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)

がん相談支援内容の分析と分類に関する研究

研究分担者 内村祐之 東京医科歯科大学病院医療情報部 (特任講師)

研究要旨

がん相談支援で過去の相談記録情報を利用したテキストマイニング技術により疾患別やカテゴリー別の傾向を分析し可視化する為、2020年度からプロトタイプの開発を進めてきた。2021年度には単語の出現頻度によるビジュアル化や単語を数値ベクトルに変換してその意味を把握する自然言語処理の手法であるWord2Vecを用い、過去の相談内容から類似度の高い順に類似相談内容を抽出する機能の追加とWebアプリケーション化を実現しがん相談支援を行っている施設での実証試験を開始した。2022年度は研究班の意見を反映し、除外ワードに加えて品詞別除外機能を追加し、出力された結果が表形式で出力されるように変更を行い、下位の出現頻度のワードも出力するようにした。また、除外品詞やワードの再調整を行いやすいように品詞も出力するように変更を行ない、引き続き実証を行なった。

A. 研究目的

現在、全国のがん相談支援センターでは、一定の 基準を満たした上で治療の内容や設備、がんに関す る情報提供などを行っており、がん相談支援におけ る相談内容をデータとして管理している。しかし、 蓄積されたデータの利活用や相談員間の知識共有 は十分に出来ていないのが現状である。将来の相談 内容の更なる多様化に備えるためには、蓄積された 相談情報を活用し、FAQの作成や相談員間の情報交 換など他施設を含めた相談員のノウハウを共有す る仕組みを作り、相談員による経験や知識に起因す る対応の差を埋めることにより、一定水準以上の相 談対応が行えるような相談対応支援の仕組みの構 築が必要であると考えられる。仕組みの構築に向け て、がん相談支援で過去に対応した相談記録情報を 利用し、テキストマイニング技術により疾患別やカ テゴリー別の傾向を分析し可視化する。さらに、現 状の相談内容や個々の相談員の対応の傾向を把握 した上で、相談内容と対応内容を分類することで、 FAQの作成支援や他施設を含めた情報共有の仕組 みを検討し、自施設の相談支援業務に活かすことを 目的とする。これまでのWebアプリケーションのプ ロトタイプに更なる改良を加え、実証試験にてシス テムの検証を行う事を目的とする。

B. 研究方法

全国の「がん相談支援センター」の相談支援内容 は施設により、がん相談支援センター「相談のため

の基本形式」(https://ganjoho.jp/med_pro/consulta tion/support/registration sheet.html)の形式を基 として、相談記録の要旨が電子カルテシステムまた はExcelのシート、FileMakerなどの独自開発シス テムなどに相談員が入力したデータとして記録さ れている。この記録を収集し、計量テキスト分析と テキストマイニングにより分析および可視化、内容 による分類を行う。2020年度は10件の相談記録「形 態素解析による分かち書きで単語を集計」「係り受 け解析」「共起ネットワークでビジュアル化」の3 種のプロトタイプをプログラミング言語Pythonに より作成した。2021年度はプロトタイプに改良を 加え、単語の出現頻度によるビジュアル化(Word Cloud)を行い、単語を数値ベクトルに変換してそ の意味を把握する自然言語処理の手法であるWord 2Vecを用い、過去の相談内容から類似度の高い順 に類似相談内容を抽出する機能を実現した。更にプ ロトタイプシステムをWebアプリケーションとし てブラウザのみで動くように改良を行い、がん相談 支援を行っている施設での実証試験を開始した。2 022年度は研究班の意見を反映し、除外ワードに加 えて品詞別除外機能を追加し、出力された結果が表 形式で出力されるように変更を行い、下位の出現頻 度のワードも出力するようにした。また、除外品詞 やワードの再調整を行いやすいように品詞も出力 するように変更を行ない、引き続き実証を行なった。

(倫理面への配慮)

データは全て実データを元にしたサンプルデータを用いた。実証試験では実証試験施設である群馬大学医学部附属病院の倫理審査委員会の承認(テキストマイニングによるがん相談支援内容の分析と分類に関する研究)を経て、倫理規定等を遵守しながら進めている。

C. 研究結果

1. 作成したWebアプリケーション概要と利用手順

Webアプリケーションはプログラミング言語「Python」のFlaskフレームワークで作成し、形態素解析エンジンには「Mecab」を使用した。Webアプリケーションは、ブラウザを開き特定のURLを入力する事で利用が可能となるが、利用の手順は以下の通り。

- 「1.ファイルの選択」で「ファイルを選択」ボタンを押して読み込むファイルを選択する。読み込むファイルには「相談内容」「対応内容」の列を必ず含める。
- 「2. 頻出ワードの取得個数」で頻出ワードを数える個数を入力する。(デフォルト100)
- 「3.除外ワードの設定」で解析から除外するワードをカンマ区切り(,または、)で入力する。
- 「4.類似相談内容の抽出」で抽出キーワードをカンマ区切り(,または、)で入力する。また、類似相談を抽出する対象の文章を入力する。

最後に「解析」ボタンをクリックすると解析処理 が開始される。

「がん相談支援	内容の分析と分類に関する研究について」評価ツール
Version: 1.1	
1. ファイルの選択	
ファイルを選択 群馬大学…プル4	l.cov
「ファイルを選択」ボタンを	リックし、読み込むファイルを選択して下さい。
2. 頻出ワードの取得	個數
20 (8)	
順出ワードの取得個数を入力	TF8U.
3. 除外品詞・ワート	の設定
○名詞 ○動詞 ○形容詞 ☎	詞 □助動詞 □感動詞 □接動詞 □趣等詞 □顧詞 □記号(例句読点、括弧など) □フィラー(例「えーと」、「なんか」など) □その他(例「あ」、「ア」のみな (
その他除外したいワードをカ	マ区切りで入力して下さい。
4. 類似相談内容の抽	出
始出するキーワードをカンマ	切りで入力して下さい。
7/4	
瞬似相談を抽出する対象の文	を入力して下さい。(なるべく長い文章を入力)
5. 解析実行	
9. M-VIX.13	
	薬が表示されるまでそのままお待ち下さい。(数分型かかる場合があります)

2. 分析結果の表示

2-1) がん相談内容、対応内容のビジュアル化 がん相談内容と対応内容に関するデータを元に頻 出語を頻度に比例する大きさで文字を並べたWord Cloudによるビジュアル化を行なった。

① 相談内容のビジュアル化



② 対応内容のビジュアル化



③ 単語の出現回数の可視化(上位、下位)

解析結果	艮	
頻出ワードの取	得個數	效: 20
除外品詞: 助詞	接頭	詞・連帯詞・記号・フィラー・その他
	135.30	(m) (全市 (m)
除外ワード:		
1. 記録内容	容の	頻出ワード分析
【相談内容 」	_位2	0ワード】
- L'	/4- W/r	1
ワード [動詞] いる	件数 14	
[動詞] いる [動詞] あり	13	
「助動詞た	9	
[動詞] し	9	
[名詞] 治療	8	
[助動詞] ない	8	
[名詞] の	7	
[名詞] 相談	7	
[動詞] れ	7	
[名詞] 通院中	6	
[名詞] 電話相談	5	
[動詞] 受け	5	
[名詞] 当院	5	
[形容詞] よい	4	
[助動詞] たら	4	
[名詞] がん患者	\vdash	
	4	
[動詞] 言わ	4	
[名詞] どこ	3	
[助動詞] で	3	

以下の下位ワード表示は2022年度に追加した新機能である。

ワード	件数	
[助動詞] で	3	
[名詞] どこ	3	
[動詞] 言わ	4	
[名詞] 他院	4	
[名詞] がん患者	4	
[助動詞] たら	4	
[形容詞] よい	4	
[名詞] 当院	5	
[動詞] 受け	5	
[名詞] 電話相談	5	
[名詞] 通院中	6	
[動詞] れ	7	
[名詞] 相談	7	
[名詞] の	7	
[助動詞] ない	8	
[名詞] 治療	8	
[動詞] し	9	
[助動詞] た	9	
[動詞] あり	13	
[動詞] いる	14	

2-2) Word2vecによる類似文章検索

単語を数値ベクトルに変換し意味を把握する自然言語処理の手法であるWord2Vecを用い、過去の相談内容から類似度の高い順に類似相談内容を抽出する機能を実現した。抽出結果を以下に示す。



D. 考察

これまでがん相談に関しては、過去の相談データを解析するなどのデータ利活用はあまり進んでいなかった。今回作成したWordCloudによるビジュアル化と形態素解析による分かち書きによる単語集計により傾向を分析し可視化することで相談内容の重要なキーワードを直観的に把握することができ、経時的にトレンドを追う事で自施設の相談支援業務に活かす事が可能になると考える。また、W

ord2vecによる類似文章検索では、相談内容の文章 を入力する事で、過去の相談から類似度の高い順に 内容を一覧で抽出する事が出来、キーワードのハイ ライトを合わせる事で、類似の相談内容に対してど のような対応を行ったのかを把握する事も出来る ので、相談内容と対応内容を分類することで、FAQ の作成支援につなげることができる事が示唆され た。2022年度は研究班や実証施設の意見を反映し、 システムの機能を追加したが、これまで相談内容や 対応内容の頻出ワードを上位から表示するのみで あったが、頻度が少ない=新しいトピックであり、 情報が整備されていない可能性があるので頻度の 低い単語にも注目する必要があることから頻度の 少ない情報から重要なワードをキャッチしやすい ように出力結果を変更した。また、分析の際に登録 する除外ワードはこれまで個別に入力を行なって いたが、品詞別に一括で登録できるように修正を行 い、今後、多施設共同で解析を行う際に、共通認識 の基で作業を進められるような工夫を行なった。

E. 結論

2021年度はプロトタイプを改良し、単語の出 現頻度によるビジュアル化を行い、単語を数値べ クトルに変換してその意味を把握する自然言語処 理の手法である Word2Vec を用い、過去の相談内 容から類似度の高い順に類似相談内容を抽出する 機能を実現した。更にプロトタイプシステムを Web アプリケーションとしてブラウザのみで動く ように改良を行い、がん相談支援を行っている施 設での実証試験を開始した。2022年度は研究班 の意見を反映し、除外ワードに加えて品詞別除外 機能を追加し、出力された結果が表形式で出力さ れるように変更を行い、下位の出現頻度のワード も出力するようにした。また、除外品詞やワード の再調整を行いやすいように品詞も出力するよう に変更を行ない、引き続き実証を行なった。今後 は実際のがん相談支援記録を用いてテキストマイ ニングを行う事で、相談内容および対応内容の分 類や FAQ の作成支援へと発展させる予定であ る。

F. 健康危険情報 特になし

G. 研究発表

(発表誌名巻号・頁・発行年等も記入)

1. 論文発表 なし

2. 学会発表

なし

H.知的財産権の出願・登録状況

(予定を含む)

1.特許取得

なし

2.実用新案登録

なし

3.その他

なし