資料７

学会発表

処方入力画面に表示される医薬品名について

木村 昌臣[1)] 鍋田 啓太[2)] 土屋 文人[3)]
芝浦工業大学[1)] 芝浦工業大学大学院[2)] 東京医科歯科大学[3)]

The name structure of medicines which should be shown on prescription entry screen

Kimura Masaomi[1)] Nabeta Keita[2)] Tsuchiya Fumito[3)]
Shibaura Institute of Technology[1)] Graduate School, Shibaura Institute of Technology[2)]
Tokyo Medical and Dental University[3)]

The safety of medical usage is one of keys to prevent from medical accidents caused by medicines. In order to ensure the safety, there have been made reminders of medicines whose name and/or visual appearance of package is similar to others, or an effort to reject such medicines. They are not enough as countermeasures, since medical accidents resulting from medicines occur in spite of the efforts. Because of this, we make a study to create database which contains data/information to ensure the safety of medical usage and to construct a prescription checking system based on the database.

In this paper, we report the result of study about the method to extract information contained in medicine names which can help to verify users' input of the medicine names to prescription input system.

Keywords: Safety of medical usage, Natural Language Processing, Medicine names

1. はじめに

医療事故を防止するためには正しい医薬品を正しく使用するためのいわゆる医薬品使用の安全性を保障する必要がある.

筆者らは医薬品の使用の安全性の観点から, 医薬品に関する医療事故の発生情報の解析およびその予防策の検討を行ってきた.独立行政法人　医薬品医療機器総合機構は医薬品に関連するヒヤリ・ハット事例を公開しているが,これを解析することにより,薬剤を取り違えた事例は医薬品の名称や外観が類似していることが要因となっていることが多いという結果を得ている[1)].この意味で、医薬品名称のうち類似しており紛らわしいものについては,取り違えられる可能性を内在していると考えることができる.実際,医薬品類似名称検索システム[2)]が構築され,新規医薬品の名称に関してはこのシステムにより類似した名称が既にないことを確認することになっている.

処方オーダリングシステムの処方入力画面において,処方されるべき医薬品を特定するためには医薬品名称を何らかのかたちで入力する必要があるが,その際の入力ミスは当然ながら医療事故に直結する原因となりうる.そのため,入力された医薬品名が正しいものかをチェックする仕組みが必要とされ,実際に筆者らは医薬品の使用の安全性を担保するための情報をまとめたデータベースおよびそれにもとづく処方のチェックを行うシステムの検討を行っているところである.

医薬品の名称に含まれる内容を構成する要素については,販売名は前修飾,語幹,後修飾,剤形,規格単位を含む23要素から構成されているという報告があり[3)],この中で医薬品の指定に関して特に重要なのは語幹・剤形・規格単位である.実際,語幹部分を指定しても複数剤形や複数規格がある場合には用法・用量が異なるため,一意に医薬品を特定するためには少なくともこれらの情報が必要となると考えられる.しかしながら,また、医薬品名称に含まれているこれらの情報の有無や表記に現れる順番には必ずしも統一性があるとはいえない[3-5)].そのため,医薬品名称に含まれるべきこれら三つの情報を処方オーダリングシステムを含むコンピュータシステムで取り扱うためには,その記述を構成部分に分解・整理し,マスタ化したうえで医薬品名称と構成部分との間で求められた対応をするなどの方法をとる必要があるが,現状ではこれを実現することができるマスタは存在しない.

また,先に挙げた医薬品類似名称検索システムでは,医薬品の製品名の語幹部分について類似したものを検索する機能が提供されているが,医薬品名称の語幹部分が自動的に取得できれば,この医薬品類似名称検索システムで利用するデータを自動抽出できる仕組みとしても活用できることが期待される.

そこで本稿では,これらのマスタを試作し,これをもとに実際の医薬品名称から語幹・剤形・規格単位を抽出することを試みることにより,医薬品名称を処方オーダリングシステム等のコンピュータシステムで扱う際の問題点について検討したので報告を行う。

2. 解析データ

本研究では医薬品各製品別に扱うため,HOT9番号を基準として件数を集計する.また,医薬品名称のマスタとしてMEDIS標準医薬品マスター2009年7月31日版[6)]を用い,販売名を解析対象とするためにその中の「公示名称」データを使用する.なお,名称の変更などに伴い変更前後のデータが共にマスタに含まれている場合があるが,その場合は名称の長さが長いもの(長さが等しい場合は先に現れるもの)のみを使用することとし,HOT9番号と医薬品名称の一対一対応を担保した.

規格単位マスタには,MEDIS標準医薬品マスターの「規格単位」データを括弧等の記号で分割し,さらに「数値+数値以外」のパターンを持つ文字列を抽出したデータを用いた.ただし,医薬品名称に現れる規格

単位には数値のみで単位がつかないものがあるため,該当パターンの文字列から数値のみを抽出したもの規格単位マスタに含めた.

剤形マスタは,H20年度厚生労働科学研究「国際化を踏まえた医薬品・医療機器の安全性情報の伝達に関する研究」の成果である医薬品辞書データにおける「基本剤形」データをもとに作成した.

本来であれば「語幹+剤形+規格単位」が望ましい組み合わせではあるが,後発品ではこれ以外に企業名が入る場合があり,また剤形として直接「注射液」と書かず「注射用」と書く場合がある.これらのものをカバーするために, 規格単位マスタ・剤形マスタ以外に,記号マスタ,括弧内文字列マスタと用途マスタを用意した.

記号マスタにはMEDIS標準医薬品マスターの「公示名称」データから漢字・片仮名・平仮名・アルファベット、数字以外の文字を抽出した文字群を,括弧内文字列マスタは記号マスタの括弧に相当するもので囲まれている文字列群(後発品の製薬企業名などが含まれる)を,用途マスタはMEDIS標準医薬品マスターの「公示名称」をデータ漢字・片仮名・平仮名・アルファベット・数字・その他の文字の5つの文字種単位で切り分け,「用」で終わる漢字文字列群(ただし「用」の直前の文字が別の文字種であればそのあとに「用」を付けた文字列とした)を収集して用いた.

また,それぞれのマスタにはデータに対応したコード番号を付加し,どの情報が医薬品名称に含まれているのかをコード番号に基づき把握できるように工夫している.

3. 解析手法

本研究では語幹部分を抽出するために,まず,医薬品名称から各マスタのデータ該当部分を抽出し,医薬品名称全体から該当部分を取り除くことによって語幹部分を取り出す方法を試みる.

3.1 マスタデータ該当部分との対応関係の取得

まず,医薬品名称に該当するマスタデータに含まれる部分文字列を抽出するが,各マスタデータのすべての組み合わせを網羅的にもとめて対応関係を求める方法を採ると,各マスタのデータ件数の積に等しい回数だけ医薬品名称との比較を行わなければならなくない.計算時間の効率化を図るために自然言語処理で用いられることが多いN-gram手法を活用し,医薬品名称に含まれる連続した部分文字列を取得したうえで,各部分文字列と一致するマスタデータを検索・取得する.ただし,例えば「サルソニン注射液」であれば,剤形マスタ内の「注」および「注射液」が含まれている部分文字列として該当するように,一方が他方の部分文字列であるが故に複数のマスタデータが該当してしまう場合がある.(「ロヒプノール注」のように注射液であることを略し「注」と記載するものもあるため,マスタはデータとして「注射液」だけでなく「注」も持たざるをえない.) このような二重の出力を避けるために,該当したマスタデータ同士を比較し、一方が他の部分文字列として含む場合には長いほうの文字列(「注」と「注射液」では後者)のみを残す操作を行った.

また,もし医薬品名称が語幹部分以外が剤形・規格単位・記号・括弧内文字列・用途から構成されているならば,以上の操作により得られた各マスタに含まれる文字列と一致した部分文字列を医薬品名称から削除すれば語幹部分が得られると考えられる.

各HOT番号に相当する医薬品の名称からマスタと一致した文字列を削除して語幹部分を抽出することを試みた.

3.2 語幹カナ部分における類似名称組の抽出

処方オーダリングシステムにおける医薬品名称入力の際の大きな問題のひとつは,類似名称をもつ医薬品を誤って入力してしまうことである.本研究では,3.2で取り出された語幹部分文字列のうち片仮名部分を取り出し類似度を計算した上で,類似度が高い組み合わせを抽出した.類似度としては既存の医薬品類似名称検索システムで利用されているエディット距離,先頭2文字末尾2文字の類似性(HTCO),先頭3文字の類似性(h3cos1)を用いた.なお,類似度の計算の際には類似している文字群「ァアヤカ」「ツツシミ」「ンソリ」「クケフワ」「エコユ」「スヌ」「ナメ」「テラ」「ィイ」「ゥウ」「ェエ」「ォオ」をそれぞれ同一文字として扱う処理を行ったのちに計算を行っている.

4. 結果・考察

4.1 マスタデータ該当部分との対応関係の取得

対象データであるMEDIS標準医薬品マスターにはユニークなHOT9番号が24,484個含まれていた.3.1で述べた操作を行い,語幹を抽出したところユニークな語幹を8,007個抽出することができた.

取得精度を評価するため,ランダムに200のHOT9番号を選択し,手作業にて正しく語幹部分が取得できているか確認した.その結果,178(89%)の医薬品名称について正しく語幹部分が取得された.正しく取得されなかった22医薬品名称のうち6はマスタデータの不足によるもの,1つは英数字だが規格単位を表さないものを規格単位として誤って取り除いたもの,15はマスタデータを部分に含むがそれを取り除くのは適していなかったものであった.

表1は,結果の一部を示したものである.以下に,結果を得て判明した問題点を挙げる.

1) 「(局)液化亜酸化窒素(エア・ウォーター)」からは「液化亜酸化窒素」が抽出されるべきだが「化亜酸化窒素」と抽出された.これは単に剤形「液」が消去されたことによるが,「液化」から「液」を除くと意味が理解できなくなる.そのため,マスタの文字列が含まれていてもその文字列を削除することを許さない文字列のリストを用意するなどの対応が必要となる.
2) 「キョーリンAP2顆粒」からは,語幹として「キョーリンAP」が抽出された.本来であれば「キョーリンAP2」が語幹として抽出されるべきであるが「2」は規格単位として抽出されてしまった.医薬品名称のアルファベット部やそれに付随する数値については統一された意味付けがなされていないため,当該部分をマスタに含めるのはコンピュータの自動処理にもとづく方法では困難である(個別に解釈してマスタに含

める必要あり).

3) 「ハイパジールコーワ点眼液0.25%」は,語幹として本来「ハイパジールコーワ」が取得されるべきだが,括弧内文字列マスタに「コーワ」が後発品の企業名部分として含まれているため,「ハイパジール」のみが抽出された.
4) 「点眼・点耳・点鼻液」には,これ自体に点眼液・点耳液・点鼻液である情報が含まれているが,剤形マスタに含まれる点眼液・点耳液とは一致しない.

表1 語幹部分抽出結果(一部)

hot9	whole_name	brand_name
100318502	1%ディプリバン注	ディプリバン
100471703	セニラン錠2	セニラン
100494601	ユーパン錠1．0mg	ユーパン
100679701	オパイリン錠	オパイリン
100760201	(局) スルピリン注射液	スルピリン
100793044	サンナックス錠	サンナックス
100937801	モノクロトン錠	モノクロトン
100957602	ニフランシロップ	ニフラン
101124101	ジブカルソー注	ジブカルソー
101219403	メシル酸ブロモクリプチン2．5mg錠	メシル酸ブロモクリプチン
101410501	ハタナジン錠	ハタナジン
101489102	(局) サイラゼパム錠0．5	サイラゼパム
101531701	インブロメン錠6mg	インブロメン
101555304	(局) ブドウ糖注10%PL「フソー」	ブドウ糖PL
101556007	(局) ブドウ糖注射液	ブドウ糖
101575102	リスパダール錠2mg	リスパダール
101600001	ネオアムノールシロップ	アムノール
101661104	(局) リドカイン注射液	リドカイン
101711308	フェネット錠250	フェネット
101732801	ロキシーン注	ロキシーン
101765601	アスポーラカプセル10	アスポーラ
101767001	ガスチロールカプセル10mg	チロール
101798401	ブスコム錠10mg	ブスコム
101930833	エペナルド錠50mg	エペナルド
101985801	アドソルボカルピン点眼液1%	アドソルボカルピン
102046501	フルコン0．1%点眼液	フルコン
102146203	FAD点眼液0．05%「ニットー」	AD
102224701	ケトチフェン点眼液0．05%「TO…	ケトチフェン
102284102	0．05%プリビナ液「チバ」	プリビナ
102306006	エスペタットAQ点鼻液	エスペタットAQ
102535403	ピンドロール錠1mg「日医工」	ピンドロール
102570504	ジソピラミド100mgカプセル	ジソピラミド
102638203	(局) トリクロルメチアジド錠2mg…	トリクロルメチアジド
102704403	アレリックス3mg錠	アレリックス
102768601	ツルセルビS	ツルセルビS

表2 剤形情報を含まない医薬品名称(一部)

hot9	whole_name
100313002	アネソキシン－50
100323918	(局) ブロムワレリル尿素
100566012	(局) フェノバルビタール
100673507	(局) アセトアミノフェン「ヨシダ」
100727531	(局) アスピリン．OY
100757217	(局) ※スルピリン(山善)
100772501	インテバンSP25
101699401	ネオペルカミン・S
104761536	(局) 重質酸化マグネシウム．OI
104835302	グリセリン浣腸「ムネ」30
105079008	(局) 複方ヨード・グリセリン「エビス」
105798033	(局) エタノール(ミツマル)
105908323	(局) ※フェノール(小堺)
106026301	山善酢酸鉛
106034801	※カンフル精(山善)

表3 規格単位情報を含まない医薬品名称(一部)

hot9	whole_name
101614706	ピーエイ錠
101646805	(局) 塩酸プロカイン注射液
101899801	フォリピロン錠
102447025	(局) 塩酸ドパミン注射液
103240601	ロニアンカプセル
103402801	(局) ベザフィブラート徐放錠
103449301	アエレックスカプセル
103798202	フスコデシロップ
103809501	アスゲン錠
103817008	(局) セネガシロップ「ケンエー」
104626702	※センブリ散(丸石)
104754705	(局) 乾燥水酸化アルミニウムゲル
104824701	(局) ※グリセリン(東海製薬)
104830804	グリセリン浣腸液東豊
104971801	スパカール細粒

また,本研究で用意したマスタには各項目にコード番号を付番しているので,各医薬品名称にどの情報が含まれれているかを抽出することが可能である.
医薬品の剤形に関する情報をもつ剤形マスタおよび用途マスタのデータを一つも含まない医薬品名称はHOT9番号の個数にして4,495個,規格単位マスタのデータを含まない医薬品名称は　HOT9番号の個数にして10,364個存在した.表2・表3にそれぞれの結果からランダムに抽出した結果を示す.本研究では,漢方薬や配合剤を分けて解析していないため,それらも規格単位マスタのデータを含まない医薬品名称に含まれてしまっているが,これを分離した上で本来あるべき規格単位が名称に含まれない医薬品を抽出する予定である.

4.2 語幹カナ部分における類似名称組の抽出

表4にhtco=1,h3cos1=1,エディット距離=1となる語幹組を示す.後発品の名称に含まれる一般名と商標名の組み合わせも多くみられるが,「タキソール」「タキソテール」・「アルマール」「アルマトール」の組も見受けられる.　今後は,医薬品の成分名をもとに一般名と商標名を切り分けた上で,商標名同士および商標名と一般名の組み合わせを明示したうえで,類似した名称組を自動抽出する仕組みを作成する予定である.また,現在稼働している医薬品類似名称検索システムではデータ作成の一部で手作業による処理を行っているが,マスタの更なる整備等により本研究の仕組みを改善していくことによりデータ作成を自動で行うことができると期待できる.

表4 類似名称をもつ語幹(カナ部分)の組み合わせ

カプトプリル	カプトリル
トリアゾラム	トリアラム
テストビロンデポー	テストロンデポー
デキストセラン	デキストラン
ラキソベロン	ラキソロン
ダイアート	ダイアコート
グリセリン	グリセロリン
ニカルジピン	ニカルピン
カルテオロール	カルテロール
タキソール	タキソテール
オメプラール	オメプラゾール
オフロキサシン	オフロキシン
エストリールデポー	エストリオールデポー
エストリール	エストリオール
アルマール	アルマトール
アラセプリル	アラセリル
クインスロン	クインロン
フラボキサート	フラボサート
スカノーゼリン	スカノーリン
フェニタレン	フェニレン
フェノール	フェノバール
ベルベゾロン	ベルベロン
フルコナール	フルコナゾール
プロピールアルコール	プロピルアルコール

本手法で注意すべき点は,語幹部分のカナに着目している点である.多くの医薬品名称では名称を特徴づけるのにカタカナの商標名を用いており本研究もそこに着目しているが,その一方でたとえば「E・A・C錠」や「5−FU錠50協和」は本研究の方法では語幹すべての文字が削除されてしまい,類似度計算の対象とならない.現在,文字の類似度を考慮に入れた類似指標を検討中であり,これを利用してカタカナ以外も対象とし文字の形状をもとに類似度の計算を行う類似名称検索システムを構築する予定である.

5. まとめ

本研究では,剤形・規格単位等のマスタを作成し,自然言語処理で用いられている手法をもとに,処方オーダリングシステムの処方入力画面に入力される医薬品名称が正しいものであることを担保する仕組みに必要な医薬品名称そのものに含まれる情報を抽出することを試みた.その結果,およそ8割以上の語幹を正しく抽出できたが,一方で医薬品名称からマスタに含まれる文字列を削除すべきではない事例がみられ,このような事例については,削除を許さない文字列リストを用意する等の改善が必要となることを指摘した.

また,語幹に含まれるカタカナ部分について,互いに類似したものの組の抽出を行った.マスタの充実を含め,本手法を改善することにより,医薬品類似名称検索システムのデータ作成を自動で行うことができると期待できる.

参考文献

[1] M. Kimura, K. Tatsuno, T. Hayasaka, Y. Takahashi, T. Aoto, M. Ohkura, F.Tsuchiya.Application of Data Mining Techniques to Medical Near-Miss Cases. Proceedings of the 12th International Conference on Human-Computer Interaction, pp. 474-483, 2007.
[2] 医薬品類似名称検索システム.https://www.ruijimeisho.jp/.
[3] 土屋文人, 川村昇, 王智瑛, 原明宏.医薬品名の標準化と類似性の検討.医療情報学, 21-1 pp.60-68, 2001.
[4] 古川裕之, 土屋文人, 大西 久, 増江俊子, 分校久志, 宮本謙一.医薬品に関連したリスクマネジメント戦略における処方オーダリングシステムの可能性についての分析.医療情報学 21-1 pp.69-76, 2001.
[5] 古川裕之.患者安全管理の視点から見た処方オーダリングシステム.医薬品情報学, 6-2, pp.106-110, 2004 .
[6] 財団法人医療情報システム開発センター.http://www.medis.or.jp/.

To be appeared in Proceedings of AHFEI2010 (Miami)

CHAPTER X

The standardization of medicine name structures suitable for prescription entry system

Masaomi Kimura[1], Keita Nabeta[1], Michiko Ohkura[1], Fumito Tsuchiya[2]

[1]Shibaura Institute of Technology
3-7-5 Toyosu, Koto City, Tokyo, Japan

[2]Tokyo Med. & Den. University
1-5-45 Yushima, Bunkyo City, Tokyo, Japan

ABSTRACT

The safety of medical usage is one of keys to prevent medical accidents caused by medicines. In order to ensure the safety, there have been efforts to avoid medicines whose name or visual appearance of package is similar to others: the efforts of naming, design and/or adoption of medicines. They are not enough as countermeasures, since medical accidents resulting from medicines occur in spite of those efforts.

Because of this, we make a study to create the database that contains information to ensure the safety of medical usage and to construct a prescription checking system based on such database.

In this paper, we report the result of study about the method to extract information contained in medicine names which can help to verify users' input of the medicine names to prescription input system.

Keywords: Medical safety, medicine name structure, prescription entry system

INTRODUCTION

The safety of medical usage is one of keys to prevent medical accidents caused by medicines.

2 THE STANDARDIZATION OF MEDICINE NAME STRUCTURES SUITABLE FOR PRESCRIPTION ENTRY SYSTEM

In order to ensure the safety, there have been many efforts to avoid medicines whose name or visual appearance of package is similar to others, including the avoidance efforts of similar naming, design and adoption of such medicines. Unfortunately, those countermeasures are not enough, since medical accidents resulting from medicines do not go away. In fact, in some hospital in Japan, muscle relaxant, Succin, was wrongly injected to a patient instead of sound-alike steroidal anti-inflammatory medicine, Saxison. Though Saxison was not adopted in the hospital as a countermeasure of confusion, the doctor confused Succin with it. As a result, the doctor prescribed Succin, and this caused a fatal accident.

In order to prevent such a wrong prescription, we need a device to check the selection of medicine is appropriate for the purpose of treatment. In Japan, computerized prescriber order entry (CPOE) systems are widely used to prescribe medicines, and there have been many efforts to prevent wrong input of medicine names: the improvement of order in the list of medicines, the emphasis of frequently confused medicine names by adding some symbol characters, and so on (Furukawa et.al., 2001, Tsuchiya et.al., 2007). Besides such efforts, we consider that it is important to provide suitable information of medicines for doctors to recognize selection errors of medicines. Because of this, we made a study to create the database that contains the information that is necessary for a prescription checking system of CPOE based on such database.

Though it is obvious that a medicine name is a significant part of prescription information, it is difficult to identify information included in the names. This is because the structures of information contained in them are not fully standardized. In the past study, it is shown that the product names are composed of 23 kinds of elements, e.g. premodifiers, stems, postmodifiers, standard units, dosage forms and so on (Tsuchiya et.al., 2001). Remember that the usage of medicines that have the same stem differs depending on standard units or dosage forms. It is obvious that wrong usage can cause an accident. Because of this, in these elements, a stem of names, standard units and dosage forms are, at least, essential to identify a medicine. Unfortunately, not all of medicines sold in Japan have these kinds of information in their name. Even for the medicine name containing all of the information in it, there is no guarantee that they are arraigned in the determined order. For instance, the order that a brand name comes first, a dosage form second and a standard unit last, such as 'トリアラム錠 0.25mg' (Trialam Tablets 0.25mg), is typical, but there also exists another orders such as 'カムリトン0.25mg錠' (Camriton 0.25mg Tablets).

For this reason, in order to treat the information included in medicine names in CPOE, it is necessary to decompose medicine names and identify the parts based on master data for each kind of information. Since there do not, unfortunately, exist such master data, we first create master databases of information including standard units and dosage forms. We next introduce the method to identify the information in medicine names by matching substrings to the ones obtained by utilizing N-grams. After that, we extract the stems in medicine names by subtract the matched substrings from them. We evaluate this method by measuring the precision of extraction.

TARGET DATA

In this study, we identified each medicinal product in Japan by HOT9 code, which is identical if and only if the medicines have the same brand name, the same standard unit, the same dosage form and are produced by the same pharmaceutical company. In the rest of this paper, we count the number of products based on this code. As a master data of original medicine names, we used official name data in 'MEDIS standard medicine master' whose version is dated July 31, 2009. This master data have HOT9 code information that has one-to-one relationship to each of 24261 medicines.

We used dosage form master data contained in the dictionary data of single ingredient medicines, collected by the Health and Labour Sciences Research project, 'The study on medicine dictionary data items and standards (ICH M5)' in 2008.

MAKING MASTER DICTIONARY

Besides the original medicine name master dictionary and dosage form master dictionary, we made dictionaries of symbol master data, standard unit master data, parenthetic expression master data and application master data.

For the symbol master dictionary, we decomposed the original medicine name data into single characters and gathered the ones other than Kana (Katakana, Hiragana), Kanji, alphabetical, and numeric characters, e.g. 「, 」, (,), 〈, 〉.

For the standard unit master dictionary, we split the standard unit data contained in MEDIS standard medicine master by characters in the symbol master dictionary, and extracted letter strings are numbers (including the decimal separator characters, namely, comma and period) followed by Kana, Kanji and alphabets, e.g. 25mg, 10000国際単位 (International Unit).

For the application master dictionary, we extracted letter strings that have a pattern 'X用'. The Kanji character '用' denotes that 'X' is the application objective, e.g. 手術用 (surgical purpose), 注射用 (injection purpose. We assumed that 'X' consists of the same sort of letters (only Kanji, only Kana and so on).

For the parenthetic expression master dictionary, we extracted letter strings in parentheses and brackets included in the symbol master dictionary. After extraction, we remove the duplicate entries in this dictionary and other dictionaries.

Each data in these dictionaries are labeled with two alphabetical letters denoting to which dictionary the data belong and a sequential number in each dictionary (Fig.1).

4 THE STANDARDIZATION OF MEDICINE NAME STRUCTURES SUITABLE FOR PRESCRIPTION ENTRY SYSTEM

Standard unit master dictionary

NewCode	data
CU-1	0.0015%
CU-2	0.003%
CU-3	0.004%
CU-4	0.005%
CU-5	0.007%
CU-6	0.01
CU-7	0.01%
CU-8	0.0143%
CU-9	0.02
CU-10	0.02%
CU-11	0.025
CU-12	0.025%
CU-13	0.025mg
CU-14	0.03%
CU-15	0.033%
CU-16	0.04%
CU-17	0.05
CU-18	0.05%
CU-19	0.05mg
CU-20	0.06%
⋮	⋮

Dosage form master dictionary

NewCode	data
CF-1	AL綿球
CF-2	うがい液
CF-3	エアゾール
CF-4	エアロゾル
CF-5	ガーゼ
CF-6	ガス
CF-7	カプセル
CF-8	キット
CF-9	クリーム
CF-10	ゲル
CF-11	ゲル細粒
CF-12	シート
CF-13	シール
CF-14	ジェル
CF-15	シップ
CF-16	シリンジ
CF-17	ドロップ
CF-18	スクラブ
CF-19	スティック粒
CF-20	スプレー
⋮	⋮

Application master dictionary

NewCode	data
CP-1	mL用
CP-2	カントップ用
CP-3	クフレックスシロップ用
CP-4	シロップ用
CP-5	スタート用
CP-6	センセファリンシロップ用
CP-7	デリバテクスヘパリンロック用
CP-8	ベミロックヘパリンロック用
CP-9	ロック用
CP-10	う過用
CP-11	医療用
CP-12	塩酸塩細粒小児用
CP-13	塩酸塩注射用
CP-14	塩酸塩点滴静注用
CP-15	塩酸塩内用
CP-16	外用
CP-17	筋動注用
CP-18	肝動注用
CP-19	関節腔内用
CP-20	含嗽用
⋮	⋮

Parenthetic expression master dictionary

NewCode	data
CB-1	123I
CB-2	131I
CB-3	133Xe
CB-4	201Tl
CB-5	51Cr
CB-6	67Ga
CB-7	80,000
CB-8	90Y
CB-9	99mTc
CB-10	AFP
CB-11	AMP
CB-12	AP
CB-13	AT
CB-14	AW
CB-15	A液
CB-16	BMD
CB-17	BT
CB-18	Bx
CB-19	B液
CB-20	CH
⋮	⋮

FIGURE 1 The master dictionaries (parts). The left column in each table shows the code corresponded to each data.

METHODS

In this study, we propose the method to extract a stem part in a medicine name by removing the matched parts to the dictionary data.

MATCHING PARTS OF MEDICINE NAMES TO MASTER DICTIONARIES

Though we may naively expect to find character substrings in medicine names that match data in the dictionaries by applying full-text search, it is obviously computationally expensive. We can evaluate its cost as O(NLM), where N is the number of medicine names, L is the (typical) length of medicine names and M is the number of registered data in the dictionaries.

There is another problem in the naïve application of full-text search for the cases that the abbreviated string of the registered data should be also registered in the dictionaries. For example, there are two words '注射液' and '注' that mean 'injection drug'('注' is an abbreviation of '注射液') and they appear in the names such as 'サルソニン注射液' (Salsonin injection) and 'ロヒプノール注' (Rohypnol injection). In such cases, both of these words should be registered in the dictionaries. The difficulty is that 'サルソニン注射液' can be regarded to be matched both '注射液' and '注', though we expect that only '注射液' is matched.

In order to overcome such difficulties, we propose the comparing method based on N-

gram, which is popular in the area of natural language processing (NLP).

First, we extract all contiguous substrings with the definite length from 1 to L (the length of the medicine name) from a medicine name and compare them to the data in the dictionaries. We note that the calculation cost is O(NLlogL · logM) under assumption that the dictionaries are indexed and their searching cost is in the order of O(logM). We can therefore expect the cost of our method is much smaller than the one for naïve full-text search, since M is usually much more than logL · logM.

Then, under the assumption that the original words of abbreviations are included in the dictionaries, we neglect the matched substring extracted in the first step, if it is included in the others as a substring. As a result, we can expect to identify the substrings in each medicine name that correspond to the data listed in the dictionaries. After the identification, by finding which dictionary the data corresponding to the substring, we can reorganize the information to find whether a medicine name lacks the information of a dosage form, a standard unit and so on.

Finally, we extract a stem part in a medicine name by removing the matched substrings. This is based on the assumption that there needs not extra information other than data in the dictionaries that we prepared for.

The schematic figure of the method mentioned above is shown in Fig.2.

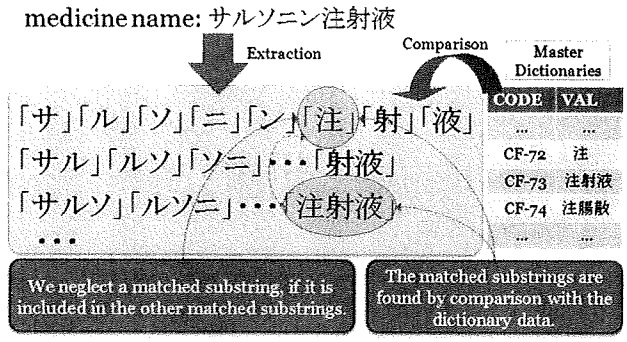

FIGURE 2 The schematic figure of the proposed method. The medicine name is decomposed into string segments whose length is from one to the length of the original medicine name. Each segment is compared to each dictionary and the matched substrings are compared to the other matched substrings to see if they are part of other substrings.

FINDING THE PAIRS OF MEDICINE NAMES WHOSE KANA STEMS ARE SIMILAR TO EACH OTHER

As we mentioned in Introduction, the problem is the confusion of medicines whose names are similar to each other. As for prescription input, this can be embodied as the input of the wrong medicine names similar to a relevant medicine. Since a name stem is usually used to specify a medicine, in this study, we extracted the name pairs whose stems are similar to

each other by calculating some similarity indices. (Since the name stems of Japanese medicines are composed by Katakana characters except for Chinese herb medicines, we focus on the Katakana part of the name stems.) As similarity indices, we use HTCO (the number of coincident characters in first two and last two characters) , H3COS1 (the number of coincident characters in first three characters) and edit distance that measures the number of operation (addition, deletion and replacement) to transform one stem to another.

Taking the similarity of Katakana characters themselves into account , we identified the characters in the groups, 'アァヤカ', 'ツッシミ', 'ンソリ', 'クケフワ', 'エコユ', 'スヌ', 'ナメ', 'テラ', 'イィ', 'ウゥ', 'エェ', 'オォ' before we calculated HTCO, H3COS1 and edit distance.

EVALUATION

MATCHING PARTS OF MEDICINE NAMES TO MASTER DICTIONARIES

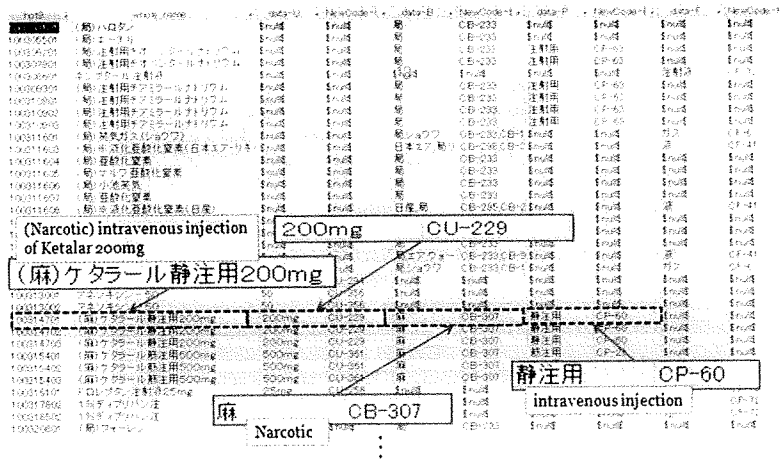

FIGURE 3 Matching result of medicine names (part). This shows that the medicine whose name is '(Narcotic) Ketalar intravenous injection 200mg' is successfully decomposed the parts, '200mg', 'Narcotic' and 'intravenous injection', which are listed in the dictionaries.

As is shown in Fig.3, we can see that our method decomposes the medicine name into some parts corresponding to the data in dictionary. For example, the original name, '(Narcotic) Ketalar intravenous injection 200mg' is successfully decomposed into the parts in the dictionaries, namely, 'Narcotic' in the parenthetic expression master dictionary, 'intravenous injection' in the dosage form master dictionary, '200mg' in the standard unit

master dictionary.

In order to evaluate the effectiveness to match medicine names to the data in the dictionaries, we measure the accuracy to find the stems in the names by removing the matched data from the original names. (The symbols registered in the symbol master dictionary such as parentheses are removed from the original names in advance of the extraction of the stems.) For 500 randomly selected medicine names, we successfully obtained 474 stems (94.8%). Fig.4 shows some examples of extracted stems. The wrong extractions for 26 medicine names were caused by the multiplicity of meanings of the parts of names. For example, we obtained 'ター'(ter) as the 'stem' of the name, 'ガスター散2%'(Gaster powder 2%). The substring 'ガス' (gas) is the only the part of the brand name 'ガスター' (Gaster) but it has its own meaning as the state of substances. Of cause, for human eyes, it is clear that the part does not mean the physical state of the medicine. However this is implicitly under the assumption that the string 'ガスター' (Gaster) is known as the brand name of a medicine. Since our objective is to identify semantics of the parts of names, we cannot assume that their meanings are already known in advance.

In the future study, we may solve this problem by taking account of the co-occurrence relationships of substrings appearing side-by-side in medicine names. If some substring invariably appears to other substring, we may presume that they suggest the existence of the stem that contains those substrings as parts.

	hot9	whole_name	name_stem
1	100310901	（局）注射用チアミラールナトリウム	チアミラールナトリウム
2	100311601	（局）笑気ガス〈ショウワ〉	笑気
3	100311604	（局）亜酸化窒素	亜酸化窒素
4	100311608	（局）※液化亜酸化窒素（日産）	化亜酸化窒素
5	100313001	アネソキシン－５０	アネソキシン－
6	100314701	（麻）ケタラール静注用２００ｍｇ	ケタラール
7	100323912	（局）※ブロムワレリル尿素（山善）	ブロムワレリル尿素
8	100328401	ユーロジン１ｍｇ錠	ユーロジン
9	100333802	ネルガート１５	ネルガート
10	100343701	ニトラゼパム錠５ｍｇ「トーワ」	ニトラゼパム

FIGURE 4 The extracted name stems (part). The column name 'hot9', 'whole_name' and 'name_stem' respectively denote HOT9 codes, the original medicine names and the stem parts.

FINDING THE PAIRS OF MEDICINE NAMES WHOSE KANA STEMS ARE SIMILAR TO EACH OTHER

Fig.5 shows the pairs of medicine names extracted under the condition HTCO=1, H3COS1=1 and their edit distance is equal to 1. Most of them are the combination of a brand name and its generic name, e.g. 'カルテオロール'(Carteolol) is the generic name of 'カル

テロール'(Cartelol). However, Fig.5 contains the pairs whose efficacies are different. For example, the pair of 'タキソテール'(Taxotere) and 'タキソール'(Taxol) and the pair of 'アルマトール'(Almatol) and 'アルマール'(Almarl) are well-known as the confusing names that cause severe accidents. It is striking that 'クインスロン' (Kuinsron) and 'クインロン' (Kuinlon) are confusing, have the different efficacies and are produced by the same pharmaceutical company. It is preferable to discuss the countermeasure against confusion of such medicines in the company.

カルテオロール	カルテロール
トリアゾラム	トリアラム
デキストセラン	デキストラン
テストビロンデポー	テストロンデポー
ラキソベロン	ラキソロン
タキソテール	タキソール
スカノーゼリン	スカノーリン
ニカルジピン	ニカルピン
クインスロン	クインロン
ダイアコート	ダイアート
カプトプリル	カプトリル
オメプラゾール	オメプラール
オフロキサシン	オフロキシン
エストリオールデポー	エストリールデポー
エストリオール	エストリール
アルマトール	アルマール
アラセプリル	アラセリル
グリセリン	グリセロリン
フラボキサート	フラボサート
ベルベゾロン	ベルベロン
フルコナゾール	フルコナール
プロピルアルコール	プロピールアルコール
フェノバール	フェノール
フェニタレン	フェニレン

FIGURE 5 The pairs of medicine names (Japanese) whose HTCO=4, H3COS1=3 and edit distance=1.

SUMMARY AND CONCLUSION

Though medicine names are a significant part of prescription information, their information structures are not fully standardized and it is, therefore, difficult to identify the information therein. Not all of medicines sold in Japan have enough information in their name and even for the medicine name containing all of the information in it, there is no guarantee that they are arraigned in the determined order. For this reason, in order for CPOS to treat the information in medicine names, it is necessary to decompose medicine names into

the parts matched to master dictionaries for each kind of information.

In this study, we proposed the master dictionaries: the original medicine name master dictionary, the dosage form master dictionary, the symbol master dictionary, the standard unit master dictionary, the parenthetic expression master dictionary and the application master dictionary. We next introduced the method to identify the information in medicine names by matching pieces of strings obtained by N-grams to the data in the dictionaries. After that, we extracted the stems in medicine names by subtract the matched substrings from them.

REFERENCES

Kimura, M., Tatsuno, K., Hayasaka, T., Takahashi, Y., Aoto, T., Ohkura, M., Tsuchiya, F. (2007), "Application of Data Mining Techniques to Medical Near-Miss Cases." *Proceedings of the 12th International Conference on Human-Computer Interaction.*

Tsuchiya, F. (2007), "Medication Errors Caused by Order Entry System and Prevention Measures." *Proceedings of the 12th International Conference on Human-Computer Interaction.*

Furukawa, H., Tsuchiya, F., Onishi, H., Masue, T., Bunko, H., Miyamoto, K. (2001), "Investigation of Possibility for Physician Order Entry System on Risk Management Strategy Related To Medication Errors." *Japan Journal of Medical Informatics*, 41, 909–996.

Tsuchiya, F., Kawamura, N., Wang C., Hara, A.(2001), "Standardization and similarity deliberation of Drug-names." *Japan Journal of Medical Informatics*, 21(1), 60–68.

The Medical Information System Development Center (1996), *http://www.medis.or.jp*

To be appeared in Proceedings of AHFEI2010 (Miami)

処方入力画面に表示される医薬品名標準化の検討

THE STANDARDIZATION OF MEDICINE NAME STRUCTURE TO BE SHOWN ON PRESCRIPTION ENTRY SCREEN

木村昌臣[1] 鍋田啓太[2] 大倉典子[1] 土屋文人[3]
Masaomi Kimura Keita Nabeta Michiko Ohkura Fumito Tsuchiya

芝浦工業大学[1]
Shibaura Institute of Technology
芝浦工業大学大学院[2]
Graduate School, Shibaura Institute of Technology
東京医科歯科大学[3]
Tokyo Medical and Dental University

1 背景・目的

筆者らは医薬品ヒヤリ・ハット事例を解析し、薬剤の取り違えは医薬品の名称や外観が類似していることが主な要因となっているという結果を得ている [1]。処方オーダリングシステムの処方入力画面における医薬品の入力ミスは医療事故に直結する原因となりうるため、入力された医薬品名の正誤をチェックする仕組みが必要とされる。医薬品名称を構成する要素のうち医薬品特定に重要なのは語幹・剤形・規格単位であるが、これらの有無や表記順には必ずしも統一性があるとはいえない。そのため、医薬品名称を処方オーダリングシステム等で取り扱うには、その記述をこれら三つの構成部分に分解・整理し、マスタ化したうえで医薬品名称との対応を求める必要があるが、そのようなマスタは現状で存在しない。

そこで本稿では、これらのマスタを試作し、これをもとに実際の医薬品名称からの語幹・剤形・規格単位を抽出を試みることにより、処方オーダリングシステム等で扱う際の医薬品名称の問題点についての報告を行う。

2 対象データ

本研究では医薬品各製品別に扱うため、HOT9 番号を基準として件数を集計する。また、マスタの元となるデータとして MEDIS 標準医薬品マスター 2009 年 7 月 31 日版を用いた。剤形名称に関しては、平成 20 年度厚生労働科学研究「医薬品辞書データ項目と基準 (ICHM5) に関する研究」医薬品辞書データをもとに作成した。

3 手法

規格単位マスタは、MEDIS 標準医薬品マスタ「規格単位」データを括弧等の記号で分割し、さらに「数値+数値以外」のパターンを持つ文字列を抽出することにより作成し、剤形マスタは平成 20 年度厚生労働科学研究「医薬品辞書データ項目と基準 (ICHM5) に関する研究」医薬品辞書データをもとに作成した。また、これらの構成要素以外に、剤型や投与対象等が「○○用」と表記されるため、用途マスタとして MEDIS 標準医薬品マスタ「公示名称」データに含まれる「○○用」(○○は同種文字列) という文字列を収集した。さらに、後発品の製薬企業名などは括弧内に含まれて表記されるため、同様に「公示名称」データに含まれる記号を抽出し、対応する括弧組に含まれる文字列を抽出して括弧内文字列マスタとした。語幹部分については、上記マスタに含まれている部分を取り除くため、「公示名称」データに含まれる各名称文字列を N-gram に分解し、これとマスタデータを比較し、含まれているマスタデータを見つける方法をとった。ただし、例えば「注射液」を部分に含む医薬品名称は「注」も含んでしまうため、医薬品名称の部分にマッチするマスタデータに含まれる部分文字列はマスタデータをマッチするものであっても対象から除外することとした。

4 結果

whole_name	name_stem
(局) 注射用チアミラールナトリウム	チアミラールナトリウム
(局) 笑気ガス (ショウワ)	笑気
(局) 亜酸化窒素	亜酸化窒素
(局) ※液化亜酸化窒素 (日産)	化亜酸化窒素
アネソキシン－５０	アネソキシン－
(麻) ケタラール静注用２００ｍｇ	ケタラール
(局) ※ブロムワレリル尿素 (山善)	ブロムワレリル尿素
ユーロジン１ｍｇ錠	ユーロジン
ネルガート１５	ネルガート
ニトラゼパム錠５ｍｇ「トーワ」	ニトラゼパム

図 1 語幹部分抽出例

医薬品 24261 品目から 500 件をランダムに抽出し語幹部分を目視で確認したところ、抽出精度は 94.8 %であった。これは、同じ文字列が複数の意味を持ちうることにより、余計に語幹部分に相当する文字列を削除されたのが原因であった。さらに抽出結果にもとづき、剤型・用途情報・規格単位の出現頻度を集計したところ、規格単位が製品名についていないものが局方品をのぞき 4954 件あり、単味剤には規格単位が名称に含められるべきだが、規格単位が製品名についていないもののなかに単味剤が少なからず存在することがわかった。このように、対象データにはあるべき情報が含まれない医薬品名が少なからず存在したが、最新のデータでは 1000 件以上の医薬品名称が改訂されている。今後はこのデータに対して改めて解析を行う予定である。

参考文献

[1] Kimura M, et al. Application of Data Mining Techniques to Medical Near-miss Cases., Proceedings of HCI Interaction 2007; 2007 July 22-27; Beijin, China, Berlin:Springer; 2007. pp.474-483.