１Ｇ０１

テキストマイニングによる医療分野の課題及びその解決のための制度体系等に関する有用知識の抽出

○内海和夫（東京工業大学），乾孝司（筑波大学），橋本泰一（東京工業大学）
村上浩司（奈良先端科学技術大学院大学），石川正道（東京工業大学）

1. はじめに

日々膨大なテキスト情報が公開される現代社会にあって，大量の電子テキストから様々な社会課題や，その課題を解決するための科学技術，さらにそれらを規制する社会制度を，課題毎に構造化して分析する手法を構築することは，合理的な科学技術の研究開発推進や政策立案のために取り組むに値する方法論的課題である．我々は，共語分析の適用範囲を拡大することを目的に，すでに社会課題と科学技術や制度との関係情報を豊富に含む新聞記事を対象として，俯瞰的アプローチによる社会課題と技術的対策に関する用語の自動抽出と，共語分析による用語の視覚化を行うテキストマイニング技術の開発を行い，その有用性を報告した[1),2)]．これらの分析を通じて，社会課題の解決において，制度的側面は課題の構造に対応して社会が対処すべき指針を示す役割を担っていることが分かった．また，社会制度は適用対象の多様化や科学技術の進展によって，その内容も見直される．このような社会課題と科学技術及び制度の関係をテキスト情報からとらえる試みは，例えば，行政が発信する情報から制度や政策の傾向を見ようとする研究例や[3)]，特定のキーワードを研究者が設定し，新聞記事情報から当該キーワードと関連する科学技術用語や社会・制度関連用語との関係をとらえようとする研究例[4)]などがあげられるが，事例に乏しい．

我々はこのような背景を踏まえて，制度関連情報を含む大量のテキストから特定のキーワードを設定しない俯瞰的アプローチによって，個別の社会課題に対応する制度情報を抽出する手法を確立することを目的とした．特に，医療分野における課題と制度体系等の関係性に注目して，制度情報を抽出することを試みた．本発表では，医療分野の制度関連情報を多く含む記事クラスタを自動形成することによって，制度関連用語を含む特徴的な言語パターンを見出し，それを用いて自動抽出された制度関連用語の共語分析により，用語間の関係性を視覚化した結果について報告する．

2. 制度関連用語の抽出

2.1 記事クラスタの作成

新聞記事情報から社会課題の抽出を行うために，まず俯瞰的アプローチにより医療分野のトピック別の記事クラスタを形成する．本研究では，日本経済新聞記事データベースを用いて 2005 年の日本経済新聞本紙から医療分野の記事群を作成し，その記事群に対するクラスタリングにより 200 個の記事クラスタを形成した[1)]．さらに，各クラスタに付与されている要約キーワードを参照し，「医療制度改革」に関するトピックを多く含むクラスタ（記事数：372，以下「医療制度改革」クラスタと呼ぶ）を選定し分析対象とした．

2.2 専門家による基準用語の抽出

まず「医療制度改革」クラスタより，専門家による制度関連用語の抽出を行い，自動抽出された用語との比較に用いる参照データを作成する．抽出にあたっては，原則として 1)制度を実施する主体を表す語は抽出しない，2)年次や比率等を表す数字や固有名詞を含む語は抽出しない，3) ２語以上がまとまって１つの意味をなしている用語はまとめて抽出する，といった基準を設けた．

2.3 制度関連用語を自動抽出するための指標の算出

抽出しようとする制度関連用語は，次の要件を満たすことが望ましい．

①社会課題と強い関係性を有する
②制度と強い関係性を有する

そこで，①，②に対してそれぞれ課題関連度と制度関連度という指標を導入し，これらの指標の積算

値にしたがって用語候補を順位付け，順序の上位に位置する用語候補を制度関連用語として抽出する．

2.3.1 課題関連度

社会課題と関連する特定のトピックで特徴付けられるクラスタから，課題関連度の強い用語を抽出する処理は，クラスタから特徴的な用語を抽出するクラスタ・ラベリングとほぼ等価な処理であると考えられる．本研究では，クラスタ・ラベリングで適用される基本的な指標であるカイ 2 乗値を，課題関連度を測る指標として採用した．

2.3.2 制度関連度

(a) 言語パターンマッチングによる制度表現文の抽出

新聞記事では，制度的な内容を含む文は特徴的なパターンで表現されることが多い．例えば，「医療制度改革」クラスタでは，共通的な名詞表現として「～制度」，「～制」，「～法」，「～案」，「～策」といった用語が高頻度で見られ，これらに加えて「方針」，「指針」，「基準」，「改革」，「仕組み」などの名詞がよく用いられる．また，これらの名詞表現に対応する動詞としては，「検討する」，「決定する」，「導入する」，「見直す」などがよく用いられ，例えば「･･･方針を･･･固める」や「･･･改革を･･･進める」は固定的なパターンとして使われている．このような制度的な内容を表現する特徴的な名詞と動詞のパターンが含まれている文（以下「制度表現文」と呼ぶ）を「医療制度改革」クラスタから抽出したところ，316 の制度表現文が抽出された．これらの制度表現文のなかで，高頻度で用いられる言語パターンを表１に示す．

表１　制度表現文に高頻度で出現する言語パターン（上位 10 パターン）

名詞	動詞	頻度	名詞	動詞	頻度
仕組み	導入する	12	改革	検討する	9
方針	固める	12	関連法案	提出する	9
改革	進める	11	医療制度改革試案	盛り込む	8
改革	議論する	10	見直し	検討する	8
改革案	まとめる	10	方針	決める	8

(b) 制度関連度の算出

専門家抽出された制度関連用語の位置及び分布状況から，技術的対策用語の抽出のときに用いた技術関連度と同様の手法[1)]が適用できると判断し，制度表現文内での用語の位置に基づく指標と，用語を含む文と制度表現文の間の距離に基づく指標の積により，課題関連度を算出することとした．前者の指標は，上述の言語パターンの名詞，及びそれを修飾する名詞のスコアが他と比べて 100 倍になるような重みとする．なお「100 倍」の値は，主に課題関連度の値の大きさを勘案して定められた．後者の指標は，２つの文間の相対距離を x（制度表現文内の用語は 0，２つの文が隣接している場合は 1，間に n 個の文を挟む場合は n＋１），各文に含まれる制度関連用語数を y としたときの分布式 $y=\exp(-0.13x)$ より算出し，各用語のスコアの重みとする．なお式中の -0.13 は，専門家抽出された制度関連用語の分布から指数回帰式を導出して求められた係数である．

2.4 制度関連用語の抽出

以上の準備のもとに，「医療制度改革」クラスタに含まれる記事内の各用語に対し，課題関連度及び制度関連度の積をスコアとして計算し，スコア上位の用語を制度関連用語として抽出する．なお，記事ごとに抽出される制度関連用語数は，各記事の文数（最大 60）までとした．

3. 自動抽出手法の評価

専門家抽出された制度関連用語を正解とし，上述した提案手法により自動抽出された制度関連用語との乖離度を評価する．評価尺度としては F 値，すなわち適合率（自動抽出された制度関連用語数〔N〕に対する自動抽出された正しい制度関連用語数〔R〕の比率）と再現率（専門家により抽出された制度関連用語数〔C〕に対する自動抽出された正しい制度関連用語数〔R〕の比率）の調和平均〔2R/(N+C)〕を用いる．提案手法に対して算出した F 値を表２に示す．

表２　自動抽出手法の評価

	F 値（制度）	F 値（がん）
提案手法	0.373	0.532
tfidf 法	0.267	0.470

表中の tfidf 法（用語 t の文書中出現頻度〔tf〕と全文書数に対する用語 t の出現文書数比率の逆数に基づく指標〔idf〕の積をスコアとして用語抽出する方法）は，文書からの特徴的な用語抽出における汎用的な手法であるが，ベースラインのスコア計算法として採用する．「制度」における両者の F 値の比較により，新聞記事から制度関連用語を抽出する場合，提案手法は tfidf 法よりも有効に働くことが確認できる．また，参考値として「がん」に関する技術的対策用語を同様な手法で自動抽出したときに算出した F 値[1), 5)]も掲載した．両者を比較したところ，「制度」の F 値は「がん」の F 値よりもかなり低かった．これは「制度」の場合は自動形成した「医療制度改革」クラスタをそのまま用いて制度関連用語を自動抽出したのに対し，「がん」の場合は正確な比較評価を行うことを目的として，分析対象クラスタからノイズとなる記事を取り除いたことが主な理由と考えられる．

4. 自動抽出された制度関連用語の共語分析結果

抽出された制度関連用語に対し，共語関係の強度を表す代表的な指標である Jaccard 指標を用いて共語分析を行う．図１，２は，それぞれ自動抽出及び専門家抽出による制度関連用語に対し，共起関係にある用語の Jaccard 指標を算出し（次式参照），各用語間を辺（エッジ）で結び，ネットワーク状に配置して可視化したものである（以下，「共語マップ」と呼ぶ）．

$$\text{Jaccard 指標：} J_{ij}=C_{ij}/(C_j+C_i-C_{ij})$$

〔 C_i：語 t_i を含む記事数　C_{ij}：語 t_i 及び t_j の両方を含む記事数〕

共語マップの半径方向の長さは用語の出現記事数を表している．各用語の位置を角座標として見たときの角度成分には意味はない．用語間を結ぶ辺の線種や用語の色は凡例に示す基準にしたがっているが，共語マップに記載されている用語は，凡例に示されている Jaccard 指標の範囲にあり，かつ共起頻度が５以上のものである．

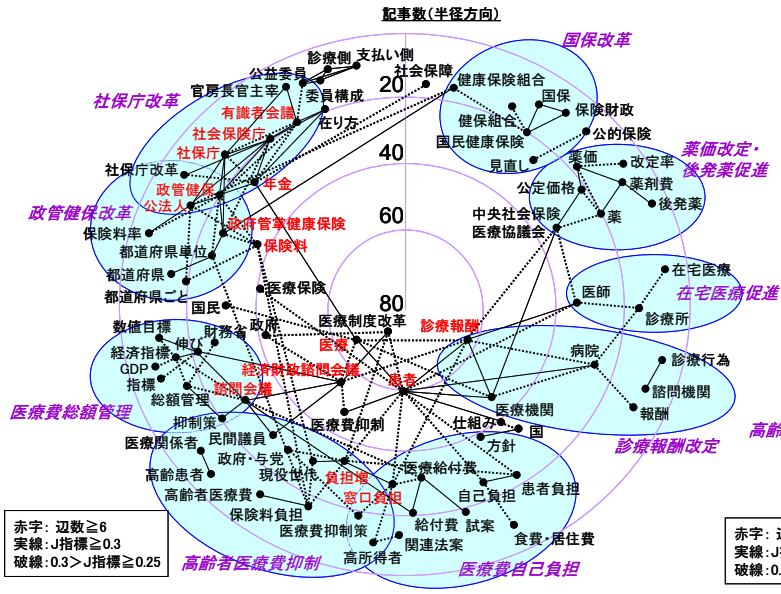

図１　自動抽出された制度関連用語の共語マップ

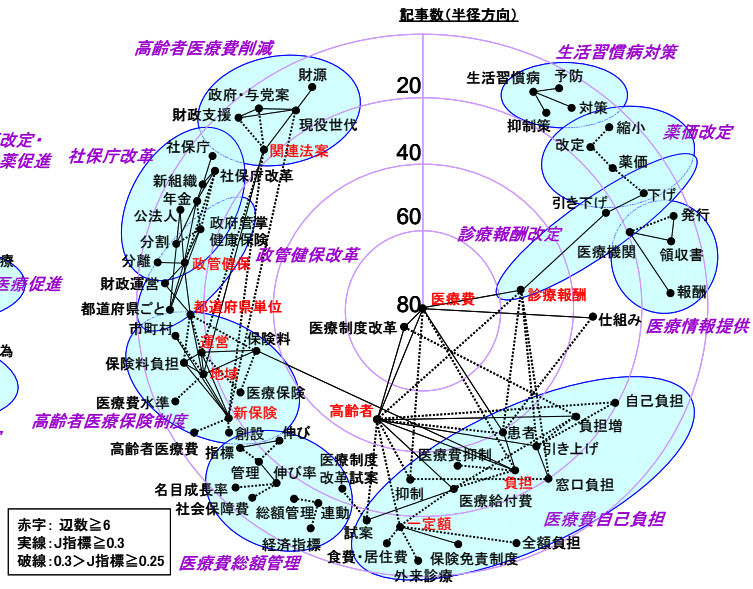

図２　専門家抽出された制度関連用語の共語マップ

自動抽出による制度関連用語の共語マップ（図１）を見ると，社会課題と制度・施策の階層構造を読み取ることができる．まず「医療制度改革」という社会課題が中心に位置付けられ，その周辺に「医療費抑制」，「医療保険」，「診療報酬」といった大きな制度分野が位置付けられている．同時にこのレベルには，制度の施行対象である「患者」や，実施主体である「政府」，「経済財政諮問会議」，「中央社会保険医療協議会」も位置付けられている．さらに外周部には具体的な制度・施策の用語の集合が９つ見ら

れる．各集合を特徴付ける内容を便宜上ネーミングすると，『医療費自己負担』，『高齢者医療費抑制』，『医療費総額管理』，『政府管掌健康保険改革』，『社会保険庁改革』，『国民健康保険改革』，『薬価改定・後発薬促進』，『在宅医療促進』，『診療報酬改定』となる（図２の紫色斜字）．これらのうち一部の集合は，上位の制度関連用語と階層構造を形成せず，独立した位置付けとなっている．また，集合間の関係に強弱が見られる．例えば，医療費の自己負担には高齢者による負担も含まれるため，『医療費自己負担』と『高齢者医療費抑制』との関係が強いことが読み取れる．また，社会保険庁改革では，社保庁所管の年金と政管健保の機能を分割することが基本となっていることから，『社保庁改革』と『政管健保改革』の関係が強いことも把握できる．

共語関係の強さを表す辺の太さやネットワークの密度から読み取れることは，「医療制度改革」が特に医療費抑制に関係する制度・施策と強く関係づけられているが，一方で『社保庁改革』や『診療報酬・薬価の改定』等は「医療制度改革」との関係は薄く，「医療制度改革」の文脈ではなく個別に注目されていたということである．また，『医療費総額管理』は結局廃案となった制度であるが，議論の過程ではかなり注目されていたということが把握できる．

専門家抽出による制度関連用語の共語マップ（図２）では，図１とほぼ同様の階層構造を確認することができる．ただし，図２の中心付近にある「医療費」や「高齢者」といった用語は図１では見られず，また周辺部では『生活習慣病対策』，『医療情報提供』といった制度・施策の集合が図１では見られなかった．前者については，自動抽出では「医療費」や「高齢者」を含む複合語を優先的に抽出しているためと考えられる．後者については，「生活習慣病」や「領収書」といった用語の課題関連度の値が低い，すなわち「医療制度改革」クラスタ以外のクラスタにも当該用語が多く含まれていることが原因と考えられる．これらの点は，今後提案手法を改善していく上での検討課題である．なお，専門家抽出ではあえて制度の実施主体を抽出しなかったが，自動抽出では実施主体を表す用語も抽出され，共語マップ上に位置付けられた．実施主体が示されることにより，用語の関係性が理解されやすくなることもあるので，その点は自動抽出のメリットの１つと言える．また，表２で示されたＦ値から，共語マップへのノイズの出現が懸念されたが，図１にはそのようなノイズは見られなかった．Jaccard 指標の下限値を下げていくとノイズが出現する可能性はあるが，下限値の取り方を工夫することによりノイズの出現を避けることができることも確認できた．

5. 結語

本研究では，新聞記事を用いて，俯瞰的アプローチにより医療分野の課題とそれを解決するための制度に関連する用語を自動抽出し，それらの記事内での共起の関係を共語分析の手法により視覚化することを試みた．この結果，制度の階層構造や具体的な制度用語群の位置付けを把握することが可能となり，提案手法の有用性について明らかにすることができた．今後は，医療制度と科学技術との関係，医療制度改革の社会受容プロセスなどを対象として，共語分析に時系列的分析の手法を組み合わせ，高齢化社会に向けた医療課題と科学技術及び制度との関係に関わる諸課題の分析に取り組んでいきたい．

謝辞

本研究は，文部科学省科学技術振興調整費「戦略的研究拠点プログラム」の支援のもとに実施した．

参考文献

1) 内海和夫，乾孝司，橋本泰一，村上浩司，石川正道 2009.03：「社会課題とその解決に結びつく科学技術に関する有用知識の抽出」『社会技術研究論文集』6，187-198

2) 橋本泰一，村上浩司，乾孝司，内海和夫，石川正道 2008.03：「文書クラスタリングによるトピック抽出および課題発見」『社会技術研究論文集』5，216-226

3) 松本浩和 2007.06：「調査研究実績からみた行政課題の抽出方法に関する研究－行政文書を素材とするテキストマイニングアプローチー」『土木計画学研究・講演集』35，163

4) Leydesdorff, L., Hellsten, I. 2006: “Measuring the meaning of words in contexts: An automated analysis of controversies about ‘Monarch butterflies’, ‘Frankenfoods’ and ‘stem cells’” Scientometrics 67(2), 231-258

5) 乾孝司，内海和夫，橋本泰一，村上浩司，石川正道 2008：「新聞記事からの社会課題に対する技術的対策情報の抽出」『第７回情報科学技術フォーラム 講演論文集第２分冊』169-170