テキストアナリティクスの動向と特許情報処理
―人間の言葉を機械で読み解く―

Trends in Text Analytics and Patent Information Processing

日本アイ・ビー・エム株式会社　東京基礎研究所主席研究員

那須川　哲哉

1989 年日本アイ･ビー･エム株式会社に入社。東京基礎研究所に配属以後、IBM T.J. ワトソン研究所での 1 年間の勤務、コンサルティング部門への 1 年半の出向などを経験しつつ、一貫して機械翻訳やテキストマイニング、評判分析、会話マイニングなど自然言語処理関係の研究に従事。博士（工学）。

nasukawa@jp.ibm.com

1　はじめに

人間の知的能力を計算機で実現しようという人工知能の研究において、人間が使う言語を扱う自然言語処理は、その重要な一分野であり、この自然言語処理の応用の一つがテキストアナリティクスである。従来、人が読んで解釈するしかなかったテキストデータを計算機で処理して活用するための技術であり、1990 年代に大量データを活用するためのデータマイニングが盛んになった頃から取り組みが盛んになってきた。当初はテキストマイニングやテキストデータマイニングと呼ばれていたものが、2000 年代に入ってから米国を中心にテキストアナリティクスと呼ばれるようになってきている。日本では今でもテキストマイニングと呼ばれることが多い[1,2,3]。

テキスト情報の一種である特許データは、比較的早い時期から電子化されていたこともあり、1990 年代にテキストアナリティクスの取り組みが本格化した当初から、その適用対象として、様々な試みが行われてきた。

テキストアナリティクスの本質は、人間が読み切れないような膨大なテキストデータの活用を可能にする点にあり、人間が通常行っている作業を代行するという観点での人工知能、Artificial Intelligence としての AI とは異なり、人間の能力を拡張する、Augmented Intelligence としての AI であるというのが筆者の考えである。

近年注目されている深層学習などの仕組みで自動化される処理は、ある入力データに対して、特定の結果を付与するものである。例えば、画像データに対して、（「食べ物」や「人物」「笑顔」「男性」「女性」といった）画像内容のラベルを付与したり、入力されたテキストを他言語に翻訳したりする処理であり、入力と出力のデータが大量に与えられると、その処理を模倣するような処理系を自動的に構築することができるようになってきている。それに対し、テキストアナリティクスは、従来、人手で行うことができなかった処理を実現するものであり、深層学習などの仕組みで自動化されるようなものではない。

テキストアナリティクスを導入するということは、特許も含めた、多様なテキストデータの新たな活用法を実現し、従来できなかったことを可能にするという点で大きな可能性を秘めている半面、今まで誰も担当していなかった新たな業務を作り出すことになる。従って、効果がありそうに感じられても、手間がかかりそうだったり、新しい応用なので投資対効果の見通しが困難だったりすることから、大きな投資は躊躇されるケースが多かった。

筆者は、1990 年代から 20 年以上にわたって、テキストアナリティクスに取り組み、基礎技術の研究開発に加え、企業研究者として、社内外の多くの現場での適用を進め、効果検証を行ってきた。成功事例を積み重ね

ることで、着実に国内外での普及が進み、普及の速度も上がってきている。特に近年では、ビッグデータやAIに対する関心の高まりから、導入しなければ取り残されるという意識が生じてきているようで、特許分析も含めたテキストアナリティクスの需要が急速に高まっていると感じている。

本稿では、まず、テキストアナリティクスの概要とその発展の動向を示した上で、テキストアナリティクスを用いた特許情報処理の可能性を紹介する。

2 テキストアナリティクス

1990年代、筆者らがテキストアナリティクスに取り組み始めた当初は、インターネットやPCの普及率が低く、現在のようにネット上にテキストデータが溢れているような状況ではなかったため、まずは研究材料としてのテキストデータを探すところから始まった。

計算機でアクセス可能な電子化されたテキストデータを収集・蓄積する企業が増えたのは、2000年問題でシステム更改が進んでからのようであり、1990年代には、テキストアナリティクスを活かせそうなデータが無いかを訊いて回っても「紙なら大量にあります」と言われることが多かった。技術的にはOCRで電子化すれば良いものの、実証実験を行う前に電子化及び人手によるエラー修正のコストが必要となれば、なかなか話が進展しないのが現実であった。

筆者らにとって幸いだったのは、当時IBMがPCの製造販売をしており、その顧客を電話でサポートするためのPCヘルプセンターで、全ての顧客対応の概要のテキストを、オペレータがシステムに入力して蓄積していたことである。日本IBMのPCヘルプセンターでは、毎週一万件程度の記録が蓄積されており、ヘルプセンターの管理者から数十万件規模のデータの提供を受けることができた。

最初から数十万規模のデータを対象にすることにより、ビッグデータ時代に先行する形で、現在でも高い競争力を維持する技術の開発につなげることができた。この技術の重要な点が「人が読みこなすことのできない大量のテキストデータの活用」にある。元々人が読みこなすことはできないデータなので、人が読みこなす場合とは違う処理を適用し、新たな価値を生み出すことになる。

2.1 PCヘルプセンターの顧客対応データ

顧客対応データは、日時や機種名などの定型情報と、問い合わせ内容の概要を自由に記述した数十文字から数百文字程度のテキストデータで構成されている。このデータは、顧客の疑問や要望などを把握するための貴重な情報源であるが、テキスト情報を定型情報のように集計することができないため、充分な活用ができていないのが実態であった。

問い合わせ内容が多種多様であるため、どのような問い合わせを受け、どのように回答したかをテキスト化して記録していたが、毎週一万件のテキスト全体に人が目を通して把握することはできない。そこで、全体の傾向を捉えるため、特許文書におけるIPCやFタームのような分類コードを設定し、オペレータが選択したコードを集計できるようにしていた。分類コードとしては、『CALL 種別』『問題種別』『回答・対応種別』など複数種類が存在し、各々に、数十程度の選択肢が付けられていた。実際に選ばれたコードを見てみると、大半が、「製品説明/情報が必要」「情報提供」といった極めて抽象度の高いコードであった。具体的な内容を示すコードを用意しようとすると、コードの数が多くなり、多くの候補から適切なコードを選択するのが難しくなる。また、一回の問い合わせに複数の内容が含まれていることも実際に多い。結果的に、抽象度の高い無難なコードが選ばれる傾向が強くなり、分類コードを集計しても、何らかのアクションにつながるような知見は得られないのが実状であった。結局、テキスト部分を読まなければ問い合わせの内容は把握できないが、量が多いため、200件から300件程度のテキストデータにざっと目を通し、気になる内容があれば報告する程度であり、せっかくのデータが活かされていない状況であった。

筆者らがこのPCヘルプセンターの顧客対応データに対してまず取り組んだのがキーワード抽出である。日本語のテキストでは、英語のように空白で単語が区切られていないため、単語を切り出すだけでも自然言語処理の技術を適用する必要がある。逆に言えば、空白で単語を区切ることのできる英語のテキストデータの処理では、自然言語処理を使わずに単語の分布を集計することがで

きるため、テキストを（空白で区切られた文字列としての）単語の集合体として統計処理をする傾向が強い。そのため、単数形と複数形が別の単語として扱われたり、可能性を示す助動詞の "can" と缶を示す名詞の "can" が同じ単語として扱われたりすることになり、意味を扱うという点では、ノイズの多い浅いレベルの処理にとどまってしまう。

日本語のテキストからキーワード抽出を行う場合、例えば、キーワードとなる表現の一覧を定義しておき、定義された表現の文字列が存在するかを調べるだけで処理しようとすると、「東京都」という文字列から「京都」を抽出するようなエラーにつながってしまう。そこで、自然言語処理で日本語を扱う場合、一般的には、形態素解析という処理を適用し、単語よりも細かい形態素という単位で文字列を切り分け、各形態素の品詞を特定する。この処理によって、単語を切り出すだけではなく、品詞の情報も利用できるようになる他、活用している動詞や形容詞などの終止形を特定することも容易になる。すなわち、自然言語処理を適用することで、単に文字列を切り出す以上の処理が可能になる。

キーワード抽出により、どのような表現がどれだけのデータに含まれているかの集計が可能になる。例えば、日本 IBM の PC ヘルプセンターの約 32 万件のデータにおいて頻出する名詞は表 1 のようになった。

表 1　PC ヘルプセンターの顧客対応データに含まれる名詞

名詞	件数
Windows95	40,603
設定	35,385
WIN95	29,648
画面	24,602
方法	23,899
APTIVA	20,187
Windows	19,693
お客様	18,638
ドライバ	18,260
状態	16,900

表 1 のデータから、例えば、35,385 件のデータのテキスト中に「設定」という名詞が含まれていることが分かる。しかし、このような雑多な表現の分布の情報だけでは、アクションに結びつくような知見の獲得にはなかなかつながらない。大量のテキストデータがあるからと、単にテキストアナリティクスのツールにかけただけで得られるのはこの程度の情報に過ぎない。

そのため、分析対象のデータから何を調べたいか、データのどこに着目すべきか、という観点を分析者が設定する必要がある。そこで、PC ヘルプセンターの顧客対応データに対しては、辞書登録という形で、雑多な表現を『ソフトウェア』、『ハードウェア』、『技術用語』といったカテゴリに分類したり、同義表現を集約したりできるようにした。その結果得られた『ハードウェア』カテゴリの表現の分布が表 2 になる。

表 2　PC ヘルプセンターの顧客対応データに含まれる『ハードウェア』カテゴリの表現

表現	件数
画面	24,599
APTIVA	20,187
電話	13,854
モデム	13,423
電源	12,626
PC	12,438
ハードディスクドライブ	11,703
FAX	8,079
カード	6,425
プリンタ	5,882

表 2 のデータから、例えば、「プリンタ」よりも「モデム」や「ハードディスク」に関する問い合わせが多いといったことが推測されるものの、これだけでもやはり活用が難しい。

より深い分析を目指すにあたり、PC ヘルプセンターのデータにおいて着目する価値が高いと考えられたのが不具合情報である。顧客が困っている状況を把握すれば、何らかの改善につながりそうである。

日本語の場合、不具合は、「〇〇できない」「〇〇しない」「〇〇してしまう」といった語尾を伴う述語表現で示されることが多い。さらにその主語や目的語がハードウェアであれば、ハードウェア系の不具合を示す可能性が高そうである。そこで形態素解析に加えて構文解析（係り受け解析）も適用することで、ハードウェア系の不具合の分布として表 3 のような結果が得られるようになった。

表３　PC ヘルプセンターの顧客対応データに含まれる『ハードウェア』系の不具合関連表現

表現	件数
電源 ... 切れる ... ない（「電源が切れない」等）	398
電話 ... つながる ... ない（「電話がつながらない」等）	312
電源 ... 入る ... ない（「電源が入らない」等）	262
モデム ... 使える ... ない（「モデムが使えない」等）	229
電源 ... 落ちる ... ない（「電源が落ちない」等）	194
電話 ... ある ... ます ... ない（「電話がありません」等）	190
画面 ... 出る ... ない（「画面が出ない」等）	182
モデム ... 応答する ... ます ... ない（「モデムが応答しません」等）	181
モデム ... 応答する ... ない（「モデムが応答しない」等）	178
画面 ... 動く ... ない（「画面が動かない」等）	175

表３の件数には、形態素解析や構文解析（係り受け解析）の出力を活かして、助詞や活用語尾の多様性などを吸収した結果が反映されており、「電源 ... 切れる ... ない」の 398 件には、「電源も切れなくなる」「電源が自動的に切れない」といった表現の件数も含まれている。

表３のデータから、ハードウェア系の不具合としてどのようなものがどの程度発生しているかを把握したうえで、頻度の高い不具合から対応していくようにしようといったアクションにつながりそうである。

しかし、人間の言葉には、「同じ表現で異なる内容を示すことができる」多義性（例えば「コード」で符号や電線を示すことができる性質）に加え、「異なる表現で同じ内容を示すことができる」同義性（例えば「自宅」を「うち」とも表現できる性質）が備わっているという特徴がある。例えば、表３における「電源が切れない」と「電源が落ちない」は同じ内容を示していると推測される。そのため、表３のデータだけからでは、実際にどの内容が最も多いのかが把握できない。それを調べるには、同じ内容の表現を集約する必要があり、それには大きな手間がかかってしまう。

そこで筆者らが見出した方向性が分布の比較による偏りの検出である。PC の機種別にどのようなハードウェア系の不具合が発生しているかを比較できるようにしたのが表４である。

表４　PC ヘルプセンターの顧客対応データにおける機種別の『ハードウェア』系の不具合関連表現の件数（相対頻度）

機種名（データ件数）	電源 ... 切れる ... ない (398)	電話 ... つながる ... ない (312)	電源 ... 入る ... ない (262)	モデム ... 使える ... ない (229)	電源 ... 落ちる ... ない (194)
機種 A (16,446)	20 (0.6)	18 (0.6)	6 (0.1)	14 (0.6)	15 (0.8)
機種 B (9,979)	7 (0.2)	9 (0.4)	4 (0.1)	2 (0.0)	3 (0.1)
機種 C (9,814)	18 (0.8)	17 (1.0)	4 (0.1)	17 (1.3)	11 (0.9)
機種 D (6,814)	18 (1.2)	5 (0.2)	1 (0.0)	20 (2.4)	11 (1.3)
機種 E (6,628)	11 (0.6)	9 (0.6)	4 (0.1)	6 (0.4)	5 (0.3)
機種 F (5,825)	7 (0.3)	10 (0.8)	8 (0.6)	2 (0.0)	2 (0.0)
機種 G (5,337)	2 (0.0)	8 (0.5)	17 (2.1)	5 (0.3)	3 (0.1)
機種 H (4,372)	13 (1.2)	5 (0.3)	6 (0.5)	3 (0.1)	8 (1.0)
機種 I (4,251)	2 (0.0)	3 (0.1)	2 (0.0)	0 (0.0)	5 (0.5)
機種 J (4,233)	0 (0.0)	5 (0.3)	3 (0.1)	1 (0.0)	0 (0.0)
機種 K (4,157)	23 (2.7)	1 (0.0)	9 (1.2)	1 (0.0)	11 (2.1)

表４において、機種 D に関する問い合わせ 6,814 件のうち、20 件のデータにおいて「モデム ... 使える ... ない」という表現が含まれている。「モデム ... 使える ... ない」という表現が含まれているデータは、機種 A では 14 件、機種 C では 17 件存在することから、20 件という件数自体は突出しているわけでない。しかし、機種 A への問い合わせが 16,446 件存在し、機種 C への問い合わせが 9,814 件存在することなどから、機種 D への問い合わせにおける「モデム ... 使える ... ない」という表現が含まれているデータの割合は、他の機種と比較すると高い。

「モデムが使えない」「電源が切れない」といった不具合は、どの機種においても同じように発生する可能性がありそうだが、顧客対応データにおける機種別の出現分布を見ると、発生の割合に偏りが生じていることが分かった。

この発生の偏りを示す指標として表４の各セルの () 内に示しているのが、他の機種と比較して何倍程度多いと判断できるかを示す相対頻度である。筆者らの経験上、他の機種よりも２倍以上高い割合で不具合が発生している場合には何らかの原因が存在する可能性が高い。

表４においてハイライトされている、相対頻度が２

以上のセルを見ることで、機種 D でモデムが使えない不具合や、機種 G で電源が入らない不具合の可能性に気付くことができる。また、機種 K においては、「電源が切れない」という不具合と「電源が落ちない」という不具合の可能性が示されている。「電源が切れない」と「電源が落ちない」は同じ現象と考えられ、システム的な不具合の原因が存在する場合、表現が異なっても同じ不具合に関する件数の割合は同様に高くなる。実際に機種 G において「電源が切れない」と「電源が落ちない」という表現を含むデータを見てみると、どのデータも同じような内容であり、「通常使うプリンタの設定」を変更するようにガイドしていることから、初期設定に問題があった可能性がうかがわれる。

このように分布の偏りをとらえることで、機種に特徴的な不具合を検出できるようになった。こういった気付きを得ることで、不具合に対する対策を迅速に取ることが可能になる。設定変更で済む話であれば、その情報をネットなどで公開することで、顧客のトラブルを減少させることができる。

IBM においては、日米の PC ヘルプセンターにおいて、この仕組みを用いることで、製品不具合の早期発見や、WEB サポートの満足度向上、コール数の削減といった成果を上げることができた。

2.2 テキストアナリティクスの基本的なアプローチ

PC ヘルプセンターの顧客対応データの具体例で示した通り、テキストアナリティクスの難しさは、人間の言葉の多様性、及び、その表現の多義性や同義性から生じる曖昧性にある。言葉の組み合わせは無限であり、同じことを冗長にも簡潔にも表現できる。また、照応表現や省略表現を読み手が補完する必要が生じる場合などもあり、テキストに記述されている情報のみではその内容を正確に解釈できないことも多い。そのため、少なくとも現時点では、機械が人間の言葉を人間以上に正確に扱える見通しは立っていない。

したがって、キーワードや係り受け表現などの情報の抽出を機械的に行った場合、その結果には誤りが含まれることを前提とした分析を行えることが望ましい。

情報抽出結果には誤りが含まれることを前提とすると、情報抽出結果を集計した件数そのものはあてにならないことになる。

例えば、前述の「電源が落ちない」という現象がどれだけの顧客で発生しているかを確認したくても、同じ現象が「電源が切れない」と表現されていたり、「電源が落ちないという問題は発生していない」と否定されていたりする可能性があるため、「電源が落ちない」という表現の出現回数を調べるだけでは不十分である。

そもそも、「電源が落ちない」現象を、PC ヘルプセンターに連絡してこない場合や、オペレータが記録していない場合もある。そのため、ある製品の全ユーザーのうちのごく一部を対象としたサンプリング的な分析を行っていると考えるのが適切である。

サンプリング的な分析であっても、傾向の比較などから有益な知見を得られることは多い。例えば、前述のように、どの PC でも発生し得る問題が、他の機種と比較して、特定の機種で顕著に多く発生していたり、特定の問題が急に増加していたりするなら、そこには何らかの原因が存在する可能性が高い。

人手では読み切れない、膨大なテキストデータから、このような偏りや変化を捉え、そこから気付いた内容を有効なアクションに結び付けることがテキストアナリティクスの重要なポイントである。

テキストアナリティクスで得られるのは気付きであり、気付きの内容を検証するためには、テキストデータ以外の情報が必要となることが多い。例えば PC の不具合であれば、製品を実際に調べてみる必要がある。ハードウェアの不具合であっても、ソフトウェアの問題や操作の難しさなどに起因している可能性があり、顧客対応データの内容のみからでは、その原因まで読み取ることはできない場合が多い。さらに詳細な調査を行う必要性に気付かせてくれるのがテキストアナリティクスの効果であり、より多くの有効な気付きをより早く得られるように工夫することで効果が大きくなる。

3 テキストアナリティクスの発展動向

テキストアナリティクスの技術開発は、分析対象となるテキストデータの多様化や、そこから把握したい内容の多様化に応じて進展してきている。以下では、その多

様化に対応するための取り組みを紹介する。

3.1 評判分析 [4,5]

2000 年代に入って、インターネットの普及率が急速に高まり、ブログやレビューサイトなどで消費者が情報発信するようになると、その情報を活用するための技術が研究されるようになった。例えば、膨大なレビューサイトの情報から、何が高い評価を受けていて、何の評価が低いのかを把握するためには、好不評を示す評価表現に着目する評判分析の技術が必要となる。

評判分析を実現するためには、好不評を示す評価表現を辞書登録して、特定商品名と評価表現との共起関係を調べるだけでは不十分である。「A は良い」であれば好評を示していると解釈できるが、「A は良くない」と否定されていれば評価が反転する。「A が良いとは思えない」のように否定の表現は多様であり、「A は良くなくはない」「A が良くないとは感じない」のように二重否定になることもある。したがって、高い精度で評判分析を実現するためには、言葉と言葉の関係を特定するための構文解析が必要となる。

構文解析を適用することで、商品名など評価対象となる表現と評価表現との関係や、評価表現と否定表現などとの関係が特定できるようになり、好不評の判定が精度良く実現できるようになる。

評判分析の技術を用いることで、例えば購買を検討している商品群に関して、どの商品の何が良くて何が悪いと評価されているかを集計することができる。

このような評判分析も一種の情報抽出であり、完全な精度を期待できるものではない。好不評が表現されていても把握できなかったり、好不評を誤って判断したりするエラーはつきものである。

エラーが含まれていても、膨大なデータを対象として大量の評価情報を収集し、複数の評価対象に関する評価を比較することができれば、相対的にどれが最も良さそうか、特定の評価対象が、どのような点で好不評と評価されているかといった特徴を捉えることができるようになる。

但し、精度があまりに低いと役に立たない。例えば、否定表現の解析ミスなどによって好不評を逆に捉えてしまう割合が高かったり、好不評の言及対象を誤って認識していたりすると、集計結果があてにならなくなってしまう。

そのため、テキストアナリティクスでは、情報抽出結果と抽出元の原文との紐づけを保持しておき、適宜原文を参照し、抽出結果の妥当性を確認できるようにしておくことが重要である。

3.2 会話分析 [6]

前述の PC ヘルプセンターの顧客対応データは、オペレータがキーボードで入力したテキストであった。人手によるテキスト化は、労力がかかる上に、有益な情報が必ずしも入力されていない可能性がある。そのため、筆者らがテキストアナリティクスに取り組み始めたころから、自動音声認識技術を用いて、顧客との会話内容をすべてテキスト化して分析したいという要望が存在した。

近年、深層学習によって音声認識技術が進化し、発話内容が高い精度でテキスト化されるようになってきたことから、コールセンターにおける顧客とのやり取りを全てテキスト化して活用しようという動きが活発化している。

やり取りの全てをテキスト化した結果は、一見すると冗長なテキストになることが多い。言い淀みなどに加え、挨拶や雑談など、分析対象としてはあまり意味をなさないように思える文字列が含まれることになる。それでも、どの情報が意味を持つかは分析してみないと分からない。

例えば、筆者らが米国のレンタカー会社の予約業務のコールセンターの顧客対応の録音データを書き起こしテキスト化して分析した事例では、電話をかけてきた顧客の第一声が、「車を借りたい」「予約したい」のようにレンタル希望を伝えているか、あるいは、「値段を知りたい」「価格を教えて欲しい」のように代金を問い合わせているかのどちらかに分かれ、代金の問い合わせで始まる顧客は予約しても無断キャンセルする割合が高いことが分かった。さらに、代金の問い合わせで始まった顧客でも、何らかの会員サービスなどで割引を適用されるとキャンセルせずに車を借りてくれるという有益な知見を得ることができた。

最初の発言が要望か問い合わせかどうかという区別は従来オペレータが意識していなかったことであり、それ

を意識して、代金の問い合わせに対しては積極的に割引適用を図るといった形で、会話分析で得られた知見を、ビジネスの向上につながるアクションにつなげることができた。

3.3　マルチモーダル分析

テキストデータには図や写真など画像データと結び付いているものも多い。例えば、ソーシャルメディア上では、画像と共にテキストが書き込まれるケースが多くなっている。このような画像付きのテキストデータに対しては、深層学習で精度が向上した画像認識を適用し、画像に何が写っているかの情報を活用した分析が可能になってきている。

例えば、特定の商品ロゴが写っている写真と紐づいたテキストデータを対象とすることで、その商品に対するイメージや意見の分析が可能になる。

このように音声や画像などの情報を取り込む形で分析対象を拡大する方向性に加えて、テキストを異なる観点からより深く分析するような方向性も存在する。

その一つの方向性が、書き手の性別や年齢層などの情報を読み取るプロファイル分析である。書き手の属性を把握したうえで評判分析などを行えば、商品開発やマーケティングにより有効な知見を得ることができるようになる。

また、心理学の分野では人間の基本的な性格の次元が下記の５つであるというビッグ・ファイブ理論[7]が確立され、人の性格が５次元空間の一点として数値化できるようになり、世界的に共通の枠組みで、性格に関する科学的研究が進められるようになっている。

- Openness to Experience（経験への開放性）
- Conscientiousness（誠実性）
- Extraversion（外向性）
- Agreeableness（協調性）
- Neuroticism（情緒安定性）

その結果、書かれたテキストに筆者の性格が反映されることが分かってきており[8]、テキストから筆者の性格を推定する仕組みが作られてきている[9]。この仕組みを用いることにより、どのような性格の人が書いたテキストであるかを考慮して、例えば、性格が似ている人の特徴を分析するようなことも可能になってきている。

4　テキストアナリティクスを用いた特許情報処理[10]

冒頭に記した通り、特許公報は早い時期から電子化が進んでいたため、1990 年代にテキストアナリティクスの取り組みが本格化した当初から、その適用対象として、様々な試みが行われてきた。

２節で示した基本的な仕組みを用いるだけでも、例えば、出願人別の特許の特徴や傾向を把握することができるようになる。特許データ全体における名詞の出現分布と、特定の出願人が出願した特許データにおける名詞の出現分布を比較することで、その出願人がどのような分野の技術に力を入れているかが分かる。さらに、その時間的な傾向を分析すれば、力を入れている分野の変化している様子も見えてくる。また、特定技術に関しては、どのような出願人が力を入れているかなど、競合関係や補完関係になる可能性のある出願人の分析などに応用することが可能になる。

重要特許の分析という観点から、全国発明表彰[11]で表彰された特許を調査対象として、産業界における業界内外で重要性が高いと評価されている特許の出願方法を調べた結果、重要特許は、近い出願日で複数の類似特許の固まりとなって出願される可能性が高いという傾向が見出されており、このような出現傾向を捉えることで、出願人が力を入れている特許を認識することが可能になってくる[10]。

特許情報に特化した分析としては、情報抽出という観点から、特長表現の抽出[12]を挙げることができる。ある特許技術によって何が可能になるのかを把握するため、「既存の技術、または既存の類似した製品・商品と比較して改善される事柄、または新たに実現される好ましい事柄を示した表現」を特長表現として抽出する取り組みであり、「〇〇を向上する」「〇〇を高める」といったパタンを用いて、「光の利用効率を高める」といった表現を抽出したり、「〇〇を防止する」「〇〇を抑制する」といったパタンで「画像の劣化を防止する」といった表現を抽出したりできるようになる。

このような特長表現を抽出し、一覧できるようにすることで、特定の技術課題の解決につながる特許を発見することが容易になる。

また、特許請求項の構造解析を用いて、請求項から発

明の新規性や進歩性に関わるキーワードを精度良く抽出できるようにもなってきている。審査請求後、一旦新規性・進歩性がないという拒絶理由のみで拒絶され、補正後に最終的に登録されたという経過情報を持つ特許において、公開公報と登録特許を比較し、登録特許で追加されたキーワードには新規性・進歩性が示されているものと仮定し、新規性・進歩性に関するキーワードの近似的な正解データを大量に収集した結果として、抽出精度の妥当性が検証できるようになっている[13]。このような新規性・進歩性に関わるキーワードの抽出は，請求項の可読性を高めるだけでなく、類似特許検索や、特許の評価への応用も期待できる。

5 おわりに

以上のように、従来、人が読んで解釈するしかなかったテキストデータを計算機で大量に処理し、新しい活用方法につなげる取り組みが進展している。個々のテキストを解読して判断する能力に関しては、機械が人間の能力を上回る見通しは立っておらず、エラーの発生を前提として利用することが重要である。テキストアナリティクスを利用するにあたっては、人が行っていた作業を置き換えるのではなく、今までできなかった処理を実現することが大きな成果につながるというのが筆者の考えである。

なお、本稿に記載されている内容は、全て筆者個人の見解に基づいている。

参考文献

[1] Ronen Feldman、James Sanger（著）/ 辻井潤一（監訳）/ IBM 東京基礎研究所テキストマイニングハンドブック翻訳チーム（訳）、テキストマイニングハンドブック、東京電機大学出版局、2010

[2] 那須川 哲哉（著）、テキストマイニングを使う技術 / 作る技術 - 基礎技術と適用事例から導く本質と活用法 -、東京電機大学出版局、2006

[3] 菰田 文男（編集）、那須川 哲哉（編集）、ビッグデータを活かす 技術戦略としてのテキストマイニング、中央経済社、2014

[4] Tetsuya Nasukawa and Jeonghee Yi, "Sentiment analysis: Capturing favorability using natural language processing," Proceedings of the 2nd international conference on Knowledge capture, ACM, p.70-77, 2003.

[5] Hiroshi Kanayama, Tetsuya Nasukawa, and Hideo Watanabe, "Deeper sentiment analysis using machine translation technology," Proceedings of the 20th international conference on Computational Linguistics, p.494-500, 2004.

[6] 竹内 広宜、那須川 哲哉、渡辺 日出雄、" コールセンターにおけるビジネス会話のマイニング "、人工知能学会論文誌、Vol.23, No.6, p.384-391, 2008.

[7] Lewis R. Goldberg, 'An alternative "description of personality": the big-five factor structure,' Journal of personality and social psychology, 59.6: p.1216-1229, 1990.

[8] Francois Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text," Journal of Artificial Intelligence Research, 30: p.457-500, 2007.

[9] 那須川哲哉、上條浩一、" 日本語における筆者の性格推定の取組み "、言語処理学会第 23 回年次大会予稿集、p.807-810, 2017.

[10] 鈴木祥子、那須川哲哉、" 特許品質評価及び特許からの情報抽出における自然言語処理のアプローチ "、月刊パテント　2016 年 12 月号 , Vol. 69 No. 15, p.19-23.

[11] 全国発明表彰 . http://koueki.jiii.or.jp/hyosho/zenkoku/zenkoku.html.

[12] 西山 莉紗 、竹内 広宜 、渡辺 日出雄、那須川 哲哉、" 新技術が持つ特長に注目した技術調査支援ツール "、人工知能学会論文誌、Vol.24, No.6, p.541-548, 2009.

[13] Shoko Suzuki and Hiromichi Takatsuka, "Extraction of Keywords of Novelties from Patent Claims," Proceedings of the 26th International Conference on Computational Linguistics, p.1192-1200, 2016.