患者に学ぶリスク発見

Risk discovery learned from patients

山口 広樹[1]　大澤 幸生[1]

Hiroki Yamaguchi[1], Yukio Ohsawa[1]

[1]東京大学大学院工学系研究科

[1]School of Engineering, Univ. of Tokyo

Abstract: Generally speaking, patients tend to be remarkably sensitive to the surrounding situations. Patients pay attention to the very details of the circumstances, and who comes to be aware of hidden risks which may harm their living condition. In this paper, we present a method of risk discovery by analyzing textual data on the conversations of patients and medial experts. Utterances rejected for the reason that they sound trivial or apparently meaningless are extracted, based on our original criteria of the utterances' significance. In contract to experts' discounting, the utterances really are addressed to latent and significant risks. The experiment shows the potential utility of this method.

1. はじめに

科学技術が飛躍的に発展してきたとはいえ，全ての問題を科学で解決するのは不可能である．特に，人の感情を計算機で扱うことは非常に困難であり，実現可能性が低いと言える．本論ではこうした不確定性の高いものへのアプローチとして，計算機のみで答えを出すのではなく，大澤らが開発した可視化ツールであるKeyGraph[1]を使用したチャンス発見のプロセス「二重らせんモデル」[2]に倣って，計算機の能力を利用して人が答えを導き出す支援をする．そのためにはリスクに対して敏感な人物からその能力を引き出すのが近道である．

本論では，リスクに敏感な人物として患者を対象にする．自らの病気を治そうと考えない患者はいないため，患者は自分の周囲の環境への感度が非常に高くなっているという傾向がある．「病気になったことが生活習慣を改善するきっかけとなった」という経験は多くの人にあるであろう．本研究の目的は，患者の磨かれた鋭い感覚を利用して潜在的なリスクを発見することである．

上述の目的を達成するために，患者の座談会の会話内容から，テキストマイニング技術を利用してリスクを発見していく．ここでは，発言の真意が周囲に理解されずに排除されてしまっている例外発言を抽出するという手法を提案する．

テキストマイニング技術の従来研究としては，用いられている単語群をベクトルで表して処理するベクトル空間法[3]など多くの研究がなされており，情報検索等に広く利用されている．ベクトル空間法はシンプルであるが故に汎用性が高いと考えられ，構造が無秩序である会話を扱うために本研究ではこれを採用した．従来のテキストマイニング技術には，文書の要約[4][5]や構造解析[6][7]を目的としたものが多い．また重要と思われるキーワードを抽出する手法として，繰り返し言及される語は重要な概念を表すと仮定して出現頻度の高い語をキーワードにする方法[8]や活性伝搬法を利用した方法[9]などがある．また，例外に注目した従来研究としては，情報量を用いてデータベースから例外的知識を抽出するもの[10]や例外から得られる情報を基にプログラムの解析やテストをするもの[11]などがある．これらの従来手法ではデータベースやプログラムなどの型にはまったデータを対象としており，本研究で解析対象とする無秩序な構造を持つ自然言語を解析することは不可能であると考えられる．

そこで本研究では，これらの従来研究とは異なる観点からのアプローチをとった．従来手法の多くは，データとして残っているテキスト，いわばすでになんらかの効果を既に発揮している情報を抽出するもので，文書の要約や会話構造の解析には有効である．本研究では，発言として直接的に現れてはいないがその裏には何ら

かの想いが潜んでいると考えられるような発言から，潜在的なリスクを発見する手法を提案する．実験では，患者座談会を解析対象として扱う．この座談会は患者と医者とインタビュアーによるもので，本来重要視すべき患者の発言を医者とインタビュアーが排除してしまっていることがあることを解析によって発見することができた．

本論では，発言から想起されるリスク（発言内リスク）と，コミュニケーション自体が内包する人間関係におけるリスクについて患者座談会の解析結果から論じる．また，従来手法との比較として，TFIDF 法[12]と IDM 法[9]で同データを解析した．

2. 提案システム

図 1 に提案システムの概要を示す．本システムは計算機によるテキスト処理を核とした部分と人による会話とを含み，この二つを繰り返し行うプロセスによって成り立っている．

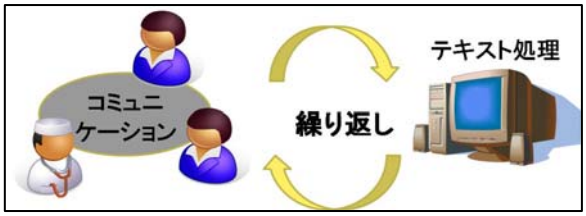


図 1 提案システム概要

3. テキスト処理

提案システムの核となるテキスト処理部のフローチャートを図 2 に示す．発言ベクトルを基に，類似度から例外を抽出する．抽出した例外発言の評価を選択平均情報量という情報量を定義して行い，例外発言にランク付して出力する．処理の詳細については口述する．

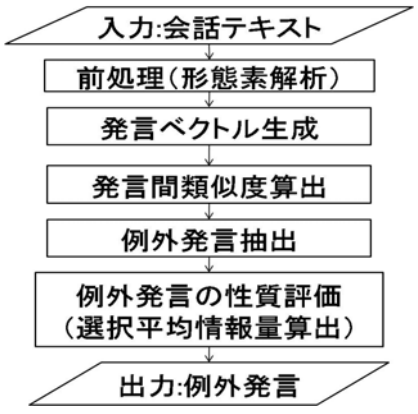


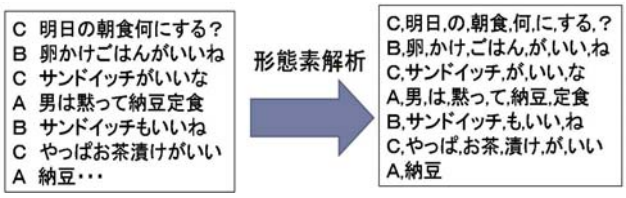
図 2 テキスト処理フローチャート

3.1. 入力：会話テキスト

本システムへの入力データとなる会話テキストは，発言者名と発言内容が一対となっているものを一発言とし，連続する複数の発言が集まったものを会話テキストとして用いた．会話テキストの例は次項 3.2. の図 3(a)に示す．

3.2. 前処理（形態素解析）

入力された会話テキストを前処理として，形態素解析して，各発言を構成する単語に分けて発言を単語の集合としてあつかえるようにする．形態素解析の例を図 3 に示す．形態素解析器には茶筅[13]を用いた．



(a)入力:会話テキスト　　(b)形態素解析後

図 3 形態素解析の例

3.3. 発言ベクトル生成

前処理によって得られた結果を利用して(1)式に示すように発言ベクトルを生成する．これによって自然言語を数値として扱えるようになる．

$$\mathbf{U}_k = (t_0, t_1, \Lambda\ , t_n), \left|\mathbf{U}_k\right| = 1 \qquad (1)$$

但し，t_k は k番目の単語の出現頻度，nは総単語種類数である．また，後の処理で扱いやすくするために正規化を行っている．

3.4. 発言間類似度算出

発言ベクトルを基に各発言とその前後の発言群（コンテクスト）との類似度を(2)式に示すように定義する．ここでは，before 類似度(2.a)と after 類似度(2.b)の二つの類似度を定義した．

$$S_{before \cdot k} = \mathbf{W}^T_{before \cdot k} \cdot \mathbf{U}_k \qquad (2.a)$$

$$S_{after \cdot k} = \mathbf{U}_k \cdot \mathbf{W}^T_{after \cdot k} \qquad (2.b)$$

$$\text{但し,} \left(\begin{array}{l} \mathbf{W}_{before \cdot k} = \sum_{i<n} \mathbf{U}_i, \left|\mathbf{W}_{before \cdot k}\right| = 1 \\ \mathbf{W}_{after \cdot k} = \sum_{i>k} \mathbf{U}_i, \left|\mathbf{W}_{after \cdot k}\right| = 1 \end{array} \right)$$

前記の類似度は，ベクトルの内積を用いるためコサイン類似度と呼ばれ，文書の内容比較にしばしば用いられる手法である．ここでは，事前に正規化してあるため，一般的なコサイン類似度とは一部異なっている．また，発言と発言群の類似度が 0 から 1 の範囲の値をとるように発言群のベクトルも正規化を行った後に内積をとっている．

3.5. 例外発言抽出

二つの類似度に閾値を設けることによって，例外発言を抽出する．例外発言は周囲から排除されている発言であるため，閾値よりも小さい値を持ち，前後の会話と内容が食い違っていると考えられる発言を例外発言として抽出した（(3)式）．なお，会話毎に妥当な閾値を定められるように閾値はユーザーが決定できるように設計した．

$$S_{before \cdot k} < \alpha \quad \mathrm{I} \quad S_{after \cdot k} < \beta \qquad (3)$$

（α，β は 0 から 1 の範囲の任意の値）

3.6. 例外発言の性質評価

(4.a)，(4.b)式によって定義する情報量を，選択平均情報量とし，抽出した例外発言の評価として利用した．

$$E_{\mathbf{W}_{before \cdot k}\mathbf{U}_k} = \sum_{t_i \in \mathbf{U}_k} p_{\mathbf{U}_k t_i} I_{\mathbf{W}_{before \cdot k} t_i} \quad (4.a)$$

$$E_{\mathbf{W}_{after \cdot k}\mathbf{U}_k} = \sum_{t_i \in \mathbf{U}_k} p_{\mathbf{U}_k t_i} I_{\mathbf{W}_{before \cdot k} t_i} \quad (4.b)$$

$$\text{但し,} \left(\begin{array}{l} \mathbf{W}_{before \cdot k} = \sum_{i \le n} \mathbf{U}_i \\ \mathbf{W}_{after \cdot k} = \sum_{i \ge k} \mathbf{U}_i \end{array} \right)$$

上式で定義した情報量を各発言に対して算出すると，二つの情報量は表 1 に示す性質を持つ．

表 1　選択平均情報量の性質

Before 値	大	以前の会話とは異なる内容
	小	以前の会話と似通った内容
After 値	大	以後の会話とは異なる内容
	小	以後の会話と似通った内容

また，定義した情報量を使って例外発言の性質を評価するため，(5)式のように評価関数を定義した．

$$\overline{E} = E_{\mathbf{W}_{before \cdot k}\mathbf{U}_k} - E_{\mathbf{W}_{after \cdot k}\mathbf{U}_k} \qquad (5)$$

(5)式の評価関数によって得られた評価値は，次表に示す特徴を持つ．

表 2　評価値の特徴

評価値 大 (before 大且 after 小)	新たな話題を導入している発言
評価値 小 (before 小且 after 大)	会話に特に影響を与えていない発言

表 2 に示した特徴から，評価値が大きい発言ほど，前後の会話コンテクストと食い違っている例外発言でありながら会話に影響を与えた発言であるといえる．こうした発言は，発言の裏側に何らかの意図や知識（暗黙知や経験知）が潜んでいる発言であると考えられる．こうした知識を引き出すことによって潜在的リスクを発見することが本研究の目的である．

3. 7. 出力：例外発言

抽出した例外発言を(5)式の評価値でランク付して表 3 に示す項目を出力する．

表 3　出力項目

Rank	発言 No	発言者	発言内容	Score

ユーザーは上記の出力結果を確認しながらパラメータの調整をすることで，より妥当な結果を追究することができる．

4. 患者座談会の解析

4.1 患者座談会データ概要

解析対象に用いた患者座談会の会話ログデータは，全 362 発言でおよそ 2 時間程度のものであった．参加者は，ペースメーカー植え込み手術経験のあり説教的に患者会に参加している患者と医師とインタビュアーの 3 人で行われた．主にインタビュー形式で行われたものであり，手術前から手術後の患者の生活に関する内容で，医療技術や患者会に関することやブログや日常生活についてなど多岐にわたるものである．また，インタビューに応じた患者の背景は，マラソン大会に積極的に参加するようなスポーツを好む人物であり，患者会に参加していて，かつ

技術者として働いている人物である.

4.2 抽出した例外発言

パラメータの設定は,窓 w を 10,閾値はα=0.2,β=0.25 として，27 発言が抽出された．抽出した発言の一部を表 4 に示す．ここで，インタビュアーの発言は話を進行するためのものが多く本研究の目的に沿うものではないと判断したため除外した.

表 4　例外発言抽出結果（一部抜粋）

rank	No.	Score	name	Contents
1	187	1.15	患者	「なんで」という感じは言葉としてはそれしかないですね。「なんで」という感じ
2	331	1.01	患者	この話はブログには一切書いていません。
3	87	0.98	患者	退院はそれからあと 1 週間です。
4	336	0.88	医師	誰も分からないです。
19	332	0.87	患者	ICD は誤作動があるから本当に大変だと思います。
23	328	0.32	患者	大したことは書いていなくても裸のところが分かりますね。

4.3 実験結果の考察

本実験で発見したリスクは，大きく分けて 2 種類に分けられる．一つは，発言に含まれる内容から想起して発見できるリスク i)であり，もう一つは，コミュニケーションそれ自体に潜む,対話における人間関係のリスク ii)である.以下に得られた知見を挙げる.

i)発言に潜むリスク

- 発言 No.187 から,「なんで自分が病気に？」という患者であれば誰しもが持っているであろう疑問を言及している．ここには発言者固有の生活背景があり，他にも発言したいことがあったのではないかと想像できる.
- 発言 No.332 から,ICD の改善につながるような知識・要求を得られる可能性がある.
- 発言 No.328 は患者の体験記に関する発言であり，役立つ情報であるにも関わらず，病院には置いていないという問題を指摘した発言であった.

ii)コミュニケーションに含まれるリスク

抽出された例外発言の発言者に注目してみると,全 27 発言中の 15 発言が患者の発言であり，過半数を占めている.

本データは患者へのインタビューが主な目的であって患者の発言が最重要であるにも関わらず，こうした結果が出てきてしまった．全くといっても過言ではないほど無視されてしまった患者発言もあったのが事実である.

ここで発見したリスクとして,「本来重要視されるべき人の発言を排除してしまっていることが多々ある」といえる．本座談会では，医療に関する知識の差から「患者＜医者」という関係が暗黙に作られていたと考えられる.

また本実験では,提案システムの繰り返しプロセスを経ることはできていない．解析結果を本人に提示して再度インタビューを行うことによってより深い理解が可能であり，本システムを利用する上では欠かすことのできないプロセスとなる.

4.4 他手法との比較

本手法の特徴と有効性を明らかにするために,従来手法と解析結果を比較する．本手法は発言を抽出するものであるが，ここではキーワード抽出法である TFIDF 法[12]と IDM 法[9]を用いて比較実験を行った.2 つの手法で抽出したキーワードを表 5 に示す．ここで,TFIDF の計算には適当なコーパスが手に入らなかったため，解析用データを基に一発言を一ドキュメントとして DF を計算し，(6)式を使用して求めた．ここで,$TF(t)$ は単語の出現頻度であり，$DF(t)$ は出現ドキュメント数である.

$$TFIDF(t) = TF(t) \times \log \frac{1}{DF(t)} \qquad (3)$$

表 5 に TFIDF 法と IDM 法によって解析した結果を示す．表にはそれぞれキーワードを 5 単語載せた．一つの発言に含まれるキーワードのスコアから発言のスコアを設定することもできるが,妥当な結果を得ることはできなかったため，キーワード抽出結果を比較対象とする.

表 5 TFIDF 法,IDM 法による解析結果

	TFIDF	IDM
1	そういう	ペースメーカー
2	自分	自分
3	ペースメーカー	アブレーション
4	病院	病院
5	いろいろ	入れ

前述の通り,本手法は発言を抽出する手法であるのに対して,TFIDF 法と IDM 法はキーワード抽出法である．また，これらの手法は，文書の要約や文書における著者（発言者）の意図を抽出する手法であるため，本手法とは全く異なる結果が得られた．

TFIDF 法では，会話の中心的なトピックがわかるようなキーワードが抽出された．しかし，適当なコーパスを用いることができなかったためか,IDM と似通った結果にはなったものの建設的な知見をえることはできなかった．

IDM 法は，後の会話に影響を与えた単語をキーワードとして抽出する手法であり,TFIDF 法では抽出されなかった「アブレーション」という単語が抽出された．このキーワードは，会話の中心的なトピックではないが，この単語をきっかけに会話が活性化した（話が膨らんだ）と考えられる．

これらの結果が示すように，TFIDF 法では会話の中心的なトピックの理解に役立ち,IDM 法ではその後の会話に大きく影響を与えたキーワードから会話の話題構造を知る手がかりをえることができることが分かった．これに対して本手法では，会話の要旨や話題構造を知る手がかりを得ることはできない．しかし，特に掘り下げられなかった発言に潜む真意を探ることができる．発言者は抽出した発言の後にも続けて主張したいことがあったのではないかと考えられ，本システムのプロセスを経ることによって発言者の感情をより深く理解することができる．

5. おわりに

コンピュータから正確な結果出力を得られるようになった昨今では，例外的な事象は排除され，その価値が葬られてしまう傾向にある．特に，会話における例外的な発言は軽んじられることが多々あり，その発言の裏に潜む主張や感情などが埋もれてしまっている．

従来のテキストマイニング技術では，文書の要約や構造解析をすることが主目的となっており，解析結果をいかにして有効利用していくかという方法を深く追究しているものは僅かである．

対して本研究は，計算機では非常に扱いにくいとされる感情をより深く掘り下げていくことができるシステムであり，通常見落とされがちな主張に焦点をあてることで今まで取り入れられることのなかった概念を導入することができるようになる．

本論文では患者座談会を解析対象として医療におけるリスクを発見した．人の生活にコミュニケーションを欠かすことはできず，商品開発における要求獲得のためのヒアリングや国会や県議会での会議，教師から生徒への指導のためのミーティングなどから，日常会話まで，非常に多くの場面に本研究を適用できる可能性がある．今後の展望として，実験的に様々な場面に本手法を適用して有効性を検証していく．

また本論で解析した患者座談会データに関して，参加者本人に対して解析結果をフィードバックして，再度インタビューを行うといった実験を今後行い，さらに解析を進める予定である．

謝辞

患者座談会のデータを提供していただき，さらに数々の貴重な助言を頂いた東京大学医科学研究所助教の田中祐次先生，座談会に参加していただいた野口氏，広多氏，インタビュアーの北澤氏に感謝申し上げます．また，貴重なプログラムを提供くださった大阪大学大学院経営学研究科准教授の松村真宏先生に感謝申し上げます．最後に，研究を進めるにあたって議論していただいた大澤研究室助教の西原陽子先生並びに研究室の学生の方々に感謝致します．

参考文献

[1] 大澤幸生，N. E. Benson，谷内田正彦，KeyGraph：語の共起グラフの分割・統合によるキーワード抽出，電子情報通信学会論文誌 J82-D-1, No.2, pp. 391–400 (1999).

[2] **大澤幸生 監修，チャンス発見の情報技術，東京電機大学出版局**, 2003.

[3] G. Salton, A. Wong, and C. S. Yang, A Vector Space Model for Automatic Indexing, Communication of the ACM, Vol.18, No.11, pp613-620, 1975.

[4] R. Barzilay, M. Elhadad, Using lexical chains for

text summarization, Advances in Automatic Text Summarization, pp.1-12, The MIT Press, London, 1999.

[5] 相良直樹, 砂山渡, 谷内田正彦, サブピックを考慮した重要文抽出による報知的要約生成, 電子情報通信学会論文誌 Vol.J90-D No.2 pp.427-440,2007.

[6] 住田一男, 知野哲郎, 小野顕司, 三池誠司, 文書構造解析に基づく自動抄録生成と検索提示機能としての評価, 電子情報通信学会論文誌 D Vol.J78-D2, No.3, pp.511-519, 1995.

[7] 横山憲司, 難波英嗣, 奥村学, Support Vector Machineを用いた談話構造解析. 情報処理学会自然言語処理研究会NL-155, pp. 193–200, 2003.

[8] H. P. Luhn, A Statistical Approach to the Mechanized Encoding and Searching of Literary Information, IBM, Journal of Research and Development, Vol.1, No.4, pp.709-317, 1957.

[9] 松村真宏, 大澤幸生, 石塚満, 語の活性度に基づくキーワード抽出法,人工知能学会論文誌 17 巻 4 号 F, 2002 年.

[10] 鈴木英之進, 志村正道, 情報理論的手法を用いたデータベースからの例外的知識の発見, 人工知能学会誌, Vol.12,No.2, pp.305-312, 1997.

[11] Saurabh Sinha, Mary J. Harrold, Analysis and Testing of Programs with Exception Handling Constructs, IEEE TRANSSACTION ON SOFTWARE, ENGINEERING, Vol.26 No.9, pp.849-871, 2000.

[12] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[13] 形態素解析ソフト『茶筅』 http://chasen.naist.jp/hiki/ChaSen/ .