厚生労働科学研究費補助金（がん対策推進総合研究事業）
分担研究報告書

ソーシャルメディアを用いた病院の医療提供体制に関する評判・風評調査

研究分担者　荒牧英治　奈良先端科学技術大学院大学 先端科学技術研究科 教授
研究分担者　若宮翔子　奈良先端科学技術大学院大学 先端科学技術研究科 准教授
研究協力者　勘場　大　奈良先端科学技術大学院大学 先端科学技術研究科 大学院生

研究要旨

患者報告アウトカム（Patient Reported Outcome）など，患者の声を医療に活かす研究は多い．本研究では，本邦の代表的なウェブ上のQAサービスである「Yahoo！知恵袋」を用いて，医療に対する不平，不満，疾患に対する悩みなどの情報の収集を試みた．この結果，6993件の悩み発言が抽出され，これを自動分類した結果，「自覚症状でがんを心配」が多いことが判明した．今後，さらに，これまで知られていなかった不安（アンメットニーズ）に絞って抽出する方法を開発する予定である．

A．研究目的

近年のWebの発達に伴い，ブログ執筆が活発な本邦において，すでに膨大な病いに関する悩みが集積されている．なかでも治療が長期化しがちながんに関する闘病記やブログの数は多く，例えば，TOBYO[1]において乳がんは全体の1割以上（約6,600件）を占めており，代表的なquestions and answers（QA）サイトであるYahoo!知恵袋（以降はYahoo Japan QA; YJQAと呼ぶ）においても，「乳がん」を含む質問は約60,000件も存在している．

このような背景の中，近年では集積された情報を利活用しようとする研究も多い．例えば，Rosenblum and Yom-Tov [2] は，Microsoft Bing検索エンジンとWeb QAサイトであるYahooAnswersを使って，Attention Deficit Hyperactivity Disorderに関する情報を人々がどのように検索するかについて調査し，インターネットを使って情報を収集していることを明らかにした．Park et al.[3] はブログやQAサイトのテキストデータからDiabetesに関係するmedical conceptsの利用法を調査した．このようにWeb上の患者記録の利用は，患者視点のニーズを質的にタイムリーに把握することができる．

一方，集積された情報の利用には限界もある．もっとも大きな問題はその膨大なデータを精査することが困難な点である．静岡分類[4,5]のような分類がされていないため，求める情報，例えば，副作用に関する情報，などを集めるためには，一つ一つに目を通し，人手で課題をピックアップすることになり限界がある．そこで，テキストデータの自動分類が必須となる．

本研究では，自然言語処理の技術を用いYJQAのテキストデータから乳がんの悩みを抽出し自動分類する．

B．研究方法

【材料】

本研究では静岡分類及び2018年1月1日から2020年7月31日の間にYJQAに投稿された7,993件の質問のテキストデータを用いる．静岡分類とは，がん患者の悩みや負担を体系化したものであり，16の大カテゴリ，631の小カテゴリからなる．YJQAは質問のみのテキストデータであり，カテゴリは付与されていない．悩みを抽出し自動分類するために，以下の2種類のコーパスを用いて学習する．

A　　静岡分類コーパス

B　　YJQA コーパス

静岡分類コーパスは静岡分類のデータベースから取得した，静岡分類カテゴリコードと静岡分類カテゴリ名（以降，この 2 つを総称して静岡分類カテゴリと呼ぶ）がラベル付けされたがん体験者の悩みのコーパスである．

YJQA コーパスとは,2018 年 1 月 1 日から 2020 年 6 月 9 日までに YJQA に投稿された乳がんに関する質問からランダムに 1,000 件を抽出し，１件につき最大 3 種類の静岡分類カテゴリを人手で割り当てたラベル付きコーパスである．

【方法】

学習データを用いて，分類対象データを静岡分類カテゴリに分類するアルゴリズムを示す．

1. 2 種類のコーパスを静岡分類カテゴリごとに TF-IDF（Term Frequency - Inverse Document Frequency）で重み付けした単語ベクトルに変換する
2. 分類対象データを YJQA の質問ごとに TF-IDF で重み付けした単語ベクトルに変換する
3. 1. と 2. のコサイン類似度を計算し，最も関連度の高い静岡分類カテゴリに分類する

単語ベクトルへの変換には，形態素解析辞書 mecab-ipadic-NEologd を用いて，名詞，動詞，形容詞を抽出した．記号や数字は除外した．また， sklearn.feature_extraction.text モジュールに含まれる TfidfVectorizer 関数を TF-IDF の計算に利用した．本関数におけるオプションはデフォルトの設定を利用した．

以降，静岡分類コーパス，YJQA コーパス及びそれらを合わせたコーパスの 3 つを用いて構築した分類手法のことをそれぞれ Description-based (D-based), Example-based (E-based), Description and Example combination-based (D+E-based)と呼ぶ．

C．研究結果

各手法の精度評価のために，学習データに対する正解率を以下のように計算した：

D-based の正解率計算アルゴリズム：

1. 静岡分類コーパスの単語ベクトルと YJQA コーパスの単語ベクトルのコサイン類似度が最大のカテゴリ（トップ 1)，及び，上位 10 カテゴリ（トップ 10）を求める
2. 1. のカテゴリが，YJQA コーパスの正解カテゴリセット（最大 3 カテゴリ）に含まれている割合を計算する．この割合をトップ 1 精度（Acc@1 と表記する）とする．トップ 10 については，トップ 10 のいずれか 1 カテゴリでも正解カテゴリセットに含まれていれば正解とみなし，正解の割合を計算する．この割合をトップ 10 精度（Acc@10 と表記する）とする．

E-based では 5-fold Cross Validation を用いて正解率を計算する．具体的には，YJQA コーパスを学習データと検証データに 5 分割し，それらのコサイン類似度を計算し，D-based の正解率計算アルゴリズムと同様に静岡分類カテゴリに含まれている割合を計算し,それらの平均，不偏標本標準偏差，中央値を求める．

D+E-based では E-based の学習データに静岡分類コーパスを含めて，E-based と同様のアルゴリズムで正解率を計算する．

表 1 に上記のアルゴリズムで計算された精度を示す．D-based の正解率は約 10%であり，Acc@10 での結果でも約 30%であった．一方，E-based 及び D+E-based での正解率は平均値及び中央値の両方で約 50%であり，Acc@10 は約 70%であった．E-based は YJQA の質問を分類するために最適化された分類方法であるが，全ての静岡分類のカテゴリを網羅しているわけではない．一方で，D+E-based は全ての静岡分類のカテゴリを網羅しており，正解率においても E-based と大きな違いはなかった．そのため，本研究においては D+E-based の結果で解釈を与えることとする．

表 1 各手法の正解率

評価指標	統計量	D-based	E-based	D+E-based
Acc@1	平均値	0.1096	0.4891	0.4781
	SD	-	0.017	0.018
	中央値	-	0.4835	0.4725
Acc@10	平均値	0.2946	0.6960	0.7062
	SD	-	0.030	0.201
	中央値	-	0.7015	0.7106

SD: Unbiased Sample Standard Deviation

D+E-based の分類結果例を示す．まず，分類対象データを静岡分類カテゴリに分類した結果のうち，頻度上位カテゴリを表 2 に示す．上位 10 カテゴリで全体の 61.9%を占めていた．最も多くの質問が分類されたカテゴリは「自覚症状でがんを心配（1,661 件)」であり,全体の 23.8%であった．D+E-based で分類したカテゴリは全部で 448 カテゴリであった．上位 1 位から上位 2 位の変化率は 57.7%で最も大きく，上位 20 位までの変化率は 40%から 20%，それ以降は 10%程度となり，裾の長い分布となった．

表 2 D+E-based による分類結果（上位 10 カテゴリ）

静岡分類カテゴリコード	静岡分類カテゴリ名	件数
16.3.1.1.	自覚症状でがんを心配	1661 (23.8%)
16.2.1.1.	がん検診に関すること	702（10%）
16.3.2.1.	がんの疑いに関すること(その他)	494 (7.1%)
12.2.4.1.	がんに対する知識	419（6%）
	不足による不安	
3.1.3.5.	受けた治療（選択）が正しかったか	255 (3.6%)
3.2.2.2.	結果やその動向が心配	234 (3.3%)
12.1.1.1.	再発・転移するかもしれない不安	225 (3.2%)
9.1.2.2.	医師に質問や心配事を言い出しにくい	137（2%）
12.3.2.3.	がんのことが頭から離れない	111 (1.6%)
9.1.1.1.	医師の言葉や態度	93（1.3%）

D．考察

本研究の応用先として，薬害有害事象の抽出（シグナル検出）とアンメットニーズ抽出について検討する．

【副作用シグナルへの応用可能性】

投稿された質問から副作用を抽出することは，薬の安全性の情報を多く集めることができることから製薬会社や患者にとって非常に有益であると考えられる．例えば，静岡分類の大カテゴリ11「症状・副作用・後遺症」に分類された質問には副作用の情報を含む質問もあると考えられるため，その質問文に対して固有表現抽出を適用することにより，副作用の情報を抽出できるであろう．

表3に,大カテゴリ11に分類された質問を示す（コサイン類似度が高いものを抜粋).最初の質問には白血球が下がる,2番目には手足の痺れという情報が含まれていたが，薬の情報がないためどの薬の副作用かを特定することができない．一方，3番目の質問にはFEC治療で脱毛という副作用が発生したという情報が含まれている．こ

のように，副作用は抽出できるが，医薬品の情報について粒度が足りない場合がある．D+E-basedで大カテゴリ11「症状・副作用・後遺症」に分類された割合は6.5% (470件) であり，そのうち100件をランダムサンプリングし,具体的な薬の名称があったものは15件であった.

表 3　静岡分類の大カテゴリ 11 に分類された質問とその分類カテゴリ

質問	薬剤	副作用
私は乳がん治療中で、副作用で白血球が下がっていて免疫力が高くありません。それなのに義実家にて 3 世帯、8 人が集まり誕生日会をするみたいで、コロナウイルスに感染するのではないかと不安で、私は不安で行きたくありません。（中略）私だけではなく夫にも行ってほしくないですが、夫は全く気にしていないようです。行かなくてすむ方法はないでしょうか？	乳がんの抗がん剤	白血球減少
抗がん剤治療の副作用でおこるしびれは改善できますか？私の妹が、乳がんで抗がん剤治療を行っているのですが手足のしびれがひどく辛いようです。何かしびれを和らげる方法はありますか？やはり抗がん剤治療を中止するしかないですか？	乳がんの抗がん剤	しびれ
抗がん剤治療をしています。3 週間に一回の点滴の FEC 治療で、乳がんです。4 クール終了し、次はまた違う抗がん剤が始まると言うところで、脱毛について、質問します。髪の毛は赤ちゃんみたいな感じで残っていて、脱毛してると言えるのですが、他の部分のまつ毛や眉毛、すね毛、下の毛も抜けると聞きますが、髪の毛以外のところが、脱毛していません。2 回目ぐらいに主治医にも不安で聞いたのですが、抜けてくるよ～と言われましたが、4 回打って、変わりません。薬の効き目が悪いのかなと不安です。経験のある方、もしくわ何かわかる方、ご助言お願いします。	FEC 治療	脱毛

【アンメットニーズへの応用可能性】

患者のアンメットニーズは社会において大きな課題になりつつある．特に，アンメットニーズについて，誰がそのニーズに答えるかについては,まだ十分に検討されているとは言えない．一部の有料のQAサイト[6, 7]を除き，一般的なQAサイトでは，一般の参加者が回答するが，中には医師による回答が望まれるものがある．

アンメットニーズは，サービスやリソースが不足しているために対応しきれていないニーズや，これまで存在しなかったニーズと考えることができる．前者は高頻度カテゴリを医療スタッフと議論することにより見つけることができるかもしれない．多くの患者が訴えているのにも関わらず，これまでの対応が不十分だったニーズを洗い出すために役立つと考えられる．一方，後者の新たなニーズについては，低頻度カテゴリや話題語（例えば「コロナ」）で検索することにより抽出できると考えられる．

表4には低頻度カテゴリ及び「コロナ」で検索したときの質問とその分類カテゴリの一部を示

す.1つ目は骨転移したがん患者の車の運転についてのアンメットニーズである.2つ目は誤診に対する訴訟であり,3つ目はCOVID-19に関係するアンメットニーズである．本研究においては実際に質問文を読みながらアンメットニーズを抽出したが，アンメットニーズの自動分類モデルの構築は今後の課題としたい.

表4　アンメットニーズと考えられた質問とその分類カテゴリ

質問	カテゴリコード カテゴリ名
母が骨転移ありの乳癌．骨転移は骨折のリスクも大きいそうですが、今後車の運転とかさせないようにするべきですか？また80歳の祖母も、未だに運転しています。しかし高齢者の事故も多いですし、若い人より事故を起こすリスクもまた高いでしょう。最悪のケースを想定したら、「祖母から車取り上げるのは可哀想」とか言っている場合ではなく、運転をさせないようにすべきでしょうか？また運転を辞めてもらう場合、どのような状況・場で言うべきですかね？外で働くのが生きがいの祖母から、今の職場に行く手段である車をとりあげるのも酷で…さらに私は免許を持っていながらペーパーなのですが、母や祖母の分も家族で出かける時に運転するために教習所に通い直すべきですかね？ブランクありなので、いきなり一般道を運転出来そうにもなく…	8.2.1.1. 交通事情が悪い
がんの誤診は訴えられるか教えて下さい。2年前マンモグラフィーで要再検になり、ある病院を受診しました。エコーで疑わしいものがあったので、その場で細胞診をしました。細胞診は検体不良で結果が出ず、組織診をお願いしました。マンモトーム生検などは紹介して別の病院で1泊になると言われ、白黒するのに時間をかけたくなかった私は、その病院でできる確定診断「外科生検」をしました。その結果「乳腺症」と診断され、定期的に病院で診察する旨言われましたが、組織検査の結果で納得できないものがあり、別の病院で検査をしていただきました。結果は乳癌・・確定診断で外すなんてありえないです・・もし乳腺症で納得し発見が遅れたら・・ぞっとします。今も普通に診察・診療している医師を訴えたいのですが、医療裁判などは難しいとお聞きします。訴える事は可能でしょうか？勝ち目はあるでしょうか？	5.3.1.3. セカンドオピニオンを受けるべきか
2月初旬に乳ガンの温存手術をして、4月から放射線治療が始まります。ですが、今まさにコロナで大変な時期で、毎日病院に通院する事に不安を感じています。個人的な対策をする以外には、なかなかできることはないのでしょうか。	8.2.1.3. 頻回の通院が大変

今年の 2 月に乳がんの温存手術を受け、放射線治療の予定でしたが、コロナの影響で延期になっています。状況が改善するのはまだまだかかりそうですが、放射線治療を開始するのは今の時期避けた方がよいのでしょうか。ホルモン治療はしていますが、放射線治療をこのまま受けずにいるのも不安が募ってしまいます。	11.1.3.6. 放射線による副作用の症状に関すること(その他)
88 歳の母が特養に入所していて、37.5 の発熱がありました。乳がんなので、乳がんが原因の発熱なのか、コロナなのか風邪なのかわかりません。乳がんの発熱の場合、どのような症状がありますか？かかりつけの専門医の診察を受けたほうがいいですか？また乳がんが原因でない場合、特養で高熱が出るということは、職員さんからウィルスが移ったということでしょうか？3 月からずっと面会禁止です。	11.2.1.5. 発熱

E．結論

本研究で構築した手法を用いて，YJQAに投稿された質問を約70%の正解率で静岡分類に分類することが可能となる．分類するだけでは患者の悩みを解決したことには繋がらないが，まずはどのような悩みがどれくらいあるかを俯瞰的に見ることは，患者視点の悩みを解決するサービスの優先順位付けのために重要な情報になるであろう．今後はこれらの情報をもとに受け皿となるサービスを検討したい．

F．研究発表

1．論文発表

なし

2．学会発表

勘場大，眞鍋雅恵，若宮翔子，荒牧英治：自然言語処理を用いた WebQA サイトからのがん患者の医療ニーズ抽出,第 40 回医療情報学連合大会，2020.

G．知的財産権の出願・登録状況

該当なし

引用文献

1. INITIATIVE INC. URL: https://www.tobyo.jp [accessed 2021-03-01]
2. Rosenblum S, Yom-Tov E. Seeking Web-Based Information About Attention Deficit Hyperactivity Disorder: Where, What, and When. J Med Internet Res 2017; 19 (4):e126 https://www.jmir.org/2017/4/e126/
3. Park MS, He Z, Chen Z, Oh S, Bian J. Consumers' Use of UMLS Concepts on Social Media: Diabetes-Related Textual Data Analysis in Blog and Social Q&A Sites. JMIR Med Inform 2016;4(4):e41 https://medinform.jmir.org/2016/4/e41/
4. Shizuoka Cancer Center. URL; https://www.scchr.jp/cancerqa/start_shizuoka.html [accessed 2021-03-01]
5. 「がんの社会学」に関する合同研究班. がん体験者の悩みや負担等 に関する実態調査報告書 がんと向き合った 7,885 人の声. 2006.
6. M3,Inc. AskDoctors. https://www.askdoctors.jp/ [accessed 2021-03-01]
7. JustAnswer. https://www.pearl.com/ [accessed 2021-03-01]