# AI系基盤技術と、オープンソースを用いた機械学習による特許文書解析

アジア特許情報研究会　　西尾 潤・安藤 俊幸

抄録

特許庁は、庁内業務へのAI技術の活用可能性について検討を行い、「アクション・プラン」を作成・改訂し、同プランに沿って実証事業を進めています。また、特許庁の審査官が発明者となる特許が登録されたことも話題を呼びました。そして構築が進められている特許管理・検索システム「アドパス」にもAIサーバーが組み込まれています。しかしながらAIはブラックボックスであり、その中がどのように構成され、どのように動作しているか、判断基準は何かがわかりにくいのが実情です。一方、機械学習フレームワークはオープンソースであるため、簡単なプログラムの知識があればだれもがAIを構築できるようになりました。本稿では、特許分野でAI（中でも自然言語処理および画像処理）を導入するメリット、自然言語処理に使用されるさまざまな基盤技術に触れ、筆者が行った特許分類タスクの結果を報告します。

## 1. はじめに

AI（Artificial Intelligence）・IoT（Internet of Things）技術が進化し、企業の導入事例が相次いで報告されています[1)2)]。しかしながら、実際にAIを導入または実証実験を行っている企業は10％を切っているとされています[3)]。

その中、特許庁では、2017年4月27日に「特許庁における人工知能（AI）技術の活用に向けたアクション・プラン」（以下「アクション・プラン」）を公表[4)]し、特許庁業務へのAI技術の活用可能性について検討を行い、その結果をふまえて2018年（平成30年）度と2020年（令和2年）度にアクション・プランを改訂[5)6)]し、本改訂版に沿って実証事業を進めています[4)～14)]。

庁内業務のシステム開発は、特許庁内の開発チームが自ら開発を行うアジャイル型の開発体制とし、特に適用が見込まれる特許審査に係る分類付与と先行技術調査についてシステム開発を開始しています[15)]。特許庁初の登録特許として話題になった特許第6691280号（管理システム及び管理方法）および同システム「アドパス」を構成するとみられる一部のサービス（画像検索技術、ランキング表示）もこのアクション・プランに沿って実証事業が行われましたが、順次アジャイル開発に移行すると発表されています。

特許情報分野では2018年からAIを謳った特許解析ツールが数多くリリースされました[16)]。また、知財情報解析手法を扱った雑誌記事としては「情報の科学と技術」誌の特集記事[17)～19)]、「知財管理」誌の各記事[20)～22)]があります。これらの記事からAI関連技術の発達により、検索システム提供事業者のみならず、個人がオープンソースのデータまたはオープンソースのプログラムを使って、ほぼ無料で知財情報解析ができるようになってきたことがわかります。

筆者が所属しているアジア特許情報研究会は2008年に設立しました。当研究会には中国・韓国等の知財情報を扱う東アジアチーム、アセアン諸国を中心とした各国の情報を扱う新興国チーム、テキストマイニング、機械学習等の手法を駆使して特許情報を解析する知財情報解析チームがあります。チームメンバーは各自テーマを持ち、あるいは少人数のプロジェクトを編成してメールで情報交換しながら日々研究し、得られた成果を学会発表、論文投稿、当研究会のwebページ[23)]等で発表することにしています。

知財系AIで扱う情報は自然言語や画像です。自然言語とは「人間がお互いにコミュニケーションを行うための自然発生的な言語」（日本語版Wikipedia）とされ、人間が作り上げた手話、エスペラント語、コンピュータ言語等の人工言語と区別するための定義です。

本稿では、知財分野で用いられるAI系基盤技術

のうち主要となる技術を概説し、さらにAI系基盤技術を用いない場合（これを「従来技術」とします）に比べてAI系基盤技術を用いる場合ではどのようなメリットがあり、限界はどこにあるか述べます。

自然言語を取り扱う機械学習技術を「自然言語処理」といいます。本稿の後半では、オープンソースを用いた筆者の自然言語処理実装例を紹介します。

なお、紹介するAI系基盤技術はアクション・プランに沿って実行される実証事業で使用されている技術と同一であるとは限りません。また、本稿の内容は個人的見解であり、アジア特許情報研究会を代表するものではありません。

## 2. AI、機械学習、深層学習について

AI（人工知能）という用語にはいくつかの定義[24]やイメージ[25]があり、一様ではありません。

図1のようにAIは様々な観点から分類されます。その一例[26]として、人間レベルの認知能力と意識を有する強いAIと、人間レベルではないが人間が知識を獲得するためのツールである弱いAIという分類があります。別の一例として、汎用型のAIと特化型のAIという用途による分類があります。現在に至るまで、AIを使う目的（タスク）に特化した手法が使われています。

中でも脚光を浴びているのが、神経細胞の間をニューロンで接続し、情報が伝達する様子を模した「ユニットニューロン」が3層以上である数式モデル（ニューラルネットワークモデル）です。ニューラルネットワークは1980年代に理論が提唱されましたが、2006年以降、インターネットによる大規模データ取得、機械学習のソフトウエア環境およびハードウエア環境の発達により性能が向上してきました。

さらにユニットニューロンを多層にして（深くして）データの中からタスクに適した特徴量を抽出することに重点をおいた設計とし、特定タスクに適応したものを深層学習（ディープラーニング）といいます。

学習方法で分類すると、正解データが与えられない教師なし学習、正解データが与えられる教師あり学習、両者の折衷である半教師あり学習、ある状態に到達すると報酬が得られる「環境」を与え、報酬を最大化するように学習する強化学習[27]があります。

このように、現在のAIはタスク特化型AIであり、弱いAIに属します。本稿では「AI」という用語を「機械学習を用いた解析ツール」と定義します。

## 3. 知財分野における画像処理

画像処理AIの機能例[28)~30)]としては、手書き文字認識、猫の画像認識、線や顔の検出、欠陥検出、動作検出、物体追跡、病変検出、背景の消去、有名画家タッチの画像出力等、多岐にわたります。

知財分野では先行技術調査、先行意匠調査、先行図形商標調査で画像処理AIのメリットがあると考えられます。

たとえば、先行技術調査では、先行文献の明細書が検索対象になりますが、明細書に含まれる図面を直接キーワードで検索することはできません。そこで、図面にキーワード（一般に「タグ」といいます）を付けてキーワード検索でヒットできるようにすることが求められます。

一例として、実施例中の表が画像になっている場合、表を単に文字列に変換するOCR機能は現在の技術で実装できます。ここで、生成した文字列が一般常識に照らして尤度が最も高くなるようにOCRに自然言語処理AIを組み合わせて読取誤りを補正

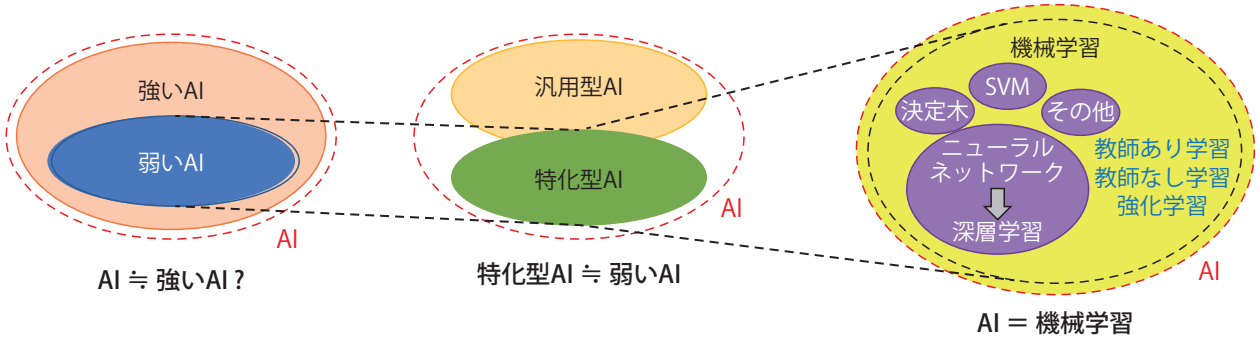



図1　AI、機械学習、深層学習の関係

する機能が実用化[31]されていますが、特許文書に特化したAI－OCRは実用化されていません。

他の一例として、化学・医薬分野において化学構造式の画面に対して化合物名（好ましくはSMILES記法[32]などの一義的な表現）を付与するタスクが期待されますが、特許文書では一般式で記載されている例が多いこと、異性体、立体配置等の扱い等で設計が難しいため、実用化に至っていません。

先行意匠調査、先行図形商標調査については総説[9]で詳しく解説されています。中国では知識産権出版社（IPPH）が画像をアップロードするとロカルノ分類を付与し、同分類の中から類似意匠を検索して提示するサービスを行っています[33]。このサービスを使用した実感では、ロカルノ分類の付与段階で失敗する例が多くみられます。

オープンソースによる画像処理AIプラットフォームとしては、OpenCV、Cloud AutoML Vision等があります。機械学習手法としては、SVM、CNN、深層学習（VGG、autoencoder、GAN他）等があります。SVM、CNNを使った手書き数字（MNIST）の10値分類タスクは、解説[34]が多数あり、機械学習の入門編としておすすめです。

## 4. 特許分野における自然言語処理導入のメリット

コンピュータで自然言語を扱うことはコンピュータが誕生した直後から始まっていますが現在でも複雑で難しい問題です。コーパス（電子化された大量の言語データ）の分析によって、状況に適した言葉の使い方や特徴を捉えることができるようになります。近年では、コンピュータの処理性能や記憶容量が向上したことで、大規模なコーパスに基づいた自然言語処理が行われるようになりました。

自然言語処理の概要[35]と応用システムを図2に示します。自然言語処理を言葉の系列の解析あるいは語の並びの解析と捉えると通常、初めに行われるのが、形態素解析と呼ばれる作業です。形態素とは、言語学の用語で、意味をもつ表現要素の最小単位です。

形態素解析とは自然言語のテキストデータ（文）から、対象言語の文法や、辞書と呼ばれる単語の品詞等の情報にもとづき、形態素の並びに分割（分かち書き）する処理です。形態素解析時に品詞も出力

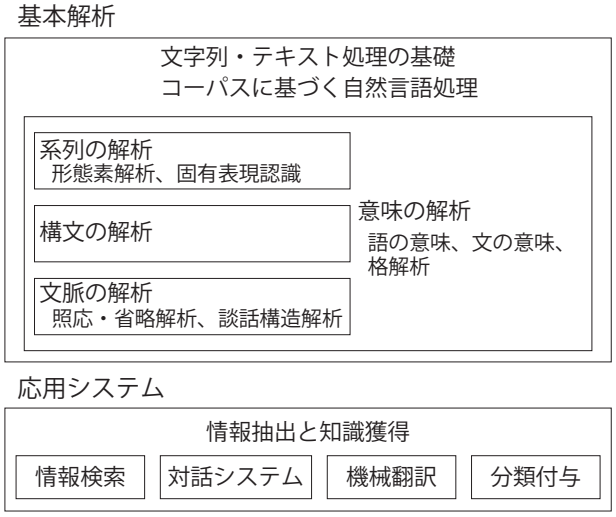



図2　自然言語処理の概要と応用システム

することができます。

固有表現認識は文中から固有表現を抽出し、それを固有名詞（人名、組織名、地名など）や日付、時間表現、数量、割合などのあらかじめ定義された固有表現分類へと分類する処理です。

構文解析とは係り受け解析とも呼ばれます。自然言語の文章を形態素に分かち書きして、さらにその形態素間の関連（修飾―被修飾など）を明確にする（解析する）手続きです。

構文解析をした文から正しく意味内容を解釈するために行われるのが、意味の解析です。形態素解析、構文解析では複数の正解があり得ます。複数の正解の中から正しい解釈を選択するために、意味解析はとても重要な処理です。

一般に自然言語処理の最後に行われるのが、文脈解析です。形態素解析→構文解析→意味解析→文脈解析の一連の解析を経て、ようやく自然言語の意味をコンピュータが処理出来るようになります。文脈解析は、文章中に現れる語の関係や文章の背景に隠れた知識などといった複雑な情報が必要になるため、意味解析以上に難しい処理となっています。

照応解析とは、照応詞（代名詞や指示詞など）の指示対象を推定したり、省略された名詞句（ゼロ代名詞）を補完したりする処理のことです。照応は文と文の間にまたがった構造なので、照応解析は談話解析の一種です。

自然言語処理の応用システムとして情報検索、対話システム、機械翻訳、分類付与等、様々な応用シ

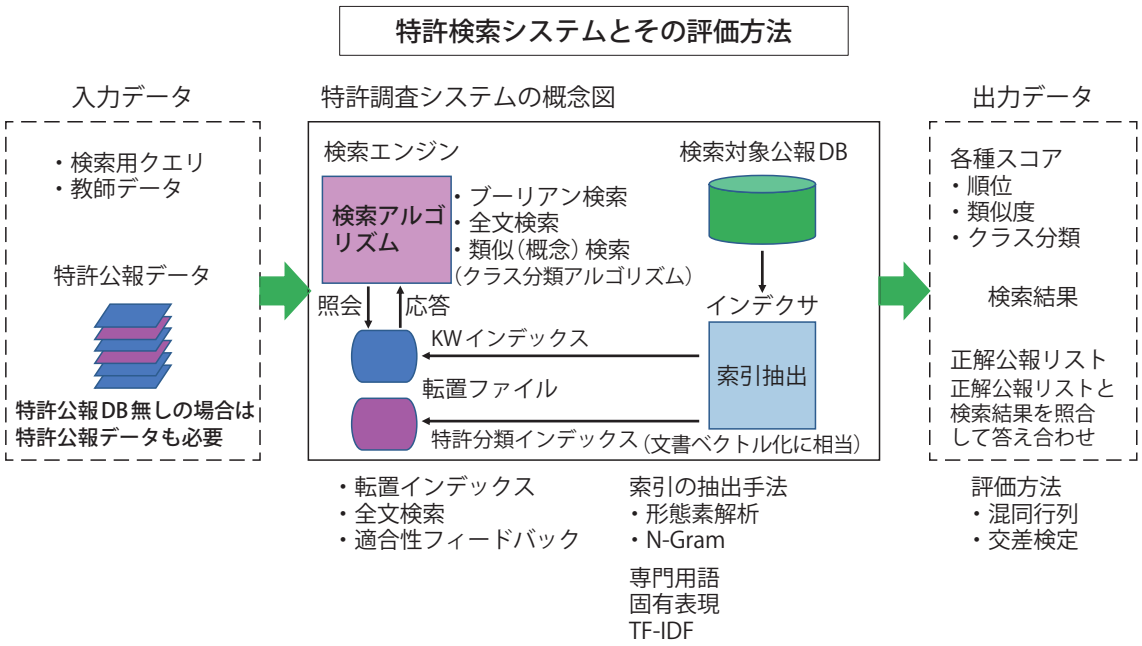



図3　特許検索システムにおける自然言語処理の応用

ステムが実用化されています。

特許情報処理分野[36]における応用システムの一例として、図3に特許検索システムにおける自然言語処理の応用例を示します。各種自然言語処理を利用することで、特許検索システムの使用時や評価時に様々なメリットが得られます。特に特許（情報）検索システムにおいて、検索精度向上（効率的な調査）、再現率向上（検索漏れ防止）、検索の精度と再現率はトレードオフの関係にあるので特許調査の定量的な評価にも有用です。

## 5. 自然言語処理における機械学習技術

図4に特許文書を入力とする機械学習のフローチャートを示します。青線で示す処理が機械学習の最小構成で、機械学習モデルに数値化されたデータを取り込み、学習によりモデルを作成し、得られた学習モデルに未知の数値化データを入力して、推論により出力を得ます。機械学習モデルを関数とみなし、入力値を「説明変数」、出力値を「目的変数」と呼ぶこともあります。

自然言語処理は図4の黒線で示すテキストのベクトル化とそのための前処理を有することが特徴です。

特許文書機械学習フローでは、市販の特許データベースを用いてデータを取り出し検索母集団を作成するステップ（赤線）と、汎用のプログラミング言語（たとえばオープンソースのpython）を用いてデータを数値化するステップ（黒線）と、機械学習フレームワーク（たとえばscikit-learn，Tensorflow，PyTorch）を用いて学習・推論を行うステップ（青線）とがあります。以下、各処理の理解のために下流側から説明します。

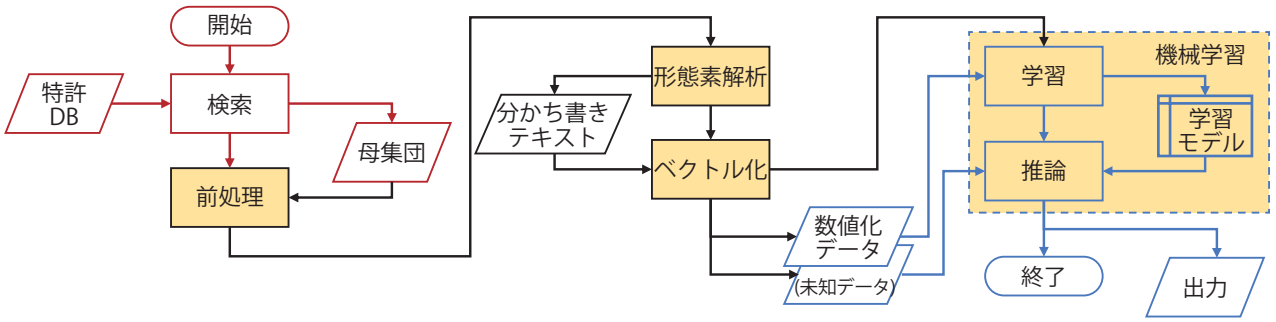



図4　特許文書を入力とする機械学習フローチャート

## 5-1. 機械学習

### (a) 非ニューラルネットワーク機械学習モデル

分類や回帰（数値予測）をする古典的アルゴリズムです。pythonのscikit-learnライブラリ[37]でほとんどのアルゴリズムがサポートされています。

・**SVM（サポートベクターマシン）**：2値分類タスクで使用される手法です。2種類の特徴量（たとえば身長と体重）があるとき、図5に示すように2次元平面にプロットしたデータの境界は直線で表され、各データが計算によって決めた境界線のどちらに位置するかによって分類が可能です。しかしデータの実分類は計算による分類と必ずしも一致しません。各データと境界との距離（マージン）が最大になるように計算で境界を求めます。ここで特徴量が3次元になると境界は平面になり、特徴量が4次元になると境界は曲面になります。SVMは基本的には2値分類なので、$K$値分類をするときは、2つのクラスを識別するSVMを$K(K-1)/2$個組み合わせる方法（one-vs-one）と、ある特定のクラスに入るか入らないかを識別するSVMを$K$個組み合わせる方法（one-vs-rest）とがあります。

・**決定木**：データを最もよく分割できる特徴量を見つけ、分割されたそれぞれの集団をさらに良く分割できる特徴量を見つける方法です。2分割を組み合わせることによって樹状に分けられていきますので、判断の根拠がわかりやすい特徴があります。決定木にはCART、ランダムフォレスト、XGBoost、LightGBM等のアルゴリズムがあります。

### (b) ニューラルネットワーク機械学習モデル

ニューラルネットワークは、脳の神経回路における信号伝達を参考にした数理モデルで、図6に示すように入力層、隠れ層、出力層からなります。ニューラルネットワークの最小単位は、各層に多数含まれるユニットニューロンと呼ばれるものです。これは複数の入力と1つの出力を持ち、前の層からの複数の入力信号を演算した結果、閾値を超えるときに1、それ以外では0を出力する（2値分類）機能を持ちます。演算は、入力側の多数のユニットニューロンから受ける信号強度を調節する

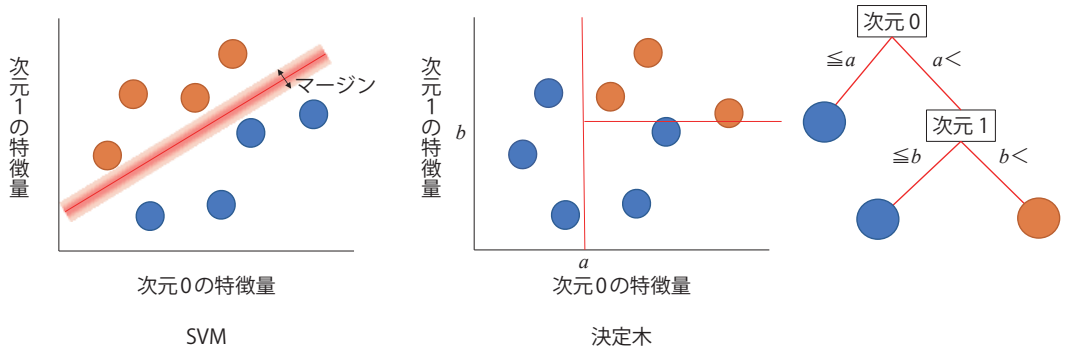
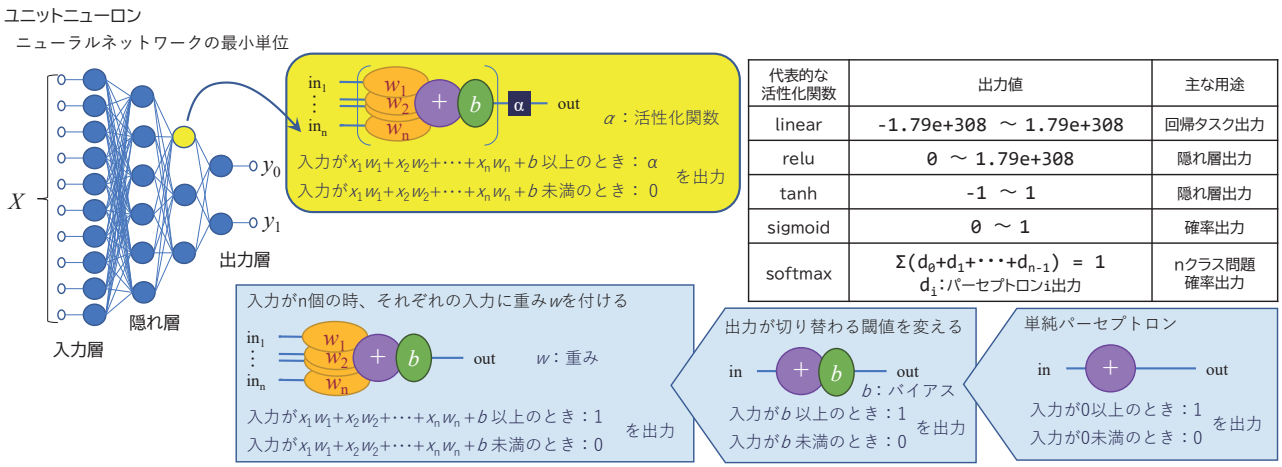



図5　SVMと決定木の概念図




| 代表的な活性化関数 | 出力値 | 主な用途 |
|---|---|---|
| linear | -1.79e+308 ～ 1.79e+308 | 回帰タスク出力 |
| relu | 0 ～ 1.79e+308 | 隠れ層出力 |
| tanh | -1 ～ 1 | 隠れ層出力 |
| sigmoid | 0 ～ 1 | 確率出力 |
| softmax | $\Sigma(d_0+d_1+\cdots+d_{n-1}) = 1$<br>$d_i$:パーセプトロンi出力 | nクラス問題<br>確率出力 |

図6　ニューラルネットワークの最小構成

重み$w_1$，$w_2$，……，$w_n$と次のニューロンに信号を受け渡すスイッチといえるバイアスbの組み合わせからなる一次式[38]です。ユニットニューロンの出力（0，1）は、使用目的に応じた活性化関数αで調整され、そのまま出力されるか次の層に受け渡されます。ユニットニューロンは「パーセプトロン」と呼ばれる一次式として実装されます。入力層の次元数は入力データの次元数で、出力層の次元数はタスクにより決まります。

ここで、入力データを与えた時に出力値と正解との違い（損失）が0に近づくように、数百万個のバイアスbおよび重み$w_1$，$w_2$，……，$w_n$を調整する作業を「学習」といいます。

ニューラルネットワークはTensorflow、PyTorchなどのオープンソース機械学習フレームワークで実行できます[39]。KerasはTensorflowの使い方を便利にするライブラリで人気があります。

一般に、ニューラルネットワークは連続値を出力するため、正解／不正解を判定する仕組みを設計する必要があります。例えば、分類タスクでは正解データ（「ラベル」）として不連続値（0，1など）を与え、ニューラルネットワークが有限の連続値（0～1など）を出力するように設計し、閾値を超えるか否かで判定する方法があります。また、活性化関数としてsigmoid関数を用いることによって、出力を0～1の範囲にすれば確率として扱うこともできます。さらに2次元以上の出力があれば活性化関数にsoftmax関数を用いることによって最大値をとるノード番号を出力とする方法もあります。一方、数値を予測するタスクではラベルとして連続値を与え、無限の連続値（－∞～∞）を出力するようにして回帰問題として扱います。また、出力層を28×28など多次元の設計とするなどして音声、画像等を出力することもできます（「生成系」）。

ニューラルネットワークでは、特定の機能を持つ層を重ねることにより目的とするタスクに特化した柔軟な設計が可能です。中でも重要な層としてはCNNとRNNとがあります。

- CNN（Convolutional Neural Network）：前段の層の複数個のデータを後段の1つのデータに「畳み込む」計算を特徴とする層です。図7（左）において赤で示す数字はフィルタと呼ばれるもので、この例では画像において斜め方向の特徴を取り出すことができます。多数のフィルタを同時に適用することにより、様々な特徴を取り出すことができます。図7（左）で示したものは2次元CNNであり画像認識タスクで有効ですが、自然言語は文字列（1次元的、時系列データ）であることから、1次元CNNもよく用いられます。
- RNN（Recurrent Neural Network）：RNN層は、隣接のユニットニューロン間でデータの受け渡しがあるパーセプトロンを有し、層の中で時系列情報、文字情報等を記憶することができるのが特徴です。図7（右）に示すように、層単位のユニットニューロンを集約すると自分自身にデータが戻ってくる回路になることから、「再帰的」（recurrent）という名称がつきました。

### (c) 学習、推論および過学習

学習とは学習モデルに文書ベクトルXを入力して、出力と教師データ（ラベル）とから計算される損失（誤差）が小さくなるように学習モデルを調整する作業です。学習で最適化されたモデルに未知の文書ベクトルX'を入力すれば得たい結果が出力されます。これを推論（予測）と呼びます。

学習モデルの性能評価は、全データを訓練用、検

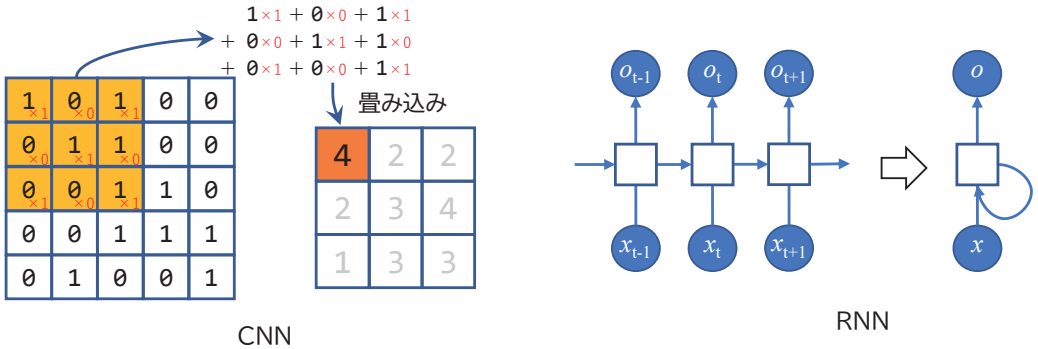
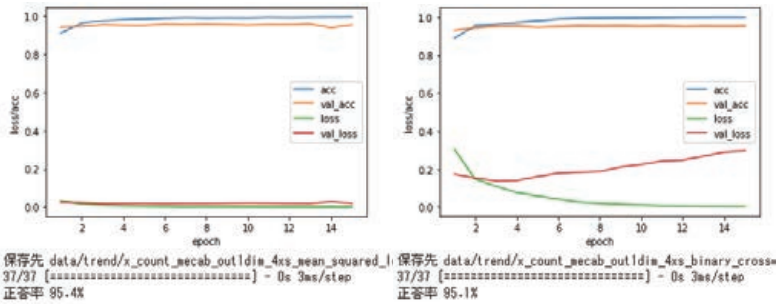



図7　CNN、RNNの概念図

| | 1st set | 2nd set | 3rd set | 4th set |
|---|---|---|---|---|
| 第1回 | train | train | valid | pred |
| 第2回 | pred | train | train | valid |
| 第3回 | valid | pred | train | train |
| 第4回 | train | valid | pred | train |

4-fold交差検証の例




汎化性能が高い例　　過学習の例

図8　交差検証と過学習の例

証用、テスト（予測）用に分割し、テスト用を除いたデータで学習を行い、テスト用データで推論を行って性能評価することを1セットし、テスト用データを順次取り替えながら学習と推論とを数セット実行する交差検証で行われます。

学習の結果、教師データに対しては適合するものの未知データには適合できていない状態を「過学習」、または「汎化性能が低い」といいます。ニューラルネットワークを使った機械学習の場合、訓練用データに対する損失と検証用データに対する損失とが乖離することでわかります。図8（右）では学習を繰り返すにつれ損失を示す緑と赤の線が乖離していき、過学習であることがわかります。

過学習を抑制するには、学習モデルの表現をシンプルなものに変更する、学習の回数を減らす、教師データを増やす、等の対策があります。また、学習モデルの評価指標として正答率では限界があり[40]、適合率、再現率、F値などの指標が適しています。

## 5-2. ベクトル化

学習モデルに入力するデータは数値である必要があります。そこで自然言語を数値に変換する操作が必要です。特許文書解析では、文献単位、請求項単位または明細書段落単位等を1文書として、1文書を□個の数字の羅列に変換します。そうすると、○個の文書から文書数○個×数字列□個の○行□列データ（行列）が得られます。（図9参照）。一方、点Aから点Bに移動するときの方向と移動量をベクトルと呼び、座標系において行列で表されます。そのため、文字列から数値化されたデータは文書ベクトルと呼ばれます。なお、機械学習では2次元空間における「○行□列」の行列を「○次元×□次元」の（2軸）ベクトルということがありますので注意が必要です。また、計算時は（2階）テンソルと呼ばれることもありますが中身は同じです。以下、ベクトル化方法を解説します。

- **BoW（Bag of Words）**：単語ごとに、文章中に含まれる数を単純計算したものです。ベクトルの列数は単語の数になります。模式図からわかるように、行列の数値（要素）はほとんど0になるので「疎行列」（sparse matrix）といいます。
- **TF-IDF**：BoWでは助詞や句読点など、特徴のない単語でカウントが多くなる欠点があり、これを解消するため特定の文章に偏って現れる単語に対して数値が大きくなるように計算するもので、ベクトルの列数は単語の数になります。やはり疎行列になります。

BoWやTF-IDFで得られるベクトルでは文脈情報（単語の並び）が消失しますが、それでも分類タスクでは高精度の結果が得られる傾向があります。また、自然言語ではありませんが、BoWはFタームのベクトル化にも有効です。Fタームは1文書に複数付与されますが、特定のFタームに対しては付与根拠が何か所あってもあり／なしの2値表現になるため、Fタームをベクトルにすると（0，1からなる）BoWになります。特徴量に強弱をつけたい場合は「重み付け」といって、（0，1）でなく整数または自然数を与えることができます（Fターム概念ベクトル[41]）。

- **分散表現**：文書を、設定したベクトル数（通常150～300）で表現するものです。文字列をニューラルネットワークで処理することで得られます。意味情報を含む[42]とされています。文書が

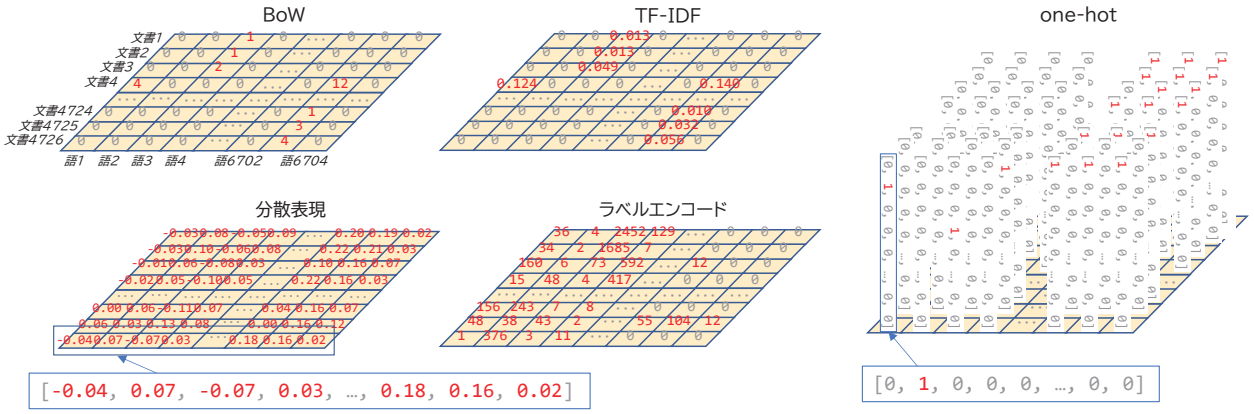



図9　文書ベクトルの模式図（赤い数字は0以外の値であることを示す）

行列の各要素に振り分けられるため「分散表現」といいますが、文書が行列の中に数値として埋め込まれるように見えることから、「埋め込みベクトル」とも呼ばれます。ニューラルネットワークの設計により、Word2Vec、FastText、GloVe、SCDV等の文脈情報が消失するものから、ELMo、BERT等の文脈情報を持ち、多義語に対応できるとされるものに進化しています。また、公開されている分散表現の学習済みモデルからニューラルネットワークを再現し、これに新たな層を追加してネットワークを構成する、蒸留、ファインチューニング、転移学習という手法で高い精度が得られるようになりました。

- **ラベルエンコード**：各単語にIDを割り振り、単語の並びの順にIDを配していくものです。文脈情報は維持されます。ニューラルネットワークを使うときは各文書ベクトルの大きさを揃える必要があるため、通常、データ不足部分を0で埋め、最も長い文書に揃えます（ゼロパッディング）[43]。ラベルエンコードで得られた文書ベクトルは、要素の数値同士に物理量の関係がないため、特殊な数理モデルを使用して機械学習モデルに入力します。
- **one-hot**：各単語のID（数値）を、（ID）番目の要素のみが1で残りの要素がすべて0のベクトルに置換したものです。ラベルエンコードベクトルを2次元空間で表現できるとすれば、one-hotベクトルは3次元空間で表現されます。文脈情報を維持したまま機械学習モデルに入力することができます。しかし、データ容量が増大し、文書数やタスクによってはコンピュータのメモリに収まらないことがあります。

## 5-2. 形態素解析

文書を小さな単位（形態素）に分割することを形態素解析といいます。形態素解析と同時に品詞解析や文法解析を行ってその情報を利用することもあります。

英語やハングル表記等は単語間を空白で区切る「分かち書き」文書のため、形態素解析は必ずしも必要ではありませんが、日本語や中国語は分かち書きがないため、形態素解析が必要になります。以下、形態素解析器の仕組み別に分類します。

- **辞書による方法**：予め単語と品詞を登録した辞書に基づいて自然文を切り出します。辞書にない語（未知語）はそのまま切り出されます。日本語形態素解析器としてはMeCab、Janome、JUMAN、Sudachi等があります。
- **サブワード法**：長い文字列に含まれる一部の文字列（サブワード）が他の単語にも共有されているとき、共有しているサブワードを長いものから順次切り出していく手法です。たとえば文章中に「カチオン性界面活性剤」、「アニオン性界面活性剤」、「親水性」、「添加剤」があるとすると、まず共通部分の「界面活性」を切り出します。MeCabと同じ作者によるSentencepieceが有名です。辞書による方法に比べて形態素の数が少なくなります。また、Sentencepieceでは予め形態素の数をおおよそ設定（例えば8000）して処理することができます。
- **ニューラルネットワーク法**：文脈から次の単語を予測するタスクをニューラルネットワークで学習

| | |
|---|---|
| MeCab | 着色␣剤␣、␣アラビア␣ガム␣、␣炭素␣数␣１␣２␣以上␣、␣２␣０␣以下␣の␣高級␣アルコール␣、␣炭素␣数␣１␣以上␣、␣４␣以下␣の␣低級␣アルコール␣、␣および␣水␣を␣含む␣インクジェット␣インク␣。 |
| Sentencepiece | ▁␣着␣色␣剤␣、␣アラビア␣ガム␣、␣炭素␣数␣12␣以上␣、␣20␣以下の␣高級␣アルコール␣、␣炭素␣数␣1␣以上␣、␣4␣以下の␣低␣級␣アルコール␣、␣および␣水␣を含む␣インク␣ジェット␣インク␣。 |
| モノグラム | 着␣色␣剤␣、␣ア␣ラ␣ビ␣ア␣ガ␣ム␣、␣炭␣素␣数␣１␣２␣以␣上␣、␣２␣０␣以␣下␣の␣高␣級␣ア␣ル␣コ␣ー␣ル␣、␣炭␣素␣数␣１␣以␣上␣、␣４␣以␣下␣の␣低␣級␣ア␣ル␣コ␣ー␣ル␣、␣お␣よ␣び␣水␣を␣含␣む␣イ␣ン␣ク␣ジ␣エ␣ッ␣ト␣イ␣ン␣ク␣。 |
| MeCab+固有表現 | 着色剤、␣アラビアガム、␣炭素数１␣２␣以上、␣２␣０␣以下の␣高級アルコール、␣炭素数１␣以上、␣４␣以下の␣低級アルコール、␣および␣水␣を␣含む␣インクジェット␣インク。 |

※ ␣ は形態素の区切りを示す記号（筆者付与）

図10　形態素解析例（固有表現の取り出しにはTermExtractを使用）

し、得られたモデルの推論結果を出力する方法です。RNN（再帰的ニューラルネットワーク）がよく使われます。JUMAN＋＋、nagisa等があります。

・**n-gram**：文法に関係なくn文字単位で切り出していく方法で、文脈情報を保持することができます。プログラム言語の文字列処理で簡単にできます。nが大きくなるほど同じ形態素の出現頻度が減ると同時に形態素の数が増加していきます。

・**固有表現抽出**：元々は未知語を抽出するものですが、技術用語が形態素解析器でサブワードに分割されてしまい、折角の特徴が失われることがあるため、固有表現抽出でこれを取り出し、解析に用います。固有表現抽出器としてはTermTxtractやGiNZA（の一部機能）等があります。筆者の研究では、TermTxtractで重要度が高い語を抽出し、MeCabで分かち書きした文書の当該箇所を置き換える（図10参照）ことで分類タスクの精度が向上することがわかっています（本稿6－2参照）。

### 5-3. 前処理[44)]

データクレンジングとも言います。特許文書に含まれる「【】」等を正規表現で除去したり、小文字と大文字と、全角と半角とを文字列処理関数を使って統一したりする処理です。ノイズが減少するため機械学習精度向上の効果があります。

## 6. オープンソースを利用した自然言語処理の実装例

機械学習モデルに入力するデータとして、インクジェット分野で、「インキ」、「機構」、「受容シート」、「用途」のカテゴリに対応するそれぞれ複数のFIを設定し、市販特許データベースからテーマコード2H186（インクジェット記録方法及びその記録媒体）で検索して母集団を得て、筆頭FIがそれぞれのカテゴリのFIである公報を抽出しました。

ここでは、インキ：1459件、機構2726件、受容シート726件、用途325件、合計5236件の公報を独立請求項のテキストデータで分類するというタスクを扱います。筆頭FIと独立請求項とは強い正の相関関係があると考えられるため、この組み合わせとしました。

形態素解析はMeCab、sentencepiece、GiNZAの出力をそのまま使用した場合に加え、TermExtractで得た固有表現（たとえば「カチオン/性/界面/活性/剤」に対する「カチオン性界面活性剤」）で元の出力結果を置換した出力（TExMeCab）を使用しました。

文書ベクトル化は、Bag of Words、TF-IDF、分散表現としてDoc2Vec（dm＝0）、ラベルエンコードを使用しました。Bag of Words、TF-IDFでは1-3gramを設定することにより最大連続する3形態素までを計算しています。独立した形態素のみカウントするより精度が1.5%程度向上しますが、デメリットとしては文書ベクトルのサイズが大きくなり計算資源が増加します。また、ラベルエンコードをニューラルネットワークに入力するときは、入力層にEmbedd層を用います。

### 6-1. SVMを用いた2値分類、XGBoost4値分類

SVM2値分類は、機械学習モデルとして図5（左）に示すSVM（サポートベクトルマシン）を用いてインキ、機構のどちらに属するかを判定するone-vs-oneタスクです。図11（左）に示すように、形態素解析器の出力をTF-IDFでベクトル化する方法が高精度で、97%近くの正答率が得られました。

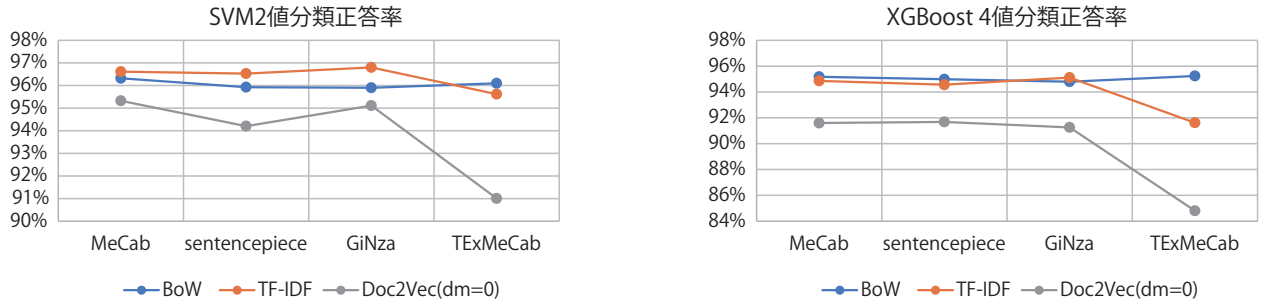



図11　（左）SVMによる2値分類　（右）XGBoostによる4値分類

XGBoost4値分類は、機械学習モデルとして図5（右）に示す決定木を用いて、母集団をインキ、機構、受容シート、用途のそれぞれに分割していくタスクです。SVM2値分類とは異なり、Bag of Wordsがわずかにに TF-IDFを凌いでいます。正答率は95％強が出ました。

この2種類の機械学習には、scikit-learnライブラリを使用しました。

## 6-2. CNNを用いた2値/4値分類

図12に分類タスクに用いたニューラルネットワークを示します。入力層の後にウィンドウサイズ2，3，4，5の4枚のフィルタをもつCNN層をconcat層で集約して、全結合層で次元を減らしていき、出力層から出力します。出力層の次元と同じ次元のラベル（y_）を与えて学習させます。機械学習フレームワークにはTensorflow/Kerasを使用しました。

2値分類の場合の出力層は、図13に示すように4通りの設計が可能です。出力にsigmoid関数を使うと、出力を確率として扱うことができるので、分類1の確率が95%、分類2の確率が5%といった出力も可能です。このように出力をある程度自由に設計できることがニューラルネットワークの強みといえます。

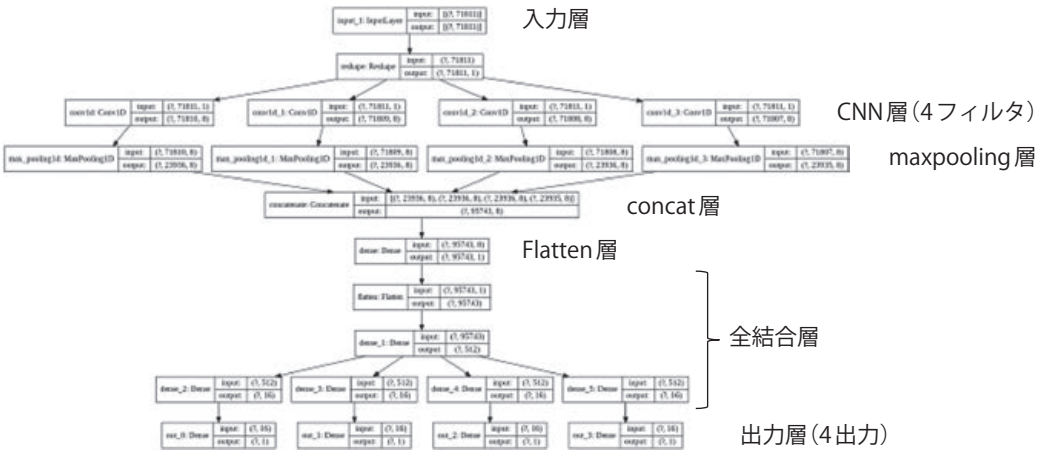



図12　CNNを含むニューラルネットワーク例

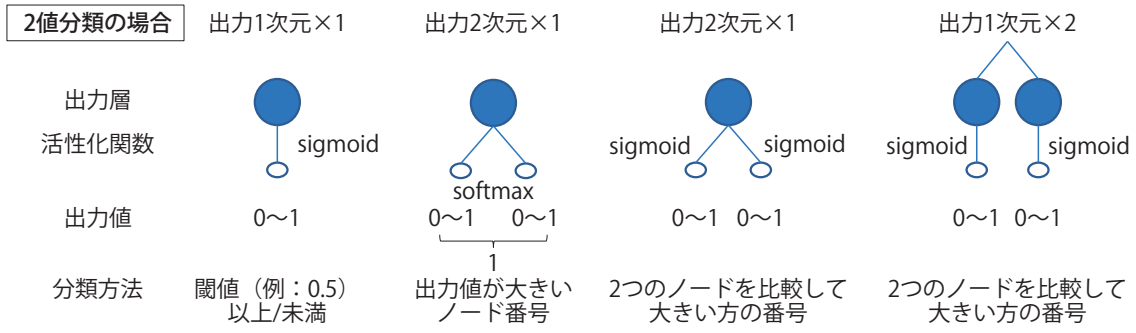



図13　ニューラルネットワークの出力設計

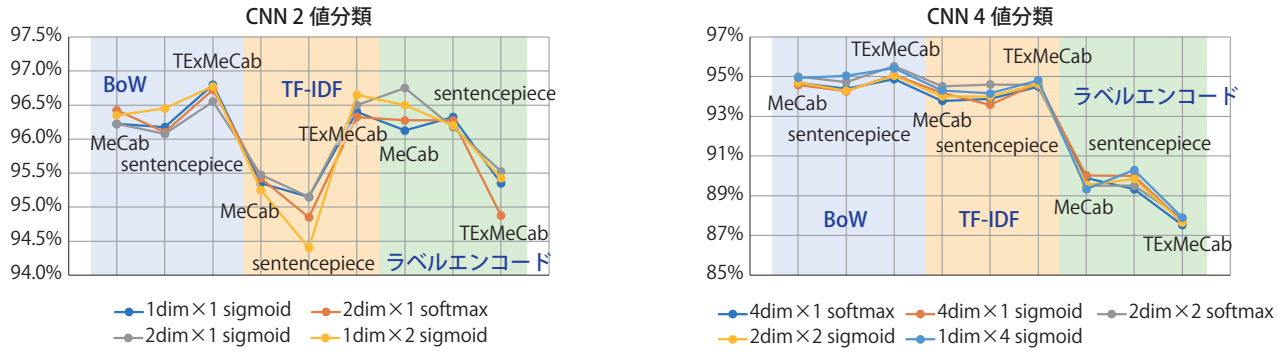



図14　ニューラルネットワークによる分類結果

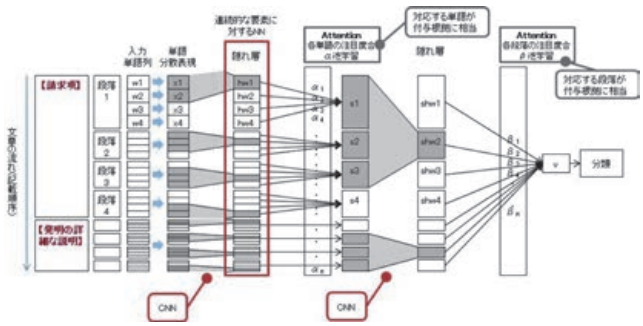
結果は、TermExtractで固有表現を抽出してMeCab出力の当該単語を置換し、BoWでベクトル化する方法が最も高精度でした。MeCabより精度が高いと言われたsentencepieceを凌駕したことが特筆されます。ただし、ニューラルネットワークの設計で精度が変動するので、TF-IDFやラベルエンコードにも最適なネットワーク構造があることでしょう。

### 6-3. 他のタスクへの拡張

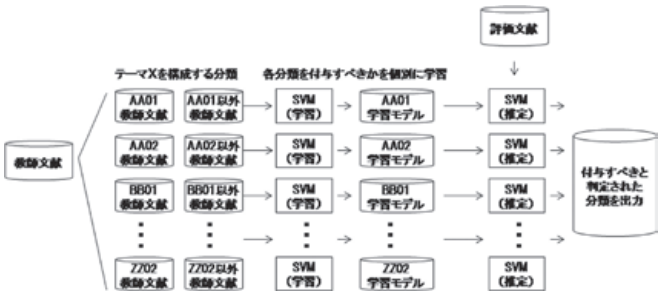
機械学習を用いて特許文献にFタームを付与するタスクを考えます。テーマコードがわかっている状態で、そのテーマコードに100個のFタームが存在すると仮定したとき、100値分類タスクをニューラルネットワークで推論することが考えられますが、実際は大変困難です。

その理由は、要約や請求項だけでは付与根拠が不十分で、明細書の段落を入力文とする必要があること、ラベルの次元が増加するとともに正例（1）に対して負例（0）が指数関数的に増加して、学習過程でラベルに合わせるよりもニューラルネットワークがすべて0を出力する方の損失が少なくなること、にあります。とくに後者の0と1との数のバランスが釣り合わないデータを「不均衡データ」といいます。

「アクション・プラン」の平成29年度実証事業において、特許文献に付与すべきFタームおよび、段落単位での付与根拠を機械学習で推定し、その精度を検証する[12), 45)]ことが行われました。この実証事業は12のテーマごとに20個のFタームを選定し、推論するタスクです。その一方法はRNNを応用したアテンション層を2つ用いるタンデム型のニューラルアテンションモデルです。20値分類のため出力層は20次元で、アテンション層が最も注目したノードの出力が最大となる仕組みです。アテンションモデルは時系列データである単語列を出力するタスクにおいて次に出力すべき単語に「注目」する機能を有し、機械翻訳タスクで飛躍的に精度を上げたことで知られます。別の一方法はFタームごとに入力文がそのFタームであるか否かを2値分類するSVMをFタームの数だけ備える、one-vs-rest方式です。またこれらをベースに改良したモデルも検討されています。




CNN-NAMによる付与根拠箇所推定の構成




SVM によるFターム付与方式：One-vs-rest 方式

図15　Fタームを推定するニューラルネットワーク例（文献45より転載）

## 7. おわりに

機械学習導入のメリットと、個人がオープンソースのデータまたはオープンソースのプログラムを使って、ほぼ無料で知財情報解析ができるようになった事例とを、基礎技術を中心に紹介しました。本稿では取り上げませんでしたが、紹介した基礎技術を用いて数百～数万の文書ベクトルを2～3次元に圧縮してプロット[46]したり、固有表現の共起関係をグラフ化[47]したりすることで、従来のテキストマイニングツールとしての使い方もできます。

本稿は2019年及び2020年の「アジア特許情報研究会」におけるワーキングの一環として報告するものです。なお、本稿の第4章は安藤が執筆を担当し、上記以外は西尾が執筆を担当しました。


### 謝辞

特許庁総務部総務課特許情報室の皆様には情報交換の場で大変有意義なアドバイスをいただきました。ここに感謝申し上げます。


### 参考文献

紙幅の関係で本稿に収録できなかった解説記事、実装例をアジア特許情報研究会のwebページに収録していますのでご参照いただければ幸いです。


1) 独立行政法人情報処理推進機構 社会基盤センター，AI白書2020～広がるAI化格差（ギャップ）と5年先を見据えた企業戦略～，株式会社角川アスキー総合研究所，2020年3月2日発行，p.381-555，https://www.ipa.go.jp/ikc/publish/ai_hakusyo.html
2) 日新税理士事務所，中小企業におけるAIの活用事例，https://ns-1.biz/report/k-201707.pdf、ほか多数
3) 文献1，pp.559
4) 特許庁，特許庁における人工知能（AI）技術の活用に向けたアクション・プランの公表について，https://www.jpo.go.jp/system/laws/sesaku/ai_action_plan/ai_action_plan.html
5) 特許庁，特許庁における人工知能（AI）技術の活用に向けたアクション・プランの平成30年度改定版について，https://www.jpo.go.jp/system/laws/sesaku/ai_action_plan/ai_action_plan-fy30.html
6) 特許庁，特許庁における人工知能（AI）技術の活用に向けたアクション・プランの令和2年度改定版について，https://www.jpo.go.jp/system/laws/sesaku/ai_action_plan/ai_action_plan-fy2020.html
7) 麻川倫広，特許庁における人工知能技術の活用に関する取組について―人工知能技術の業務への適用可能性の検討及び電話等の質問対応に関する実証事業を中心に―，Japio YEAR BOOK2018，p.6-11，https://www.japio.or.jp/00yearbook/files/2018book/18_a_01.pdf
8) 近藤裕之，特許文献への分類付与と付与根拠の推定，Japio YEAR BOOK2018，p.16-22，https://www.japio.or.jp/00yearbook/files/2018book/18_a_03.pdf
9) 渡邉潤，平成29年度 商標業務におけるAI活用に関する実証的研究事業―「先行図形商標調査」、「不明確な指定商品・役務調査」の高度化・効率化の可能性調査―，Japio YEAR BOOK2018，p.24-27，https://www.japio.or.jp/00yearbook/files/2018book/18_a_04.pdf
10) 藁谷智雄，人工知能技術を活用したデータエントリー業務の高度化・効率化について，Japio YEAR BOOK2018，p.28-33，https://www.japio.or.jp/00yearbook/files/2018book/18_a_05.pdf
11) 多賀和宏，特許庁におけるAI技術を活用した業務支援ツール導入に向けた取組について，Japio YEAR BOOK2019，p.22-27，https://www.japio.or.jp/00yearbook/files/2019book/19_a_02.pdf
12) 特許庁委託事業 外国特許文献へのFターム等付与に関する機械学習活用可能性調査事業調査報告書，日立製作所，2017年3月
13) 特許庁委託事業 Fターム等付与支援システム実用化に向けた実証的研究事業調査報告書，日立製作所，2018年3月
14) 特許庁委託事業 先行技術調査支援システム実用化に向けた実証的研究事業報告書，日立製作所，2018年3月
15) 特許庁，特許行政年次報告書2020年版第2部第6章，p.160 https://www.jpo.go.jp/resources/report/nenji/2020/document/index/honpenall.pdf
16) 野崎篤志，知財情報調査・分析を取り巻く人工知能とその周辺動向－AIツール・RPAツールとの協働・共創時代へ－，Japio YEAR BOOK2019，p.58-65，https://www.japio.or.jp/00yearbook/files/2019book/19_a_08.pdf
17) 特集：テキストマイニング技術の活用に関する研究，https://www.jstage.jst.go.jp/browse/jkg/69/7/_contents/-char/ja
18) 特集：インフォプロのためのプログラミング事例集 https://www.jstage.jst.go.jp/browse/jkg/70/4/_contents/-char/ja
19) 特集：AI時代のインフォプロ，https://www.jstage.jst.go.jp/browse/jkg/70/7/_contents/-char/ja
20) 情報検索委員会第3小委員会，テキストマイニング技術の活用に関する研究，知財管理，69巻12号（2019），p.1426-1440
21) 情報検索委員会第4小委員会，特許およびオープンな非特許情報を活用した特許分析手法の検討，知財管理，69巻9号（2019），p.1706-1719
22) 丸山 宏，深層学習と知財－技術とビジネスの観点から－，知財管理，70巻4号（2020），p.538-550
23) アジア特許情報研究会，https://sapi.kaisei1992.com/
24) 人工知能の定義：John McCarthy，WHAT IS ARTIFICIAL INTELLIGENCE?，http://www-formal.stanford.edu/jmc/whatisai/whatisai.html、松尾豊，人工知能は人間を超えるか，p.45 など
25) 総務省，平成28年版情報通信白書，p.233，https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/pdf/n4200000.pdf
26) John Searle，Minds，Brains and Programs，Behavioral and Brain Sciences，3 (3)，417-457
27) 久保隆宏，機械学習スタートアップシリーズ Pythonで学ぶ強化学習，講談社，ISBN978-4065142981
28) たとえば、Google ResearchのInceptionism関係の論文

を流し読み，http://cygx.mydns.jp/blog/?arti＝563、Deep Learningとは，http://ibisml.org/archive/ibis2013/pdfs/ibis2013-okatani.pdf、ほか多数
29) 岡谷貴之，画像認識分野でのディープラーニングの研究動向，http://ibisml.org/archive/ibis2013/pdfs/ibis2013-okatani.pdf
30) 斎藤康毅，ゼロから作るDeep Learning――Pythonで学ぶディープラーニングの理論と実装，第8章，ISBN978-4-87311-758-4
31) たとえば、リコー株式会社，https://www.ricoh.co.jp/service/cloud-ocr/column/aiocr/、キヤノン株式会社，https://cweb.canon.jp/solution/biz/trend/ai-ocr.html、など
32) たとえば、化学の新しいカタチ，https://future-chem.com/smiles-smarts/
33) IPPH，外观设计专利智能检索系统，http://wg.cnipr.com/、http://dscope.cnipr.com/fenlei/first?lang＝jo_JP
34) たとえば、scikit-learnのSVMでMNISTの手書き数字データを分類，https://note.nkmk.me/python-scikit-learn-svm-mnist/、kerasのmnistのサンプルを読んでみる，https://qiita.com/ash8h/items/29e24fc617b832fba136、など
35) 黒橋禎夫，改訂版 自然言語処理，放送大学教育振興会，ISBN978-4595319587
36) 奥村学，特許情報処理:言語処理的アプローチ（自然言語処理シリーズ5），コロナ社，ISBN978-4339027556
37) scikit-learnのアルゴリズムチートシート：http://scikit-learn.org/stable/tutorial/machine_learning_map/
38) 学習時に誤差が小さくする（収束する）ためには微分可能な式である必要がある
39) 機械学習フレームワークの関数として実装されるのでパーセプトロンを直接記述する必要はない
40) たとえば、0が多いラベルに対して0のみを出力するネットワークモデルは、正答率が高いが実用には適さない
41) 目黒光司ほか，Fターム概念ベクトルを用いた特許検索システムの改良，言語処理学会 第21回年次大会発表論文集（2015），https://www.anlp.jp/proceedings/annual_meeting/2015/pdf_dir/D5-1.pdf
42) 意味情報：王様－男＋女＝女王、パリ－フランス＋日本＝東京 などの演算例が有名
43) 特許文書、特に請求項では1文書の前半部分に特徴となる単語が含まれることが多いので、文書を前から特定の長さでカットしても良い。精度とトレードオフの関係になるが、ベクトルサイズが縮小し、計算資源を有効活用することができる
44) 本橋智光，前処理大全［データ分析のためのSQL/R/Python実践テクニック］，技術評論社，ISBN 978-4774196473
45) 富永泰規ほか、特許文献への分類付与における付与根拠箇所推定，情報の科学と技術 68巻7号（2018）p.338-342，https://www.jstage.jst.go.jp/article/jkg/68/7/68_338/_pdf/-char/ja
46) 安藤俊幸，機械学習による予備検索を考慮した効率的な特許調査，pp14-15，https://sapi.kaisei1992.com/wp-content/uploads/2020/07/bd1c796e71bfde143b69ae262af724b5.pdf
47) Pythonで共起ネットワークを作成する｜いるかのボックス，https://irukanobox.blogspot.com/2019/10/python.html

## profile

**西尾 潤**（にしお じゅん）

2011年　（株）ユポ・コーポレーションにて知財業務に従事
2018年　同社開発部門
2015年　アジア特許情報研究会入会、中国特許調査および知財情報解析の研究

## profile

**安藤 俊幸**（あんどう としゆき）

1985年　花王株式会社入社、研究開発に従事
1999年　研究所の特許調査担当（新規プロジェクト）
2009年　知的財産部
2011年　アジア特許情報研究会入会
2019年度「特許情報普及活動功労者表彰」日本特許情報機構理事長賞【技術研究功労者】受賞
情報科学技術協会、人工知能学会、データサイエンティスト協会 各会員