ポスター | 薬剤情報システム
# ポスター12
# 薬剤情報システム
2019年11月24日(日) 10:00 ～ 11:00 ポスター会場1 (国際展示場 展示ホール8)

[4-P1-2-02] 医薬品名の形態素解析アルゴリズムの構築

○佐藤 弘康[1]、蝦名 勇樹[1]、真井 雄規[1]、島津 智行[1]、喜多 力[1]、田村 広志[1]、渡辺 浩明[1] （1. JA北海道厚生連　帯広厚生病院）

キーワード：Drug Information, Morphological Analysis, Natural Language Processing

【背景】近年、医療記録等のテキスト情報を用いた自然言語処理 NLP等の研究が多く行われている。医療テキスト情報の中には医薬品名が存在することも多いが、その記載のゆらぎは大きく、機械的に医薬品を特定することは困難である。我々は、医療テキスト情報内の医薬品名情報から医薬品を特定するアルゴリズムの構築を計画した。本発表では、医薬品名に関する形態素辞書の開発を目的とし、承認医薬品の名称について医療薬学上の形態素への分解を試みたので報告する。

【方法】2018年12月9日に診療情報提供サービスのホームページよりダウンロードした約2万品目の承認医薬品リストを用いて、その医薬品名を形態素に分解するアルゴリズムを Visual Basic for Applicationにて構築し、Excelのユーザー定義関数として実装した。形態素としては、「一般名」、「ブランド名」、「規格量」、「濃度」、「剤型」、「屋号」等を設定した。構築アルゴリズムにより生成された形態素解析結果について、2名の薬剤師がランダム抽出した500品目を評価した。発見された不適切解析結果を基にアルゴリズムを改変し、これを繰り返すことにより解析精度を向上させた。

【結果】99.2%（496/500）の精度で医薬品名を適切に形態素解析できるアルゴリズムを構築した。当初は「ブシ」、「キナ」、「リン」等の文字数の短い一般名に誤判定する事例が確認されたが、除外規定等を設けることにより精度を向上した。

【考察】適切な判定ができなかった医薬品名には、臨床使用頻度の少ないものも含まれており、医療記録等に登場する医薬品名に適応する場合には、さらに高い精度になると思われる。今後は、今回生成したアルゴリズムを医療記録等のテキスト情報に適応し、抽出精度を評価するとともに、医薬品標準コードを一意に付与できるかどうかについても検討していく予定である。

医薬品名の形態素解析アルゴリズムの構築

佐藤弘康、蝦名勇樹、真井雄規、島津智行、喜多力、田村広志、渡辺浩明
JA 北海道厚生連　帯広厚生病院　薬剤部

Development of Morphological Analyzer for Drug Name

Hiroyasu Sato, Yuki Ebina , Yuki Sanai, Tomoyuki Shimazu, Chikara Kita, Hiroshi Tamura, Hiroaki Watanabe
Pharmaceutical Department, Obihiro Kosei General Hospital

In recent years, research on natural language processing (NLP) using text information such as medical records have been actively conducted. In medical text information, drug name often exists, but the fluctuation of the expression is large. It is difficult to simply specify the drug. We planned to construct an algorithm to identify a drug from the drug name information in the medical text information. For the purpose of developing a morpheme dictionary for drug names, this study tried to distinguish the names of approved drugs from medical pharmacological morphemes.

The algorithm was built with Visual Basic for Application and implemented as an Excel user-defined function “Drug MA”. As the morpheme, “general name”, “brand name”, “drug strength”, “drug concentration”, “drug form”, “maker name” and the like were set. Two pharmacists evaluated 500 items randomly extracted from the morphological analysis results. The analysis accuracy was improved by modifying the algorithm based on inappropriate cases and repeating this.

The accuracy of the constructed algorithm is 99.2% (496/500). In the future, we plan to create a dictionary for identifying medicines from medical record text information using the algorithm we have constructed.

Keywords: Drug Name, Morphological Analysis, Natural Language Processing (NLP).

1. 背景

近年、医療記録等のテキスト情報を用いた自然言語処理 NLP 等の研究が多く行われている。医療テキスト情報の中には医薬品名に関する情報が存在する場合も多いが、その記載のゆらぎは大きく、剤型や規格に関する情報を省略したり等、承認医薬品名がそのまま記載されることは少ない。そのため、承認医薬品名との単純な文字列の比較により医薬品を特定することは困難である。我々は、医療テキスト情報内の医薬品名情報から医薬品を特定するアルゴリズムの構築を計画した。本発表では、医薬品名に関する形態素辞書の開発を目的とし、承認医薬品の名称について医療薬学上の形態素への分解を試みたので報告する。

2. 方法

2018 年 12 月 9 日に診療情報提供サービスのホームページよりダウンロードした約 2 万品目の承認医薬品リストを用いて、その医薬品名を形態素に分解するアルゴリズムを Visual Basic for Application にて構築し、Excel のユーザー定義関数「DrugMA」として実装した（図 1）。医薬品名における医療薬学上の形態素としては、「一般名」、「ブランド名」、「規格量」、「濃度」、「剤型」、「包装・デバイス」、「屋号」等を設定した。

「一般名」は、厚生労働省がホームページ上で公開している「薬価収載医薬品リスト」との照合を行い、文字列が一致した場合に「一般名」として抽出した。ただし、誤判定の多い語については除外語リストを設け、「一般名」として抽出しなかった。抽出した「一般名」については、「一般名修飾語」マスタを作成し、「一般名語幹」と「一般名修飾語」に細分化した。

「規格量」、「濃度」については、正規表現を用いて該当する部分文字列がヒットした場合に判定した。また、それぞれ抽出後に数値部分と単位部分に細分化した。最終的に「規格量」が未判定であり、残存文字列に数値部分が存在した場合には、それを「規格量」とした。

「包装・デバイス」および「剤型」については、それぞれのマスタを作成し、医薬品名の残存文字列内に一致する部分文字列が存在した場合に判定した。

「屋号」については、医薬品名に「」がある場合、その中の文字列を切り出し判定した。また、それ以外であっても、作成した「屋号」マスタに一致する文字列が医薬品名内に存在した場合には、それを「屋号」と判定した。

最終的に「一般名」が抽出されていない場合、どの形態素にも割り当てられていない残存文字列を「ブランド名」として抽出した。「ブランド名」は、作成した「ブランド名修飾語」マスタを用いて、「ブランド語幹」と「ブランド前句」、「ブランド後句」にさらに分解した。

DrugMA は、抽出された各形態素を該当文字列の後ろに{}付きで表現する形式で出力するユーザー定義関数とした。

構築アルゴリズムにより生成された形態素解析結果について、2 名の薬剤師がランダム抽出した 500 品目を評価した。発見された不適切解析結果を基にアルゴリズムを改修し、これを繰り返すことにより解析精度を向上させた。

1．一般名名称マスタによる「一般名」の判定
2．「」内の切り出し及び屋号マスタによる「屋号」の判定
3．剤型マスタによる「剤型」の判定
4．正規表現による「濃度」の判定
5．正規表現による「規格量」の判定
6．デバイスマスタによる「包装・デバイス」の判定
7．修飾語マスタによる「修飾語」の判定
8．「一般名」非特定時、残存文字列を「ブランド名」

図 1　医薬品名の形態素解析に関する主な手順

3. 結果

99.2%（496/500）の精度で医薬品名を適切に形態素解析できるアルゴリズムを構築した（図 2）。

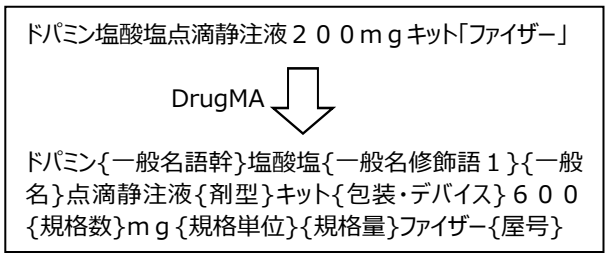

図 2 医薬品名の形態素解析結果の一例

当初は「ブシ」、「キナ」、「リン」等の文字数の短い一般名に誤判定する事例が複数確認された（図 3）が、これらが医療テキスト情報の中で、医薬品名を示す単語として記述されることは稀であると判断し、除外規定等を設けることにより精度を向上した。

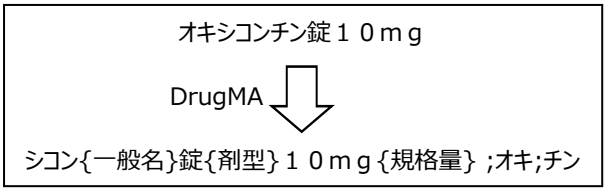

図 3 文字数の短い一般名への該当により誤判定となった一例（現在は除外規定により解消）

また、「ケフレックスシロップ用細粒200」等、剤型候補が複数存在する医薬品名については、適切に判定することができなかった。しかしながら、「剤型」を複数許容するようにアルゴリズムを改修にした場合には、新たな誤判定を生む懸念もあり、本事例への対応は実施しなかった。

4. 考察

今回構築したアルゴリズムは 100%の判定率ではないものの十分に臨床応用可能な精度であると考える。誤判定医薬品名には、臨床使用頻度の少ないものも含まれており、医療記録等に登場する医薬品名に適応する場合には、さらに高い精度になると思われる。

また、本アルゴリズムでは、「一般名」の抽出において、「薬価基準収載品目リスト」における成分名と文字列が完全一致した場合に「一般名」として判別している。そのため、「ジクロフェナクナトリウム錠」は「一般名」として判定されるが、「ジクロフェナク Na 錠」は「一般名」として認識されず、「ブランド名」として判定される。我々の知る限り、一般名の基幹語部分のみがリスト化され、かつ新薬に対応するメンテナンスが行われている公開データベースは存在しない。そのため、この問題の解決は現時点で困難である。しかしながら、最終的な医療テキスト情報の識別においては、本アルゴリズムは医薬品名の基幹語が「ブランド名」であるのか「一般名」であるのか意識することなく判定可能であるため、本限界点については臨床応用時には大きな問題にならないと考える。

今回、構築したアルゴリズムは、汎用性を考えて新たな新規成分、新規製品に対しても可能な限り適応可能なように構築したが、新しい「剤型」や「デバイス」が登場した場合には、それぞれのマスタリストに追加していく必要がある。

今後は、承認医薬品名がどのような形態素から構成されているかを分析し、どの形態素の組み合わせが存在すれば医薬品標準コードを一意で特定することが可能か、基幹語ごとに明らかにすることで、医薬品名形態素辞書が構築可能になると考える。この医薬品名形態素辞書を医療テキスト情報に適応することで、剤型や規格が省略された記載である医薬品名情報から医薬品を特定できる可能性がある。

医薬品名には、「名詞」、「形容詞」といった通常のテキスト情報の品詞に関する形態素ではなく、「規格」、「剤型」、「デバイス」といった医療薬学上の特殊な形態素に判別する必要があり、既存の形態素解析ツールではその判定が困難である。今回、我々は医薬品名を医療薬学上の形態素へ解析することに特化した形態素解析ツール DrugMA を構築した。今後は、アルゴリズムの解析精度向上を目指すとともに、本ツールの臨床応用を進めていく予定である。