MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1020

Auto-selection of DRG Codes from Discharge Summaries by Text Mining in Several Hospitals: Analysis of Difference of Discharge Summaries

Takahiro Suzuki[a], Shunsuke Doi[b], Gen Shimada[c], Mitsuhiro Takasaki[d], Toshiyo Tamura[b], Shinsuke Fujita[e], Katsuhiko Takabayashi[a]

[a] *Department of Medical Informatics and Management, Chiba University Hospital, Japan*
[b] *Graduate School of Engineering, Chiba University*
[c] *Medical Information Center, St. Luke's International Hospital, Japan*
[d] *Division of Medical Informatics, Saga University Hospital, Japan*
[e] *Department of Welfare and Medical Intelligence, Chiba University Hospital*

Abstract

Recently, electronic medical record (EMR) systems have become popular in Japan, and number of discharge summaries is stored electronically, though they have not been reutilized yet. We performed text mining with Tf-idf method and morphological analysis in the discharge summaries from three Hospitals (Chiba University Hospital, St. Luke's International Hospital and Saga University Hospital). We showed differences in the styles of summaries, between hospitals, while the rate of properly classified DPC (Diagnosis Procedure Combination) codes were almost the same. Beyond different styles of the discharge summaries, text mining method could obtain proper extracts of proper DPC codes. Improvement was observed by using integrated model data between the hospitals. It seemed that huge database which contains the data of many hospitals can improve the precision of text mining.

Keywords:

Text mining, Discharge summary, Electronic medical record.

Introduction

With the spread of recent hospital information systems, the discharge summary begins to be saved electronically at many hospitals in Japan. However, every hospital has its own style of discharge summaries. In addition, an inconsistent form is used in every hospital, so there is no suitable study samples comparisons of discharge summaries among different hospitals.

In Chiba University Hospital, full text computerization of electronic discharge summaries began in 1999. We reported to past MEDINFO congress that useful information was obtained from summaries by text mining [1-3]. However, we were not able to conclude that the results were generally valid, since the styles of summaries are different. Therefore, it is necessary to review whether our conventional method can be applied to other hospitals. We tried a cross-sectional comparison experiment for the discharge summary of the St. Luke's International Hospital and the Saga University hospital. All these hospitals are major hospitals in Japan.

Using text mining we performed an experiment to select DPC (Diagnosis Procedure Combination) from the discharge summary. DPC or so called Japanese version of DRG (Diagnosis Related Group) is a classification that has become the basis of an inclusive evaluation system of hospitalization health care cost.

Figure 1 shows the structure of the DPC code. The first two digits are the Major Diagnostic Category (MDC) and the next four digits stand for the disease name. The remaining eight digits indicate the admission purpose, age, operation, treatment, complication, and severity.

DPC code is given to the discharge summaries of all hospitals. Therefore, we consider that we can compare between different hospitals using a DPC code. We examined the differences in the structure of the discharge summaries of each hospital by examining the type and the frequency of the extracted terms.

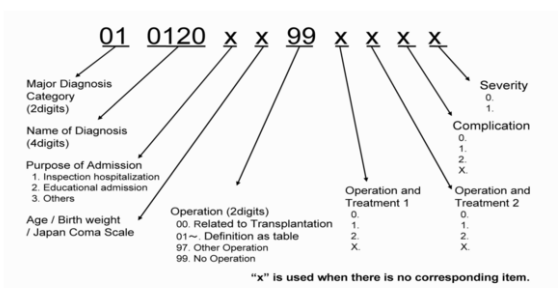

Figure 1 –Sample of DPC code "Idiopathic mono neuropathy; no operation"

Materials

Discharge Summary

We have analyzed discharge summaries of patients, who had been discharged from Chiba University Hospital, St. Luke's International Hospital and Saga University Hospital between April, 2006 and March, 2008. All analyzed discharge summaries had been written and stored electronically and had been

assigned a DPC code. The numbers of discharge summaries extracted from each hospital are as follows.

1. Chiba University Hospital: 24,594
2. St. Luke's International Hospital: 24,002
3. Saga University Hospital: 1,188

Problems with discharge summaries in each hospital

Each of the three hospitals has original discharge summary formats. We describe the general conditions and problems with the discharge summaries of the hospitals in the following.

Chiba University Hospital

A chief physician inputs the discharge summary as free text within two weeks after patient's discharge. The entry item is standardized, but there is no clear standard about the length of a summary, and a significant difference occurs between departments.

St. Luke's International Hospital

Both entry item and length are standardized. Because the length tended to get too long before, they set up one page limit. Their aim was a concise summary which could be read within twenty seconds.

Saga University Hospital

A chief physician can extract necessary information from the electronic patient record semi-automatically to make a discharge summary and convenient for physicians.

Methods

Morphological Analysis and Reconstruction of the Dictionary

Morphological analysis that decomposes a character string into elements such as a noun, adjective, and particle, is necessary for the analysis of Japanese sentences. Chiba University Hospital and St. Luke's Hospital have improved the precision of the Japanese medical dictionary conventionally for the process of morphological analysis. We use the user's medical dictionary of both hospitals together in this study.

The PHYXAM dictionary was used as a Japanese medical technical dictionary for the physical examination field [4]. Terms from the master table of both hospitals were added to the dictionary. The unknown terms from the discharge summaries were also added to the dictionary. Then our dictionaries included about 320,000 terms.

We used Mecab ver0.96 developed at Kyoto University information science graduate course as a morphological analysis system for Japanese sentences [5].

Investigate the characteristics of the summaries

We compared the following points among hospitals. We extracted only the nouns to characterize a sentence, and counted the number of terms and their type that appear in each summary.

Number of terms included in summary

We compare a difference of the summaries by the number of terms which were extracted by morphological analysis. We examined dispersion of the number of terms by values of mean, median, standard deviation, maximum and minimum. Normality was certificated by the Kolmogorov-Smirnov official.

Comparison by MDC

Discharge summary may be different between different medical departments, because the day of hospitalization, surgery and treatment varies according to a disease.

MDC is called major Diagnostic Category, and is expressed with the first two digits of the DPC code. MDC divides all diseases into 16 categories of a macrotaxonomy.

The hospitals we intended for this study were general hospitals with a vast array of medical departments. Therefore we could extract MDC from all discharge summaries (Table 1).

Table 1 - Disease of MDC

MDC	Disease
01	Nervous system disease
02	Ophthalmologic disease
03	Otorhinolaryngological disease
04	Respiratory disease
05	Circulatory disease
06	Digestive system disease
07	Musculoskeletal system disease
08	Skin and Subcutanea disease
09	Breast disease
10	Endocrine and Metabolism disease
11	Urogenital system disease
12	Gynopathy and Obstetrical disease
13	Blood and Immunological disease
14	Anomary and Newborn infant disease
15	Pediatric disease
16	Injury, Toxicosis and Other diseases

Comparing the precision of DPC selection

Grouping of data

We arranged the summaries of each hospital according to a discharge date and divided them into two groups in a ratio of 7:3, each of which had at least 10 cases in St. Luke's International Hospital and Chiba University Hospital. The first group was collected to generate document vector space model according to the DPC, the second group was collected as a test group to verify automatic DPC selection. In Saga University Hospital, all the summaries were assigned to a second group. In this way, we selected 20,013 cases for this study. The cases contained 97 different DPC codes. We show the number of the cases in Table 2.

Table 2 - Number of summaries

	Data for model	Data for verify
Saga university hospital	0	218
St. Luke's International Hospital	7,421	3,197
Chiba University Hospital	6,416	2,761
Total	13,837	6,176

Vector Space Model

The vector space model, which converts the characteristics of a document into a multidimensional vector, is a technique widely used in the field of information retrieval [6, 7]. In this study, we calculated the weights of each term to convert the characteristics by term frequency–inverse document frequency (tf-idf) method. The targeted discharge summary set is assumed to be $\boldsymbol{D}$, and the discharge summary sets of each disease are assumed to be $\boldsymbol{d_1, d_2, \cdots, d_j, \cdots, d_n}$. Next, $\boldsymbol{m}$ pieces of index terms extracted from $\boldsymbol{D}$ are assumed to be $\boldsymbol{\omega_1, \omega_2, \cdots \omega_i \cdots \omega_m}$. Thus, the weight in the discharge summary $\boldsymbol{d_j}$ added to index term $\boldsymbol{\omega_i}$ is assumed to be $\boldsymbol{\alpha_{ij}}$. Then, discharge summary $\boldsymbol{d_j}$ can be expressed by a matrix composed of $\boldsymbol{m}$ pieces of element $\boldsymbol{\alpha}$. Consequently, the sets of discharge summary $\boldsymbol{D}$ can be defined as a collection of matrices of $\boldsymbol{d_n}$, as shown in Figure 2.

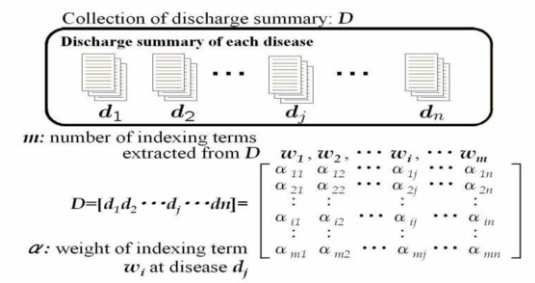

Figure 2 - Vector space model

tf-idf Method

In tf-idf method, weight $\boldsymbol{\alpha_{ij}}$ is shown by $\boldsymbol{\alpha_{ij}=(l_{ij}\times g_i)/n_j}$.

$$\boldsymbol{\alpha_{ij}=(l_{ij}\times g_i)/n_j} \quad (1)$$

$$\boldsymbol{l_{ij}=log(1+f_{ij})} \quad (2)$$

$\boldsymbol{l_{ij}}$ is local weight, and it is calculated based on the frequency data in document D_j of index term $\boldsymbol{\omega_i}$ and a big value is given to the index term that appears frequently in the document. Frequency data $\boldsymbol{f_{ij}}$ in discharge summary $\boldsymbol{d_j}$ of DPC classification $\boldsymbol{j}$ of index term $\boldsymbol{\omega_i}$ was used this time.

$$\boldsymbol{g_i=log(n/n_i)} \quad (3)$$

$\boldsymbol{g_i}$ is global weight, and it is calculated based on the index term $\boldsymbol{\omega_i}$ distribution over the entire document set, and a big value is given to the index term that appears only in a specific document. $\boldsymbol{n}$ is total number of the discharge summary, and $\boldsymbol{n_i}$ is a number of discharge summaries including index term $\boldsymbol{\omega_i}$.

$\boldsymbol{n_j}$ is a document normalization coefficient to which removes the influence by the length of the discharge summary.

DPC Selection

We determined the DPC code from discharge summaries based on the calculated vectors. We calculated the inner products of the vectors to compare the similarity between each summary and the DPC code of a model group. When we select a DPC code of the summary for inspection, we use a DPC code from models that have high similarity. However, as for just using a DPC code of the summary that shows a highest similarity, is influenced by accident. Therefore we judged a DPC code of a test summary by a weighting point calculation of multiple summaries (Figure 3)

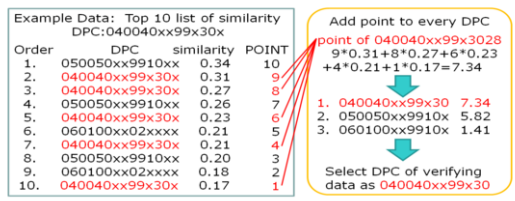

Figure 3 - DPC selection by weighting point

At first we select the top 10 summaries from a model group by similarity, and calculate a weighting point based on order and ratio. We expressed the weighting point (P) by the product of order (Rank) and similarity ratio (S (A, B)).

$$\mathbf{P\ (Rank) = S\ (A, B) \times (11-Rank)} \quad (4)$$

Then we selected a DPC code of which sum of points was the greatest among DPC code test data. By this technique, precision was improved from 2% to 5% (Data not shown). Next, we counted the cases in two groups. Selected code was exactly the same as the original code of test case in the first groups, and it matched the first 6 digits of the code in the second group.

Cross match selection

We replaced or integrated the data of multiple hospitals and carried out the selection experiment as follows.

- We verified test data with their own model data. (experiment No.1,2)
- We verified test data with other hospital's model data. (experiment No. 3 - 6)
- We verified test data with integrated model data (experiment No. 7-9)

Results

Comparison by morphological analysis

Comparison by number of terms

We have shown the number of terms in the discharge summaries from each hospital in Table 3. The average number of terms is almost the same between St Luke's International hospital and Saga university hospital, however about twice that of Chiba university hospital. St. Luke's International Hospital has a small difference of the median and the mean, and the standard deviation is small, too, Chiba University Hospital has a large standard deviation and the median is smaller than the mean. In Saga University Hospital, the gap is smaller than that of Chiba University Hospital, but standard deviation is the largest among the three hospitals.

The normality of all hospitals was dismissed by the Kolmogorov-Smirnov certification, however St. Luke's International Hospital is the nearest in normal distribution.

Table 3 - number of terms in discharge summary

	Chiba University Hospital	St. Luke's International Hospital	Saga university hospital
mean	136.4	285.0	281.7
median	79	295	247
SD	164.4	123.5	181.5
minimum	1	24	5
maximum	1921	830	1261

As shown in Figure 4, the numbers of cases according to the number of terms in one summary are in inverse proportion at Chiba University Hospital. By contrast, they are distributed in the neighborhood of a mean at St. Luke's International Hospital. Saga University Hospital was in the middle.

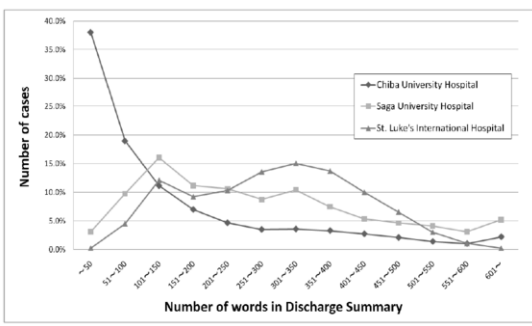

Figure 4 - Number of cases according to the number of terms

Comparison among number of terms by MDC

Figure 5 shows the comparison of the average number of terms according to MDC. Pattern of St. Luke's International

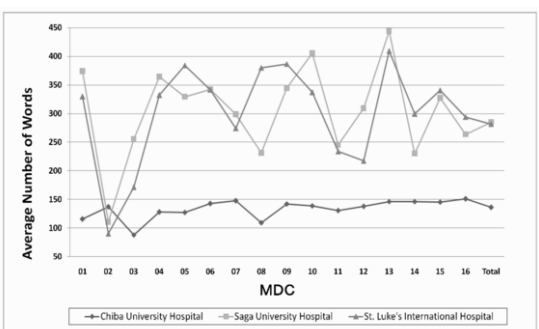

Figure 5 - Comparison among number of terms by MDC

Hospital resembles Saga University Hospital except 06 (Digestive system disease). Chiba University Hospital has a smaller number of terms and smaller difference between diseases than the other hospitals.

Comparison by precision of DPC selection

Selection by the data of same hospital

At first, we show a DPC selection rate where we used the model data of the same hospital for verification (see Table 4). Selection rate at Chiba University Hospital and St. Luke's International Hospital is approximately of equal value. These results show that the application of the method is independent of each institute.

Table 4 - DPC selection by same hospital data

No.	Data for selection		Precision of DPC selection	
	verify	model	Match 14 digit	Match 6 digit
1	Chiba (2761)	Chiba	75.9% (2097)	85.8% (2370)
2	St. Luke's (3197)	St. Luke's	78.9% (2524)	84.7% (2708)

Selection by the data of different hospital

Next, we show a cross match selection where we used the model data of different hospitals (see Table 5). For Chiba University and St. Luke's International Hospital, it seems that the selection rate falls about 10-20% in comparison to experiment 1, 2.

Table 5 - Selection by the data of different hospital

No.	Data for selection		Precision of DPC selection	
	verify	Model	Match 14 digit	Match 6 digit
3	Chiba (2761)	St. Luke's	62.5% (1725)	73.4% (2029)
4	Saga (3197)	Chiba	61.9% (1979)	68.2% (2182)
5	Saga (218)	St. Luke's	56.4% (123)	72.0% (157)
6	Saga (218)	Chiba	50.9% (111)	63.7% (139)

Selection by the integrated data

Finally, we show the DPC selection rate when using the integrated model data of Chiba University and St. Luke's International Hospital and verify data of each hospital (see Table 6).

Precision of St. Luke's International Hospital was the same as the result of experiment 2 where only its own data was used. The results of Chiba University hospital slightly improved in comparison with experiment 1.

The data of Saga University hospital also improved by using integrated data. The ratio extracted from the two hospitals was almost the same (Chiba 54%, St. Luke's 46%).

Table 6 - Selection by the data of integration

No.	Data for selection		Precision of DPC selection	
	verify	model	Match 14 digit	Match 6 digit
7	Chiba (2761)	Chiba + St. Luke's	76.5% (2114)	85.9% (2374)
8	Saga (218)	Chiba + St. Luke's	56.8% (124)	70.6% (154)
9	St. Luke's (3197)	Chiba + St. Luke's	77.1% (2465)	82.7% (2644)

Discussion

In the medical field, text mining applied to clinical contents are still rare [8]. One of the reasons is the lack of accumulated electronic medical documents to be analyzed, although there is a suitable target, Medline, which integrates numerous medical abstracts [9,10].

We described the different and common characteristics among the three hospitals. The length of the summary at Chiba University Hospital is half of that at other hospitals. At St. Luke's international and Saga University Hospital, about 30% of the summary surpasses 400 words. In Saga University hospital about 30% of the summaries have less than 150 words, indicating a big difference of the summaries. The length of the discharge summaries was also different by MDC. For example, 03 (otolaryngology diseases) have fewer words, adversely 13(blood/immune organ diseases) has many words. It was considered that these were common characteristics of the diseases and independent from hospitals.

The DPC selection rate at Saga University hospital is not high, implying that the terms of their summaries were not included in a dictionary. Difference of term structure of the summaries of each hospital decreased the precision of the DPC selection when we use the model data of another hospital. However the difference is not so large that we can consider the vector space model by tf-idf method selects a DPC independent from an institution. Improvement of the precision was observed when we use integrated data of hospitals. It suggests the possibility to improve precision by correct summaries from many hospitals.

When an integrated database of discharge summaries beyond hospitals is available, text mining will provide us with various application, such as acquisition of knowledge [11], similar case search [12], automatic coding of findings [13], extracts cancer staging from pathology reports automatically [14], and making classification automatically [15, 16] as well as comparison of many quality indicators between facilities.

Conclusion

Using a vector space model by the text mining method, we carried out a DPC selection based on the discharge summary from multiple hospitals. We have shown by morphological analysis that there was a difference in term structure among the discharge summaries of each hospital.

We were able to carry out a DPC selection independently of hospitals. Furthermore, improvement was observed by using integrated model data between the hospitals. A giant database which contains the data of many hospitals could improve the precision of text mining.

References

[1] Ono H, Takabayashi K, Suzuki T, Yokoi H, Imiya A, Satomura Y. Extraction of diagnosis related terminological information from discharge summary. Medinfo. 2004; (CD): 1786.

[2] Suzuki T, Yokoi H, Fujita S, Takabayashi K. Discharge Summaries can be diagnosed from extracted index terms by text mining. Medinfo. 2007; (CD): 1786.

[3] Suzuki T, Yokoi H, Fujita S, Takabayashi K. DPC Code Selection from Electronic Medical Record -Text Mining Trial of Discharge Summary-. Methods Inf Med 2008; 47:541-548

[4] Fujita S. The interaction of the reason for encounter (ICPC-2) and standardized physical findings (PHYXAM). Proc. 2nd Annu Conf JAMI 2004;1;908-909

[5] Kudo T, Yamamoto K, Matsumoto Y. Applying conditional random fields to Japanese morphological analysis. Proc. EMNLP, 2004:230–237

[6] Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing. CACM 1975; 18:613-620.

[7] [Goldman JA. Chu WW. Parker DS. Goldman RM. Term domain distribution analysis: a data mining tool for text databases. Methods Inf Med. 1999;38:96-101

[8] Collier N, Nazarenko A, Baud R, Ruch P. Recent advances in natural language processing for biomedical applications. Int J Med Inform, 2006, 75:413-417.

[9] Srinivasan P. MeshMap: A textmining tool for Medline. Proc AMIA Symp. 2001;642-646.

[10]Mendonca EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. Proc AMIA Symp. 2000; 575–579.

[11]Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. JAMIA. 2008; 15:87-98.

[12]Takemura T, Sato J, Kuroda T, Nagase K, Takada A, Tanaka K, Guo J, Yoshihara H. Development of the retrieval system of similar discharge summary in MML (Medical Markup Language) instance. Proc 5th Annu Conf JAMI 2004;1:464-465

[13]Mamlin BW, Heinze DT, McDonald CJ. Automated Extraction and Normalization of Findings from Cancer-Related Free-Text Radiology Reports. AMIA Annu Symp Proc. 2003. 2003; 420-424.

[14]McCowan IA. Moore DC. Nguyen AN. Bowman RV. Clarke BE. Duhig EE. Fry MJ. Collection of Cancer Stage Data by Classifying Free-text Medical Reports. JAMIA. 2007;14:736-745.

[15] Pakhomov SV, Ruggieri A, Chute CG. Maximum entropy modeling for mining patient medication status from free text. Proc AMIA Symp. 2002;587-91.

[16] Iwahashi Y, Ohe K. Trial of automating classification of incident reports. Proc 2nd Annu Conf JAMI 2004;1:804-805

Address for correspondence

Takahiro Suzuki, MD, PhD, FJSIM
1-8-1 Inohana, Chuo-ku, Chiba, 260-8677 Chiba, Japan
E-mail: suzuki@ho.chiba-u.ac.jp