ディープラーニングによる特許文献からの技術用語抽出

Technical term extraction from patent documents by deep learning

株式会社 NTT データ数理システム　データマイニング部グループリーダー・主任研究員
岩本　圭介

1999 年株式会社数理システム（現：株式会社 NTT データ数理システム）入社。データマイニング・テキストマイニングに関わるツール・手法開発及び分析業務に従事。現職はデータマイニング部グループリーダー・主任研究員。

iwamoto@msi.co.jp

1　はじめに

昨今、ディープラーニング技術が機械学習技術のブレイクスルーとして盛んに注目を浴びており、画像認識、音声認識、自然言語処理など様々な分野での成果が報告されている。本稿では、言語処理の分野に焦点を当ててディープラーニング技術の解説を行い、また特許情報のデータからディープラーニングにより構築した言語モデルを用いた技術用語の自動抽出への適用可能性について紹介する。以下、２章で機械学習とディープラーニング、及びディープラーニングの基礎概念であるニューラルネットワークについて概観し、３章では自然言語処理への適用が広く試みられている RNN と呼ばれるモデルの解説を行う。４章で実際にこれを技術用語抽出に適用した例を解説し、最後に５章でディープラーニングに関わる当社株式会社 NTT データ数理システムの取り組みを紹介する。

2　機械学習とディープラーニング

ディープラーニングは、機械学習の一手法であるニューラルネットワークをより発展的に利用すべく開発された、強力なテクニックの集合体である。ディープラーニングを理解するためには、ニューラルネットワーク及び機械学習の基本的な概念を押さえておく必要があるため、本節ではまずそれらの点について解説する。

2.1　ニューラルネットワークの基礎

ニューラルネットワークは、神経回路網を模した一種の計算ロジックであり、神経細胞の働きをシミュレートした仮想的な素子であるニューロンから構成される。このニューロンの模式図を図 1 に示す。このニューロンは入力値 $x_1, x_2, \cdots, x_n$ を受け付け、それらの値に結合荷重 $w_1, w_2, \cdots, w_n$ を掛けて足し合わせたものからバイアス θ を減じた値に活性化関数と呼ばれる単調増加関数 f を適用した結果を出力とする。これは、実際の神経細胞が、他の神経細胞からの信号が加わることで膜電位が上昇し、その電位が閾値を超えると接合部位であるシナプスを通じて他のニューロンに対して出力される（発火する）ことを模したものになっている。

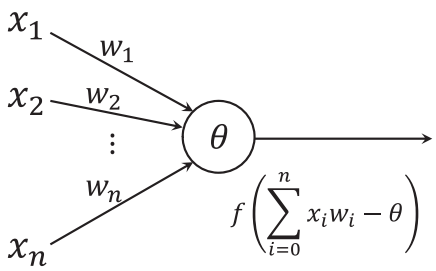

図 1　ニューロンの模式図

このニューロン 1 つのみでは、入力の重み付き総和の大小に従って出力値が決まるのみの機構でしかないが、このニューロンの数を増やし、あるニューロンの出力を他のニューロンが受け付けるというように層状のネットワークをなすよう連結していったものがニューラルネットワークである。

ここで、ニューラルネットワークが何らかの目的を達成するように動作させることを考える。例えば、図２

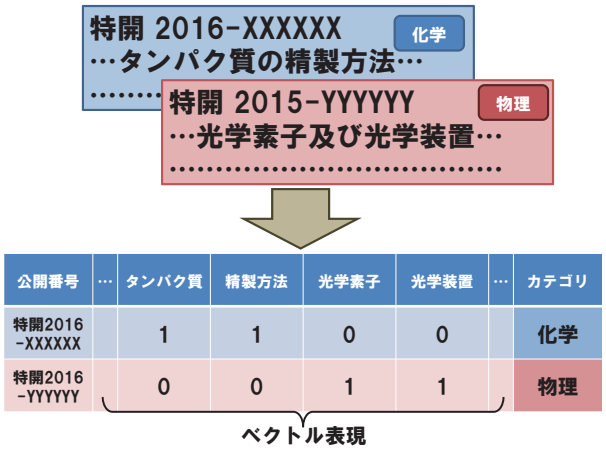

図２　文書のカテゴリ分類問題

のような文書のカテゴリ分類といった例を考える。文書のベクトル表現を作成し、カテゴリ「化学」に対応するような入力ベクトル $\mathbf{x}=\{x_1, x_2, \cdots, x_n\}$ を与えた場合には「化学」に相当する出力層のニューロンが発火し、「物理」に対応するような入力を与えた場合には「物理」に相当するニューロンが発火するようにすることで、ニューラルネットワークが分類器として機能するようになる。この様子を図３に示す。

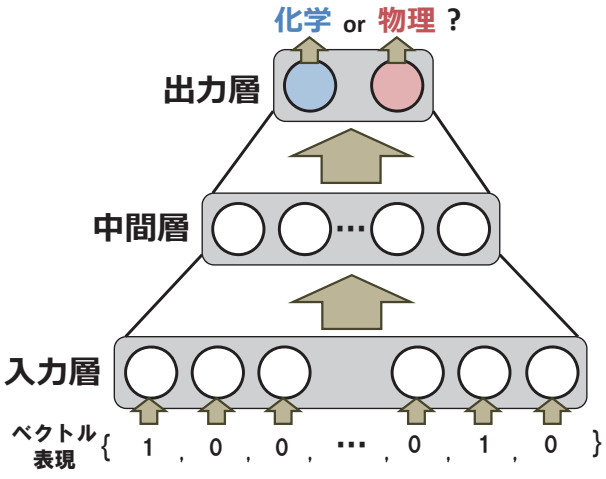

図３　文書の分類問題

これは、問題としては文書に対してどのカテゴリに割り振られるべきかという正解が与えられているような性質のものであり、こういった場合の入出力関係を獲得するような行為を教師あり学習と呼ぶ。そして、ニューラルネットワークの場合は、望むべき入出力の対応関係が得られるように各ニューロンの結合荷重 $\mathbf{w}=\{w_1, w_2, \cdots, w_n\}$ 及びバイアス θ の値を逐次調整していくことがこの学習の過程に相当する。実際には、そのとき入力層から順次情報を伝播させ出力層に得られた予測値と、本来あるべき正解値との誤差を評価し、この誤差が小さくなるよう $\mathbf{w}$ の値を修正する。この修正方法としては、一般的には確率的勾配降下法（Stochastic Gradient Decent）と呼ばれる最適化手法を用いる。また、出力層から入力層までの誤差を伝播する方法は、逆誤差伝播（Back Propagation）と呼ばれる。

また、前述の教師あり学習に対して、正解値といったものが存在せず、入力データのみからモデルを作成するタイプの学習は教師なし学習と呼ばれる。これについては、例を３章で述べる。

2.2　ディープラーニングの特徴

従来は、ニューラルネットワークの構造として１層の入力層、１層以上の中間層、１層の出力層を持つものが多く用いられてきたが、これよりも層の数を増やし、複雑な構成を持つネットワークを用いてモデルを構築することがディープラーニングである。ネットワークが「深く」なればそれに応じて調整しなければならない結合荷重 $\mathbf{w}$ の個数は飛躍的に増えるため、ネットワークの学習には非常に時間を要するが、近年の並列計算や GPU といった高速計算環境の発展に伴い様々な場面に適用されるようになってきた。

また、深い層を持つニューラルネットワークがデータの抽象的な内部構造を発見しているという報告がなされ、注目を浴びている。画像を入力とした、顔認識への適用例[1]を挙げると、入力となるデータそれ自体は濃淡を持つピクセルの集合でしかないが、中間層において、顔の輪郭やパーツといった抽象化した特徴を自ら獲得していることが判明した。この抽象化した特徴を自ら発見することで、認識の精度は従来手法とは比べ物にならないほど向上した。この例では、畳み込みニューラルネットワーク（Convolutional Neural Network, CNN）とプーリング（Pooling）と呼ばれる技法の組み合わせによって、画像の局所的な領域の特長を次の層に対応付け、層を経るにしたがってデータの抽象化を高めていく働きが得られている。

3　RNN と言語モデル

言語処理の分野で成果を挙げ、注目されているアルゴリズムが再帰型ニューラルネットワーク（Recurrent Neural Network, RNN）である[3]。

言語モデルとは、一般に単語列の確率分布を表すものである。著名なモデルとして n-gram 言語モデル があり、これは（$n-1$）個の単語 w_1, w_2, …, w_{n-1} が与えられたときに、次の単語 x が出現する確率 $P(x|w_1, w_2, \cdots, w_{n-1})$ を与えるものである。（2 章では文字 w を結合荷重：weight の意味で用いたが、本章以降では単語：word の意味で用いることとする。注意されたい。）

RNN はこの言語モデルを構築することが可能なネットワークであり、図 4 のような構造を持つ。

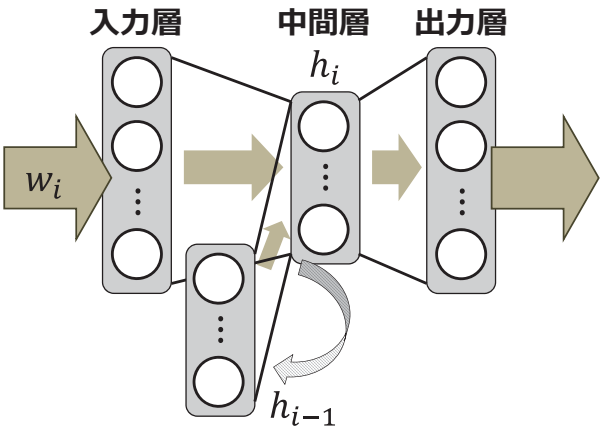

図 4　RNN の構造

まず、入力層は単語の語彙数に相当する次元をと持つものとし、そのときの単語 w_i に相当する成分だけが 1、その他の成分が全て 0 であるようなベクトルとして入力を与える。そして、入力データは層間の演算を順次受け、入力層よりも低次元な表現 $\mathbf{h}_i$ を中間層に得る。ここで、$\mathbf{h}_i$ はその直前の層の出力だけではなく、1 つ前の単語を与えた時の値 $\mathbf{h}_{i-1}$ も同時に入力することが RNN の特徴である。ここで、$\mathbf{h}_i$ は入力 w_i（を表す各成分が 1 or 0 のベクトル）の次元を層間の演算を経て圧縮したものであるといえるが、それを求める際の計算に $\mathbf{h}_{i-1}$ が寄与し、またその $\mathbf{h}_{i-1}$ を求める際には $\mathbf{h}_{i-2}$ が寄与し…というように過去の出現単語の影響を再帰的に受けている点が特色である。したがって、この低次元表現は、過去どういう単語列を経て今 w_i が出現しているのか、言い換えればどんな文脈の中で w_i が出現しているかという情報も折り込まれていることになる。

さらに、ニューラルネットワークの学習方針としては、この $\mathbf{h}_i$ から次の単語 w_{i+1} が得られるように結合荷重を調整していくものとする。今度は、$\mathbf{h}_i$ から順次出力層に向かって次元を拡大し、最終的には入力層と同様なベクトルの 1 or 0 表現に戻し、w_{i+1} に相当する成分が発火

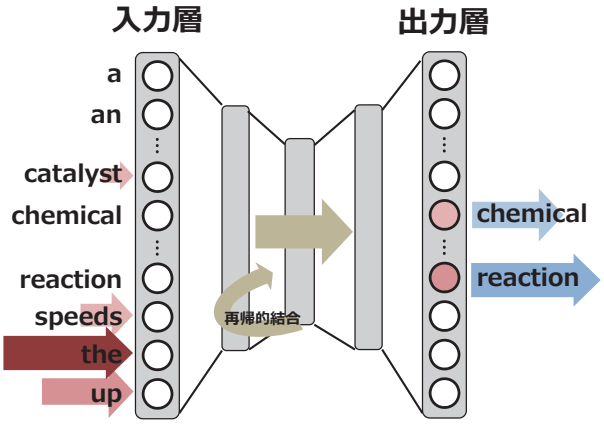

図 5　RNN 言語モデルの学習

するようにする。結果、このネットワークは、w_i までの単語列…, w_{i-2}, w_{i-1}, w_i を順次入力したときに、次の単語 w_{i+1} の予測確率 $P(w_{i-1}|w_i, w_{i-1}, \cdots)$ が得られるようなものになる。このネットワークそのものが言語モデルを表すといえるが、先の n-gram 言語モデル とは異なり、RNN 言語モデルでは明示的に n の値を決めなくてよい点が特色である。学習に用いた文書中に

- catalyst speeds up the reaction …
- catalyst speeds up the chemical reaction …

というような表現が多く存在していたと仮定すると、図 4 のように catalyst, speeds, up, the といった単語を順次入力していった場合、その時点で chemical や reaction といった「文脈を踏まえたうえで、次に出現する確率が大きい単語」が大きな出力値を持つようになることが期待できる。

この過程では、w_i という入力に対して w_{i+1} という正解を設定はするものの、入力データのみからその内部的な構造（言語モデル）を学習するものであるため、教師なし学習を行っている例にあたる。また、RNN モデルの構築時には、実用的には過去の情報をどの程度保持するか、もしくは忘れるかといった点も含めて学習を行う LSTM（Long Short-Term Memory）といった仕組みをあわせて利用することが多い。

4　特許情報への応用

4.1　RNN 言語モデルの応用

RNN 言語モデルを作成した場合、その使い道は大きく分けて 2 種類存在する。一つは、次の単語の出現確

率を予測するという性質をそのまま使うものである。例えば、モデルに基づいた文章の自動生成、また学習に用いたテキストとは異なるテキストを用いてその文書における単語出現状況の正しさや確からしさを評価する、というような適用が考えられる。[2] では文章の自動生成や校正といった事例が示されている。

もう一つの使い道は、文章を RNN 言語モデルに入力した際の中間層の表現 **h** を、文章の次元圧縮表現として用いることである。一般に文章は可変長の単語列 w_1, w_2, …, w_n といった情報で与えられるが、これらの単語の系列を反映させた抽象的な表現を、固定長のベクトルである **h** に落とし込めるという点でこの手法は特に価値がある。単語 w を用いるかわりにこの **h** を用いて教師あり学習を行う、となどといった適用が考えられる。

4.2　技術用語の抽出例

ここでは、次の単語の出現確率を予測するという RNN 言語モデルの性質をそのまま使った適用を考える。特許文献等、英文の技術情報に対してテキストマイニングを行いたいといった場合、1 単語単位での集計や傾向把握のみでは、その語がどのような文脈で用いられているかが不明であるためアウトプットから具体的な状況を読み取ることは難しい。一般に、技術用語とみなせる単語は複数の単語の連接からなることが非常に多いが、事前にこういった連語を網羅的に抽出して辞書化しておくことは現実的ではない。

そこで、RNN 言語モデルを用いて、単語間の確率的な結合度合いを評価し、まとまって出現する傾向の強い語群･連語を自動的に抽出することを試みた。学習には、機械学習分野の PCT 特許 122 件の要約部分を用いた。データ諸元を表 1 に示す。

表 1　利用データ

文章数	496 文章
延べ単語数	13,439 語
単語種別数	2,568 種類

このデータを 2 層の LSTM 層を持つ RNN で学習した。**h** の次元数は 100 とした。そして、学習後のネットワークに対し、次の手順で連語の抽出を試みた。

1. ネットワークに、適当な初期単語列を与える。
2. それまでの単語列…, w_{i-2}, w_{i-1}, w_i から、次に生じる単語 w_{i+1} の確率分布 $P(w_{i+1} | w_i, w_{i-1}, \cdots)$ をネットワーク出力より求める。
3. 乱数を生成して w_{i+1} から 1 つの単語を選択する。w_{i+1} が文末でなければネットワークへ再度入力し、2.～3. を繰り返す。
4. 生成された単語列の中から、閾値以上の予測確率が連続した文断片のみを抽出する。
5. 初期単語列を変えて、1.～5. を繰り返す。

上記のうち、2.～3. が RNN 言語モデルを用いた文生成の過程であるといえる。1.の適切な初期単語列は、学習に用いたデータから適度に乖離したものを与えることが望ましいと考え、原文中のランダムな位置から連続する 3 単語のみをランダムに抜き出して与えることとした。4.の閾値は 0.99 とした。5.の繰り返し数は 2000 とした。

結果として得られた連語のリストを表 2 に示す。生成文中に出現した頻度の大きい 20 例を示している。

表 2　抽出された連語表現

one or	according to
based on	actuation of
associated with	machine learning 222
at least	portion of
interpretation of	concept using the configuration
learned function 502	existence of one or
suitabilities of	metrics based on
components of	prediction problem
included in	rejected sentences
thereof; and a software module	snps in

前置詞で終わる定型句が多く抽出されたため、更に、以下のようなストップワード設定を行うことで、表 3 の結果を得た。こちらに関しても上位 20 例を示している。機械学習分野の技術用語とみなせる連語が抽出できているといえる。

- 末尾の前置詞は含めない。
- 先頭の冠詞・be 動詞は含めない。
- 先頭・末尾の接続詞は含めない。
- アルファベット以外の文字が含まれる語は含めない。

頻度下位のものには、複数の節からなるような、単語

表３　ストップワード設定を施した連語表現

at least	metrics based
machine learning	modeling procedures
prediction problem	rejected sentences
learned functions	threshold charge level
software module	training data
selected modeling procedures	computing units
concept using	determined based
existence of one	diagnosis or other
heterogenous set	feature data
magnetic resonance data	functional interpretation

数が 10 以上と非常に長い「文章そのまま」とみられるような連語表現が抽出される現象もみられたが、これは RNN 言語モデルの内部状態が原文をそのまま与えた状態と近いものになり、原文をほぼそのまま出力するサイクルに入ってしまったためと推測する。

5　当社の取り組みと展望

本稿では、言語処理の分野で広く適用されているディープラーニングの学習モデルについて解説を行い、構築した言語モデルを用いて技術用語の自動抽出が行える可能性について論じた。

最後に、当社におけるディープラーニングの活用について述べる。当社 NTT データ数理システムは、データ分析や機械学習を実現する分析ツールのディベロッパー・ベンダーであり、また分析受託業務やツールを実際に活用しての分析コンサルティング、システム化への支援を行い顧客の問題解決に実績を挙げてきた。分析受託業務においては、画像認識などの分野でもディープラーニングが成果を上げている。これを、当社製品群に加えるべく、データマイニングツール **Visual Mining Studio** へのディープラーニング機能の搭載に向けた開発が現在進行中である。一般に、世間のディープラーニングのフレームワークを利用するには特殊な環境設定や Python 言語などでのプログラミングを要するものが多いが、現在開発中の機能はツールの GUI からビジュアルに、プログラミング不要で利用可能なものになる予定である。

参考文献

[1] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng：“Unsupervised learning of hierarchical representations with convolutional deep belief networks”, Communication of the ACM, 54(10) pp.95-103 (2011)

[2] 高橋 諒：“Deep Learning で始める文書解析入門（終）：LSTM と Residual Learning でも難しい「助詞の検出」精度を改善した探索アルゴリズムとは - @ IT”, http://www.atmarkit.co.jp/ait/articles/1611/11/news016.html (2016)

[3] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, S. Khudanpur.：“Recurrent neural network based language model”, Proceedings of Interspeech (2010)