平成22年度厚生労働科学研究費補助金
（医薬品・医療機器等レギュラトリーサイエンス総合研究事業）
分担研究報告書
収集したデータ（副作用情報）の処理・解析方法の検討

分担研究者　土橋　朗　東京薬科大学薬学部医薬品情報解析学教室　教授

研究要旨

本研究は、患者自らが医療用医薬品に関わる副作用発症を報告するための副作用自発報告システムを開発することを目的とする。分担研究テーマは、インターネット等を介する電子的な副作用報告受理システムが構築された後、収集されたデータをテキストマイニング等の手法を用いて解析し、患者報告の有効活用の可能性について検討するものである。

昨年度は、Yahoo!Japanのポータルサイトから利用できるサービスのうち、ブログと質問投稿サイトを対象に医薬品副作用に関する記事を収集し解析を行った結果、患者が自らの副作用を語る用語を収集することができた。さらに、質問投稿サイト上には患者による副作用自発報告が数多く集積されていることを明らかにした。

そこで、本年度は食物アレルギー(FA)をテーマにしたWebアンケートを実施し、その自由記述文を解析し、FAの症状に関する自発的発話は、男性よりも女性のほうが情報量が多く、かつ、家族にFAがいない方が多くなる傾向を明らかにした。情報の含有率の比較では、原因食物（97.6%）や症状（87.8%）に比べ、調理加工(23.6%)、発症年齢（22.8%)、その他のアレルギー疾患の有無（3.3%）などの内容が含まれにくいことを明らかにした。

患者の自発的発話の中に含まれる情報を効率良く収集するために、形態素解析により単語の意味を加味した抽出法を先述のWebアンケートの自由回答文に対して検討し、目視による分類と比較した。情報の含有率は、アレルギー疾患の有無（4.7%）については、目視を少しだけ上回る含有率となった．食品調理加工状態（70.1%)、検査結果(15.6%)に関しては、検索用語の見直しにより改善する可能性が示唆されたが、その他の分類では検索用語の見直しによる改善が困難であると結論づけた。

上記の問題を克服するために、少量サンプルより作成した手本をもとに出現用語とその頻度から分類用スコアを算出するナイーブベイズモデル（ベイズ学習法）の利用による自動分類法を検討した。先述のWebアンケートの必要なFA情報に関する自由回答文150件を用いて、8つの分類を行うための手本を作成し、学習を行い、正解率92.7%のモデルを構築した。質問投稿サイトであるYahoo!知恵袋に投稿された「食物アレルギー」に関する質問文を収集し、得られたモデルを用いて分類したところ、病院や専門医、検査についてなどの診断(50.6%)に関するニーズが最も大きく、啓蒙活動(31.9%)、原因因子などについての機序(13.8%)に関するものが続いた。

A. 研究目的

本研究は、患者自らが医療用医薬品に関わる副作用発症を報告するための副作用自発報告システムを開発することを目的とする。分担研究テーマは、インターネット等を介する電子的な副作用報告受理システムが構築された後、収集されたデータをテキストマイニング等により解析し、患者報告の有効活用の可能性について検討するものである。

副作用報告データの解析に先立ち、アレルギー情報、特に食物アレルギーに関する情報を題材にして、３つの検討を行った。

I.食物アレルギー患者の自発的発話に関する解析では、患者自身が、自身や家族の食物アレルギーに関する情報をどのように自発的に発話するのかを明らかにすることを目的として、自由記述文（自発的発話）の言語量の解析およびFA情報の充足率、内容の一致度および FA 情報の認知度を検討した。

II.FA 患者の自発的発話のコーディングルールを用いた自動分類法の検討では、患者の自発的発話の中に含まれる情報を効率良く収集するために、形態素解析により単語の意味を加味した抽出法の有用性を明らかにすることを目的に、自由記述文内の出現単語の解析とコーディングルールを用いた自動分類を検討した。

III.FA患者の自発的発話のベイズ学習法を用いた自動分類法の検討では、質問投稿サイトであるYahoo!知恵袋への質問投稿記事の中から、アレルギーに関する質問を抽出し、Web アンケートの自由記述文を基にした少量サンプルから作成した手本により、出現単語とその頻度から分類用スコアを算出するナイーブベイズモデル（ベイズ学習法）の利用による大量サンプルの自動分類法を検討した。

B. 背景

医薬品による有害事象の発生のうち 6～10%はアレルギーによる（文献 1） ものであるといわれているが、その発症機序は必ずしも明確に定義できるものばかりではない。アレルギー体質を有する患者が医薬品を服用する際には、その発症率は増加すると考えられ、服薬の際には患者自身のアレルギーに関する情報を聞き取っておくことが、その発生を防ぐ方法の一つである。しかし、アレルギーという用語の一般的な認知には幅があると考えられ、定式化された Closed Question では有効な情報を得ることが難しく、その聞き取り方には工夫が必要である。自発報告システムでは Open Question であっても患者の発話にあわせて質問を変えていくには限度があるため、アレルギーに関する自発的発話の特性を明らかにすることには意義がある。

既に生活者からの副作用報告制度の導入は、1973年にスウェーデンで開始されたのをはじめ2005年までに米国、カナダ、オーストラリア、デンマーク、オランダ、英国で導入されている（文献2）。こうした副作用報告制度の検討では、生活者と医療に日常的に携わる者との間の用語理解の差が問題になることがある。また、自発的発話にはその人の想いが影響を与え、同じ用語であってもその使い方は一定ではないことが予測される。こうした想いを文脈から読み取るための技術として、近年テキストマイニングやデータマイニング手法が検討され

ている。最も基本的なテキストマイニング手法は、キーワードによる検索であるが、自然言語の解析では形態素解析とこれを基にした構文解析が基礎となる（文献3）。収集したテキストデータ中から言葉を自動的に取り出して統計解析を行うことで、テキストデータを探索・提示し、その中から明示的にテキストデータ中のコンセプトを明示的に取り出して分析を深めることが可能になると考えられる（文献4）。さらに、テキストに対して人間が与えた正解カテゴリを組とした学習用データをもとに、各カテゴリの文書の特徴を自動学習する（教師あり学習）ナイーブベイズ分類器を用いたテキスト分類は、日常的に応用されている（文献5）。

C. 方法、結果、考察

以下、I, II, IIIの検討ごとに、方法、結果、考察を詳述する。

I. 食物アレルギー患者の自発的発話に関する解析

患者自身が、自身や家族の食物アレルギーに関する情報をどのように自発的に発話するのかを明らかにすることを目的として、自由記述文（自発的発話）の言語量の解析およびFA情報の充足率、内容の一致度およびFA情報の認知度を検討した。

1. 解析対象データの収集

1-1. アンケートA：食物アレルギーの症状に関するアンケート

ネットユーザーを対象とし、Webアンケート調査を実施した。調査実施日は2010年5月30日～6月4日とした。タイトルは「食物アレルギーの症状に関するアンケート」とし、Webアンケートの質問内容が表示される前のページに、集計結果を研究に利用し公開する旨について明記することで、倫理的配慮を行った。調査対象者はネット上のインテージネットモニターに登録している20－70歳の男女とした。該当者36,167名に3問からなるスクリーニング調査を行い、本人あるいは同居家族の食物アレルギーの有無により4群にわけ、それぞれの群の回答数が50を越えた時点で、調査を終了とした。総回答数は287、スクリーニング対象者に対する回収率は69.8%であった。回答者属性、受療行動、問診、FAの知識に関する調査項目（択一式）およびFA歴に関する調査項目（自由記述式）15問を用意した。回答は回答者属性などを含むCSV形式で取得した。

1-2. アンケートB：生活に関するアンケート

ネットユーザーを対象とし、Webアンケート調査を実施した。調査実施日は2011年1月25日～1月27日とした。タイトルは「生活に関するアンケート」とし、Webアンケートの質問内容が表示される前のページに、集計結果を研究に利用し公開する旨について明記することで、倫理的配慮を行った。調査対象者はネット上のインテージネットモニターに登録している10－70歳の男女とした。該当者1,392名に依頼し、回答数が400を越えた時点で調査を終了とした。有効回答数は435、回収率は31.7%であった。調査項目14問を用意した。回答は回答者属性などを含むCSV形式で取得した。

2. 解析方法

2-1. アンケートA-Q3

「あなたあるいは同居家族の食物アレルギーについて、できるだけ詳しく教えてください。原因となった食物や症状、原因と思われる食物を食してからアレルギーが出るまでの経緯、その後の治療や対応、その頻度や時間経過なども書いてください。」とする設問文に対する自由記述式回答文のうち、176名の回答を言語量解析に用いた。さらに、本人が食物アレルギー（以下、FA）の経験がある127名と、15歳未満の同居家族がFAを経験している49名の回答を、「原因食物について」「症状について」「経過について」「食品の調理加工状態について」「発症年齢について」「検査結果について」「アレルギー疾患の有無について」のそれぞれの項目について記載があるかどうかを目視により検討し、全ての回答者に対する該当項目の自発的発話者の人数を充足率として

算出した。

2-2. アンケート A-Q5

「食物アレルギーがありますか」とする設問文に択一式回答で「いいえ」とした回答者に「何か特定のものを食べると下痢をする、だるくなる、口の中がかゆくなるなど、過去にご自身や同居家族の食物アレルギーを疑った経験はありますか。」とする設問文を提示し、これに対する択一式回答で「はい」とした回答者に次の設問を行った。「その時のことをできるだけ詳しく教えてください。原因となった食物や症状、原因と思われる食物を食してからアレルギーが出るまでの経緯、その後の治療や対応、その頻度や時間経過なども書いてください。」とする設問文に対する自由記述式回答文のうち、本人にFA経験がない126名と、同居家族にFA経験がないとした101名の回答を解析に用いた。得られた結果について、事例を目視により分類解析した。

2-3. アンケート A-Q19

「生卵を食べると食物アレルギー症状が出るが、ゆで卵なら大丈夫、というように原因食物であっても調理加工によりアレルギー症状がでないことがあります。卵の他に同じような例を知っていますか。」とする設問に対する287名の択一式回答を用いた。得られた結果について、本人および同居家族のFA経験の有無と、調理加工の影響の認知の関係をクロス集計した。

2-4. アンケート A-Q17

「一般的な食物アレルギーの症状として、あなたが思いつくものを5つ自由に書いてください」とする設問に対する287名の自由記述式回答で得られた940件の回答を解析に用いた。得られた結果について、その内容を症状別に分類した。

2-5. アンケート B-Q10

「次の項目の中で病院に行くことを選択するものを選んでください」に対する435名の複数選択式回答を用いた。

3. 結果

3-1. FA に関する自発的発話における情報量の検討

FAの発症に関する自発的発話の言語量の解析結果をTable 1に示す。

成人患者（回答者本人）の症状に関する内容では、男性回答者の記述量が53±37文字（平均±SD）であるのに対し、女性は89±63文字と大きくなる傾向が見られた。

小児の同居家族がいる群の記述量が67±56文字であるのに対し、いない群では76±59文字と大きくなる傾向が見られた。さらに、FA患者が同居家族にいる群の記述量が71±63文字であるのに対し、いない群では76±59文字であったが、中央値は逆転していた。一方、小児患者（回答者の家族）については、回答者本人にFA経験がある群の記述量が62±52文字であるのに対し、本人にはない群では103±115文字と大きくなる傾向が見られた。

次に、「原因食物について」「症状について」「経過について」「食品の調理加工状態について」「発症年齢について」「検査結果について」「アレルギー疾患の有無について」の7項目について、それぞれの項目ごとの充足率を検討した結果をFig. 1に示す。

原因食物については、専門医であっても特定が難しい症例もあるため、原因食物の特定に至らなかった旨が記載されている場合には、特定の食品名等の記載がない場合でも「記載あり」とした。

原因食物についての充足率は最も高く、成人患者（回答者本人）で97.6%、小児患者(回答者家族)で92.3%と、いずれも90%を超えていた。症状については80%、臨床症状の経過については60%程度の充足率であり、成人患者と小児患者の間で大きな違いは見られなかった。一方、食品の調理加工状態、発症年齢、検査結果、アレルギー疾患の有無については、成人患者の充足率は小児患者のそれに比べて低かった。検査結果についてとアレルギー疾患の有無について、成人患者群では充足率が 10%に満たなかった。

3-2. FA に関する自発的発話の不一致の検討

Closed Questionにより、「FAはない」と回答した人に対して、Closed Question と Open Question を組み合わせて再質問することで、FAの既往ありと疑われた症例数の割合をFig. 2に示す。

成人患者群（回答者本人）では126人のうち24名（19.0%）が、食物アレルギーがあるかもしれないと不安になったことがあると回答した。小児患者群（回答者家族）では、35人のうち3人（8.6%)、成人患者群(回答者家族)では66人のうち5人(7.6%)において、同様の回答が存在した。食物アレルギーではないと判断した理由を、Table 2に示す。

最も多い理由は、症状が軽度であったことで、32 人中 13 人(40.6%)であった。毎回体調が悪くなるわけではないを理由とした人も多く、全体の28.1%であった。

なお、理由に関する自由記述文に対する、3-1.で述べた情報項目充足率は、原因食品については 89.7%に減少し、症状については100%に増加した。食品の調理加工状態や検査結果の有無などについては大きな違いが見られなかった(data not shown.)。

3-3. アレルゲン性に対する調理加工の影響に関する認知度の検討

回答者本人と同居家族の FA の有無による、調理加工の影響の認知度をクロス集計した結果をFig.3に示す。

調理加工がアレルゲン性に与える影響が最も認知されていると考えられるのは鶏卵の例である。また鶏卵アレルギーはリゾチーム製剤やワクチンの接種の際などに禁忌となるため、情報の非対称性は他のアレルゲンよりは低いと考えられる。

鶏卵を調理加工するとアレルゲン性が低下することを知っていると回答した割合は、鶏卵の例だけを知っていると回答した30.7%と他の例も知っていると回答した17.8%をあわせて約半数(48.5%)であった。家族に患者がいる I 群(60.7%)、III 群(57.8%)の認知は高く、家族の中に患者が全くいない IV 群は 32.4%の認知率と最も低い値となった。

3-4. FA の症状の認知度の検討

調査対象とした952件の回答のうち、487件が症状ではなく、食品名や動植物名をあげたものであった。これらは除外回答とし、残りの465件の回答を症状別に12分類した

ものをFig. 4に示す。

最も多かった回答は湿疹であり、同様の皮膚症状である蕁麻疹と合わせると全体の31.6%になった。全体的に皮膚症状の回答数が多く、消化器症状や不快感などの軽症状、痛みやしびれ等は認知が低い傾向にあった。

3-5. FAに対する自己判断に関する検討

調査対象とした435名の回答をまとめたものをFig.5に示す。

骨折(84.1%)や検査値の異常(56.8%)、呼吸器症状(53.6 %)の3項目は、全体の半数以上の人が、自己治療ではなく、医療機関の受診を選択した。一方、アレルギー症状として代表的な、食物アレルギー(24.1%)、花粉症 (26.9%)、アトピー・かぶれ (42.8%) は中等度であった。

4. FA患者の自発的発話に関する考察

FAの症状に関する自発的発話は、男性よりも女性のほうが情報量が多く、かつ、家族に FA がいない方が多くなる傾向を明らかにした。情報の含有率の比較では、原因食物 (97.6%) や症状 (87.8%) に比べ、調理加工(23.6%)、発症年齢 (22.8%)、その他のアレルギー疾患の有無 (3.3%) などの内容が含まれにくいことを明らかにした

FA患者は、経験した症状を語る時に、食品名のみを示したり、症状のみを示したりすることが多いことが明らかになった。副作用報告システムの目的は、発生頻度の低い副作用を発見することのほかに、背景因子や使用法、期間等も考慮した解析を実現することにある。従って、ある程度の概念語の使用は、情報を整理する際に有効であるが、概念語の多用により、重要な情報の漏れが危惧される。本検討の結果、本人や身近にFAを経験した人がいるほど、自発的発話の情報量が減少することが示された択一式の設問により、調理加工がアレルゲン性に与える影響の認知度を検討したところ、家族に FA 患者がいるほどその認知は高いことから、情報量の減少は、関連情報の摂取機会の多寡により、特定の用語が概念語に変化している可能性を示唆している。また、用語に対する概念の獲得は、日常的に医療に接する者から受ける可能性が高い。骨折のように医療技術の必要性が明らかなもの、呼吸器疾患のように命の危険を感じさせるもの、検査値の異常のように客観的な指標があるものに関しては、自発的に医療機関を受診する可能性が高いことは容易に予測できる。アレルギー症状は、生活習慣やしつけの一環として捉えられることも多く、骨折のように医療技術の必要性が明示的ではない。一方で、呼吸器疾患のように命の危険を感じるものから、かゆみといった一時的な症状として捉えられるものまで、その症状には幅がある。さらに、近年では患者血清を利用した抗原抗体反応による判定テストが保険適用されるようになり、客観的な指標が生まれつつある。こうした背景から、一定の行動パターンが生まれにくく、個別情報の重要性が示唆された。

II. FA患者の自発的発話のコーディングルールを用いた自動分類法の検討

患者の自発的発話の中に含まれる情報を効率良く収集するために、形態素解析により単語の意味を加味した抽出法の有用性を明らかにすることを目的に、自由記述文(自発的発話) 内の出現単語の解析とコーディングルールを用いた自動分類を検討した。

1. 解析対象データの収集

1-1. アンケートA:食物アレルギーの症状に関するアンケート

ネットユーザーを対象とし、Web アンケート調査を実施した。調査実施日は2010年5月30日～6月4日とした。タイトルは「食物アレルギーの症状に関するアンケート」とし、Web アンケートの質問内容が表示される前のページに、集計結果を研究に利用し公開する旨について明記することで、倫理的配慮を行った。調査対象者はネット上のインテージネットモニターに登録している20－70歳の男女とした。該当者36,167名に3問からなるスクリーニング調査を行い、本人あるいは同居家族の食物アレルギーの有無により4群にわけ、それぞれの群の回答数が50を越えた時点で、調査を終了とした。総回答数は287、スクリーニング対象者に対する回収率は69.8%であった。回答者属性、受療行動、問診、FAの知識に関する調査項目(択一式) およびFA歴に関する調査項目(自由記述式) 15問を用意した。回答は回答者属性などを含むCSV形式で取得した。

1-2. 質問投稿サイトからの質問文の抽出

Yahoo!デベロッパーネットワークで提供されているAPIにより、Yahoo！知恵袋内の質問投稿文を検索した。検索結果は自作のPerlプログラムを用いて、質問者属性を含むCSV形式にて取得した。検索用キーワードには「アレルギー」「食物アレルギー」を用いた。2010年3月1日に取得した質問文の中から、FAの発症を断定できていない42件の質問文を抽出した。

2. 解析方法

2-1 アンケート A-Q3 回答文における頻出用語の解析

「あなたあるいは同居家族の食物アレルギーについて、できるだけ詳しく教えてください。原因となった食物や症状、原因と思われる食物を食してからアレルギーが出るまでの経緯、その後の治療や対応、その頻度や時間経過なども書いてください。」とする設問文に対する自由記述式回答文のうち、本人が食物アレルギー(以下、FA) の経験がある127名の回答を回答者ごとに識別できるようCSVファイルに保存し、言語解析ソフトウェアKHCoder(ver.2 beta 22) の基本モジュールおよびTermExtractモジュールを用いて形態素解析による出現用語などの解析を行った。

2-2. 自動分類用コーディングルールを用いた分類

1-2.の方法で収集した42件の質問文を参考にして、自動分類を行うためのコーディングルールを作成した(Table 3)。コーディングルールを外部変数ファイルとして保存し、KHCoder上で利用した。

2-3. コーディングルールを用いた自動分

類

2-2. で作成したコーディングルールを、言語解析ソフトウェア KHCoder (ver. 2 beta 22) のコーディングモジュールを使用して、2-1. で解析したアンケート A-Q3 の回答文に回答者毎に適用し、各分類への自動分類率を検討した。さらに、強制抽出語の有無による自動分類率の比較を行った。

3. 結果

3-1. アンケート A-Q3. 回答文における頻出用語の解析

解析対象は 127 段落（1段落を1人の回答とする)、307 文であった。総抽出語数は 5,839 語、異なり語数は 1,040 語、そのうち、助詞や助動詞などを除外した分析使用語数は 797 語であった。これらの用語は複合語として出現する可能性が高い。そこで、専門用語辞書を用いた専門用語自動抽出システム「TermExtract」を用いて、複合語の検出を行った結果を Table 4 に示す。

複合語としてのスコアが最も高かったのは「卵アレルギー」であり、続いて「アレルギー症状」、「アレルギー反応」と続いた。「アレルギー」は後述する Table 5 に示した通り、最も出現頻度の高い用語であるとともに病態を表す概念語であり、スコア 5 以上の 17 語中 11 語において複合語として使用されていた。

そこで、これらの用語を強制検出語に指定して、再度解析を行った。解析対象は 127 段落（1段落を1人の回答とする)、307 文であった。総抽出語数は 5,823 語、異なり語数は 1,055 語、そのうち、助詞や助動詞などを除外した分析使用語数は 805 語に増加した。

Table 5 に品詞別単語出現頻度を、Tabale 6 に 10 件以上の頻出語を示す。最も出現頻度の高い用語は「食べる」という動詞であった。続いて「出る」「アレルギー」「症状」「病院」「治療」「卵」など、FA 症状を説明するのによく用いられると思われる言葉が挙げられた。

3-2. アンケート A-Q3. 回答文のコーディングルールを用いた自動分類

コーディングルールを用いた自動分類は、「症状について」「経過について」「食品の調理加工状態」「検査結果について」「アレルギー疾患の有無」について検討した。「原因食品名について」および「発症年齢」については、3-1. の結果から、出現が予測される語が非常に多岐にわたり、検出が困難であったため、コーディングルールを用いた検討対象から除外した。コーディングルールを用いた自動分類の結果を Table 7 に示す。

単語の出現の有無による自動分類の結果、「症状について」は目視による分類では 87.8%であったものが、コーディングルールを用いた自動分類においても 70.1%をカバーすることができた。目視に比べて抽出率が増加したのは、「食品調理加工状態」と「検査結果について」「アレルギー疾患の有無」の3分類であった。一方、「経過について」は 67.5%であったものが、22.0%まで低下してしまった。

なお、コーディングルールに含有する語と TermExtract によりスコア 5 以上にて抽出された複合語を強制抽出語として検討した結果、その抽出率には大きな差はなかった。

4. FA 患者の自発的発話のコーディングルールを用いた自動分類法の検討に関する考察

形態素解析およびコーディングルールに基づく、FA 患者の自発的発話に含まれる情報の自動分類を試みた。形態素解析により、頻出用語の抽出などを改善したが、FA 情報の含有率は、アレルギー疾患の有無(4.7%)については、目視を少しだけ上回る含有率となった．食品調理加工状態（70.1%)、検査結果(15.6%)に関しては、コーディングルールに含める検索用語の見直しにより改善する可能性が示唆されたが、その他の分類では検索用語の見直しによる改善が困難であると結論づけた。

アレルギー疾患の有無は、自発的発話が予測より少ないが、副作用報告を解析する上では欠くことのできない重要な情報である。これらの情報については、従来の形態素解析手法により、ピックアップできる可能性が示唆された。

当初、目的としていたようなコーディングルールを用いて、正しく自動分類することはできなかったが、頻出語などを手掛かりに、情報を有する可能性の高い標本を作成することはできると考えられる。さらに、形態素解析により、個々の自発的発話に対して、いずれの検索語が含まれているかを明らかにすることで、質的データを数的データに変換することが可能となると考えられる。

III. FA 患者の自発的発話のベイズ学習法を用いた自動分類法の検討

質問投稿サイトであるYahoo!知恵袋への質問投稿記事の中から、アレルギーに関する質問を抽出し、Web アンケートの自由記述文（自発的発話）を基にした少量サンプルから作成した手本により、出現単語とその頻度から分類用スコアを算出するナイーブベイズモデル（ベイズ学習法）の利用による大量サンプルの自動分類法を検討した。

1. 解析対象データの収集

1-1. アンケートA：食物アレルギーの症状に関するアンケート

ネットユーザーを対象とし、Web アンケート調査を実施した。調査実施日は2010年5月30日～6月4日とした。タイトルは「食物アレルギーの症状に関するアンケート」とし、Web アンケートの質問内容が表示される前のページに、集計結果を研究に利用し公開する旨について明記することで、倫理的配慮を行った。調査対象者はネット上のインテージネットモニターに登録している 20－70 歳の男女とした。該当者 36,167 名に 3 問からなるスクリーニング調査を行い、本人あるいは同居家族の食物アレルギーの有無により４群にわけ、それぞれの群の回答数が 50 を越えた時点で、調査を終了とした。総回答数は 287、スクリーニング対象者に対する回収率は 69.8%であった。回答者属性、受療行動、問診、FA の知識に関する調査項目（択一式）および FA 歴に関する調査項目（自由記述式）15 問を用意した。回答は回答者属性などを含む CSV 形式で取得した。

1-2. 質問投稿サイトからの質問文の抽出

Yahoo!デベロッパーネットワークで提供されている API により、Yahoo！知恵袋内の質問投稿文を検索した。検索結果は自作の Perl プログラムを用いて、質問者属性を含む CSV 形式にて取得した。検索用キーワードには「アレルギー」「食物アレルギー」を用いた。2010 年 11 月 8 日に検索を実行し、各検索ごとに最新のものから 900 件の質問文を取得した。

2. 解析方法

2-1. アンケート A-Q20

「食物アレルギーについて不足していると感じている情報がありますか？あなたが食物アレルギーについてもっと知りたいと思っている事などを具体的に教えてください」とする設問に対する 287 名の自由記述式回答文のうち、得られた 150 件の回答文を解析に用いた。得られた回答文について、「病院、専門医、検査について（診断）」「原因になる食物の種類と症状について（種類）」「啓蒙活動について（啓蒙）」「原因因子について（機序）」「食品表示法について（表示）」「治療法について（治療）」「対処法について（対処）」「リスク回避の方法について（回避）」「その他」の９分類に目視により分類した。1 人の回答文に複数の項目の内容を含んでいる際には、最も強いニーズをあらわす分類に分類した。

2-2. ベイズ学習による FA 患者のニーズ分析

言語解析ソフトウェア KHCoder（ver.2 beta 22）の「ベイズ学習による分類」モジュールを用いて、2-1. で得られた分類を学

習させた。2-1.で得られた回答文と正解の分類の組を、学習見本としての外部変数として取り扱い、ランダムにデータを欠損させたものを学習用の外部変数として取り扱った。

1) ベイズ学習モジュールの「外部変数から学習」メニューを用いて、比較的分類が容易であった 89 件の回答文の分類を学習させて「学習結果ファイル 1」を得た。このとき、交差妥当化を 10 回繰り返してその妥当性を検討した。

2) ベイズ学習モジュールの「学習結果を用いた自動分類」メニューを用いて、残りの 61 件の回答文に対して「学習結果ファイル 1」による自動分類を行った。

3) 不正解だった回答文に正解をあたえながら、再学習を繰り返して「学習結果ファイル 3」を作成し、これを用いて最終的に 150 件の回答文に対して、自動分類を行った。

4) Yahoo!知恵袋から「アレルギー」を検索語として収集した 900 件の質問投稿文に対して、「学習結果ファイル 3」を用いて、自動分類を行った。得られた質問投稿文から、4,280 文、106,986 語を抽出し、6,231 語を用いて解析した。

5) Yahoo!知恵袋から「食物アレルギー」を検索語として収集した 900 件の質問投稿文に対して、「学習結果ファイル 3」を用いて、自動分類を行った。得られた質問投稿文から、4,753 文、114,633 語を抽出し、5,660 語を用いて解析した。

3. 結果

3-1. 外部変数からの学習と学習結果を用いた自動分類結果

89 件の回答文を用いた学習により自動分類した結果、1 回目の学習における正解率は 52.8%であり、学習を繰り返して得られた最終的な正解率は 95.6%であった。1 回目の学習における交差妥当化の結果を Fig. 6 に示す。Kappa 統計量は 0.447 であった。学習に用いた回答文は、診断 15、種類 11、対処 7、啓蒙 11、回避 9、表示 16、治療 13、機序 7 であった。交差妥当化の結果、最も正解率が高かったのは表示の 87.5%であり、対処は 1 問も正解することができなかった。

61 件の回答文を、学習結果ファイル 1 を用いて、自動分類した結果、その正解率は 44.3%であった。最終的に、150 件の自由回答文を用いた再学習の結果、自動分類の正解率は 92.7%となった。150 件の内訳は、多い順に、啓蒙 30、診断 25、機序 24、表示 19、種類 18、治療 17、回避 10、対処 7、であった。

最終の学習結果ファイルの概要を Fig. 7 に示す。回答文と自動分類による得点分布の例を Fig. 8 に示す。質問文ケース 71 では、「病院」という語が出現するので、Fig. 7 より「診断」分類に 2.07 が加算され、「祖父」「世代」という語により「啓蒙」分類に 1.10、1.39 が加算され、最終的に「啓蒙」分類のスコアが 74.90 と最も高いことから、「啓蒙」分類に自動分類される。

3-2. 質問投稿サイトの質問文の自動分類結果

2010 年 10 月と 11 月の質問投稿サイトにおける「アレルギー」および「食物アレルギー」に関する投稿数を比較した結果を Table 8 に示す。さらに、「食物アレルギー」に関する投稿数の一年間の変化を Fig. 9 に

示す。変化を見る限り、アレルギーの中の一分野として定常的に質問が行われていた。また、食物アレルギーに関する質問は、健康、美容とファッションカテゴリでの 341 件よりも、子育てと学校カテゴリの 407 件の方が若干多かった。アレルギーに関する質問では逆転し、健康、美容とファッションカテゴリでの 514 件となり、子育てと学校カテゴリは 148 件と少なかった。

「アレルギー」を含む 900 件の質問投稿文に対して、学習ファイル 3 を用いて自動分類を行った結果を Fig. 10 に示す。さらに、「食物アレルギー」を含む 900 件の質問投稿文に対して、学習ファイル 3 を用いて自動分類を行った結果も同時に示す。

いずれの検索結果に対しても、最も多く分類されたのは「診断」に関する内容であった。食物アレルギーに関する投稿文では、アレルギー全般を対象にした時と比較すると、「診断」に関する内容が、67.7%から 50.6%へと少なくなると同時に、「啓蒙」に関する内容は 27.3%から 31.9%に増加した。さらに「機序」に関しても 7.8%から 13.8%に増加した。

4. FA 患者の自発的発話のベイズ学習法を用いた自動分類法の検討に関する考察

少量サンプルより作成した手本をもとに出現用語とその頻度から分類用スコアを算出するナイーブベイズモデル（ベイズ学習法）の利用による自動分類法を検討した。

先述の Web アンケートの必要な FA 情報に関する自由回答文 150 件を用いて、8 つの分類を行うための手本を作成し、学習を行い、正解率 92.7%のモデルを構築した。質問投稿サイトである Yahoo!知恵袋に投稿された「食物アレルギー」に関する質問文を収集し、得られたモデルを用いて分類したところ、病院や専門医、検査についてなどの診断(50.6%)に関するニーズが最も大きく、啓蒙活動(31.9%)、原因因子などについての機序(13.8%)に関するものが続いた。

KJ 法や親和図法では、集められたカードを分類し、島をつくり、島に名前を与える作業が行われる。多くの場合、同じ語が出現するカードを島にしたうえで、調整していくことになる。最終的には「青い海」と「青い空」から「夏の思い出」という島が生まれてくるが、このことは「青い」＝「夏の思い出」という訳ではない。本検討で用いた正解カテゴリは人的な親和図法や KJ 法作成と同じ思考回路で行われていると考えられ、本検討で用いたナイーブベイズモデルによる分類は、いわば、事象を整理するために使われる親和図法や、問題解決のために用いられる KJ 法を自然言語処理技術を使って具体化するものと捉えることができる。また、本手法は、お手本用のサンプルが大きくなるほど、自動分類の精度があがると考えられている。本検討では、自発的報告ではなく、要望文に対して適応したが、サンプル数が増えることで、自発的報告に応用することも可能であると考えられる。

D. 研究発表

1）論文発表

なし

2）学会発表

土橋　朗, 中ノ堂　ひとみ, 倉田　香織, 岡崎　光洋、質問ができる Web サイトからの食物アレルギー情報の抽出、第 20 回日本医療薬学会年会、2009 年 11 月、千葉、ポスター発表・

倉田　香織, 中ノ堂　ひとみ, 岡崎　光洋, 土橋　朗、食物アレルギー歴の聞き取りで想定される患者発話に関する意識・実態調査、第 20 回日本医療薬学会年会、2009 年 11 月、千葉、 ポスター発表・

「倉田　香織 , 中ノ堂　ひとみ , 岡崎　光洋 , 土橋　朗、親和図法に基つ▯く食物アレルキ▯ー情報に関するニーズ分析、日本薬学会 131 年会、2011 年 3 月、静岡、 ポスター発表。

E. 参考文献

文献 1　岡田 正人著、レジデントのためのアレルギー診療マニュアル、医学書院、2006.

文献 2　一般用医薬品セルフメディケーション振興財団平成 20 年度調査研究報告書「一般用医薬品による副作用の生活者からの自発報告システムの開発」主任研究者　望月真弓　、http://www.otc-spf.jp/symposium/pdf/b_05.pdf

文献 3　那須川哲哉著、テキストマイニングを使う技術/作る技術、東京電気大学出版局、2008.

文献 4　樋口耕一著、 KHCoder 2.x チュートリアル、http://khc.sourceforge.net/

文献 5　C.M. ビショップ著、元田浩ら訳、パターン認識と機械学習上下巻、シュプリンガー・ジャパン株式会社、2008.

F. 表

Table 1 FA の発症に関する自発的発話の言語量の比較

a) 自らの FA 症状に関する記述

	男性 回答者	女性 回答者
データ数	38	89
平均文字数	53	82
標準偏差	37	63
最大文字数	151	303
最小文字数	1	2
最頻文字数	78	66
中央値	50	67

	小児の同居 家族 FA あり	小児の同居 家族 FA なし
データ数	40	87
平均文字数	67	76
標準偏差	56	59
最大文字数	225	303
最小文字数	1	2
最頻文字数	94	33
中央値	55	59

	同居 家族 FA あり	同居 家族 FA なし
データ数	61	66
平均文字数	71	76
標準偏差	57	59
最大文字数	303	270
最小文字数	1	2
最頻文字数	78	22
中央値	63	59

b) 小児の同居家族の FA に関する記述

	男性 回答者	女性 回答者
データ数	21	45
平均文字数	50	95
標準偏差	49	104
最大文字数	188	536
最小文字数	1	1
最頻文字数	1	145
中央値	33	55

	回答者本人 FA あり	回答者本人 FA なし
データ数	15	34
平均文字数	62	103
標準偏差	52	115
最大文字数	188	536
最小文字数	1	1
最頻文字数	–	28
中央値	52	58

Table 2　FAを疑いながら、自己否定した理由

理由	回答数
症状が軽度だった	13
毎回体調が悪くなるわけではなかった	9
体調不良が原因だと思った	5
ただの食あたりだと思った	4
アレルギー体質ではなかった	4
その食物が原因食物となるとは考えられなかった	3
アレルギー検査の結果がマイナスだった	2
家族に食物アレルギーの人はいなかった	1
その他	3

Table 3 自動分類用コーディングルール

分類	キーワード
症状について	体調 or かゆい or 口腔 or アナフィラキシー or 発疹 or 腫れる or 湿疹 or じんましん or 蕁麻疹 or ジンマシン or 赤い or 痛い or 吐く or 嘔吐 or おう吐 or 吐き気 or 痒い or ブツブツ or　大丈夫
経過について	今 or 以前 or 昔 or 現在 or 子供 or 昨夜 or 昨晩 or 昨日 or 朝 or 今日 or 前 or 夜 or 食後
食品の調理加工状態	加熱 or 生 or 火 or 新鮮 or 鮮度 or 缶詰 or 調理 or レトルト or 卵白 or 白身 or 避ける or さける or 食べる
検査結果について	ＲＡＳＴ or RAST or ラスト or MAST or ＭＡＳＴ or マスト or テスト or 結果 or 測定値 or クラス or 数値 or アレルギー検査 or アレルギー or 検査 or 陽性
アレルギー疾患の有無	花粉症 or ぜんそく or ぜん息 or 喘息 or アトピー or ハウスダスト or ＨＤ or イネ科 or シラカバ or ネコ or 猫 or 犬 or イヌ or ダニ or ホコリ or セフゾン or アクディーム or ザジテン or カビ or だに or ほこり or クラリス or メイアクト

Table 4　TermExtract によりスコア 5 以上にて抽出された連結語

抽出複合語	スコア	抽出複合語	スコア
卵アレルギー	36.04217	青魚	6.640092
アレルギー症状	27.26926	アレルギー原因	6.344228
アレルギー反応	24.24619	アレルギー食品	6.344228
皮膚科	21.40695	そばアレルギー	6.061547
甲殻類	13.55403	牡蠣	6.000000
アレルギー剤	12.12309	ジンマシン	6.000000
かにアレルギー	11.46531	学校給食	6.000000
アレルギーテスト	10.66968	アレルギー体質	5.732657
		小麦アレルギー	5.334838

Table 5 品詞別単語出現頻度*

名詞	頻度	サ変名詞	頻度	動詞	頻度	形容詞	頻度
アレルギー	39	治療	22	食べる	131	痒い	13
症状	30	発疹	20	出る	66	悪い	10
病院	29	下痢	11	飲む	20	赤い	6
蕁麻疹	20	注射	9	行く	20	酷い	4
原因	15	点滴	8	思う	14	小さい	4
全身	14	摂取	7	治る	11	多い	4
湿疹	12	嘔吐	7	治まる	8	痛い	4
牛乳	9	対応	5	受ける	7	強い	3
食品	9	発症	5	出来る	7	軽い	3
皮膚	9	反応	5	言う	6	古い	3

*強制抽出語の指定をしないで抽出した結果を示す。

Table 6　出現数 9 以上の頻出語*

抽出語	出現数	抽出語	出現数	抽出語	出現数
食べる	131	その後	19	治る	11
出る	66	薬	17	特に	11
アレルギー	39	原因	15	悪い	10
症状	30	思う	14	口	10
病院	29	時間	14	牛乳	9
治療	22	全身	14	食品	9
卵	21	体	13	注射	9
飲む	20	痒い	13	皮膚	9
行く	20	湿疹	12		
発疹	20	下痢	11		
蕁麻疹	20	顔	11		

*強制抽出語の指定をしないで抽出した結果を示す。

Table 7　コーディングルールを用いた自動分類の結果

分類	目視による分類	コーディングによる自動分類	
		強制抽出語 なし	強制抽出語 あり
原因食物について	97.6%	-	-
症状について	87.8%	68.5% (-19.3)	70.1% (-17.7)
経過について	67.5%	19.7% (-47.8)	22.0% (-45.5)
食品調理加工状態	23.6%	66.1% (42.5)	70.1% (46.5)
発症年齢について	22.8%	-	-
検査結果について	4.1%	25.2% (20.9)	19.7% (15.6)
アレルギー疾患の有無	3.3%	3.9% (0.6)	4.7% (1.4)
該当なし	-	5.5%	7.9%

() 内は目視からの増減ポイント数を示す。- 検討なし

Table 8　1ヶ月間の投稿数の変化

	アレルギー (件)	食物アレルギー (件)	食物アレルギーの占める割合
2010.10.8	41310	1772	4.3%
2010.11.8	42921	1828	4.3%
増加数/月	1611	56	3.5%

G. 図

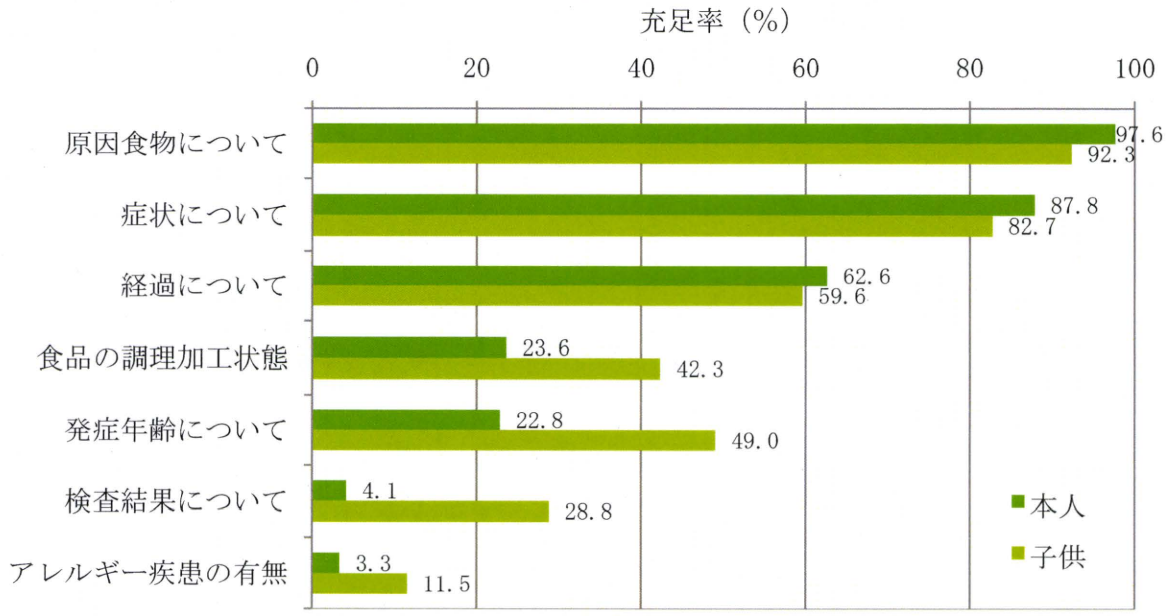

Fig. 1. 自発的発話における FA 情報項目の充足率

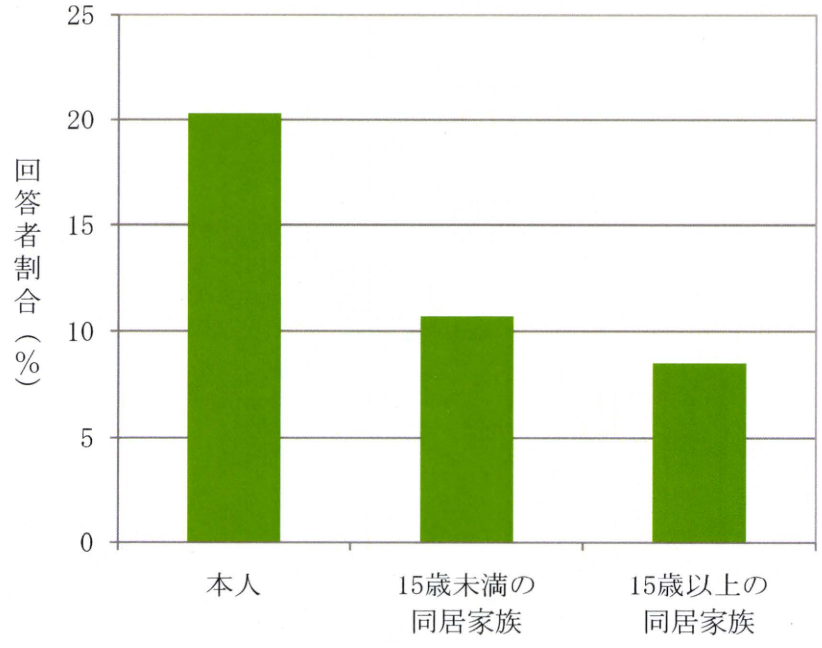

Fig. 2 FA 経験に関する自発的発話の不一致率

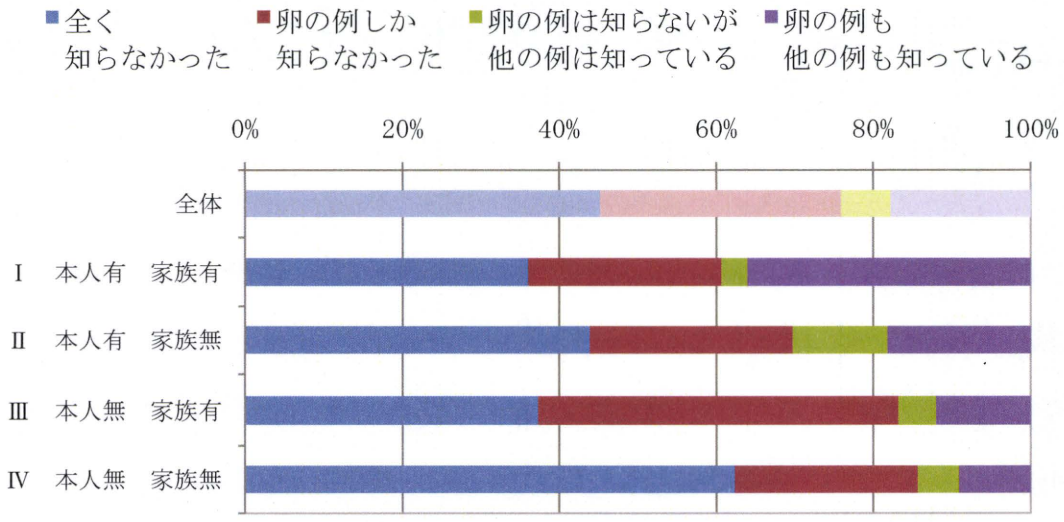

Fig. 3　食物アレルゲン性に調理加工が与える影響の認知度

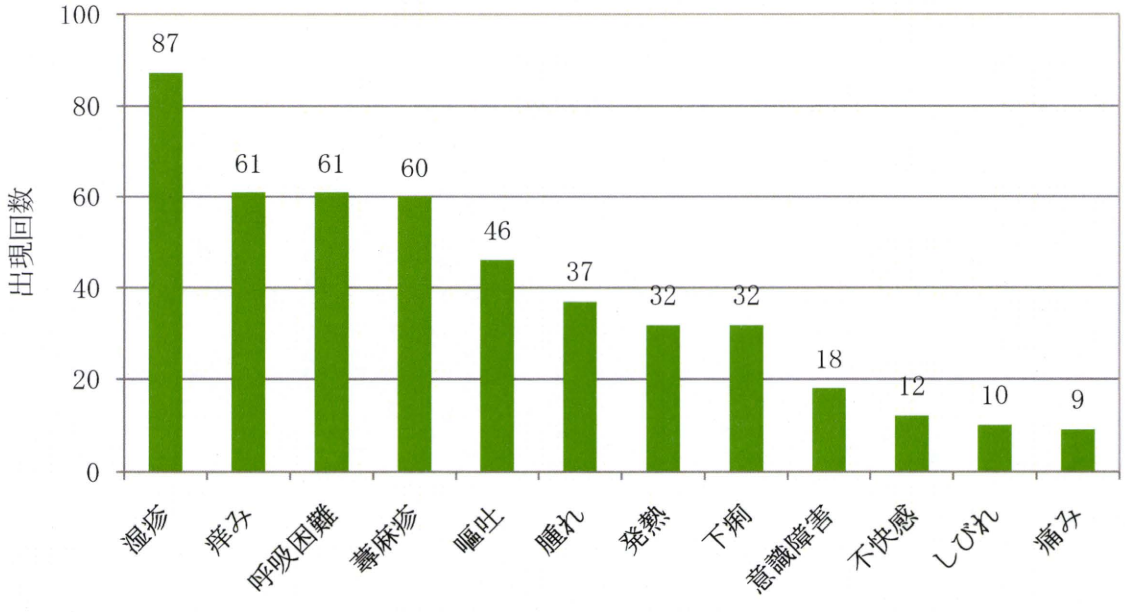

Fig. 4　代表的なFA症状の認知度