情報処理学会論文誌　数理モデル化と応用　Vol. 3　No. 2　108–118 (Mar. 2010)

単語の重要度評価基準の検討と医療関連文書への適用評価

末永高志[†1]　松永務[†1]
関根純[†1]　村松正明[†2,†3]

医療機関で作成される関連文書の電子化により，投薬ミスを防止するシステムや保険審査業務の高度化などが実現されている．医療保険を例にすると保障対象などで扱う疾患が異なるため，審査業務を行う前に疾患の分野といったカテゴリの観点での文書の振り分けが必要である．この振り分け処理は保障対象に関わる単語がもととなり，業務の実現のためにはカテゴリに応じた単語辞書が必要とされる．この辞書に登録すべき単語の条件として，まれな疾患であっても保障対象に関わるならば無視できず，候補となる単語が膨大に存在するため辞書の構築をいかに効率化するかが課題となっている．このような特徴を持つ医療関連の単語辞書構築において，単語の重要度を評価する基準の適用による業務効率化の検討を行った．この結果，カテゴリと単語の関係を考慮することに加え，多くの単語を用いて説明される単語を重要な単語と見なし，単語ごとに共起する単語の交互作用の効果を加算した基準の適用が有効であることが分かった．

A Study on Term Selection Measures and Applying to Medical Document Data

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

A method of checking a huge amount of electronic medical documents is becoming a key technology for enabling efficient prevention of medical accidents, insurance screening processes and so on. Since the method is executed in each health insurance policy and the documents should be categorized using medical terms, a dictionary including the terms is important. A variety of candidates for the term is a huge amount because rare-occurrence terms related to the policy should not be ignored. For the reason, how to effectively and semi-automatically construct the dictionary is required. We proposes a term selection measure by considering a statistical interaction of a co-occurrence term pair in a specific category, and a 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 performs good term ranking.

1. はじめに

医療事故の防止や保険審査業務の高度化といった社会的な要請の実現において，医療機関により作成される関連文書の電子化が重要な要素となっている[1),2)]．医療文書の電子化によって，たとえば，医療事故につながりかねない投薬ミスを防止するシステムが実現されている[3)]．このシステムでは，診療報酬請求書（レセプト）[⋆1]のデータ入力とあわせて，罹患した疾患と処方される医薬品の適合性チェックを自動的に行っている．そのほか，保険審査業務において保障対象に応じた分類ごとに関連する単語を含む辞書を用意し，この辞書をもとに支払いの可能性のある診断書として検索されたものを，審査担当者が優先的に確認することで顧客により早く案内を送付する，請求勧奨業務を実現するシステムが報告されている[4)]．このシステムによって，保険商品の複雑さに起因した顧客による保障適否の判断が困難な現状に対して，提供するサービスの品質向上の期待が持たれている[⋆2]．

医療関連文書のデータ入力にあたっては，医療に関連する単語には表記揺れなどの慣例表現や類似する医薬品の名称の存在が知られており[5),6)]，入力結果の信頼性向上のために医療に関連する単語辞書を充実させることが必要である．また，請求勧奨業務のように保障対象と関連する単語を含む辞書を用意するにあたっては，保障対象によって用いられる医療用語が異なるため担当分野ごとにチーム分けされており[⋆3]，担当分野の分類（カテゴリ）で重要とされる単語の登録された辞書が必要となる．これに加え，まれな疾患だとしても保障

†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.

⋆1 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で，患者が受けた診療に対して診療にともなう検査や処方薬の費用が傷病名とともに記載されている．

⋆2 診断書の入手には一定額の経費がかかるため，保障されない可能性を考慮するととりあえず保険会社に請求してみるという行動を，顧客はとりにくい現状にある．

⋆3 ニュースリリースにより請求勧奨業務の各社の取り組みが報告されている．
http://www.dai-ichi-life.co.jp/company/news/pdf/2007_036.pdf
http://www.jbaudit.go.jp/report/summary19/pdf/yokyu_36.pdf
http://www.sumitomolife.co.jp/news/090731.pdf など．

対象に関わる単語であれば顧客に提供するサービスの公平性の観点から無視できないため，まれな疾患を表す単語も網羅的に含めた辞書が期待される．

上記の単語辞書の構築においては一般に専門家の協力が不可欠で，実在の蓄積文書をもとに専門家による単語の選定が行われている[4),7)]．たとえば，疾患名の 1 つである「肝がん」という直接的な単語や，間接的に関連する「肝硬変」や「ウィルス性肝炎」といった単語のレベル間の基準を決め，専門家の語感を信用した選定を行うことで妥当性の高い単語辞書が構築可能であることが報告されている[7)]．しかしながら，蓄積文書に含まれる単語の種類は膨大で電子化される文書数の増加にともなって，人手による単語選定だけではなく，自動処理による単語選定の業務効率化支援が期待されている[4)]．そこで本稿では，網羅的な単語辞書の構築における業務効率化支援技術の検討を行う．

業務効率化支援の実現にあたり医療関連文書を対象とする単語選定の要求を整理すると，「消化器疾患」や「小児疾患」などの専門分野や，保険における保障対象の担当分野といった，カテゴリに分類される文書から重要とされる単語を適切に提示することあげられる．また，上記のまれな疾患も含めた網羅的な単語辞書の実現においては，頻出するなどの基本的な単語とあわせて，低頻度にしか出現しないような単語の中でもまれな疾患といえる単語を適切に選定対象とすることが期待される．

上記の要求に対して，単語の重要度を何かしらの基準で評価し，評価値の高いものから優先的に選定対象とするアプローチが知られている．カテゴリの付与された文書については，情報利得の単語評価基準を用いた単語ランキング方式が広く利用されている[8),9)]．この基準では低頻度に出現する単語が適切に評価されない場合があることが指摘されており[10)]，まれな疾患も含めるような網羅的な辞書を構築する場面においては課題が残る．また，単語間の関係を評価の対象とし，ある単語が出現する文書に含まれる単語の，出現頻度の分布構造に着目する基準が提案されている[11),12)]．この基準では，ある単語の出現する文書に含まれる単語の出現頻度の分布と文書全体の単語の出現頻度の分布の差が大きくなる単語を，重要な単語として評価している．これらの基準は，蓄積された文書の中で高頻度に出現する単語を抽出することが対象で，文書の概要把握を目的とするキーワード抽出への適用が報告されている．しかしながら，基準の適用を高頻度に出現する単語に限定したり単語のクラスタリングを行ったりするなど，対象とする単語の種類を減らす工夫が必要なことから[11),12)]，本稿で対象とする網羅的な単語辞書を構築する場面における効果は明確ではない．

一方，専門用語が単語の組合せ（複合語）で構成されることに着目し，複合語を抽出するアプローチが提案されている[13)–15)]．このアプローチでは，複合語の専門用語らしさを評価するために，C-value[13)]，連接統計情報[14)]，部分文字列のパープレキシティ[15)] といった基準を用いている．しかしながら，複合語を構成する単語の組合せがまったく異なる場合においては，専門用語らしさを評価するのみで，蓄積文書における重要さを比較する基準とはなっていない．蓄積文書における重要さの評価にあたっては，文献 14) では専門用語らしさの基準と単語の出現頻度の積を評価基準とするなどのアドホックな対策がなされている．

本稿では，医療関連文書から単語を選定するにあたって，支援技術の要件としてあげたカテゴリの観点の導入可能性と，低頻度に出現する単語の適切な評価の可能性を検証することを目的とし，複合語抽出のアプローチではなく，単語の重要度を評価するアプローチに基づいた技術評価を行う[⋆1]．技術評価においては，3 つの既存の単語の重要度評価基準を取り上げ，基準の形式をもとに課題の整理を行い解決策を提案する．さらに既存法と提案法の基準を用いた各単語ランキング方式について，カテゴリで重要とされる単語を選定する業務を想定した実データによる業務効率化試算による比較評価を行う．具体的には，疾患の分野にカテゴリ分けされた医療関連文書に対しカテゴリごとに単語ランキングを実施し，標準的な傷病を表す単語を集約した標準病名データに含まれる単語が上位にあげられる割合を求める．これにより，カテゴリで重要とされる単語がより先に確認可能となることを検証する．さらに，比較検証の結果により提案法が業務効率化に効果があることと，提案法は単語辞書を構築する状況で求められる特性を満たす基準であることを，それぞれ示す．

以上の結果から，カテゴリと単語の関係を考慮することに加え，提案する，多くの単語を用いて説明される単語を重要な単語と見なし共起する単語の交互作用の効果について加算した基準の適用が，有効であることが分かった．

以下，2 章で既存の単語評価基準を概観する．3 章では既存法の課題を整理し，解決策を提案する．4 章では，実データをもとに各基準の効果検証を行う比較結果を報告する．5 章では，単語辞書構築における要件整理を通して単語選定に関する指針を考察する．6 章はまとめと今後の課題である．

2. 既存の単語評価基準

本章では，単語選定における単語の重要度を評価するアプローチを対象に，ここで用いられる単語の評価基準について，基準の定義と選定にあたって重視する単語の性質を概観す

⋆1 実業務を考慮すると，単語選定支援システムにより提示された単語の登録有無を判断する際には，文書中の単語の前後関係を確認したうえで決定することから複合語の抽出は専門家の判断によるものとし，単語の重要度を評価する基準の検討を優先した．

表 1　本稿で用いる記号の定義
Table 1　Definitions of notation in this paper.

記号	定義
w	単語
$\bar{w}$	文書データに出現する w 以外のすべての単語
w_p	w と同一文書に出現する単語
c	選定にあたるカテゴリ
$\bar{c}$	c 以外のすべてのカテゴリ
$P(w)$	w の文書集合全体での出現確率
$P(w,c)$	w と c の同時確率
$P(w\|c)$	c の条件のもとでの w の周辺確率
N	全文書数
N_w	w の出現する文書の数（文書頻度）
N_c	c の文書数
N_{w,w_p}	w と w_p の同一文書に出現する文書の数（共起頻度）
$N_{w,w_p,c}$	c における単語 w と w_p の共起頻度
$W_{pair}(w)$	w と同一文書で共起する単語の集合
$W_{pair}(w,c)$	c に属する文書の中で w と同一文書に出現する単語の集合

る．既存の単語評価基準については，文書のカテゴリと単語の関係を考慮する情報利得の基準と相互情報量の基準，および単語間の関係をもとに，出現する文書での単語の分布構造を考慮する単語共起の χ^2 適合度統計量の基準[12)]（以下，単に χ^2 統計量と呼ぶ）を取り上げることにする．

単語ランキングの実施にあたっては，蓄積さ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，カテゴリ単位で定義に従って単語ごとの基準値を算出し，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カテゴリごとに実施することを想定している．

ここで，本稿で用いる記号の定義を**表 1** に示す．なお，本稿では共起頻度に対する検証を行っているため，単語の出現確率もそれにあわせ文書頻度をもとに $P(w) = N_w/N$ と定義する．

2.1　カテゴリと単語の関係を考慮する基準

蓄積された文書データにおける単語選定の要件に，文書に付与されたカテゴリごとに重視する単語が異なることがあげられる場合において，一般に単語の出現するカテゴリの偏りを評価する基準が用いられる[⋆1]．本節では，この考えで広く使われている情報利得の基準と相互情報量の基準[10)] を取り上げる．

情 報 利 得

情報利得の基準 $IG(w,c)$ は単語 w とカテゴリ 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基準である．これは，単語とカテゴリの独立性を考慮した基準といえ，

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

と定義される．

相互情報量

相互情報量の基準 $MI(w,c)$ は，単語のカテゴリに対する相互依存の尺度を表す量で，選定にあたるカテゴリ c に特化した基準であり，

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

と定義される．

上記 2 つの基準を用いた単語ランキング方式の特徴として，情報利得の基準は高頻度に出現する単語が優先的に上位にあげられ，逆に，相互情報量の基準は低頻度に出現する単語が優先的に上位にあげられることが報告されている[16)]．また，これらの方式では単語の出現するカテゴリの偏りにのみ着目しており，カテゴリで重要とされる単語とそれ以外の単語との，直接的な比較の観点が原理的には盛り込まれていないことが分かる．

2.2　単語の出現頻度の分布構造に着目する基準

単語間の関係を考慮するにあたり，「ある単語の意味は，共起する単語の組合せにより理解できる」という直感的な指摘[11)] をもとに，同一文書に出現する単語の出現頻度の分布構造に着目する基準が提案されている[11),12)]．具体的には，ある単語の出現する文書集合における単語の出現頻度の分布と，文書全体での単語の出現頻度の分布の差を評価し，この差が大きくなる単語を重要とする基準となっている．分布の差の評価にあたっては，文献 11) ではカルバック・ライブラー情報量[17)] を，文献 12) では χ^2 適合度統計量[18)] をもとにした基準が用いられている．

⋆1 文書中に多様な話題が含まれるトピックモデルの検討も考えられるが，今回の検討対象である医療関連文書における影響は確認できなかったため，1 文書あたりに割り当てられるカテゴリは 1 つとして扱う．

本稿では，形式が簡便である χ^2 適合度統計量について基準の定義を示し，選定にあたって重視される単語の性質を概観する[*1]．

χ^2 適合度統計量

χ^2 適合度統計量を用いる基準 $\chi^2(w)$ は，ある単語 w について共起する単語 w_p の共起頻度と各々の単語の出現確率を用いて，

$$\chi^2(w) = \sum_{w_p \in W_{pair}(w)} \frac{(N_{w,w_p} - N_w P(w_p))^2}{N_w P(w_p)} = \sum_{w_p \in W_{pair}(w)} \frac{\left(N_{w,w_p} - NP(w)P(w_p)\right)^2}{NP(w)P(w_p)} \quad (3)$$

と定義される．この式から N_{w,w_p} と $NP(w)P(w_p)$ の差の絶対値が大きい場合に $\chi^2(w)$ の値が大きくなる傾向にあることが分かる．この差が正の値をとる場合は単語間で依存的な関係，負の値をとる場合は単語間で排他的な関係にそれぞれあるといえ，式 (3) はこれら両方の関係を重視する基準となっている．

ここで，式 (3) を低頻度に出現する単語に適用することを想定すると，そのような単語の出現する文書の数はそもそも少ないため，それらの文書に含まれる単語の出現頻度の分布は文書全体の一部しか反映されない．そのため，低頻度に出現する単語であるほど上記の分布の差は大きくなる性質があり，低頻度に出現する単語がより重要と評価する傾向がある[*2]．この傾向は，文献 12) の共起する単語の種類が少ない単語に着目する基準という主張と合致する．

この傾向に対して，文献 12) では，前処理として出現頻度の 30%程度を占める高頻度の単語に限定したり，類似すると予想される単語のクラスタ分けによる統合を行ったりと，単語の種類数を減少させるいくつかの工夫が盛り込まれている．一方で，形式からも明らかなようにカテゴリの観点は含まれていない．

3. 既存法の課題の整理と解決策の提案

本検討の医療関連文書を対象とする網羅的な単語辞書の構築の要件に，各々の分類（カテゴリ）の観点が含まれることと，高頻度に出現する単語に加えて低頻度にしか出現しない単語をあわせて評価できることをあげた．これに対し，カテゴリと単語の関係を考慮する基準は，1 つ目の要件を満たしているが，2 つ目の要件においては適切でないとの指摘がある[*3]．この基準は単語間の差違を直接的に評価するものではないことから，単語間の関係の考慮による高度化の余地が残されている．一方，単語間の関係を考慮するにあたり出現頻度の分布構造に着目する基準では，1 つ目の観点の導入が必要であり，かつ 2 つ目の要件について課題が残る．

本章では，最初に出現頻度の分布構造に着目する χ^2 適合度統計量にカテゴリの観点を導入した基準を示し，この基準の課題を整理する．さらに，単語間の関係を考慮するにあたり新しい着目点を導入し，高頻度に出現する単語と低頻度に出現する単語をあわせて評価する基準を提案する．

まず，式 (3) の形式へのカテゴリの観点の導入を考える．N_{w,w_p} で記述される共起頻度（観測度数）はある単語の含まれる文書集合に対する単語の出現頻度の分布を意味することから，観測度数についてはカテゴリに限定した共起頻度を用いればよい．一方，$NP(w)P(w_p)$ で記述される共起の期待値（理論度数）は文書全体での単語の出現頻度の分布を表していることから，カテゴリに依存せず独立と仮定した場合の理論度数を用いればよいと考えられる．これは，当該カテゴリの文書数と文書全体での単語の出現確率の積で求められる．

具体的には，表 1 で定義した記号を用いて，

$${\chi_c}^2(w,c) = \sum_{w_p \in W_{pair}(w,c)} \frac{\left(N_{w,w_p,c} - N_c P(w)P(w_p)\right)^2}{N_c P(w)P(w_p)} \quad (4)$$

と定式化される．しかしながら，2 章の式 (3) と同様に出現頻度の分布構造に着目する基準であるため，この形式は低頻度に出現する単語がより重要と評価する傾向を持つ[*4]．

次に，単語間の関係を考慮するにあたり単語の主題性に着目する新しい基準を提案する．ここで主題性の意味する内容であるが，カテゴリに対応する文書の中で重要な単語を主題となる単語と見なした場合，この主題となる単語は他の単語を用いて詳述されるものと考えられる．この考えに基づくと，2.2 節で述べた文献 12) の主張とは逆に，共起する単語の種類が多い単語である方がより主題性のある重要な単語と評価することになる．一方で，高頻度

⋆1 なお，文献 11) の基準は選定される単語において χ^2 適合度統計量と類似する傾向が見られた．

⋆2 詳細は 4 章の効果検証で述べるが，実際のデータに式 (3) を用いる単語ランキング方式を適用すると，当該カテゴリの出現頻度が 1 である極端に出現頻度の低い単語が最上位にあげられた．

⋆3 なお，この具体例を 4.4.3 項で示す．

⋆4 詳細は 4 章の効果検証で述べるが，実際のデータに式 (4) を用いる単語ランキング方式を適用すると，当該カテゴリの出現頻度が 1，それ以外のカテゴリでの出現頻度が 0 となる，極端に出現頻度の低い単語が最上位にあげられた．

に出現する単語の中で主題とならないいわゆる一般的な単語も，出現頻度の高さを起因として偶発的により多くの単語と共起することになる．この一般的な単語との共起である可能性を排除するために，交互作用の効果[18]，すなわち共起する単語の観測度数と個々の単語が独立に共起する期待値の差を用いて単語間の関係を評価することで，回避可能となると考える．

以上の考えに基づいた単語の主題性に着目する基準は，各々の単語の組合せについて交互作用の効果を算出し，単語ごとに共起する単語の交互作用の効果を加算した値を評価値として用いることで実現される．具体的には，カテゴリと単語の関係に加え単語間の交互作用の効果を用いる単語の重要度評価基準 $CI(w,c)$ は，

$$CI(w,c) = \sum_{w_p \in W_{pair}(w,c)} \left(N_{w,w_p,c} - N_c P(w) P(w_p) \right) \tag{5}$$

と定式化される．

これにより，多くの種類の単語と共起する単語ほど，加算対象となる評価値の数を多く持つことになる．また，各々の評価値の重みについては交互作用による共起頻度の高い単語との関係を重視し，独立関係にある単語は加算の対象とせず，さらに，排他的な関係にある単語はペナルティとして扱うことを意味する[⋆1]．これにより，低頻度に出現する単語が他の単語と共起したとしても共起頻度の値が低く抑えられるため，出現頻度の分布構造に着目する基準とは異なり低頻度の単語を重視する傾向は弱くなるといえる．

単語ランキングを行う場合は，既存の方式と同様に式 (5) の定義に従って単語ごとに基準値を算出し，降順に並べ直したランキングを提示する．式 (5) の形式により定義された単語の主題性に着目する単語評価基準を，カテゴリに基づく交互作用付き単語対法（category-oriented word pair interaction method；以下 CI もしくは提案法）と呼ぶ[⋆2]．

4. 実データによる単語評価基準の効果検証

本章では，カテゴリに対応する疾患の分野に分けられた医療文書データを対象に，2 章と 3 章で取り上げた単語評価基準を用いた単語ランキング方式の導入による効果検証を行う．具体的には，カテゴリで重要とされる単語をより先に選定対象としてあげることが，各基準を用いることで可能となるかを比較検証することを目的とする．

以下，評価の具体的な方法および，検証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，各基準の比較結果について述べる．

4.1 評価方法

検証には，疾患の分野にカテゴリ分けされた医療文書データと，表記の統一を目的として医療の専門家により定められた標準病名データを用いる．医療文書データはカテゴリで重要とされ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カテゴリで重要とされる単語として扱う．

具体的な手順としては，疾患の分野をカテゴリと見なし，文書データに含まれる単語を取得した後に単語評価基準を適用し，カテゴリごとに選定順に単語を提示する方式を用いる．このランキング結果に対し，上位ランキングに含まれる標準病名の数を算出する．なお今回の検証にあたっては，概要把握のための単語抽出技術の評価で行われるような，出現頻度による単語の制限は行わない．

単語ランキングには，カテゴリを考慮する単語ランキングの既存法である情報利得の基準を用いる方式（information gain measure；以下，IG）と相互情報量の基準を用いる方式（mutual information measure；以下，MI），および式 (3) の χ^2 適合度統計量の基準をカテゴリ別に適用する方式（以下，CHI），式 (3) にカテゴリの観点を導入した式 (4) を用いる方式（以下，CHIc）を取り上げる．これらと，3 章で提案した CI 法（以下，提案法もしくは CI）の比較を行う．

各基準の比較において，最初に具体的な単語の例をもとに各基準で重視される単語の傾向を確認する．具体的には，単語ランキングの最上位にあげられる単語を抽出し，それらの単語の出現頻度の傾向の違いを比較する．

次に，各基準の業務効率化についてゲインチャート[21]を用いた比較評価を行う[⋆3]．ゲインチャートは，カテゴリに対応する文書で出現する標準病名の全数に対するランキング上位にあげられる標準病名の割合（以下，ゲイン率）を算出し，それらを単語ランキング方式

⋆1 ただし，交互作用の効果の絶対値，すなわち排他的な関係にある単語をペナルティとして扱わず加算の対象とする評価法でも，今回の実験による傾向の違いは見られなかった．

⋆2 なお，連接する単語の組合せ w_1，w_2 の重要度を評価する基準として，以下の

$$z(w_1,w_1) = \frac{P(w_1,w_2) - P(w_1)P(w_2)}{\sqrt{P(w_1)P(w_2)(1-P(w_1)P(w_2))}}$$

という類似する形式の評価法が提案されているが[19]，これも式 (3) と同様に，低頻度に出現する単語を重要視する傾向があることが指摘されている[20]．

⋆3 データマイニング分野において顧客ターゲティングと呼ばれる課題での性能評価によく使われるもので，リフトカーブとも呼ばれる[22]．

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種類の単語に対して，単語ランキング方式を用いて選定対象とする単語の数を制限した場合の効果試算に対応し，単語ランキングの適用により，多くの標準病名が上位にあげられるようになることを確認する．

さらに，提案法と既存法の単語ランキングの差違を確認するために，不一致となる単語の具体例を用いて比較を行う．ここでは，提案法と既存法により異なる単語が選定される割合により傾向の違いを確認し，同一ランキングでの異なる単語の例から定性的な評価を行い，提案法によって一般的な単語がランキングの下位に下げられた例を示す．

なお，上記のゲイン率の算出にあたっては，作業効率を確認するためにカテゴリ全体で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，カテゴリで重要とされる単語はカテゴリ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[23]を取り上げる．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．

文書データの概要を**表 2** に示す．カテゴリに対応する疾患の分野の数は 23 個で，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 1 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．共起する単語は同一文書に出現する単語と単語の組合せのすべてとした．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，単語のすべての組合せの 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[24]（以下，病名マスタ）を用いる．これは，レセプトに記述する傷病名の標準化に向け電子カルテなどに利用されることを目的に，医療の専門家により構築されたものである[⋆2]．

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．

表 2　医療関連データの疾患の分野ごとの文書数，異なり語数，異なり単語対数
Table 2　The numbers of document, unique term and unique term pair according to disease areas in experiments of medical document data.

疾患の分野	文書数	異なり語数	異なり単語対数
栄養障害	115	2,127	261,560
内分泌・代謝疾患	331	3,359	746,125
消化器疾患	351	3,377	521,200
肝・胆道疾患	160	2,131	323,262
筋骨格・結合組織疾患	283	2,915	478,985
呼吸器疾患	250	3,372	766,097
耳鼻咽喉疾患	123	1,641	139,359
眼疾患	172	1,644	161,175
歯科・口腔疾患	55	1,511	213,684
皮膚疾患	217	2,354	239,063
血液疾患と腫瘍	354	3,297	554,811
免疫；アレルギー疾患	140	2,442	452,546
感染症	756	5,071	914,607
神経疾患	249	3,623	736,016
精神疾患	179	3,141	641,161
循環器疾患	317	3,607	901,222
泌尿生殖器疾患	235	2,663	449,734
産婦人科疾患	233	3,372	735,262
小児疾患	806	5,840	1,204,932
物理的要因による疾患	59	1,639	199,800
特定な諸分野	109	2,460	289,678
臨床薬理	78	1,663	167,363
中毒	29	1,206	140,892

表 3 に，疾患の分野ごとに含まれる標準病名の数，異なり語数の再掲，および標準病名の割合を示す．割合を示す列の太字は最大値，下線は最小値を表している．標準病名の割合から，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 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 3 つのテーブルが用意されており，標準病名データとして「索引テーブル」を用いることにする．これは，病名基本テーブルに含まれない同義語や，病名基本テーブルに含まれる語に修飾語を付与した語，およびフリガナが含まれ，3 つのテーブルの中で最も多くの単語が登録されている．

表 4　単語ランキング方式ごとの最上位にあげられた単語 5 件の例
Table 4　Examples of top 5 terms using the term measures.

提案法			IG			MI			CHI			CHIc		
単語	N_c	$N_{\bar{c}}$	単語	N_c	$N_{\bar{c}}$	単語	N_c	$N_{\bar{c}}$	単語	N_c	$N_{\bar{c}}$	単語	N_c	$N_{\bar{c}}$
欠乏症	68	56	欠乏症	68	56	欠乏症	68	56	BUN	1	41	イソプレニル	1	0
ビタミン	63	149	ビタミン	63	149	GI	19	5	albus	1	0	イソプレン	1	0
摂取	53	423	栄養	49	219	栄養素	21	8	めまい	1	78	エポキシド	1	0
栄養	49	219	栄養素	21	8	RDA	12	0	アプローチ	1	72	クリスマス	1	0
代謝	44	469	GI	19	5	ミネラル	15	9	アルギニン	1	5	グルタミルカルボキシラーゼ	1	0

表 3　疾患の分野ごとに含まれる標準病名の数と異なり語数
Table 3　The numbers of standard disease term and unique term according to disease areas in experiments of medical document data.

疾患の分野	標準病名の数	異なり語数	標準病名の割合
栄養障害	1,054	2,127	0.496
内分泌・代謝疾患	889	3,359	0.265
消化器疾患	1,553	3,377	0.460
肝・胆道疾患	1,097	2,131	0.515
筋骨格・結合組織疾患	1,460	2,915	0.501
呼吸器疾患	1,422	3,372	0.422
耳鼻咽喉疾患	887	1,641	0.541
眼疾患	907	1,644	0.552
歯科・口腔疾患	877	1,511	**0.580**
皮膚疾患	1,146	2,354	0.487
血液疾患と腫瘍	1,465	3,297	0.444
免疫；アレルギー疾患	1,096	2,442	0.449
感染症	1,968	5,071	0.388
神経疾患	1,641	3,623	0.453
精神疾患	1,130	3,141	0.360
循環器疾患	1,522	3,607	0.422
泌尿生殖器疾患	1,320	2,663	0.496
産婦人科疾患	1,519	3,372	0.450
小児疾患	1,780	5,840	0.305
物理的要因による疾患	809	1,639	0.494
特定な諸カテゴリ	1,084	2,460	0.441
臨床薬理	677	1,663	0.407
中毒	604	1,206	0.501

4.4　単語評価基準を用いるランキング方式の比較

本節では，各々の単語評価基準を用いるランキング方式（以下，単語ランキング方式）により上位にあげられる単語の傾向を把握し，その結果をもとに既存法と提案法の比較を行う．

4.4.1　単語評価基準で重視される単語の傾向

各単語ランキング方式で最上位にあげられた単語の出現頻度を確認し，各々の基準で重視される単語の傾向を把握する．**表 4** は，「栄養障害」のカテゴリを対象に最上位 5 件の単語を抽出した結果である．表中の N_c は「栄養障害」における単語の出現頻度，$N_{\bar{c}}$ は「栄養障害」以外のすべてのカテゴリの文書における出現頻度を意味する．

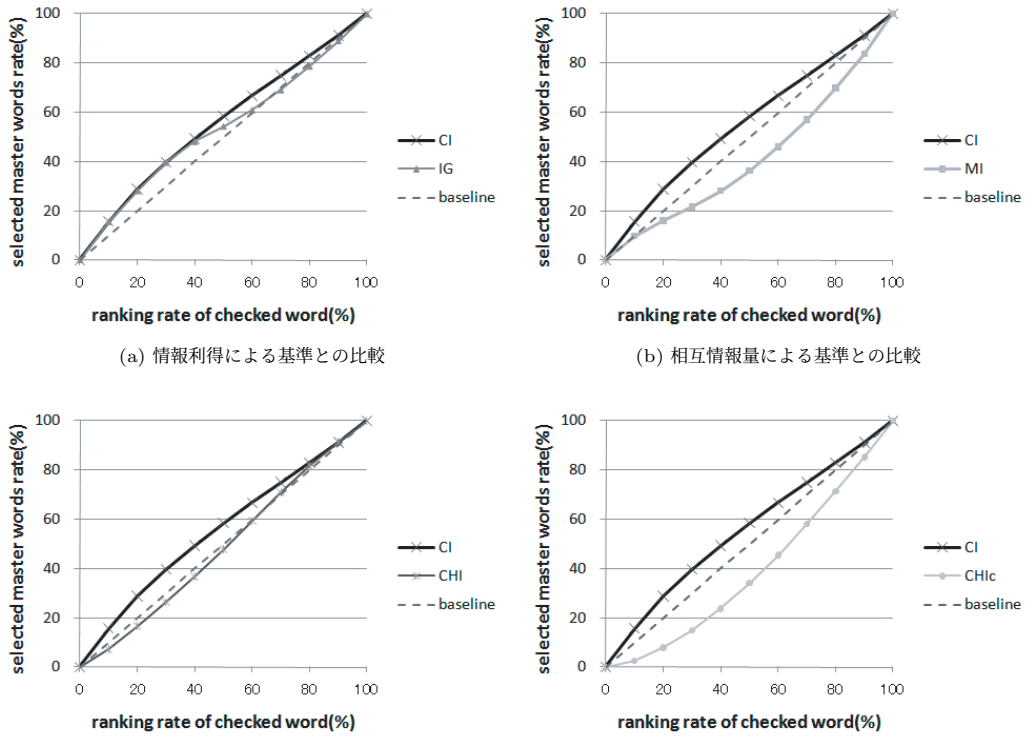
この表から，提案法は比較的高頻度に出現する単語，IG は高頻度の単語と低頻度の単語の両方を上位にあげる傾向があるといえる．MI は提案法や IG と類似する単語が上位にあげられているが，これらの中では比較的低頻度に出現する単語が上位にあげられているといえる．一方，CHI と CHIc では当該カテゴリにおける単語の出現頻度が 1 と極端に低い単語を上位にあげる傾向があることが分かる．このような単語は信頼に足る結果であるかの判断が困難であり，少なくとも最優先で選定対象とすべきとは考えられない．

4.4.2　ゲインチャートによる比較

各々の単語ランキング方式による業務効率化の効果を検証するために，**図 1** に示すゲインチャートを用いた比較を行う．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．なお，ゲイン率は

$$Gain(m,r) = \sum_{c \in C} \frac{w_{st}(c,m,r)}{w_{st}(c)} \Big/ |C|$$

の定義により算出する．ここで，$w_{st}(c)$ はカテゴリ c の文書に含まれる標準病名の異なり



(a) 情報利得による基準との比較　(b) 相互情報量による基準との比較

(c) χ^2 適合度統計量による基準との比較　(d) カテゴリを考慮した χ^2 適合度統計量による基準との比較

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.

語数，$w_{st}(c, m, r)$ はカテゴリ c の文書集合に適用した基準 m による，単語ランキング上位 r%に含まれる標準病名の異なり語数，C はカテゴリの集合，$|C|$ は C の要素の数とする．

図 1 (a) の結果から，ランキング上位 40%までは提案法と IG は同等の 50%程度のゲイン率が得られているが，それ以上になるとゲイン率の差が開き始めることが分かる．また，IG はランキング上位 60%以降で無作為に選定した場合とほぼ変わらなくなっている．次に，図 1 (b)，(c)，(d) の結果から，MI，CHI，CHIc のそれぞれでゲイン率が無作為の場合と同等もしくは低下する傾向があることが分かる．なお，4.4.1 項で，MI の最上位にあげられる単語は CI や IG と類似することを示したが，図 1 (b) を詳細に確認すると，上位 10%程度は無作為と同等であっても，それ以降の上位 20%程度から適切な評価ができていないことが分かる．

以上から，今回の検証の目的である単語辞書の構築のための単語評価基準の適用による業務効率化の実現においては，提案法と IG のみ期待できるといえる．

4.4.3　提案法と既存法で選定される単語の差違

本項では，単語辞書の構築業務に効率化が期待できる提案法と IG を対象に，ランキング上位にあげられる単語の違いをもとに比較を行う．具体的には，これらのランキング上位に含まれる単語の中で，異なる単語の割合の確認と具体的な単語の違いを用いた定性的な比較

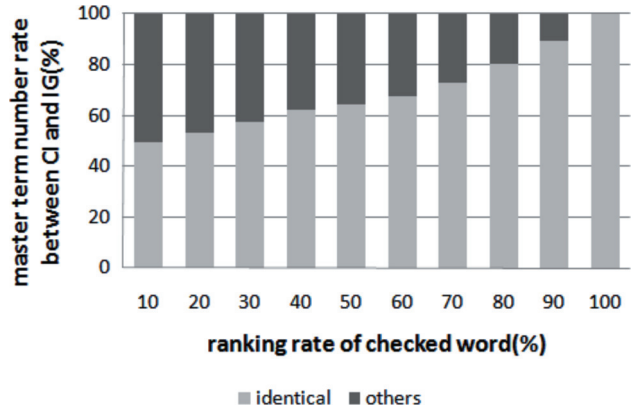


図 2　提案法と IG により選択された標準名称の一致割合

Fig. 2　Identical master term number rates between the ranks selected by the proposal method and IG.

を行う．

まず，ランキング上位にあげられた単語について，一致した単語とそれ以外の割合を算出した結果を**図 2** に示す．これは，提案法と IG の両方で上位にランキングされた単語と，提案法でのみ上位にランキングされたものに分け，それらの割合を算出した結果である．identical が一致した単語，others は提案法でのみ選択された単語，横軸が単語ランキング，縦軸が単語の種類の割合を意味する．ゲイン率で同等であったランキング上位 40%の時点では 40%弱の異なる単語が選定され，ゲイン率で明らかな差が見られたランキング上位 60%の時点では 30%強の異なる単語が選定されている．このことから，選定される単語には傾向の違いがあることが示唆される．

次に，提案法でのみ上位にあげられた単語と，IG でのみ上位にあげられた単語の具体例を**表 5** に示す．この表は，「栄養障害」の分野を対象に提案法と IG のそれぞれのランキング上位 60%に含まれる単語の中で，一致する単語を除いてランキング最下位から 5 件抽出した例である．これらは，いずれの方式でも当該カテゴリ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，提案法の単語の主題性に着目する単語評価基準を用いることで，このような抽象的な単語はランキングの上位にあげられなかったと考えられる．

なお，上記に示した IG の性質は，文献 10) で指摘された低頻度に出現する単語を評価する性能が劣ることの実例といえる．

表 5　提案法と情報利得で上位にあげられた単語の違いの例

Table 5　Examples of different terms ranked by the proposal method and IG.

提案法のみ選択			IG のみ選択		
単語	N_c	$N_{\bar{c}}$	単語	N_c	$N_{\bar{c}}$
アナフィラキシー	2	35	適当	1	53
落屑	1	27	ろう	1	0
光線	1	42	エナメル	1	5
弾性	1	34	カドミウム	1	5
混濁	1	38	シリコン	1	2

5. 考　　察

本章では，単語ランキング方式を適用する単語選定業務で想定される状況を取り上げ，単語評価基準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．

実業務においては，与えられた文書データをもとにカテゴリで重要とされる単語の辞書を新規に構築する場合のほかに，すでに単語辞書が構築済みでこの辞書の高度化，すなわちまれな疾患やまれな表記も含めて重要な単語を辞書に追加する場合が考えられる．このような単語辞書の有無による状況の違いにおける，単語評価基準に求められる特性を考察する．

まず，既存の単語辞書が存在しない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を漏らさず選定の対象とすべきで，このような単語を上位にあげる特性が期待される．一方の，既存の単語辞書に対して高度化が求められる段階においては，まれな疾患やまれな表記である単語も含めてカテゴリで重要とされるものを選定の対象とする必要がある．しかしながら，初期段階において選定の対象から漏れた単語とは，出現頻度が低く信頼性についても疑問が残るため直接的な評価は期待できない．そこで，カテゴリに関連せず出現する一般的な単語をランキング下位にし，一般的でない単語をランキング上位にする特性が期待される．

これらの特性に対し，情報利得の基準では，原理的に単語間の比較の観点がなく 4.4.3 項で示したとおり低頻度に出現する一般的な単語をランキング下位にする特性は持たない．また，χ^2 適合度統計量をもとにした基準では，一般的な単語をランキング下位にする特性は，基準の形式から有しているといえる．しかしながら，高度化の段階であってもきわめて低頻度に出現する単語は膨大に存在し，4.4.1 項で示したようにきわめて低頻度の単語を上位に

あげるため，適切なランキングとなることが期待できない．これらに対し，提案する単語の主題性に着目する基準では，4.4 節で示したとおり，初期段階で求められる高頻度の単語を重視し，一方の高度化の段階で求められる一般的な単語をランキング下位にする効果が認められた．

以上より，提案法は医療関連文書からの単語辞書構築の状況の違いにおける，単語ランキングに求められる特性を有する基準であることが示された．

6. まとめと今後の課題

本稿では，医療関連文書の電子化を対象とした単語辞書の構築における，支援技術の検討を目的として，単語評価基準に求められる要件の整理と既存法の課題の整理，および単語の主題性に着目する基準による解決策の提案を行った．医療に関連する文書データを用いた各基準の比較結果から，情報利得の基準と提案法によってカテゴリで重要とされる単語が上位にあげられ，業務効率化に貢献することを示した．さらに，単語ランキング方式を適用する単語選定業務を想定した状況から基準に求められる特性を検討し，提案法のみがこの特性を満たすことを示した．

単語選定の支援を行うにあたっては，カテゴリで重要とされ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いっそう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あげられる．今回は，カテゴリ分けされた医療関連文書の検討を行ったが，専門書や科学技術論文とは異なる多様な話題を対象とする文書からの，単語の重要語評価基準の検討も課題としてあげられる．

謝辞　本研究の機会を与えていただいた株式会社 NTT データ技術開発本部上島康司部長，論文化にあたり貴重な事例，意見をいただいた株式会社 NTT データ第一金融事業本部並河悠介氏，日本電信電話株式会社コミュニケーション科学基礎研究所坂野鋭博士，ならびに日頃議論していただいている同僚諸氏に感謝いたします．

参　考　文　献

1) 小沼　敦：医療保険制度改革の動向—平成 18 年度改革法案の主要論点，国立国会図書館調査と情報，Vol.519 (2006).
2) 日経 BP：保険業界における不払い対策—ビジネスルールを “見える化” する，金融 IT イノベーション，Vol.3 (2008).
3) 日経 BP：疾患名と薬品名の適合をチェックし処方の際の入力ミスを未然に防止，日経ヘルスケア 21，pp.99–101 (2006).
4) 日経 BP：【プロジェクト完遂の軌跡】第一生命保険不払に撲滅目指しシステムを刷新 全社プロジェクトを IT 部門が主導，日経コンピュータ，No.696, pp.64–68 (2008).
5) 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).
6) 土屋文人：薬剤のリスク管理—医薬品関連医療事故防止のために，第 127 回日本医学会シンポジウム記録集，日本医学会 (2004).
7) 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).
8) Sebastiani, F.: Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, Vol.34, pp.1–47 (2002).
9) 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).
10) Yang, Y. and Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization, *Proc. ICML-97, 14th International Conference on Machine Learning*, Nashville,TN, pp.412–420 (1997).
11) Hisamitsu, T., Niwa, Y. and Tsujii, J.: A method of measuring term representativeness: baseline method using co-occurrence distribution, *Proc. 18th conference on Computational linguistics-Volume 1*, Association for Computational Linguistics Morristown, NJ, USA, pp.320–326 (2000).
12) 松尾　豊，石塚　満：語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム，人工知能学会論文誌，Vol.17, pp.217–233 (2002).
13) Franzi, K. and Ananiadou, S.: The C-value/NC-value method for ATR, *Journal of NLP*, Vol.6, No.3, pp.145–179 (1999).
14) 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).
15) 三浦康秀，増市　博：部分文字列のパープレキシティを利用した低頻度専門用語抽出，情報処理学会研究報告，pp.139–144 (2007).
16) 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).
17) Cover, T. and Thomas, J.: *Elements of information theory*, Wiley (1991).
18) 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).
19) Barnbrook, G.: *Language and computers: A practical introduction to the computer analysis of language*, Edinburgh Univ Pr (1996).
20) 小池生夫，井出祥子，河野守夫，鈴木　博，田中春美，田辺洋二，水谷　修：応用言語学事典 (2003).
21) 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).
22) 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).
23) 福島雅典（総監修），日経メディカル（翻訳・編集）：メルクマニュアル第 17 版日本語版，日経 BP 社 (1999).
24) 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).

(平成 21 年 11 月 19 日受付)
(平成 22 年 1 月 6 日再受付)
(平成 22 年 1 月 13 日採録)

末永　高志

1997 年早稲田大学理工学部経営システム工学科卒業．1999 年同大学大学院理工学研究科修士課程修了．同年株式会社 NTT データ入社．パターン認識，データ分析技術の実用化研究に従事．人工知能学会，電子情報通信学会各会員．

松永　　務

1988 電気通信大学大学院通信工学専攻修士課程修了．株式会社 NTT データ技術開発本部主任研究員．博士（工学）．データ/テキストマイニングの研究・開発に従事．

関根　　純（正会員）

1982 年東京大学大学院工学系研究科計数工学専攻修士課程修了．同年日本電信電話公社．2005 年 NTT データ技術開発本部副本部長．博士（工学）．データベース，BI の研究開発に従事．日本データベース学会，ACM 各会員．

村松　正明

1982 千葉大学医学部卒業，1989 東京大学大学院医学研究科修了，医学博士．ゲノム情報の臨床応用に向けた SNP（一塩基多型）解析に関する研究に従事．東京医科歯科大学難治疾患研究所教授．ヒュービットジェノミクス株式会社取締役．