
ポスター

[PO-1～8、P-1～16] ポスター立会

2019年6月7日(金) 15:00 ～ 16:00 ポスター会場 (熊本市民会館 2F ホワイエ)

[P-5] 機能性表示食品の一般消費者向け情報に含まれる機能性および安全性に関する難解語抽出手法の提案

種村 菜奈枝 (慶應義塾大学 薬学部 医薬品開発規制科学)

機能性表示食品の一般消費者向け情報に含まれる 機能性および安全性に関する難解語抽出手法の提案

種村 菜奈枝^{*1}, 木村 正吾^{*2}, 町井 湧介^{*1}, 長 雄一郎^{*1}, 柿崎 真沙子^{*3},
小野寺 理恵^{*4}, 漆原 尚巳^{*1}

^{*1} 慶應義塾大学薬学部, ^{*2} ヘッドウォーターズ IT インキュベーション事業本部, ^{*3} 名古屋市立大学大学院医学研究科, ^{*4} 札幌医科大学医学部附属フロンティア医学研究所

Automatic Elusive Term Recognition in Functionality and Safety Information Targeted at General Consumers Using Several Glossaries

Nanae Tanemura^{*1}, Shogo Kimura^{*2}, Yusuke Machii^{*1}, Yuichiro Cho^{*1}, Masako
Kakizaki^{*3}, Rie Onodera^{*4}, Hisashi Urushihara^{*1}

^{*1} Keio University, ^{*2} Headwaters, ^{*3} Nagoya City University,
^{*4} Sapporo Medical University

抄録

ミス・コミュニケーションが生じる原因として、共通基盤の違いがある。この違いを平準化するため、難解語を特定し、平易化することが重要である。しかし、機能性表示食品の一般消費者向け情報に含まれる難解語を特定するための形態素解析用辞書はない。そこで、形態素解析用辞書の作成およびその精度評価を目的とした。医学または臨床試験用語を収録する10種類の用語集（取得用語数973,895語）を用い、203,095通りの辞書を作成した。その後、機能性表示食品の一般消費者向け情報1,310件からランダムに66件（4.65%）抽出し、作成した全辞書の精度評価を行った（データ取得日2018年7月）。その結果、最も再現率が高い辞書（検出数:MeCab 1,725, マニュアル 529, 精度:適合性 0.283, 再現性 0.924, F値 0.434）で、F値は最大0.961となると推定された。今後、さらに実証研究の実施が望まれる。

キーワード

難解語抽出, 形態素解析, リスク・ベネフィットコミュニケーション, 機能性表示食品

1. はじめに

機能性表示食品は、2015年4月1日に施行された食品表示法に基づく食品表示基準により、事業者等の責任の下で、科学的根拠に基づき機能性を表示することができるものである[1]。

機能性表示食品の機能性および安全性の情報を消費者向けに加工した情報（以下、一般消費者向け情報）が、消費者庁のホームページにて一般公開されている。しかし、この一般消費者向け情報を作成するのは専門家であるため、消費者にとって理解の妨げになる医学用語や臨床試験用語といった日常生活では使用頻度が少ない難解語を特定し、その用語を平易にすることが重要である。この現状を把握するためには、一般消費者向け情報の文章を形態素解析することで把握可能であるが、専門領域ごとに語彙の種類およびそれらの使い方が異なるため[2]、領域毎に

形態素解析用専用辞書を作成する必要がある。

そこで我々は、一般消費者向け情報に含まれる難解語候補（以下、難解語）を抽出するための形態素解析用辞書を作成することを目的とした。

2. 方法

1) 形態素解析用辞書の作成

本研究における難解語の定義は、「医学または臨床試験用語から健康食品の素材情報データベースに含まれる機能性関与成分の用語を除外した用語」とした。

この定義と対応した形で、「医学または臨床試験用語」を抽出するため一般公開されている9種類の辞書を、さらに「機能性関与成分の用語」を除外するために1種類の辞書を使用した。全203,095通りの形態素解析用辞書を作成した。

2) 形態素解析用辞書の精度評価

(1) 解析対象

一般消費者向け情報 1,408 件のうち旧届出様式を除いた 1,310 件を精度評価のための解析対象情報とした(データ取得日 2018 年 7 月)。その後、1,310 件からランダム抽出した 66 件(4.65%)の情報を対象に、精度評価を行った。

(2) 正解用語集合の作成

疫学、医学および薬学分野に精通した 2 名が独立して、解析対象 66 件の一般消費者向け情報から難解語をマニュアル抽出(TP: True Positive+FN: False Negative)した。その後、意見が分かれた用語については 2 名で討議を行い、疑義がなくなるまで繰り返し検討を重ねた。

(3) 形態素解析の実行

形態素解析用辞書 203,095 種類を用いて、解析対象 66 件の一般消費者向け情報の形態素解析を行い、難解語を抽出した。形態素解析エンジンは、オープンソースの MeCab 0.996 を使用した。形態素解析の結果は、“TP: True Positive+FP: False Positive”で示された。

(4) 精度評価

形態素解析用辞書の精度評価の指標として再現率、適合率、F 値を採用した。

(5) 形態素解析用辞書の採用

精度評価の結果、再現率または F 値が最も高い辞書(以下、前者の辞書を用いた検討を「条件 x1」、後者の辞書を用いた検討を「条件 x2」とした)を一般消費者向け情報に含まれる難解語を抽出するための候補辞書とした。その後、「条件 x1」または「条件 x2」のどちらの辞書を選択した場合、最終的に難解語の抽出精度がより向上するのかを予測するため試行実験を行った。

3) 難解語候補の抽出

最終的に選択した形態素解析用辞書を用いて、一般消費者向け情報 1,310 件を対象に形態素解析器 MeCab で難解語を抽出した。

3. 結果

最終的に、再現率が最も高かった「条件 x1」の形態素解析用辞書(抽出用語数: MeCab 1,725, マニュアル 529, 精度: 適合性 0.283, 再現性 0.924, F 値 0.434)を用いた場合、F 値は最大

0.961 となると推定された。この形態素解析用辞書を用いて、一般消費者向け情報 1,310 件を対象に難解語の抽出を行った結果、例えば、摂取(15,232 回)、評価(7,315 回)、健康(4,885 回)といった用語の他、計 541 種類の難解語が抽出された。今後、実証研究の実施が望まれる。

4. 考察

一般消費者向け情報に含まれる難解語を抽出するために 203,095 通りの形態素解析用辞書から最も再現率の高い辞書を選択した。しかし、我々が選択した辞書の場合、現状では、再現性が 0.924 と高い一方、適合性 0.283 かつ F 値 0.434 といったように、やや精度は低値であった。

これは、既存の医学または臨床試験の用語集の中には、これらの分野に無関係な用語が一定数以上含まれていることが原因として考えられた。

しかし、形態素解析用辞書の抽出精度の向上予測のための試行実験を行った結果、我々が作成した辞書に含まれる用語集データから医学または臨床試験とは無関係な用語を継続的にクリーニング精査することにより、辞書精度の向上見込みを推定できたことは大きな意義があった。

5. 結語

機能性表示食品の機能性および安全性に関する一般消費者向け情報に含まれる医学または臨床試験用語といった難解語を抽出するための新たな形態素解析用辞書を作成した。今後、食のリスクコミュニケーション推進に向け、本研究で作成した辞書の実社会での利活用が期待される。

(平成 30 年度厚生労働科学研究補助金(食品の安全確保推進研究事業)助成を受け実施した。)

参考文献

- [1] 種村菜奈枝, 濱舘直史, 漆原尚巳: レギュラトリーサイエンスの視点からみた医薬品と保健機能食品における有効性又は安全性の科学的根拠に必要な規制やその考え方の相違, 日本補完代替医療学会誌 14(2)47-60, 2017.
- [2] 相良かおる, 小野正子, 小作浩美, 他: 分かち書き用辞書 ComeJisyo の評価, 医療情報学 32(6)301-307, 2013.