自然言語処理技術の進化：AI による「ことば」の処理から汎用 AI へ
最近の動向について

概要

コンピュータ登場の最初期から人間の言葉そのものによるやり取りは夢の技術[1]となっていた。しかし人間の言語、すなわち自然言語は数学として定式化することが非常に難しいことがわかり、それはすなわちプログラミングはもとよりアルゴリズムとして実現することが非常に困難であることを物語っている。数々の技術発展があったものの自由に使えるような成果が得られたとは言えないまま今日に至っている。

ヒトの場合に言語の学習、例えば母国語や外国語学習を考えると、確かに学習方法がある程度重要ではあるものの基本的には「習うより慣れる」という表現がしっくりくることは認めざるを得ないところだろう。近年の深層学習は現在の AI 技術の中核を担っているといっても過言ではないが、これを自然言語処理に応用したところ、これまでにない成果が得られるようになった。例えば明示的な学習なしに四則演算が可能となり、複雑な画像の説明、または説明文から対応する画像生成が可能になるなど、深層学習を用いた膨大なテキストデータの学習結果からは当初予想していた以上の成果が報告されている。まさに学習器が膨大なテキストデータにより「習うより慣れる」を実践した結果とも見える。

本報告書では最新の自然言語処理が如何に「習うより慣れる」を実現したのか、技術の解説を行うのと同時に、ビジネスへの展開例、そして日々更新されているこの技術領域について可能な限りキャッチアップし紹介するものとなる。

第 1 章では近年の自然言語処理の概要を示す。膨大なテキストとパラメータが何を可能にしつつあるのか、発表後に大きな話題となった GPT-3[2]を例に説明する。特にキー技術の一つである Transformer について概要を読み解くことで直感的な理解を目指す。

第 2 章では、GPT-3 を使用した実際のサービスをいくつか紹介することで大規模言語モデルの実用的な側面を取り上げた。いくつかのサービスは実際にローンチしており AI の使用を明示しているとはいえ、その内容は、これまでにない言語でのやり取りが実現されており従来のサービスとは一線を画している。

第 3 章では、国内の取り組みに関して取り上げる。大規模言語モデルが英語中心に進んでいる中、国内の状況はどうなっているのか？最近の対話コンペティションの結果も合わせて簡単に紹介した。

---

[1] 1956 年に人工知能について初めて話し合いが行われたダートマス会議において自然言語処理について取り上げられている。

[2] GPT-3: Generative Pre-Training-3 の略で Open-AI が開発している汎用言語モデルの第 3 世代となる。https://beta.openai.com/docs/engines/gpt-3、 https://openai.com/blog/openai-api/

第 4 章では、大規模言語モデルが汎用 AI の基盤モデルの礎になるとの報告がある。人の活動はそのほとんどを言語で説明することができると言えるため、汎用言語モデルは人の活動のほとんどをモデル内に記述していることになる。そこで汎用 AI へつながる大規模自然言語モデルの可能性について米国 Stanford 大学の論文を基に紹介する。さらに実世界との直接のインタラクションという観点に注目し大規模言語モデルとロボティクスの関係について記す。

第 5 章では GPT-3 以降の自然言語処理技術の動向について海外の動向を中心に取り上げた。より大きなモデル構築やモデル規模を小さくしながらも性能を上げる方法等について日々更新があるなかスナップショットをまとめた。

第 6 章では、自然言語処理・自然言語モデルに関する重要な課題について述べる。言語は社会に大きな影響をもたらすことから改めて見直す。

第 7 章はまとめと今後の展望について記した。

執筆　専門委員　山本雅裕　2022 年 6 月 7 日

内容

1　はじめに 1
1.1 大きな進歩を支えるのは急激に増大したパラメータスケール 2
1.2 効率的な総当たりを可能にした Transformer によるブレークスルー 4
1.3　Transformer による自然言語処理の特徴 7
2 自然言語処理技術アプリケーション 12
2.1 GPT-3 のビジネス応用例 13
2.2 GPT-3 の応用技術 16
言葉による柔軟な画像生成 DALL･E と柔軟なキャプション付加 CLIP 16
Power Apps 18
3 国内の動向 19
3.1 日本語の対応 19
3.2 国内リソース 20
3.3 国内の話題 22
4　汎用 AI につながる可能性を秘めた大規模自然言語モデル 23
4.1 Foundation Models 23
4.2 自然言語モデルの本質 25
4.3 ロボット工学と大規模自然言語モデル 26
4.4 現実世界との対応 27
5 GPT-3 以降の自然言語モデル関連トピックの紹介 29
5.1 Microsoft と NVIDIA の強力な自然言語モデル MT-NLG 30
5.2 Microsoft Project Florence-VL 31
5.3 BigScience T0 32
5.4 Google の対話アプローチ LaMDA 33
5.5 DeepMind RETRO（Retrieval Enhanced TRansfOrmers） 34
5.6 韓国 LG EXAONE 35
5.7 中国 BAAI Wu Dao 2.0 37
5.8 OpenAI GLIDE、DALL･E2 37
5.9 MLP 再考、Transformer 以降 39
6 重要な課題 41
7 最後に 43
付録 46

1　はじめに

2018 年 Google は同社の開発会議において Duplex[3]と名付けた AI アシスタントを披露し、会場で実際にレストランを予約するというデモンストレーションを行った。Duplex のアプリを起動し、音声でレストランの指定、人数、日時を伝えると自動でそのレストランに電話予約を行ってくれる。ネット予約を導入していないレストランも数多くある中、Duplex が代理で電話予約を行うことで、すべてのレストランに予約が可能となる。

実際の一連のやり取りを聞けば、アプリとの対話は自然であり、その先、つまりレストランとの予約電話内での応答も十分自然であることが分かる。予約を受けたレストランはなんらとまどうことはなく、つまり Duplex が自動で一連のやり取りを行っていることを感じさせるようなことはない。実際、レストランで対応した人は相手が人間ではないと知ると驚くようである。

Duplex が発表された 2018 年には、それでも 25%は実際には人間が部分的な補助を行う必要があったのだが、2020 年の更新報告では、これまでに 100 万件の予約を完了させていることが紹介されると同時に、99%が AI のみの自動応答で予約が完了していることが報告された。

さらに同社は Duplex を Google マップと Google 検索のビジネス情報の自動更新にも利用していること、つまり、店舗の営業時間やテイクアウトをビジネスオーナーが提供しているかどうかなどの詳細な情報などが、もはや手動で更新することなく自動で更新されていることを報告したのである。

Duplex は予約するときの情報を Google 検索から調べているだけではなく、使用する情報そのものも自ら更新していることになる。今では映画のチケットの購入やレンタカーなどの予約など、より広範囲の手続きを簡単にするために、さらなる応用機能が追加されている。人間はアプリに自然な言葉で投げかける、すると AI はこれまでにない自然な言葉で応答を返しながらすべてを行ってくれるのである。

この Duplex を支えるキー技術の一つが自然言語処理である。日常の言語を用いてシステムとあたかも対話し、システムは自動で目的を果たし、結果を自然な言語で報告する、そのための技術の総称が自然言語処理である。実際、Duplex の例からは、自然言語処理の影響はインターフェースからビジネスロジックまで広範囲に及ぶことが分かる。

現在の自然言語処理では、言語による入力から始まり、処理、結果までの自動化はいうに及ばず、最新の技術ではたったひとつの単語から自然な文章生成すらできるようになっている。人の創造という作業さえ置き換えることが近い将来の現実として見えてきていると

---

[3] Google Duplex: A.I. Assistant Calls Local Businesses To Make Appointments
https://www.youtube.com/watch?v=D5VN56jQMWM

も言える。図 1 に自然言語処理の例（概略）を示す。

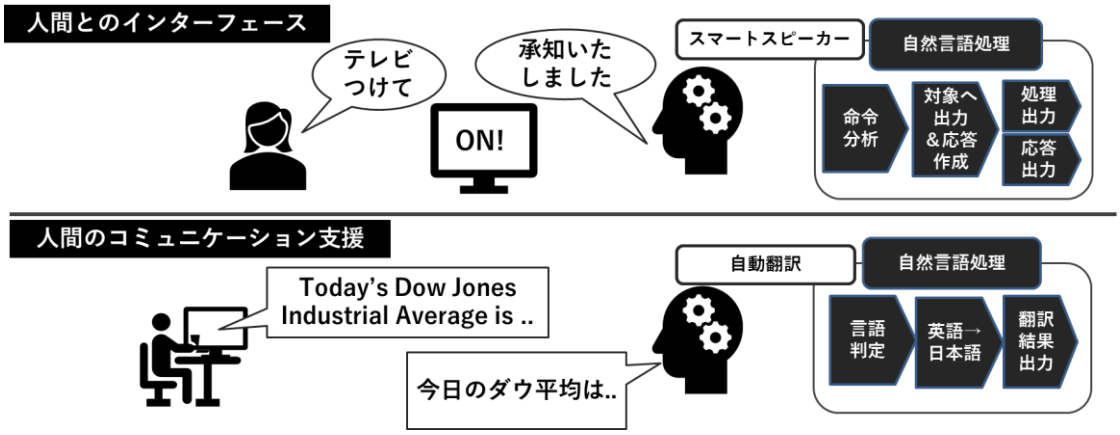


図 1 自然言語処理の例 （出典:DX 白書 2021 付録 AI 技術）

1.1 大きな進歩を支えるのは急激に増大したパラメータスケール

Duplex の機能や用途が拡大した背景の一つが新しい自然言語モデルにある。BERT[4]と名付けられたこの言語モデル[5]は、膨大なパラメータ[6]をもつ。自然言語処理の急激な発展の理由として、ひとつにはこの学習パラメータ数の急激な増大が挙げられる。近年の AI の中心技術である深層学習では膨大な数のパラメータを、それもとてつもなく巨大な数のパラメータを、これまでにない膨大なデータで学習させることにより性能を上げることが特徴となっている。

ここ数年の自然言語処理の性能は、特にこのパラメータ数の増大そのものが大きな進化をもたらしている。図 2 に技術エポックとなる自然言語処理の技術名と発表された時点でのパラメータ数を示した。きれいに指数関数で増大していることが確認できるように縦軸は対数とした。実にこの 3 年で言語モデルが使用するパラメータは 10 万倍に達している。現在 Google T5+においては 1.6 兆のパラメータとなっている。最新の Microsoft + NVIDIA の発表に至っては 5.3 兆ものパラメータに達した。

では、これだけのパラメータ数のニューラルネットワークをどのように学習させ最適化するのか。学習データをどのように用意するのか。これらの点は直感的にも、そして技術的にも非常に重要な点となることに異論はないだろう。例えば 2020 年にその性能で大きな話題をさらった GPT-3 においては 4 兆語のテキストデータで学習、そして数十億円の費用がかかったと見積もられている。もちろん計算リソースも非常に巨大であり、V100[7]クラスの

---

[4] BERT: Bidirectional Encoder Representations from Transformers., Google が 2019 年に Google 検索に導入した言語モデル用機械学習技術の名称

[5] 言語モデル:自然言語処理のための機械学習モデル。現在は確率に基づくモデルが主流

[6] パラメータ:機械学習で設定すべき変数

[7] A100、V100 は NVIDIA の GPU となる。それぞれ Ampere アーキテクチャと Volta アーキテクチャとなる。内部ユニットの形式、データバンド幅、設計ルールが異なる

1420GPU で約 1 年間という見積もりとなる。

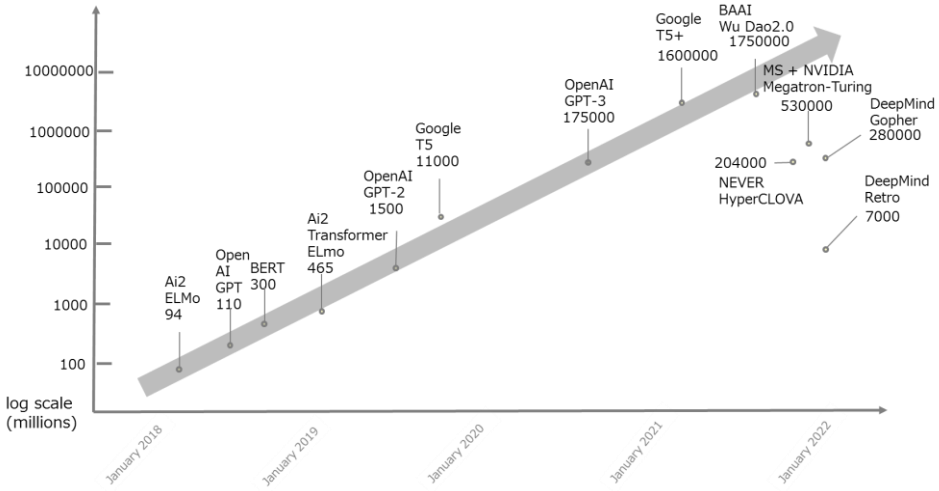


図 2 言語モデルのパラメータ数の変化[8]

（出典: https://ja.stateofaiguides.com/20200914-future-of-nlp/、
https://ourworldindata.org/grapher/ai-training-computation を基に IPA にて作成）

計算リソースに関しては現在も年々大きな進化をしており 2021 年に投入された GPU A100 では V100 の 20 倍程度の性能が期待でき、より現実的な時間、計算規模による学習が期待できることになる。2022 年に NVIDIA が発表した H100[9]では、A100 の 30 倍で自然言語モデルが構築可能としており、GPT-3 と同程度の巨大言語モデルを作成するとしても概算では一週間程度で可能になることになる。もちろん言語モデルそのものをさらに今の数十倍、数百倍に拡大することも容易となる。半導体チップの進化とともに計算性能は今後も大きく上がっていくことからタイミングが合えばコストパフォーマンスが高い計算リソースを手に入れることも可能であり設備計画が非常に重要になるともいえる。

また、実は 2021 年末に相次いで発表された自然言語モデルは必ずしもパラメータ数が増加していない。これは工夫によりある程度十分なパラメータ数がある場合には、より大きなパラメータ数のモデルと同等以上の性能を示すこともできるということを報告している。計算リソースの観点も含めて新しい流れとして捉えておく必要があるだろう。

GPT-3 がパラメータ数の増大により興味深い結果が得られており、演算ロジックを明示的に教えていない自然言語モデルが四則演算を獲得することに成功している[10]。図 3 に自然言語モデルのパラメータ数と獲得した計算能力を示している。ここで重要なのは学習には

[8] https://www.youtube.com/watch?v=G5lmya6eKtc, https://ja.stateofaiguides.com/20200914-future-of-nlp/

[9] H100 も NVIDIA の GPU となり、GTC2022 で発表された最新の Hopper アーキテクチャを採用している。https://www.nvidia.com/ja-jp/gtc/keynote/

[10] Language Models are Few-Shot Learners. Tom B. Brown et al. 2020, https://arxiv.org/abs/2005.14165

テキストデータのみを使用しているにもかかわらず自然言語モデルパラメータ数が 100 億を超えると計算を行う機能を持つようになったということである。もちろん膨大なテキストデータ内には、計算に関しての記述があり、また、1,2 問の簡単な例示が必要ではあるが、明示的に計算を学習していなくても自然言語モデルが計算そのものを学習したという結果は驚異的な結果といっていいだろう。膨大なパラメータを持つ自然言語モデルをテキストデータで学習させさえすれば、自然言語モデルは記述された手続きそのものを学習し、機能として取り込むという可能性を示したのである。当然プログラムの記述も学習する。

膨大な学習パラメータの効率的な学習、そして機能の発現、これらを可能にしたのが BERT で注目された技術である Transformer となる。

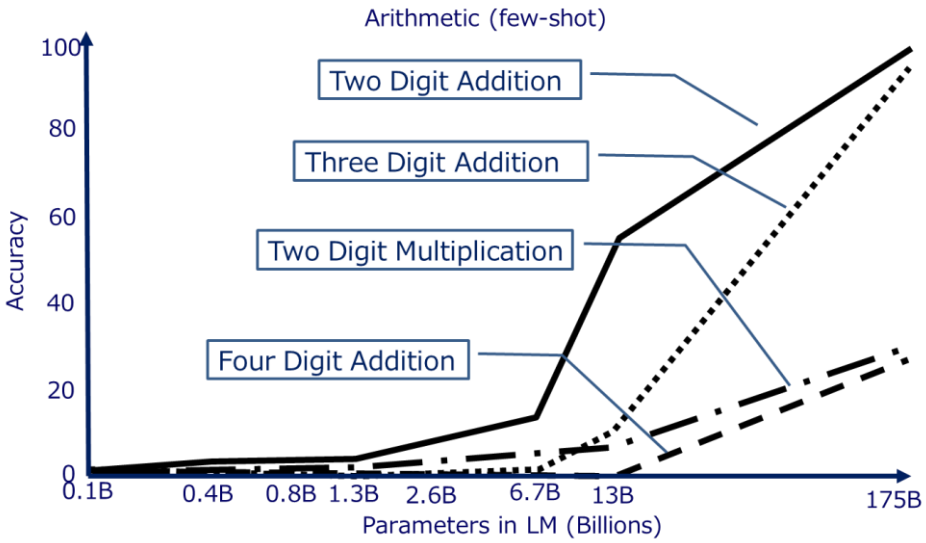


図 3 計算能力と自然言語モデルのパラメータ数[8]

（出典：Language Models are Few-Shot Learners　の内容を基に IPA にて作成）

1.2 効率的な総当たりを可能にした Transformer によるブレークスルー

自然言語処理の研究開発の歴史は長く、AI の本命技術の一つとして常に期待された領域の一つであった。深層学習が AI を席巻してからも様々な取り組みがなされたもののすぐには大きな進歩がなかった。これは自然言語の特性が邪魔をしていたともいえる。深層学習の発展のひとつに GPU[11]を画像描画以外に利用する、いわゆる GPGPU(General-purpose computing on graphics processing units)がある。GPGPU では GPU 上にある多数（数千個程度）の積和演算ユニットを並列かつ効率的に利用して計算を実行する。

ところが自然言語処理では、前後に行きつ戻りつしながら意味や指示語、代名詞等を扱う

[11] GPU: Graphics Processing Units 画像処理用演算装置。これを汎用計算に使うことを GPGPU（General-purpose computing on graphics processing units）と呼ぶ

必要がある。とりわけ文章理解ではこのような処理が必要となる。こうした処理では、特定部分を一度記憶しながら全体として時系列で処理する必要があるために並列計算と一般的には相性がよくない。自然言語では RNN[12]と呼ばれる時系列を扱うための構造を用いるのが普通なのはこのためであり、順序による依存関係を考慮する必要があるため、複数の依存関係間での調節(待ち)が発生し並列化しづらくなるのである。結果としてGPGPU が効率的に使用できないことになる。現在主流の GPGPU による計算はとにかく機械的に並列に計算できることがなによりも好ましいため、依存関係がある時系列計算は本来向いていないことになる。

そこで登場したのが Attention[13]機構である。文章中の単語に焦点をあてることでその単語と関係するあらゆる繋がりを文章中でみつけるという方法である。その際に繋がりのみに注目し文法的な性質を考慮しないことが重要となる。図 4 に概念的に Attention の概要を示す。「高い富士山と海が美しい」を例にすると「高い」に Attention が当てる。それ以外の文の要素である「富士山」「と」「海」「が」「美しい」とどのくらいの重みでつながればいいのかを計算する。図では実線、点線、線の太さで簡易的に表してある。これを文章中のすべての要素に対して繰り返す。メモリ上に配置できる文章全体に対して行うため、原理的にはかなり離れた要素間であっても重みを計算することが可能になる。

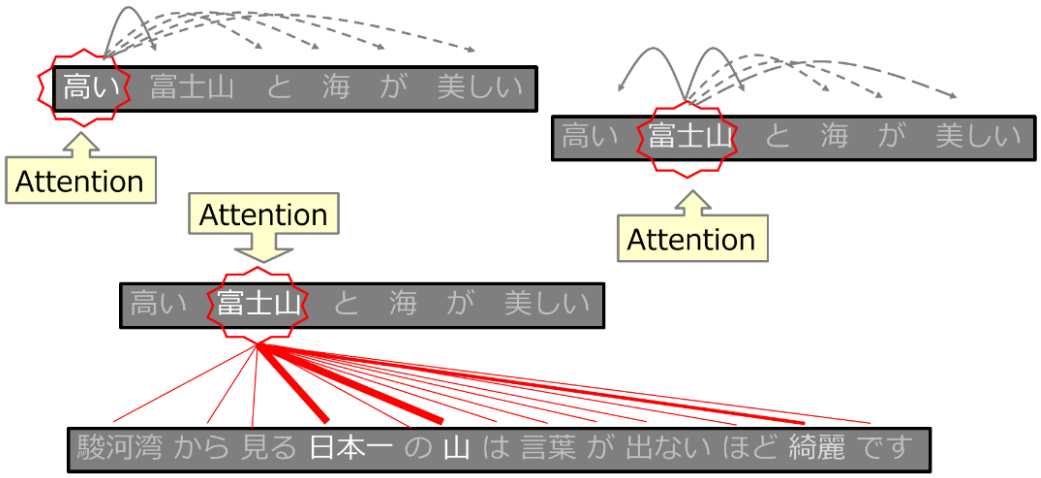


図 4 Attention 機構の概念図(Attending to all the words)

そして注目すべきはメモリ上の文章は、あたかも画像として扱えることから、その中の一部分の特徴量を他の場所の特徴量と比較するかのような並列計算を当てはめることができる。つまり本来は並列計算にそぐわない自然言語処理であるにもかかわらず並列処理で扱

[12] RNN: 再帰型ニューラルネットワーク（Recurrent Neural Network)。出力の一部が入力につながるネットワーク構造を持つ。Transformer 以前は RNN の一種である LSTM が自然言語処理では主流であった。

[13] Attention Is All You Need., A. Vaswani et al. 2017, https://arxiv.org/abs/1706.03762

えることになる。見方を変えればメモリ上の全要素の繋がりを計算することになるので、実はある意味で初期の全結合ネットワークに戻ったともいえる。

この並列処理が可能となった効果は絶大である。GPGPU の利用という点で効率的な並列計算が可能になったのはいうまでもないが、全結合によりテキストの相互参照による重み関係のみを学習すればよくなったため一般的な教師あり学習[14]が必ずしも必要ではなくなったことが挙げられる。自然言語の学習は後述するようにヒトが行うように簡単に評価ができないため、これまでに書かれたテキストを題材に学習を行う必要がある。この時に学習データの一部を教師データとし、その中であたかもテストを行い問題と正解を教える必要があり、このペアを作成するコストが非常に高いのが難点であった。ところが全結合が利用できれば、いわゆる穴埋め問題に帰結させられるため基本的にあらゆるテキストがほぼコストゼロで学習データに成り得るのである（これを自己教師あり学習と呼ぶ）。Transformer を利用した BERT では主に二つの簡単な学習方法をとる。いわゆる穴埋めを行う Masked Language Modeling(MLM) と文章の近さを評価する Next Sentence Prediction(NSP)である。

図 5 に MLM の概念を示す。Attention が相互参照に基づくことから、相互参照そのものを学習ともいえる。この方法にはさらに利点がある。つまり普通のテキストデータさえあれば、その一部をマスクすることで正解がある学習データとなるのである。このようにデータそのものを変形・加工することで正解ラベルを元データに与える方法を自己教師あり学習と呼び、明示的に教師データを与えるわけではないため代表的な教師無し学習として注目されている。実際に試験によくあるような一部をマスクすることによるマスク推定課題が自己教師あり学習の枠組みとして成立することが示され大きな進歩となったのである。

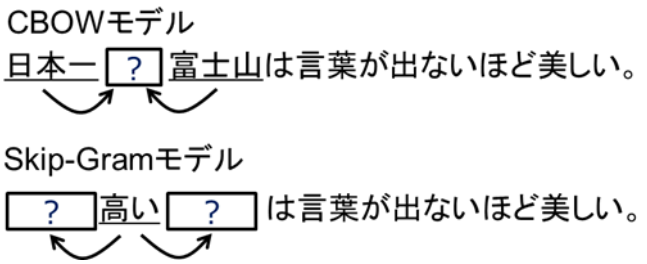


図 5　典型的な穴埋め学習方法(MLM)

さらに NSP では、例えば二つの文章を比べて次に来る文章として適切か否かを判定する学習を行う。相互参照情報からこの判定は簡単に行うことができる。BERT で実際に確かめ

[14] 機械学習には、大きく分けて教師あり学習と教師なし学習、強化学習があり、教師データは前者に必要である。詳細は DX 白書 2021 付録 第 1 部 AI 技術 3 学習を参照のこと

られた貴重な学習データに関する知見は、普通のテキストデータさえあれば、それが良質な学習データにも成りえることである。つまり大量の学習データは特にテキストデータに手を加えることなく自然発生することになる。自然言語モデルの学習の結果は文章の学習と同時に関係性をも学習することに成功し、さらに計算機能までを学習するという成果さえモデルにもたらした。

Sentence Pair	Label
母はパスタをつくった。 それはとてもおいしかった。	isNext
クジラが泳いでいた。 私は数学の問題を解いた。	notNext

図 6 NSP の例

このような Attention を中心とした基本的な機械学習は Transformer と呼ばれている。Transformer では、普通のテキストを学習データに、そして計算資源を効率よく並列使用することで膨大なパラメータを処理することを可能にしたのである。

1.3 Transformer による自然言語処理の特徴

Transformer を理解するための追加のキーワードをいくつか紹介する。これらのキーワードは現在の自然言語処理技術を特徴づけている以上のものにもなっている[15]。

パラメータスケール

パラメータが多ければ、例えば学習時に参照範囲を広げることによるモデルへの反映により有用な結果を期待できる。つまりパラメータスケールが大きくなることにより、現在のモデルの限界を押し上げることになり、同時に性能も上昇することが期待できる。実際、Transformer においてパラメータスケールの拡大による性能保証は確かめられており[16]、パラメータスケールの拡大はそのまま性能上昇につながる。基本的にはより大きなパラメータ数を持つことがより良い自然言語モデルとなるため、現在さらに大きなパラメータスケールが試されている。一方で、大きくなったモデルを圧縮して効率化する手法もまた研究開発されている。すなわち、モデルの限界は更新され続け、同時に効率化する術が研究開発されているのが現状となる。

---

[15] Transformer は自然言語処理のための機械学習手法であったが画像応用等でも有用性が示されている。この分野は Vision Transformer(ViT)と呼ばれている

[16]Scaling Laws for Neural Language Models., J. Kaplan et al. 2020, https://arxiv.org/abs/2001.08361

自然言語モデル

自然言語モデルは一般的には自然言語処理を行う上で機械学習の中核部分となり自然言語処理の基盤を担う。自然言語処理要素の一つは、文章のどこにどの単語が出現するのが自然なのかを扱うことであるが、Transformer では文章中の特定の位置での単語の出現確率を前後の単語との相互参照を含めて決めることにより確率ベースで扱うことでモデルは構成されている。

膨大な数のパラメータを有し、膨大なテキストデータにより学習された大規模な事前学習[17]後の自然言語モデルは学習データから驚くほど多くの知識を獲得している。この知識は主に単語間の関係性を学習することで獲得されているが、これらのモデルは知識獲得に成功したのみならず、単語間の関係性だけでは得られない記述内容の論理的な繋がりさえ学習している。そのためテキストの要約や会話応答といった文章生成タスクにおいて正確かつ構成上も問題ない水準に押し上げるだけでなく、単独の自然言語モデルから言語タスク以外の機能、例えば計算や仕様からの自動プログラミング等も可能にする。

Zero(Few)-Shot 学習

前述したように事前学習モデルには、学習データ内のテキストに記述されたすべての参照関係が含まれるため、例示が無いもしくは一例二例の提示のみでドメインチューニングが可能であり、幅広い範囲での任意のタスク出力を可能にした。

Few-shot 学習はタスク（推論時）に少数（10 から 100）の例を示す場合を、One-shot 学習では一つの例のみを示す場合となる。そして Zero-shot 学習では例示無しとなる。これらは事前学習モデルの性能に大きく依存することになるが現在では、実際タスク例がないか、一つのタスク例を示すことで、従来必要であったドメインごとの学習は必要なく、タスク達成性能のベンチマークでは、ほぼ最高のパフォーマンスを上げることに成功している。

特に GPT-3 ではプロンプト学習[18]を使用しており、その性能の高さからその後の自然言語モデルのチューニング法としてほぼデファクトとなっている。プロンプト学習では対話 UI 画面上での簡単なやり取り、その多くは簡単なセンテンスの組を shot 数に合わせて例示することで調整する。一般的に追加学習と呼ばれるものであり、この場合は自然言語モデル内にある膨大なテキストペアの関係性に対してドメインとしてはどれを優先させるべきかを追加で調整していると理解することができる。

---

[17] 汎用的な言語能力の獲得を目的として，大規模コーパスで学習をすることを指す。目的タスク用の学習に対してあらかじめ行っておくことになる。これにより目的タスクのデータが少量であっても高性能となることが期待される。

[18] 対話形式による短い文章のやり取りを利用して学習器に教示する方法

言語ベンチマーク

自然言語モデルを評価することは、自然言語モデルを使用する上で非常に重要な点であることは言うまでもない。例えば人であれば簡単に書かれた文章の自然さは分かるが、その自然さを評価のための数値にするのは元々簡単ではない。そもそも評価が数値化できれば、それをアルゴリズムに落とし込むことにより自然言語処理が可能になるわけであり、それが難しいため今日に至っている。こうした背景がある中で自然言語処理用にいくつかの評価指標としてベンチマーク[19]が考えられてきた。基本的には人が書いた教師データを基にすることで、簡単に言えばテストを行い客観的な評価を可能にする方法となる。例えば国語の試験のように質問応答や文書生成、多段階の論理的な繋がり自然言語推論、参照による解決、語義曖昧性解消や翻訳といった視点となる。

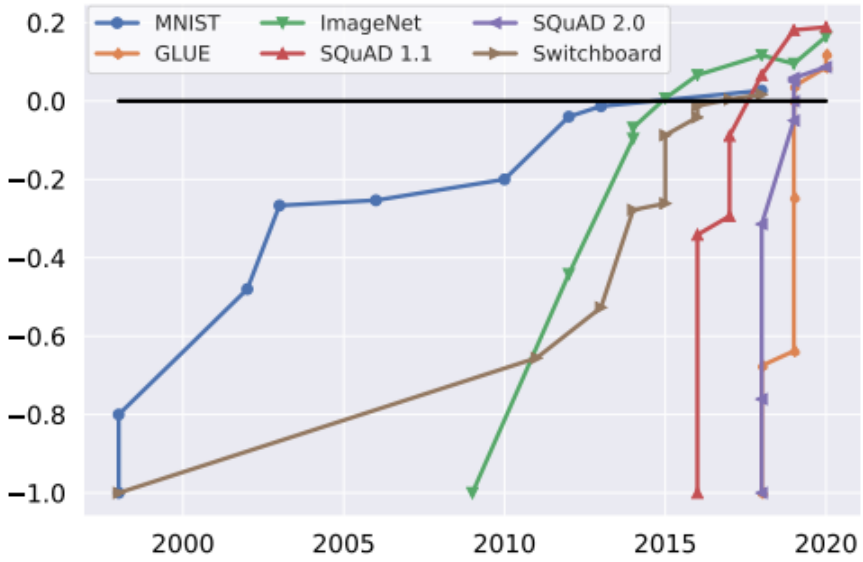


図 7 言語ベンチマークの数値の変遷（抜粋）

出典: Dynabench: Rethinking Benchmarking in NLP., D. Kiela etal. 2021, https://aclanthology.org/2021.naacl-main.324.pdf

図 7 にいくつかの自然言語処理のベンチマークのスコア（比較のため正規化済み）を示す。2010 年代深層学習がメジャーになってからのスコアの伸びは激しく、人のスコアを超えていることが分かる。そのため従来はベンチマークの性能向上を狙って自然言語処理システムを作り上げ、数値を単に追い求める研究開発もあり得たが、現在はより汎用的な自然言語処理システムを構築して、その上で種々のベンチマークに

[19] https://paperswithcode.com/area/natural-language-processing

より検証する形に変化している。つまりテストのための学習から一般的な広範囲の学習を行いテストに挑むという方向になりつつある。これは技術的には個別のタスク、ドメインに特化していない現在の汎用的な自然言語処理システムでも特化システムの性能に匹敵していることを示している。

現在は、ここ数年の自然言語処理技術の発展によって、これまでのベンチマークが役に立たなくなりつつあり、新しいベンチマークが提案されている[20]。加えて現在では言語ベンチマークは単に優劣を見定められるだけでなく、動作の詳細な診断を行うことにも繋がりつつある。これは万能なベンチマークの作成が困難なために特定のタスク、ドメインにターゲットを絞ったものが主流であることをうまく利用した方法となる。

また、すでにベンチマーク上ではヒトのスコアを超えているものもいくつか確認できるため、すぐに実用的なアプリケーションが登場すると期待できる一方、社会実装上で問題となることも少なくはない。そこで従来のベンチマークの欠点を補い、より社会実装に即した新たなベンチマークを研究開発することにより自然言語処理系をさらに洗練させていくことを目指す試みもある。同時に新しいベンチマークには別の側面が期待されている。ベンチマークの中で自然言語モデルのリスクについて本格的に考えなければならない状況なのである。

現在のベンチマークツールは、いくつかの重要なリスクを評価するには不十分だとする人も多い。例えば自然言語モデルを使用したアプリケーションにおいてある入力に対して出力が誤っている場合を考える。その情報を人々が真実であると信頼することが本当にないのか。例えば、通常ほぼ100%信頼できる情報の提示がなされている場合には、仮に例外として出力された情報が間違っていても、普通はこれまでの信頼度が横滑りするため、それが信頼バイアスとなり人々は信用してしまう。この場合、必ずしも悪意があるわけではないにも関わらず問題が内在することになる。つまり自然言語モデルの使用は基本的には大きなリスクを内包することになる。

現状、解決方法としては、自然言語モデルをベンチマークツールのみで評価せず、実際に人が試し、確認していくという人間を含めた精査以外に方法がないため、リスクの軽減に関してより多くの考察を行うことが重要である。先ごろ自然言語モデルが有害な社会的ステレオタイプを再現することで大きな問題として取り上げられたが[21]、実はこういった問題に関しては初期段階の研究がなされているにすぎないとも言える。リスクをより評価するベンチマークの改良が引き続き行われていくことに期待す

[20] https://ruder.io/nlp-benchmarking/、https://github.com/kwchurch/Benchmarking_past_present_future

[21] 例えば以下。米マイクロソフトがAIの発言に関して陳謝した例(https://jp.reuters.com/article/tay-idJPKCN0WU056)、米アマゾンのAI採用ツールにおけるジェンダー差別に関する例(https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G)

る向きも多い。

多言語対応

言語の多様性は、自然言語モデルにおいても非常に重要な視点となる。言語の構成はその地域のモノ、コトの説明性と強く関係していると考えるのが自然である。事前学習モデルでは、単に一言語での相互参照にとどまらず、多言語にまたがる汎用的な関係性を学習することが期待されている。そのため学習された多言語にわたる汎用性は本質的に自然言語モデルに内包されるべき分散表現、言い換えれば表現ベクトルが獲得すべき性質となる。分散表現が汎用性を獲得しているとすれば、つまり、それはより基本的な言語レベルでのモノとコトに対応する様々な関係を獲得していることになるため言語が国ごとに異なっている場合でも言語モデル上は特に問題が生じることはないと考えてもいいだろう。そのためできるだけ広範囲の言語データを使用し自然言語モデルを構築することにより、双方向の多言語対応機械翻訳を目的とするだけではなく、自然言語モデルそのものをより高品質にすることに繋がる。

本来翻訳には翻訳元と翻訳先の双方向のテキストデータが必要であるが、自然言語モデルが汎用的な分散表現を獲得することで、必ずしも双方向の十分な学習用のテキストデータが無い場合でも機械翻訳が可能となることが期待される。実際、Google は 2022 年 5 月の開発会議において十分な学習データがない場合でも言語モデルが新しい言語の翻訳を学習する単一言語アプローチが可能であり、新規に 24 の言語を翻訳サービスに追加したことを発表している[22]。この 24 の言語はメジャーではないとされているものの使用者の合計は 3 億人の人口にのぼるため、今後この次世代翻訳による広がりは大きな期待がもたれる。

世界には 7,000 以上の言語[23]があることからも（厳密には言語を区別するのはかなり難しい）現在、翻訳が可能な言語はそのごく一部のみとなる。世界の言語をマップにしたものが図 8 になる。各ドットは 1 つの言語を表し、色は各言語の最上位の言語族を示す。これまでにモデル化された言語はまだ少数ではあるが、それでも分散表現の改良に影響がある。今後言語数が増えるにつれてさらなる普遍性を獲得することが期待できるのではないだろうか。さらにより多くの言語への対応が、未知の言語に対応するためにも有用になることがわかっている。世界の言語すべてへの機械翻訳の対応は、それ自体が重要な技術的挑戦の一つであるのと同時に、現在の自然言語モデルの位置づけからは言語そのものを理解するための知的な挑戦にもなる。

---

[22] Google I/O 基調講演, https://io.google/2022/program/8e80903f-955f-4a5b-9118-b0ce4acdb0e6/intl/ja/

[23] https://www.ethnologue.com/

図 8　世界の言語

（出典: Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources., S. Nordhoff el al. 2011, First International Workshop on Linked Science 2011-In conjunction with ISWC 2011.）

以上のようなキーワードが Transformer による自然言語処理を理解する助けになるだろう。すべての特徴をまとめると実は習うより慣れるという言葉が当てはまるのではないだろうか。現在の自然言語処理は結果として言語の意味理解を行っているように見えるため現在では自然言語理解というフレーズも使われ始めている。これは対象のテキストの内容を理解した上で、その意味を含めた処理結果が得られるように見えるためである。システムが真に理解しているかどうかは、議論の的になっているところではあるが非常に高度な処理が可能になったのは事実である。

2 自然言語処理技術アプリケーション

2020 年 7 月に GPT-3 は自然言語処理の成果として OpenAI による 3 番目の大規模自然言語モデルとして紹介された。このインパクトは時とともに大きくなり[24]何年後かに振り返ったときには人工知能そのものの転換点だと認識される可能性が高いといわれている。GPT-3 は何兆もの単語で訓練された 1750 億のパラメータを有した Transformer による自然言語モデルとなる。

GPT-3 の自然言語処理の性能は非常に高く、まるで人間のようにテキストを処理、そして生成する。例えばテキストによる入出力を行うプロンプト画面でユーザーがテキストを

[24] 例えば https://maraoz.com/2020/07/18/openai-gpt3/

入力すれば、GPT-3 は、用途に応じて、一貫性のあるトピックを電子メール、ツイート、雑学クイズなどに合わせた出力を返す。つまり、電子メールの作成、顧客とのやり取り、ソーシャルメディア、さらにはニュース記事でさえ、控えめに言えば部分的な文章を、基本的には全文を自動で作成できるような環境が突如として用意されたのである。

OpenAI による GPT-3 の利用方針は商業応用を支援しており、それを受けて即座にビジネス応用が始まっている。このような自然言語モデルを利用さえすれば、新しいスタートアップとしてアプリケーションが立ち上るため GPT-3 および同様のモデルは、世界中の自然言語処理の活用を検討している人々の手に巨大な AI の力を手軽にもたらすことになる[25]。実際、Amazon で著者として GPT-3 を指定して検索する[26]と共著ではあるが GPT-3 が著者である書籍が確認できる。GPT-3 の文章をチューニングするためには、まだ人の手が必要であるようだが、売られているという事実は非常に重要な転換点を示している。

2.1 GPT-3 のビジネス応用例

例えば、OthersideAI[27]ではメールでは、本文のキーもしくは文頭を書きさえすれば、そこから完全なメールの文面を生成する。こういったキーワードによる文章生成には他にも広告コピーやキャンペーンコンテンツの Broca[28]や Snazzy[29]がある。両社とも試すことが可能で文章自動生成のこのイノベーションをすぐにでも体験できる。以下、これらの例を解説する。

E-mail はインターネットの初期から現在まで重要なコミュニケーションツールとして重要な位置を占めているが、同時に近年は特に生産性低下の代表的な例、つまり原因として扱われている。OthersideAI は E-mail による生産性低下を吹き飛ばし、むしろ生産性を上げるサービスのために GPT-3 を活用し、この問題を克服した。基本的には質問と回答の組み合わせを考える。これを都度生成するのはもちろんであるが、同じような内容であっても細かな調整をシナリオごとに行うことを可能とし、シナリオに対して少しずつ変化し必要な情報を盛りこむことを可能としている。そのため販売、サポート、営業等、従来であれば別々に組み直す必要がある複数のシナリオ・状況であっても共通のエンジンとして何の問題もなく対応できることになる。図 9 では、キーワード入力するだけで完全な文面が生成されるメールの例を示している。必要なのは必要十分な単語の記述だけであり、あとは自動でメール文面が作成されるのが分かる。

この文面自動生成の機能は、別の見方をすることでタイピング補助としての使用法が浮かび上がった。元々すべての単語、文節を完全に自身の言葉で表現することは、メールのや

---

[25] https://gpt3demo.com/

[26] https://www.amazon.co.jp/advanced-search/books　において著者名に GPT-3 と入力して確認できる

[27] https://www.othersideai.com/

[28] https://www.usebroca.com/

[29] https://snazzy.ai/

り取りではまれであるが、特にビジネスに限れば、かなりの割合を定型文章が占めるだろう。いずれにせよ大部分は自動生成のみの文面でも大きな問題はないのである。このことは例えば視線入力で文章を作成する場合には、最初の単語入力から文章の生成が行われ、後は確認と若干の修正のみで文章が完成することになり、その恩恵は計り知れない。

興味深いのはこのような入力補助としての文章生成はサービス開始時点から想定されていたわけでは無く、自動生成例の確認、解析によりその可能性に気付き、応用領域として取り上げたとのことである。実際これまでの入力補助ツールのなかでも最も使いやすいものの一つであるといっても過言ではないだろう。

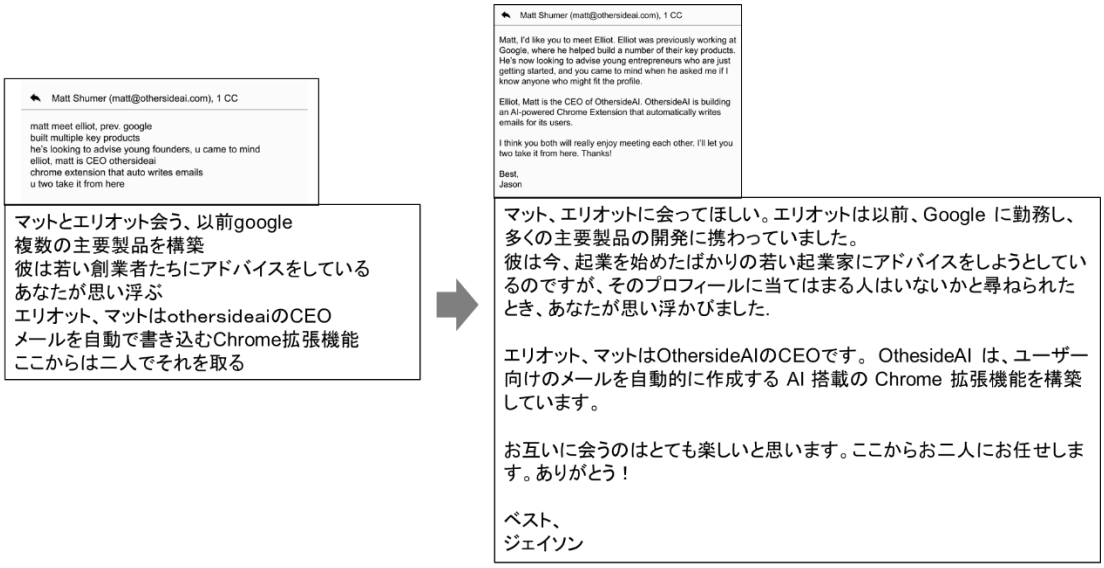


図 9　OthersideAI による e-mail 自動生成例 [16]
(出典: OthersideAI のデモ画面(https://techcrunch.com/wp-content/uploads/2020/11/otherside-Demo-3.gif)より IPA にて作成)

広告サービスを展開している Broca や Snazzy では、キーワード入力による文章生成を最大限生かしている。たった一言の入力、もしくは限りなく短く、しかしインパクトがある広告コピーから文脈を考慮したキャンペーンコンテンツを自動で生成することができるのはこの上なく便利なサービスとして映る。

この AI による自動生成コンテンツを武器にして Broca はコンテンツマーケティングの拡大を図ろうとしている。従来数時間、時には数日かかっていたコンテンツがわずか数分で、しかも高品質で生成可能となることが売りとなる。そして本当に品質が高いのだろうか、という疑問の声にこたえるために、Broca のサイトのすべてのコンテンツを Broca のサービスを使用して作成することでその品質の高さを保証している。もちろんこれは広告からソーシャルメディアまで提供する種類に応じたコンテンツの品質への自信の表れなのである。

例えば新製品のキャンペーンコンテンツを作成する場合、図 10 に示すようにたった 3 ス

テップのみで実現できる。実際には、最初のステップに入力する 2、3 のセンテンスだけがユーザーの入力となり、たったそれだけであとは自動でコンテンツが生成され、一貫したコンテキストを保持したプロモーションコンテンツを種類を問わず手に入れることができる。

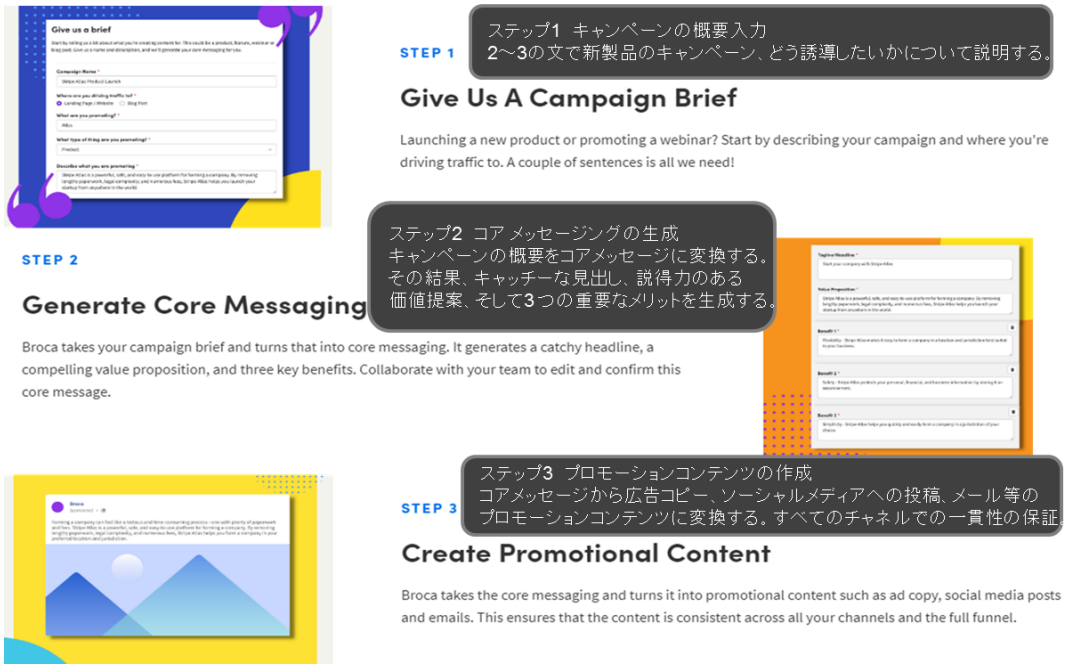


図 10　Broca でのコンテンツ自動生成の 3step
（出典:Broca の説明（https://www.usebroca.com/campaigns）より IPA にて作成）

そもそもの基本的な要素として BERT そのものの利用についても多くの例がある。FORETHOUGHT[30]は、Agatha と呼ばれる質問応答検索 AI エージェントを提供している。Agatha は、ナレッジベースの記事またはヘルプデスクテンプレートを使用して、電子メールまたは Web ウィジェットを介して一般的な質問に回答する。これにより、顧客はより迅速に回答を得ることができることになる。Agatha は、時間の経過とともに向上し続けるように機械学習と自然言語処理の継続的な学習を特徴の一つとしている。すでに 100 言語に対応しておりグローバルにソリューションを展開している。

Gong[31]は顧客との会話を音声解析する次世代型 CRM を提供している。結果として何千もの営業チームが、Gong の自然言語処理機能を利用して取引を成立させ、販売サイクルを短縮している。顧客と共有した情報の 99%は CRM に到達することはなく、1%は厳しくフィルタリングされてしまうため営業の役に立つことがない。そこで Gong では、顧客との情報に関して自然言語処理を駆使することで、実際の取引に必要な情報に結び付く高次の洞

[30] https://www.forethought.ai/
[31] https://www.gong.io/

察につながる要素を取り出す。顧客の電子メール、電話、ビデオコールを転記し、次に機械学習を使用して、顧客が製品の更新を提案する準備ができているか、取引が失われるリスクがあるか、洞察につながる要素のすべてを分析する。

Moveworks[32]は、従業員の IT 利用時の諸問題を自律的に解決できるボットを提供している。そして自然言語処理機能により、会話や曖昧な質問を理解できるようにした。2018 年の夏に 1人の顧客が人間のサポートなしで IT 利用時の諸問題の 20%を自律的に解決したことから始まり、現在、平均でボットの解決率は最大40%であり、場合によっては 65%になる。つまりＩＴ利用時の諸問題のうち 65%は自動で解決できることとなった。例を以下に示す。

例　*「質問」→「自動応答」*

「PDF ファイルを編集したい」　→「Acrobat Pro のライセンスを発行します」

「リモートワークなので新しいキーボードがほしい」　→「以下の何番のキーボードがいいですか？」

「パスワードを忘れた」　→　「新規パスワード生成はこちらです」

明らかにこの機能の適用先は、なにも IT 分野に閉じることはなく、現在はさらにこの AI 機能を人事、施設、財務分野にまで拡大している。

コンタクトセンターの業務を解析し、効率化を行う会社として OBSERVE.AI[33]がある。OBSERVE.AI では、コンタクトセンターでは、通話の 1%から 2%、またはそれ以下しか相手に伝わらないという経験から音声分析をこの未開拓の領域に適用した。音声からテキストへの処理と自然言語処理を利用することで会話内での関心ポイントを見つけだし、それを顧客体験の向上につなげることに成功したのである。顧客の気分をよりよく理解するためにコールセンターの対話に対して教師あり学習を適用することで、より良い顧客体験につなげている。

2.2 GPT-3 の応用技術

言葉による柔軟な画像生成 DALL･E と柔軟なキャプション付加 CLIP

OpenAI はテキストから画像を生成する DALL･E[34]を発表した。GPT-3 の 120 億パラメータバージョンをテキストと画像のペアで学習させている。入力されたテキストの内容を読み取り、学習データに存在しない画像であってもモデル内部での組み合わせにより生成

[32] https://www.moveworks.com/

[33] https://www.observe.ai/

[34] DALL･E: OpenAI が開発したテキストから画像を生成する AI。Zero-Shot Text-to-Image Generation., A. Ramesh et al. 2021, https://arxiv.org/abs/2102.12092

することができる。発表内容[35]によれば「アボカドのようなイス」「ベッド脇のスツールの上にある花瓶」という、ヒトが読めば一見完全な記述であるが、実際に画を描こうとすると細かな条件が与えられていないことに気付くような内容から問題なく画像(正確には画像群)を生成することができる。

Transformer の学習では、一見して無関係と思われる要素であっても、必要な関係性をあらゆる条件で重みづけを行い学習に反映される。そのため実際に画像を生成する時に必要なあらゆる条件、例えば光の方向、影、質感、ベッド脇のどの方向といった描画に必要な細かい条件は、事前学習から得られる当たり前の条件を学習された常識から取り出していると考えることもできる。そのため、不足している条件は必要十分条件として推測され、問題なく一つの絵として成り立つように描画することが可能となる。この条件には種々の変形としての擬人化、絵文字化も入っており、簡素化や詳細表示などテキストの解釈から可能な範囲でかなり柔軟な変形が可能なようである。画像生成時に補完した情報を含めて、補完したテキスト表現と画像をいわばつなげ直すため、これを高度なキャプションの自動生成機能として別の利用方法も示し例示している。

図 11　話題になったアボカドの形をしたアームチェアの DALL･E による自動生成
(出典：”an armchair in the shape of an avocado” を入力したデモより
https://openai.com/blog/dall-e/)

同時期に発表された CLIP[36]（対照言語-画像事前学習）は画像エンコーダー[37]とテキスト

[35] https://openai.com/blog/dall-e/

[36] CLIP: Contrastive Language-Image Pre-training の略 https://openai.com/blog/clip/ , Learning Transferable Visual Models From Natural Language Supervision., A. Radford et al. 2021, https://cdn.openai.com/papers/Learning_Transferable_Visual_Models_From_Natural_Language_Supervision.pdf

[37] 情報を圧縮するのがエンコーダー、復元するのがデコーダー。

エンコーダーを組み合わせた新規の事前学習により画像とテキストスニペット(キャプション)ペアを高精度に予測する。画像とそのテキスト記述に関しては、依然大きな問題が残っており、目指すのはいわゆる Zero-shot 分類であり学習に使用していない画像データであっても高精度にテキストキャプションを予測することである。

キーとしての考え方の一つは、単純だが効果が大きい方法、つまりこれまで多くの例で示されたようにより大きなデータ数による学習である。ImageNet が 1400 万画像での学習であるのに対し、CLIP は 4 億ペアともなる大規模なデータセットを使用している。また、CLIP では、Zero-Shot 機能を実現するために Internet 上の様々な画像とそのキャプションのペアを使用して学習されている。実際の学習では画像とそれに対応するキャプションを直接最適化するわけではなく、画像と最も関連性の高いテキストラベルを予測させる。

CLIP ではテキストキャプションをランダムにサンプリングされた 32,768 個から選択する。これを Transformer により効率的に行っている。わかりやすい例を示すと、例えば候補として「走り回る犬」、「犬の散歩」、「ソファーで寝る猫」、「じゃれつく犬」、「小屋で寝る犬」がある場合に写真が「猫」であれば答えとして「ソファーで寝る猫」の確率が最大となりこれが選択される。この繰り返しにより画像とテキストキャプションが紐づけされていく。画像の分散表現とテキストキャプションの分散表現が学習されるのである。単語レベル、画像の断片レベルでの結び付けから言語と画像を強固に結び付けている。CLIP を使用することで人によるラベルつけを完全に自動化することが可能になる。画像に対してのラベル付けは非常に高価となるためその応用が期待される。

Power Apps[38]

Microsoft は、OepnAI とパートナーシップを結んでおり GPT-3 の発表後は OpenAI と GPT-3 の使用に関して独占契約を結んだ[39]。OpenAI の開発プラットフォームは同社のクラウド環境 Microsoft Azure に移行している。開発者会議 Build2021 の中で、この強力な自然言語モデルである GPT-3 を Power Apps に利用することを報告し、一連のデモを行った。発表があった製品である Power Apps は、いわゆる日曜プログラマから業務でプログラミングを行っているプロの開発者までローコードアプリ開発プラットフォームとして人気がある。

その Power Apps に GPT-3 が組み込まれることによって、コードや数式を意識することなく、自然言語による記述からアプリケーションがビルド可能となったのである。英語で説明を行えば自動で Power Fx に変換する。(図 12) Power Fx[40]には GPT-3 の自然言語処理機

---

[38] Microsoft が提供するビジネスニーズに合ったカスタムアプリを構築するための環境、短時間でアプリケーションを開発できる。https://powerapps.microsoft.com/ja-jp/

[39] https://www.infoq.com/jp/news/2021/01/microsoft-license-gpt-3/

[40] Power Apps 等で使用されるローコード言語。Excel でのプログラミングで使用するような汎用、厳密

能が統合されているため、さらにサンプルを示すことでアプリケーションをビルドすることも可能になった。

GPT-3 による新たな自然言語処理機能は今後、Azure などの主力製品にも取り込まれていくことになる[41]。平易な言葉で数式による記述と全く同じ結果が得られるため、だれもがアプリケーションを、しかも AI を含んだアプリケーションを気軽に作れることになる。

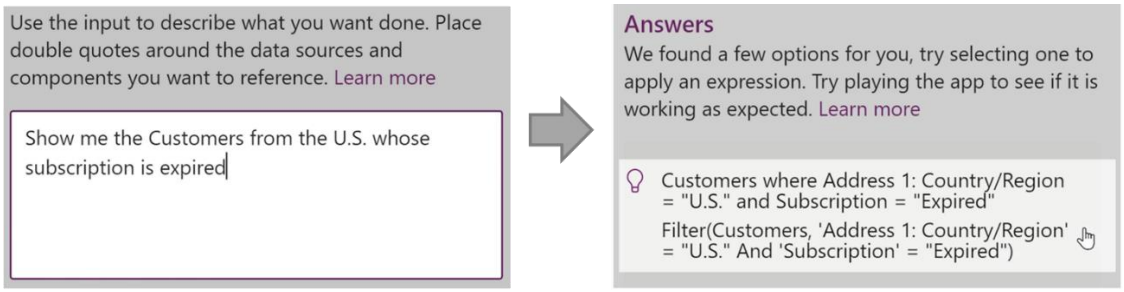


図 12　Microsoft による Power Apps での GPT-3 応用例
（出典：Microsoft によるデモから https://news.microsoft.com/ja-jp/2021/05/26/210526-microsoft-introduces-its-first-product-features-powered-by-gpt-3/）

以上のように Transformer とその中核技術の一つである Attention による自然言語処理技術は、現在の人工知能のソフトウエア面での基盤技術である深層学習と GPU をハードウエアとして効率的に使用できる GPGPU による大規模なデータによる超多数パラメータ学習を行うというものであり、衝撃的な発表であった GPT-3 を例に取り上げた。その後、競って技術の発表があり、さらに技術が進んでおり、後ほど、その一部については改めて取り上げる。

3 国内の動向

3.1 日本語の対応

GPT-3 の規模の超大型自然言語モデルを国内の一社が作り上げるのは難しいだろう。とすれば国内では自然言語処理技術はどのように扱うべきなのかを考える必要がある。例えば複数の組織が協力して超大規模自然言語モデルを作るのは一つの方法であるといえる。実際 GPT-3 規模の自然言語モデルを作り上げる場合、計算リソースを見積もれば、理研 RAIDEN を一年間、一日 24 時間そのためだけに使うことが必要となる。（先にも述べたが計算リソースの性能は、例えばこの 4 年では自然言語に限れば 600 倍となっており、GPU

な型指定、宣言型、関数型となる。

[41] https://news.microsoft.com/ja-jp/2021/05/26/210526-microsoft-introduces-its-first-product-features-powered-by-gpt-3/

関連各社のロードマップでは今後もこのペースは変わらない。そのためハードの更新が可能であるとすれば単位時間あたりの計算リソースはじきに問題なくなるという見方もできる。) ただし、計算リソース以外にも学習データをどのように集めるのかという課題もある。これは我々が日本語という日本という国に根ざした言語を使用しているため生じる課題であり、テキストさえあれば学習データとして使用できる状況でも、実用的な汎用自然言語モデルを作り上げるためには満遍なく偏りがないテキストデータがかなりの量が必要なため、特に重要な課題となる。

そこで別の視点としてはアルゴリズムの改良が挙げられる。例えば PET/iPET[42]では 2 億 2300 万のパラメータでも GPT-3 と同等の性能が確認されており、さらにベンチマークである SuperGLUE[43]では GPT-3 を上回る性能を示している。事前学習された ALBERT[44]を利用し Few-Shot の例から追加の学習データを生成する半教師あり学習手法となる。PET は、最初に入力例を cloze-style のフレーズに変換することによって機能する。これらは、自然言語モデルのアンサンブルを微調整するために使用され、次に、ラベルのない大きなデータセットに注釈を付けてソフトラベルの付いたデータセットを生成するために使用される。最終的なモデルがソフトラベルされたデータで微調整される。このソフトラベル化された学習データを使用することで本学習する。この場合にはリソースを含めた学習コストが 1/500 程度になる可能性もあり、計算機環境が同じであれば、これまでは 500 日かかっていたが、計算上 1 日以下で準備が整うことになる。学習時間を短くするのか、計算機環境の規模を小さくするのか、新たな戦略的方法論を考えることができる。

3.2 国内リソース

参考までに国内で利用できる代表的な自然言語モデル(BERT)を示す。そのほとんどが自由使用のオープンソースライセンス形態[45]をとっており自社で使用することが可能である。事前学習済みの汎用モデルが提供されているので、自社で使用する場合には Zero(Few)-Shot 学習が必要となるが、事前学習を一からたちあげる必要がないため手軽にタスクを試すことができる。現在、超大規模な GPT-3 でなくても、比較的小回りの利く BERT で事業応用を行っている例は多い。しかも PET/iPET の例からも、要素技術としてまだまだ工夫

---

[42] PET/iPET: 自然言語処理 (NLP) モデルの深層学習での学習手法。Pattern-Exploiting Training、Iterative Pattern-Enhanced Training の略。より小さなモデルで同様のパフォーマンスを達成するための学習方法となる。

[43] 自然言語処理ベンチマークの一つ。質問応答や自然言語推論、共参照解決、語義曖昧性解消など、幅広いタスクからなるテストから成り立つ。

[44] ALBERT: A Lite BERT for Self-supervised Learning of Language Representations., Z. Lan et al. 2019, https://arxiv.org/abs/1909.11942

[45] 例えば Apache ライセンス[Apache Software License]: オープンソースソフトウェアを開発・配布する際によく用いられる、利用条件などを定めた利用許諾契約書（ライセンス）の一つ。

次第であることも事実である。BERT 事前学習モデルを使用することで、各々が現在、解くべき課題とタスクに実際に適用できることは重要だろう。早稲田大学河原研究室が公開した BERT 系日本語モデル(Large)は国内で最も充実したモデルのひとつ[46]である。

表 1 日本語 BERT モデル(出典：各 URL)

製作者	フレームワーク	データソース	ライセンス	URL
Google	TensorFlow 2	日本語 Wikipedia 他	Apache 2.0	https://github.com/google-research/bert/blob/master/multilingual.md
京都大学 黒橋・褚・村脇研究室 早稲田　河原研究室	TensorFlow 2 PyTorch Transformers	日本語 Wikipedia、CC-100	Apache 2.0	https://nlp.ist.i.kyoto-u.ac.jp/index.php?ku_bert_japanese https://huggingface.co/nlp-waseda/roberta-base-japanese https://huggingface.co/nlp-waseda/roberta-large-japanese (より新しい RoBERTa-base)
東北大学 乾・鈴木研究室	TensorFlow 2 PyTorch Transformers	日本語 Wikipedia	Apache 2.0	https://github.com/cl-tohoku/bert-japanese
NICT	TensorFlow 2 PyTorch Transformers	日本語 Wikipedia	CC by 4.0	https://alaginrc.nict.go.jp/nict-bert/index.html

GPT モデル、CLIP

rinna 社 (GPT)	PyTorch Transformers	日本語 C4 、CC-100 、Wikipedia	MIT	https://huggingface.co/rinna/japanese-gpt-1b
rinna 社 (CLIP)	PyTorch Transformers	CC-100 、Wikipedia 、CC12M	Apache 2.0	https://huggingface.co/rinna/japanese-clip-vit-b-16
rinna 社 (CLOOB)	PyTorch Transformers	CC-100 、Wikipedia 、CC12M	Apache 2.0	https://huggingface.co/rinna/japanese-cloob-vit-b-16

[46] https://twitter.com/daisukekawahar1/status/1524288370767134721?cxt=HHwWgoCsrc-MrqcqAAAA

2022 年 1 月に日本語 GPT 言語モデル[47]について発表があった。Microsoft から独立した rinna 社が 13 億パラメータの日本語に特化した自然言語モデルを公開したのである。MIT ライセンスで提供されており商用利用も可能となる。BERT からさらに先に進んだ形で日本語 GPT モデルが公開されたことで、実利用の目的に合わせてプロンプト学習やファインチューニング等の Few-Shot 学習のみで比較的短期に簡単に利用できる場合もあり注目されている。2022 年 5 月 12 日に同社から CLIP およびその改良版である CLOOB の発表があった[48]。どちらも商用利用が可能なライセンスでの提供であり合わせて紹介する。

3.3 国内の話題

人工知能学会の対話システムシンポジウムコンペティションにおいて LINE の対話システムが二冠を達成し、評価の中で人間との区別がつかないと話題にのぼった。LINE は韓国 NEVER と開発を進めた日本語大規模言語モデル[49]を利用した対話システムでコンペティションに挑んだとのこと。対話システムシンポジウムのライブコンペティション 4 のオープン/シチュエーションの両トラックにおいて LINE が 1 位を獲得した。「ペルソナ一貫性の考慮と知識ベースを統合した HyperCLOVA[50]を用いた雑談対話システム」が 2 部門トップの成績を収めたのである。

LINE の 2020 年 11 月の汎用自然言語モデル発表から数えれば約 1 年をかけて開発を進めたものと思われる。発表時点では日本語の巨大自然言語モデルとしては世界初となり 700PFLOPS を超える能力のスパコンを活用するとのことであった。パラメータ数は 1,750 億以上であり、日本語データは 100 億ページ以上を用意するのこと。実際に開発経緯を見ると日本語モデル最大 390 億、日本語を含むマルチリンガルモデル 820 億のパラメータモデルが使用可能となっている。

2021 年 11 月の LINE の開発者向けの説明では Natural response;会話が滑らかに達成されているかどうか、Following a topic;各応答に関してトピックに追従しているかどうか、Providing a topic or asking a question;話題の提供を応答ごとにできているかどうか、Achievement of goals;ゴール、すなわち対話の目的を達成できているかどうかを、きめ細かく評価した。LINE の解説によれば、種々のタスクにおいて異なるパラメータ数の自然言語

---

47 https://huggingface.co/rinna/japanese-gpt-1b

48 https://rinna.co.jp/ニュース/f/rinna 社、日本語に特化した言語画像モデル clip を公開

49 LINE による超巨大言語モデル採用の日本語 AI に関する発表 https://linecorp.com/ja/pr/news/ja/2020/3508、元となる韓国版 What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers., B.Kin et al. 2021, https://arxiv.org/abs/2109.04650

50 大規模な汎用日本語言語モデルを搭載した「HyperCLOVA」の現状- 2021 日本語版- https://www.youtube.com/watch?v=V4pZulIWHpY

モデルに対して上記の評価を行った結果、390 億パラメータモデルがほぼすべてにおいて最も高得点となったのであるが、例えばネガティブな感情を持ったユーザーとの対話での応答にまだまだ改善の余地があること等、さらなる改良が必要な部分があるとのことであった。

日本語の大規模自然言語モデルの準備とその活用が開始され、十分な性能が出ることが証明された貴重な結果となる。

4　汎用 AI につながる可能性を秘めた大規模自然言語モデル

4.1 Foundation Models

Stanford 大学では、巨大な自然言語モデルを利用したマルチモーダルな取り組みを評価した結果、言語に限らず種々の AI のタスクのための基本モデルとなりうることから Foundation Models として提案している[51]。ワークショップ[52]では、技術面はもちろんのこと現実味を帯びた汎用 AI の実現を視野に入れ、今後の社会で重要な位置づけになるとして、幅広いテーマに関して議論している。

図 13 は Foundation Models の位置づけを示しており、深層学習（Deep Learning）では特徴量を中心として発展しており、Foundation Models ではさらにその流れを発展的に展開し機能(Function)に重点が置かれるようになったとして説明している。

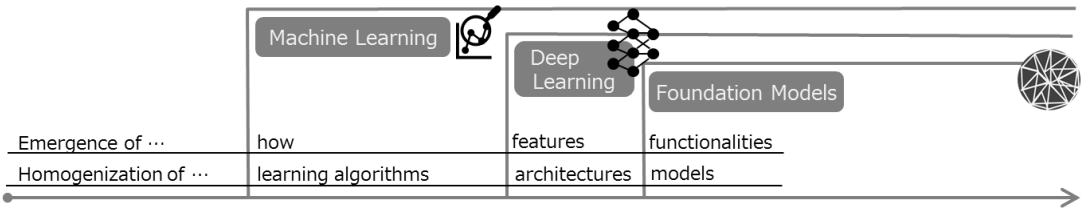


図 13 米 Stanford 大学 による Foundation Models の位置づけ
（出典：On the Opportunities and Risks of Foundation Models.,
https://arxiv.org/abs/2108.07258）

大規模かつ幅広いデータで学習された大規模自然言語モデルすなわち Foundation Models は幅広いダウンストリームタスク[53]に適応できるモデル（BERT、GPT-3、DALL-E、CLIP など）であり、明らかにパラダイムシフトを成し遂げている。もちろん Foundation Models は従来の深層学習の延長線の上にあり、自己教師学習と転移学習に基づいていると

[51] https://crfm.stanford.edu/, On the Opportunities and Risks of Foundation Models., R. Bommasani et al. 2021 https://crfm.stanford.edu/assets/report.pdf

[52] https://crfm.stanford.edu/workshop.html

[53] 分類、意図認識、固有表現抽出等の具体的な自然言語処理タスク

はいえ、一見すれば桁違いの学習データと学習パラメータ数の違い以外に目立った要素が無い。しかし、結果は新しい創発的な機能をもたらしていると感じられ、非常に多くのタスクにわたってその有効性が示されている。

Stanford の論点はまさにここからスタートする。一つのモデルがあらゆるタスクに対応でき、しかもマルチモーダルなタスク間タスクすら実現できている現状は、強力なレバレッジを提供する一方で土台となるモデルの欠陥はすべてのダウンストリーム、またはダウンストリーム用の適応モデルに継承されることになり、その影響は大きく細心の注意が必要になる。これまでの深層学習に比べてもはるかにパラメータ数が大きい Foundation Models では、なおさら、その広範なベンチマークの結果にもかかわらず、それらがどのように機能するか、いつ失敗するか、そしてそれらの創発的な特性のために何が一番重要なのかについての明確な理解が足りない状況なのである。

この点を現時点で重要視し、大きな流れのもと体系的に進めるというのが Stanford の指針の一つである。Foundation Models の機能（言語、ビジョン、ロボット工学、推論、人間の相互作用など）から技術原則（モデルアーキテクチャ、学習手順、データ、システム、セキュリティなど）に至るまで、その可能性とリスクについて検討し、アプリケーション（例えば法律、ヘルスケア、教育）および社会的影響（例えば、公平性[不公平]、誤用、経済的および環境的影響、法的および倫理的な考慮）への適用を考察するとどうなるのか？について議論されている。図 14 には議論した/すべき内容を示した。技術の面だけではなく、社会のなかでこの技術群がどのような位置づけになるのかに関しても重視しており、その可能性と影響力をよく表しているといえる。

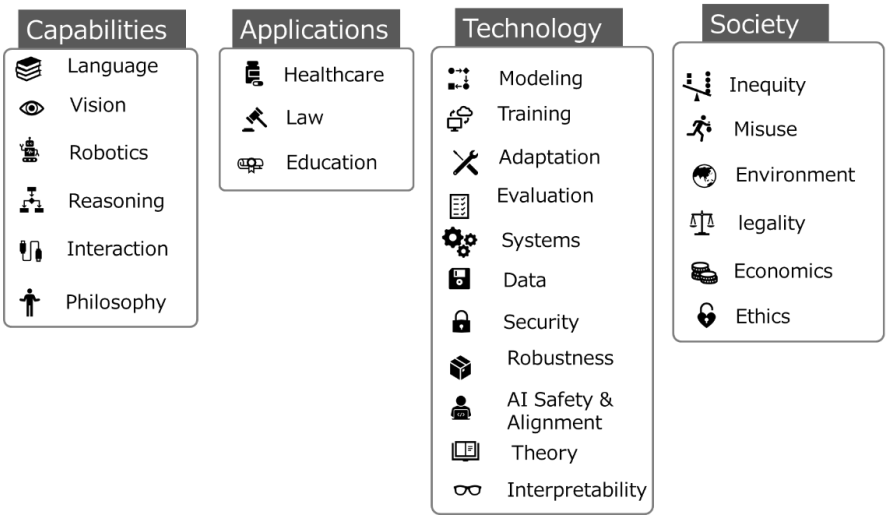


図 14　Emergence and homogenization での議論の対象範囲
（出典：On the Opportunities and Risks of Foundation Models.,
https://arxiv.org/abs/2108.07258）

4.2 自然言語モデルの本質

さて、自然言語モデルがこのようなマルチモーダルな性能を持つのか？これについてはワークショップ内でも議論になったが、基本的には非常に単純な理由によるものとコンセンサスがとれつつある。すなわち言語は人の営みにおいてほぼすべてを表現する能力を持っているので表現能力がいったんモデル化されればそれをモノ、コトに変換するのは割と簡単になるのではないかということである。つまり自然言語モデルは区別なく分散表現を学習するため学習データが多様かつ多岐にわたることで多くのモノ、コトを含むことになり、モダリティに関係なくテキスト記述できるレベルになるわけである。あとは、この分散表現をモノ、コトに戻すことができればマルチモダリティとして再現されたことになる。

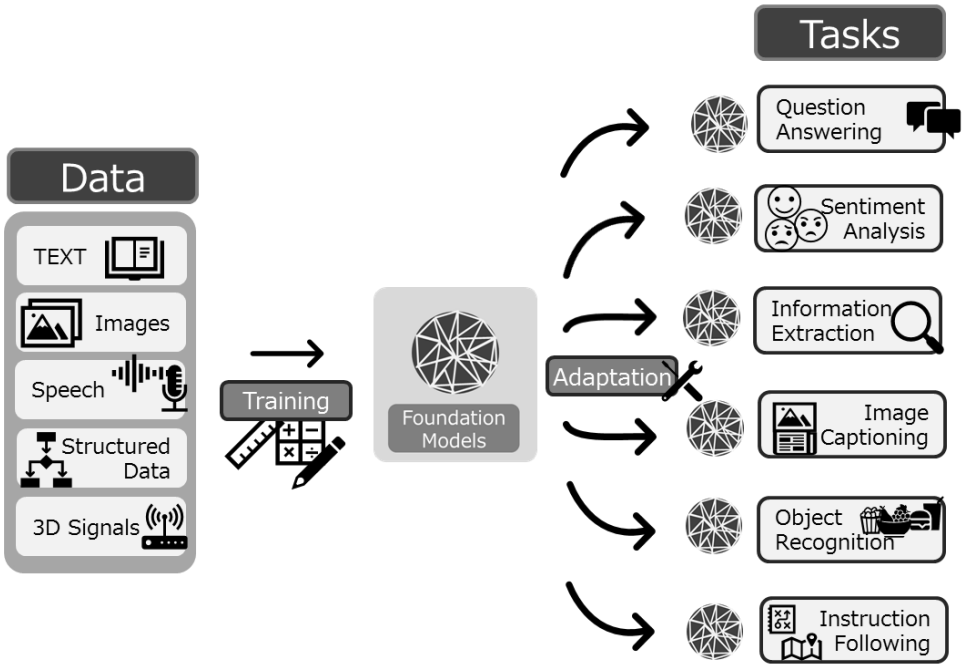


図 15　Foundation Models の概要

（出典：On the Opportunities and Risks of Foundation Models.）

Foundation Models は、明らかに人間のように機能する NLP システムを構成するために大きな前進を果たした。しかも取得した言語システムや学習プロセスは人間の言語学習とはおそらく全く異なるようにみえる。今後登場するより強力な Foundation models を有効に利用するためにも、その限界と可能性を把握するために機械学習の結果として得られる NLP と人間の言語学習の間にあるギャップの意味を理解することが非常に重要となる。例えば人間の言語習得は機械学習に比べて非常に効率的にみえる。例えば GPT-3 のような Foundation models は、ほとんどの人間が聞いたり読んだりするよりも約 3〜4 桁多くの言語データで学習する必要があり、そのような環境で育つ子供たちはいない。

4.3 ロボット工学と大規模自然言語モデル

大規模言語モデルの構築では、超多量のテキストデータを学習させる。その結果の四則演算の創発は前述したが他にも、学習データから言語以外の機能を創発する可能性を考えるのは自然である。以下では、一見、自然言語処理と遠く考えがちなロボット工学についての考察を紹介する。

図 16 に一般的な Foundation Models のデータと出力の関係とそれをロボット工学にあてはめた場合について示す。ロボット工学はさながら統合システムとして機能する必要があることが見えてくる。Foundation Models が文字通り基盤モデルとなりうるかはロボット工学への応用を考えることでより明確になることが期待できるわけである。さらにロボットが実環境とインタラクションをする必要があることも非常に重要となる。なぜなら汎用人工知能としてしばしば問題となる記号設置問題を明確に解決せざるを得ないためである。

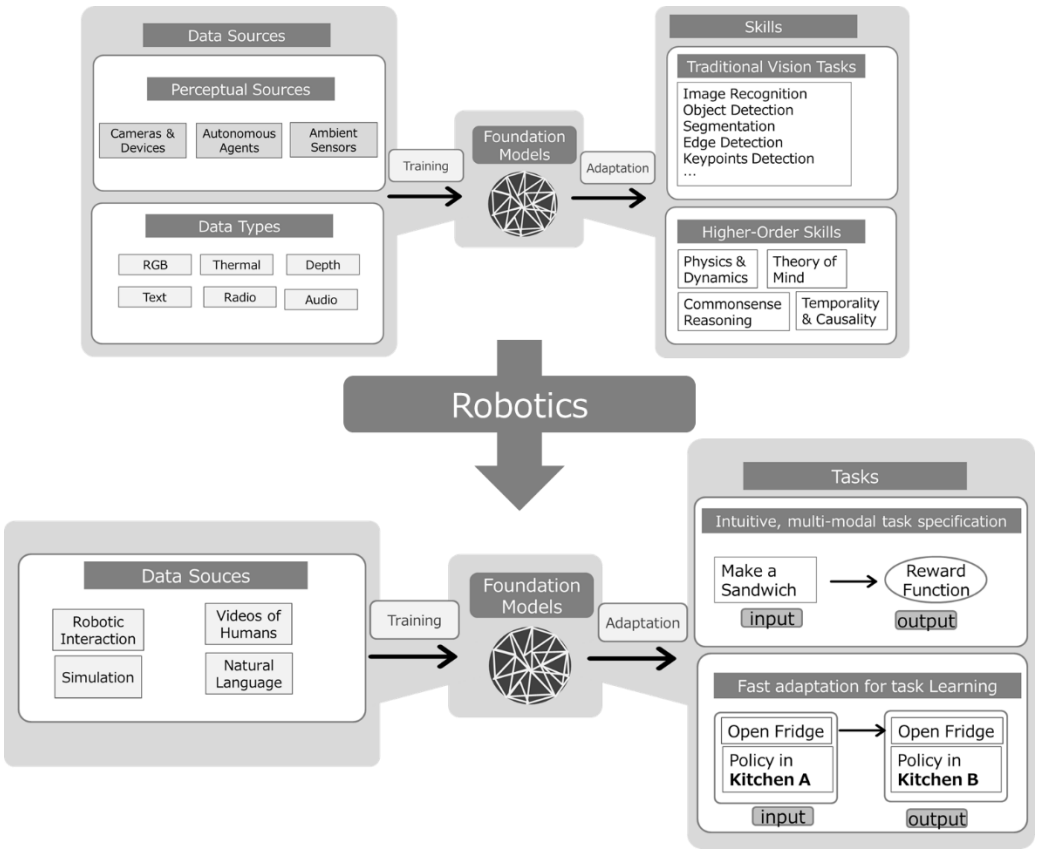


図 16　Foundation Models の活用例としてのロボット工学
（出典：On the Opportunities and Risks of Foundation Models）

ロボット工学には様々な環境と動作にまたがる大規模なデータセットが必要となる。現

在、実環境、仮想環境でどのようにデータを収集するかも含めて様々なアプローチが行われている。ロボット工学の視点を外しても基盤モデルとして、シミュレーション、ロボット同士による相互作用、人間のビデオ、自然言語での動作・環境の説明等はすべてデータとして有用になる。より複雑なデータソースから出来/不出来という評価が必要なこと、異なる環境で同一の行動を通常 Zero-Shot で行うことがタスクとなり、難易度が非常に大きくなる。

それゆえ様々なモダリティのデータは、動画を含む点で自然言語処理の枠組みを大きく超えるのであるが、基盤モデルとしての性質を考慮すれば一般的にロボット工学で用いられるタスクの仕様や教師あり/なし学習、強化学習を含む広範囲な問題の定式化につながると考えることができる[54,55]。

4.4 現実世界との対応

ロボットへの命令すなわち入力を普通の言葉で行うことは夢の技術の一つである。そのためにはヒト同士で会話する時に補完するであろう常識、前提条件をロボットが理解することが必要であり、先の DALL･E や CLIP の延長をロボットへの入力として考えてみることは非常に有用となるだろう。テキストの内容を画像に変換できるとすれば、同じようにテキストの内容をロボットの動作に変換できる可能性がある。CLIP の画像キャプチャ生成能力は実際に Room Rearrangement Challenge において深度画像や部屋配置地図などの使用モデルよりも優れた結果を示しているという報告もある[56]。つまり物体の配置は正確に把握できることになる。その上で例えとして一般的な理想としてのロボットとのインタラクションを考えてみる。

「朝食を作って」という入力で、それ以外の要素をロボットが推定・補完して朝食が用意されることを考えてみる。人間であれば、不確かな情報はその場で聞き返し補完する、そもそも食材に何があるのかを確認しつつ朝食を用意するだろう。さらに付け加えるのであれば、そもそもキッチンという場所の理解がないと朝食がつくれないことである。ロボットは実際の行動においては無数の条件を処理しつつ環境に対応しないといけないのである。この課題を解決するには、テキスト（音声）入力に対し、あらゆる情報を統合し物理的な具現化とつなげることが必要であり、自然言語処理や画像処理(コンピュータビジョン)での研究課題とは大きく異なる面を持つ。Stanford 大では Foundation Models をロボット工学の基盤モデルとすることでその可能性を確かめており、ロボット応用が日常生活の中で機能することを期待し、そのための実現アプローチの一つとなっている。

朝食を作るという高レベルのコンテキストには、非常に多くの曖昧さと相互に関係する独立ではないタスクを伴う。しかも物理世界とのインタラクションが生じるため、すべての

---

[54]Deep spatial autoencoders for visuomotor learning. J Chelsea Finn et al., 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE,

[55] Habitat 2.0:Training Home Assistants to Rearrange their Habitat., Szot et al. 2021, Neurips 2021

[56] https://github.com/allenai/ai2thor-rearrangement

タスクは細かい点まで具体性が必要となる。十分に具体的な明確さを備えたタスクとその目的のためにはサブタスクに分解、サブ目的を指定する必要がある。その上で実際に実行するタスクを指定できる。自然言語モデルが現在、すべてのタスクを具体化できることはないが、Foundation Models としての可能性は十分以上にあると言っていいだろう。ロボットのように環境とインタラクションを行う身体を持つ場合に適応できる Foundation Models はさらに進化した汎用人工知能であることは言うまでもない。

ロボットは、目的のタスクを理解することで階層的なタスク構造とそれに伴う様々な条件付けを行うことができることになる。タスク仕様は、通常は人間が提供するため、実際にロボットがタスクをどのように評価しながら完了するのか、タスクそのものを客観的に測定する定量的なメトリックが必ずしもあるわけではない。Foundation Models は人間が自然な方法で書かれたタスク仕様をロボットのタスク仕様に変換するために機能することになる。その変換には報酬関数による自己評価関数や動作に必要な信号が含まれることになる。つまりロボット自身がタスクを自ら評価することも可能になることになる。その結果ロボットの動作の最適化、障害の診断、人間へのフィードバック等も提供することが可能になると考えているようだ。

例えば、人間のオフラインビデオ、ロボットの相互作用、テキスト、および/またはシミュレーションから学んだ事前情報にはオーブンは通常壁に接しており、熱を発生させるためにオンにする必要があるという事実など、キッチンの一般的な説明をエンコードできることがわかっており、このような常識的な知識、物理的な事前確率、および視覚的な事前確率により、新しい環境への適応がより効率的にできることが期待できる。同様に、ロボットタスク学習の基礎モデルでは、学習データセット内の多数の料理動画を使用して、「卵を焼く」などの一般的なスキルのポリシーを学べるだけではなく、データの工夫により少数から特定のユーザーの好みに適合させることも可能となる。学習された Foundation Models であれば、普通に「朝食が食べたい」と記述すれば、自分好みの焼き方の目玉焼きとパンとコーヒーが出てくるのである。

更なる視点として Foundation Models のロボット工学への応用に関しては物理的な具現化に伴う大きな課題として安全性と堅牢性が挙げられる。先にも述べたがこの視点ゆえ自然言語処理やコンピュータビジョンとの大きな差がある。様々なモダリティのデータで学習を行って仮に満足のいく結果をバーチャル空間で得たとしても、ロボットの場合は実環境での動作が最終的に必要となるのはいうまでもない。その際に安全性と堅牢性を確保することが非常に重要となる一方、これが基盤モデルの開発をさらに複雑にすることになる。なぜなら、この視点は必ずしもデータに含まれているとは限らないためである。さらに問題なのは、実環境でのデータ収集は、物理的な世界で直接環境とインタラクションをとりながらなされるため、Foundation Models には存在しない情報を含んでいることである。そしてこの情報は安全性と密接に関係するだろうことは想像に難くない。通常はシステムに制約を課すことで実世界に制限を加えることになる。その環境下で致命的な失敗を行うことな

くタスクを学習する。

データを収集する前に安全性のシステム制約を指定する必要があるという鶏が先か卵が先かという問題はしばしば議論される。キッチンでの学習の際にキッチンに壊れ物がないことを確認するか、データを収集しようとしているときに破損する可能性のあるアイテムを確認して交換するべきなのか、そのままの環境で壊れることを学ぶべきか、環境内での相互作用と学習には検討要素が多い。中断のない学習を行う場合は前者を選択することになるが、新しい刺激、予期しない行動を一般化するためには後者も必要である。エージェントの因果分析[57]、安全性評価ツール、および現実的なシミュレーション環境[58,59,60]等が議論されている。

以上のようにロボット工学と Foundation Models は非常に密接な関係があると結論付けられる。同時にこの領域であればなおさらデータが重要であるということになる。実際の物理世界での多様な環境や実行状態を広い範囲でカバーするデータは、それ自身収集もしくは生成に大変な労力が必要になる。しかも、このデータはシステムの安全性と堅牢性に直接つながることになる。非常に重要な視点を与えてくれる。

では、Foundation Models をロボット工学が採用すれば、理想的なロボットが実現できるのかというと、実はそう簡単ではない。例えば服を着るときにボタンを留めるという動作を例にとると、これをロボットで実現するためには Foundation Models だけでは不十分なのである。これは Foundation Models にボタンを留めるという表現はあるだろうが具体的な手順、すなわちボタンをつまみ、ボタン位置にあった前身ごろの端部分(前立て)をつかみ、ボタン穴に裏側からボタンを概ね垂直に入れ込み、その後水平にボタンを戻すところまでの動作を行い、やっとボタンを一つ留めることになる。通常のテキストデータにはこのような記述はないだろう。同様なケースとして服をたたむという行為がある。実は人間が何気なく行っている動作に関しては Foundation Models が必ずしも有効であるとはいえない可能性があり、今後の進展に注目するしかない。

5 GPT-3 以降の自然言語モデル関連トピックの紹介

この章では GPT-3 発表以降の技術発表について簡単に触れる。GPT-3 の成功により様々

---

[57] Causal Analysis of Agent Behavior for AI Safety., G. Déletang et al. 2021, https://arxiv.org/abs/2103.03938

[58] A Survey of Algorithms for Black-Box Safety Validation., A. Corso et al. 2020, https://arxiv.org/abs/2005.02979

[59] Compositional Falsification of Cyber-Physical Systems with Machine Learning Components., T. Dreossi et al. 2017, In NASA Formal Methods, Springer

[60] Guaranteeing Safety for Neural Network-Based Aircraft Collision Avoidance Systems., K. D. Julian et al. 2019, IEEE/AIAA 38th Digital Avionics Systems Conference (DASC)

な取り組みが発表されており、直接的・間接的に GPT-3 を意識した方向での技術となっている。

5.1 Microsoft と NVIDIA の強力な自然言語モデル MT-NLG[61]

MT-NLG は Microsoft と NVIDIA によって開発された自然言語モデルであり、GPT-3 の 3 倍に当たる 5300 億のパラメータとなっている。105 層で構成されたこの現在の最大級のモデルは、発表時に既存のすべての自然言語処理モデルの性能を上回っていると発表された。これだけのパラメータの学習の場合はハードとの連携が必要であり、Microsoft と NVIDIA が協力することによって初めて実現可能となったと見ることもできる。432 個の機械学習用のテンソルコアと 80GB の RAM を備えている A100 GPU を 8 個搭載した DGX サーバーを 560 台つなぐことで構成されたスーパーコンピュータがその心臓となる。

学習データセットのサイズは 1.5TB で、Wikipedia や PubMed を含む 11 のデータベースから取得した数千億単位のテキストデータで構成されている。

MT-NLG は、意味によるテキストの完成の予測、読解、論理的結論の生成、自然言語での結論の作成、複数の意味を持つ単語の意味の区別などのベンチマークにおいて好成績を収めたことが報告されている。

また開発者のコメントでは、他の大規模自然言語モデルでの報告があるように最も単純な数学の理解について言及しており、機能の創発が確認できている。同時にいわゆるバイアス問題が存在していることも報告しており単に大規模化のみでバイアス問題が解決しないことを示唆している。参考までに以下、MT-NLG の例を示す。True or False の判定が成されていることが興味深い。文章の内容評価を行えていることになる。

Prompt:
The banker forgot that the judge advised the professors.
Question: The judge advised the professors. True or False?
Answer:(MT-NLG)True. The banker forgot that the judge advised the professors.

Prompt:
Famous professors supported the secretary.
Question: Professors supported the secretary. True or False?
Answer:(MT-NLG)True. The secretary was supported by famous professors.

---

[61] https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/、https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

5.2 Microsoft Project Florence-VL[62]

人工知能の中心的な目標の 1 つにマルチモダリティデータから効果的に学習するアルゴリズムの開発がある。繰り返しになるが画像とテキストの連携は、通常、私たちが周りの世界を理解するために行っている視覚と言語の連携と同様である。現在、この分野ではテキストクエリで最も類似した画像を検索する、あるいは画像から内容を説明するテキストを生成するというテキストと画像の双方向の変換が当たり前になりつつあり、Microsoft は一つのアプローチである Florence-VL(Florence - Vision and Language project)の報告をしている。

Web やソーシャルメディアは、画像とその注釈となるテキストのペアにあふれている。これらのテキストをラベルのソースとして使用すること、さらにビデオでは、内容を説明するオーディオチャネルを備えたものも多く、このオーディオはラベルとしてテキストに転記することも可能となる。これらのデータを視覚言語事前学習(VLP: Vision-Language Pre-training)に適用できれば、手動のデータラベル付けを必要としないという利点に加えて、特定のモダリティで学習した知識が別のモダリティでの学習に役立つ、テキスト⇔画像⇔音声でのクロスモーダル知識蒸留[63]が可能となる。

Florence-VL では事前学習としてコンテキストに基づいてマスクされた要素の予測など、従来の Transformer 系と同様の自己教師学習により Web やソーシャルメディアの大量の画像とテキストのペアデータを使用することで大規模モデルを作成した。このモデルに対してクロスモーダル表現の微調整を行い、いわゆるダウンストリームタスクに対応している。Microsoft はこの取り組みの結果を一連の論文として相次いで発表した。UNITER、OSCAR、VILLA、VinVL、VIVO、TAP とそれぞれ特徴を持っている。例えば、VIVO では、新しい画像キャプション（nocaps）タスクで人間と同等性を達成しているし、画像内で検出されたシーンテキストを使用して事前学習を強化することにより、TAP では TextCaps Challenge2021 で No.1 を達成した。

さらに Florence-VL を先に進めるためにエンドツーエンドの事前学習に関して UFO と METER の開発、統合された視覚画像モデルの UNICORN、スケーリングに関する LEMON と SimVIM、マルチモーダル Few-Shot Learning の PICa を開発している。これらの取り組みは、最終目標を念頭に置いて、互いに深く結び付いていることは以下のように示せる。例えば、エンドツーエンドの事前学習で開発されたモデルアーキテクチャは、統合された視覚言語モデリングの構成要素として機能し、さらにスケールアップするために開発した別の手法を使用することによって統合ソリューションとしてスケールアップすることにつながる。その際に Few-Shot 学習の機能が自然に使用できるようになり、その結果いくつかのコ

---

[62] https://www.microsoft.com/en-us/research/project/project-florence-vl/

[63] 知識蒸留:必要な知識要素のみを取り出すこと。深層学習の場合は主として機能する小さなネットワークを取り出す、もしくは圧縮して構成することになる。

ンテキスト内の例を使用すれば統合 VL 基盤モデルが誕生し、様々なタスクに容易に適応するわけである。

5.3 BigScience T0 [64]

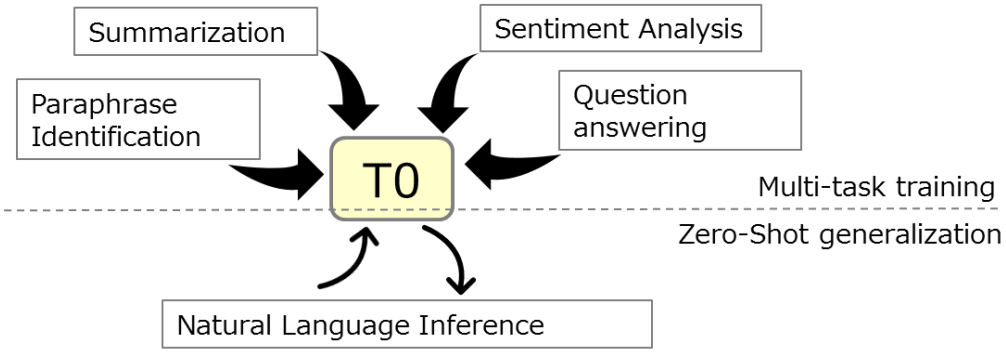


図 17　T0 モデル概要
（出典：https://bigscience.huggingface.co/blog/t0）

本報告書でも度々述べているように最近の大規模な自然言語モデルでは、多様なタスクで Zero-Shot 学習による汎化を示すことが報告されている[65]。これは、自然言語モデル学習におけるマルチタスク学習の結果であるという仮説が提案されている[66]。Zero-Shot 学習での汎化は、どのように機能するのか？マルチタスク学習による誘発はどのように機能するのか？この疑問を大規模にテストするために、一般的な自然言語タスクを人間が読めるプロンプト形式に簡単にマッピングするシステム T0 を実際に開発して詳細に分析した結果が報告された。大規模な教師付きデータセットを変換し、それぞれのデータセットには様々な自然言語を用いた複数のプロンプトが用意されている。

事前に学習したエンコーダー・デコーダーモデル[67,68]を、多種多様なタスクをカバーする

---

[64] https://huggingface.co/bigscience/T0pp
報告: https://github.com/bigscience-workshop/promptsource
論文: https://arxiv.org/abs/2110.08207

[65] Language Models are Few-Shot Learners., T. Brown et al. 2020, Advances in Neural Information Processing Systems

[66] Language Models are Unsupervised Multitask Learners. A. Radford et al., 2019, https://github.com/openai/gpt-2

[67] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer., C. Raffel et al. 2020, Journal of Machine Learning Research

[68] The Power of Scale for Parameter-Efficient Prompt Tuning., B. Lester et al. 2021

このマルチタスクプロンプトで微調整する[69]。結果は未知のテキスト課題に回答することに成功している。つまり Zero-Shot 学習が達成されており、いくつかの標準的なデータセットによる Zero-Shot 性能では、最大で 16 倍のサイズのモデルよりも優れている多くの結果を得た。また BIG-Bench ベンチマークのタスクのサブセットにおいても、最大で 6 倍のモデルを上回る性能を達成している。プロンプト学習によりモデルサイズを小型化できる可能性を示した。

5.4 Google の対話アプローチ LaMDA[70]

Google が会話に焦点をあて開発した LaMDA(Language Model for Dialogue Applications)が会話においてさらに進化した。Google の発明である Transformer は、単語と単語の繋がりを学習する。そのため次に来る単語を高精度で予測する。しかし対話においてはそれだけでは不十分であり、しばしば単語の裏に付随するニュアンスを理解したうえで予測を行う必要がある。そこで LaMDA では会話から学習する方法を取り入れた。

会話の中での応答を適切に返すためには単なる単語に関連した応答を返すだけでは不十分であり、会話のなかで現在扱っている内容に沿った文脈を理解し、その上で関連した、かつ、具体的な応答を返さなければならない。これは Google 自身が説明していたことだが「それはいい」という応答をすればかなりの確率で適当な応答となるが、人間の会話でもそうであるように、そのような具体性にかけた応答だけでは会話を続けるには不十分であり、発展的に会話を続けるためには、現在の会話の弾み具合を評価することも重要となる。

会話は最初から着地点、つまり最終的に行き着く先が決まっているわけではなく、時にはトピックが大きく変化して終了することもある。このような時、時として洞察に満ちている、機知に富んでいるといった評価をするものである。LaMDA ではまさにこの点も会話学習により、単語レベルでの関係性だけでは学習できない会話でのトピック間の関連性を学習するのである。もちろん、昨今問題となっている事実確認を含めた説得力のある回答を行う方法についても考慮しつつある。今後は会話能力をより多くの製品に組み込むために努力を続けるとしている。

2022 年 5 月の開発者会議において LaMDA2 が紹介された。LaMDA2 は数千人の同社社員により試されることでより洗練された。Imagine It と名付けられた LaMDA2 の機能デモの一つでは最初のヒトの入力から直接の関連内容の回答の後にあらかじめ設定されていないが、はっきり関連性がある話題への展開、対話を刺激する質問を経ながら、深海→マリアナ海溝（のシーン）→生物、潜水艦..が可能となっており、これらはモデルが学習データから統合し対応した結果となる。決して対話戻りが発生することはなく、発展していく対話が可能とのことである。

---

[69] プロンプト学習（https://huggingface.co/bigscience/T0pp）

[70] LaMDA: https://blog.google/technology/ai/lamda/

5.5 DeepMind RETRO（Retrieval Enhanced TRansfOrmers）

従来の大規模自然言語モデルでは、モデルサイズとデータサイズは大きければより性能があがる関係になっている。そこで RETRO では検索を利用して、直接、学習データセットそのものを検索(retrieval)することで、より小さなモデルサイズで性能を維持することに成功している。その結果、同じ数のパラメータを持つ標準的な Transformer ベースモデルと比較して大幅な性能向上を実現した。

自然言語モデルの学習には膨大な計算資源が必要であり、GPT-3 の 1,750 億のパラメータ、Microsoft の 5,300 億のパラメータのような大規模な自然言語モデルでは、学習にあまりにも膨大な計算能力が必要となる。そこで RETRO では、学習サンプルのサイズを縮小せずに学習コストだけを削減するために外部データベースを利用する。つまりモデル内で本来持つべきパラメータの一部を外部のデータベースにアウトソーシングするのである。これは Transformer の働きから単純化して考えることが可能である。注目する単語の前後関係の重みは通常大きい。とすれば、注目単語に関して重みが大きなテキストの最良の例は、その単語が含まれるテキストそのものであることは明白である。そのため注目する単語の近隣単語の重みについてはテキストデータそのものを利用して、より遠い位置関係は Transformer での学習を利用するとしても、十分に機能することは理に適っている。この場合、必要なパラメータ数は小さくなり、より少ないコンピューティングリソースの学習でも性能は十分達成されることになる。上記は、極端な例にしたがって説明しており実際に近隣を完全には無視できないため工夫をしている。

RETRO では、英語、スペイン語、ドイツ語、フランス語、ロシア語、中国語、スワヒリ語、ウルドゥー語を含む 10 言語のテキストを含むニュース記事、ウィキペディアのテキスト、書籍、GitHub のテキストで構成されるデータセットでモデルを学習する。RETRO ニューラルネットワークには 70 億個のパラメータしかない。そのため、これを約 2 兆のテキストパッセージを含むデータベースで補う方法をとっている。

RETRO がテキストを生成するとき、各文章（文書の 1 段落程度）に対して、最近傍探索を行い、学習データベースで見つかった類似のテキストとそのテキストから連続する結果を返す。これらの配列は、入力テキストから次のテキストを予測するのに役立てることになる。RETRO のアーキテクチャは、文書レベルでの通常の Self-Attention と、より細かいパッセージレベルでの検索された近傍探索との Cross-Attention を織り交ぜている。この結果、より正確で事実に基づいた継続を実現する。さらに、RETRO はモデル予測の解釈可能性を高め、テキストの正確性を向上させるために検索データベースを直接操作することで結果を改良する方法も提供している。データベースを使用して、書いたものと同様のパッセージを検索および比較する。これにより正確な予測を行うことになるのである。検索データベースのサイズを大きくするにつれて、自然言語モデルの性能が継続的に向上し、少なくとも 2 兆トークンまで確認されており、学習データとしての 2 兆トークンはそのまま外部データ

ベースとして使用された。

標準的な自然言語モデリングベンチマークである Pile での実験では、75 億パラメータの RETRO モデルは、16 データセット中 10 データセットで 1750 億パラメータの Jurassic-1 を上回り、16 データセット中 9 データセットで 2800 億 Gopher を上回る性能を示しました。パラメータサイズが 1/16 であっても十分な性能が出ることを示したのである。

さらにデータベースは、ニューラルネットワークを再学習せずに更新することもできる。これは、新しい情報をすばやく追加したり、古い情報や誤った情報を削除したりできることを意味している。DeepMind は、RETRO のような外部メモリシステムでは GPT-3 などのブラックボックスモデルと比較してより透過的であると主張している。データベースに関しては Web そのものを再利用する方法としては OpenAI の WebGPT[71]がある

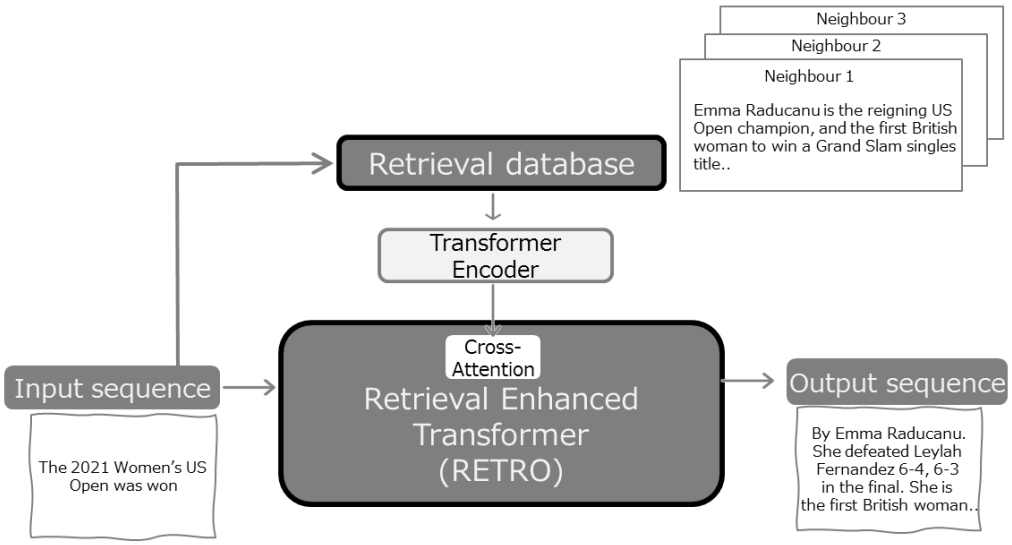


図 18　RETRO 動作概念図

（出典: http://jalammar.github.io/illustrated-retrieval-transformer/ より IPA にて作成 ）

5.6 韓国 LG EXAONE[72]

LG が Google と協力して作った韓国最大の AI となる大規模自然言語モデルを含む巨大 AI として EXAONE を発表している。LG が多業種からなるため Foundation Models で述べられている考え方のように自然言語モデルを中心に種々の AI 応用ができることがグループの戦略として理に適っており、その中心となるのが EXAONE といえる。

6000 億フレーズからなるテキストコーパスと 2 億 5000 万枚以上の画像を学習データとし、両データの結合を行っている。使用しているパラメータ数は 3000 億個となる。説明で

[71] https://gpt3demo.com/apps/webgpt

[72] LG AI Talk Concert; https://www.youtube.com/watch?v=b6B43VNW1jk

は大規模データを自ら学習し、人間のように思考・学習・判断できる AI であり、特定の用途に限定されず、多様な分野で活用できるという。EXA とはいうまでもないが 10 の 18 乗、つまり 100 京を意味する接頭語「EXA」の意味も持っており、人類がこれまでに使用したすべての単語をデータとして保存した場合にその量は約 5 エクサバイト（Exabyte）と推定されており、これが EXAONE に込められているというのが LG 側の説明だ。

LG AI Research はこれまでに段階的にニューラルネットワークのパラメータ数を変化させ、13 億個、130 億個、390 億個、1750 億個と段階的に大きくして、その性質を研究してきた。その結果やこれまでの報告からは、理論上パラメータが多いほど深層学習に基づく AI はさらに洗練された学習ができる可能性が高い。テキストとイメージの相互間学習であるマルチモーダル(multi-modality)学習では、明らかにパラメータ数が大きいことが利点となる。そこで大規模パラメータで実現したのが EXAONE である。

EXAONE も例にもれず「カボチャ模様の帽子を作ってくれ」と言えば、カボチャ模様の帽子を作り出す。LG AI Research では、こういった出力にいたる状態を、入力やデータを理解するレベルを超えて推論し、創造していると表現している。LG AI Research は、メタバース空間でクリスマスパーティーを準備する EXAONE の映像を公開しているが、その中では顧客が言う意図を理解、判断し、創造して飾る過程となっている。

LG AI Research の計画では EXAONE を製造・研究・教育・金融など、事実上すべての分野で「上位 1%水準の専門家 AI」として活躍できるようにするという筋書きがある。まず EXAONE の使用を促すために API を LG 系列社に公開し、電子・化学・通信など LG の事業全般に超巨大 AI を使用できるようにしている。各社はチャットボットの高度化や、過去 100 年間の化学分野の文献約 2 千万件に対する分析と学習による新素材・新物質の発掘などに EXAONE を実際に使用し着実に成果をあげているようだ。

LG グループは、EXAONE 完成に向け Google と密接に協力している。Google が 2021 年 5 月に発表したばかりの未発売の最新型の AI チップ「TPUv4」が EXAONE 開発に使われている。さらにソフトウエアにおいても Google Brain が協力している。今後 EXAONE は 6000 億パラメータに拡張するとともに、集団知性で超巨大 AI 生態系を生成するために国内を含むグローバル AI 連合軍を結成するという目標を掲げており、LG では AI 戦略が今後の発展のための中心的な役割を果たすと考えていることが分かる。その目標に協力しているのが Google Brain なのである。実は Google としては、LG への協力には大きな意味がある。AI 学習では NVIDIA が事実上、市場シェアの 80%を超える独占的事業者の地位を維持している。インフラ分野で LG が系列会社を基盤に B2B 事業モデルをつくりあげることができれば Google にとっては NVIDIA に挑戦する強力なリファレンスを得ることになるだろう。

5.7 中国 BAAI Wu Dao 2.0[73]

北京人工知能研究院（BAAI、中国名：北京智源人工智能研究院）は、「中国初」「世界最大」を謳うなんと 1 兆 7500 億ものパラメータを持つ事前学習済みの深層学習モデル Wu Dao2.0 を発表した。会話音声のシミュレーション、詩の作成、画像の理解はもとよりレシピの生成まで可能と発表しており、これまでの英語の自然言語処理モデルで可能なことがすべて中国語で可能になっているということである。モデルは、4.9TB の画像とテキストを学習しており、その中には中国語と英語の 2 つの言語のテキスト 1.2 テラバイトを含んでいる。すでに WuDao 2.0 には、スマートフォンメーカーの Xiaomi やショートビデオ大手の Kuaishou など、22 社のパートナーがいる。

この自然言語モデルは、BAAI によって 9 つのベンチマークでそれまでの最高レベル(SOTA: State-of-the-Art)に到達またはそれを上回ったことが報告されている。同時に Wu Dao2.0 の発表時には世界初の中国の仮想学生である Hua Zhibing の紹介があった。Hua は自ら学び、絵を描き、詩を詠うことができるという。将来、彼女はコーディングを学ぶことができるようにもなると紹介された。これらはオリジナルでかつ素の GPT-3 にはない能力であるが、一方で GPT-3 では Few-Shot Learning によりコード生成が可能となることが報告されている。Wu Dao の学習データの詳細は不明であるが、学習データの内容の差が初期の機能に現れている可能性もある。

5.8 OpenAI GLIDE、DALL･E2

CLIP は詳細なキャプションを画像に与えるため、モノの相対位置関係を画像から読み取ることができる。問題点の一つは、より抽象的な表現やある種のタスク、例えば画像内にある物体の数を数えあげる、画像の中にある自動車のうち一番近い車がどれくらいの距離にあるのか計測するという場合には限界があることがわかっている。タスクでの回答は、ランダムに推定するよりは多少上回るものの、高精度というのは程遠い状況であった。また、より細かい分類（車のモデル分類や花の種別分類など）でも十分な結果は得られていないことも問題であった。

CLIP はこれまでにない強力な Zero-Shot 分類器ではあるものの、言葉遣いや言い回しが変化すると対応できないことがある。その場合はプロンプト学習による試行錯誤が必要となり、対応はケースバイケースとなってしまうことになる。そのため従来の生成モデルである GAN[74]を含めて見直しが必要となった。

GAN は非常に性能が高い生成モデルであり、十分な生成機能の実現に様々な検討がなされてきた。現状はハイパーパラメータとレギュラライザーの高度なチューニングが必要となる状況であり、これは同時に、学習したドメインとは異なるドメインへの適用は非常に困

---

[73] https://wudaoai.cn/home

[74] GAN: Generative Adversarial Network(敵対的生成ネットワーク）深層学習を用いた画像生成。くわしくは DX 白書 2021 付録 第 1 部 AI 技術　第 1 章 9 創造を参照。

難であることを示している。そのため適用範囲内では品質が大きく向上するものの、適用範囲を越えて広げると品質が著しく低下するというトレードオフの関係が付きまとうことになる。この問題を解決するために GAN とは異なる方法として拡散モデルが提案された。

OpenAI では、この拡散モデルを改良し GAN より高い成果がでることを報告している[75]。画像生成の際に GAN ではノイズベクトルを使用するが、その際に生成される画像の保証は原理的にはない。そのため先の説明のようなチューニングが必要になるわけだが、拡散モデルでは、その点を改良するために元の画像データにガウシアンノイズを徐々に加え、データをノイズ化してゆき、元のデータが完全に失われてノイズのみになるまでマルコフ過程に従わせながら進める。生成モデルの学習は、ノイズを除去しながらデータを逆向きに復元することになる。つまり GAN では難しかった生成画像の制御をノイズと画像との関係を最初から最後まで学習することで可能にしたことになる。

GLIDE ではこの拡散モデルをテキスト→画像の対応モデルに適用した。具体的には、自然言語の記述を条件とするテキストエンコーダーを使った 35 億パラメータの拡散モデルを学習させたのだ。前モデルの DALL･E よりも GLIDE の方がフォトリアリスティックでは 87%、キャプションの類似性では 69%の好意的な評価が得られている。これらは Zero-Shot でレンダリング可能だが、複雑なテキストに対しては、フォトリアリスティックな画像生成が困難な場合が確認されたため、モデルを微調整（fine-tune）して画像の一部をインペインティングするための編集機能を持たせ、より複雑なテキストでも良質な画像生成ができるように強化した。その結果、画像の一部を塗りつぶして、テキストを入力すると、その箇所だけがテキスト通りに変わる。しかも、変わったその箇所は周囲の文脈のスタイルや照明に応じた影や反射を含み、周囲と調和し合成することが可能になった。

2022 年 4 月 6 日には DALL･E 2[76]が発表された。拡散モデルを採用した unCLIP により DALL･E との比較では一見しただけでも大きな進歩が認められる。図 19 に DALL･E 2 によるテキスト入力による画像生成例を示す。テキストの内容とそれを表す画像データからより緻密に画像が生成されていることが分かる。テキストから画像の生成機能においてはもはや想像力と創造力が実装されたのかと思わざるを得ないほどの進化となっている。さらに unCLIP では CLIP を超える表現能力を持つことが期待でき、例えばデジタルツインや世界モデルが数枚の画像から完全に作成可能となる可能性が高く、今後の検証に注目が集まる。

---

[75]Diffusion Models Beat GANs on Image Synthesis. P. Dhariwal et al. 2021, https://arxiv.org/abs/2105.05233

[76] https://openai.com/dall-e-2/ 、https://cdn.openai.com/papers/dall-e-2.pdf

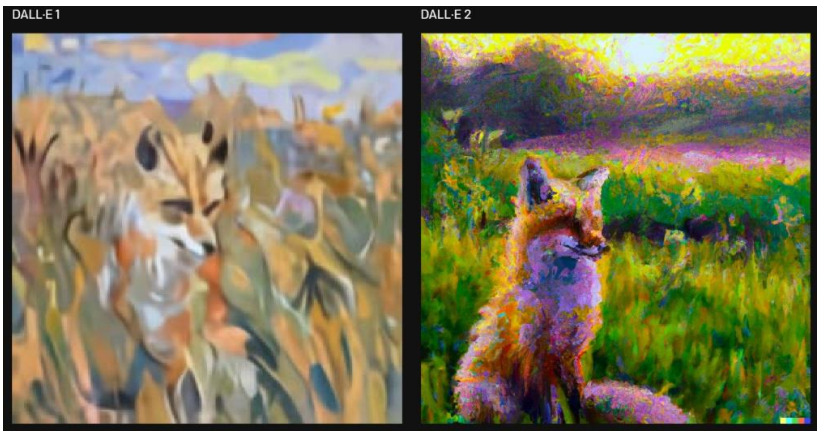


図 19 DALL･E 2 による画像生成例
“a painting of a fox sitting in a field at sunrise in the style of Claude Monet”
左：DALL･E、右：DALL･E2
（出典：https://openai.com/dall-e-2/より。付録 1 に様々な生成例を添付）

5.9 MLP[77]再考、Transformer 以降

Google からゲーティングを備えた MLP のみに基づく Attention 機構のない Transformer の改良型ネットワークアーキテクチャ gMLP[78]が提案されている。主要な言語および画像処理分野で Transformer と同様な性能を示した。最近、gMLP のような MLP がなぜ高性能なのか、詳しく解析され始めている。例えば Attention の効果の是非については議論が白熱している。例えば Attention が入力による動的なパラメータ（有効）バイアスの導入と考えれば、MLP は静的なパラメータバイアスということもできる。

バイアス（注目点の重み）の与え方で性能がどのように変わるかは、現在非常に重要な視点となっている。注目されている MLP では静的なバイアスを SGU[79]で実現している。厳密な表現とはならないが、Attention が前後関係、つまり繋がりを追うことでパラメータの重みに対する理由付けを明確にしていることに対し、SGU では単純な相関に基づいた計算となっている点が大きく異なる。画像では特にわかりやすい差となり、当然計算は後者の方が簡単である。Transformer の課題としてよく言われている計算コストが非常に高い点に対して計算コストを抑えつつ性能をあげる方法としての一つの解決方法として注目を集めており、性能がデータ数とパラメータ数でスケールできる理由の一つとなっている。明示的か否かは関係なく、注目ベクトルと他のデータベクトルとの重み関係を如何に考慮しつつ実装するのかが非常に重要となる。

---

[77] MLP: multi layer perceptron(多層パーセプトロン)ニューラルネットを順伝搬するよう多層結合した基本構造のひとつ

[78] Pay Attention to MLPs H. Liu et al. 2021, https://arxiv.org/abs/2105.08050

[79] SGU: Spatial Gating Unit　入力要素の系列方向の計算を行うゲート機構

別の視点としては Transformer の画像応用である ViT (Vision Transformer)と gMLP の比較や、その他の方法との比較の焦点の一つに、画像認識の基底の獲得がどのように行われるのかという点がある。ヒトの視覚の特徴量の基底は、生物学的[80]、DL、そしてテンソル解析[81]によるフレームレットのどれにおいても、ほぼ同様な形式が導出される。つまりモノを見て区別するための汎用的な基底は実は限られていると考えても問題ない可能性が高い。例えば我々が使用する文字、アルファベット等に関しては、もっと単純な基底群[82]であるという報告もあり、必ずしも多数のデータからの学習で得る必要がないかもしれない。

画像分野での CNN は明示的に基底を誘導するようにネットワーク構造を設計している。その基底もしくはそれに準じる特徴量フィルターがどのように抽出、構成するのか。Transformer ではポジションエンベディング層により、いわばそれがどこから来たのかを追跡することを行うが、SGU ではそういったことを行わず相関を用いることになる。そのためその分の差も計算コストに反映されるのである。

興味深いのは先に述べた脳の視覚野（低次視覚野）で確認されている基底とは別に、複数の図形から構成される相関基底がやはり脳の視覚野（高次視覚野）で確認されている[83]。脳は 20W で駆動される超高効率な計算システムであるため、様々な特徴量の有効活用をしていると考えられる。つまり、CNN や Transformer や gMLP に代表される現在有効であるといわれている方法のどれか一つではなく、おそらく今後の発展形も含めた方法を組み合わせて適応的に活用していると思われる。

最近になって HyperTransformer[84]による CNN の生成という結果の報告や MLP でのハッシュ計算による高速化の報告もあり、自然言語処理に使用された Transformer が CV 分野で新たな可能性を示し[85]、さらに他の分野の技術へ波及することが期待できる状況である。Attention に代表されるトークンミキサーモジュールを指定せずに Transformer のみを使用しても結果がよいことを示唆する別の報告では MetaFormer[86]が提案されている。MetaFormer とは、トークンミキサーモジュールが定義されていない Transformer の抽象化

---

[80] The complete pattern of ocular dominance stripes in the striate cortex and visual field of the macaque monkey. S. LeVay et al. 1985, J. Neuroscience,5

[81] Framelet analysis of some geometrical illusions.2010, H. Arai et al., Japan Journal of Industrial and Applied Mathematics

[82] The Structures of Letters and Symbols throughout Human History Are Selected to Maych Those Found in Objects in Natural Scenes., Changjzi et al. 2006, The American Naturalist

[83] Object Representation in Inferior Temporal Cortex Is Organized Hierarchically in a Mosaic-Like Structure. M. Tanifuji et al. 2013, J. Neuroscience

[84] HyperTransformer: Model Generation for Supervised and Semi-Supervised Few-Shot Learning., A. Zhmoginov et al. 2022, https://arxiv.org/abs/2201.04182

[85] コンピュータービジョン最前線 Winter2021, 2021 共立出版

[86] MetaFormer is Actually What You Need for Vision., W. Yu et al. 2022, CVPR2022, https://arxiv.org/abs/2111.11418

であり、トークンミキサーは、Attention または空間 MLP などに置き換え可能な一種のモデルアーキテクチャとなる。

例えば、単純な空間プーリング演算子に置き換えれば PoolFormer となる。PoolFormer は単純なプーリング演算子をミキサーに使用する。もちろんプーリング演算子の機能は、情報を混合するのではなく、トークンをその近くのトークンに平均的に集約するようにすること以上の機能はない。しかも PoolFormer では、プーリング演算子には学習可能なパラメータがない。それにも関わらず ImageNet-1k や DieT-B / ResMLP-B24 などのコンピュータービジョンタスクで高性能を示したことは驚きを伴う結果である。

以上は自然言語処理に使用された Transformer の画像処理への応用から派生した Transformer そのものの解析となるが、同じ流れで Attention は自然言語処理においても特定の課題には有効であるものの本質的に必要なのか問いかけをしている報告もある。

さらに最近では、モーダル間での学習方法に関しても考察があり、すべてのデータ、画像、音声、自然言語に対して同一方法でベクトル化する data2vec[87]の発表もある。完全なデータから表現学習を行う Teacher モードとマスクされているデータから完全なデータを予測する Student モードの 2 つにより、データの種類に関係なく潜在表現をえる。Transformer のどの層を予測(Masked Prediction)に使用するかを詳細に分析している。結果は従来の各モーダルに対して SOTA もしくは同様の結果となった。Transformer のその先をすでに探り始めており 2021 年はその最初の年だったといえる。

6 重要な課題

自然言語モデルに含まれるバイアスがしばしば問題とされることは先にも述べたがここで再度取り上げる。公平性や差別などしばしば社会問題として表面化している。自然言語モデルの場合には、学習データに大きく影響されることが分かっており、そのためにどのような修正が可能なのかはしばしば議論されている。アルゴリズムが公平性を有していることは前提の条件となり、その上で偏りがないか、例えば LGBTQ + ID 用語が学習データから除外されている、もしくは極めて少数である場合には、データ上では過小評価される、あるいは完全に消去される可能性があることも知らなければならない。

これはアルゴリズムが中立であれば、むしろ正しい動作の結果ではあるがモデルとしては大きな問題となる。学習データと自然言語モデル内での固有バイアスとの関係は正確には分かっていないため、小規模で体系的な調査を行いバイアスにも適用できうるスケーリング則を確立し、大規模なデータプラクティス応用できるようになるのではないかという

[87] https://ai.facebook.com/research/data2vec-a-general-framework-for-self-supervised-learning-in-speech-vision-and-language

のが現在の理解である。中立性かつ公平性が保障された学習用データセットをすべての自然言語モデルが利用しなければならないことが今後の課題である。

Google では自身の AI 原則に準拠しているかどうかを特に厳密に調査する。同社のサービスでは言語は中心的な位置を占めており、その性質、すなわち人類の最も優れたツールの 1 つであるということを前提に誤用を極力避ける工夫を随所に取り入れている。特に誤用による偏見、悪意のある表現、誤解を招く情報の複製などには十分以上の注意を払うことが要求される。例えば会話アプリケーションの場合は、会話の出力、つまり結果のみを注意深く精査したとしても、モデル自体が悪用される可能性が残るため、さらにこれを防ぐ必要もある。

そこで Google では LaMDA のようなテクノロジーを作成する際の優先事項は、リスクを最小限に抑えるようにするために行えることを十分に協議し、実行することとしている。長年にわたるこれらの技術の研究開発では、様々な問題、例えば不公平なバイアスが機械学習モデルにどのような影響を与えるのかなど、関連する問題に数多く対応している。さらにはこのような経験に基づいてモデルと学習対象のデータを分析するために使用できるオープンソースリソースも提供し、外部からのオープンな検証をも受け入れる。例えば LaMDA では、開発のすべての段階で精査していることを説明しており、同社の自然言語モデルおよび自然言語アプリケーションの公平性についての姿勢を表そうとしている。

2022 年の Google の開発者会議においても不正確、不適切、不愉快な内容への取り組みに関しては協調されており、社内ユーザーのフィードバックを積極的に活用していることを示している。同社の AI 倫理に従い、反復的規則的な原則に従った方法かつ外部の有識者の意見も取り入れていることを報告している。

2022 年 5 月に Google から Imagen[88]の報告があった。テキストから画像を生成する Google 版 DALL･E2 に相当しており、拡散モデルを採用し、いくつかの項目で SOTA を更新している。DALL･E2 とも比較しており、さらに洗練された画像生成結果が得られていることに間違いはない。ところが彼らはこれについてはデモの公開、コードの公開はしないとしている。これは次の理由によるとのこと。学習データに使用した画像データは広く使用されているデータセット、例えば LAION-400M[89]を使用しており、このセットには有毒な画像が多く含まれていることが知られている。もちろん注意深く有毒データは取り除いたものの、言語モデルに使用した Web レベルでのテキストデータは十分にキュレートされておらず、有害なステレオタイプや表現をエンコードしているリスクがのこることになるとの考察をしている。Google では今回、有毒な表現に対してもあえて評価することによって研究開発の進展とバイアス問題の両者において真摯に報告したことになる。

OpenAI においても GPT-3 の文章生成を改良しユーザーの意図をより正確にくみ取るこ

---

88 https://imagen.research.google/

89 https://laion.ai/laion-400-open-dataset/

と、有毒バイアスの処理を行った InstructGPT[90]を発表している。InstructGPT においてはヒトが強化学習にかかわる RLHF[91]と呼ばれる手法を適用する。人間の好みを報酬信号として使用することでモデルを微調整することが可能となり、文章生成での問題は複雑で主観的なため、現状の単純な自動メトリックでは完全に対応できない状況を改善するのに大きく役立っている。

7 最後に

本レポートでは自然言語処理のモデルからの発展を中心に最近の話題をまとめたが、大規模自然言語モデルはすでに汎用 AI の中心に位置づけられているといっていいだろう。今後も注視すべき代表的な AI 領域の一つである。現在のところ、大規模なモデルにはスーパーコンピュータレベルの計算資源が必要であること、より大規模なデータが必要であることから、強いところがより強くなる傾向となっており新規の参入がほぼできない状況である。しかし、Attention(Transformer)も、発表時は小規模な研究開発の結果からであったことから、アカデミックな視点での小規模な環境下での新しい計算モデルの探索はますます重要になることも事実である。一方、社会実装の面では、幸いに大半が公開されている訓練済みのモデルを如何に活用するのかという視点で開発を進める必要があるだろう。

最近の DeepMind の報告では[92]、モデルのパラメータ数と訓練データ規模との関係は必ずしも線形ではなく、実際に 700 億パラメータのモデルで 1750 億パラメータのモデルの性能をほぼすべてのベンチマークで上回るだけではなく、特定のベンチマークではヒトのスコアを上回っている。さらに 2022 年 4 月の下旬には 800 億パラメータの Flamingo[93]を visual language model として発表している。この画像内容記述能力はさらに洗練されたものとなっている(付録参照)。

Google は 2022 年 4 月 4 日に 5400 億パラメータの Pathways Language Model (PaLM)[94]について報告した。自然言語処理のベンチマーク上で実に 29 種類中 28 種において SOTA を更新したとしている。特にこれまで困難であったジョークを理解しているという結果、つまりどの点がジョークのポイントなのかを説明できていることは大きな衝撃を与えている。Chain-of-Thought Prompting と名付けられた複数の理解段階が必要な文脈に対応するためにマルチステッププロンプトによる解決を講じた。つまり複数の質問と回答のペアを与え

---

90 https://openai.com/blog/instruction-following/

91 Deep reinforcement learning from human preferences., P.F Christiano et al. 2017, https://proceedings.neurips.cc/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf

92 Training Compute-Optimal Large Language Models., J. Hoffmann et al. 2022, https://arxiv.org/abs/2203.15556

93 Flamingo: a Visual Language Model for Few-Shot Learning., J-B. Alayrac et al. 2022, https://arxiv.org/abs/2204.14198

94 https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html, PaLM: Scaling Language Modeling with Pathways., A. Chowdhery et al. 2022, https://arxiv.org/abs/2204.02311

ることで、一連の階層構造をも学習させることができる。その結果、階層的、段階的な理解が進むようになり、例えばジョークの内容を説明したり、数学の問題に対しステップを踏んだ回答を提示したり、より複雑なプログラムコードの生成を可能とした。

またスケール拡大法則[16]の観点でも PaLM の成果はパラメータ数が増大しても効率的な学習がまだまだ可能であることを示している。モデルのパラメータ数と訓練データ規模、そして性能に関してはまだまだ理解が始まったばかりであり、これからの進展と深い考察が待たれる領域である。

2022 年 4 月末にはこの分野のこれまでの進展からさらなる AI の進化を誘導するべく Useful General Intelligence を標榜する Adept[95]が設立された。メンバーは NLP、Transformer 等で名をはせており、動きの速いこの分野の象徴ともいえる。少なくとも動きについていけなければ大きな機会を失いかねないようだ。汎用知能に向けての取り組みとしては 2022 年 5 月 12 日に DeepMind からも A Generalist Agent と題して GATO の発表があった[96]。大規模言語モデルのように Tranformer で学習された単一モデルで機能するエージェントであり、マルチモーダル、マルチタスク、マルチボディのジェネラリストとして機能する。一つのモデルが、Atari、キャプション画像、チャット、実際のロボットアームとのスタックブロックなどの各タスクに対して、コンテキストに合わせる形での出力として、テキスト、関節動き、ボタンの押下、またはその他を出力するというまさにジェネラリストである。研究開発としてついにここまで来たかという成果であり Foundation Models の実現にまた一歩近づいた印象を得た。

最後に興味深い最近の取り組みとして Transformer の応用が言語とは関係ない領域でも大きな革新をうみつつあることについて付け加える。DeepMind が AlphFold[97]によりタンパク質の構造解析に大きな革新を与えた。さらに AlphaFold2[98]によりより大きな進展があり、AlphaFold が CNN をベースにしているのに対し AlphaFold2 は Transformer をベースとしていることは偶然の一致ではなく、Transformer 系のポテンシャルの高さからと考えるのは自然だろう。もちろん化学物質名から立体的な化学構造への変換という課題が記号とそれらが成す構造という観点で Transformer と親和性があったのも確かであろう。

同様に Materials Informatics (MI)はデータ駆動による材料開発の効率化・高速化する技術で大きな期待が持たれている[99]。この分野ではデータからの学習という観点においては、

---

[95] https://www.adept.ai/

[96] https://www.deepmind.com/publications/a-generalist-agent,

[97] Improved protein structure prediction using potentials from deep learning., A. W. Senior et al. 2020, Nature

[98] Highly accurate protein structure prediction with AlphaFold., J. Jumper et al. 2021, Nature

[99] 例えば Compositionally restricted attention-based network for materials property predictions., A. Y. Wang et al. 2021, npj Computational Materials, https://www.nature.com/articles/s41524-021-00545-1

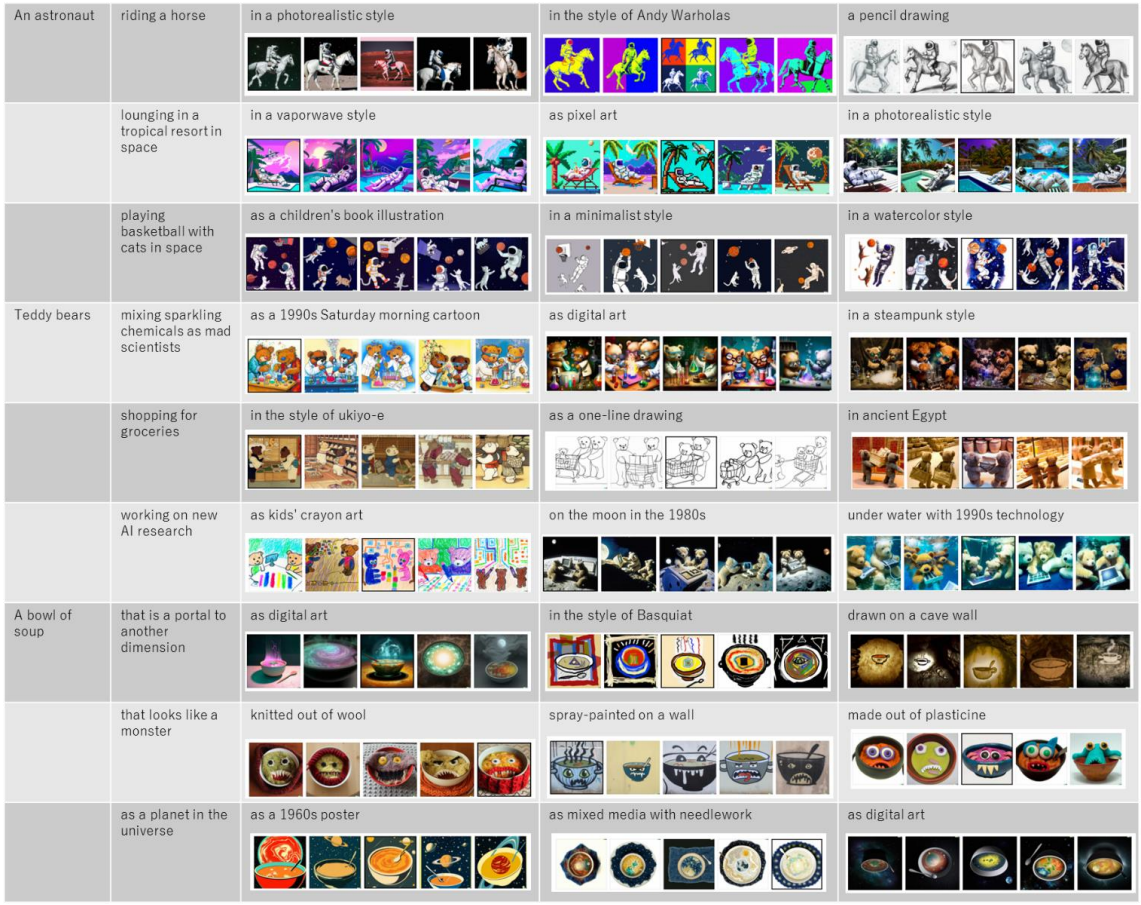
テキストデータに比較して圧倒的に少ないデータ量での工夫、極めて少ないデータからどれだけ正確に新材料の構造、性質を予測するかという特有の課題がある。最近 MI でも事前学習モデルが非常に有用であることがわかってきており、今後様々な分野へ波及していくと思われる。

【お問合せ先】
独立行政法人情報処理推進機構
社会基盤センター　イノベーション推進部　先端リサーチグループ
E-mail：ikc-ar-info@ipa.go.jp
電話：03-5978-7522

付録

自然言語処理と画像についての最新例として DALL･E2 と Flamingo の例を参考までに示す

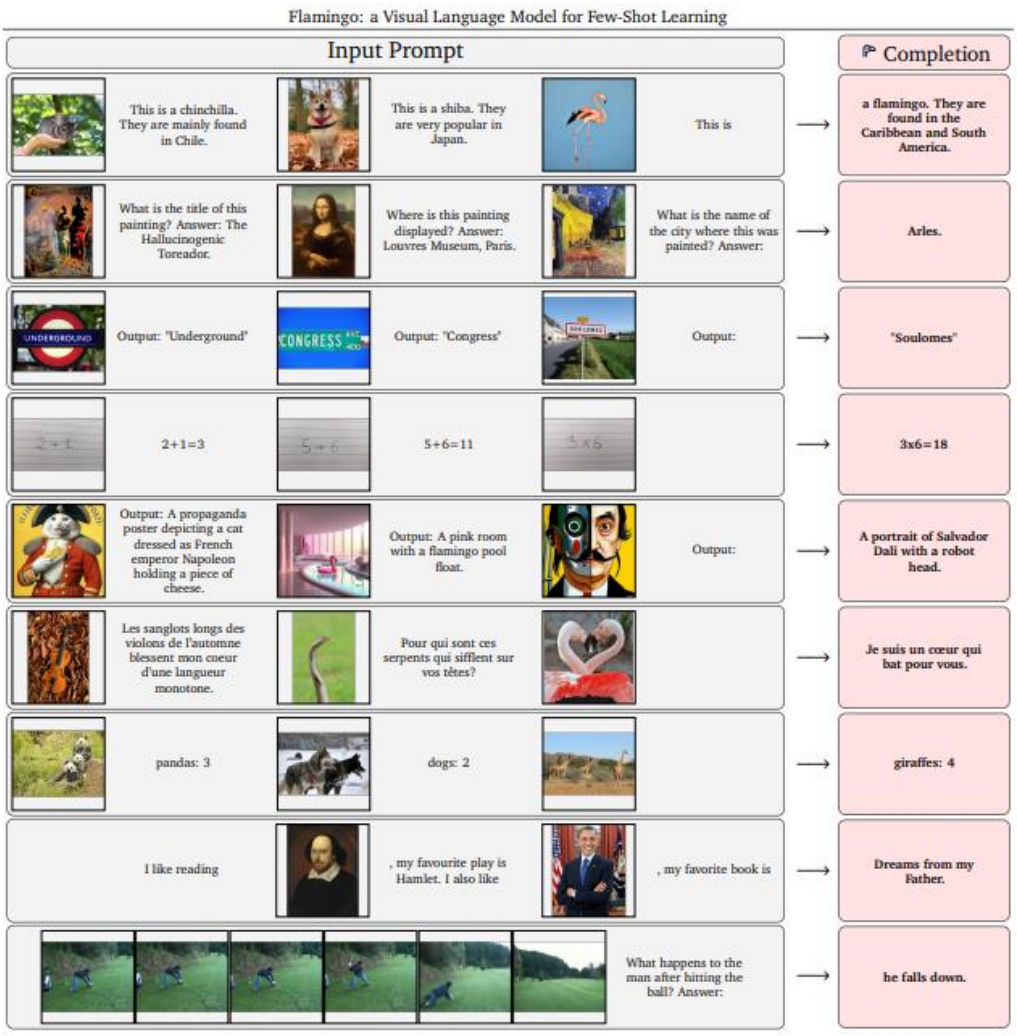
付録 1: OpenAI の DALL･E 2 の例



付録図 1 OpenAI DALL･E2 の例

（出典：https://openai.com/dall-e-2/より IPA にて作成）

付録 2: DeepMind の Flamingo の例



付録図 2　DeepMind Flamingo の例

（出典：Flamingo: a Visual Language Modelfor Few-Shot Learning.,
https://arxiv.org/pdf/2204.14198.pdf）

改訂履歴

2022/8/4　　P.3 図 2 修正