FUTURE SALES ESTIMATION USING PATENTS

Koichi Kamijo

Department of Information Technology, International Professional University of Technology in Tokyo, Shinjuku-ku, Tokyo, Japan

ABSTRACT

We propose a model to improve estimation accuracy of the future sales volume, focusing on pharmaceutical products, from their patents. Our approach is based on an analysis of patents obtained in the early development stages of the products. The development of pharmaceuticals often takes a long time (up to several decades in some cases), and the costs are huge, even exceeding one billion USD for just one product. Therefore, it is strongly desirable to estimate future sales volume at an early stage. One piece of information potentially useful for the estimation is the brand, i.e., the name of the developing company. Our model learns the sales volume and words used in multiple patent specifications and also focuses on the extent to which "seasonal" words are used. Experiments showed that our model much improved the accurately of the sales volume estimation compared with the case of just estimating from its brand name.

KEYWORDS

Sales Estimation, Pharmaceuticals, Patents, Natural Language Processing, Deep Learning.

1. INTRODUCTION

As COVID-19 vaccines are being announced by several pharmaceutical companies, the world's attention is currently focused on vaccines. These vaccines are being sold worldwide, and their sales volume is bound to be huge. For example, Pfizer expects robust COVID-19 vaccine demand in the current year and estimates a sales volume of 26 billion USD [1].

In a case such as the COVID-19 pandemic, the demand for vaccines is huge, spanning almost the entire global population, and this demand is bound to result in huge profits for pharmaceutical companies. However, excluding such special cases, estimating the sales volume of a new pharmaceutical in its development stage is not easy because we cannot accurately predict how demand will evolve.

Estimating the sales volume of new products or services in the early stages of development is very important when formulating a marketing strategy, especially for pharmaceuticals, as pharmaceuticals have a longer development period than other products and many of them have the potential to make huge profits.

Currently, estimating pharmaceutical sales volume requires knowledge of the disease and pharmaceutical market, a comprehensive understanding of the domestic and global regulatory environment, and market access. Therefore, this task is often performed by a company with appropriate expertise, such as Clarivate [2]. However, outsourcing such estimations is typically very expensive and requires pharmaceutical companies to share sensitive information with the contracting company. Accordingly, it would be desirable to make sales volume forecasts in-house without such exposure.

David C. Wyld et al. (Eds): CSE, AI & FL, NLPTT - 2021
pp. 45-58, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112404

When a company develops a new product or service, it usually applies for a patent before introducing it in the market, acquires the rights, and then starts full-fledged development and trading.

In this paper, we present a model for estimating the future sales volume of new pharmaceuticals by using specifications available in the initially acquired patents for each pharmaceutical.

Our model considers only pharmaceuticals whose prior sales volume and first patent are available. We performed morphological analysis of each patent specification provided to the patent office in the early development stage and counted usage ratio of each word in each patent. The usage ratio of each word and sales volume were used as training/test data for the model. We evaluated the model by leave-one-out (LOO) cross-validation; that is, out of *n* data, we used the *n*-1 data as training data and estimated the rest as test data. We repeated this *n* times and evaluated the sales estimation performance by averaging the *n* estimation results.

For developing the model, we also performed morphological analysis of articles related to pharmaceuticals, and the word usage ratio discussed above were weighted based on the usage of words contained in the articles.

For patent specifications, to ensure a unified patent format, we used only those patents that complied with the Patent Cooperation Treaty (PCT) [3]. We also targeted English texts only.

Section 2 of this paper presents related work, and Section 3 introduces the proposed pharmaceutical sales estimation model. Section 4 details the experiment, which are then discussed in Section 5. We conclude in Section 6 with a brief summary and mention of future work.

2. RELATED WORK

Since we could not find any research that directly discusses the relationship between sales estimation and patents, we examined studies on sales estimation and patent analysis.

2.1. Sales Estimation

Merino et al. proposed the combination of a spatial interaction model and simulation approaches for the reliable estimation of retail interactions and store sales volume on the basis of data on consumer shopping behavior in Mexico [4]. Their proposed methodology was based on the combination of a Huff model [5] and a Monte Carlo simulation [6] to reproduce shopping patterns in retail stores. Jordan et al. investigated how to improve the estimation accuracy of a firm's sales volume [7] and emphasized that rather than customer satisfaction, return on investment, or economic value, an evaluation of the quality of the firms' planning practices is the most important. Pavlyshenko et al. used machine learning for predicting sales volume and found that the use of stacking techniques could improve the performance of predictive models used for sales volume time series forecasting [8]. They noted that the use of regression approaches for sales volume forecasting could often provide better results than time series methods. Loureiro et al. investigated the use of a deep learning approach to forecasting sales volume in the fashion industry, namely, for predicting the sales volume of new individual products in future seasons, without the use of historical data [9]. They developed forecasting models by considering a wide and diverse set of variables (e.g., products' physical characteristics and the opinion of domain experts) and were able to perform highly accurate forecasting. They also found that deep neural networking outperformed other techniques such as random forest.

Note that none of the methods mentioned above are based on a patent or articles referring to the products.

2.2. Patent Analysis

Kim et al. analyzed patents to identify emerging and vacant technology areas of wireless power transfer. They extracted topic areas from patents by text mining, where topics with similar semantics were grouped together, and then applied a time series analysis and innovation cycle of technology to the grouping result [10]. The results of the clustering, time series analysis, and innovation cycle were then compared to minimize the possibility of misidentifying emerging and vacant technology areas. Guderian et al. investigated how innovation management decisions in times of crisis (e.g., the COVID-19 pandemic) could be improved through publicly available data, such as patents [11], and examined which data were valuable from the viewpoint of patent citation. Lee et al. proposed a forecasting model for new innovative product diffusion based on both technology diffusion and interest diffusion. Technology diffusion was defined on the basis of the number of patent citations, while interest diffusion was defined on the basis of web search traffic [12]. They used the model to predict the sales volume of hybrid cars and industrial robots in the US market and found that its prediction performance was better than that of the Bass model [13] and the Bass model with patent citation for both cases.

While all of the above works discuss how patent analysis can contribute to the forecasting of future business and technology trends, none of them focus on actual sales volume values and none of them use deep learning for analysis.

In a paper related to patent analysis, Suzuki et al. proposed an approach to automatically extract keywords related to novelties or inventive steps from patent claims by using the structure of the claims [14]. Hido et al. addressed the problem of assessing the quality of patent specifications on the basis of machine learning and text mining techniques. They computed a score called patentability, which indicates the likelihood of an application being approved by the patent office [15], and employed a new statistical prediction model to estimate examination results (approval or rejection) on the basis of a large dataset including 0.3 million patent specifications. While these two papers do not directly relate to sales estimation with patents, they do provide tips for analysing patents.

3. METHODOLOGY

Our objective was to construct a model that could estimate the sales volume of new pharmaceuticals on the basis of not only the names of the development companies but also patents and articles related to the pharmaceuticals. Figure1 shows our research framework. (A) through (D) below correspond to those in the list after the sentence “In total” in the next page. We obtained the pharmaceutical list and sales volume from a database of Cortellis [16] ($1^{\dagger}$and $2^{\dagger}$ in the figure). We first estimated the sales volume by using the information about the pharmaceutical, specifically, the name of the company that developed the pharmaceutical and the year in which the first patent application for the pharmaceutical was made (A). For each pharmaceutical, we collected the first patent application for the pharmaceutical that complied with the PCT and that was written in English ($3^{\dagger}$), and estimated the sales volume using the information about the patents (B). For the first patent, we used the information provided by Cortellis and Derwent [16,17], whose employees include experts on pharmaceuticals and patents. We then performed morphological analysis for each patent ($4^{\dagger}$) and estimated the sales volume

from the usage ratio of each word, that is, the number of appearances of each word divided by the number of appearances of all the words in each patent specification (C). Next, we collected articles related to the pharmaceuticals (5†), performed morphological analysis on them (6†), and calculated the

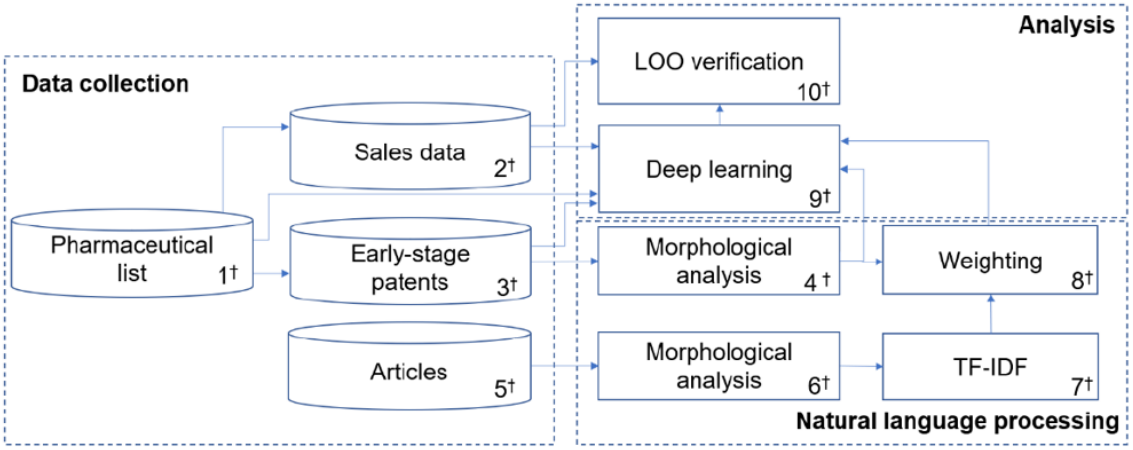

Figure 1. Research framework.

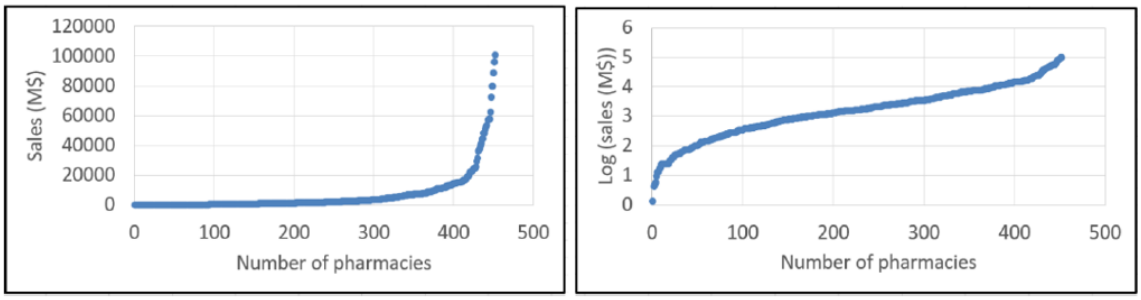

Figure 2. Plot ofthe sales volume of pharmaceuticals used in the experiment (in millions of USD, 2019 to 2027, actual sales volume + estimation by experts), smallest first (left: real; right: logarithm).

Term Frequency - Inverse Document Frequency (TF-IDF) of each word in each year (7†). We then calculated the sum of the usage ratio of each word weighted by the TF-IDF each year (8†), and estimated the sales volume (D). In our model, for sales estimation, we used deep learning (9†), and for the evaluation of the sales estimation, we used LOO cross-validation (10†).

In total, we performed the following estimations.

A. sales estimation from information about the pharmaceutical
B. sales estimation from information about the first patent of the pharmaceutical
C. sales estimation from the words used in the first patent of the pharmaceutical
D. sales estimation from C plus pharmaceutical related articles
E. sales estimation by combining A–D

For the model construction, for the sales volume s_i of pharmaceutical d_i, we used log(s_i)(base=e) instead of s_i, since the range of pharmaceutical sales volume can vary widely. Figure2 shows a plot of the sales volume of pharmaceuticals used in the experiment (in millions of USD, 2019 to 2027, actual sales volume + estimation by Clarivate [2]), smallest first. The left-side figure shows

the real data and the right-side figure shows the logarithm of the real data. Clearly, logged sales volume are well distributed and linear, and it is expected that we can make a more accurate estimation compared with the case of the real data.

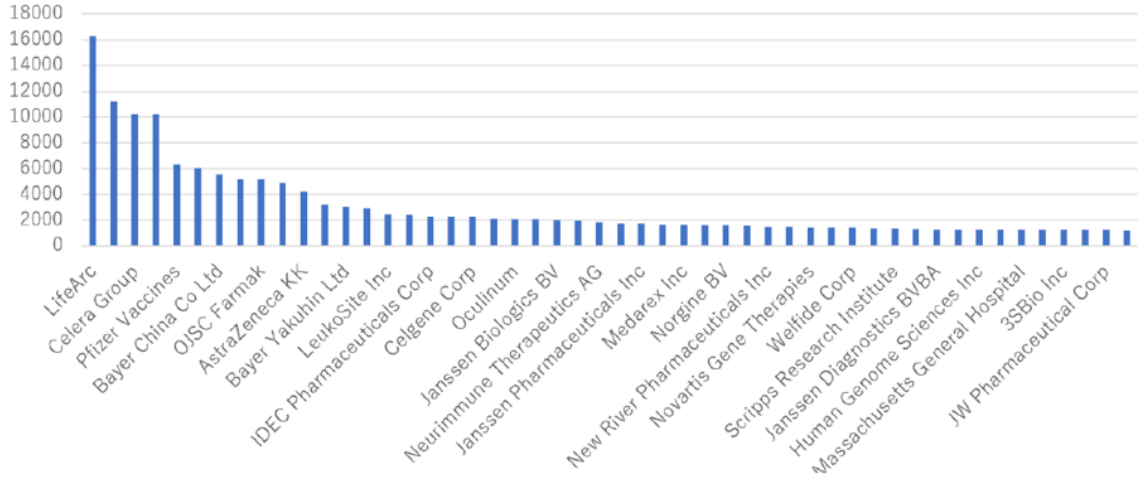

Figure 3. Sales volume values for different developing companies for the year 2019 (in millions of USD).

In this study, the sales volume we used was the sum of actual sales volume in 2019 plus expert-. estimated sales volume from 2020 to 2027.

In other words, the values to be estimated include the future sales volume value estimated by experts. This is because pharmaceuticals that were developed just before or after 2019 do not have enough sales achievement data. From the viewpoint of sales estimation research, estimating these values still has worth for research.

The following subsections describe the details of each estimation.

3.1. Sales Estimation from Information about the Pharmaceutical

In this step, we estimate sales volume by utilizing information on each pharmaceutical. This information includes the name of the company that developed the pharmaceutical and the year in which a patent application was made in the early development stage. Figure3 shows the sales volume values for each pharmaceuticals developing company for the year 2019[2]. Since the sales volume values are different for different companies, by knowing the developing company, we can roughly estimate the future sales volume of each pharmaceutical.

We could also use additional information, such as the name of the selling companies, but these companies may not have been decided at the time the pharmaceutical was developed. Therefore, we used only the developing company and the year of the first patent application.

For sales estimation from information about the pharmaceutical d_i, we used a one-hot vector $\boldsymbol{v}_i[j]$ for the developing company, defined as

$$\boldsymbol{v}_i[j] = \begin{cases} 1, & j = f_c(d_i), \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $f_c(d_i)$ is the index of a company that developed d_i. We used the year in which the first patent for the pharmaceutical d_i was applied (yp_i) as part of the training/test data. We defined a ($|\boldsymbol{v}_i|$+1) dimensional vector $\boldsymbol{x}_i^{[1]}$ as

$$\boldsymbol{x}_i^{[1]} = [v_i, yp_i]^T, \tag{2}$$

where ^{T} is a transpose.

3.2. Sales Estimation from Information about the First Patent of the Pharmaceutical

Patent specifications usually consist of a title, an abstract, claims, a detailed explanation, and so on. We analyzed the correlation between these components and sales volume and examined the sales estimation obtained with these parameters. We used the following information from the first patent of pharmaceutical d_i for sales estimation.

- pr_{1i}: Word count of title
- pr_{2i}: Word count of abstract
- pr_{3i}: Word count of the document
- pr_{4i}: Number of claims

We defined a vector $\boldsymbol{x}_i^{[2]}$ for a pharmaceutical d_i as

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

3.3. Sales Estimation from the Words used in the First Patent of the Pharmaceutical

We analyzed the correlation between words used in each patent and the sales volume of each pharmaceutical. For each patent specification a_i for a pharmaceutical d_i, we first performed morphological analysis for all the words in the patent and counted the usage ratio of each word, u_{wi}, where $\sum_w u_{wi} = 1$. We excluded stop words, numbers, and symbols in the calculation of the usage ratio, but included all other words, regardless of the part of speech to which they belonged. We defined a vector $\boldsymbol{u}_w$ for each word w as

$$\boldsymbol{u}_w = [u_{w1}, .., u_{wn}]^T. \tag{4}$$

We calculated Pearson's r-value, r_w, between $\boldsymbol{u}_w$ and $\boldsymbol{ls} = [\log(s_1), .., \log(s_n)]^T$, and then selected a set of words $\Omega(Tr, Tp)$ that satisfied the following:

$$\Omega(Tr, Tp) = \{w | |r_w| \geq Tr, p_w \leq Tp\}, \tag{5}$$

where T_r and T_p are thresholds and p_w is the p-value for r_w. Table 1 shows an example of $\Omega(0.1, 0.01)$ sorted by $|r_w|$, which are used in the experiment. Bold words in the table are related to a pharmaceutical or a disease, and "ratio" is the ratio of patents that used at least once out of 423 patents for each word.The table shows that 36% of the words were related to a pharmaceutical or a disease.

Finally, for each pharmaceutical d_i, we used the usage ratio of the words in Ω, as

$$\boldsymbol{x}_{\boldsymbol{i}}^{[3]} = [u_{w_1 i}, .., u_{w_{|\Omega|} i}]^T, \quad (6)$$

where $\Omega = \bigcup_j w_j$.

3.4. Sales Estimation from C plus Pharmaceutical Related Articles

If a patent includes "hot" keywords that reflect the high popularity of the pharmaceutical at the time an application is made for the patent, the sales volume of the pharmaceutical is likely to grow in the future. With this in mind, we weighted the word usage ratio in the patent specification based on the usage of each word in pharmaceutical related articles.

For the analysis, we first collected pharmaceutical-related articles published in year *y* and then calculated the TF-IDF of the word *w* in year *y*, tfidf(*w*,*y*), defined as

Table 1. Sample of words with |*r*|-values in 432 patents used in the experiment, and ratio of patents that used at least once out of 423 patents for each word.

word	r-value	ratio	word	r-value	ratio
coval	0.567	0.124	**epiderm**	−0.277	0.106
plasma	0.49	0.108	**cag**	0.274	0.106
tag	0.483	0.177	**cynomolgu**	−0.273	0.104
unmodi	0.482	0.119	herebi	0.27	0.195
fraction	0.427	0.261	**best**	−0.268	0.128
test	0.334	0.179	qiagen	−0.266	0.106
mainten	0.328	0.146	draw	−0.265	0.131
energi	−0.321	0.108	**precipit**	−0.262	0.122
stock	0.321	0.104	inhibitor	0.259	0.383
obtain	0.308	0.139	wherea	−0.256	0.155
lupu	−0.307	0.102	posit	0.254	0.575
copi	−0.304	0.104	examin	−0.253	0.102
germlin	0.303	0.128	complement	−0.247	0.104
second	−0.301	0.108	transform	−0.242	0.102
ctg	0.297	0.102	underlin	−0.241	0.113
dmso	0.292	0.157	prepar	−0.241	0.144
transplant	−0.287	0.111	**coloni**	−0.241	0.148
delet	0.285	0.175	**microtit**	−0.24	0.128
amino	0.283	0.126	separ	0.236	0.361
vegf	0.28	0.1	**substrat**	0.235	0.106
non-limit	0.279	0.215	principl	−0.234	0.113
isoleucin	0.279	0.117	alter	−0.234	0.162
immobil	0.278	0.177	altern	0.232	0.648
lymphoma	0.277	0.197	**polynucleotid**	−0.232	0.223
prolong	−0.277	0.10	combin	−0.231	0.142

$$\text{tfidf}(w, y) = \text{tf}(w, y)\log(\text{idf}(w)), \quad (7)$$

where tf(*w*,*y*) represents the frequency of the word *w* in articles published in year *y* and idf(*w*) denotes the inverted frequency of *w* among all of the articles for all of the years. The frequency is normalized so that we have $\sum_w \text{tf}(w, y) = 1$ for all of *y* s.

For each pharmaceutical d_i, we calculated the sum of the word usage ratio in the patent specification a_i, that is, u_{wi}, weighted by tfidf(*w*,*y*) for each year *y*, as

$$ut(y, i) = \sum_w \text{tfidf}(w, y) u_{wi.} \quad (8)$$

It is possible to increase the *ut* weighting ratio in the model, whose year of first patent specification application (yp_i) is close to the nearest article publication year. In deep learning, such a weighting ratio is automatically adjusted. From this viewpoint, we evaluated the case of

putting elements of *ut*s so that the position of each element for the year the article is published minus the year the first patent is applied (yp_i) is the same for all *i*s. Towards this, we created a new vector, $\boldsymbol{x}_i^{[4]}$, by padding 0s to the left and/or right, as

$$\boldsymbol{x}_i^{[4]} = [\mathbf{0}^{z1_i}, ut(ya_{min}, i), .., ut(ya_{max}, i), \mathbf{0}^{z2_i}]^T, \tag{9}$$

where $\mathbf{0}^j$ is a vector that consists of *j* 0s,

$z1_i = yp_{max} - yp_i, z2_i = yp_i - yp_{min}, yp_{min}$ and yp_{max} are the minimum and maximum value of yp_i, respectively, and ya_{min} and ya_{max} are the oldest and latest years of the articles for our analysis, respectively. No year is skipped between ya_{min} and ya_{max} in (9). For example, if yp_0=1996,yp_1=1999, ya_{min}=1998, ya_{max}=2020, yp_{min}=1980, and yp_{max}=2021, then,

$$\begin{aligned} \boldsymbol{x}_0^{[4]} &= [\mathbf{0}^{25}, ut(1998,0), .., ut(2020,0), \mathbf{0}^{16}]^T, \\ \boldsymbol{x}_1^{[4]} &= [\mathbf{0}^{22}, ut(1998,1), .., ut(2020,1), \mathbf{0}^{19}]^T. \end{aligned} \tag{10}$$

$\boldsymbol{x}_i^{[4]}$ included *ut*s with articles published after the first patent application. We can calculate such *ut*s some years after the first patent application year and publication year of the articles. However, immediately after the first patent application, we do not have articles published after that year, so we define another vector $\boldsymbol{x}_i^{[5]}$ that only includes *ut*s that use articles published in the year equal to or before the first patent application, as

$$\boldsymbol{x}_i^{[5]} = \begin{cases} [\mathbf{0}^{z1_i}, ut(ya_{min}, i), .., ut(yp_i, i), \mathbf{0}^{z3}]^T, & yp_i \geq yp_{min}, \\ [\mathbf{0}^{yp_{max} - yp_{min} + ya_{max} - ya_{min}}]^T, & \text{otherwise}, \end{cases} \tag{11}$$

where $z3 = ya_{max} - ya_{min}$.

3.5. Sales Estimation by Combining A–D

To obtain a more accurate estimation, we combined the training/test data defined in the previous subsections. We define

$$\boldsymbol{x}_i^{[a_1,..,a_m]} = [\boldsymbol{x}_i^{[a_1]T}, .., \boldsymbol{x}_i^{[a_m]T}]^T. \tag{12}$$

a data that combines $\boldsymbol{x}_i^{[a_1]}$ through $\boldsymbol{x}_i^{[a_m]}$, as $m \geq 1$.

4. EXPERIMENT

We evaluated the sales estimation performance of each of the methodologies discussed in the previous sections. In the experiment, we used *n*=432 pharmaceuticals whose sales volume (s_i) and first patent (a_i) (in PCT, written in English) were both available. Table 2 shows the notation used.

The left-side panel of Figure4 shows the number of companies that developed one or more pharmaceuticals; a total of 432 pharmaceuticals were considered in the experiment. Specifically, 186 companies (43%) developed only one pharmaceutical, and one company developed 13 pharmaceuticals, which was the highest number of pharmaceuticals developed by a company. The right-side panel of Figure4 shows the number of first patent applications in different years.

Since n=432 is not sufficiently large, we evaluated the accuracy of performance by using LOO cross-validation.

For the estimation model, we used Keras Regressor for multiple regression with two hidden layers, 128 nodes with relu activation each, and number of epochs = 100, unless specified otherwise. We used mean squared error for the loss and Adam for the optimizer. We normalized the input vector by z-score normalization, except for the one-hot vector. To speed up the experiments, we used a Google Colaboratory [18] TPU. For morphological analysis, stemming, and lemmatization, we used "word_tokenize" in the nltk Package [19]. In this package, sentences "We were performing maintenance. It rains cats and dogs." are converted to "We were performmainten. It rain cat and dog."Some are converted to words not in dictionaries. We dealt with case insensitive.

For each d_i, we used the following training and test sets, (X_train, Y_train), (X_test, Y_test), as

Table 2. Notation. Description of symbols and variables used in this paper.

Notation	Description
n	number of pharmaceuticals
d_i	pharmaceutical (drug) i
s_i	actual + experts' estimated sales of d_i from 2019 to 2027
a_i	first patent specification for d_i, PCT written in English
u_{iw}	ratio word w is used in a_i
yp_i	the year a_i was applied to a patent office
yp_{min}	minimum value of yp_i (=1980)
yp_{max}	maximum value of yp_i (=2021)
ya_{min}	the oldest year of the articles in the experiment (=1998)
ya_{max}	the latest year of the articles in the experiment (=2020)
v_i	one-hot vector for pharmaceutical d_i
pr_{ki}	property of a_i, $k = 1, .., 4$
r_w,p_w	r-value, p-value for word w
$r_c(T_c)$	the ratio whose difference between estimated volume and actual sales volume $\leq T_c$
$\Omega(T_r, T_p)$	set of words that satisfy $\|r_w\| \geq T_r$ and $p_w \leq T_p$
tfidf(w, y)	TF-IDF of the word w in year y
$ut(i, y)$	u_{iw} weighted by tfidf(w, y) and summed over w

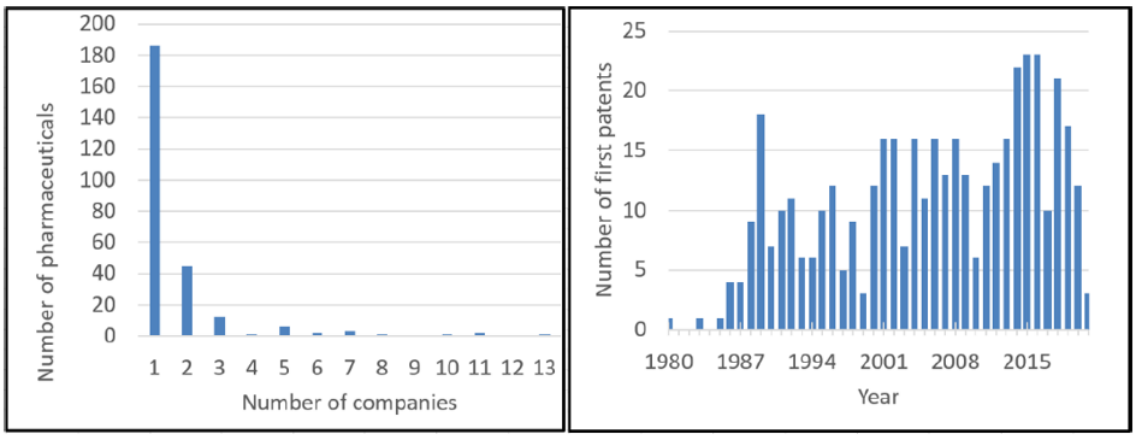

Figure 4. (Left) Number of companies that developed one or more pharmaceuticals and (right) number of first patents applied for per year, both of which were considered in our experiment.

$$(X\_train, Y\_train), (X\_test, Y\_test) = (\boldsymbol{X}_{\neg i}^{[k]}, \boldsymbol{ls}_{\neg i}), (x_i^{[k]}, \log(s_i)), \quad (13)$$

where

$$\boldsymbol{X}_{\neg i}^{[k]} = [\boldsymbol{x}_i^{[k]}, .., \boldsymbol{x}_{i-1}^{[k]}, \boldsymbol{x}_{i+1}^{[k]}, .., \boldsymbol{x}_n^{[k]}]^T,$$
$$\boldsymbol{ls}_{\neg i} = [\log(s_1), .., \log(s_{i-1}), \log(s_{i+1}), .., \log(s_n)]^T, \quad (14)$$

And k=1,..,5 or a combination of these values, as discussed in Section 3. We constructed models n times for each d_i and then calculated the root mean square error (RMSE) and mean absolute error (MAE), as

$$RMSE = (\sum_{i=1}^{n} (l\hat{s}_i - \log(s_i)^2/n)^{0.5},$$
$$MAE = \sum_{i=1}^{n} |l\hat{s}_i - \log(s_i)|/n, \quad (15)$$

Table 3. Experimental results: bold = best data, italic = worst data in RMSE,MAE, and r_c, respectively.

No.	Input(k)	No. of epochs	Node size - 1	Node size - 2	RMSE	MAE	r_c log(2)
1	1	100	128	128	2.314	1.893	*0.208*
2	2	100	128	128	*2.65*	*2.121*	0.211
3	3	100	128	128	1.921	1.489	0.289
4	4	100	128	128	2.036	1.634	0.266
5	1,3	100	128	128	1.811	1.408	0.32
6	1,3,4	100	128	128	**1.724**	**1.359**	0.326
7	1,3,5	100	128	128	1.809	1.459	0.282
8	1,3,4	10	128	128	1.994	1.604	0.241
9	1,3,4	20	128	128	1.88	1.499	0.282
10	1,3,4	50	128	128	1.773	1.378	0.317
11	1,3,4	150	128	128	1.865	1.443	0.299
12	1,3,4	200	128	128	1.774	1.361	**0.35**
13	1,3,4	100	64	128	1.802	1.38	0.34
14	1,3,4	100	256	128	1.844	1.415	0.317
15	1,3,4	100	128	64	1.749	1.365	0.319
16	1,3,4	100	128	256	1.973	1.474	0.333

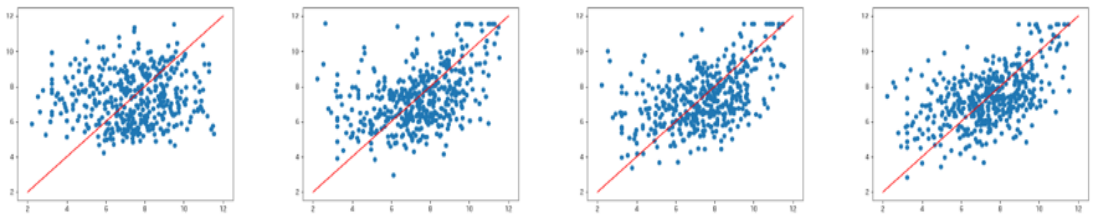

Figure 5. Scatter plots of actual (x-axis) versus estimated (y-axis) sales (logged in millions of USD, from 2019 to 2027). From left to right: Nos. 1, 3, 5, and 6 in Table 3. The diagonal line shows estimation = actual sales.

where $l\hat{s}_i$ is the estimated volume by our model for the input of (13). For outliers, we replaced $l\hat{s}_i$ with max(min($l\hat{s}_i$),max$_{j\neq i}$(log(s_j))).

We also calculated r_c, the ratio of d_i whose estimated volume was close to actual sales volume log(s_i), defined as

$$r_c(\mathrm{T}_c) = |\{i||l\hat{s}_i - \log(s_i)| \leq T_c|/n, \quad (16)$$

where T_c is a threshold.

Table 3 shows the results of all the experiments. Values in the "input(k)" column correspond to k in (14). We used T_c=log(2), which implies that the difference between the actual and estimated data is log(2); in other words, their ratio is between 0.5 and 2.0. We discuss each of the experiments below, where the various numbers (e.g., No. 1) refer to the serial number ("No.") in Table 3.

The scatter plots in Figure5 show the actual versus estimated sales for input Nos. 1, 3, 5, and 6.

Nos. 1 - 4 are the results without using any combinations. No. 5 is the results of combinations with k=1, 3, that is, information of developing company name, the year the first patent was applied for, and the words used in the first patent. Nos. 6, 7 are the results of combinations with k=1, 3, 4, and k=1, 3, 5, respectively. k=4 is the case of using words in pharmaceutical articles, and k=5 is the case where articles published after the first patent were excluded in the k=4 case.

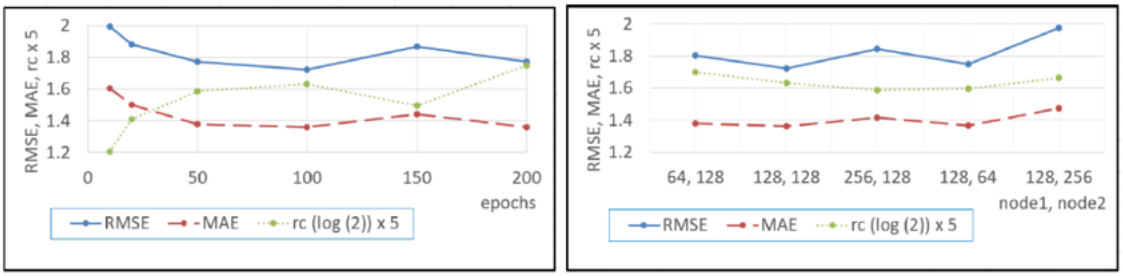

Figure 6. Estimation accuracy for different number of epochs (left) and different node sizes (right).

We performed other tests by changing the number of epochs and node sizes with combinations of k=1, 3, 4. Nos. 8 - 12 are the results when changing the number of epochs from 100 to 10, 20, 50, 150, and 200, respectively. Nos. 13 - 16 are the results when changing node sizes to (node 1, node 2) = (64,128), (256, 128), (128, 64), and (128, 256), respectively. Figure6 shows the results of Nos. 8-12 (left) and 13-16 (right).

We used pharmaceutical data from the years between 1980 and 2021, so we had yp_{min}=1980 and yp_{max} = 2021. For the 432 patent specifications, 203,632 different words were used. For k=3, we used $\Omega(0.01,0.01)$, $|\Omega|$=866, and 50 of the 866 words in Table 1. To calculate the r-value for each d_i, we used all of the patent specifications except for a_i.

As the articles for k=4, 5, we used Pharmaceutical Benefits Pricing Authority Annual Reports published from 1998 to 2020 [20, 21] and reports of the Pharmaceuticals and Medical Devices Agency from 2004 to 2018 [22]. For the years between 1998 and 2003 and between 2019 and 2020, we used reports from Pharmaceutical Benefits Pricing Authority Annual Reports. Therefore, we had ya_{min}=1998 and ya_{max} = 2020.

5. DISCUSSION

Several inferences can be made from the experimental results.

Combination of k=1, 3, 4 (No. 6 in Table 3) yielded the best performance for RMSE and MAE.

(RMSE, MAE) = (1.724, 1.359) is ×0.75 and ×0.72 of those of No. 1, respectively, which use only the developing company name and the year of the first patent application. This is a significant improvement.

One interesting fact is that, while the performance of k=4 alone (No. 4) was not as good as No. 3, it helped improve the combination of k=1, 3 (No. 5). This implies that patents containing "hot" words indicate high potential for the future sales volume of the pharmaceutical.

Using only the information of the first patent yielded the worst result for RMSE and MAE (No. 2). This implies that the length of the title, the abstract, the patent, and the number of claims contain very little or no useful information regarding future sales volume.

Using only the developing company name and the year of the first patent application (No. 1) yielded a better performance than the case of No. 2 for RMSE and MAE. This is reasonable since Figure 3 indicates that the sales volume of some pharmaceuticals depends on the developing company names. However, 43% of the pharmaceuticals were "single" pharmaceuticals with only one developing company in the training/test data, and for these, sales volume estimation was close to the average value of the rest of the companies.

In contrast, using words whose absolute*r*-values were equal to or more than 0.1 was very effective (No. 3), compared with No.1 or No. 2, even without combining with other vectors.

This indicates that patents may include words implying that the target products or services will sell in the future. It is possible that this stems from the confidence of the patent authors.

Comparing the panels in Figure5 from left to right, we can observe that dots are shifting to the diagonal line, which shows estimation = actual sales, which are the cases of k=1only (No. 1), k=3 only (No. 3), combination of k=1, 3 (No. 5), and combination of k=1, 3, 4 (No. 6), respectively.

We evaluated the case of using articles published in the same year or before the first patent application, namely, the combination of k=1, 3, 5 (No. 7). In this case, the estimation performance was worse than that for No. 6, which is reasonable since the input data were partly omitted, but the performance was still better than without using k=5 (No. 3) for RMSE.

For the case of the combination of k=1, 3, 4, we evaluated the performance by using a different number of epochs (Nos. 8-12, Figure6, left panel) and node size (Nos. 13 - 16, Figure6, right panel). For Nos. 8 - 10, the estimation performance improved for RMSE and MAE as the number of epochs increased, but the performance remained unchanged or deteriorated as the number of epochs increased beyond 100, implying that deep learning parameters overfit after 100 epochs. On the other hand, varying the node size did not influence the performance significantly, but found that node size = 256 in either node (Nos. 14, 16) yielded worse results than those with smaller nodes (Nos. 13, 15).

In this experiment, we considered only the usage of words in each patent, regardless of whether they were used in the abstract, claim, or other parts. However, as several studies have been performed on patent structure analysis [14,15,23-35] and keyword extraction analysis[14,15,26,27], there is still scope for further estimation accuracy improvement by, for example, applying weights to word usage in accordance with the location of the words (e.g., in the abstract, claims, or other parts).

6. CONCLUSION AND FUTURE WORK

We proposed a model that improves the estimation performance of future pharmaceutical sales by analyzing the first patent specification submitted for the pharmaceutical and articles related to pharmaceuticals. We performed experiments using a data consisting of 432 pharmaceuticals whose first patents and sales volume are both available and found that the best sale estimation performance was obtained by using a combination of pharmaceutical developing company name, the year the first patent was applied for, words used in the first patent specification, and TF-IDF calculated from words used in the pharmaceutical related articles to weight the word usage ratio of the first patent of the pharmaceutical.

One interesting finding was that just using words in the patent specification yielded much better estimation performance than the case of using the company name (i.e., the brand name) and the year the first patent was applied for. Also, the estimation performance was much improved by combination all of these plus pharmaceutical related articles. These are groundbreaking results because these prove that patents and related articles contain information about future pharmaceutical sales. Since patent specifications and articles can easily be obtained, this will help us significantly in building a marketing strategy.

In this paper, we focused on pharmaceuticals, but our model can be applied to other industries such as food, electrical appliances, cars, clothes, and so on.

As future work, we would like to apply NLP while taking the structure of patents and articles into consideration. We would also like to examine the use of word embedding concepts (e.g., BERT [28] or word2vec [29]) to determine similar word usage between patents and articles, and see how these concepts improve the estimation performance.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Tetsuya Nasukawa and Dr. Shoko Suzuki in IBM Research-Tokyo for their accurate advice.

This work was supported by JSPS KAKENHI Grant Number 10881998.

REFERENCES

[1] M. Erman and M. Mishra, (2021) “Pfizer sees robust COVID-19 vaccine demand for years, $26 bln in 2021 sales.” https://www.reuters.com/business/healthcare-pharmaceuticals/pfizer-lifts-annual-sales-forecast-covid-19-vaccine-2021-05-04.(Last accessed: 13.Dec.2021)
[2] Clarivate. https://clarivate.com/.(Last accessed: 13. Dec.2021)
[3] PCT – The International Patent System. https://www.wipo.int/pct/en/.(Last accessed: 13. Dec.2021)
[4] M. Merino and R. Adrian, (2016) “Estimation of retail sales under competitive location in Mexico,” *Journal of Business Research* 69.2, pp. 445-451.
[5] D. L. Huff, (1963) “A probabilistic analysis of shopping center trade areas,” *Land economics* 39.1, pp. 81-90.
[6] D. E. Raeside, (1974) “An introduction to Monte Carlo methods,” *American Journal of Physics* 42.1 pp. 20-26.
[7] S. Jordan and M. Martin, (2020) “The use of forecast accuracy indicators to improve planning quality: Insights from a case study,” *European Accounting Review* 29.2, pp. 337-359.
[8] B. Pavlyshenko, (2019) “Machine-learning models for sales time series forecasting,” *Data* 4.1, 15.
[9] A. L. D. Loureiro, V. L. Miguéis, and L. F. M. da Silva, (2018)“Exploring the use of deep neural networks for sales forecasting in fashion retail,” *Decision Support Systems* 114, pp. 81-93.
[10] K. H. Kim, Y. J. Han, S. Lee, S. W. Cho, and C. Lee, (2019) “Text mining for patent analysis to forecast emerging technologies in wireless power transfer,” *Sustainability* 11.22, 6240.

[11] C. C. Guderian, P. M. Bican, F. J. Riar, and S. Chattopadhyay, (2021) "Innovation management in crisis: Patent analytics as a response to the COVID-19 pandemic," *R&D Management* 51.2, pp. 223-239.
[12] W. S. Lee, S. C. Hyo, and Y. S. So, (2018)"Forecasting new product diffusion using both patent citation and web search traffic," *PloS one* 13.4, e0194723,
[13] F. Bass, (1969) "A newer product growth for model consumer durables," *Management Science*, Vol. 15, No. 5, January, pp. 215-227.
[14] S. Suzuki and H. Takatsuka, (2016) "Extraction of keywords of novelties from patent claims," *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*: Technical Papers, pp. 1192-1200.
[15] S. Hido, et al., (2012) "Modeling patent quality: A system for large-scale patentability analysis using text mining," *Information and Media Technologies* 7.3, pp. 1180-1191.
[16] Cortellis. https://clarivate.com/cortellis.(Last accessed: 13. Dec.2021)
[17] Derwent. https://clarivate.com/derwent/solutions/derwent-innovation/.(Last accessed: 13. Dec.2021)
[18] Google Colaboratory. https://colab.research.google.com/notebooks/intro.ipynb?utm_source=scs-index.(Last accessed: 13. Dec.2021)
[19] nltk Package. https://www.nltk.org/api/nltk.html.(Last accessed: 13. Dec.2021)
[20] Pharmaceutical Benefits Pricing Authority Annual (1998-2010).https://www.pbs.gov.au/pbs/industry/pricing/pbs-items/historical/pbpa-annual-reports.(Last accessed: 13. Dec.2021)
[21] Pharmaceutical Benefits Pricing Authority Annual (2011-2020). 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.(Last accessed: 13. Dec.2021)
[22] Pharmaceuticals and Medical Devices Agency. https://www.pmda.go.jp/english/index.html.(Last accessed: 13. Dec.2021)
[23] Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa, and Makoto Iwayama, (2003) "Patent Claim Processing for Readability: Structure Analysis and Term Explanation," *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing*, 20: pp. 56-65.
[24] P. Parapatics and M. Dittenbach, (1990) "Patent Claim Decomposition for Improved Information Extraction," *Proceedings of the 2nd International Workshop on Patent Information Retrieval*: pp. 33-36.
[25] S. Sheremetyeva, S. Nirenburg, and I. Nirenburg, (1996) "Generating patent claims from interactive input,"*Proceedings of the 8th International Workshop* on Natural Language Generation: pp. 61-70.
[26] M. Verma and V. Varma, (2011) "Applying Key Phrase Extraction to Aid Invalidity Search," *Proceedings of the 13th International Conference on Artificial Intelligence and Law*: pp. 249-255.
[27] M. A. Hasan, W. S. Spangler, T. Griffin, and A. Alba, (2009) COA: Finding Novel.
[28] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, (2018) "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv*:1810.04805.
[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, (2013) "Efficient estimation of word representations in vector space,"*arXiv preprint arXiv*:1301.3781.

AUTHOR

Koichi Kamijo received B.E. degree in Electrical Engineering from the University of Tokyo in 1985, M.E. degree in Computer Science from Cornell University in 1996, and Ph.D. degree in Engineering from the University of Tokyo in 2010, respectively. After having worked at IBM Research-Tokyo for more than 20 years, he is currently a professor of International Professional University of Technology in Tokyo, department of Information Technology, since 2020. He is interested in Image Processing, Human Interaction, Digital Rights Management, Machine Learning, and Natural Language Processing. He holds 98 issued patents all over the world.