

うつ病患者の周辺看護者に向けた構造化検索手法

齋藤 俊英¹ 荒牧 英治² 宮部 真衣³ 平田 圭二¹

¹ 公立ほこだて未来大学 ² 奈良先端科学技術大学院大学 ³ 和歌山大学

1 はじめに

近年、闘病ブログや SNS から、患者自身が主体的に情報を得る機会が増えつつある。しかし一方で、患者を気にかける人々を対象とした情報は乏しい。山下らは、うつ病患者の看護経験がある 15 名の家族看護者にインタビューを行い、コミュニケーション環境の改善や精神的および社会的ストレスを軽減するために有効と思われる技術に関する示唆として、病状変化への適用に関する支援、共有する場の提供、探究する場の提供を挙げた [3]。本研究の目的は、探究する場における探求の支援である。家族、恋人、友人や同僚といった患者を気にかける人々(以降、周辺看護者)が知りたいと考えるのは、病気に関する知識だけではなく、自分と似たような状況に置かれた他の周辺看護者についての情報である。しかし、一般的な情報検索手段である Web でのキーワード検索では、期待しない不要なページを多く含んでしまう。そこで本研究では、キーワードの一致では検索できない意味的な情報の検索を目指し、悩みの文脈と視点・状況・志向という構造による周辺看護者のための類似文章の検索を実現する。

2 機械学習による分類とタグ付け

時間や関連人物のそれぞれの立ち位置、関係性、背景といった意味的な項目を全て学習するためには膨大なデータが必要となるため、1つ1つの項目を検索のインデックスとすることは現実的ではない。そこで悩みの文章の文脈の分類と、タグ付けによる構造抽出の2つのアプローチにより類似文章の検索を目指す。文脈の分類では、分類問題を解く代表的手法の Support Vector Machine (SVM) を用い、文脈の類似した候補を得られる。タグ付けによる構造抽出では、文脈で表現しきれない詳しい情報を系列ラベリング問題を解く代表的手法である Conditional Random Field (CRF) を用い、視点・状況・志向という構造という形で抽出する。この構造のそれぞれの項目は「誰が」「どうしたのか/どうであったのか」「誰/何に対して」を表している。

2.1 取得したブログ記事の文脈

構造が一致していても、文脈によって意味が異なる場合がある。それは、後述の CRF によって抽出される「夫」や「妻」が「患者」という情報に抽象化されるためである。そこで文脈を一致させるため、SVM による判別を行う。カテゴリは 100 件の文章を分類し、その使用頻度をもとに設定した。頻度の多い「患者が夫」カテゴリには条件を付け加え分割し (X1, X2) とし、頻度

表 1 カテゴリの種類

番号	カテゴリが表す文脈
X1	患者が夫、かつ子供なし、または不明
X2	患者が夫、かつ子供あり
X3	患者が妻
X4	患者が親 (配偶者の親を含む)
X5	患者が兄弟もしくは姉妹
X6	患者がその他の人物

の少ない「患者が両親」と「患者が義両親」、「患者が兄、姉」と「患者が弟、妹」のカテゴリそれぞれを統合することにより (X4, X5) を得た (表 1)。

2.2 取得したブログ記事の構造

表 2 タグの種類

タグ	文章中の対象となる箇所
p	患者自身の情報
a	筆者自身の情報
o	その他の人物の情報
ap	筆者から患者への行動や心情
pa	患者から筆者への行動や心情
do	筆者・患者からその他の人物への行動や心情
od	その他の人物から筆者・患者への行動や心情

構造を表現する基本的なタグとして、周辺看護者の文章という前提から文中に存在が保証される筆者 (author) を表す <a>, 患者 (patient) の <p>, その他の人物 (other) の <o> を用い、全ての人物を表現した。さらに、それらを組み合わせることで関係性を <ap>, <pa>, <do>, <od> のタグで表現する (表 2)。学習効率の向上を図り、<po>, <ao>, <op>, <oa> であるタグを統合した <do, <od> を導入した。文章から抽出される構造をタグで表現すると図. 1 となる。

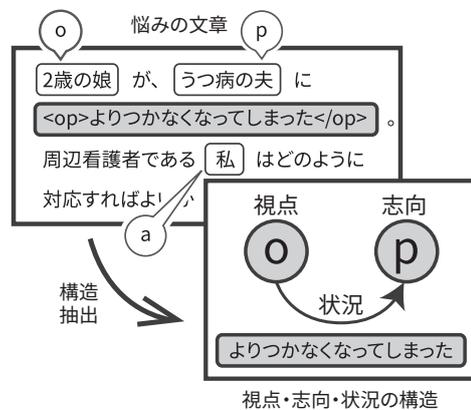


図. 1 タグ付けによる構造抽出の例

Structural Retrieval Method with Aim of Supporting Caregivers of Depressed Family Member

¹ Toshihide Saito, Keiji Hirata, Future University Hakodate

² Eiji Aramaki, NARA Institute of Science and Technology

³ Mai Miyabe, Wakayama University

表3 SVMにおける10分割交差検定の結果

	システムの出力と正解の一致数 / 正解の分類数										AVE
	A	B	C	D	E	F	G	H	I	J	
X1	4/9	5/8	4/8	5/8	2/8	5/8	1/8	4/8	5/8	4/8	3.9/8.2
X2	5/7	5/7	2/7	3/7	2/7	4/7	1/7	4/7	3/8	4/8	3.3/7.4
X3	7/8	4/8	6/8	5/8	4/7	6/7	4/7	4/7	4/7	6/7	5.0/7.7
X4	8/10	9/10	8/10	9/11	8/11	9/11	8/11	8/11	10/11	8/11	8.5/11.1
X5	2/7	3/7	3/7	3/7	2/7	4/7	3/7	2/7	5/7	1/7	2.8/7.5
X6	0/5	2/5	2/5	1/5	1/5	1/5	2/5	0/5	1/5	1/5	1.1/5.6
SUM	26/46	28/45	25/45	26/46	19/45	29/45	19/45	22/45	28/46	24/46	24.6/45.4

適合率 54 % (= 24.6 / 45.4)

表4 CRFにおける10分割交差検定の結果

	A	B	C	D	E	F	G	H	I	J	Average
GOAL	1181	1501	1472	1357	1198	1389	1282	1433	1169	1374	1335.6
SYS	666	773	868	737	660	726	728	797	594	787	733.6
MATCH	322	324	370	337	312	329	352	334	277	367	332.4

適合率 45.3 % (= 332.4/733.6), 再現率 24.9 % (= 332.4/1335.6), F 値 0.321

3 実験と評価

3.1 正解データ

実験に用いる文章は、2014年8月26日時点でOK-Wave [4] の夫婦家族カテゴリに存在した投稿3577件中1000件を取得した。正解データは、1000件中周辺看護者の投稿であると判断した449件(平均文字数1656字)であり、分類・タグ付けを行い作成した。

3.2 分類の評価

文書への前処理として形態素解析を行い、その結果を二値ベクトルとして表現した。分類では、オープンソースのSVMソフトウェアであるTinySVM [1] を使用し、2次の多項式カーネルを用いた。分類する上では、one-versus-rest法によりカテゴリを選択した。

学習データに対する10分割交差検定の結果は表3となった。10分割にしたそれぞれのデータのまとまりはA~Jとした。X4の値が最も高く、X6を除くと、X5が最も低い値となった。いずれの値もランダムで分類する確率1/6 (0.167)を上回る結果となった。

3.3 タグ付けの評価

文書への前処理として形態素解析を行い、その結果を系列とし、またそれぞれの素性に対するIOB2タグ [2] をラベル列として与え、学習を行った。正解を判定する上で、IOB2タグを用い特定した範囲の重なりを利用した判定基準を設定した。

学習データに対する10分割交差検定の結果は表4の通りである。SVMの時と同様で、10分割にしたそれぞれのデータのまとまりはA~Jとしている。

3.4 考察

分類では、「患者が両親」と「患者が義両親」が統合されたX4が最も高い値となった。これは「義母」が文中では「お母さん」と表記されるなど、文章が表記的に非常に似ているためである。一方、同様に「患者が兄、姉」と「患者が弟、妹」が統合されたX5の値はそれほど高くない。これは「兄、姉」と「弟、妹」が表記として異なることが原因として考えられる。

タグ付けの評価において、今回は品詞によらずタグ付けを行ったが、一般に範囲が長くなるほど期待するタグ付けを行うのは難しい。そのため、範囲が長くなる傾向の動詞などは対象とせず、名詞にのみを対象とすることで精度の向上が見込まれる。

4 おわりに

本研究では、周辺看護者の支援に向けた類似文章の検索を目指し、検索に用いるSVMとCRFの学習データに対しての10分割交差検定を行った。文脈の類似した文章をある程度判別することが出来たが、詳細な情報を抽出する上では、特に再現率の向上が課題である。

今後は、それぞれの精度向上に加え、検索クエリに対して類似文章の候補を出力するシステムを実装し、検索手法についての評価を行う。

謝辞

本研究はJSPS科研費26280116の助成を受けたものです。

参考文献

- [1] Kudo, T.: TinySVM: Support Vector Machines, ChaSen.org (online), available from <<http://chasen.org/~taku/software/TinySVM/>> (accessed 2016-01-06).
- [2] Tjong Kim Sang, E. F. and Veenstra, J.: Representing Text Chunks, *Proc. 9th Conf. European chapter of the Association for Computational Linguistics (EACL 1999)*, pp.173-179 (1999).
- [3] Yamashita, N., Kuzuoka, H., Hirata, K. and Kudo, T.: Understanding the conflicting demands of family caregivers caring for depressed family members, *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI 2013)*, pp.2637-2646 (2013).
- [4] 株式会社オウケイウェイヴ: 質問・疑問に答えるQ&AサイトOKWAVE (オンライン), 入手先 <<http://okwave.jp>> (参照 2016-01-06).