人工知能と言語処理の現状と展望

Present and Future of AI and Natural Language Processing

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

人工知能の研究分野は、過去 10 年、大きな技術革新の時代を迎えている。この技術革新は、機械学習、とくに深層学習の進展によって引き起こされてきた。本稿では、この変革とは、どのようなもので、また、それが言語処理の技術にどのような影響を与えてきたかを考えよう。

人工知能やロボットの技術は、(1)外界からのセンシングデータを解釈して、それらを意味に結び付ける認識処理、(2)外界の状況と自らの目的とを結びつけて推論、問題解決を行う思考処理、(3)外界に対して働きかける行動処理、の３つに分けて考えることができる。言語の処理も、単語列という入力データを意味に結び付け、その意味と背景知識に基づく推論処理を行い、その結果を出力する。対話システムの場合には、回答を生成し出力する部分が、外界への働きかけを行う行動処理に対応する。機械翻訳の場合は、外界への働きかけ部分は、相手言語の文を生成する処理である。

言語処理の場合には、意味に結び付けられる入力データが文やテキストで、その基本単位が単語という記号的なものであるために、画像や音声データの処理とは違った性質のものとなる。また、文にはその文法に基づいた構造がある。前述のように、結果としての行動も回答文であったり、相手言語での翻訳文であったりで、物理世界での行動というよりは、記号的なものとなることが多い。

2　人工知能技術の枠組み

古典的な人工知能システムの枠組みを図 1 に示す。この枠組みにうまくあっていて、理解しやすいのは自分

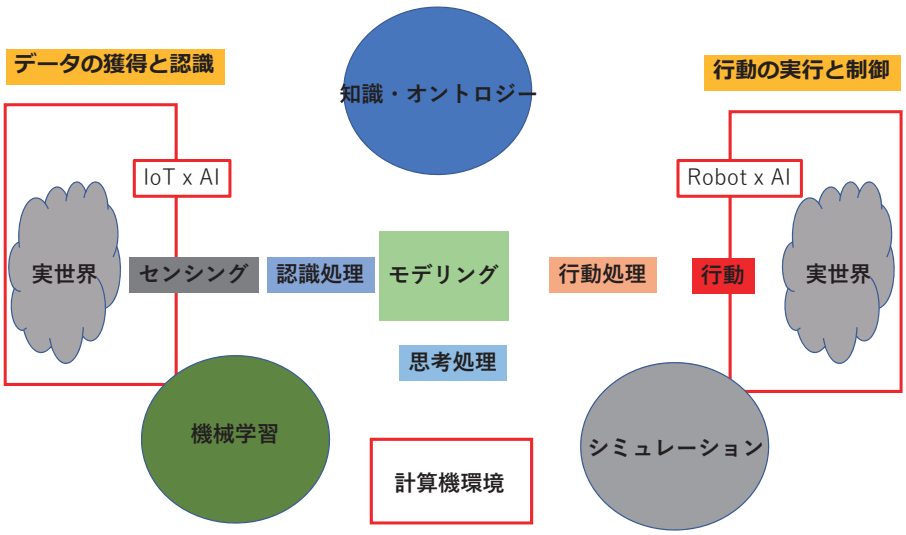


図 1　人工知能の要素と技術的な基盤

で自律的に行動するロボットであろう。たとえば、自走ロボットは、目、耳などの感覚器を持っており、この感覚器からのデータを解釈して、自分が置かれている環境を認識する。目からの感覚データは、単純化すると、写真を計算機に入力したようなものを考えればよい。入力データは、例えば、2 次元の画面の各点ごとに三原色の明るさが数値化されたものである。

入力データは、3 つの数値データの組が画面上の各点に与えられたものである。この入力データから、画面のどの部分に人が写っているか、どの部分に壁があるか、などを認識するのが認識処理である。単なる数値の組（ベクトル）の羅列であったものに、画面の領域ごとに「人」、「壁」、「通路」、「ドア」といった意味を与える。これが認識処理である。

感覚器からのデータが解釈され意味が与えられると、ロボットは、たとえば、移動している「人」を避けて行動し、「壁」ではなく「通路」を移動して、自分の目的地へと移動することができる。自分の目的と自分の置かれた環境を考えて、どう行動すべきかを思考できる。これが図の中央にある思考処理、推論処理である。思考処理の結果として作られたプランに従って、実際に手や足を動かすのが行動系の処理である。

3 認識処理と深層学習、End-to-End のシステム

現在の人工知能の発展の原動力となっているのは深層学習の技術であるが、これが最初に大きな成功を収めたのは、認識処理の部分であった。画像を集めて、それらの画像に何が写っているかを教えるお手本データ（訓練データとよぶ）を大量に作ると、深層学習のシステムは、そのお手本データから学習し、新たな画像に何が写っているかを認識することができる。

画像データを入力、その画像に写っているもののカテゴリ（たとえば、車、人、壁、など）を出力とし、このような入力と出力の対の集合が訓練データとなる。この訓練データを与えると、学習器が入力と出力をつなぐ潜在的な規則性（関数）を学習する。この学習の結果、新たな入力が与えられると、その学習された関数に従って、対応するカテゴリ名が出力できる。

入力と出力の対の訓練データを人間から与えられるだけで、それを結びつける関数がいかなるものかについては、人間からの教示や介在なしで学習できる。このようなシステムは、入力と出力の対の訓練データが与えられるだけで、学習器が自力で対の間にある関数関係を学習するという意味で、End-to-End のシステムと呼ばれる。

深層学習が広まる以前の画像認識では、例えば、「人」を認識するためには、どのような特徴の集合をまず見つけるのがよいか、次に、その特徴群がどのような関係で結びついているときに「人」と判断すべきかなど、データからどのような特徴を取り出し、それを認識に使うかを人間が設計することで、認識処理のシステムを作るのが普通であった。入力と出力の間を結ぶのに有効な特徴のセットを人間があらかじめ設定し、その特徴を抽出するコンポーネントと特徴を使うコンポーネントとを結ぶという構成を人間が設計するという意味で、End-to-End のシステムではなかった。

何層にも層を積み重ねる深層学習は、どのような特徴群が役に立つか、その特徴群間の相互関係がどのようなものであれば、次のレベルでの特徴群になるかを自力で学習する。人間の介在なしで、例としての大量の対の集合だけから、入力と出力を結ぶ関数を学習できる（図 2）。

この End-To-End のシステムという考え方は、必ずしも、認識系だけに適用されるわけではない。図 1 では、人工知能の一般な構成として、認識処理―思考処理―行動処理という 3 つの過程があると説明した。しかし、

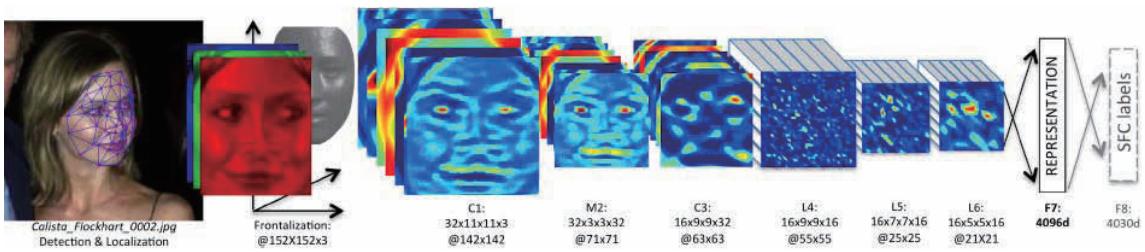


図 2　多層のニューラルネットによる画像処理
［Taigman+: Deep Face: Closing the Gap to Human-Level Performance in Face Verification, 2014 より引用］

入力として認識処理への入力データ、出力として行動処理の結果としての行動を考えて、この対を訓練データとして与えることで、入力データと出力行動とを結びつける関数関係が学習させることができれば、自走ロボット、対話システム、機械翻訳システムにも、End-To-End のシステムとして構成できることになる。

実際、ニューラル機械翻訳の場合には、原言語文（入力文）と相手言語文（出力文）の対を大量に用意するだけで、その間にある関数関係をシステム側が学習する。あるいは、対話の場合には、人間同士の現実のやり取りを入力―出力対の訓練データとして用意することで、一方の人間と同じような応答を出力する End-to-End な対話システムを構成することも考えられる。

このような End-To-End のシステムは、原言語の文の構造をまず認識し、それをもとに相手言語の構造を計算、次にその構造に従って出力文を生成するという、言語構造を使った移行方式の機械翻訳システムとは大きく異なる。また、原言語と相手言語の対応の確率的な関係、および、相手言語の確率的な性質を学習し、この２つの確率モデルを結合することで、入力文に対して最適となる相手言語の文（翻訳文）を選択する統計的な機械翻訳とも異なったものになっている。

4 End-to-End システムとその設計

End-to-End のシステムでは、入力と出力の対を訓練データとして与えることで、人間の介在は必要がない、とした。この言い方は、多少、誤解のある言い方になっている。実際には、画像の認識処理に使われる CNN（Convolution Neural Network）でも、たとえば、ニューラルネットの層を実際に何層にすればよいのかなどは、解くべき画像認識のタスクに応じて、システム設計者が決める。

言語のように基本要素（たとえば、単語）の並び方に意味のあるデータ、すなわち、一次元の要素の並び方に意味のあるデータの認識処理には、画像処理で有効であった CNN よりも、RNN（Recurrent Neural Network）が有効である。

機械翻訳や対話システムでは、この傾向はさらに顕著になる。機械翻訳の場合、原言語の文（入力文）を認識する部分と相手言語の文（翻訳文）を生成する二つの性

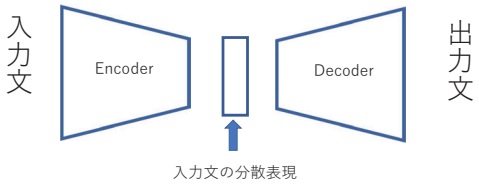


図３　Encoder-Decoder モデル

質の違う処理がある。これの性質の違いは、Encoder-Decoder モデルという、認識と生成を受け持つ２つのニューラルネットの結合でとらえられる（図 3）。

さらには、翻訳の場合には、入力のある部分と出力文のある部分とが対応するといった規則性をとらえる必要があり、Encoder-Decoder モデルにおける Encoder 部分と Decoder 部分は完全に独立ではない。このような Encoder-Decoder の内的な関係をとらえるために、LSTM という別のニューラルネットとそのうえでの Attention 機構という機構が持ち込こまれる。この Attention の機構は、翻訳だけでなく、入力画像からそのキャプション（テキスト）を出力するタスクのように、入力の部分と出力の部分との相互関係をとらえたいタスクに有効である。画像からそのキャプションを出力するシステムでは、入力の画像データの部分と出力のテキスト・データの部分とが、相互関係している。

このようなデータの構造を取り扱う必要は、言語に関係するタスクには普通に見られる。言語には、単語が句にまとまり、句はさらにそれを含むより大きな単位としての句や節に、そして、最終的に文の単位にまとまる。「言語に文法がある」という意味は、文には、単語並びが小さな単位からより大きな単位へとまとまっていく階層構造があること、また、任意の単語並びや句の並びがより大きな単位にまとまるわけではなく、まとまるための規則があること、をいう。

単語並びの持つ文法的な階層構造は、単語並びの意味とも関係している。たとえば、

The premier of the western Canadian province of British Columbia

という単語の並びは、図４のような階層構造を持ち、この階層構造から、句全体としては、階層構造の最上位にある Premier が句全体の意味を規定し、句全体としては、特定の人物を指しているが分かる。

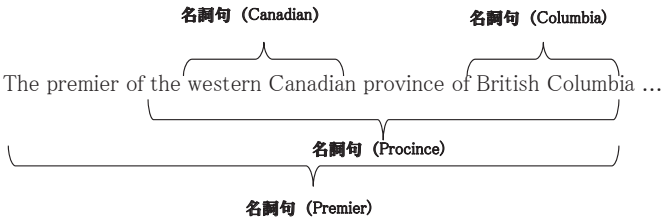
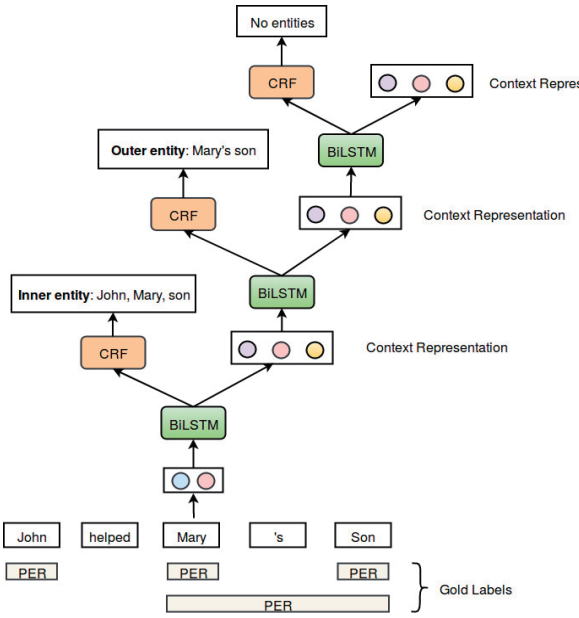


図４　単語列の持つ階層構造



Entity Extraction Order: inside to outside

Layer Stacking Order: bottom to up

Dynamics: depth of the model is determined by nested level of entities in the input

Terminate Stacking Layer Condition: no entities are detected

図５　NER の構成

言語の文がこのような階層構造を持つことを、文中から人物を指示している句を取り出すプログラム（NER-Named Entity Recognizer という）では、図５のようなニューラルネットを処理中に動的に積み重ねる形で利用する。

このように、End-To-End のシステムも、入力と出力の対という訓練データを準備すれば、あとは全能で汎用なシステムが学習してくれるというわけではなく、対象とするタスクによって、どのようなニューラルネットワークを定義すればよいかを人間（システム開発者）があらかじめ設計する必要がある。

深層学習でネットワークの構造を定義することは、深層学習という計算機構を前提として、一種のプログラムを記述することに対応している。

システム開発者は、実行させたい個々のタスクに関する自らの知見を反映したネットワーク構造を定義する。実際、深層学習のネットワーク構造を定義するツールは、Google の TensorFlow をはじめとして、多くのものが流通している。これらのツールは、深層学習という計算の枠組みを前提とした一種のプログラム言語とみなすことができる。

従来のプログラム言語は、機械語・アセンブリ言語から出発して、計算過程に関する様々な抽象化を経て、高級なプログラム言語に至った。その抽象化のやり方に応じて、様々なシステム開発手法やエンジニアリング手法も作られてきた。

これに対して、深層学習という計算の枠組みは、これまでの von-Neuman 型の計算原理、あるいは、関数型や論理型の計算原理とは大きく異なるものであり、あとの節で述べるように、そのためのプログラム言語やエンジニアリング方法論も、その発展途上にあると考えてよい。

5　End-to-End のもつ限界

最近、話題になっているトピックに、D. Kahneman のいう Slow な思考と Fast な思考の議論がある。前者は熟考型、後者は反射的・直観的な思考といってよいだろう。人間が何か判断をする場合にも、自覚的な説明はうまくできないが非常に素早く直観的に行う判断と、自

覚的に、いわば、論理的で系列的な推論を重ねながら行う判断とがある。

ある画像を見て、そこに写っているもののカテゴリを認識する認識処理は、反射的・直観的で非常に素早く行われる。あるいは、自転車に乗るとか、泳ぐといった行動の場合も、試行錯誤を繰り返すことによって、他人にはうまく説明はできないが、反射的に行えるようになる。

これに対して、例えば、物理の問題を解くといった場合を考えると、問題文を理解し、それを数学的に定式化し、数式を操作することで問題を解く。この過程では、反射だけでは説明できない、自覚的な思考が関与している。

言語による対話、大きなテキストからの要約、翻訳といったタスクには、この２つのプロセスが複雑に絡み合いながら関与しているように見える。

図１の人工知能の全体図で、認識や行動という部分は思考というものとは一応別のものとしていた。

深層学習以前の人工知能の分野は、認識や行動の部分は、パターン認識やロボティクスの分野として人工知能の中核分野とは考えられていなかった。言語処理も、単語の並びから文が持つ階層構造を認識する処理は、言語処理・計算言語学という独立の分野を作り、人工知能からは周辺的な分野とされた。入力データが意味に結び付いた後、その意味の世界で人間の心が行う計算の過程を対象とするのが人工知能の中核的な研究、とみなされていた。このような立場は、心の中での計算過程を研究するという意味で、認知主義の人工知能と呼ばれる。

深層学習の発展が契機となった現在の人工知能は、認識と行動の処理という人工知能の周辺と考えられてきた分野を復権し、これを人工知能の中核におくこととなった。入力と出力の対という、外から観察できることから、心の計算を解明しようという、行動主義の人工知能といってよい。認知主義の人工知能は Slow な思考、行動主義の人工知能は Fast な思考に重点をおいてきたといってよいだろう。

翻訳は、Slow と Fast のいずれのプロセスで行われているのだろうか？　また、End-to-End のシステムの限界はあるのだろうか？

よく似た言語同士の翻訳をプロの翻訳家が行う場合、例えば、日本語と韓国語、イタリア語とフランス語といった翻訳を行う場合には、タイピングのスピードが翻訳のスピードを決める、といわれる。これらの似た言語同士の翻訳をプロの翻訳家が行う場合には、ほとんど、反射的に翻訳が実行でき、入力の文を意味に結び付け、その意味を理解した上での Slow な自覚的な処理を経って翻訳を行う必要はない。

これに対して、英語と日本語の翻訳の場合には、プロの翻訳家の場合でも、一般的には、タイピングのスピードで翻訳を行うことはできない。文を読んで、それを相手言語で言い換える間には、かなり自覚的なプロセスが介在する。これには、動詞のような述語が最後に来る日本語と主語のすぐ後に来る英語の語順のために、文の構造自体をかなり変えないといけない、という事情もあろう。ただ、このような構造に関する差異の調整だけでなく、入力文の意味を理解し、どのような事柄が記述されているのかを理解し、その結果を相手言語で表現することで、翻訳文をつくるという処理が行われている。このような処理が介在している場合には、背景知識という入力データにはない情報の活用が必要で、End-to-End の現在のニューラル翻訳では不可能な処理となる。

例えば、日本語を英語に翻訳する場合に、日本語では省略されている語句を前後の文脈から補う必要がある場合など、一文単位の境界を越えた情報を参照する場合には、このような明示的な理解が必要となる。

また、語族が近い言語同士の場合には、２つの言語で単語の語源を共有している場合が多い。このため、２つの言語で同じことを言い表す、語源を共有した単語があり、翻訳家が文で記述されている内容を理解していなくても、相手言語で対応する単語に置き換えるだけで、読み手は困難なく理解できる。これに対して、日本語と英語のように語族が異なる場合には、単語の語源自体が全く異なっているために単語同士の対応がかなり複雑なものとなる。翻訳家は、文で記述されている内容を理解して、その内容を適切に言い表す相手言語での単語を選択しなければならない。この訳語の選択には、理解という過程が関与する。

6　専門テキストの翻訳

現在の End-to-End の翻訳システムを使ってみると、テキストの種類によって、翻訳の質がかなり異なることに気が付く。

特許のような専門用語が多い場合には、専門用語は言語族が異なっても実は語源を共有し、一対一対応するものが多い。この場合には、訳語の選択に理解というプロセスを経る必要がない。したがって、理解を経ない翻訳を行う現在のニューラル翻訳で取り扱いが容易な翻訳になっている。

ただ、現在のニューラル翻訳の場合には、訓練データに出現しない専門用語にはうまく対応できないという欠点を持っている。このため、現状のニューラル翻訳は、専門用語の訳出について技術的な困難を抱えているが、この困難は、翻訳に理解プロセスが関与する本質的な困難さとは別種のものである。したがって、現在の技術の延長上で解決できる問題だと考えられる。実際、このような未知語に対する処理は、現在、活発に研究が進められている。

しかしながら、専門性の高いテキストの翻訳であっても、本当に質の良い翻訳が現在の技術の延長でできることには、私は、否定的である。熟練した翻訳家の場合でも、取り扱う分野についての知識に欠けている場合には、質の高い翻訳を行うことはできない。質の高い翻訳を作る場合には、翻訳家はその分野についての背景知識を丹念に調べたり、分野専門家からの情報提供を受けたりする必要がある。わかりやすい、質の高い翻訳を作る場合には、明らかに意味と背景知識に基づいた理解のプロセスが関与している。

現在の対話システムやテキスト抄録についても、機械翻訳と同様なことが言える。コールセンターなどでのユーザからの質問とそれへのオペレータの回答の対を大量に収集すれば、訓練データを作ることができる。こうして作られた End-to-End のシステムは、新たなユーザからの問い合わせに対して、それなりの回答を作りだすことができる。基本的な考え方は、ニューラル翻訳と変わらない。入力文に対応するものがユーザからの質問文、翻訳文に相当するものが回答文となる。

このような対話システムも、ユーザからの質問内容を理解しているわけではない。したがって、込み入った質問内容、たとえば、それまでの応答過程で状況が議論された後で、その状況下での質問が出された場合に、状況に応じた適切な回答を返す、といったことはできない。

人間のオペレータの場合には、ユーザとのやり取りの過程で、ユーザの状況を理解し、その理解に基づいて回答を行う。

人間のオペレータと同様な振る舞いをするシステムを構成するためには、現在の End-to-End のシステムに、当該の Call センターが対象としている業務に関するモデルとそのモデルとの関係でユーザとのやり取りを理解するといった、現在の End-to-End のシステムにはない、かなりの追加機能を持たせる必要がある。

7 今後―その 1　AI エンジニアリング

画像認識システム、人物同定システム、ニューラル翻訳システム、対話システムなどは、5 年前のそれらと比べると長足の進歩を遂げている。この方向での研究をさらに進展させていくことは、今後の技術開発の一つの方向であろう。

5 節、6 節では、語族が異なる言語の翻訳において、一つの単語に対して多くの訳語が存在する場合、訳語選択の問題があると述べた。ただ、この訳語選択の問題も、必ずしも理解プロセスを経なければ解決できないものだけではない。文中でその単語の周辺に表れている単語に手がかりがある場合には、入力文と翻訳文の対を大量に集めた訓練データから、当該単語の周辺情報と適切な訳語の関係が学習され、現在の技術で解けることも多い。日本語では省略されている情報を補う問題も、一文中のほかの箇所や直前の文といった局所的な文脈中に省略された語が現れている場合には、現在の技術の延長で解ける。

すなわち、4 節で述べた NER のモデルのように、我々が持っている言語に関する知識、すなわち、言語が持つ構造や規則性をニューラルネットワークの構造設計に反映することで、明示的な理解のプロセスを経ることなく、翻訳や対話システムの性能を向上させることができる。

4 節に述べたように、深層学習でのニューラルネットワークの構造を記述するツール、Google の TensorFlow などのツールは、深層学習という計算原理を基にした一種のプログラム言語だとみなすことができる。翻訳、対話システムを構築するためには、このプログラム言語を使って、翻訳や対話の研究者が自らの知見をシステムの構造として記述していくことになる。

現在の深層学習のためのツールは、対象タスクに関して専門家が自らの知見を簡単に表現できるという段階ま

で成熟してはいない。従来のプログラム言語が、さまざまな抽象化によって高級言語になっていたように、これらのツールを土台にしてより抽象度の高いモデル記述言語を作り出していく必要がある。

また、従来のプログラム言語には、あらかじめ開発された様々なパッケージやモジュールが用意され、これらのパッケージを組み合わせることでより複雑なシステムを作っていくことができる。

入力と出力の対だけから、内部の構造を学習で作り上げる End-To-End のシステムでも、あらかじめ構築されているモジュールを組み合わせて全体を作り上げるための方法論、すでに構築された学習ずみモデルを再利用していく有効な手法を確立していく必要があろう。

また、現在の深層学習では、モデルの構造だけでなく、様々なハイパーパラメータの調整によって、その性能が大きく変化する。この調整は、現在のところ、システム設計者の経験と勘による部分が大きい。従来のプログラム言語では、効率上の最適化をシステム側に任せられたように、これらの経験と勘によるハイパーパラメータやモデル構造の最適な調整がより系統的に整理され、可能であれば、この性能の最適化をシステム側が自動的に行うようにする必要があろう。

8 今後―その２　明示的な理解

自走ロボットの例で述べたように、感覚器からのデータは認識処理を経て、意味と結びつけられる。この認識器は、入力として感覚器からのデータ、出力として認識結果のカテゴリを出力する処理である。これの認識処理は、End-to-End のシステムとして構成できる。この認識処理の結果を使って、自走ロボットでは、自らの目的地に人や壁との衝突を避けながら、移動していく別のシステム・コンポーネントが使われる。

では、感覚器からのデータから自走ロボットの行動まで、その全体を End-To-End のシステムとして構成できるであろうか？

実際には、自走ロボットの動作環境には、非常な多様性がある。環境にいる人の人数、その配置、壁などの位置などは、さまざまに変化する。また、その環境下でとる最適な行動も、周辺の環境と目的によって変化する。

周りにいる人間の動きを予測し、それとの衝突を避ける動きを取るためには、当該の人間の進む方向や視線の方向を考慮した予測処理が必要となる。この予測処理自体が直前の人間の移動方向と視線とを入力として、その移動先の予測を出力とするシステムとして定式化できる。このよう人の動きを予測する End-to-End の予測器を使うためには、人間の存在を認識する処理が先行していなければならない。言い換えると、ロボットが、自らが置かれた環境の構造を認識し、その構造を前提に周辺の人の動きを予測する必要がある。ここでは、周辺環境の明示的な理解とその理解に基づく推論処理が必要となる。

この例のように、明示的な理解の構造的な表現（例えば、自分の周辺環境のどの位置に人がいるか、自分の周辺に壁があるか、など）を前提にして、その構造中の個々の要素についての推論処理を行う必要があるタスクでは、現在の End-To-End のシステムの構成をそのままでは使えない。

入力データの多様性と出力の多様性が非常に大きな場合には、入力データの多様性を作る要因を部分とその構造に分け、それぞれに別の推論処理を行う必要がある。

このような明示的な理解の表現の必要性は、機械翻訳、対話、テキスト要約などの言語処理の場合にも、いえることである。より質の高い翻訳やユーザとの過去のやり取りに従って適切な応答を返すシステムを構築するためには、入力の単語列を文脈・意味・知識をそれぞれ内部的に表現し、その表現の上での推論処理とその統合を考えなければ、入力が持つ多様性と出力の多様性、その複雑な対応関係をとらえることはできないであろう。

End-to-End の言語処理の限界を超えるために、このような明示的な理解の表現をシステムの中にどのように入れ込むことができるか、これが次の大きな研究課題となろう。

9 おわりに

本稿では、現在の人工知能をけん引している深層学習と End-to-End のシステム構築が、機械翻訳、対話システム、自動抄録の技術にどのような変革をもたらしつつあるかを議論した。特に、深層学習に代表されるニューラルネットの技術は、機械翻訳や対話システムの性能を大きく進展させた現在、これらの技術は実世界で広く使われるようになっている。

この技術の延長で、性能はさらに向上していくことが期待できるが、その反面、End-To-End のシステム構成の限界も見え始めている。それをどのような技術で解消していくかが、次の大きな飛躍につながると思っている。認知主義の人工知能と行動主義の人工知能との融合が求められている。