【論　文】

製薬特許の分析による 2 つの知見
―医薬品の売上予測とパテントファミリー内の特許成立の予測

上條浩一・大関和夫

Two Findings from the Analysis of Pharmaceutical Patents
-Sales Estimation and Patent Grant Prediction within Patent Family

Koichi Kamijo, Kazuo Ohzeki

Abstract：We report two cases of findings obtained by analyzing the specification of pharmaceutical patents using machine learning. The first is about pharmaceutical sales estimation. Many pharmaceutical companies take a huge amount of time and money to develop a pharmaceutical, and it is strongly desired to estimate future sales volume at an early development stage. The second is about the patent family. Many pharmaceutical companies apply for the same patent in multiple countries as a patent family, but it is often the case that the same patent is granted in one country but not in another. In both cases, it turns out that deep learning provides good predictive performance by using the words used in the patent and the “hot” words that were also published in the journal at the time of filing.

Keywords：sales estimation, pharmaceuticals, patent specification, deep learning, natural language processing.

Ⅰ．はじめに

医薬品は他の製品よりも開発期間が長く、多くの製品が大きな利益を生む可能性がある。そのため、特に医薬品のマーケティング戦略を策定する際には、開発の初期段階で新製品や新サービスの販売量を見積もることが非常に重要となる。

一方、多くの企業が複数の国で同じ特許を申請している。これは、商品やサービスの権利をグローバルに取得することを目的として行われており、これらの特許はパテントファミリーと呼ばれる。しかしながら、各国の特許庁の方針の違いなどにより、同じ特許であるにも関わらず、ある国では成立するが、他のある国では成立しない、ということがしばしば起こる。

本論文では、対象を製薬に限定し、特許明細書の分析による 2 つの知見の報告を行う。1 つ目は、医薬品の売上予測であり、2 つ目は、パテントファミリー内の国毎の特許成立の予測である。2 つ目に関しては、対象国を米国、インド、およびブラジルに絞る。これらの 2 種類の予測に関しては、何れも深層学習を用いる。

Ⅱ．関連研究

ここでは、2 つの知見に共通となる、特許の解析に関する関連研究を紹介する。

Suzuki らは、特許申請書類の構造を利用して、特許クレームから新規性または進歩性に関連するキーワードを自動的に抽出する方法を提案している [1]。Kim らは、特許申請書類を分析して、ワイヤレス電力伝送の新しい技術分野と、未開拓の技術分野を特定する。彼らは、テキストマイニングによって特許からトピック領域を抽出し、さらに、時系列分析により類似のセマンティクスを持つトピックをグループ化した [2]。次に、クラスタリングおよび時系列分析の結果を比較して、新規または未開拓のテクノロジー領域を誤認識する可能性を最小限に抑えた。Guderian らは、COVID-19 パンデミックなどの際に、特許を含めた公開データを用いて、如何にして有益な情報を提供できるかに関しての調査を行った [3]。

これらは全て、特許分析によるものであるが、売り上げ予測やパテントファミリーとしての特許予測を行うものではない。また、何れの物も、深層学習は利用していない。

Ⅲ．特許の扱いの国ごとの違い

本章では、2 つ目の知見である、パテントファミリー内の国毎の特許成立違いを解析する準備として、まず、3 つの各対象国の特許庁の特徴を分析し、次に、各国の特許における特徴に対しての定量的な比較を行う。

A．特許庁の比較

特許庁の方針は国によって異なる。たとえば、米国特許庁（United States Patent and Trademark Office=USPTO）の特徴として、2011 年の America Invents Act に従って、USPTO が特許の優先日を決定する際に先願主義を採用していることが挙げられる [4], [5]。USPTO のもう 1 つの特徴は、経験豊富な審査官が特許を付与する可能性が高いことが挙げられる [6]。インドとブラジルは Trade Related aspects of Intellectual Property Rights（TRIPS）[7], [8] に準拠しており、各国は国の状況に応じて柔軟な対策を講じることができる。インドの特許法は、“evergreening”、つまり法律で一般的に許可されている期間よりも長い期間特許を延長することを防ぐことを目的とした Section 3(d) の存在を特徴としている。“evergreening” を防ぐためのこの対策は、“public health safeguard” と広く見なされている [9]。セクション 3(d) は、既知の物質の新しい形態の単なる発見は、有効性に関して特性が大幅に異なる場合を除いて、特許を受けることができないということを示している。Agência Nacional de Vigilância Sanitária（ANVISA：国家衛生監督庁）による審査は、ブラジルの特許法の特徴である [10]。ANVISA は、ブラジル特許庁による出願の審査の前に、医薬品に関する特許出願を審査するが、これにより、2 段階の審査プロセスが発生し、出願から特許が付与されるまでの待ち時間が長くなる。

これらの違いを考えると、同一の特許が出願されている場合でも、特許成立の比率が米国、インド、ブラジルで異なることが理解できる。

B. 特徴量に伴う特許成立率の比較

1999 年から 2014 年の間に少なくとも米国、インド、ブラジルで同じ特許が出願された、医薬品に関するパテントファミリー 12,499 件を、Cortellis [11] および Derwent Innovation [12] データベースより収集した。つまり、12,499 × 3 = 37,497 件の特許を収集したことになる。因みに、これらの 2 つのデータベースは、特許に関するエクスパートが多く在籍している Clarivate 社 [13] により運営されている。

実験を行う前に、これらの 37,497 の特許に対して、“ 成立 ”、“ 拒絶 ”、“ 不明 ” の何れかのラベルを付与する。“ 成立 ” は、実際に国の特許庁によって “ 成立 ” となった場合、またはまだ審査中のものであっても、Clarivate が成立の確率を 100% と判断した場合に付与される。“ 拒絶 ” は、実際に国の特許庁によって “ 拒絶 ” となった場合、または、まだ審査中のものであっても、Clarivate 社が成立の確率を 0% と判断した場合に付与される。また、審査中の特許で、出願より、その国の特許の審査の期間が、その国の承認された特許の 95% 以上が承認が終了している期間を超えた場合も、その特許には “ 拒絶 ” のラベルを付与する。それ以外の審査中の特許は、“ 不明 ” とする。

Fig.1 は、パテントファミリー内の 37,497 の特許に関して、“ 成立 ”、または “ 拒絶 ” のラベルがついたものに対して、各国において “ 成立 ”、または “ 拒絶 ” となった特許における、平均発明者数、平均クレーム数、平均被引用特許・文献数、及び平均引用特許・文献数を表したものである。ここで、これらの 4 つの平均値に関しては、米国に出願されたものを用いた。これは、それ以外の国におけるこれらの 4 つの特徴量が取得できるもの、できないものが混在しているためである。

さらに、この後で述べる、特許明細書の解析においても、米国の明細書を、そのパテントファミリーの代表として用いた。これは、本研究が、出願された内容に対応した特許成

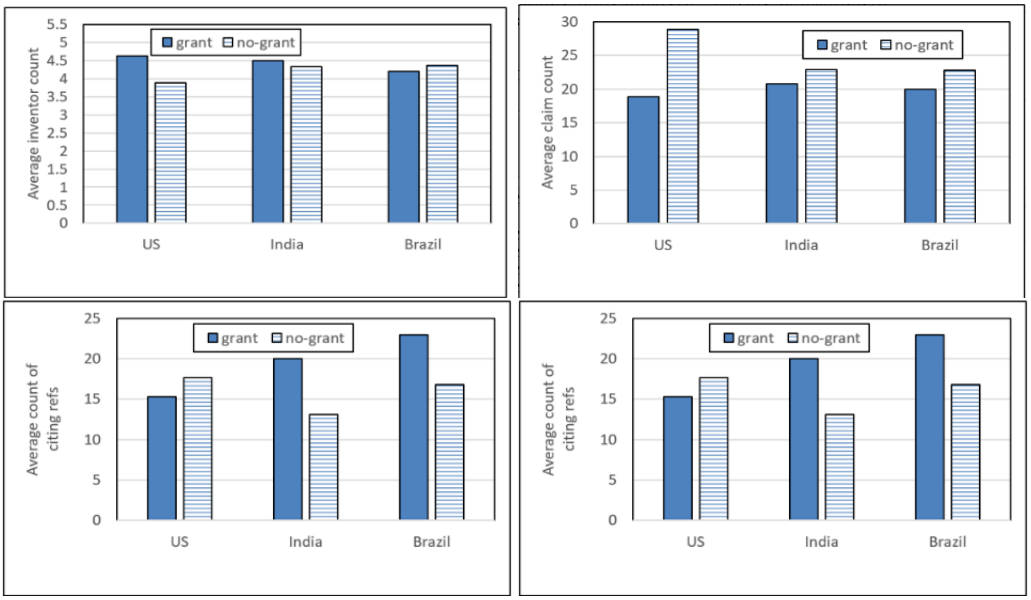

Fig. 1　12,499 のパテントファミリを利用した、特許成立可否の国ごとの特徴量の比較
左上：平均発明者数、右上：平均クレーム数、左下：平均被引用文献数、右下：平均引用文献数。

立可否の解析を行うことが目的であり、各国における、明細書の違いから生じる成立可否の違いを解析するものではないからである。また、ブラジルの特許明細書はポルトガル語で書かれており、それを翻訳して使うことにより、本研究の趣旨ではない、翻訳エンジン自体のノイズが加わることも理由として挙げられる。

これらの図を見ると、特に平均被引用特許・文献数において国毎の特徴が認められ、特に米国において、成立特許の場合に大きく、拒絶の場合小さいことが判る。

Ⅳ．予測モデル

予測モデルにおいては、医薬品に関連する特許や記事をもとに、新薬の売り上げ、及びパテントファミリーの特許の成立の可否の予測を行う。Fig.2 は予測モデルの概要である。以下に各機能の説明を行う。Sales data は、医薬品の品名、開発会社名、売り上げ実績などの医薬品の売り上げに関するデータを蓄積するデータベースであり、Cortellis のデータベースを利用する。このデータベースは、売り上げ予測のみで利用される。Patents は、医薬品に関連する特許申請書類を蓄積するデータベースであり、Derwent Innovation の特許データベースを利用する。Articles は、医薬品関連の記事のデータベースであり、具体的には、“Pharmaceutical Benefits Pricing Authority Annual Reports published” (1998 - 2020) [14], [15] および “Reports of the Pharmaceuticals and Medical Devices Agency” (2004 - 2018) [16] を格納している。**Morphological analysis-1,2** では、**Patents** に格納された特許申請書、及び Articles に格納された記事の形態素解析（stemming, lemmatization を含む）を行う。具体的には、nltk Package [17] に格納された “word tokenize” を用いて、形態素解析を行う。これを用いることにより、例えば、“We were performing maintenance. It rains cats and dogs.” という文章は、“We were perform mainten . It rain cat and dog .” という文章に変換される。ここでは、大文字、小文字の区別は行わず、数字やストップワードは無視される。**Word count** では、形態素解析された単語の利用頻度が、各特許申請書毎にカウントされる。**TF-IDF**（Term Frequency Inverse Document Frequency）では、全記事の形態素解析された単語の年ごとの TF-IDF が計算され、**Weighting** で、**Word count** で計算された単語数の重みづけが年毎に行われる。**Deep learning** では、上記で得られた情報を深層学習で解析し、**Cross-validation** で、交差検証により、これらの結果の評価を行う。

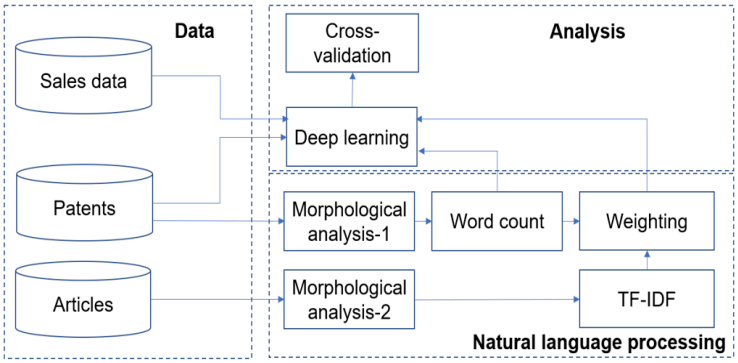

Fig. 2　予測モデル

以下で、“ 売り上げ予測 ”、“ 特許成立予測 ” の各々の手法に関して、詳細に議論を行う。

A. 売り上げ予測モデル

売り上げ予測モデル [18] においては、以下の 1) ～ 4) の 4 つのアプローチに加え、それらのコンビネーション (5)) を用いて売り上げ予測を行った。

本論文での売り上げ予測においては、ある年の単年度での売り上げではなく、対象製薬の累積売り上げの予測を行う。

1) 医薬品の開発会社、及び開発年からの売り上げ予測
2) 医薬品の特許情報からの売り上げ予測
3) 医薬品の特許申請書類に使われた単語からの売り上げ予測
4) 医薬品関連記事を用いた売り上げ予測
5) 1) ～ 4) のコンビネーション

医薬品を d_i,（$i = 1, 2, ..$ はインデックス）、その実際の売り上げの合計値を s_i とする。医薬品リスト、及びその売り上げは、Cortellis のデータベースから得られたものを利用する。Fig. 3 は、実験で使用された 439 の医薬品の医薬品売り上げ合計を示している。この売上合計は、医薬品が販売されてから 2019 年までの累計に加え、2020 年から 2027 年までの、Clarivate 社の専門家による売り上げ予測の累計を加えている。これは、2019 年の直前または直後に開発された医薬品の売上実績データが不足しているためであり、このようにすることで、各薬の生涯売り上げに近い値が得られる。このように一部予測値が入ったものを深層学習で予測、評価することになるが、売上予測研究の観点から、これらの値を推定することは研究としての価値を損なわない。この図を見てわかるように、売り上げは指数関数グラフに近い形をしているため、本論文では、売り上げ予測に対しては、s_i ではなく、$\log(s_i)$（底 $=e$）を使用する。申請書に関しては、特許フォーマットの統一の観点より The Patent Cooperation Treaty（PCT）[19] に準拠した英語で書かれた特許申請書のみを使用した。以下に、1) ～ 4) の詳細を述べる。

1) 医薬品の開発会社、及び開発年からの売り上げ予測

COVID-19 により、製薬会社の売り上げに大きな差がつき、例えば、ファイザー社は 2021 年の COVID-19 ワクチンの需要が旺盛で、売上高は 260 億米ドルと予測されている [20]。2021 年は特別としても、このように、製薬会社によって薬の販売量は異なり、製薬会社から、薬の売り上げ予測がある程度可能であることが推定される。ここでは、医薬品を開発した会社の名前と、開発の開始年より、その薬の売り上げを予測する。開発の

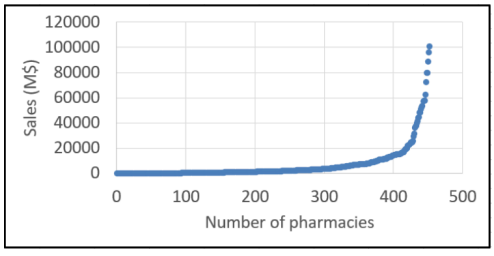

Fig. 3　実験で用いられた 439 の医薬品の累計売上（予測値を含む）（単位：百万米ドル）

開始年は、その薬に対する最初の特許が出願された年を利用する。売り上げ予測モデルの **Deep learning** への入力としては、d_i に対して、製薬会社名を one-hot vector v_i で定義し、さらに、製薬の開発年 yp_i を用いて、以下の $x_i^{[1]}$ を利用する。

$$x_i^{[1]} = [v_i, yp_i]^T . \tag{1}$$

ここで、^{T} は転置を意味する。

2)　医薬品の特許情報からの売り上げ予測

特許明細書は通常、タイトル、要約、クレーム、詳細な説明などで構成される。これらのコンポーネントと販売量の相関関係を分析し、売り上げ予測行う。具体的には、まず、医薬品 d_i に対して最初に出願された特許申請書 a_i より以下の情報を取得する。

- pr_{1i}：タイトルの単語数
- pr_{2i}：アブストラクトの単語数
- pr_{3i}：明細書全体の単語数
- pr_{4i}：クレーム数

これらを用い、医薬品 d_i に対し、以下のベクトル $x_i^{[2]}$ を定義し、**Deep learning** への入力とする。

$$x_i^{[2]} = [pr_{1i}, pr_{2i}, pr_{3i}, pr_{4i}]^T . \tag{2}$$

3)　医薬品の特許申請書類に使われた単語からの売り上げ予測

各特許明細で使用されている単語と各医薬品の売り上げの相関関係より売り上げ予測を行う。具体的には、医薬品 d_i の特許明細書 a_i で使用されている全ての形態素解析後の単語 w に対して、形態素解析を実行したものに対し、各単語の使用率、すなわち、各単語

TABLE I

実験で使用された 439 件の特許の $|r|$ 値が大きい単語のサンプル、および各単語の 439 件の特許のうち少なくとも 1 回使用された特許の比率。

word	*r*-value	ratio
lupu	−0.304	0.103
germlin	0.305	0.128
ctg	0.301	0.103
dmso	0.292	0.155
transplant	−0.285	0.109
amino	0.284	0.123
vegf	0.290	0.098
isoleucin	0.259	0.114
lymphoma	0.269	0.196
epiderm	−0.275	0.105
cag	0.273	0.106
cynomolgu	−0.273	0.105
best	−0.266	0.128
precipit	−0.263	0.121
coloni	−0.243	0.146
microtit	−0.241	0.128
substrat	0.230	0.105
polynucleotid	−0.232	0.221

の出現頻度をその明細書での全ての単語数で割ったもの、u_{wi}（$\sum_w u_{wi}= 1$）、を算出する。ここでの算出においては、本章上記に既に述べた通り、ストップワード、数字、記号を除外したが、品詞や、出現場所（タイトル、アブストラクト、図、等）に関係なく、算出を行った。そして、各単語 w に対して、u_w を次のように定義する。

$$u_w = [u_{w1}, .., u_{wn}]. \tag{3}$$

ここで n は対象となる医薬品数（明細書数）の総数である。次に、u_w と $l_s = [\log(s_1), .., \log(s_n)]$ の間のピアソンの r 値、r_w, を求め、以下を満足する単語の集合 $\Omega(Tr, Tp)$ を作成する。

$$\Omega(T_r, T_p) = \{w||r_w| \geq T_r, p_w \leq T_p\}. \tag{4}$$

ここで、T_r、T_p は各々 r 値、p 値の閾値である。Table I は、実験で使用される 439 の医薬品特許明細書の中から、$\Omega(0.1, 0.01)$ を満たす医薬品用語を $|r_w|$ の大きい順にソートしたものである。ここで、“ 比率 ” は、各単語が、439 件の特許のうち少なくとも 1 回使用された特許明細書の比率を表す。

ここで、各医薬品 d_i に対して、以下のベクトル $x_i^{[3]}$ を定義し、**Deep learning** への入力とする。

$$x_i^{[3]} = [u_{w1i}, .., u_{w|\Omega|i}]^T. \tag{5}$$

4) 医薬品関連記事を用いた売り上げ予測

特許出願書類に、ある年の医薬品の注目度を反映した「旬」なキーワードが含まれている場合、将来的な医薬品の販売総数が大きくなる可能性がある。そこで、3) で計算された単語使用率 u_{wi} を「旬」の具合に基づき重みづけをした合計値を各年毎に計算し、売り上げ予測に利用する。重みづけに際しては、y 年に発行された医薬品関連の記事 [14]–[16] を収集し、形態素解析後、そこで利用された単語 w に対する tf-idf, tfidf(w, y) を以下のように定義する。

$$tfidf(w, y) = tf(w, y)\ log(idf(w)). \tag{6}$$

ここで、tf(w, y) は、y 年に発行された全ての対象記事での w という単語の出現頻度を表し、idf(w) は、対象となるすべての年のうち、w が 1 回以上 1 つ以上の記事で利用された出現回数（年の数）の逆数を表す。各 y に対し、tf(w, y) は正規化されており、Σ wtf(w, y) = 1 が成立する。

次に、各医薬品 di に対して、y 毎の、単語使用率を tf-idf で重みづけした合計、ut (y, i) を以下のように計算する。

$$ut\ (y, i) = \Sigma_w \text{tfidf}(w, y)u_{wi}. \tag{7}$$

$ut(y, i)$ を年毎に並べたベクトルを深層学習の入力に入れ学習させることにより、特許が申請されてからの年数に応じた重みづけを行い、その特許の売り上げを予測することが可能となる。その際、i に対する各ベクトルを、医薬品の記事に対し、医薬品の特許が最初に申請されてからの年数で揃える必要がある。この観点から、ut (y, i) に対し、記事が公開さ

れた年から最初の特許が出願された年（yp_i）を引いた年の各要素の位置がすべての i で同じになるように、ベクトルの左、右、もしくはその両方に 0 をパディングし、以下のベクトル $x_i^{[4]}$ を定義し、**Deep learning** への入力とする。このようにパディングを行うことにより、各特許の出願年に関わらず、特許が出願されてから y年経った年の $x_i^{[4]}$ における要素の位置が左から $yp_{max} - ya_{min} + y'$ と同じ位置になる。単に年次順ではなく、このように、出願してからの経過年順でベクトルの要素をそろえた理由は、年次順の場合に比べ、経過年順で各 $x_i^{[4]}$ を揃えた方が、推定精度が高いためである。

$$x_i^{[4]} = [0^{z1_i}, ut(ya_{min}, i), .., ut(ya_{max}, i), 0^{z2_i}]^T . \tag{8}$$

ここで、0^j は、j 個の 0 で構成されるベクトル、$z1_i = yp_{max} - yp_i$、$z2_i = yp_i - yp_{min}$、yp_{min}、yp_{max} は各々 yp_i の最小値と最大値、ya_{min}、ya_{max} は各々記事の最も古い年と最も新しい年である。
例えば、$yp_0 = 2000$, $ya_{min} = 1998$, $ya_{max} = 2020$, $yp_{min} = 1980$, $yp_{max} = 2021$ の場合、

$$x_0^{[4]} = [0^{21}, ut(0, 1998), .., ut(0, 2020), 0^{20}]^T , \tag{9}$$

となる。

B. パテントファミリにおける特許成立予測

パテントファミリにおける特許成立予測モデルにおいては、以下 2) 〜 4) に加え、それらのコンビネーション (=5)) を用いて特許成立予測を行った。下記においては、売り上げ予測の 1) を除いたものと基本的なアプローチが同じのため、売り上げ予測場合と同じ番号に揃える。ただし、売り上げ予測と区別するために、ベクトル等に全て′をつけて区別する。

2) 医薬品の特許情報からの特許成立予測
3) 医薬品の特許申請書類に使われた単語からの特許成立予測
4) 医薬品関連記事を用いた特許成立予測
5) 2)-4) のコンビネーション

以下、各項目の説明を行う。2) 医薬品の特許情報からの特許成立予測売り上げ予測同様、特許申請書 $a'i$ の特徴量を利用するが、Fig.1 で紹介した以下の内容を用いる。

- pr'_{1i}：発明者数
- pr'_{2i}：クレーム数
- pr'_{3i}：被引用文献数
- pr'_{4i}：引用文献数

これらを用い、医薬品 d'_i に対し、以下のベクトル $x_i'^{[2]}$ を定義し、**Deep learning** への入力とする。

$$x_i'^{[2]} = [pr'_{1i}, pr'_{2i}, pr'_{3i}, pr'_{4i}]^T . \tag{10}$$

3) 医薬品の特許申請書類に使われた単語からの特許成立予測

ここにおいては、売り上げ予測同様、各特許明細で使用されている形態素解析済みの単語 w の利用頻度を用い、各単語 w に対する利用頻度を $u'_w = [u_{w1}, .., u_{wn'}]$ と定義する。ただし、n'は、パテントファミリにおける特許成立予測で利用するパテントファミリの数である。

次に、各パテントファミリ i, c に対して以下の sc, i を定義する。

$$s_{c,i} = \begin{cases} 1, & a'_i \text{ は } c \text{ では成立、他の全ての国で拒絶} \\ 0.5, & a'_i \text{ は } c \text{ と他の 1 つの国のみで成立} \\ 0, & a'_i \text{ は全ての国で成立、もしくは拒絶} \\ -0.5, & a'_i \text{ は } c \text{ では拒絶、他の 1 つの国のみで成立} \\ -1, & a'_i \text{ は } c \text{ では拒絶、他の全ての国で成立} \end{cases} \tag{11}$$

ここで、例えば、特許成立の可否に貢献している単語を選ぶ、という目的であれば、単純に、各単語の 3 つの国の成立／拒絶の数に応じて重みづけを行う方法も考えられる。しかしながら、本論文では、パテントファミリ内で、国毎の成立 / 拒絶の差を生む可能性が高い単語を選ぶ目的で、式 11 の方法を採用した。

そして、$s'_c = [s_{c,1}, .., s_{c,n'}]^T$ とし、s'_c と u'_w との間のピアソン値 r'_{wc} を、w が国 c での成立 / 拒絶の予測にどの程度関与しているかを示す値として計算する。さらに、以下の式を満たす w の集合 $\Omega(T'r, T'p)$ を選出する。

$$\Omega\ (T'r, T'p) = \{w \| r'_{wc}| \geq T'r, p'_w \leq T'p, c = US, IN, BR\}. \tag{12}$$

ここで、$T'r$ と $T'p$ はしきい値、p'_w は r'_{wc} に対応する p 値、US, IN, BR は各々米国、インド、ブラジル、である。このように単語を選択することにより、国 c において特許成立に関与した単語には大きい、拒絶に関与した単語には小さい（負の絶対値の大きい）r 値の単語を選ぶことができる。次に、売り上げ予測同様、特許明細書 $a'i$ ごとに、Ω'内の単語の使用率をリストするベクトルを定義する。

$$x'^{[3]}_i = [u_{w1i,}, .., u_{w_{|\Omega'|}i}]^T. \tag{13}$$

4)　医薬品関連記事を用いた特許成立予測

売り上げ予測の時と同じ医薬品関連の記事を用い、各特許ファミリ i に対し、年毎の「旬」の言葉の利用頻度に基づき重みづけをした合計値、$ut'(y, i)$ の計算を行う。また、売り上げ予測同様、これらの値を年毎に並べたベクトルを **Deep learning** の入力に用いる。ただし、利用する特許文書が異なるため、まず、使用される特許の申請年の最小値、最大値を yp'_{min}、yp'_{max} とする。また、0^j を、j 個の 0 で構成されるベクトル、$z1'_i = yp'_{max} - yp_i$、$z2'_i = yp_i - yp'_{min}$ とし、ベクトルの左、右、もしくはその両方に 0 をパディングした、以下のベクトル $x'^{[4]}_i$ を定義し、**Deep learning** への入力とする。

$$x'^{[4]}_i = [0^{z1'_i}, ut'(ya'_{min}, i), .., ut(ya'_{max}, i), 0^{z2'_i}]^T. \tag{14}$$

Ⅴ．実験

売り上げ予測、特許成立予測に関して、各々実験を行った。評価は交差検証で行った。

A. 売り上げ予測

前章までで議論した各方法論の売上予測の評価を行った。実験では、医薬品 d_i に対する販売量（s_i）と最初に申請された特許明細書（a_i）の両方が利用可能な $n = 439$ の医薬品を使用する。$n = 439$ は十分に大きい数ではないため、Leave one out (LOO) 交差検定（$n - 1$ を学習データに用い残り 1 つを推定することを n 回の全ての組み合わせで行い、その平均値を計算）で評価を行った。推定モデルでは、2 つの隠れ層を用い、各々 128 ノードで、relu activation を用い、エポック数 = 100 で実験を行った。

損失には平均二乗誤差を使用し、オプティマイザーには Adam を使用した。入力ベクトルは、one-hot vector を除き、z-normalization で正規化を行った。

各 d_i に対し、学習データ、テストデータ (X_train, y_train), (X_test, y_test) は以下のようにして実験を行った。

$$\text{(X\_train, y\_train), (X\_test, y\_test)} = (X_{\neg i}^{[k]}, ls_{\neg i}), (x_i^{[k]}, \log(s_i)), \tag{15}$$

ただし

$$\begin{aligned} X_{\neg i}^{[k]} &= [x_1^{[k]}, .., x_{i-1}^{[k]}, x_{i+1}^{[k]}, .., x_n^{[k]}]T, \\ ls_{\neg i} &= [\log(s_1), .., \log(s_{i-1}), \log(s_{i+1}), .., \log(s_n)]^T. \end{aligned} \tag{16}$$

ここで、$k = 5$ は、$k = 1$ から $k = 4$ の全てのベクトル結合したベクトルを表す。実験では、(16) のモデルを各 i で n 回構築し、以下で計算される root mean square error (RMSE) と mean absolute error (MAE) により評価を行った。

$$\begin{aligned} \text{RMSE} &= (\Sigma_{i=1}^{n}(\hat{ls}_i - \log(s_i))^2/n)^{0.5}, \\ \text{MAE} &= \Sigma_{i=1}^{n}|\hat{ls}_i - \log(s_i)|/n. \end{aligned} \tag{17}$$

ここで、$\hat{ls}_i$ は、(15) の入力に対するモデルによる売り上げ推定値である。外れ値に関しては、$\hat{ls}_i$ を $\max(\min(\hat{ls}_i, \max_{j\neq i}(\log(s_j))), \min_{j\neq i}(\log(s_j)))$ とする。また、推定値と実際の値の比率がある閾値 T_c 以下である医薬品 i の個数 $r_c(T_c)$ の比率の評価も行った。$r_c(T_c)$ 以下の式で定義される。

$$r_c(T_c) = |\{i||\hat{ls}_i - \log(s_i)| \leq T_c\}|/n. \tag{18}$$

Table II に実験結果を示す。“input(k)” 列の値は、$x_i^{[k]}$ の k に対応し、$Tc = \log(2)$ とする。これは、実際のデータと推定データの比率が 0.5 ～ 2.0 であることを表す。

TABLE II
実験結果（売り上げ予測）：太字 = 最高値。

No.	Input(k)	RMSE	MAE	r_c
1	1	2.324	1.893	0.206
2	2	2.646	2.121	0.211
3	3	1.903	1.489	0.288
4	4	2.045	1.634	0.265
5	1,3	1.812	1.408	0.319
6	1,3,4	**1.726**	**1.360**	**0.326**

No.1-4 は、各ベクトル単独での実験結果である。No. 5 は、k = 1, 3 の組み合わせによる結果、つまり、製薬会社名、最初の特許が申請された年、特許申請書で利用された単語を用いて売り上げを推定したものである。No. 6 は、k = 1, 3, 4 を組み合わせたもの、つまり、No.5 に、さらに医薬品記事を用いて売り上げを推定したものである。

実験で用いられた特許申請書の申請年は、1980 年から 2021 年の物を用いたので、yp_{min} = 1980, yp_{max} = 2021 である。439 の特許申請書中全体において、207,403 の単語が抽出された。k = 3 の実験においては、Ω (0.1, 0.01) とする。|Ω| = 887 であり、その中の例は、Table I で示したものである。各 d_i に対する r 値の計算においては、a_i を除いて計算を行った。

k = 4 における記事に関しては、1998 年から 2020 年の物を用いたため、ya_{min} =1998, ya_{max} = 2020 とする。

B. 特許成立予測

前章までで議論した方法で、特許成立予測の精度の評価の実験を行った。実験では、IV-B 章で議論した合計 12,499 のパテントファミリーのうち、米国、インド、ブラジルの全ての国で " 成立 "、もしくは " 拒絶 " のどちらかの判定ラベルがついている、n = 7,043 のパテントファミリーを使用する。

形態素解析は、売り上げ予測と同じライブラリを利用する。

実験で使用したパテントファミリーに含まれる特許明細書は、1999 年から 2014 年に出願されたものを使用したため、yp_{min} = 1999、yp_{max} = 2014 とする。

深層学習においては、2 つの隠れ層を使用する。各々のノード数は 128、モデルの評価には、成立 / 拒絶の実際の値と予測値に対する Recall, Precision を用いて計算される、F 値 =2 ×Precision×Recall/(Precision+Recall)、を用いる。**Deep learning** には、Tensorflow / Keras[21] の python パッケージを使用する。すべてのケースで、5 分割交差検定（trainsize = 0.8）によってパフォーマンスを評価し、各結果の平均を計算する。r 値を計算する際は、、各交差検定におけるテストデータのみを使用する。

各パテントファミリー内のパテントについて、次のケースの評価を行った。

Case-1：国 c_0 における特許の成立可否の予測
Case-2：国 c_1 での付与 / 非付与ステータスがわかっている条件で、国 c_0 での特許の成立可否の予測

Case-1 と 2 について、2), 3), 4) 及びこれら 3 つのコンビネーション（=5)）の 4 通りに関して、深層学習による学習結果による合否の正解率を、F 値を用い、評価を行った。Case-1 では、次のデータセット X[k] を使用し、学習データとテストデータに関して、次のようにラベル g_{c0} を用いた。

$$X^{[k]} = [x_1^{[k]}, .., x_n^{[k]}]^T,$$
$$g_{c_0} = [g_{c_0 1}, .., g_{c_0 n}]^T \quad . \tag{19}$$

ここで、$x_i^{[k]}$ はセクション IV のものに対応し、特許 i が国 c_0 に対して「成立」ラベルが付与されている場合は $g_{c_0 i}$ = 1、「拒絶」ラベルが付与されている場合は $g_{c_0 i}$ = 0 となる。

TABLE III
実験結果（特許成立予測）：太字 = 最高値。

	Case-1		
c_0	US	IN	BR
2) Patent info	0.929	0.574	0.573
3) Words	0.974	0.613	0.681
4) Hot words	0.898	0.598	0.767
5) 2) 〜 4)	**0.988**	**0.643**	**0.863**

	Case-2　c_1 成立					
c_1	US	US	IN	IN	BR	BR
c_0	IN	BR	US	BR	US	IN
2) Patent info	0.618	0.606	0.883	0.873	0.887	0.610
3) Words	0.619	0.868	0.960	0.865	0.972	**0.642**
4) Hot words	0.582	0.876	0.835	0.888	0.829	0.610
5) 2) 〜 4)	**0.641**	**0.900**	**0.982**	**0.898**	**0.978**	0.628

	Case-2　c_1 拒絶					
c_1	US	US	IN	IN	BR	BR
c_0	IN	BR	US	BR	US	IN
2) Patent info	0.618	0.880	0.879	0.875	0.865	0.597
3) Words	0.630	0.853	0.967	0.856	0.974	0.619
4) Hot words	0.598	0.871	0.830	**0.907**	0.830	0.617
5) 2) 〜 4)	**0.633**	**0.889**	**0.976**	0.877	**0.975**	**0.654**

Case-2 の場合、c_1 のラベルが g（$\in 0, 1$）として知られているときに、c_0 の i のラベルを予測する。
各テストでは、インデックスが $J = \{i | g_{c_0 i} = g\}$ である明細書を利用する。
実験においては、X に対し、z-normalization による正規化を行った。
Table III は、Case-1,2 の実験結果である。太字は、各国または国のペアでの F 値の最高スコアである。

VI. 議　　論

売り上げ予測、特許成立予測の両方の実験結果に関して、幾つかの興味深い知見が得られた。

A. 売り上げ予測

$k = 1, 3, 4$（TableII の No.6）の組み合わせは、RMSE と MAE で最高の精度であった。（RMSE,MAE）=（1.726,1.360）は、それぞれ開発会社名と最初の特許出願年のみを使用する No.1 の× 0.74、および× 0.72 であり、大幅な改善である。
興味深い事実の 1 つとして、$k = 4$ のみ（No. 4）の精度は No. 3 の精度ほど良くはないが、$k = 1, 3$（No.5）と組み合わせることにより（No.6）、No.5 の精度が改善することが挙げられる。これは、「旬」な言葉を含む特許が、医薬品の販売予測に貢献していることを示していることを示している。
特許の情報のみを使用する場合（No.2）においては、RMSE, MAE 共に最低の精度であっ

た。

これは、タイトルの長さ、要約、特許、およびクレームの数は、他の情報に比べ、売り上げ予測にはあまり貢献していないことを示している。

開発会社名と最初の特許出願年のみを使用する場合（No.1）は、No.2 の場合よりも RMSE と MAE の両方において精度が良かった。これは、No.1 の 2 つの情報が、ある程度の将来の売り上げの情報を含んでいることを示唆している。

一方、No. 3 において、|r| 値が 0.1 以上の単語を使用すると、No.1 と比べても売り上げ推定精度が高いことが判り、特許申請書に製品の売り上げが将来上がることを示唆する言葉が特許に含まれている可能性があることを示している。これは、特許権者の自信に起因している可能性も否定できない。

この実験では、要約、クレーム、またはその他の部分で使用されているかどうかに関係なく、各特許申請書の単語の使用のみを考慮している。ただし、特許構造分析 [1], [22]–[25] およびキーワード抽出分析 [1], [22], [26], [27] についていくつかの研究が行われており、効果が確認されているため、それらを考慮することにより、推定精度をさらに向上させられる可能性がある。

B. 特許成立予測

特許成立予測に関しても、幾つかの興味深い点が発見された。Case-1（単一国の特許成立予測）では、Table III から確認できるように、米国とブラジルの特許成立 / 拒絶を各々 F 値 0.988、0.863 と高精度で予測する事ができた。インドの場合、F 値は 0.643 となった。単一のデータ入力（つまり、2), 3), 4)）に関しては、3)（特許の単語）は米国とインドで最高のスコアを生成し、4)（特許明細書の「旬」な単語を利用した予測）では、ブラジルが最高のスコアであった。米国の場合、2) が 2 番目であった。Fig.1 に示すように、米国では被引用文献数と特許成立可否に大きな差があったため、これら納得感のある結果である。ただし、3) と 4) のスコアが 2) よりも優れていることは新しい発見である。

Case-2（成立 / 拒絶が国（c_1）で既知の場合に、国（c_0）の成立可否を予測）では、F 値は、c_1 に関係なく、国 c_0 の Case-1 の F 値に近いことが判った。具体的には、c_0 =U.S.、の場合の 5) (=2),3),4) のコンビネーション）の F 値は、c_1 が既知か否か（Case-1,2) に関係なく、全て 0.975 以上であった。また、c_0 = ブラジル、の場合においても、全て 0.877 以上であった。ただし、c_0 = インド、の F 値は最高（c_1 = ブラジル、拒絶）でも 0.654 にとどまった。このように、米国、ブラジルに対しては高精度で、インドでは中精度で特許成立予測が可能であることが判った。インドの特許成立予測精度が他の 2 国と比べて相対的に低いのは、Fig.1 から判るように、特徴量による成立可否の差が小さいこと、III-A 章で議論した “evergreening” に対する特許庁の判断基準と、特許明細書に記述されている単語やそこに使われている「旬」な単語との関連性が小さいことが可能性として挙げられる。Case-2 の結果から、ある国の特許の成立 / 拒絶がわかれば、米国またはブラジルでの成立 / 拒絶は高精度で予測可能であるが、インドでの成立可否の予測精度は中程度であることが判った。

Case-1,2 の両方で、特許明細書 3),4) で使用されている単語は、特許情報 1) よりも予測に貢献している。これは、特許明細書自体に特許成立 / 拒絶の情報が含まれていることを

意味する。また、全ての情報を使った場合 (5)) の精度が両方の Case に関して一番良いことも判明した。

因みに、特許成立予測の実験において、他国の成立可否が判っている Case-2 より、判っていない Case-1 の方が結果が良い場合が幾つかある（例えば、Case-1 の US の 5) は 0.988 であるのに対し、Case-2 の c_0 =BR, c_1 = US は 0.975）。これは、Case-2 の場合、Case-1 と比べ、学習に使われるデータの数が少ないことや、学習に使われる単語を式 (11) のように選んだことが原因として考えられる。

Ⅶ. むすび

本論文においては、製薬特許の仕様を深層学習で分析することにより、医薬品の売上予測、およびパテントファミリーにおける特許成立 / 拒絶予測を行い、一定の成果を得ることができた。

興味深い発見として、特許明細書に書かれた単語を解析するだけで、売り上げ予測、特許成立予測の両方に対してある程度の予測が立てられることが判ったこと、「旬」の単語が多く含まれる特許の予測の精度が更に向上することが挙げられる。これらは、特許および関連記事に将来の医薬品販売に関する情報が含まれていることを証明しているため、画期的な結果といえる。特許の明細書や記事は簡単に入手できるので、これは将来のマーケティング戦略の構築に役立つことが期待できる。その一方で、両方の予測にはまだ改善の予知があることも否定できない。

今後の取り組みとして、特許や記事の構造を考慮した上で NLP を適用していきたいと考える。また、単語埋め込みの概念（BERT [28] や word2vec [29] など）を使用して、特許と記事の間で類似した単語の使用法を特定し、これらの概念によって推定パフォーマンスがどのように向上するかを確認したい。また、本論文では、医薬品に焦点を当てたが、このモデルは、食品、電化製品、自動車、衣類などの他の業界にも適用可能であるため、それらの解析も行いたい。

Ⅷ. 謝辞

本研究は JSPS 科研費 JP20H04424 の助成を受けたものである。本研究を行うにあたり、大阪大学高等教育入試研究開発センターの三森八重子招聘教授、日本大学法学部の加藤浩教授、加藤暁子准教授、日本 IBM 東京基礎研究所の那須川哲哉 Senior Technical Staff Member、および鈴木祥子 Research Staff Member に多大なる助言を頂いたことに感謝する。

参考文献

[1] S. Suzuki and H. Takatsuka, “Extraction of keywords of novelties from patent claims,” Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1192–1200, 2016.

[2] K. H. Kim, Y. J. Han, S. Lee, S. W. Cho, and C. Lee, “Text mining for patent analysis to forecast emerging technologies in wireless power transfer,” Sustainability 11.22, 6240, 2019.

[3] C. C. Guderian, P. M. Bican, F. J. Riar, and S. Chattopadhyay, “Innovation management in crisis: Patent analytics as a response to the COVID-19 pandemic,” R&D Management 51.2, pp. 223–239, 2021.
[4] J. Matal. “A Guide to the Legislative History of the America Invents Act: Part I of II.” Fed. Cir. BJ 21: 435, 2011.
[5] M.A. Lemley, and C. V. Chien. “Are the US patent priority rules really necessary.” Hastings LJ 54: 1299, 2002.
[6] M. A. Lemley, and B. Sampat, “Examiner characteristics and patent office outcomes,” Review of economics and statistics 94.3: pp. 817–827, 2012.
[7] Agreement, Trips. “Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr. 15, 1994.” WTO Agreement, Annex C 1, 1994.
[8] Y. Mitsumori, “An Analysis of the Impact of TRIPS’ Special Exemption for LDCs on the Bangladesh Pharmaceutical Industry,” 2018 Portland International Conference on Management of Engineering and Technology (PICMET), pp. 1–6, doi: 10.23919/PICMET.2018.8481873, 2018.
[9] R. Gabble, and J. C. Kohler, “To patent or not to patent? the case of Novartis ’cancer drug Glivec in India,” Globalization and Health 10.1: pp. 1–6, 2014.
[10] L. L. Mueller and S. M. T. Costa, “Should ANVISA be permitted to reject pharmaceutical patent applications in Brazil?,” Expert opinion on therapeutic patents 24.1: pp.
[11] Cortellis, https://clarivate.com/cortellis/.
[12] Derwent, https://clarivate.com/derwent/.
[13] Clarivate, https://clarivate.com/cortellis/.
[14] Pharmaceutical Benefits Pricing Authority Annual (1998–2010). [Online]. Available: https://www.pbs.gov.au/pbs/industry/pricing/pbs-items/historical/pbpa-annual-reports.
[15] Pharmaceutical Benefits Pricing Authority Annual (2011–2020). [Online]. Available: https://www.health.gov.au/about-us/corporate-reporting/annual-reports?utm source=health.gov.au&utm medium=callout-auto-custom&utm campaign=digital transformation.
[16] Pharmaceuticals and Medical Devices Agency. [Online]. Available: https://www.pmda.go.jp/english/index.html.
[17] nltk Package. [Online]. Available: https://www.nltk.org/api/nltk.html.
[18] K.Kamijo, “ Future Sales Estimation using Patents,” 2nd International Conference on NLP Trends & Technologies, Sydney, 2021.
[19] PCT – The International Patent System. [Online]. Available: https://www.wipo.int/pct/en/.
[20] M. Erman and M. Mishra, “Pfizer sees robust COVID-19 vaccine demand for years, $26 bln in 2021 sales.” [Online]. Available: https://www.reuters.com/business/healthcare-pharmaceuticals/pfizer-lifts-annual-sales-forecast-covid-19-vaccine-2021-05-04/, 2021.
[21] Tensowflow : https://www.tensorflow.org/.
[22] S. Hido, et al., “Modeling patent quality: A system for large-scale patentability analysis using text mining,” Information and Media Technologies 7.3, pp. 1180–1191, 2012.
[23] Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa, and Makoto Iwayama, “Patent Claim Processing for Readability: Structure Analysis and Term Explanation,” Proceedings of the ACL-2003 Workshop on Patent Corpus Processing, 20: pp. 56–65, 2003.
[24] P. Parapatics and M. Dittenbach, “Patent Claim Decomposition for Improved Information Extraction,” Proceedings of the 2nd International Workshop on Patent Information Retrieval: pp. 33–36, 1990.
[25] S. Sheremetyeva, S. Nirenburg, and I. Nirenburg, “Generating patent claims from interactive input,” Proceedings of the 8th International Workshop on Natural Language Generation: pp. 61–70, 1996.

[26] M. Verma and V. Varma, “Applying Key Phrase Extraction to Aid Invalidity Search,” Proceedings of the 13th International Conference on Artificial Intelligence and Law: pp. 249–255, 2011.
[27] M. A. Hasan, W. S. Spangler, T. Griffin, and A. Alba, COA: Finding Novel, 2009.
[28] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

上條浩一　東京国際工科専門職大学　工科学部　情報工学科　教授
大関和夫　東京国際工科専門職大学　工科学部　情報工学科　教授