D-010

Web テキストから医療の悩みを抽出する：がん体験者の悩みの自動分類

Extracting Patients' Distress of Their Medical Care from Web Texts: The Automatic Classification of Cancer Patients' Distress

宮部真衣 † 島本裕美子 † 荒牧英治 †
Mai Miyabe Yumiko Shimamoto Eiji Aramaki

1. まえがき

現在，ソーシャルメディアが急速に普及し，個人での情報発信に対する敷居が下がっている．これに伴って，ブログなどを介した個々人による情報発信が活発化している．本研究では，ソーシャルメディアの医療応用に着目する．

医療において，患者自身の訴えを聞くことの重要性が再認識されている [1]．近年，がん化学療法において，患者報告アウトカム（Patient Reported Outcome; PRO）と呼ばれる，患者から直接得られた健康面についての情報を取り入れた臨床試験も行われるようになった [2]．PRO は面接，自己記入式質問票，生活・健康状態・治療についての日誌などのデータ収集ツールを介して得られるものである．このような方法は，PRO 作成のための様々なコストがかかることから，患者が従来用いている媒体から患者自身の訴えを抽出できることが望ましい．

Web 上で公開されているブログの中には，疾病経験を共有する患者 SNS や患者の闘病経験を共有するブログ（以降，闘病ブログと表記する）が存在する．つまり，医療機関の受診や診療時の悩みや負担を医療者に直接伝えるのではなく，ソーシャルメディア上で執筆・公開する患者が存在すると考えられる．このような患者が普段用いているブログから，患者の抱える悩みや訴えを抽出することができれば，PRO として活用可能な情報を少ない負担で得ることができる可能性がある．

そこで本稿では,「静岡分類」と通称されるがん体験者の悩みの分類を用いて，闘病ブログに含まれる患者の悩みの抽出・分類を行い，その精度について報告する．

2. がん体験者の悩みの自動分類

2.1 静岡分類：がん体験者の悩みの分類

静岡分類とは,「がんの社会学」に関する合同研究班が，がんに関する不平・不満，悩み 7855 件を整理し，16 の大分類から始まる階層構造に体系立てた指標である [3]．

表 1 に静岡分類における 16 の大分類と，それぞれに該当する悩みの例を示す．本研究では，Web テキストをこの 16 分類に分類することを目指す．

2.2 コーパス

本研究では，以下の 2 種類のコーパスを用いる．

コーパス A： アンケート調査データ（総数：22651 文）
Web 版がんよろず相談 Q&A サイト[1]で提供されている，がん体験者のアンケート調査データを用いる．

表 1: 静岡分類と悩みの該当例

大分類		例
01	外来	自分の将来と、これから治療を受けるためにはどの病院がよいか選択に悩んだ。
02	入院・退院・転院	診断を受け、早く入院して治療を受けたかったが、すぐに入院の連絡がなく不安だった。
03	診断・治療	抗がん剤で治療中で将来手術予定だが、本当に手術したら再発し難いのか。
04	緩和ケア	余命あと僅かとなったとき、延命治療を受けるべきかどうか。
05	告知・インフォームドコンセント・セカンドオピニオン	食道手術の時咽頭も切除したが、音声が出なくなるとは思わなく、説明もなかった。
06	医療連携	複数の病院で治療を受けたが、病院や医師間での報告、連絡、相談で嫌な思いをした。長期間の闘病で心身のダメージと闘い、前向きな姿勢を貫くことの悩み。
07	在宅療養	自宅にベッドをレンタルしようと思うが、どのような手続きをしていけばよいのか。
08	施設・設備・アクセス	薬の副作用に悩むとともに、週 1 回診察、注射があり、通院が大変である。
09	医療者との関係（現在の病院）	心の悩みを聞いてもらえる相談室のようなものがあればよいと思う。
10	医療者との関係（以前の病院）	以前から精密検査を受けていながら、早期発見できなかったことが残念だった。
11	症状・副作用・後遺症	これから老いて寝込んだとき、言葉が出にくい。
12	不安などの心の問題	診断をされたときは重い病気にかかったと思いがっかりしてした。この病気は一時よくなっても 2〜5 年位の命と思った。
13	生き方・生きがい・価値観	まだ若いので、ベッドの上で横になっているときに、将来に対する不安もあった。
14	就労・経済的負担	高額療養費のことを悩む。
15	家族・周囲の人との関係	1 人暮らしなので、入院すると自宅を閉めきりになり、一般世間との付き合いが遮断され、情報や交際に支障をきたすことが苦になった。
16	がんの予防・がん検診・がんの疑い	大腸ポリープがあって、次回２〜３年後に検査予定になっているが、最近便に血が混じっているので不安。

コーパス B： 闘病ブログデータ（総数：2498 文）
Google 検索および闘病サイトライブラリである TOBYO[2]から，がんの告知に関する記述を含む 100 ブログ（各ブログ 1 エントリ）を抽出した．

コーパス A に含まれる各文は，静岡分類上のコードが付与され，提供されているため，既存の分類結果をそのまま用いた．

コーパス B については，100 ブログに含まれるテキスト 2499 文に対し，人手で分類作業を行った[3]．

各コーパスの静岡分類による分類結果を表 2 に示す．

†京都大学学際融合教育研究推進センター

[1] http://cancerqa.scchr.jp/

[2] http://www.tobyo.jp/

[3] 10%のデータに対して 2 名で分類作業を行った結果，十分な一致が確認できた．そのため，残りは 1 名の作業者が分類を行った．

表 2: 各コーパスにおける分類結果

静岡分類	コーパス A: アンケート調査データ（文）	コーパス B： 闘病ブログデータ（文）
01	112	9
02	182	18
03	1696	84
04	158	12
05	250	128
06	2	1
07	15	3
08	48	1
09	413	27
10	146	1
11	3391	26
12	10914	33
13	922	139
14	2019	-
15	2369	13
16	14	-
該当なし	-	2129

闘病ブログは，1 テキストが複数のカテゴリに属する場合もある．

2.3 分類器

2.2 節で述べたコーパスを用いて，悩みの自動分類器を構築した．今回は，素性として形態素（1-gram）を用いた．SVM[4]による学習には，多項カーネル（d=2）を用い，パラメータはデフォルト値を用いた．

3. 実験

3.1 検証内容

構築した分類器の精度を検証するため，2.2 節で述べたデータを用いて 3 種類の実験を行う．それぞれの実験において，以下の内容を検証する．

実験 1： アンケート調査データをもとに，16 種類の悩みを正しく分類できるか？

実験 2： 闘病ブログデータにより，悩みの有無を分類できるか？

実験 3： アンケート調査データをもとにした分類器は，闘病ブログも正しく分類できるか？

実験 1 では，コーパス A の各分類に該当するものを正例とし，10 分割交差検定を行う．実験 2 では，静岡分類 01〜16 のいずれかに該当した場合に悩みを含むテキストとみなし，コーパス B における悩みの有無の判定が可能かどうかを 10 分割交差検定により検証する．実験 3 では，コーパス A をトレーニングデータ，コーパス B をテストデータとして精度を確認する．

3.2 実験結果

まず，実験項目 1 について検証する．図 1 に，10 分割交差検定の結果を示す．図 1 より，分類 07（在宅医療），11（症状・副作用・後遺症），12（不安などの心の問題），14（就労・経済的負担），15（家族・周囲の人との関係）については，比較的高い精度（F 値 0.7 以上）で判定可能であることが分かった．

次に，実験項目 2 について検証する．表 2 におけるコーパス B の「該当なし」2129 件を負例，何らかの分類に該当したテキスト 369 件を正例とし，10 分割交差検定を行った結果，F 値は 0.280 となった．今回はいずれかの大分類に該当すれば，悩みを含むとみなしたが，分類によって性質が異なる可能性がある．今後，大分類の違いを考慮した上で，精度を向上させる方法を検討する必要がある．

[4]TinySVM を利用した．
http://chasen.org/ taku/software/TinySVM/

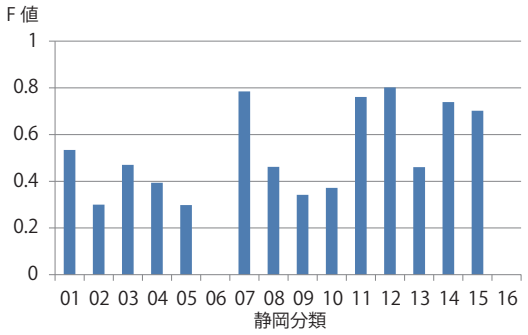

図 1: 10 分割交差検定の結果

最後に，実験項目 3 について述べる．コーパス A（アンケート調査データ）をトレーニングデータ，コーパス B（闘病ブログデータ）をテストデータとして検証を行った結果，静岡分類 01〜16 のいずれも，判定精度（F 値）は 0.1 を下回った．コーパス A と B ではテキストの性質が異なる可能性があるため，今後，コーパスの性質について詳細な分析を行う．

4. むすび

本稿では，「静岡分類」というがん体験者の悩みに関する 16 の分類を用いて，アンケート調査データおよび闘病ブログデータを対象とした悩みの分類器を構築した．分類器の精度を検証した結果，(1) アンケート調査データについては，いくつかの分類項目については高精度に判定可能であること，(2) 大分類のすべてを用いて悩みの有無を分類しても，高精度な判定はできないこと，(3) アンケート調査データと闘病ブログデータとではテキストの性質が異なる可能性があることが明らかとなった．

今回，闘病ブログに関しては高い分類精度が得られなかったため，今後，精度向上方法の検討を行う．

謝辞

闘病ブログの分類にあたり，久保圭氏，四方朱子氏に多大なる御協力をいただいた．ここに深く感謝の意を表する．本研究は，JST 戦略的創造研究推進事業の助成による．

参考文献

[1] KE Lasch, P Marquis, M Vigneux, L Abetz, B Arnould, M Bayliss, B Crawford, and K Rosa. Pro development: rigorous qualitative research as the crucial foundation. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation*, Vol. 19, No. 8, pp. 1087–1096, 2010.

[2] Nick Black. Patient reported outcome measures could help transform healthcare. *BMJ*, Vol. 346, 2013.

[3] 「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.