Analysis on Drug Dosage Form Name Based on N-gram Technique and Network Analysis

Masaomi Kimura[1] and Fumito Tsuchiya[2]

[1] Shibaura Institute of Technology, 3-7-5 Toyosu, Koto City, Tokyo 135-8548, Japan
[2] International University of Health and Welfare, 2600-1 Kitakanemaru, Otawara City, Tochigi, 324-8501, Japan

Abstract. In this paper, we analyzed drug dosage form names. We created the network structure whose nodes are dosage form names. Its edges between dosage form names denote that they share some of sub-strings generated based on N-gram technique. We employed Simpson coefficient to define the weight of an edge. We proposed a new clustering method and applied it to the network. The results showed that "dosage forms" can be categorized based on not only physical form information but their application site,purpose,processing and so on.

Keywords: Medical safety, Dosage form, N-gram, Network analysis.

1 Introduction

Drugs have many kinds of forms, such as tablets, capsules, powder, intradermal injection and so on, which are categorized in "dosage forms". The information related to dosage forms is important to ensure the safety of medicinal usage, because the drugs with different dosage forms make their usage and effectiveness strength different, even if they have the same active ingredient.

Naively, we expect the word "dosage form" to indicate physical shapes of drugs. However, the names of dosage forms usually contain other information such as the routes of administration (*e.g.* oral powder) and properties (*e.g.* sustained-release tablets).

In Japan, there is a series of drug spec definition books called"Japanese Pharmacopoeia" [1]. It includes general rules for preparations, which define drug dosage forms. It is described that they are primarily categorized by the routes of administration / application sites of a body, and secondly categorized by physical forms, functions and properties. Though the categories are defined by medical experts but are not necessarily used as standard categories of dosage forms. In fact, there are no standard dosage form categories. This might originate in the variety of information contained in the names of dosage forms.

In this paper, we show the results of analyses applied to the names of dosage forms to identify the structures of the contained information. We employed the technique based on N-gram method and a novel network clustering technique.

We regarded the name of dosage forms as character strings and obtained a set of their sequential substrings. One reason why we used N-gram method is

C. Stephanidis (Ed.): Posters, Part II, HCII 2013, CCIS 374, pp. 178–182, 2013.

that the most of the length of each dosage form is short and, therefore, it can be regarded to carry much information per letter. Another reason is that the physical forms are usually expressed in short letter strings. Our strategy is to regard dosage form names as nodes in a network and the pairs of sharing the same sequential substrings as edges. We also introduced the weight of edges based on Simpsons coefficient.

The existing clustering methods have some difficulties. The traditional ones are maximization of modularity and spectral clustering. The difficulty of modularity maximization is known as resolution limit and the one of spectral clustering comes from the interpolation of discrete cluster label (integer) to real number. We propose the method that assigns the quantities to (nearly) diagonalize the adjacent matrix of a network. The nodes that have the same value of the quantities constitute a community of the network.

2 Target Data

We used 158 dosage form names listed in a code table used for Japanese drug approval, which is disclosed by Japanese authority, Ministry of Health, Labour and Welfare [2]. For simplicity, we omitted "日局" (official recipes for Japanese drugs), "その他の" (others) , and words in parentheses, which show additional information. We identified the same dosage form names after the omission, and separated plural drug names if they are bundled in one name. As a result, we obtained 119 target dosage form names.

3 Methods

We utilized N-gram method and extracted all substrings in each of dosage form names. Since a dosage form name consists of plural components which indicate physical forms, applied body parts and so on, we assumed that the more common substrings the dosage forms have, the more similar they are. Based on this policy, we propose the analysis method that connect a pair of dosage form names sharing the components with an edge and apply graph partitioning (community findings) to the resultant network [1].

Naive application of graph partitioning to the network does not reflect the commonality extent of components. Therefore, after connecting dosage form names with edges, we assigned weight to each edge. In order to measure the commonality, we used Simpson coefficient,

$$w_{ij} = \frac{|X_i \cap X_j|}{\min(|X_i|, |X_j|)}, \tag{1}$$

where X_i denotes the set of substrings generated by N-grams. We set $w_{ij} = 0$, if $\min(|X_i|, |X_j|) < \epsilon$ for some threshold ϵ.

[1] We should note that our target network is undirected and simple.

The standard methods of graph partitioning/community finding are Spectral clustering and Modularity maximization proposed by Newman et. al [3, 4]. The former needs to define the number of communities in advance and the latter has un-resolved problem, named as "resolution limit". We, therefore, propose another graph partitioning method. In order to get good partition, we arrange the well-connected nodes to belong to the same community. In terms of an adjacency matrix, its rows/columns corresponding to well-connected nodes should be neighbor, and this requires that non-zero elements of the adjacency matrix of the network get together to diagonal elements. In order to quantify this, we assigned the position x_i to Node i and defined covariance and variance of x_i as followings:

$$\sigma_{xx} = \sum_{i,j} \frac{A_{ij}}{2M}(x_i - \bar{x})(x_j - \bar{x}), \tag{2}$$

$$\sigma_x^2 = \sum_{i} \frac{k_i}{2M}(x_i - \bar{x})^2, \tag{3}$$

where A_{ij} is an adjacency matrix, k_i is the degree of Node i ($\sum_i k_i = 2M$), and $\bar{x} = \sum_i \frac{k_i}{2M} x_i$.

We assigned a series of $\{x_i\}$ to maximize the correlation coefficient,

$$r = \frac{\sigma_{xx}}{\sigma_x^2}. \tag{4}$$

As for the resultant series $\{x_i\}$, the value x_i for the well-connected nodes should be identical, since the position of non-zero elements of the adjacency matrix, (x_i, x_j), are almost along the line $y = x$, namely, $x_i \simeq x_j$. Therefore, after sorting the series $\{x_i\}$, we can find communities as "plateaus" of x_i values. We generalized this by utilizing weight matrix w_{ij} instead of adjacency matrix A_{ij}.

4 Results

Table 1 shows the resultant $\{x_i\}$. This shows that the similar dosage form names have similar values of x_i.

What is interesting is "絆創膏" (adhesive bandage) and "軟膏剤" (ointment) respectively belong to different clusters, though they share the Kanji character "膏" (plaster).

Contrary to this, "凍結乾燥注射剤" (freeze dry injection) and "粉末注射剤" (powder injection) belong to the same community but belong to the other community of liquid injections. It is interesting that the two dosage forms are in the same community, though they do not share the same substrings other than "注射剤"(injection). We note the fact that the word "性" (property) is contained in the dosage form names of liquid injections. This seems to divide liquid injection from other type of injections. This suggests that we can interpret the community which freeze dry injection belongs to as the community corresponding to the non-liquid injections.

Table. 1. The dosage form names and their x_i values. The values of x_i are normalized so that their maximum is 1 and minimum is 0. "Dosage Form (J)" in the header indicates original Japanese dosage form names, and "Dosage Form (E)" indicates corresponding English dosage form names.

Dosage Form (J)	Dosage Form (E)	x_i	Dosage Form (J)	Dosage Form (E)	x_i
気体	gas	1.00	腸溶性細粒	enteric-coated fines	0.08
液体	liquid	0.97	徐放性顆粒	extended-release granule	0.08
輸液	transfusion	0.93	腸溶性顆粒	enteric-coated granule	0.08
液絆	liquid plaster	0.87	注入剤	injectable filler	0.08
鉱物生薬	metal crude drug	0.83	浣腸剤	enema	0.08
薬品付絆創膏	medicated adhesive plaster	0.67	徐放カプセル	extended-release capsule	0.07
絆創膏	adhesive bandage	0.66	腸溶カプセル	enteric-coated capsule	0.07
ガーゼ付絆創膏	gauze plaster	0.64	コーティング細粒	coated fines	0.07
ハッカゴム膏	mint gum plaster	0.59	コーティング顆粒	coated granule	0.07
硬膏	emplastrum	0.57	徐放錠	extended-release tablet	0.07
軟膏剤	ointment	0.57	腸溶錠	enteric-coated tablet	0.07
軟稠エキス	soft extract	0.52	バッカル錠	buccal tablet	0.07
軟カプセル	soft capsule	0.52	内用細粒	internal fines	0.07
眼軟膏	eye ointment	0.46	コーティング錠	coated tablet	0.07
洗口うがい剤	mouth rinse collutorium	0.20	内用素顆粒	internal granule	0.07
洗浄・清拭剤	cleaner	0.20	かみ砕き錠	chewable tablet	0.07
洗眼剤	eye wash	0.20	重層錠	multi-layered tablet	0.07
尿道坐剤	urethral bougie	0.18	有核錠	pressure-coated tablet	0.07
膣坐剤	vaginal suppository	0.18	外用顆粒	external granule	0.06
肛門坐剤	rectal suppository	0.18	内用素錠	internal tablet	0.06
コロジオン剤	collodion preparation	0.18	内用発泡錠	internal effervescent tablet	0.06
グリセリン剤	glycerin preparation	0.18	顆粒	granule	0.06
乳剤性点眼剤	emulsion eye drop	0.17	外用発泡錠	external effervescent tablet	0.06
水性点眼剤	aqueous eye drop	0.17	外用錠	external tablet	0.06
非水性点眼剤	nonaqueous eye drop	0.17	内用	internal	0.06
懸濁性点眼剤	suspension eye drop	0.17	内用エアゾール	internal aerosol	0.06
懸濁剤	suspension agent	0.14	外用	external	0.06
芳香水剤	aromatic water	0.13	内用散剤	internal powder	0.06
皮膚用水剤	dermatological water	0.13	外用散剤	external powder	0.05
懸濁性注射剤	suspension injection	0.11	散剤	powder	0.05
非水性注射剤	nonaqueous injection	0.11	体外ガス剤	external gas	0.05
水性注射剤	aqueous injection	0.11	外用エアゾール剤	external aerosol	0.05
乳剤性注射剤	emulsion injection	0.10	組み合わせ剤	combination drugs	0.04
粉末注射剤	powder injection	0.09	噴霧・吸入剤	inhalation	0.03
凍結乾燥注射剤	freeze dry injection	0.09	ペンシル剤	pencil	0.03
粉末剤	powdered drugs	0.09	吸入ガス剤	inhalation gas	0.03
粉末状エアゾール	powdered aerosol	0.09	吸入型エアゾール剤	inhalation aerosol	0.02
乾燥エキス	dry extracts	0.09	綿吸着剤	cotton adsorbent	0.00
注射錠皮下埋没用	implantation injection tablet	0.09	紙吸着剤	paper adsorbent	0.00
注射錠	injection tablet	0.09	合成樹脂吸着剤	synthetic-resin adsorbent	0.00
徐放性細粒	extended-release fines	0.08	ガーゼ吸着剤	gauze adsorbent	0.00

We should also note another point. The community whose $x_i = 0.06$ contains several different physical dosage forms, such as tablets, capsules and fines. The dosage form names in the community commonly contain the substrings "外用" (outer) and "内用"(inner). This suggests that view points to categorize dosage forms are not necessarily limited to physical dosage forms but can be others, such as body parts which drugs are administered to.

These results suggest that Japanese dosage form names do not necessarily focus on drugs' physical forms. This is consistent with the categorization policy

of Japanese Pharmacopoeia, primarily categorization by the routes of administration and secondly categorization by physical forms.

5 Conclusion

In this paper, we analyzed the dosage form name of Japanese drugs. We used the technique based on N-gram method and a novel network clustering technique.

We tend to associate the word "dosage form names" with physical form of drugs. However, our results suggest that dosage form names should be categorized by body parts which drugs are administered to or their properties prior to categorization by physical drug forms.

In future works, we will propose the coding system of dosage form names based on our results.

References

Ministry of Health, Labour and Welfare: Japanese Pharmacopoeia (2012), http://www.mhlw.go.jp/topics/bukyoku/iyaku/yakkyoku/

Ministry of Health, Labour and Welfare: FD Application Related to Approve of Drugs, Cosmetics and Medical Equipments (2012), http://web.fd-shinsei.go.jp/download/software/index.html

Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. 103, 8577–8582 (2006)

Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74, 036104 (2006)