Vol.2009-MPS-76 No.23
Vol.2009-BIO-19 No.23
2009/12/17

テキストマイニングのための ドメイン別単語辞書の構築方法

末永高志[†1]　松永務[†1]
関根純[†1]　村松正明[†2,†3]

企業に蓄積される文書情報の増加と共に，その有効活用に向けて，内容理解に基づく知識の集約のニーズが高まっている．知識の集約を行うテキストマイニングでは，一般に文書に出現する単語を基に意見の集計や関連文書の収集が行われている．この集計や収集に用いる単語辞書の構築にあたっては意見や分野（ドメイン）を代表する単語が選定されることが重要である．本稿では，共起の対となる単語の数と，共起する単語対と分野の関係の二つの着目点を持つ，分野を代表する単語を選定するための単語ランキング方式を提案する．具体的には，多くの単語により詳述される単語を代表的な単語とみなし，分野に起因する統計的な交互作用の効果による単語の組合せの評価を共起する単語について加算した基準を用いる方式である．新たな単語辞書を構築する作業を想定した実データによる評価実験の結果から，提案法によるランキング上位 10% に含まれる代表的な単語の数が，無作為に選定する場合に比べて 57% 増加することがわかった．さらに，単語辞書を構築する際の要件を考察し，提案法は要件を網羅的に満足するものであることを明らかにした．

A Term Selection Method for Domain-oriented Thesaurus in Text Mining

Takashi Suenaga ,[†1] Tsutomu Matsunaga ,[†1]
Jun Sekine [†1] and Masaaki Muramatsu[†2,†3]

There is a need to integrate knowledge extracted from electronic document data, since volume of the data is increasing in each company. The knowledge integration, such as survey of customer opinions and collection of relevant documents, is practically handled based on terms. Therefore, the terms contained on a text mining dictionary should be domain-oriented in the data. In this paper, we propose a term ranking method for selecting representative term by considering a statistical interaction of co-occurrence term pair in a specific domain and co-occurrence term number from a point of view that a representative term is described using a variety of other terms. Experimental results using real medical documents show that our method of term ranking performs good term ranking and representative term number included in our method's rank of top 10% increases by 57% than one in a random sampling. Additionally, our method is shown that it is well-suited for requirement in developing a domain-oriented thesaurus.

1. はじめに

企業に蓄積される文書情報は規模を増し，これら資源の内容理解を元に意思決定につなげる，知的処理支援を行うテキストマイニングの技術が広く使われるようになってきている[1)]．これは，文書を対象とした自動処理技術の課題が，個々の文書の内容を把握することから，顧客の意見が記述された文書から評判傾向を把握したり[2)]，事故報告レポートから頻発する課題を探索するなど[3)]，内容理解に基づく知識の集約にシフトしていることを意味する．

文書から知識を集約するにあたっては文書中に出現する単語が基となる．例えば，評判情報の分析における満足や批判などの意見を表す単語の辞書が，様々な分析を行うための基礎的な知識として必要とされることが指摘されている[4)]．また，Web 上に散在した医療関連の文書を集約し一般向けに適切な情報を提供する試みとして，医療機関に蓄積された文書データを元にした疾患に関する単語辞書構築の検討が挙げられる[5)]．

上記の評判や疾患を表す単語は文書の主旨や分野（ドメイン）に由来するものと考えられ，上記の評判情報の分析や医療関連文書の集約の実現においては，文書の主旨や分野ごとに対応する単語を集約した辞書を用意しなければならない．この辞書の作成にあたっては，上記の主旨や分野を代表する単語が含まれていることが望ましく，様々な場面で蓄積された文書を元にした専門家による単語の選定が一般的に行われる．選定の対象となる単語の数は膨大となるため，自動処理による選定支援が要求されてきている[5)]．

選定支援の課題に対して，文書に含まれる単語と分野の関係をある着目点で基準化し，その基準を用いて単語をランキング形式で提示することで，分野を代表する単語がより先に確認されるようにして効率化を図る方法が検討されている．具体的には，医療関連文書のように疾患の分野に対応する文書が収集されるような場合において，ある分野とそれ以外の

†1 株式会社 NTT データ 技術開発本部 R&D Headquarters, NTT DATA CORPORATION
†2 東京医科歯科大学 難治疾患研究所 Medical Research Institute, Tokyo Medical and Dental University
†3 ヒュービットジェノミクス株式会社 Research Institute, HuBit Genomix Inc.

Vol.2009-MPS-76 No.23
Vol.2009-BIO-19 No.23
2009/12/17

分野を比較評価する情報利得や相互情報量などの単語選択基準を用いた単語ランキング方式が一般に利用されている[6),7)]．これらの方式は，単語の出現する分野の偏りに着目したものである．この着目点は個々の単語を個別に評価するものであるが，単語と単語との違いを直接的に評価するものではない．そのため，これらの単語間の関係を考慮することによる，分野を代表する単語のランキング方式の高度化の余地が残されていると言える．単語間の関係を考慮する方式として，ある単語が出現する文書に含まれる単語集合の頻度分布と文書全体に出現する単語集合の頻度分布の違いに着目した方式が提案されているが[8)]，これは分野の観点がないため分野を代表する単語を選定する場合には適さない．その他に，専門用語らしさに着目する専門用語抽出方式が提案されているが[9)]，蓄積された文書から専門用語と思われる単語の抽出は可能であっても，専門用語らしさの着目点が必ずしも分野を代表するものとはならないため，単語のランキングを行うにあたっては異なる検討が要求される．

本論文では，分野を代表する単語の選定を行うにあたり，分野を代表する単語を多くの単語を用いて詳述される主題とみなし，共起の対となる単語の数が多くなることと，分野に由来する単語の組合せにおける統計的な交互作用[10)]の効果を考慮する，分野と単語対の関係に着目した単語ランキング方式を提案する．いくつかの疾患について記述された医療文書データと，標準的な傷病に対する単語を集約した標準病名データを用いて，提案の有効性を実験により検証する．具体的には，分野を代表する単語を選定する業務において，文書データに単語ランキング方式を適用することで，分野を代表する単語に対応する標準的な病名が，上位に挙げられる割合を求め標準的な病名がより先に確認される割合を試算する．これらを基に，既存の方式と提案する方式の効率化の向上率を評価し提案法の効果を示す．さらに，単語辞書の作成に対する要件整理を行い，各々の要件に応じた辞書作成のための指針を示す．

2. 分野と単語の関係に着目した既存のランキング方式

本章では，これまでに知られる分野と単語の関係に着目した既存の単語ランキング方式について，方式の適用方法と，これらの基となる基準の定義および選択される単語の着目点を概観する．

単語ランキングの実施にあたっては，分野ごとに分けられた文書データを用いて，各々の文書に含まれる文章を単語単位に分割した後に，定義に従って単語ごとに基準値を算出し降順に並べ直すことでランキングを行う．ランキング上位の単語から辞書に登録するか否かの判断を各々の分野ごとに実施すること想定している．

以下，与えられた文書データ[⋆1]に対して単語選定の対象とする分野が複数あることを想定し，選定にあたる分野を c，それ以外の分類を $\bar{c}$，単語を w，文書データに出現する w 以外の全ての単語を $\bar{w}$ とする．また，$P(w)$ は単語 w の文書集合での出現確率，$P(w,c)$ は単語 w と分野 c の同時確率，$P(w|c)$ は分野 c の条件の元での単語 w の周辺確率とする．

情報利得[6),7)]

情報利得 $IG(w,c)$ は単語 w と分野 c の各々の出現確率 $P(w)$，$P(c)$ に対する同時確率 $P(w,c)$ との違いを対数尤度比で評価する指標の一つであり，単語と分野の独立性を考慮した基準となっている．これは，次式で定義される．

$$IG(w,c) = \sum_{C\in\{c,\bar{c}\}} \sum_{W\in\{w,\bar{w}\}} P(W,C) \log \frac{P(W,C)}{P(W)P(C)} \tag{1}$$

相互情報量[7)]

相互情報量 $MI(w,c)$ は，単語の分野に対する相互依存の尺度を表す量で選定にあたる分野 c に特化した基準であり，次式で定義される．

$$MI(w,c) = \log \frac{P(w,c)}{P(w)P(c)} \tag{2}$$

上記二つの基準を用いた単語ランキング方式の特徴として，情報利得は高頻度に出現する単語が優先的に上位に挙げられ，逆に，相互情報量は低頻度に出現する単語が優先的に上位に挙げられることが報告されている[11)]．このように，既存の基準を用いる方式では単語の出現する分野の偏りにのみ着目しており，分野を代表する単語とそれ以外の単語との比較の観点が原理的には盛り込まれていない．

3. 分野と単語対の関係を考慮したランキング方式の提案

本章では，ある分野を代表する単語とそれ以外の単語の関係，および分野にまたがって出現する一般的な単語，すなわち分野の代表とならない単語とそれ以外の単語の関係の検討を基に，分野と単語対（単語と単語の組合せ）の関係を考慮した分野を代表する単語を選定するための単語ランキング方式を提案する．

ある分野を代表する単語とそれ以外の単語の関係においては，分野を代表する単語をある分野での主題に対応しそれ以外の単語を用いて詳述されるものとみなすと，このような詳述

⋆1 一文書あたりに割りあてられる分野は一つとして扱う．

Vol.2009-MPS-76 No.23
Vol.2009-BIO-19 No.23
2009/12/17

の対象となる単語は共起の対となる単語がより多く存在すると言える．一方で，分野にまたがって出現する一般的な単語とそれ以外の単語の関係を考えると，出現頻度の高いものが存在することが想定され，分野を代表する単語と同様に，共起の対となる単語が多く存在すると考えられる．このような一般的な単語を含む単語対は，出現頻度の高さに起因して分野に関係なく偶発的に共起した結果と言える．

このことから，共起の対となる単語の数に加え，ある分野に出現する単語対の頻度から偶発的に共起する可能性を排除した，いわゆる交互作用[10] の効果に着目することが，分野と単語対の関係を考慮するにあたっては必要と言える．上記の共起の対となる単語の数は分野の主題性につながり，交互作用の効果は単語対による分野の特定性につながると言え，分野の主題性と分野の特定性の二つに着目する基準を提案する．具体的な手順は以下の通りである．

まず，分野の特定性を表す単語対の交互作用の効果 $DI(w_1, w_2, c)$ を以下のように定義する．

$$\begin{aligned} DI(w_1, w_2, c) &= N_{w_1,w_2,c} - N_c P(w_1) P(w_2) \\ &= N_{w_1,w_2,c} - N_c N_{w_1} N_{w_2} / N^2 \end{aligned} \tag{3}$$

ここで，w_1 と w_2 はそれぞれ異なる単語，$P(w)$ は文書全体の単語 w の出現確率である．また，$N_{w_1,w_2,c}$ は分野 c の文書に含まれる単語対 w_1 と w_2 が共起する文書頻度，N_c は分野 c の文書数，N_w は単語 w の文書全体の頻度，N は全文書数とする．この式の右辺第二項が一般的な単語の偶発的に共起する可能性を排除する処理に対応する．

次に，分野の主題性を表す出現の共起の対となる単語の評価 $DWPI(w, c)$ を，(3) 式を用いて以下のように定義する．

$$DWPI(w, c) = \sum_{w_p \in W_{pair}(w,c)} DI(w, w_p, c) \tag{4}$$

ここで，$W_{pair}(w, c)$ は分野 c において単語 w と共起する単語の集合を表す．式 (4) では，多くの単語と共起する単語は加算される値が多く存在することを意味し，このような単語は上位にランキングされることになる．

単語ランキングを行う場合は，既存の方式と同様に式 (4) の定義に従って単語ごとに算出された基準値を用いて降順に並べた単語列を提示する．

4. 実データによる提案法の効果検証

本章では，疾患ごとに分けられた医療文書データを対象に提案の単語ランキング方式の評価を行う．これは，蓄積された文書データから分野の代表となる単語の選定を行う際に，単語ランキング方式を導入することを想定したもので，分野を代表する単語をより先に選定対象とすることの効果検証を目的とする．

以下，実験の評価方法および，実験に用いる医療文書データと標準病名データの概要を説明し，提案法と既存の単語選択基準を用いる方式との比較結果について述べる．

4.1 評価方法

実験に用いるデータは，疾患の分野ごとに分けられた医療文書データと，表記の統一を目的とし医療の専門家により定められた標準病名データである．医療文書データは分野を代表する単語の選定対象として用い，標準病名データに登録された標準病名は単語をランキングした結果の性能評価に用いる．すなわち，標準病名データと一致するものを選定すべき分野を代表する単語として扱う．

具体的な手順としては，疾患の分野に分けられた文書データに含まれる単語を取得し，分野ごとに単語ランキング方式を適用し選定順に単語を提示する．それらに対し，標準病名が上位ランキングに含まれる数を算出する．

単語ランキングに用いる方式は，分野を考慮する単語ランキング方式である，提案法（domain-oriented word pair interaction method；以下 DWPI），情報利得を用いる方式（information gain measure；以下 IG），相互情報量を用いる方式（mtual information measure；以下 MI）の三つとする．

各方式の評価にあたってはゲインチャート[12] とリフト率[13] を用いる．ゲインチャートは，分野に対応する文書に含まれる標準病名の全数に対する，ランキング上位に挙げられる標準病名の割合を意味するゲイン率をランキング方式ごとにプロットしたものである．これは文書データに含まれる膨大な数の単語に対して，選定対象とする単語の数をランキング方式に基づいて制限した場合に選定される標準病名の割合を表し，提案法によってより多くの標準病名が選定対象となることを確認する．リフト率は，ゲイン率と同様の想定で，無作為に確認する場合と比較してランキング上位に標準病名が含まれる数の向上する割合を表し，提案法による効率化の効果の最大値を確認する．

さらに，提案法と既存法で上位にランキングされた単語の不一致となる割合から，ランキング方式により選定される異なる単語の例を示す．最後に，一般的な単語が単語がランキング上位に含まれなくなる例を示し，提案法により選択される単語が分野を代表するものを優先付けていることを確認する．

また上記のゲイン率やリフト率の算出にあたっては，作業効率を確認するため分野全体で

Vol.2009-MPS-76 No.23
Vol.2009-BIO-19 No.23
2009/12/17

平均した値を用いることにする．実作業においては，疾患の分野ごとに専門チームに分かれて選定することになると考えられ，その場合における全体的な効率化の期待値を評価することに相当する．なお，代表となる単語は分野間での重複を認めることとする．

4.2 医療文書データ

医療文書データはメルクマニュアル[14] を採り上げる．これは，主要な疾病を網羅し，症状から診断，治療法に至るまで医療従事者向けに総合的に記載されたものである．ここで用いられる単語を選定の対象とする．

文書データの概要に関しては，疾患は「栄養障害」,「内分泌・代謝疾患」,「消化器疾患」といった 23 個の分野に分かれており，それぞれに記載される病因，症状，診断，治療法などの段落を単位に，便宜上一つの文書として取り扱う．選定の対象とする単語は，形態素解析ツール Mecab [⋆1] を用いて形態素に分割し名詞と判断されたものとした．共起する単語対は同一文書に出現する単語と単語の組合せの全てとした．文書数は 5,601 件で，異なる単語の数（以下，異なり語数）は 16,424 個，文書中に出現する異なる単語対の数（以下，異なり単語対数）は 5,762,352 個あり，単語の全ての組合せの 4.27% に相当する．

4.3 標準病名データ

評価に用いる標準病名の単語データは，ICD10 対応電子カルテ用標準病名マスター第 2 版[15]（以下，病名マスタ）を用いる．これは，医療従事者により作成される診療報酬請求書（レセプト）[⋆2]に記述する傷病名の標準化に向けて，電子カルテなどに利用されることを目的に構築されたものである[⋆3] ．

登録語数は 86, 331 個であるが，これらの語は「1 型糖尿病」の「1 型」といった修飾語が多く含まれるため，個々の語に対して文書データと同様に Mecab を用いて形態素に分解して得られた異なり語数 45, 629 個の単語を，評価に用いる標準病名データとした．なお，4.2 節で述べた医療文書データ全体に含まれる単語と一致する標準病名の異なり語数は 3, 766 個である．疾患の分野ごとに含まれる異なり語数における標準病名の割合は，最小値は 0.265 で最大値は 0.552 と多少のばらつきが見られた．

⋆1 http://mecab.sourceforge.net/
⋆2 患者が受けた診療に対して，医療機関が健保組合などの保険者に請求する医療費の明細書のことで，診療にともなう検査や処方薬の費用が傷病名と共に記載されているものである．
⋆3 病名マスタには「病名基本テーブル」,「修飾語テーブル」,「索引テーブル」の三つのテーブルが用意されており，標準病名データとして登録語数の最も多い「索引テーブル」を用いることにする．

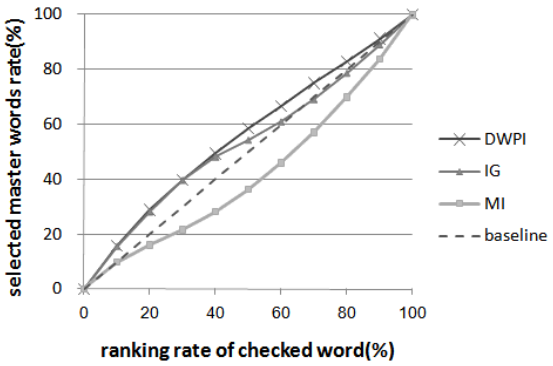

図 1 単語ランキング方式ごとのゲインチャート．
Fig. 1 Gain charts of the proposal method and the exiting measures.

4.4 単語ランキングによる効率化試算

提案法の評価を**図 1** に示すゲインチャートにより行う．図中の横軸がランキング上位 r%，縦軸がゲイン率を表す．また，図中の baseline は無作為に単語を選定した場合に含まれる標準病名の期待値を意味する．この結果から，無作為に選定する場合に比べて，提案法と IG ではゲイン率が向上し，MI では低下している．また，ランキング上位 40% までは提案法と IG は同等で 50% 程度のゲイン率が得られるが，それ以上になるとゲイン率が開き始める様子が示されている．また，IG はランキング上位 60% から無作為に選定した場合とほぼ変わらなくなっている．以上から，提案法では広い範囲でゲイン率が向上することがわかる．

次に，各々の単語ランキング方式のリフト率を**表 1** に示す．これは，ゲインチャートと同様に上位ランキングに含まれる標準名称の数を評価するもので，無作為に選定を行った場合との比較を意味する．提案法はランキングの全ての状況において既存の方式を上回るか同等で，リフト率の最大となるのは，提案法によるランキング上位 10% の単語を選定したリフト率 57% であることが示されている．

提案法と IG では，ランキング上位 40% までは同等であったことから，これらのランキング上位に含まれた標準名称の中で異なった単語の割合を確認する．具体的には，提案法でのみ上位にランキングされた標準名称と，提案法と IG の両方で上位にランキングされたものに分け，それらの比率を算出した．**図 2** はその結果である．DWPI が提案法でのみ，identical が共通して選択された単語の割合である．横軸がランキング上位 r% から選定し

Vol.2009-MPS-76 No.23
Vol.2009-BIO-19 No.23
2009/12/17

表 1 単語ランキング方式ごとのリフト率（%）.
Table 1 Lift rates (%) of the proposal method and the exiting measures .

ランキング	DWPI	IG	MI
10	57	55	-3
20	44	41	-20
30	32	32	-28
40	24	20	-30
50	17	9	-27
60	11	2	-24
70	7	-1	-19
80	4	-2	-13
90	1	-1	-7

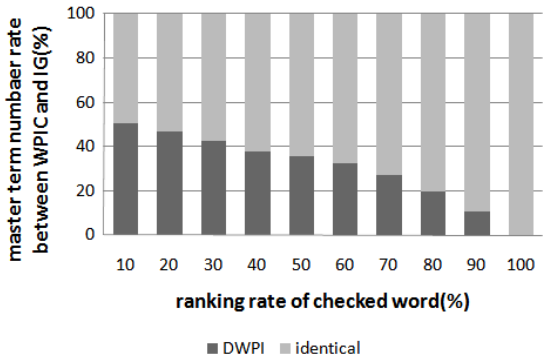

図 2 提案法と IG により選択された標準名称の一致割合.
Fig. 2 Identical master term number rates between the ranks selected by the proposal method and IG.

た割合，縦軸が単語数の比率を意味する．ランキング上位 40% の時点では，およそ 40% 弱程度の異なる単語が選定され，ランキング上位 60% の時点では，およそ 30% 強程度の異なる単語が選定されている．このことから，選定される単語に傾向の違いがあることが示唆される．

上記のように，提案法と IG で異なる単語がランキング上位に挙げられた．**表 2** は提案法でのみ含まれる標準名称と，IG でのみ含まれる標準名称の具体例を用いてランキングされた単語の違いの例である．この表は,「栄養障害」の分野を対象にランキング上位 60% まで選定を行った際の，いずれかの方式でのみ選定された単語の中でランキング最下位から 5 件抽出した例である．いずれの方式でも該当する分野での出現頻度は低い．また，IG で選定の対象となった「適当」という単語は一般的な意味で用いられる単語と見られるが，提案の共起する単語対による分野の特定性を考慮することによって，このような抽象的な単語はランキングの上位に挙げられなかったと考えられる．

表 2 提案法と情報利得で選択された素性の違いの例.
Table 2 An example of identical terms selected by the proposal method and IG.

提案法のみ選択	分類 栄養障害	それ以外
アナフィラキシー	2	35
落屑	1	27
光線	1	42
弾性	1	34
混濁	1	38
情報利得のみ選択	**分類 栄養障害**	**それ以外**
適当	1	53
ろう	1	0
エナメル	1	5
カドミウム	1	5
シリコン	1	2

5. 考　　察

本章では，提案する単語ランキング方式を実業務に適用することを想定して，単語辞書の構築に求められる要件と，単語ランキング方式に要求される特性について考察する．さらに，提案法がこの特性を満たすことを示す．

第 4 章の実験では，与えられた文書データを元に分野を代表する単語の選定を新規に行う作業を想定して検証を行ったが，実業務においては既に単語辞書が構築済みでこの辞書の高度化が求められる場合が考えられる．ここから，要件として既存の単語辞書の有無が挙げられる．また，辞書に登録する単語に関して，評判情報で用いる単語辞書と文書の収集のための単語辞書の構築の場合では要件に違いがある．具体的には，評判情報で用いる単語は意見の集計という観点で選定されるためまれな表現のものは必要性が低いと考えられる一方で，文書の収集に用いる単語では，例えば，他言語からの借用が多い医学的用語ではカタカナ表記による表記揺れが数多いことが指摘されており[16]，概念としては一般的であっても

Vol.2009-MPS-76 No.23
Vol.2009-BIO-19 No.23
2009/12/17

まれな表記方法が数多く存在するため，まれな表現であっても必要性が高いものが存在すると考えられる．ここから，要件のもう一つに単語の出現の多寡が挙げられる．

上記二つの要件を整理すると，既存の辞書が存在しない辞書構築の初期段階においては，頻出するなどの基本的な単語は漏らさず選定の対象とすべきである．一方の，高度化が求められる場合ではランキング上位に挙がるような単語は登録済みと考えられるため，まれな表現（単語）も含めて分野の代表となるものを選定する必要がある．

ここから，単語ランキング方式に求められる特性を考察すると，辞書構築の初期段階においては，ランキング上位の単語が先に選定対象とされると考えられ，ここに含まれる単語が適切にランキングされることが求められる．一方の高度化が求められる場合においては，初期段階において選定の対象から漏れた概念としては一般的であってもまれな表記の単語は，出現頻度が低く信頼性に期待できないことから，分野を問わず出現する一般的な単語をランキング下位にすることで，一般的でない可能性のある単語をランキング上位にする特性が期待される．

これらの機能に対して 4.4 節で示した通り，初期段階で求められる単語については，提案法も情報利得を用いる方式のいずれもランキング上位 40% 程度までゲイン率の向上が示されたことから，いずれも要件に合致する特性を有すると考えられる．一方の，高度化の段階で求められる特性については，提案法は「適当」のような抽象的な単語を挙げなかったことから提案法のみ要件に合致すると考えられる．以上より，提案法は単語辞書構築の要件に基く単語ランキング方式に求められる特性を有すると考えられる．

6. まとめと今後の課題

本論文では，分野の主題性を表す共起の対となる単語の数と分野の特定性につながる単語対の交互作用の効果に着目した，分野と単語対の関係を考慮した分野を代表する単語選定のための単語ランキング方式を提案した．医療に関連する文書データによる単語ランキングの実験結果から，提案する方式によって標準病名データに登録された単語が上位に挙げられることを示した．さらに，このような単語辞書を構築する際の要件を検討し，提案する方式がこれらの要件を網羅的に満たすことを示した．

テキストマイニング技術の適用先が広がるにつれ，今回のように分野を代表する単語を提示するだけでなく，単語の選定を行う際に根拠となる情報を付与するなど，要因となる背景の抽出や理解が一層期待されるようになると考えられる．今後の課題としては，このような根拠となる情報を提示する理解支援のための技術の検討が挙げられる．

参 考 文 献

1) 小池麻子：テキストマイニングによる潜在的知識の発見支援，情報処理， Vol.48, No.8, pp.824–829 (2007).
2) 倉島健，藤村考，奥田英範：大規模テキストからの経験マイニング，電子情報通信学会論文誌 D， Vol.92, No.3, pp.301–310 (2009).
3) 齊藤孝広，薬師寺あかね，渡部勇，松井くにお，佐々木敏宏，寺田昭，齋藤隆：航空安全情報分析ツール–因果関係に着目したレポート分析手法の提案–，飛行機シンポジウム講演集，日本航空技術協会 (2008).
4) 乾孝司，奥村学：テキストを対象とした評価情報の分析に関する研究動向，自然言語処理， Vol.13, No.3, pp.201–241 (2006).
5) 中川晋一，内山将夫，三角真，島津明，酒井善則：コーパスに基づくがん用語集合の作成と評価，自然言語処理， Vol.16, No.2 (2009).
6) 田中牧郎，金愛蘭，桐生りか，近藤明日子：コーパスによる難解語・重要語の抽出–医療用語を例に–，社会言語科学会第 21 回大会，社会言語科学会 (2008).
7) Sebastiani, F.: Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, Vol.34, pp.1–47 (2002).
8) Hisamitsu, T., Niwa, Y. and Tsujii, J.: A method of measuring term representativeness: baseline method using co-occurrence distribution, *Proceedings of the 18th conference on Computational linguistics-Volume 1*, Association for Computational Linguistics Morristown, NJ, USA, pp.320–326 (2000).
9) 中川裕志，森辰則，湯本紘彰：出現頻度と連接頻度に基づく専門用語抽出，自然言語処理， Vol.10, No.1, pp.27–45 (2003).
10) 東京大学教養学部統計学教室編：自然科学の統計学，東京大学出版会 (1992).
11) 末永高志，松永務，関根純：相補的な素性選択基準の関係を考慮した文書分類のための素性選択方式，*MPS*, Vol.73 (2009).
12) 佐藤栄作：マーケティング・サイエンス III: 顧客ターゲティング分析: データマイニング手法の活用，オペレーションズ・リサーチ， Vol.48, No.3, pp.210–215 (2003).
13) 鶴田育雄，後藤正輝，香田正人：リレーションシップ・データへのデータマイニングの適用，オペレーションズ・リサーチ， Vol.47, No.9, pp.581–587 (2002).
14) 福島雅典：総監修（翻訳・編集：日経メディカル）：メルクマニュアル第 17 版日本語版，日経 BP 社 (1999).
15) 医療情報システム開発センター：ICD10 対応電子カルテ用標準病名集，日経 BP 社 (2002).
16) 荒牧英治，今井健，美代賢吾，大江和彦：Support Vector Machine を用いた医学用語の表記ゆれ解消，言語処理学会年次大会発表論文集，言語処理学会，pp.135–138 (2008).