深層学習による自然言語処理の急進化と事業サービス応用における課題

菅（かん）　和聖（かずとし）／吉田光男（よしだみつお）

要　旨

深層学習モデルによる自然言語処理は、多くのタスクにおいて、それ以前の自然言語処理の性能を凌駕することが経験的に知られている。高い性能を発揮する深層学習モデルは、自然言語を効果的に取り扱うためのモデル構造の工夫を巧みに組み合わせて実現されている。近年では、大規模データをあらかじめ学習させた汎用モデルの普及、一連のデータ処理工程の構築の容易化などから、同分野への参入障壁が低下している。また、前処理ツールや言語データセットなどのリソースを無償で公開する慣習や、データの自動収集に関する法規制の緩和も、モデルの研究開発および普及を後押ししている。深層学習モデルを事業サービスに活用する際には、モデルがプライバシー情報や誤情報、その他の有害な表現を出力しないよう倫理的な配慮が求められるが、倫理に関する普遍的な規範はなく、万人に適した対応は困難である。加えて、深層学習モデルが自然言語を操るメカニズムは人間のそれとは原理的に異なるため、モデル性能の不確実性や限界を認識しておく必要がある。このほか、自然言語処理を念頭においた機械学習モデルに特有の情報セキュリティ・リスクにも注意すべきである。モデルの活用では、適用するタスクに即したモデル性能の評価を継続し、各種のリスク緩和策を講じていくことが求められる。

キーワード：　自然言語処理、深層学習、事前学習済みモデル、倫理、コーパス、セキュリティ

本稿は、日本銀行からの委託研究論文である。本稿の作成に当たっては、荒瀬由紀氏（大阪大学）と坂地泰紀氏（北海道大学）から有益なコメントを頂戴した。ここに記して感謝したい。ただし、本稿に示されている意見は、筆者たち個人に属し、日本銀行や筑波大学の公式見解を示すものではない。また、ありうべき誤りはすべて筆者たち個人に属する。

菅　和聖　日本銀行金融研究所企画役（E-mail: kazutoshi.kan@boj.or.jp）
吉田光男　筑波大学ビジネスサイエンス系准教授
（E-mail: mitsuo@gssm.otsuka.tsukuba.ac.jp）

1. はじめに

金融分野においてテキスト・データの活用が広がっている。金融市場の予測ではセンチメント分析が行われ、ニュース記事などをもとに自動的に売買の判断を行うアルゴリズム・トレードの枠組みが提案されている（例えば、Nuij *et al.* [2014]、Hu *et al.* [2018] を参照）。個人や企業への融資判断や信用スコアの算定では、有価証券報告書はもちろん、ソーシャル・メディアの投稿等の定性的な記述から抽出された有用な情報が活用されている。顧客対応では、顧客が入力したテキスト・データなどを分析して適切な回答を表示するチャットボットの導入が進んでいる。Araci [2019] は、金融業界用に深層学習モデルをカスタマイズすることで、文書分類の性能が向上することを報告した[1]。

こうしたテキスト・データの活用には自然言語処理（natural language processing）が用いられる。自然言語処理とは、人間が使う自然言語をコンピュータによって演算・処理することである。近年、機械学習、とりわけ、深層学習モデルを適用することによって、自然言語処理の性能が著しく向上しており、人間にしか担えなかった高度な言語処理を自動化できるようになってきた。また、自然言語処理に用いることができる良質なテキスト・データの蓄積、データベースの整備、API（application programming interface）サービスの提供を通じたデータへのアクセスの改善といった要因も、自然言語処理の性能向上に寄与している。すべての応用例に深層学習モデルが適しているとは限らないが、今後、データの蓄積が一段と進むと予想されるなかで、難度の高いタスクを解決する手段として深層学習モデルを活用するケースが増えると見込まれる。

深層学習モデルを適用した自然言語処理は、2018 年に BERT（Bidirectional Encoder Representations from Transformers、Devlin *et al.* [2018]）が提案されて以来、その性能が飛躍的に向上した。そして、機械翻訳や文書要約といった難度の高いタスクにおいて、深層学習モデルを適用した自然言語処理が性能面で伝統的な自然言語処理を凌駕するようになった。BERT は深層学習モデルの一種であり、提案当初、多くの機械学習のタスクで従来モデルの性能を大幅に上回ったことから注目を集めた。BERT は、今日のデファクト・スタンダードとなっている一連のモデル（「トランスフォーマー」と呼ばれるモデル構造を基盤に持つもの、詳細は 2 節（2）、（3）を参照）の 1 つである。

現在では、深層学習モデルを自然言語処理に適用する際には、以下に掲げる要素

1 その他の研究として、例えば、Kim and Yoon [2021] は、経営に関する定性情報から深層学習モデルを用いて企業の倒産確率を予測した。また、気候変動がマクロ経済や物価に与える影響に関して、金田・坂地［2023］は、気候変動に関する経済記事を分類して、その間の因果関係を分析した。

技術や概念、およびそれらの組合せが活用される。

① 自然言語という単語の記号列を数値のベクトル形式で表現することで、コンピュータによる処理が可能になった。とりわけ、自然言語を分散表現に変換することにより、数値ベクトルを入出力データとする深層学習モデルでの効果的な取扱いが可能になった。分散表現とは、単語を高次元のユークリッド空間に埋め込んだものである（詳細は 2 節（2）イ．を参照）。

② 確率を用いて自然言語を数理的に表現することにより、「コンピュータに自然言語を習得させる」という漠然としたタスクを機械学習モデルが求解可能な問題として定式化することができた。言語モデルは「文中での単語の発生確率」、系列変換モデルは「文から文への変換確率」を自然言語の生成過程であると仮定する。こうした数理的表現で自然言語を捉えたうえで、機械学習モデルはこれらの確率を推定することにより言語を「理解」[2] する。

③ 自然言語処理に適した深層学習モデルを作成するためのアーキテクチャ（モデル構造）の工夫が蓄積された。

a. 従来の深層学習モデルでは入力として固定長のデータを取り扱っていたが、可変長のテキスト・データを取り扱えるようになった。例えば、再帰型ニューラル・ネットワーク（recurrent neural network: RNN）や LSTM（long short-term memory）は可変長の単語列（文）を扱うことのできる深層学習モデルである（詳細は 2 節（2）ハ．を参照）。

b. 注意機構（attention mechanism、Bahdanau, Cho, and Bengio [2014]）を取り込んだトランスフォーマーを代表例として、文脈に関する情報をより深く処理できるようになった。自然言語では、文の中で離れた場所にある単語を参照することが多い。こうした参照関係は、深層学習モデルに注意機構を組み込むことで効率的に捉えられる。注意機構とは、深層学習モデルが単語や文字列（「トークン〈token〉」と呼ばれるが、2 節（1）ハ．のとおり、本稿ではこれらを厳密に区別しない）の情報を処理する際に、どのトークンを重視して処理するか（注意を振り向けるか）を自動的に学習する仕組みである（詳細は 2 節（2）ホ．を参照）。このほか、歴史的には、遠い参照関係の情報を深く処理する工夫は、LSTM でも取り入れられている。

さらに、以下の理由から、深層学習モデルを用いた自然言語処理への参入障壁が低くなっている。

2 ここでの「理解」とは、機械学習モデルが、あたかも言語を理解しているかのように自然言語処理のタスクを解ける状態になることを指す。6 節でも述べるように、機械学習モデルが人間のように言語を理解することを意味しない。

④ 深層学習モデルの訓練（training）を事前学習とファイン・チューニングに分離することによって、データが少ないタスクにも適用できるようになった。事前学習では、巨大な言語データセット（コーパス）を用いて汎用の言語モデル（事前学習済みモデル、詳細は2節（3）イ．を参照）を作成する。生成された事前学習済みモデルは、ユースケースに応じた小規模なデータセットによって再度訓練される（ファイン・チューニング）。このように、事前学習済みモデルは深層学習による自然言語処理の共通基盤として利用されている。

⑤ 伝統的な自然言語処理では、前処理から特徴量エンジニアリング（feature engineering）を経て後処理に至る各段階における分析技術をモジュールとして適切に組み合わせる。それら一連の処理の流れ（パイプライン）の設計では、分析実施者の熟練に頼る部分が大きかった。

深層学習モデルによる自然言語処理では、入力データから出力データまでの一連の処理を一気に学習するエンド・ツー・エンド（end-to-end）学習が可能になるため、パイプライン構築が容易である。深層学習モデルでは、パイプライン構築に代わり、モデル・アーキテクチャの設計が重要となる。例えば、系列変換モデルを用いた自然言語処理では、全体処理をエンコーダ部分とデコーダ部分に分離したエンコーダ・デコーダ・モデル（詳細は2節（2）ホ．を参照）が主流である。タスクに応じたモデルの設計では、エンコーダやデコーダのうち必要な部分のみを取り入れて、全体のアーキテクチャを決定する。

⑥ 深層学習モデルの研究開発に有用なコーパスや形態素解析ツールなどのリソースが無償で公開されている（詳細は2節（4）および3節を参照）。自然言語処理分野では、研究開発に有用なリソースを囲い込むのではなく、共有する文化が根付いている。

金融機関などの事業法人が深層学習モデルを用いた自然言語処理を活用する際には、情報セキュリティ・リスクについて留意する必要がある。一般に、深層学習モデルの挙動には不確実性があり、出力の品質を保証することが容易でない。例えば、顧客対応を行うチャットボットについては、想定外の応答をする可能性があるほか、顧客対応を通じて学習を進めるものについては、適切な文章を出力する精度を意図的に低下させる攻撃手法が存在し、そうしたリスクに晒されていることを意識しておく必要がある。また、非開示であるべき情報を引き出そうとする攻撃も想定される。こうした深層学習モデルの限界や弱点を理解したうえで、情報セキュリティ・リスクの軽減策を講じることが求められる。

情報セキュリティ・リスクに加えて、深層学習モデルが出力する文章の表現に倫理的な観点で問題がないかを事前に検査することも必要である。モデルの予期しな

い出力が、それを受け取る個人の人格を損なうものであったり、虚偽の情報を含むものであったりした場合、当該個人や金融機関のレピュテーションに悪影響を及ぼす可能性もある。そうした可能性がある用途においては、モデルを事前に検査することに加えて、その出力をそのまま使用するのではなく、人間によるチェックを経たうえでの使用を検討するなどの対応が望ましい。

本稿では、深層学習モデルによる自然言語処理の技術的な発展とその要因を解説するとともに、さまざまなサービスに適用される際に留意すべき事項を説明する。2節では、深層学習による自然言語処理を支える技術要素を、歴史的な位置付けとともに述べる。3節では、テキスト・データの収集方法や倫理面で配慮すべき事項を指摘する。4節では、自然言語処理を行う深層学習モデルに関する情報セキュリティ・リスクについて解説し、5節では、その他の留意点や今後の展望を述べる。

2. 深層学習による自然言語処理の基礎

画像データ処理を中心に発展してきた深層学習は、簡単には自然言語処理に適用できない。これを可能にするため、さまざまな技術的な工夫が開発または再評価されて深層学習に取り入れられてきた。本節では、これらの基本的な事項を説明する。詳細については、岡﨑ほか［2022］、坪井・海野・鈴木［2017］、近江ほか［2021］を参照されたい。

(1) 自然言語処理と機械学習

イ. 自然言語処理とは何か

人間がコミュニケーションなどに用いている自然言語は、文法構造と文脈および意味を持ち、単語が順序をもって並んだ列である点に注目すると、前後参照関係のある一次元の離散的な系列データと捉えることができる。しかも、距離が離れた単語間に強い参照関係が生じることが多々あるという特徴を持つ。こうした自然言語をコンピュータで演算処理することを自然言語処理と呼ぶ。自然言語処理で取り組むことが求められる典型的なタスクには、文書分類（document classification）、文書要約（text summarization）、機械翻訳（machine translation）、質問応答（question answering）、対話（dialogue）などがある。

文書分類とは、あらかじめ決められたカテゴリに文書を分類するタスクである。例えば、有価証券報告書のテキストから、当該企業に投資すべきか否かを自動判断する場合には、1つの方法として、「投資する」、「投資しない」という2つのカテゴ

リを用意し、有価証券報告書という「文書」をそれらのいずれかのカテゴリに分類する問題を解けばよい。このような分類処理をコンピュータに実行させる単純な方法は、人間が文書分類のルールを用意してプログラムで指示することである。例えば、有価証券報告書の「事業の状況」のテキストに「売上高が伸びている」が含まれていれば、「投資する」のカテゴリに入れるといったルールが考えられる。しかし、「投資する」と適切に判断を下すためには、検証すべき記述パターンに膨大な組合せが存在し、すべてのルールを網羅的に記述することは現実的ではない。こうしたルール・ベースのアプローチの限界を映じて、現在の自然言語処理では、コーパスを用いた統計的自然言語処理が主流になっている。

ロ. コーパス

コーパスとは、自然言語の文章を構造化し、大規模に集積した言語データセットである。コーパスには、文章中の単語の品詞や意味に関する情報、あるいは文書全体に関する付加的な情報が注釈として付与されているものがある。このようなコーパスは、特に注釈付きコーパス（またはタグ付きコーパス、ないしラベル付きコーパス）と呼ばれる。これに対して、付加情報のない大規模なテキスト・データは注釈無しコーパス（またはラベル無しコーパス、ないし生コーパス）と呼ばれる。

ハ. 形態素解析

典型的な自然言語処理では、処理対象となる文章に前処理を施し、それをトークンと呼ばれる細かい単位に分解する。英語の場合には、単語の境界がスペースで区切られているため、文章を単語に分解することは容易である。他方、日本語では単語の境界が自明ではないため、形態素解析を行う必要がある。形態素解析とは、文章を形態素（morpheme、意味を持つ言語の最小単位）に分解し、それぞれの形態素の品詞を判定する処理をいう。単に形態素の境界を挿入するだけの処理は「分かち書き」と呼ばれる。なお、本稿では「単語」、「トークン」、「形態素」を厳密には区別せず、詳細な違いには立ち入らない。以降では、トークンは概ね単語に相当するものとして扱う。

ニ. 統計的自然言語処理

統計的自然言語処理とは、計量的な手法を用いた自然言語処理の総称である。現在では、統計的自然言語処理が、単に自然言語処理と呼ばれる場合が多くなっている。統計的自然言語処理ではコーパスが用いられ、例えば統計的自然言語処理によって文書分類を行うケースでは、一般的に注釈付きコーパスが用いられる。

個々の有価証券報告書に対して「投資する」または「投資しない」というラベルが付与されたものも、注釈付きコーパスの 1 つである。統計的自然言語処理による文書分類では、注釈付きコーパスをもとに、「投資する」のラベルが付与された有

価証券報告書の集合と、「投資しない」のラベルが付与された有価証券報告書の集合を比較して、出現する単語の傾向の違いなどを定量的に学習し、それぞれの出現傾向によって自動的に分類する。このような自動分類の精度は、それに用いられる機械学習モデルの性能と密接に関係している。

ホ. 機械学習

機械学習とは、モデル[3]にデータを学習させることにより、そのデータのパターンを自動的に習得させる計算パラダイムの総称である。この学習の過程を訓練と呼び、訓練に用いるデータを訓練データ（training data）と呼ぶ。訓練データにラベルが付与されているか否かにより、機械学習は、教師あり学習（supervised learning）と教師なし学習（unsupervised learning）に二分できる。

教師あり学習は、入力データとラベルのペア（訓練データ）から両者の関係を学習する。そうして得られたモデルは、入力データを与えるとラベルを出力する。例えば、ラベル付きの有価証券報告書のデータを学習し、学習に用いていない有価証券報告書に対して「投資する」または「投資しない」というラベルを出力するモデルは、教師あり学習によって獲得できる。テキストの分類などに用いられる代表的な機械学習の手法として、教師あり学習であれば、サポート・ベクター・マシン（support vector machine）やランダム・フォレスト（random forest）、単純ベイズ法（naive Bayes）が挙げられる。

他方、教師なし学習では、ラベルを使わずにモデルを訓練する。テキスト分類に類似する教師なし学習の例として、クラスタリングを挙げることができる。クラスタリングによる文書分類では、ラベルなしの有価証券報告書の集合を入力し、これらをいくつかのクラスタに分割する。代表的な教師なし学習の手法として k 平均法（k-means）が知られている。機械学習モデルに入力されるデータは、一般にベクトルとして表現される。自然言語処理のモデルでは、トークンの列をベクトルとみなし、入力データとして用いるケースが多い。

ヘ. Bag-of-Words モデル

統計的自然言語処理では、トークンの集合または系列を入力データとして受け取り、さまざまな処理を実現している。トークンの集合を入力データとして取り扱うモデル（または処理）は、Bag-of-Words モデル（BoW モデル）と呼ばれる。このモデルでは、文章に出現する単語（words、概ねトークンに相当）をあたかもばらばらに袋（bag）に入れるように取り扱うもので、トークンの出現順序を考慮せず、出現回数のみを考慮する。トークンの出現順序を考慮しないシンプルなモデルであ

3 ここでのモデルとは、入力データと出力データとの関係を表現したものを意味する。モデルに入力データを与えると、出力データが得られる。その際、出力データはモデルに含まれるパラメータに依存する。

るにもかかわらず、文書分類や感情推定（emotion estimation）などで高い精度を示すことが知られている。

ト．言語モデル

文章や文が生成される確率をモデル化したものは確率的言語モデル（probabilistic language model）または単に言語モデル（language model）と呼ばれる。言語モデルは、文章または文としての自然さ（例えば、単語の出現順序が正常であるなど）をモデル化したものとみなすことができる。「本日は晴天なり」と「本日なり晴天は」を例に取ると、より自然な文である前者に高い確率が付与される言語モデルが高性能なモデルとされる。

言語モデルは文の自然さを測定できることから、機械翻訳や音声認識など、出力として自然な文が求められるタスクで用いられてきた。このようなタスクでは、一般的に、*n* 件の出力候補文（*n*-best）が生成され、その中から言語モデルのスコア（文の自然さの度合いを示す数値）が高いものを出力する。

チ．系列変換モデル

系列変換モデルは、文を受け取り、別の文へと変換する確率をモデル化したものである。これは、入力文を所与とした出力文の条件付き生成確率であり、言語モデルを拡張したものと捉えることができる。自然言語処理の多くのタスクは、この系列変換モデルによって表現できる。例えば、機械翻訳は、翻訳元の言語の文から翻訳先の言語の文への変換とみなせる。同様に、質問応答は質問文から回答文への変換、対話は相手の発言から自分の発言への変換、文法誤り訂正は文法に誤りのある文から正しい文への変換とそれぞれみなせる。

(2) 深層学習による自然言語処理

自然言語処理においては、単語が持つ概念の数学的な表現方法、そして、単語の系列である文の数学的な表現方法が問題となる。一般的に、前者は単語のベクトル化により、後者は言語モデルや系列変換モデルの利用により実現される。とりわけ、文（または文章）の表現方法に関しては、より難度の高いタスクを解決するには単語の並びが意味を持つ文脈を考慮することが必要と考えられるため、文中の単語の出現回数のみを利用する BoW モデルより言語モデルや系列変換モデルの方が有用である。深層学習の導入は、これらの有用なモデルを用いて、単語や文の情報を高度に処理することを可能とする。

イ．単語ベクトル表現と分散表現・単語埋込み

統計的に単語をベクトル化する際には、分布仮説（distributional hypothesis）が暗黙的に受容されている。分布仮説とは、「ある単語の意味は、その単語の周辺に出現する単語（周辺単語）によって表される」というアイデアである。そして、単語ベクトル表現の基本的なアイデアは、周辺単語を要素として単語をベクトル化するというものである。

例えば、「ある単語の意味は、その単語の周辺に出現する単語によって表される」という文を例に、「あるトークンは、その前後2つのトークン（合計4トークン）によって表現される」という分布仮説を採用したときの周辺単語を考えてみる。

まず、形態素解析によって文をトークンに分割する。トークンの境界にスラッシュ記号「／」を挿入すると、トークン分割された文は次のように表される。

> 「ある／単語／の／意味／は、／その／単語／の／周辺／に／出現／する／単語／に／よって／表さ／れる」

次に分布仮説に基づいて周辺単語を抽出する。上記の文には3つの「**単語**」が含まれており、それぞれの前後2つの単語を周辺単語として抽出する。具体的には、以下の下線が付された単語が周辺単語に相当する。

> 「ある／**単語**／の／意味／は、／その／**単語**／の／周辺／に／出現／する／**単語**／に／よって／表さ／れる」

こうして抽出された周辺単語（ここでは、「ある」「の」「意味」「は」「その」「周辺」「出現」「する」「に」「よって」）を用いて単語ベクトルを生成する。

分布仮説による単語ベクトル表現の獲得にニューラル・ネットワークを取り入れたモデルとしてWord2Vecが挙げられる。同モデルは、「king」、「man」、「woman」、「queen」という単語に対してそれぞれ生成された単語ベクトルにおいて、「king – man + woman = queen」という意味的な演算が概ね成立することから関心を集めた。このWord2Vecには、CBoW（continuous bag-of-words）とSkip-Gramという2つのモデルが存在する。このうち、CBoWは周辺単語から中心単語（上記の例の場合は「**単語**」）を予測する。Skip-Gramは中心単語から周辺単語を予測する2層ニューラル・ネットワークである。このようなニューラル・ネットワークで獲得された単語ベクトル表現は、特に、分散表現または埋め込み表現（embedding）と呼ばれる。

Word2Vecは、ニューラル・ネットワークとして2層に過ぎず、議論は分かれるところであるが、一般的には深層学習とはみなされていない。また、Word2Vecは、多義的な単語の意味が文脈によって変化するという言語的性質を正確に捉えることはできない。それでもなお、このモデルの成功が、自然言語処理におけるニューラル・ネットワーク活用の再評価と深層学習の活用につながる起点となったため、重

要なマイルストーンとみなされている。

ロ．深層学習

深層学習モデルは、多数のニューラル・ネットワークの層（layer）から成る。ニューラル・ネットワークとは、脳の神経回路網に学んだ機械学習モデルである。深層学習モデルのそれぞれの層では、活性化関数により非線型変換が施される。深層学習モデルは、深い層、多数のニューロン、活性化関数によって、モデルの入出力データの関係（関数）について高い表現能力を持つことが経験的に知られている[4]。この表現能力の高さにより、当初は画像処理のタスクで高い性能を発揮して、いわゆる第 3 次 AI（artificial intelligence）ブームの引き金となった。近年では、自然言語処理においても深層学習モデルの有用性が確認されている。

ハ．深層学習と系列データ処理

深層学習が広く注目されるきっかけとなったのは、Google LLC による画像認識であった。画像認識における深層学習モデルでは、畳み込みニューラル・ネットワーク（convolutional neural network: CNN）が頻繁に用いられるが、自然言語処理の分野では、現在、トランスフォーマー（本節（2）ヘ．を参照）を用いるのが一般的であり、CNN が用いられることはあまりない。文章はトークンの並び順が重要な意味を持つ系列データであり、CNN はこうした系列データを取り扱うことができないからである。そうした中で、系列データを取扱い可能なモデルとして、再帰型ニューラル・ネットワークやその改良型である LSTM が注目されるようになり、ニューラル・ネットワークの自然言語処理への応用の道がひらかれた。

画像処理と自然言語処理の違いとして、系列データの取扱いが重要であることに加え、可変長データの取扱いの重要性も指摘できる。この点、RNN や LSTM では、系列データの先頭から末尾まで再帰的に入力ベクトルの合成を繰り返す処理を行うため、系列データだけでなく、可変長データも取り扱うことができる。

ニ．言語モデルの発展

従来、言語モデルといえば、n 個の連続する単語の列を用いて文の生成確率を表す n グラム言語モデル（n-gram model）が主流であった。2010 年代になると、深層学習の発展とともに、ニューラル・ネットワークを用いて文の生成確率を表す（予測する）ニューラル言語モデル（neural language model）が n グラム言語モデルに取って代わった。

4 理論的にも、深層学習モデルの表現能力の高さを示す結果として普遍近似定理（universal approximation theorem）が知られている。同定理は、（層が 2 つあり、一層当たりのパラメータが十分に多い）深層学習モデルは、どんな連続関数でも（無視できる誤差の範囲で）表現できることを主張する。連続関数以外を対象とした拡張版の定理も発見されている。詳しくは、今泉［2021］を参照されたい。

ニューラル言語モデルは *n* グラム言語モデルと比較して、モデルの学習に要する計算負荷（計算量）は非常に重いものの、生成されるモデルが従来よりもパープレキシティ（perplexity: ppl）[5] などの指標で優れていることが実験で示されている。計算負荷が非常に重くなる問題については、GPU（graphics processing unit）の活用やコンピュータの性能向上により、徐々に解消されつつある。また、ニューラル言語モデルは、類似する文脈のデータから未知の文脈に対する確率を計算できるという特徴があり、*n* グラム言語モデルで問題とされた単語のゼロ頻度問題[6] を解消できるという大きな利点がある。代表的なニューラル言語モデルとしては、順伝播型ニューラル言語モデル（feed forward neural network language model）や前出の RNN を用いた再帰型ニューラル言語モデル（recurrent neural network language model）が挙げられる[7]。

ホ．系列変換モデルの発展と注意機構

Sutskever, Vinyals, and Le [2014] は、機械翻訳を対象として、入力と出力のそれぞれに LSTM を用いる系列変換モデル（sequence-to-sequence model）を提案した。系列変換モデルのアーキテクチャは、エンコーダ（符号化器、encoder）、注意機構、デコーダ（復号器、decoder）から成る構成が主流である。エンコーダは、入力文を数値データに変換し、これを合成して中間生成物である特徴量ベクトルを作成する。デコーダは、特徴量ベクトルから出力文を生成する。エンコーダとデコーダを分離するモデルのアーキテクチャを、エンコーダ・デコーダ・モデルと呼ぶ（図 1）。

注意機構は、エンコーダから受け取った特徴量ベクトルをデコーダが効率的に利用できるように、重み付けなどの情報処理を行う[8]。注意機構は、どのトークンにどの程度の注意を払うかを、入力された文章に応じて適応的に決定する。モデルに注意機構を組み込むことで、離れた位置にあるトークンの情報を適切に取り込み、深い文脈を考慮できるようになると考えられている。一般的に、注意機構は、膨大な情報の中から一部の重要な情報を選び出し、それ以外の情報を捨てるフィルタの役割を担っていると考えられており、自然言語処理以外の分野でも利用されている。

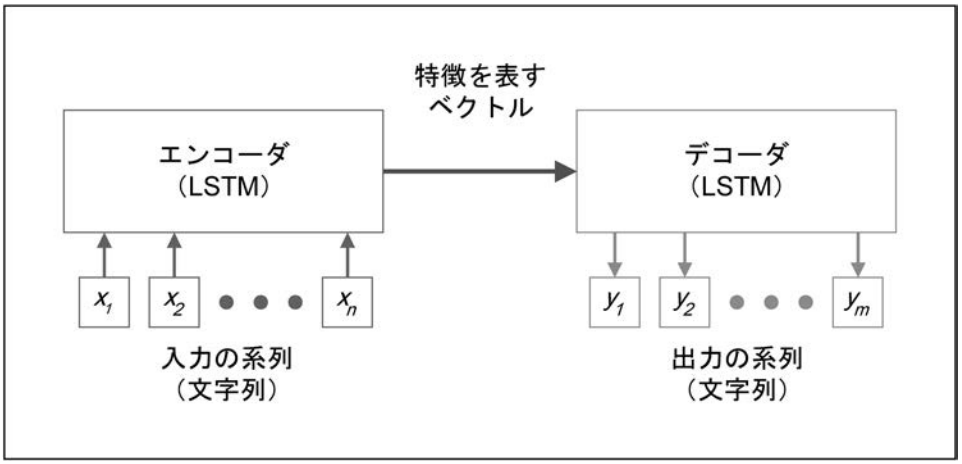
5 パープレキシティは言語モデルの複雑さを情報量の観点から捉える指標であり、低い値を持つほど良い言語モデルであるとされる。

6 ゼロ頻度問題とは、単語の生起確率をコーパスから推定する場合に、コーパスに出現しない単語の生起確率が厳密にゼロになり、そうした単語に関する推定を行うことができない問題を指す。

7 順伝播型ニューラル言語モデルは *n* グラム言語モデルと同様に、*n* 個の連続する単語の列を用いて次の単語を予測するモデルである。再帰型ニューラル言語モデルは再帰型ニューラル・ネットワークに基づいた言語モデルであり、ある単語より前に出現したすべての単語を考慮する点で順伝播型ニューラル言語モデルとは大きく異なる。

8 注意機構にはさまざまなバージョンがあり、それらは特徴量ベクトルの重み付けを行うという点では概ね共通するものの、必ずしもエンコーダとデコーダの橋渡しを行うわけではない。

図１　エンコーダ・デコーダ・モデルの概念図



ヘ．トランスフォーマーの登場

トランスフォーマーは、機械翻訳の性能向上のために考案された深層学習モデルであるが、機械翻訳のみならず、自然言語処理全般において高い性能を示すことが経験的に知られており、自然言語処理では欠かせない存在となっている。トランスフォーマーが初めて提案された論文（Vaswani *et al*. [2017]）の題は「Attention Is All You Need」（必要なものは注意機構だけ）である。この題名が示すとおり、トランスフォーマーにおいては注意機構が重要な役割を果たしているが、その高い性能は注意機構のほかさまざまな技術と経験知に基づく工夫によって実現されている。なお、トランスフォーマーでは、エンコーダとデコーダのそれぞれに、注意機構を特殊化した自己注意機構（self-attention）が組み込まれている[9]。

(3)　自然言語処理活用の障壁の低下

トランスフォーマーの登場は事前学習済みモデルの発展につながった。深層学習モデルの成熟は処理のパイプライン構築を簡略化させ、その結果、自然言語処理への参入障壁が低下した。

イ．事前学習済みモデル

自然言語処理における事前学習済みモデル（pre-trained model）は、膨大なパラ

9 より正確には、複数の注意機構を重ねたマルチヘッド注意機構（multi-head attention）を採用している。注意機構の詳細については、岡野原［2022a］、岡崎ほか［2022］などの基本書を参照されたい。

メータ数を持ち、膨大な言語データを学習させた巨大な汎用モデルである[10]。自然言語処理における代表的な事前学習済みモデルとしてGPT（generative pre-trained transformer、Radford *et al.* [2018]）やBERTが挙げられる。このうちGPTは執筆時点でGPT-4までバージョンが上がっている。事前学習済みモデルを作成するための訓練方法の1つに、文章（トークン列）のうちの一部のトークンを隠し、その隠されたトークンを予測するというタスクを処理する訓練がある。他の方法として、文章の流れを学習させるために、2つの文のペアを与えて、連続する文か否かを判定させるという訓練もある。事前学習済みモデルは、自然言語処理のさまざまな課題に応用できることから、言語の普遍的な特徴を捉えていると考えられている。

それぞれのタスクに特化したモデルは、事前学習済みモデルを拡張したモデルの一部のパラメータを再訓練することによって構築される。この再訓練の過程はファイン・チューニングと呼ばれる。あるタスクを解く際に獲得した知識（訓練済みモデル）を別のタスクを解くための学習に利用することを転移学習（transfer learning）と呼び、ファイン・チューニングも転移学習の一種に相当する。事前学習済みモデルを利用することで、教師データが少ないタスクであっても、高い精度のモデルを構築できることが知られている。

ロ．パイプライン構築とモデル・アーキテクチャの設計

1節⑤で述べたように、伝統的な自然言語処理では、モジュール化された処理を組み合わせてパイプラインを構築する必要があり、熟練の経験を要していた。これに対して、エンコーダ・デコーダ・モデルをはじめとした深層学習モデルでは、パイプラインの構築が大幅に簡略化され、入力から出力までの処理を一気に学習するend-to-end学習が可能になった。

深層学習モデルではアーキテクチャの設計が重要となる。設計に当たっては、エンコーダやデコーダなどのモデルの部品を選定し、モデルに組み込まれる注意機構などの演算処理を改良するなど、主流となる方法論はある程度標準化されている。これにより、自然言語処理のモデル開発の参入障壁が大幅に低下することとなった。

トランスフォーマーは、機械翻訳の性能を向上させたモデルとして注目を集め、現在までに、トランスフォーマーを基盤のモデル構造に採用してアーキテクチャを設計したさまざまな派生モデルが提案されている。これらの一部には、トランスフォーマーを原型として、エンコーダとデコーダのうち必要な部分を取り入れて改良を施したモデルも存在している。例えば、エンコーダを事前学習したBERT、デコーダを事前学習したGPT、エンコーダとデコーダを訓練したBART（Bidirectional

10 事前学習済みモデルは、分野ごとに高度かつ汎用的な処理を行う基盤として利用される。自然言語処理分野以外にも、例えば、コンピュータ・ビジョン分野では、事前学習済みモデルは、画像から物体を認識する基盤を提供する。

and Auto-Regressive Transformers、Lewis *et al.* [2020]）などが知られている。このうち、BERT は、トランスフォーマーのエンコーダのみを採用し、文全体を参照しながら複数のタスクで事前学習を行うように改良が施されたモデルである。2022 年時点において、利用される事前学習済みモデルのほぼすべてがトランスフォーマーを基盤のモデル構造として採用している。

(4) 有用なリソース

自然言語処理の活動コミュニティでは、伝統的に、論文やツールなどのリソースに無償でアクセスできる環境が整えられている。こうしたリソースの公開も、同分野の研究開発への参入障壁の低下に寄与している。リソースの利用方法については、例えば、榊ほか［2022］を参照されたい。なお、本稿で紹介するリソースの大半が商用利用可能である[11]。

イ. 論文

自然言語処理に関する学術論文の多くはオープン・アクセス（無償で誰でもアクセス可能）であり、主要な論文は ACL Anthology[12] で公開されている。同様に、日本の言語処理学会が発行している学会誌「自然言語処理」[13] も無償でアクセスできる。なお、自然言語処理に関する学術発表では、他の情報工学分野と同様、査読付き学術雑誌論文よりも、査読付き国際会議論文を優先する傾向がある。ACL Anthology で公開されている論文の大半が、査読付き国際会議論文である。研究者は分析対象として英語の言語データを選択しやすい土壌にある。

ロ. 形態素解析器・日本語解析ツール

自然言語処理において日本語のテキストを取り扱う場合、事前に形態素解析を行う必要がある場合が多いが、これを支援するツールもオープン・ソースで公開されている。従前は、速度や精度の面から MeCab[14] が広く利用されていた。近年では、形態素解析に加えて、人名や地名などを抜き出す固有表現抽出（named entity recognition）や単語同士の係り受け構造の解析などの機能も持つ GiNZA[15] が利用さ

11 筆者らが確認した限り、国立情報学研究所の情報学研究データリポジトリ（Informatics Research Data Repository: IDR）を通じて公開されているリソースは利用目的を限定しているものの、そのほかのリソースには限定はなかった（2022 年 12 月時点）。ただし、クリエイティブ・コモンズ・ライセンスの CC BY-SA 4.0 を適用するなど、成果物の公開方法に一定の制限のかかるリソースもあるため、使用前にライセンスを確認されたい。

12 https://aclanthology.org/

13 https://www.jstage.jst.go.jp/browse/jnlp/-char/ja

14 https://taku910.github.io/mecab/

15 https://megagonlabs.github.io/ginza/

れる機会も増えている。

MeCab や GiNZA による解析処理の一部は、人手により作成・更新される辞書に依存している[16]。この手法は、品詞や読み方など、辞書に登録されている詳細な情報を出力できる反面、辞書に登録されていない未知語が出現すると、文を適切に分割できない場合がある。品詞の情報などを利用せず、分かち書きのみが必要であれば、辞書に依存しない Sentencepiece[17] が用いられることもある。

ハ. コーパス

統計的自然言語処理はコーパスによって支えられている。Wikipedia[18] の記事データや青空文庫[19] に掲載された作品のテキスト・データは、機械学習の訓練データとしてよく使われる。よく使われるが故に周辺ツールや分析例も豊富であり、自然言語処理の練習用データとしても適している。

これらのほかに、下記の公開データも分析の練習に使うことができる。

a. livedoor ニュースコーパス（NHN Japan 株式会社）[20]
b. 死傷災害（死亡・4 日以上）データベース（厚生労働省）[21]
c. ニコニコデータセット（国立情報学研究所）[22]
d. 地方議会会議録コーパス（地方議会会議録コーパス・プロジェクト）[23]
e. 企業分析用データセット「CoARiJ」（TIS 株式会社）[24]

より大規模なデータベースとして、ウェブから収集した自然言語（テキスト）をもとに作成されたコーパスがある。フランス国立情報学自動制御研究所による OSCAR（Open Super-large Crawled Aggregated coRpus）[25] と Google LLC による mC4（Multilingual Colossal Clean Crawled Corpus）[26] は、いずれも代表的な多言語ウェブ・コーパスである。mC4 は、Multilingual T5（Text-to-Text Transfer Transformer）と呼ばれる大規模言語モデルの学習に利用された実績がある。

日本における情報学に関する主要な研究データは国立情報学研究所の情報学研究

16 辞書に依存する場合、その辞書が更新されない限り、新語を上手く抽出できないという問題を抱える。MeCab で標準的に用いられる辞書でも、近年の新語が含まれておらず、mecab-ipadic-NEologd などのユーザ辞書を併用することもある。
https://github.com/neologd/mecab-ipadic-neologd
17 https://github.com/google/sentencepiece
18 https://ja.wikipedia.org/
19 https://www.aozora.gr.jp/
20 https://www.rondhuit.com/download.html#ldcc
21 https://anzeninfo.mhlw.go.jp/anzen_pgm/SHISYO_FND.html
22 https://www.nii.ac.jp/dsc/idr/nico/
23 http://local-politics.jp/
24 https://github.com/chakki-works/CoARiJ
25 https://oscar-project.org/
26 https://www.tensorflow.org/datasets/catalog/c4

データリポジトリ[27]を通じて公開されており、利用申請が必要ではあるものの、研究にも使われている信頼性の高いデータを取得できる。

ニ．事前学習済みモデル

単語分散表現などのモデルを構築するには大規模なコーパスが必要となるが、すでに構築済みのモデルが多数公開されている。日本語 Wikipedia のテキスト・データをもとに構築されたモデルとしては、例えば東北大学 BERT[28] がある。また、先に取り上げた mC4 に含まれる日本語データで学習した ELECTRA（BERT を改良した手法）の学習済みモデルも公開されており[29]、GiNZA の中で ja_ginza_electra として利用できる。汎用的なモデルのほかにも、特定のドメインに特化したモデルもあり、東京大学の和泉研究室は金融に特化した学習済みモデルを公開している[30]。

ホ．ツール・ライブラリ

Hugging Face[31] という主に自然言語処理を対象にした大規模なオープン・ソース・コミュニティでは、さまざまなデータやモデルが公開されており、Hugging Face Transformers[32] を用いると最先端の深層学習モデルを利用することができる。また、有志による日本語の自然言語処理に関する Python（パイソン）ライブラリ、学習済みモデル、辞書、およびコーパスの厳選リストも公開されている[33]。

既に独自のテキスト・データを保有していて、そのデータをもとに分散表現を獲得したい場合には、比較的軽い負荷のもとで作動し、一定の性能が確保されているオープン・ソースの fastText[34] を利用することが考えられる。fastText では、分散表現の獲得から一気通貫でテキスト分類（文書分類）まで実行可能である。

3. 自然言語処理に関するデータの収集・倫理

本節では、深層学習を用いた自然言語処理に必要となるデータを収集する方法や、収集に際して倫理面などで留意すべき事項を説明する。特に、ウェブ上で公開されているテキスト・データを自動的に収集する場合に焦点を当てる。

27 https://www.nii.ac.jp/dsc/idr/
28 https://github.com/cl-tohoku/bert-japanese
29 https://huggingface.co/megagonlabs/transformers-ud-japanese-electra-base-discriminator
30 https://sites.google.com/socsim.org/izumi-lab/tools/language-model
31 https://huggingface.co/
32 https://github.com/huggingface/transformers
33 https://github.com/taishi-i/awesome-japanese-nlp-resources
34 https://fasttext.cc/

(1) 公開データの利用

統計的自然言語処理、とりわけ深層学習を用いた自然言語処理には、大規模なテキスト・データが欠かせない。同分野における研究開発は、大規模なデータセットを作成・公開することに価値を認める文化的価値観によって支えられている。

近年のオープン・サイエンス、オープン・ガバメントの潮流や、コンペティション形式のアルゴリズム開発[35]の普及などにより、さまざまなデータが公開されるようになった。また、いくつかの研究グループや研究機関では、自前で収集・整備した研究用のデータを公開している。これらの公開データを用いることで、自らデータを収集せずとも、自然言語処理の研究開発に取り組める環境が整っている。具体的なリソースについては、2節（4）を参照されたい。

(2) ウェブからのデータ収集

自身の研究開発に適した公開データが存在しない場合、自身でデータを収集する必要がある。ウェブ上で公開されている情報は、ウェブ・ブラウザを介してアクセスすることで収集することができる。その作業を人手で行うことも不可能ではないが、人間の作業効率の限界から大規模化することは難しい。コンピュータを使って自動的に収集できれば、大規模なデータを効率的に収集することも可能となる。

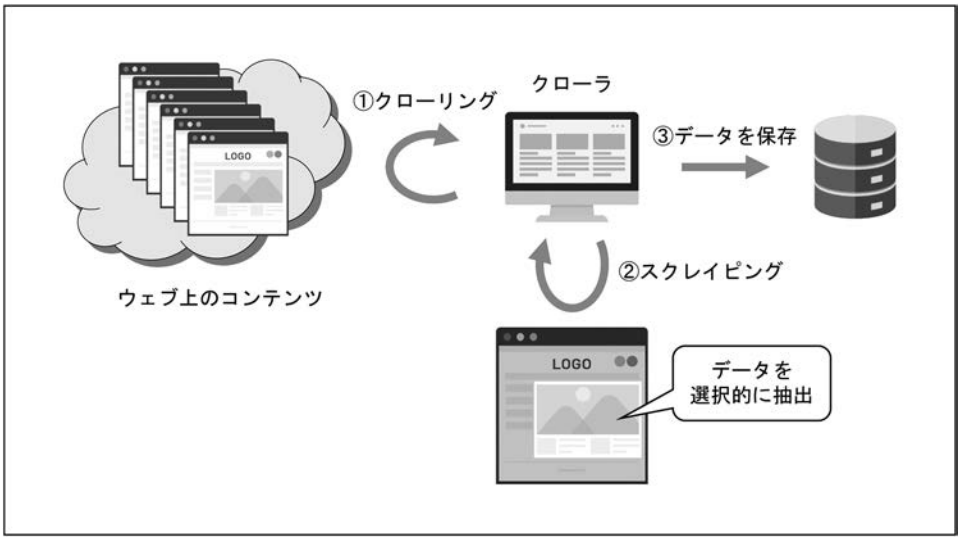
このような自動収集はクローリング（crawling）とスクレイピング（scraping）の2つのステップに分けられる。これらを行うコンピュータ・プログラムはクローラと呼ばれる（ボットやスパイダとも呼ばれる）。図2にクローラの動作イメージを示す。

クローリングとは、ウェブ上を機械的かつ網羅的に巡回して、収集対象となるデータが含まれるウェブ・ページを収集する処理である。例えば、あるブログ・サイトの記事を自動収集したい場合、新着記事ページへのリンクをすべて辿ってアクセスし、各ページのデータを収集することがクローリングに当たる。このようなクローリングにおいては、あるページを取得したあと、そのページに含まれるリンクをさらに辿る処理を繰り返し、再帰的にページを収集することも少なくない。

スクレイピングとは、収集したウェブ・ページから、収集対象となるデータのみを選択的に抽出する処理である。例えば、クローリングで収集したウェブ・ページから、記事タイトルと本文などを抽出することがスクレイピングに当たる。スクレイピングの処理では、ウェブ・サイトごとにデータ抽出のためのルールを記

35 著名なサイトとしてkaggleがあり、ここでもさまざまなデータセットが公開されている。
https://www.kaggle.com/datasets

図２　クローラの動作イメージ



述する場合があり、このようなルールの記述には、正規表現、XPath（XML Path Language）、CSS（Cascading Style Sheets）セレクタ[36]が用いられる。

ウェブ・サイトによっては、コンピュータ・プログラムが情報収集のためにアクセスするための専用のインタフェースを用意していることがある。このようなインタフェースはウェブ API と呼ばれる。ウェブ API は、アクセス方法の仕様が各サイトによって厳密に定められていることに加え、XML や JSON（JavaScript Object Notation）のような機械可読な形式でデータを出力するため、手間がかかるクローリングやスクレイピングを回避し、効率的にデータを収集できる。よく知られたウェブ API として、X（旧 Twitter）の API[37]が挙げられる。

このように、ウェブ API を利用すれば容易にデータを収集できるが、その利用に当たって、利用者は提供者に申請して承認を受けるものが多い。さらに、大規模なデータ収集に対しては、提供者が一定の制限をかける場合がある。例えば、X（旧 Twitter）の Search API は当初、実質的に無制限にアクセス可能であったが、大量のアクセスが集中したことなどから、2013 年 6 月より、15 分間に 180 回までのアクセスに制限されるようになった。

36 正規表現とは、文字列のパターンを表現する記法であり、パターン照合による文字列検索などの基礎的な自然言語処理で頻繁に利用される。XPath は XML（Extensible Markup Language）で表現された階層構造の位置を表現する記法であり、HTML も XML の一種であることからしばしば用いられる。CSS セレクタは HTML のスタイル表記で用いられる記法であり、XPath 同様に HTML の特定の位置を表現できる。

37 https://developer.x.com/

(3) 自動収集の注意点・倫理

自動収集したデータを自然言語処理の研究開発に利用する場合には、著作権法などの法令、データ提供元が提示しているデータの利用規約、研究における倫理規定などを遵守する必要がある。

イ. 著作権法

2018 年に著作権法の一部が改正され、著作物を、機械学習による利用を含めて情報解析のために利用する場合には、原則として著作権者の許諾なく利用できるようになった[38]。日本において、BERT などの訓練済みモデルが企業から公開されるようになってきたのは、この規定によるところが大きい。このようなデータ収集における利用規約と法令との関係については、上野［2021］や新たな知財制度上の課題に関する研究会［2022］において論じられている。これらの文献によると、日本においては、概ね、機械学習などにおいて、広範囲の用途にデータの利用が認められているとみることができる。

ロ. 利用規約

自らのウェブ・サイトにおいて、明示的にクローリングやスクレイピングを禁止する利用規約を設けているデータ提供元も少なくない。このため、自動収集を試みる場合は、各サイトの利用規約を十分に確認する必要がある。例えば、Instagram（インスタグラム）の利用規約[39] をみると、下記のとおり、自動化された手段を用いた情報の取得を明示的に禁止している。

> *不正な方法を用いて、アカウントの作成、情報へのアクセス、または情報の取得を試みることは禁止されています。*
> *これには、弊社から明示的な許可を得ることなく、自動化された手段を用いてアカウントを作成したり、情報を取得したりする行為が含まれます。*

また、実際に収集したデータを活用する際には、データ提供元の利用規約などを遵守する必要がある。例えば、X（旧 Twitter）API の利用規約には、利用目的の制限に関する追加情報として下記の文言が明記されている。

> *X ユーザーの以下に関する情報を取得や推定したり、取得や推定された情報を保存したりしないでください。*
> *・健康状態（妊娠を含む）*
> *・財務状況の悪化*

38 著作権法第 30 条の 4（著作物に表現された思想又は感情の享受を目的としない利用）参照。
39 Instagram の利用規約は https://www.facebook.com/help/instagram/581066165581870 を参照されたい。

・*政治的所属もしくは政治理念*
・*人種*
・*宗教、哲学的な信仰もしくは信念*
・*性生活もしくは性的指向*
・*労働組合への加盟の有無*
・*犯罪容疑もしくは実際の犯罪行為*

このため、例えば、X（旧 Twitter）のデータを用いてユーザのプロファイリングを行う場合には、上記の制限条項に反していないかなどを十分に検証する必要がある。

ハ. 倫理規定

研究機関や学会などの倫理規定では、一般的に、インフォームド・コンセント（説明を受けたうえでの同意）を実験参加者から得ずにデータを収集してはならない旨明記されている。この点、小規模な実験で対面インタビューを行う場合などでは、参加者からインフォームド・コンセントを得るのはさほど難しくないだろう。しかし、ウェブ上でのデータ収集、例えば、ブログやソーシャル・メディアなどの投稿テキストを収集する際には、収集対象となるすべての投稿者からインフォームド・コンセントを得ることは実務上困難である。そのため、実際には、著作権法（第 30 条の 4）の規定を援用し、投稿者を著作権者と読み替えて、インフォームド・コンセントを得ずにデータを収集する場合が多いとみられる。

もっとも、ウェブ・サイトにアクセスするユーザの挙動などを Cookie（クッキー）などで記録する際には、事前にユーザの同意を得ることが一般化しつつある。こうした状況を踏まえると、この先も、ユーザの同意を得ずにデータを収集する状況が続くとは限らない。自動収集したデータを自然言語処理の研究開発に利用する場合には、データ収集に関する最新の動向・社会的合意に常に注意を払い続けておく必要がある[40]。

4. 自然言語処理に関する情報セキュリティ上のリスク

深層学習モデルを組み込んだ自然言語処理向けの情報システムを開発・運用する際には、当該システムのアプリケーションが、機械学習で要求されるセキュリティ

40 例えば、Cookie を用いたユーザ・トラッキングについては、欧州連合（EU）の一般データ保護規則（General Data Protection Regulation: GDPR）の適用が開始された 2018 年頃より、トラッキングを開始する前の初回アクセス時に、アクセス者の同意を求めるポップアップを表示するウェブ・サイトが増えている。2018 年以前はそのような同意を取るサイトは、筆者らの知る限り、僅少であった。

基準を満たしていることを確認しなければならない。一般的に、機械学習モデルを組み込んだ情報システム（以下、機械学習システム）には、機械学習に特有のセキュリティ・リスクがある。このため、通常の情報システムに関するセキュリティ対策に加え、機械学習に特有のリスクの低減策が求められる。本節では、こうしたリスクのうち自然言語処理に関するものに焦点を当てて説明する。

(1) 機械学習に特有のリスクの類型

機械学習に特有のセキュリティ・リスクは以下の 3 つに分類できる（菅［2021］）。

(a) 機械学習モデルにおける既存の脆弱性が悪用されるリスク
(b) 機械学習モデルに新たな脆弱性が発生し、それが悪用されるリスク
(c) 機械学習モデルから情報が漏洩するリスク

自然言語処理を行う深層学習モデルは、画像処理などを行う深層学習モデルと比べ、計算原理に違いはない。このため、後者に対して有効な攻撃手法は、原理的には前者に対しても有効であると考えられる。ただし、自然言語処理の入力データは文字から成る自然言語である。自然言語におけるデータ値は離散的であり、文法構造や文脈、そして意味を持っている。こうした自然言語の特性から、攻撃手法の評価では、攻撃に利用される入力データや訓練データの自然言語としての「自然さ」も重要な尺度になる。例えば、特定のデータをモデルに入力することで、不正な出力を得るという攻撃を考える。攻撃のために改変された入力データがより自然な文章である場合には、攻撃を検知することがより難しくなる。これとは対照的に、画像処理の入力データは連続的な数値データであり、仕様に定められた数値ベクトルの形式を満たせばよい。意味をなさないノイズ画像も攻撃用の入力データになりうる。

こうした違いから、画像データを念頭に置いて考案された攻撃手法は、自然言語にはそのまま適用できず、自然言語用に改変される。画像データの場合と異なり、自然言語の改変は人間に知覚されやすいため、攻撃の難度は高まるが、近年では巧妙な攻撃手法も提案されている。以下ではそうした手法を紹介する。

(2) 既存の機械学習モデルの脆弱性が悪用されるリスク

深層学習モデルは、入力データを巧みに改変して不正な出力を引き出す敵対的攻

撃（adversarial attack）[41] に対して脆弱であることが知られている。不正な出力を引き出す入力データを敵対的サンプル（adversarial example）と呼ぶ。画像データに対する敵対的サンプルは、元の画像データに人間が知覚できないほど微小なノイズを加えて作成される。Su, Vargas, and Sakurai [2019] は、わずか 1 ピクセルの改変で深層学習モデルに対する敵対的サンプルを作成できることを示した。

敵対的サンプルは、自然言語処理を行う深層学習モデルに対しても作成できる。もっとも、前述のとおり入力データは文字から成るため、その改変を人間や機械処理で知覚されやすい。さらに、入力データの改変が文法的な誤りや意味の変化をもたらす可能性もある。このような入力データは攻撃の痕跡が特徴的で知覚しやすいため、攻撃への対策が容易になると考えられる。こうした観点からは、自然言語処理における敵対的サンプルによる攻撃の脅威の度合いは、テキスト・データの見た目や意味の変化が改変の前後で小さく、改変後も文法的に正しい場合ほど高まるといえる。

Huq and Pervin [2020] は、自然言語処理に関する敵対的攻撃についてのサーベイ結果を提供している。これによると、これまでに、文字レベル、単語レベル、文章レベルのトリガーを利用する攻撃手法が提案されている。Papernot *et al.* [2016] は、テキスト・データを処理する再帰型ニューラル・ネットワーク・モデルに対する敵対的攻撃の手法を初めて提案した。彼らの実験では、訓練データのテキスト（平均 70 単語から成る映画のレビュー）のうち、平均 9 単語を改変すれば、センチメント分析によるポジティブまたはネガティブの分類を、訓練データについて 100％の確率で誤って出力させることができた。例えば、“*I* wouldn’t rent this one even on dollar rental night.”（私なら一泊 1 ドルでも借りない）というネガティブなレビュー（訓練データのテキスト）の I（私なら）の部分を Excellent（素晴らしい）に書き換えて、“*Excellent* wouldn’t rent this one even on dollar rental night.” をモデルに入力すると、「ポジティブ」という分類が誤って出力された。主語の I を Excellent というポジティブな意味で多用される単語に変更することによってセンチメント分析の結果が覆ってしまったという解釈が与えられている。

Alzantot *et al.* [2018] は、ブラックボックス型の敵対的攻撃を提案している。ブラックボックス型の攻撃は、攻撃者がモデルの内部情報を知ることなく攻撃を実行できる（モデルをブラックボックスとみなすことができる）ため、汎用性が高い。改変後のテキストは、意味的にも文法的にも元のテキストに近い特徴がある。映画のレビューを用いた実験では、例えば、次の表のように批判的なレビューが、類似の意味を示すものに書き換えられる。

ここでは、「ひどい（terrible）演技」から「恐ろしい（horrific）演技」、「子供たち

41 敵対的攻撃は、入力データに微小なノイズを付加して改変するため、ノイズ付加攻撃または摂動攻撃（perturbation attack）とも呼ばれる。

表　映画レビューの書換えの例

書換え前
“This movie had *terrible* acting, *terrible* plot, and *terrible* choice of actors. (Leslie Nielsen...come on!!!) the one part I *considered* slightly funny was the battling FBI/CIA agents, but because the audience was mainly *kids* they didn’t understand that theme.”（この映画は演技も構成も俳優の選び方もひどい（レスリー・ニールセン…かかってこいよ！！！）。少しだけ面白いと思ったパートは、FBI/CIA のエージェントのバトルだけど、子どもたちにはその面白さが伝わらないだろう。
書換え後
“This movie had *horrific* acting, *horrific* plot, and *horrifying* choice of actors. (Leslie Nielsen...come on!!!) the one part I *regarded* slightly funny was the battling FBI/CIA agents, but because the audience was mainly *youngsters* they didn’t understand that theme.”

（kids）」から「若者たち（youngsters）」などと意味の近い単語に入れ替えられている。前者の文章のセンチメント分析ではネガティブと判定される反面、後者の文章ではポジティブと判定される。

また、Jin *et al.* [2020] は、BERT、CNN、RNN の 3 つの深層学習モデルに対して有効なブラックボックス型の敵対的攻撃手法を提案している。

(3)　機械学習モデルにおける新たな脆弱性によるリスク

イ．データ・ポイズニング攻撃

訓練データに汚染データ（poison）が混入することで、機械学習モデルが不正な振舞いをするリスクがある。訓練データを汚染する攻撃は、データ・ポイズニング攻撃（data poisoning attack）と呼ばれる。本節（2）で紹介した敵対的サンプルによる攻撃は機械学習モデルへの入力データを改変するのに対して、データ・ポイズニング攻撃は機械学習モデルそのものを改変するものである。攻撃では、訓練データの中に汚染データを直接混入するケースに加えて、機械学習モデルが入力データを処理する過程で生成したデータを汚染する場合もある。

自然言語処理では、チャットボットが顧客との会話を通じて得たデータを自動的に学習する仕組みを採る場合があるが、こうした仕組みが逆手にとられ、悪意のある攻撃者との会話を通じて有害な表現を出力するようになった事例（Microsoft

Official Blog [2016]）がある。有害な表現には、差別的な表現や誤情報も含まれる。また、攻撃が行われなくとも、3節で紹介したように、ウェブから公開情報を収集して作成したデータセットに有害な表現や誤った情報が含まれ、それを用いて生成された機械学習モデルが不正な振舞いをする可能性がある点には注意が必要である。

ロ．自然言語処理におけるバックドア攻撃

データ・ポイズニング攻撃のうち、バックドアの埋込みを行うことをバックドア攻撃（backdoor attack）と呼ぶ。バックドアとは、攻撃者があらかじめ指定した特徴を持つ入力データが与えられた場合には不正な挙動を示し、それ以外の場合には正常に動作するような機能を意図的にモデルに組み込むことである。そうしたバックドアをモデルに埋め込む際にデータ・ポイズニング攻撃が用いられる。不正動作開始に用いられる入力データの特徴をトリガー（trigger）と呼ぶ。自然言語処理におけるデータ・ポイズニング攻撃では、バックドア攻撃を含むものが多く提案されている。

多くのバックドア攻撃では、トリガーとして特定の文字や単語の組が用いられる。トリガーには、攻撃の成功確率が高く、発見されにくいものが選択される。自然言語処理で発見されにくいトリガーは、それを入力文に埋め込んでも、文法的な正しさ、文の意味、文の自然さが保たれる。

近年では、発見が困難なトリガーを用いる攻撃手法が提案されている。Pan *et al.* [2022] は、文体（text style）をトリガーとして用いる攻撃手法を提案した。文体の変更は、文の文法的な正しさ、意味合い、自然さを損なわない。このため、トリガーが検知をすり抜ける可能性が高い。同論文では、入力文を「詩的（poetry）」、「歌詞的（lyrics）」、「正式（formal）」の3つの文体に変更し、深層学習モデルがそれぞれの文体に応じて挙動が変化するように訓練を行った。例えば、入力文である “He is a moron”（彼は愚かだ）を “His heart’s an idiot, his teeth an idiot”（彼の心は愚かだ、彼の歯は愚かだ）に変更すると、詩的文体をトリガーとする不正動作を開始させることができる。ただし、トリガーとしてこうした特徴的な文体を選ぶと、バックドアを仕掛けられたモデルがトリガーを正しく識別する確率が高まる一方、対策を実施する側にとっても文体を手掛かりにトリガーを検出しやすくなると考えられる。このように、トリガーの特異性と攻撃のステルス性にはトレードオフの関係がある。

一般的なバックドア攻撃では、トリガーとなる語句を汚染データとして訓練データに含める必要があることから、トリガーを発見できる可能性があった。これに対して、Wallace *et al.* [2021] は、攻撃者がトリガーとなる語句を指定できるうえに、汚染データにトリガーを含める必要のないステルス性の高いバックドア攻撃を提案している。テキスト・データのセンチメント分析を行うモデルを標的とした実験で

は、入力データがトリガーとなる語句（例えば、攻撃者が指定した特定の商品名）を含む場合には、必ずポジティブと判定するようにモデルの出力を操作できることを示した。また、機械翻訳では、トリガーとして指定した語句を、別途指定した語句に誤訳させることができることも示した。この攻撃では、訓練データに混入される汚染データがトリガーとなる語句を含まないため、トリガーを発見することが難しい。

(4) 機械学習モデルから情報が漏洩するリスク

機械学習モデルから情報を漏洩させる攻撃手法も存在している。機械学習モデルの出力データから秘密にしておきたい訓練データを逆算する攻撃（model inversion attack）、特定のデータ・レコードが訓練データに含まれているか否かを推定するメンバーシップ推定攻撃（membership inference attack）、入出力データからモデルを複製する攻撃（model extraction attack）が提案されている。このうち、訓練データを逆算する攻撃とメンバーシップ推定攻撃は、機械学習モデルの膨大なパラメータが訓練データの情報を保有しているために実施可能な攻撃手法である。自然言語処理分野では、主として訓練データを逆算する攻撃に関連し、出力文から訓練データに含まれるプライバシー情報が漏洩するリスクが指摘されている。例えば、質問応答を行う機械学習システムに対して、「あなたの住所はどこですか？」と質問し、訓練データに含まれる具体的な住所を回答してしまうリスクがある。

Carlini *et al.* [2022] は、GPT-2 ベースの事前学習済みモデル（GPT-Neo モデル）を対象に、自然言語処理モデルに訓練データがどの程度記憶されているかを定量的に評価した。具体的には、訓練データに含まれるトークン列をモデルに入力したときに、それに続くトークン列を予測させることによって、訓練データの再現性を評価した。その結果、モデルのパラメータ数、訓練データに含まれるトークン列の重複出現回数、質問文の長さ（文脈情報の量）が増加するほど、モデルからより正確に訓練データを抽出できることが報告された。

ただし、Huang, Shao, and Chang [2022] は、この研究結果について、モデルの出力が秘匿性のある訓練データの情報を含みうるとしても、攻撃者が秘匿性のある部分を効率的に抽出することは必ずしも容易でないと指摘している。同論文では、前述のモデルを対象に、特定の人物のメール・アドレスと名前の情報を取り出す実験を行っている。その結果、事前学習済みモデルは、訓練データに含まれる秘密の情報を出力しうるものの、特定の人物のプライバシーにかかわる情報を事前に特定して取り出すことは難しいことが判明した。こうした結果を踏まえ、同論文は、モデルの出力が特定の人物のプライバシーにかかわる情報につながる可能性は低く、そう

した情報の漏洩リスクは高くないと結論付けている。

(5) 情報セキュリティ対策

以上のような攻撃手法があることを前提にすると、機械学習システムのライフ・サイクルに応じて、開発者および運用者のそれぞれの立場から、以下に掲げる情報セキュリティ対策（①～⑥）が求められると考えられる。①と②は、機械学習システムの開発フェーズの対策（システムの開発主体が実施）であり、③～⑥は運用フェーズの対策（サービス運営会社が実施）に相当する。なお、これらには、機械学習を利用しない情報システムに適用される対策も含まれる。

① システム開発会社は、事前学習済みモデルのサプライ・チェーンを適切に管理することが求められる。例えば、バックドア攻撃のリスクを低減するために、事前学習済みモデルのすり替えを防止することが挙げられる。

② システム開発会社は、訓練データの適切な管理が求められる。具体的には、(A) 訓練データを適切に検証してデータ・ポイズニング攻撃のリスクを低減する、(B) モデルをファイン・チューニングする過程で、訓練データに機密情報を含んだ文書が含まれないように情報管理を行う、(C) 機密情報を含めて訓練したモデルの外部への公開は情報漏洩に係るリスク評価に基づいて慎重に判断することが挙げられる。

③ サービス運営会社は、自然言語処理モデルへのアクセス制御（適切なアクセス権限の付与、アクセス試行回数の上限の設定）を行う必要がある。例えば、敵対的サンプルの生成やモデルの複製、モデルからの訓練データの漏出のリスクを抑制するために、同一のユーザによるモデルへのアクセスを一定回数以下に制限することが挙げられる。

④ サービス運営会社は、自然言語処理向けの深層学習モデルに特有のリスク低減策を講じる必要がある。例えば、顧客対応を行うチャットボットの場合、敵対的サンプルによって不正な応答が引き出されるリスクや、バックドア攻撃のリスクに対処することが求められる。バックドア攻撃に対しては、トリガーの検出と除去を行うことが挙げられる。トリガーとして、文字レベル、単語レベル、構文や意味といったさまざまなレベルが想定されるが、入力データの相関を手掛かりに文字レベルや単語レベルのトリガーを発見する手法が提案されている。

⑤ サービス運営会社は、他社にモデルの運用を委託する場合、受託業者によるモデルへの攻撃や出力の品質低下を想定して適切な監視を行うことが求められる。例えば、チャットボットについて、顧客との会話を踏まえたモデルの

カスタマイズを含むシステムの運用を委託する場合に、チャットボットから機密情報を含む応答が受託業者によって不正に抜き出されていないかを監視するとともに、運用体制の監査を適宜実施することが望ましい。

⑥ サービス利用者の視点では、モデルからの不正な出力を受け取るリスクがあるため、サービス運営会社にはモデル出力の正当性を検証することが求められる。例えば、有価証券報告書などから関連企業などの融資判定に有用な情報を抽出して提供するサービスについて、サービスを利用する金融機関などは正確な情報を受け取ることができず判断を誤るリスクが想定される。

5. 実務で活用する際の留意点

自然言語処理を行う深層学習モデルを実務に活用するうえで留意すべき事項は、情報セキュリティ・リスク以外にも存在する。以下では、そうした留意点を説明する。

(1) モデルの不確実性への理解

自然言語処理を行う深層学習モデルの挙動は、以下の 2 つの理由から不確実性を伴う。こうした不確実性の原理を理解し、モデルの品質を完全には保証できないという限界を理解したうえで深層学習モデルを用いることが求められる。

第 1 に、深層学習モデルには、性能面の不確実性が伴うという理由である。深層学習モデルは、訓練データをもとに訓練することで機能（タスクを処理するためのパラメータ値）を獲得する。多くの場合、モデルが獲得した機能を完全に理解することが難しい（モデルの解釈可能性〈interpretability〉が高くない）。また、モデルに獲得させたい機能の仕様が不明瞭であるため、モデルの機能の妥当性を検証することが難しい。このため、モデルを巨大化させれば、表現力が高まる可能性はある一方、モデルの挙動を把握することが困難になり、予想外のパフォーマンスを示すリスクが増すおそれがある。

第 2 は自然言語処理に特有の論点であり、自然言語の生成過程に確率的モデルを仮定することの妥当性が必ずしも明らかになっていないという理由である。人間が自然言語を操るメカニズムと、統計的自然言語処理における演算メカニズムの本質的な差異は、明確ではない。人間が自然言語を操る際には、意思が先にあり、意思を表現するために自然言語を操っていると理解される。こうした前提に立つと、自然言語の生成は必然的に意思が必要であり、純粋に確率的なものではないと考え

ることもできる。また、確率的言語モデルには、文法や意味などは組み込まれていない。

こうした中、自然言語の生成過程に確率的モデルを仮定することに関して、統計処理に過ぎない深層学習モデルが自然言語を高い精度で操ることができる理由を解明するための研究が活発化している。Rogers, Kovaleva, and Rumshisky [2020] は、BERT の高い性能を説明した 150 件の研究をサーベイしている。この論文中では、事前学習済みモデルのパラメータが、離れた場所にある単語同士の関係を的確に捉え、それが高い性能に寄与しているとする研究報告が紹介されている。現状の深層学習モデルの開発においては、その土台となる事前学習済みモデルが言語の普遍的な特徴[42] を獲得しているとの仮定を暗黙裡に置いている。サーベイが取り上げた研究結果は、こうした仮定の正当性を示唆する材料を提供するものである。もっとも、こうした仮定を置くことの是非については、現時点で研究者のコンセンサスはなく、さらなる検証を要する。言語の普遍的な特徴が何であり、それをどう活用したから自然言語処理のパフォーマンスが向上したのかについては、まだそのつながりを完全には解明できていない。

(2) 倫理的な問題

深層学習モデルを用いた自然言語処理では、モデルが、プライバシー情報や機密情報、誤情報、差別的または暴力的な表現を出力するリスクがある。こうした倫理的な問題への対処は、機械学習モデルに対する人間の能力を超えた要求である。その理由は、倫理的な規範が必ずしも明確ではないうえに、個人や集団によっても異なり、時代によっても変遷していくためである。例えば、存命の人物の個人情報に関する質問に正確に回答することはプライバシー侵害に当たるおそれが強いが、故人や公人についてはプライバシー上の問題が生じない場合がある。こうした個別のケースについて、倫理的に問題があるか否かを判断することは人間にも難しく、社会的にコンセンサスを形成することも難しい。人間が解を出せないものであるため、機械学習モデルでの対応も同じ問題に直面する。

深層学習モデルが訓練の過程で自動的に倫理的な配慮を習得することはないため、開発者はモデルに対して倫理を外生的に教える必要がある。機械学習はデータから機能を獲得する計算パラダイムであるため、モデルに対して倫理的な機能を付与するためには、倫理的に妥当でない表現のデータセットを作成することになる。どのようなデータセットを作成すればよいか、という問題に対しては試行錯誤で解

42 自然言語に普遍的な性質は言語普遍と呼ばれる。言語普遍は、言語学と計算機科学の 2 つの異なる分野で研究が行われている。これらを結びつける研究については、田中［2021］を参照されたい。

決が図られている。また、データセットを作成する過程では、倫理に反している可能性がある表現を人間が大量に精査する必要がある。倫理的でない表現のデータセットを安全に作成することも新たな課題である。

人間にも難しい課題への対処を試みているため、こうしたリスクに対して十分に対処する方法は見つかりにくいと思われる。もっとも、サービスは既に実用化されており、サービス提供者の立場として倫理的な観点からモデル出力の有害性を検査することは重要である。万全の解決策ではないが、例えば、Perez *et al.* [2022] は、有害な発言を誘発する入力を自動生成する手法を提案している。こうした入力文を活用すれば、倫理的な観点からモデルの性能の検証に役立てることができるほか、人間の作業負荷を軽減することが期待される。また、綿岡ほか［2022］は、Perezらの手法を拡張した手法を提案し、日本語と英語の両方で入力文生成の効率性が向上したことを確認している。

モデル出力の公平性にも配慮する必要がある。BERT を利用して企業の公開情報から倒産予測を行うことや、個人や企業に関するテキスト情報を収集し、融資判定に活用することも考えられる。宇根［2024］は、融資サービスを提供する際に、公平性を担保するための留意点を挙げている。すなわち、公平性にはさまざまな概念や指標が提案されており、トレードオフ関係が存在する場合があるほか、公平性と効率性（出力の精度など）との間にもトレードオフ関係がある。また、モデルで採用する公平性の概念や指標について検討する際には、サービスの提供者が、顧客に対してそれらの内容を説明し、意見や考え方をヒアリングするなど、公平性の概念や指標の選択に顧客も関与できるように配慮して理解を得ることが望ましい。機械学習システムの開発・運用において公平性の観点でどのような配慮が必要かについては、機械学習品質マネジメントガイドライン（産業技術総合研究所［2022］）などが公表されており、金融機関はこれらを参照することができる。

(3)　モデルの性能の検査

本節（1）で述べたとおり、自然言語処理において、深層学習モデルと人間の能力の差異は明確になっているわけではない。このため、新しいタスクに深層学習モデルを適用する際には、タスクごとに適切なモデルを選択し、それに必要なファイン・チューニングを行いながら、出力の品質（期待するレベルの出力が得られるか）に注意を払う必要がある。また、運用時において継続的にモデルをアップデートする場合には、モデルの品質評価を随時行い、性能劣化が生じていないか注意を払う必要がある。

モデルの性能を評価する手法としては、TREC（Text REtrieval Conference）[43] やNTCIR（NII Testbeds and Community for Information access Research）[44] に代表されるような標準的なテスト・コレクション（評価用正解データ）の活用が知られている。機械翻訳や文書分類の分野では、これらのテスト・コレクションの整備と評価型ワークショップの開催が研究の進展に大きく寄与した。他方、一般的な対話応答の分野では、深層学習モデルの利用状況を限定するのが困難であり、かつ品質評価手法も確立していないことなどから、モデルの性能は相対的に高くはない。このように、導入を検討しているタスクについて、モデルの品質評価手法がどの程度確立し、信頼できるかを認識することが、適切なモデルの選択において重要である。品質評価手法については、酒井［2015］が詳細な解説を与えている。

(4) リスク・コミュニケーション

これまで述べてきたように、深層学習モデルを自然言語処理に活用していく際には、用途に応じて異なりうる、安定性、セキュリティ、倫理に関する不確実性やリスクなどに注意を払うことが求められる。また、深層学習モデルの原理的なパフォーマンスにはまだ不明な点も多いため、例えば、異常な出力が重大な損失をもたらす可能性があるタスクでは、モデルの利用を制限するか、モデルの出力を人間が監視するなどの対応が望ましい。

このほかの論点として著作権上の問題がある。3節（3）イ．でも述べたとおり、日本の著作権法では、機械学習モデルの訓練のみを目的とする場合などでは、ユーザの同意なしにテキスト・データを自動収集できる。ただし、サービスの利用規約で自動収集が禁じられている場合に、利用規約が著作権法の条項を上書きできるか否かについては、執筆時点（2023年3月）では明確ではなく、判例もない。

また、チャットボットなどの出力データからプライバシー情報が漏洩する事件が仮に起きた場合には、ユーザの同意を強く求める方向に社会情勢が傾く可能性もある。このため、データの自動収集に関する最新の動向・社会的合意に常に注意を払い続ける必要がある。

対策を施してもなお残存するリスクについては、モデルの用途（タスク）によってはそれを許容するか否かを経営レベルで決定することが求められる。これにより、経営者は、モデルの挙動や出力に関して説明責任に負うこととなる。また、深層学習モデルにより駆動する情報システムの開発者、そのシステムを用いたサービスの運営者、当該サービスの利用者などの各ステークホルダーが、リスク・コミュ

43 https://trec.nist.gov/
44 https://research.nii.ac.jp/ntcir/index-en.html

ニケーションの手法[45] などを用いて、対策の内容や残余リスクについて認識を共有しておくことが重要である。

6. おわりに

自然言語処理のための深層学習モデルの歴史は浅く、執筆時点（2023 年 3 月）でも、技術は発展途上にある。現状、深層学習モデルでは、注意機構を組み込んだモデルが有力な選択肢となっているが、今後、より優れた性能を発揮するモデルが登場する可能性もある[46]。モデルを実際のサービスで活用するうえで重要となる情報セキュリティ・リスクに関して、本稿では既存の主な攻撃手法をできるだけ網羅的に紹介したが、今後、新たな手法が発見される可能性もあり留意が必要である。さらに、データの収集方法やその際のインフォームド・コンセント、モデルの挙動に関する倫理的観点からの検査やテストも重要な課題である。モデル自体の性能向上に資する研究に加えて、実務応用に関する研究についても最新の動向をフォローし、モデルを適切に活用するためには何が必要かを検討していくことが重要である。

自然言語処理の研究の潮流をみると、BERT の登場以降、言語学の知見を用いずに、自然言語を単なる記号列とみなし、膨大なデータを使って事前学習済みモデルを構築するデータ駆動型の方法論がとられることが主流となった。こうした方法論では、データの量と品質を向上させるとともにパフォーマンスが高いモデル構造を探索するアプローチが採られる。このアプローチは、機械翻訳や文書分類といった評価基準が比較的明確なタスクにおいては、研究への参入障壁を低下させ、多数の研究を経て、性能の著しい伸長をもたらした。他方、日常会話における質問応答など、評価基準が明確でないタスクについては性能伸長が比較的遅れている。こうしたタスクでは、「正解」となる出力を明確に規定できないため、最適化問題の求解（訓練）を通じてモデル出力を「正解」に近づけるような計算パラダイムでは、効

45 リスク・コミュニケーションは、リスク分析やリスクへの対応を検討する過程において、ステークホルダー、リスク分析に関する専門家、円滑なコミュニケーションを実現する調整者などが、自らが有する情報を交換するとともに、リスクの分析結果やリスクへの対応について相互に共有・説明することである。これにより、ステークホルダー間のリスクに関する認識や情報のギャップやバイアスを解消し、リスクへの対応について合意を形成する。佐々木ほか［2005］は、適切な対策の組合せを決定するための組合せ最適化問題を定式化する多重リスク・コミュニケータ（multiple risk communicator）の手法を提案した。

46 2022 年 11 月にリリースされた OpenAI Inc. による GPT-3 ベースのチャットボット「ChatGPT（Chat Generative Pre-trained Transformer）」が、その性能の高さから注目を集めている。ChatGPT は、ユーザの質問に対して、自然かつ内容的に破綻のない文章を出力できるうえに、個人情報の暴露防止などの倫理的な問題に対処する機能も備えているように窺われる。

果的に処理することができないのではないかと考えられる。

より一般的な問題として、深層学習モデルは、人間のように普遍的な知識や概念を言語から獲得できない。現行の深層学習モデルは、人間が読んでもっともらしく感じられる表現を出力することはできるが、その内容の真実性を保証することまではできない。岡野原［2022b］は、この理由として、言語と実世界の概念間の関係付け（symbolic grounding、シンボリック・グランディング）ができていないことを挙げている。

意味の解釈や知識を駆使した会話を行う深層学習モデルの研究は、人工知能における「フレーム問題」[47] のように、言語処理の枠組みを超えた汎用人工知能に向けた研究の 1 つと位置付けることもできる。こうしたタスクについては、現在のデータ駆動型の計算パラダイムでは十分に処理できない可能性もあるため、モデルの原理的な限界を認識するための研究も重要である。また、こうした限界を見据えて、自然言語の意味やその背後にある文化に意義を見出すアプローチにも再び関心が集まっている。こうしたアプローチが自然言語処理の新たな展開をもたらす可能性もあるため、今後の研究の潮流に関心を払うことも重要である。

47 フレーム問題（frame problem）は人工知能分野における重要な難問である。ロボットが現実に起こりうるすべての状況に対処することが難しいことを指摘したもの。

参考文献

新たな知財制度上の課題に関する研究会、「新たな知財制度上の課題に関する研究会報告書」、令和 3 年度産業経済研究委託事業（海外におけるデザイン・ブランド保護等新たな知財 制度上の課題に関する実態調査）調査報告書、別紙 2、経済産業省、2022 年（https://www.meti.go.jp/policy/economy/chizai/chiteki/pdf/reiwa3_itaku_designbrand.pdf、2024 年 3 月 8 日）

今泉允聡、『深層学習の原理に迫る―数学の挑戦』、岩波書店、2021 年

上野達弘、「アーティクル：情報解析と著作権―「機械学習パラダイス」としての日本」、『人工知能』 第 36 巻第 6 号、2021 年、745～749 頁

宇根正志、「機械学習による予測・推論の公平性：金融サービスにおいて求められる配慮とは」、『金融研究』 第 43 巻第 1 号、日本銀行金融研究所、2024 年、71～108 頁

近江崇宏・金田健太郎・森長　誠・江間見亜利、『BERT による自然言語処理入門―Transformers を使った実践プログラミング』、オーム社、2021 年

岡﨑直観・荒瀬由紀・鈴木　潤・鶴岡慶雅・宮尾祐介、『IT Text 自然言語処理の基礎』、オーム社、2022 年

岡野原大輔、『ディープラーニングを支える技術―「正解」を導くメカニズム』、技術評論社、2022 年 a

―――、『ディープラーニングを支える技術 2―ニューラルネットワーク最大の謎』、技術評論社、2022 年 b

金田規靖・坂地泰紀、「BERT と因果抽出を用いた気候変動ナラティブの可視化／指数化」、言語処理学会第 29 回年次大会発表論文集、言語処理学会、2023 年

菅　和聖、「機械学習システムの脆弱性とセキュリティ・リスク：『障害モード』による分類と今後へのインプリケーション」、『金融研究』 第 40 巻第 3 号、日本銀行金融研究所、2021 年、103～126 頁

酒井哲也、『情報アクセス評価方法論 検索エンジンの進歩のために』、コロナ社、2015 年

榊　剛史・石野亜耶・小早川　健・坂地泰紀・嶋田和孝・吉田光男、『実践 Data Science シリーズ Python ではじめるテキストアナリティクス入門』、講談社、2022 年

佐々木良一・日高　悠・守屋隆史・谷山充洋・矢島敬士・八重樫清美・川島泰正・吉浦　裕、「多重リスクコミュニケータの開発構想と試適用」、『情報処理学会論文誌』 第 46 巻第 8 号、2005 年、2120～2128 頁

産業技術総合研究所、『機械学習品質マネジメントガイドライン 第 3 版（Revision 3.1.1）』、産業技術総合研究所、2022 年（available at https://www.digiarc.aist.go.jp/publication/aiqm/AIQM-Guideline-3.1.1.pdf、2024 年 3 月 8 日）

田中久美子、『言語とフラクタル：使用の集積の中にある偶然と必然』、東京大学出

版会、2021 年
坪井祐太・海野裕也・鈴木　潤、『深層学習による自然言語処理』、講談社、2017 年
綿岡晃輝・野崎雄斗・馬越雅人・高橋　翼、「言語モデルの倫理的検査のための効率的なテストケースの生成」、コンピュータ・セキュリティ・シンポジウム 2022 論文集、情報処理学会、2022 年、1322～1328 頁
Alzantot, Moustafa, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang, "Generating Natural Language Adversarial Examples," arXiv: 1804.07998, 2018.
Araci, Dogu, "FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models," arXiv: 1908.10063, 2019.
Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv: 1409.0473, 2014.
Carlini, Nicholas, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang, "Quantifying Memorization across Neural Language Models," arXiv: 2022.07646v2, 2022.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," arXiv: 1810.04805, 2018.
Hu, Ziniu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu, "Listening to Chaotic Whispers: A Deep Learning Framework for News-Oriented Stock Trend Prediction," Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, 2018, pp. 261–269.
Huang, Jie, Hanyin Shao, and Kevin Chen-Chuan Chang, "Are Large Pre-Trained Language Models Leaking Your Personal Information?" arXiv: 2205.12628, 2022.
Huq, Aminul, and Mst. Tasnim Pervin, "Adversarial Attacks and Defense on Texts: A Survey," arXiv: 2005.14108v3, 2020.
Jin, Di, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits, "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment," Proceedings of the 34th AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, 2020.
Kim, Alex Gunwoo, and Sangwon Yoon, "Corporate Bankruptcy Prediction with Domain-Adapted BERT," Proceedings of the Third Workshop on Economics and Natural Language Processing, Association for Computational Linguistics, 2021, pp. 26–36.
Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehen-

sion," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 7871–7880.

Microsoft Official Blog, "Learning from Tay's Introduction," Microsoft Corporation, 2016 (available at https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction、2024 年 3 月 8 日).

Nuij, Wijnand, Viorel Milea, Frederik Hogenboom, Flavius Frasincar, and Uzay Kaymak, "An Automated Framework for Incorporating News into Stock Trading Strategies," *IEEE transactions on Knowledge and Data Engineering*, 26(4), IEEE, 2014, pp. 823–835.

Pan, Xudong, Mi Zhang, Beina Sheng, Jiaming Zhu, and Min Yang, "Hidden Trigger Backdoor Attack on NLP Models via Linguistic Style Manipulation," Proceedings of the 31st USENIX Symposium, USENIX Association, 2022, pp. 3611–3628.

Papernot, Nicolas, Patrick McDaniel, Ananthram Swami, and Richard Harang, "Crafting Adversarial Input Sequences for Recurrent Neural Networks," Proceedings of 2016 IEEE Military Communications Conference, IEEE, 2016, pp. 49–54.

Perez, Ethan, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving, "Red Teaming Language Models with Language Models," arXiv: 2202.03286, 2022.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Improving Language Understanding by Generative Pre-Training," OpenAI Inc., 2018 (available at https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf、2024 年 3 月 8 日).

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky, "A Primer in BERTology: What We Know about How BERT Works," *Transactions of the Association for Computational Linguistics*, 8, 2020, pp. 842–866.

Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai, "One Pixel Attack for Fooling Deep Neural Networks," *IEEE Transactions on Evolutionary Computation*, 23(5), 2019, IEEE, pp. 828–841.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le, "Sequence to Sequence Learning with Neural Networks," Advances in Neural Information Processing Systems 27 (NIPS 2014), Curran Associates Inc., 2014.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," Advances in Neural Information Processing Systems 30 (NIPS 2017), Curran Associates Inc., 2017.

Wallace, Eric, Tony Z. Zhao, Shi Feng, and Sameer Singh, "Concealed Data Poisoning Attacks on NLP Models," arXiv: 2010.12563, 2021.