったものを使って、業界用語を理解したり、あるいは推論したりするといったことについて、いろいろなデータセットを作って取り組もうとしているができていることは、ものすごくシャローだ。まさにこれからの分野だ。本当の推論のようなことを目指している人はそれほど多くない。したがって、研究者がちゃんと集まって今やっていけば、大きなゲインが得られるというふうに思っている。

それから、教育のアセスメントのような話というのは、本当にショートタームなゴールも、非常に難しいロングタームなゴールも設定できる、非常にいいフィールドだと思っている。どの省庁の政策における「AI プラス何とか」の中に「教育」はあまり出てこない。また、教育についてはデータの流通が未だにほとんどない。GAFA が握っているわけでもない。そのようなデータを手に入れさえすれば、できる領域というのはまだまだたくさん残っている。教育は一つの例にすぎないが、そういうところをしっかり狙っていくことが重要ではないか。ただし、今できている AI で何かをするのではなく、本ワークショップで議論してきたように、ロボットや推論といった、これからやらなくてはいけない技術を用いて、ゲインが出そうなところを狙っていくべきだと思っている。

福島　きょうの議論の中でも、知覚や行動から言語につながることで、共有が進むといった効果もあるということなので、教育やスキル伝播にもつながるのではないか。ロボットや自然言語も含めた対話などは、この手の技術が生きるところだと思っている。教育も確かに大きい分野かもしれない。AI が社会産業インパクトをもたらすに分野として、教育、ロボット、対話といったものが挙げられたが、それらについて補足なり、特に重要だといったコメントがあれば頂きたい。

橋田　自動運転の研究が盛んだが、割とプリミティブな内容であり、自然言語対話みたいなところまで進んでいない。将棋とか囲碁のようなゼロサムゲームと異なり、自然言語対話や自動運転は、お互いにウイン・ウインの関係をつくるものだ。車同士がぶつからないようにお互い道を譲るとか、ほかの車が自分の意図を酌みやすいように急ハンドルを避けるとか、そのようなコラボレーションとかゲーム理論的なところまでは、まだまだ行っていない。恐らくそこまで行かないと、特に市街地などでの自動運転は実用化できないのではないかと思う。そうなってくると、今の自動運転の研究でやっているように、いろいろな機能モジュールと、ほかのいろいろなモジュールとの連携が必要となり、先ほどの議論と関連が出てくる。産業上重要なところとしては、そういうのもあるのではないか。

福島　そのようなエージェント間交渉については、NEC の森永聡氏らが中心になって立ち上げた企業横断の活動があり、標準化も含めた現実的な取り組みを進めているが、さらに上のランクとして、「意味」まで含めた交渉や対話といったものもあると思う。

麻生　先ほど話題に挙がった、教育が重要というのはそのとおり。熟練者、エキスパートをたくさん育てるのも広い意味での教育であり、社会的にとても重要なことだ。私の発表では、社会の中の知識の増殖について述べたが、あれはまさにそういうエキスパート教育を加速するための技術。実現までにはまだ距離があるが、それがサービス化でき、そこからデータが取得できるようになると面白い。乾先生が進めていることはテキストの範囲かもしれないが、良いモデルになると思う。

木村　乾先生が言ったような教育というのは、利益にならないので GAFA は絶対やらないところ。そこを日本がやればいいと感じる。それをサービス化し、そこからデータをとってくると、かなり継続的に面白いことができるのではないか。

橋田　関連する話題として、教育に e ポートフォリオ、電子学習録を導入すると文科省が言い出している。小中高から、大学卒業後、就職後まで継続的に、各個人が自分の学習データを持ち、生涯学習等いろいろなことに活用できるようにしようというものだ。もしそれがうまくいけば、e ラーニングや、普段のいろいろな行動データ等も含めて本人のところに集約されることになり、本人の同意のみでデータを収集・分析できるようになる。ぜひそういうふうにしたい。JST から、省庁間連携まで上手く持っていけると良いと思う。

木村　教育関係の会社は、小学校から高校、大学までの教育データを持っている。これを活用すべく随分取り組んだが、なかなか難しかった。

橋田　教育データはあるところにはある。文科省には、まず教育データのポータビリティーを国策として提唱して欲しい。

木村　乾先生が言われたことは、現在問題となっている大学入試の品質を高めるためにすごく役に立つと思う。例えば、人間は問題の内容が分からなくても、日時を聞いている問題だと思えば、日時に関連する回答を選ぶ。東ロボはそのようなやり方で、センター試験で高得

点をとっている。これに対して記述式の試験を導入しなくても、本当に考えなければならないような問題を設定すれば解決できる。教育とか医療というのは、儲からないが、国としてはやらなくてはならない。

会場　すばらしい教育システムができたときに、どのような教育をするのだろうか。フェイクニュースなど、いろいろ問題があるが、ぜひ陰の部分も考えていただきたい。

会場　NEDO では人とともに進化する AI について公募予告を行ったところだ。AI と人がお互いに高め合い、共生していくことにより、よい社会をつくりたいという内容だが、今日の議論とすごく関連があると思った。AI については多様性がある研究が進んで、そこに理論的な研究もサポートしていけば、とても面白いのではないか。

福島　最近、JST と NEDO との間で、頻繁に議論、交流をしている。JST では次世代 AI への飛躍をもたらすような基礎的な研究を中長期的に推進することにウエイトを置きつつ、その成果が NEDO プロジェクトで広く展開・活用できるように連携していけるとよいと思っている。

木村　今日の発表での示唆や議論を戦略プロポーザルに活用していきたい。

4．まとめ

科学技術未来戦略ワークショップ「深層学習と知識・記号推論の融合による AI 基盤技術の発展」を開催した。5 件の発表と総合討議を通して、次世代 AI の中核的な技術チャレンジや社会・産業インパクトについて議論した。その要点は、以下の通りである。

1 件目の麻生英樹氏による発表「自然知能と人工知能の共進化－「知の創生」に向けて－」では、これまでの AI 研究の 2 つの流れや課題を概観しつつ、データから学習する即応的な知能と、記号的知識を使って推論・試行する熟考的知能の融合に向けて、注目する取り組みや考え方が幅広く紹介された。

2 件目の尾形哲也氏による発表「深層学習モデルのロボットへの応用」では、ロボットへの深層学習の適用を通して知能を考えるというスタンスから、模倣予測学習や連続動作学習等の具体的な応用事例を交えつつ、ロボットにおける記号接地問題や動作の分節化の現状と課題が示された。

3 件目の乾健太郎氏による発表「自然言語理解への道のり」では、現在の自然言語処理技術は、文章読解のベンチマークで一見高いスコアが得られていても、文脈理解・常識推論を含め、真の理解はいまだできておらず、今後、生のテキストの処理と知識ベースによる高次の処理を共通構造上に融合させる方向への発展が必要なことが示唆された。

4 件目の橋田浩一氏による発表「知能の基本構成素としてのサイクル＝価値＝意味」では、仮説検証サイクルが、意味（価値）の維持・向上・創造に重要な役割を果たし、このサイクルを組み込んだ循環ニューラルネットワークや、サイクルをベースとしたモジュラリティーが今後向かう方向になるとの示唆がなされた。

5 件目の長井隆行氏による発表「マルチモーダル確率的生成モデルと汎用知能」では、記号世界の正解ラベルに実世界のものを結びつけるという記号接地問題の捉え方は適切でなく、実世界とのインタラクションや社会的な共有の中で記号が創発されるという記号創発ロボティクスの考え方、汎用的なロボット・知能へのアプローチ、そこでキーとなるマルチモーダル確率的生成モデルが示された。

そして、総合討論では、まず、深層学習と知識・記号推論の融合、あるいは、知覚・運動系の即応的知能（System 1）から言語・論理系の熟考的知能（System 2）までを統一的に扱う仕組みをどう捉えるかを論じた。次に、この方向で研究開発を推進するうえで、新たにエクスペリエンス（経験）のデータ化が重要になることが指摘された。エクスペリエンスデータは、実世界における行為の連鎖に関わるデータの束（五感や発話を含む行為者や環境の状態の観測データ群）である。さらに、国際競争力という面に関して、まだ皆がスタートラインに立った段階であり、日本にもチャンスがあること、および、ロボット、自然言語、教育といった切り口から産業応用が広がる可能性が示唆された。

ここで示された方向性や知見を取り込み、戦略プロポーザルをまとめていく所存である。

付録

付録１　開催概要

ワークショップ名称：

科学技術未来戦略ワークショップ
「深層学習と知識・記号推論の融合による AI 基盤技術の発展」

開催趣旨：

深層学習が第 3 次 AI ブームを牽引している。画像認識、文書読解、囲碁・ポーカー等、さまざまな特定タスクで人間を上回る結果を出して話題になり、分類・予測・異常検知等の用途で、さまざまな産業応用も広がっている。その反面、大量の教師データや計算リソースが必要であること、想定外の状況での振る舞いが分からないこと、理由の説明が難しいこと等、深層学習の限界も指摘されるようになった。

このような限界を克服する次世代 AI の方向性として、深層学習に記号推論を融合し、認識・行動のレベルから言語・推論のレベルまでを扱える仕組みが検討されつつある。ここでは、深層学習と知識・記号処理の単純な組み合わせ・併用ではなく、より本質的な融合が必要になる。

そこで、本ワークショップでは特に、次の 2 つの論点について掘り下げる。

【論点 1】　次世代 AI の中核的な技術チャレンジは何か？
（深層学習＋記号推論 ～ 記号接地問題へのアプローチ等）

【論点 2】　論点 1 がもたらす社会・産業インパクトは何か？
（ロボットおよび自然言語処理が有望と見込まれる等）

開催日時・場所：

2020 年 1 月 30 日（木）13:30～18:30
東京都千代田区五番町 7　K's 五番町　JST 東京本部別館　2 階セミナー室

プログラム：

13:30～14:00　CRDS からの趣旨説明
14:00～14:30　麻生英樹先生からの発表と質疑
14:30～15:00　尾形哲也先生からの発表と質疑
15:00～15:15　休憩
15:15～15:45　乾健太郎先生からの発表と質疑
15:45～16:15　橋田浩一先生からの発表と質疑
16:15～16:45　長井隆行先生からの発表と質疑
16:45～17:00　休憩
17:00～18:30　総合討議

付録２　参加者

招聘者：

麻生英樹	産業技術総合研究所人工知能研究センター 副研究センター長
尾形哲也	早稲田大学基幹理工学部表現工学科 教授、産業技術総合研究所人工知能研究センター 特定フェロー
乾健太郎	東北大学大学院情報科学研究科 教授、理化学研究所革新知能統合研究センター 自然言語理解チームリーダー
橋田浩一	東京大学大学院情報理工学系研究科附属ソーシャル ICT 研究センター 教授
長井隆行	大阪大学大学院基礎工学研究科 教授、電気通信大学大学院情報理工学研究科機械知能システム学専攻 教授

チームメンバー：

奥付記載の通り

傍聴参加者：

文部科学省	3 名
経済産業省	2 名
新エネルギー・産業技術総合開発機構（NEDO）	5 名
科学技術振興機構（JST）	6 名（上記チームメンバー以外）

■ワークショップ企画・報告書編纂メンバー■

総括責任者：	木村 康則	上席フェロー	（システム・情報科学技術ユニット）
チームリーダー：	福島 俊一	フェロー	（システム・情報科学技術ユニット）
チームメンバー：	東 良太	フェロー	（システム・情報科学技術ユニット）
	井上 貴文	フェロー	（ライフサイエンス・臨床医学ユニット）
	奥山 隼人	主査	（未来創造研究開発推進部）
	小林 容子	主任調査員	（戦略研究推進部）
	坂本 明史	主任調査員	（戦略研究推進部）
	嶋田 義皓	フェロー	（システム・情報科学技術ユニット）
	的場 正憲	フェロー	（システム・情報科学技術ユニット）
	山田 直史	主査	（経営企画部）

※お問い合せ等は下記までお願いします。

CRDS-FY2019-WR-08
科学技術未来戦略ワークショップ報告書

深層学習と知識・記号推論の融合による ＡＩ基盤技術の発展

令和 2 年 3 月　March 2020
ISBN 978-4-88890-674-6

国立研究開発法人科学技術振興機構　研究開発戦略センター
Center for Research and Development Strategy, Japan Science and Technology Agency

〒102-0076　東京都千代田区五番町 7　K's 五番町
電　話　03-5214-7481
E-mail　crds@jst.go.jp
https://www.jst.go.jp/crds/

ISBN 978-4-88890-674-6