複数の作成者情報付き文書からの専門用語抽出

立石健二[†] 久寿居大[†]

本稿では，人や部門といった作成者情報が 1 文書に複数付与された文書から専門用語を抽出する手法を提案する．企業内で誰がどの技術，製品，顧客に詳しいといった社員の専門領域をデータベース化して検索しやすくするような情報共有システムを開発するためには，製品名，技術名，顧客名といった専門用語を企業内文書から自動的に抽出する専門用語抽出が必要である．本稿では，製品名等の専門用語はそれを管理/担当する少数のカテゴリ（作成者）が存在するという仮定に基づき，少数のカテゴリに関連が深い用語を専門用語として抽出する．本方式は，カテゴリを利用して専門用語を抽出する従来方式を利用するが，さらに，1 文書に複数のカテゴリ情報が付与された文書にも対応できるように改良する．評価実験により，企業内文書において提案手法を従来方式と組み合わせることよって専門用語の抽出精度を向上できることを示した．

Terminology Extraction from Documents Labeled Plural Categories by Authors

Kenji Tateishi[†] and Dai Kusui[†]

This paper proposes a method for terminology extraction from documents where plural categories by authors are labeled. It is important to extract many kinds of terminologies such as product names, technical names, and customer names when we develop a knowledge-sharing system that stores expertise of employees. Our proposed method extracts a term as a terminology if there is totally a few authors labeled on the documents including the term. The proposed methoed is based on the assumption that a terminology such as a product name tends to have a few employees or departments responsible for the terminology. The proposed method utilizes a previous method that extracts terminologies using a category labeled on a document, and improves it in order to address plural categories on a document. The experimental result shows that the system which combines the proposed method with previous methods achieves high accuracy on two document sets.

1. はじめに

企業内で誰がどの技術，製品，顧客に詳しいといった社員の専門領域を検索できるようにする情報共有システムは，企業が大規模になるにつれて業務を円滑に進めるうえで重要な手段となる．このようなシステムを開発するためには，部門や人と，技術，製品，顧客等の専門領域との関係をあらかじめデータベース化する必要がある．このデータベースを人手で日々更新する運用形態は，社員への負担が大いため定着が難しく，社内の報告文書等の企業内文書から自動的に構築できることが望ましい．そのためには，製品名，技術名，顧客名といった専門用語を企業内文書から自動的に抽出する専門用語抽出が必要である．

† NEC インターネットシステム研究所
Internet Systems Research Laboratories, NEC Corporation

従来の専門用語抽出方式は大きく（従来 1）抽出パターンを用いて固有表現を抽出する方式[1)]，(従来 2) tf/idf 等の頻度と文書に対する用語の出現頻度の偏りを用いる方式[2)]，(従来 3) エントロピーやカイ二乗値[3)]を利用して文書に付与されたカテゴリに対する用語の出現頻度の偏りを用いる方式が提案されている．しかしながら，(従来 1) や（従来 2）の手法は，製品等の専門用語抽出には必ずしも精度が十分でなく，特に高頻度な不要語が残る問題があった．高頻度の不要語は，情報共有システムでは多くの入力キーワードと結び付き性能低下の原因となりやすい．また，(従来 3) の手法は，1 文書に 1 つのカテゴリが付与された文書を前提としており，1 文書に複数のカテゴリが付与された企業内文書に対しては適用できない問題があった．

本稿では，企業内文書に多数存在する人や部門といった作成者情報がカテゴリとして 1 文書に 1 つ以上付与された文書から専門用語を抽出する手法を提案す

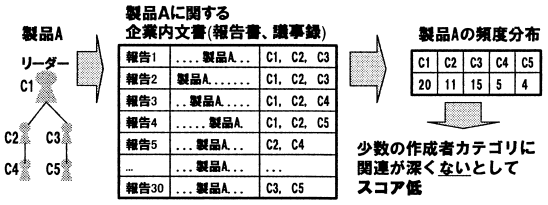

図 1 複数の作成者カテゴリ付き文書
Fig. 1 Documents labeled plural categories by authors.

る．製品名等の専門用語はそれを管理/担当する少数の作成者カテゴリ（人や部門）が存在すると考えられる（企業内では製品名等の担当者や部門はある程度決まっている）ため，少数の作成者カテゴリに関連が深い用語が専門用語である可能性が高いと考えられる．本手法は，前述の（従来 3）の手法を利用するが，さらに，1 文書に複数の作成者カテゴリ情報が付与された企業内文書にも対応できるように改良する．

ここで，企業内文書における複数カテゴリの問題とそれに対する提案手法を概説する．企業において製品等は人や部門のチーム（グループ）で管理/開発されることが多く，これらの開発状況等を記載した報告書，議事録等の企業内文書は，チームメンバの連名により記載されることとなる（**図 1** 参照）．この報告書等が連名で記載される状態が 1 文書に複数の作成者カテゴリが付与された状態を意味する．このとき，製品 A の専門用語としてのスコアをエントロピーやカイ二乗値を用いて計算するためには，作成者カテゴリごとの製品 A の出現頻度の分布を求める必要がある．しかしながら，1 文書に複数の作成者カテゴリが付与されている状態では，1 つの文書が複数の作成者カテゴリの頻度計算に用いられるため，1 文書に付与された作成者カテゴリの数が多ければ多いほど，出現頻度の分布は一様になり，エントロピーやカイ二乗値では専門用語らしくないと判定されることになる．すなわち，複数カテゴリの問題とは，チームで管理/担当される専門用語は，それを含む文書に複数カテゴリが付与され（連名で記載され）ており，1 文書に付与される作成者カテゴリの数が多いほど出現頻度の分布が一様になるため，従来方式では抽出が難しいことを指す．

提案手法は，この問題に対して，出現頻度の分布を計算する前に作成者カテゴリ選択処理を加え，各文書に付与された作成者カテゴリのうちの 1 つのみを代表カテゴリとして選択し，出現頻度の計算に用いるようにする．このときの代表カテゴリの選択基準は，各文書から最もリーダ的な役割を果たす作成者カテゴリを選ぶ．この方法により，1 文書に複数の作成者カテゴリが付与されている企業内文書から少数のリーダ的なメンバが他のチームメンバを率いて管理/担当する製品 A のような用語を専門用語として抽出できる．

評価実験では，2 つの企業内文書セットを対象として，上記の（従来 1）–（従来 3）の従来手法とそれらと提案手法を組み合わせた手法を比較した．その結果，どちらの文書セットに対しても提案手法を含む手法は最も高い抽出精度を示し，企業内文書において提案手法を従来方式と組み合わせることよって専門用語の抽出精度を向上させることができ，特に高頻度の不要語を削除できる利点があることが判明した．また，カテゴリを用いた専門用語用語抽出においてカテゴリ選択処理を用いる提案方式がカテゴリ選択処理を用いない従来方式よりも有効であることが分かった．

本稿の構成を以下に説明する．2 章では，本稿で取り扱う企業内文書の定義と，企業内文書から抽出すべき専門用語の範囲を明確化する．3 章では，本章で提案する専門用語抽出方式を従来方式と比較しながら説明する．4 章では，従来方式に 3 章の提案方式を組み合わせた場合の精度向上を評価し，5 章では従来研究について述べ，6 章でまとめる．

2. 企業内文書のモデルと専門用語

本稿が目指す社員の専門領域をデータベース化し検索できるようにする情報共有システムのために必要な専門用語抽出システムは，企業内文書を入力として専門用語リストを出力とする．

企業内文書の特徴として，図 1 で説明したように人や部門といった作成者情報がカテゴリとして 1 文書に 1 つ以上付与された文書（作成者カテゴリ付文書）が多いことがある．このような企業内文書の例として社外向けの広報記事や，研究部門の報告記事がある（4 章の評価実験の図 5，図 6 参照）．1 文書に複数の作成者情報が付与される場合とは，複数の人や部門が連名で報告書等を作成した場合が該当する．本稿では，このような作成者カテゴリ付き文書を企業内文書の代表として，専門用語抽出システムで取り扱うことにした．一方，企業内での情報共有に必要という観点からは，**表 1** のような専門用語の種類が考えられる．本稿では，これらの範囲に含まれる語を専門用語として扱うことにした．

3. 専門用語抽出方式

本稿で提案する専門用語抽出方式は，少数の作成者カテゴリ（以下，カテゴリ）と関連が深い用語を専門用語として抽出する．製品名，技術名，顧客名といった専門用語はそれを管理/担当する少数のカテゴリが

表 1 専門用語の種類
Table 1 The classes of terminologies.

種類	内容	例
製品名	製品化されている物品/物質/サービス/システムの名称	IP8800/R シリーズ，電子カルテソリューション/システム
開発成果物名	製品化前の開発成果としての物品/物質/システムの名称	分子動力計算専用サーバ，カーボンナノチューブ
技術名	製品等に使われた特定の技術領域の名称	テキストマイニング，直径制御技術
機能名	製品等に固有の機能	3D 表示，おまかせ録画
顧客企業名	製品等の顧客企業の名称	「○○」殿に導入
提携先企業名	製品等を実現するために提携した企業名	「○○」と共同で

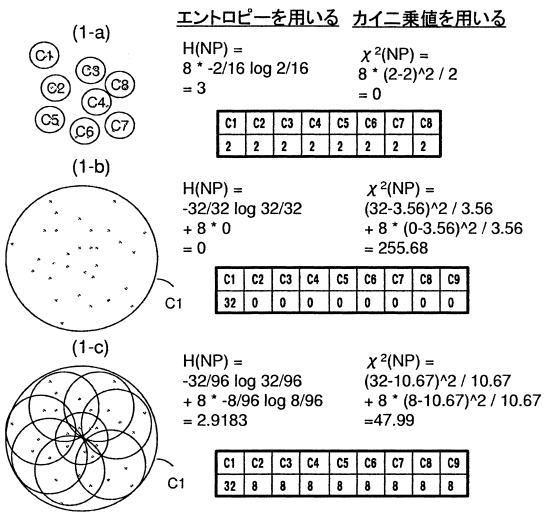

図 2 エントロピーとカイ二乗値の計算例
Fig. 2 Examples of entropy and χ^2 values.

存在する（企業内では製品名等の担当者や部門はある程度決まっている）と考えられるため，少数のカテゴリに関連が深い用語が専門用語である可能性が高いと考えられる．このような企業内文書の特徴を利用した方法を従来手法と組み合わせることにより専門用語の抽出精度向上が期待できる．

本方式は，1 章で述べた従来手法の（従来 3）の文書に付与されたカテゴリに対する用語の出現頻度の偏りを用いる手法を利用する．しかしながら，従来手法は 1 文書に 1 つのカテゴリが付与されていることを前提としており，そのまま適用したのでは 1 文書に複数のカテゴリが付与された文書を取り扱う場合に抽出精度が低下する恐れがある．その問題に対応できるように改良したものが提案方式である．以下，従来方式と，その課題，および改良方式について具体的に説明する．

3.1 従 来 方 式

少数のカテゴリと関連が深い用語を抽出する方式としてエントロピーを用いる方法とカイ二乗値を用いる方法がある．以下それぞれを簡単に説明する．

エントロピーを用いる方法は，用語のカテゴリに対する出現の偏りをエントロピー関数を用いて計算し，偏りが大きい用語を専門用語とする．用語のエントロピーは，下記の式で定義され，この値が大きいほどカテゴリに対する用語 NP_i の偏りが小さく，逆に小さいほど少ないカテゴリに用語 NP_i が偏って出現していることになる．

$$H(NP_i) = \sum_j -p(NP_i|c_j)\log_2 p(NP_i|c_j)$$

$$p(NP_i|c_j) = \frac{f(c_j \cap NP_i)}{\sum_k f(c_k \cap NP_i)}$$

ここで，$p(NP_i|c_j)$ はカテゴリ c_j が付与されている文書に NP_i が出現する確率であり，$f(c_j \cap NP_i)$ はカテゴリ c_j が付与された文書に NP_i が出現する回数である．

図 2 はある用語 NP の分布を示した図である．各点が NP の出現を，各円がカテゴリを示す．各点がどの円内に存在するかにより各点がどのカテゴリに所属するかが分かる．この図の (1-a) の NP のエントロピーは NP が c_1 から c_8 のカテゴリについてそれぞれ 2 回出現していることから (1-a) の $H(NP) = 3$ となる．また，図 2 の (1-b) の NP のエントロピーは NP が c_1 で 32 回，c_2 から c_9 で各 0 回出現していることから (1-b) の $H(NP) = 0$ となる．

次に，カイ二乗値を用いる方法[3] は，用語のカテゴリに対する出現の偏りをカイ二乗値を用いて計算し，偏りが大きい用語を専門用語とする．用語のカイ二乗値は，下記の式で定義される．カイ二乗値は，カテゴリごとの出現頻度が期待値からどの程度大きく離れているかを意味し，この値が小さいほどカテゴリに対する用語 NP の偏りが小さく，逆に大きいほど NP が少ないカテゴリに偏って出現していることになる．

$$\chi^2(NP_i) = \sum_j \frac{[(f(c_j \cap NP_i) - E(j,i)]^2}{E(j,i)}$$

$$E(j,i) = \frac{\sum_k f(c_k \cap NP_i) \times \sum_l f(c_j \cap NP_l)}{\sum_k \sum_l f(c_k \cap NP_l)}$$

ここで，$f(c_j \cap NP_i)$ はカテゴリ c_j が付与された文書に NP_i が出現する回数であり，$E(j,i)$ は $f(c_j \cap NP_i)$ の期待値であり，$E\{f(c_j \cap NP_i)\}$ の略表記である．

図 2 の (1-a) の NP のカイ二乗値は，今，c_1 から c_8 の NP の期待値がどれも等しいと仮定すると $E\{f(c_j \cap NP)\} = \frac{2\times 8}{8} = 2$ となる．この場合，カイ二乗値は (1-a) の $\chi^2(NP) = 0$ となる．また，図 2 の (1-b) の NP のカイ二乗値は，今，c_1 から c_9 の NP の期待値がどれも等しいと仮定すると $E\{f(c_j \cap NP)\} = \frac{32}{9} = 3.56$ となる．この場合，カイ二乗値は (1-b) の $\chi^2(NP) = 255.68$ となる．

3.2 従来方式の課題

従来方式の課題は，チームで管理/担当される専門用語は，それを含む文書に複数カテゴリが付与され（連名で記載され）ており，1 文書に付与される作成者カテゴリの数が多いほど出現頻度の分布が一様になるため，抽出が難しいことである．この点について以下詳細に説明する．

従来の方法では，1 文書に複数のカテゴリが付与されている文書に対しては，文書に含まれる NP が複数のカテゴリで数えられるため，文書に付与されたカテゴリの数が多いほど NP はカテゴリに対して一様に出現するとして扱う．これは，ある文書に c_1 から c_n の n 個のカテゴリが付与されている場合，頻度計算上は n 個の文書それぞれに c_1 から c_n のカテゴリが 1 つずつ付与されている場合と同じに扱うからである．

たとえば図 2 の (1-c) は，1 つの文書に複数のカテゴリが付与されている場合の NP の分布の例である．NP の総出現頻度は 32 回で，それらが 9 つのカテゴリで出現している．この場合エントロピーを用いる方法では，NP がカテゴリ c_1 に対して 32 回，c_2 から c_9 に対して各 8 回出現していることから (1-c) の計算式のようになり $H(NP) = 2.918$ となる．カイ二乗値を用いる方法では，今，c_1 から c_9 の NP の期待値が等しいと仮定すると $E\{f(c_j \cap NP)\} = \frac{32+8\times 8}{9} = 10.67$ となり $\chi^2(NP) = 47.99$ となる．これらのスコアは (1-b) よりも高く（カイ二乗値の場合は低く），閾値によっては (1-b) は専門用語となるが (1-c) は専門用語とならない場合がある．

しかし，このような一様に出現すると扱われる用語は，少数のカテゴリに偏って出現すると解釈することもできる．上記の例では，(1-c) はすべての NP がカテゴリ c_1 に所属しており，NP はカテゴリ c_1 に偏って出現するとも解釈できる．企業内では複数の部門や人がチームを組み，リーダ的な存在が同一のチームの他の部門や人を率いて専門用語を管理/担当する場合が多く存在するため，この場合でも少数のカテゴリに偏って出現すると解釈するほうがより高精度で専門用語抽出が可能と考える．

図 3 提案方式の処理の流れ
Fig. 3 Process of the proposed method.

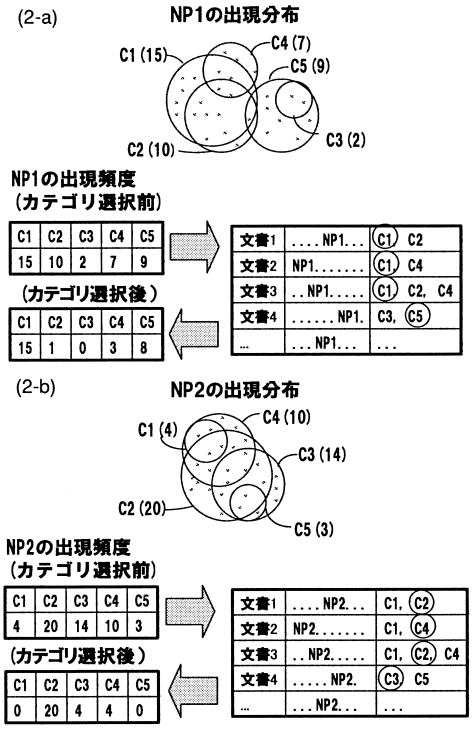

図 4 提案方式
Fig. 4 Proposed method.

3.3 提 案 方 式

そこで，改良方式では，出現頻度の分布を計算する前にカテゴリ選択処理を加え，各文書に付与されたカテゴリのうちの 1 つのみを代表カテゴリとして選択し，出現頻度の計算に用いるようにした（**図 3** 参照）．このときの代表カテゴリの選択基準は，各文書に付与されたカテゴリのうちでその用語に関して出現頻度が最も高い 1 つを選ぶ．すなわち，文書 d に付与されたカテゴリの集合を C としたとき，d の用語 NP に関する代表カテゴリ $c_{rep}(d, NP)$ は下記の式で求める．

$$c_{rep}(d, NP) = argmax_{c_i \in C} f(c_i \cap NP)$$

これは，概念的には企業内文書においてその用語に最も中心的にかかわるリーダ的な部門や人を優先的に選択することを意味する．たとえば，**図 4** の (2-a) の $NP1$ の分布の場合は，$f(c_i \cap NP)$ が $c_1 \to c_2 \to c_5 \to c_4 \to c_3$ の順序で高い．このとき，各文書の代表カテゴリは○のついたカテゴリとなる．その結果，$NP1$ のカテゴリごとの出現頻度は，$(c_1, c_2, c_3, c_4, c_5) = (15, 1, 0, 3, 8)$

となる．この提案方式に従うと，カテゴリごとの出現頻度は，図 2 の (1-b) と (1-c) は同一となり（$c_1 = 32$, $c_2 - c_9 = 0$），その結果エントロピーも (1-b) と (1-c) は同一の $H(NP) = 0$, $\chi^2(NP) = 255.68$ となり，(1-b) の NP が専門用語の場合は，(1-c) も専門用語とすることが可能である．

提案方式を導入することにより，少数のリーダ的なカテゴリが他の従属的なカテゴリを率いて管理/担当する専門用語を抽出することができる．図 4 の (2-a) では，c_1 と c_5 という 2 つのカテゴリが率いて管理/担当する専門用語と解釈でき，(2-b) は c_2 が率いて管理/担当する専門用語と解釈できる．これらの専門用語は従来手法では c_1 から c_5 のすべてのカテゴリに一様に出現する用語として解釈され，提案手法よりもスコアが高く（カイ二乗値の場合は低く）見積もられ，専門用語とできない可能性が高い．

なお，上記の式では，代表カテゴリは同じ文書であっても着目する用語に応じて変化する．たとえば，図 4 の (2-b) の $NP2$ の場合は (2-a) と同じ文書であるが $NP1$ と異なる代表カテゴリを選択する．一方で，代表カテゴリを用語の種類にかかわらず固定する方法もある．すなわち，c_i が付与された文書の数を $f(c_i)$ としたとき，

$$c_{rep}(d) = argmax_{c_i \in C} f(c_i)$$

で代表カテゴリを求める方法もある．今回この式ではなく最初の式を選んだのは，開発プロジェクトごとにリーダが割り当てられる企業内における実際の製品名等の担当/管理方法に柔軟に対応できると考えられるからである．

4. 評　　価

本章では，従来手法とそれに 3 章の提案手法を組み合わせた手法を比較することにより，提案方式の有効性を評価する．

4.1 実験対象の記事

2 種類の企業内文書を用いる．1 つは，社外向けの広報記事である．広報内容としては，新製品発売や製品の導入事例，新技術開発がトピックとして記述されている．記事にはその広報を発行した部門名がカテゴリとして付与されている．複数の部門が共同して広報を発行することがある．大規模な企業ではどの部門がどのような製品や技術を保有しているかを各社員が把握するのは容易でなく，このような広報記事を解析した情報共有システムの利用価値は高いと考える．2004 年度の 1 年分である 441 記事を使用し，各記事の平均文字数は 1663 文字で，平均文字数の分散は 532 文字，平均カテゴリ数は 1.36 である．**図 5** に記事の例を示す．

作成者 カテゴリ （部門名）	○○推進本部 ○○ソリューション事業部
本文	**「GPS を利用した選手位置情報表示システム」を新たに開発**: NEC では，このたびの実証実験の結果をふまえ，「GPS を利用した選手位置情報表示システム」をイベント運営事業者や広告代理店，放送事業者等を中心に幅広く提案していく考えであります．…

図 5 社外向けの広報記事の例
Fig. 5 Example of a press release.

作成者 カテゴリ （部門名）	○○マネージャー ○○主任 ○○主任研究員 ○○主任研究員
本文	**A 社と H システムの実証実験実施へ**: A 社に新しい H システムの技術を提案し，実証実験を行うことで合意した．…

図 6 研究部門の報告記事の例
Fig. 6 Example of a report on research section.

もう 1 つは，研究部門の報告記事である．研究成果物に関する事業部門とのミーティング内容や，社外の企業等への研究の紹介等がトピックとして記述されている．記事には記事に関係する人物名（社員名）がカテゴリとして付与されている．複数の人が記事に関係する場合がある．研究成果を事業として成功させるためには研究成果の事業部門との連携が不可欠であり，このような広報記事を解析した情報共有システムの利用価値は高いと考える．2003 年度の 1 年分である 1390 記事を使用し，各記事の平均文字数は 589 文字で，記事の平均文字数の分散は 157 文字，平均カテゴリ数は 3.23 である．**図 6** に記事の例を示す．人名は苗字+職位で表し，1 年間の間に職位が変更した人物も存在する．今回は，人名の同義語辞書を外部定義せず，このような場合は別のカテゴリ値として扱うようにした．

このように，2 種類の文書集合を用いたのは主に次の理由による．1 つは，2 つの異なる文書集合を用いることで，評価の信頼性を高めるためである．次に，

付与されたカテゴリ数が多い文書と少ない文書で提案手法の効果を比較するためである．また，付与されたカテゴリの種類が人と部門で異なる文書集合で提案手法の効果を検証するためである．さらに，研究部門の報告記事のようなカテゴリ値にゆれが存在する場合に提案手法の有効性を検証することある．

4.2 前 処 理

専門用語の候補となる用語は下記の基準で企業内文書から抽出した．原則としては，連続する品詞「名詞」「未知語」「記号–一般」「記号–アルファベット」の形態素列を専門用語候補の用語とした（例．情報+検索+システム）．形態素解析ツールとしては「茶筅[4]」を使用した．ただし，例外として「名詞–固有名詞」以外の 1 形態素の用語は専門用語候補から除外した（例．開発，メール）．これは，専門用語は複合名詞が多いという指摘がある[7] ことに加え，対象文書を目検した場合でもこの傾向が確認できたためである．

4.3 評価に用いる計算式（評価方式）

ある用語の専門用語らしさのスコアを求める計算式は下記の 4 つの評価式を組み合わせて用いる．

(1) **抽出パターンを用いる手法**

専門用語候補の用語の内部及び前後の文字列が抽出パターンを満足する回数 tf_m をスコアとする．この方式は 1 章の従来方式の（従来 1）を代表する方式として選択した．**表 2** の抽出パターンを用いた．表 2 の EXP_S は用語の開始位置を，EXP_E は用語の終了位置を示す．抽出パターンは正規表現で記述され，ある用語がこれらのいずれかを満足する場合はスコアを 1 つ増やす．これらの抽出パターンは 4.1 節の 2 種類の実験対象の記事のそれぞれ 30 記事から人手で抽出した．

(2) **頻度と文書に対する用語の出現頻度の偏りを用いる方法**

用語の tf/idf 値をスコアとする．この方式は従来方式の（従来 2）の方式として選択した．tf/idf 値は下記の式を用いた．ここで，tf は用語の出現頻度を示し，N は文書数，df は用語を含む文書数を意味する．

$$tf \times \log \frac{N}{df}$$

(3) **文書に付与されたカテゴリに対する用語の出現頻度の偏りを用いる方式（カテゴリ選択処理なし）**

用語のカテゴリに対するエントロピーをスコアとする．この方式は 1 章の従来方式の（従来 3）を代表する方式として選択した．具体的な方式は，3.1 節と同様である．$H(NP)$ で表す．

(4) **文書に付与されたカテゴリに対する用語の出現頻度の偏りを用いる方式（カテゴリ選択処理あり）**

用語のカテゴリに対するエントロピーをスコアとする．この方式は 1 章の従来方式の（従来 3）にカテゴリ選択処理を加えた提案手法である．具体的な方式は，3.3 節と同様である．$H_{cat}(NP)$ で表す．

上記を組み合わせた下記の（方式 a）から（方式 d）の 4 つの式を比較する．これにより，従来方式に提案方式を組み合わせることによる精度向上を評価できる．ただし，tf_m と tf は，専門用語候補が専門用語である場合に相関が強く，両方を掛け合わせると tf の影響が強くなりすぎるため，組み合わせる際には tf_m のみを用いることにした．この補正を行ったほうが抽出精度が高いことを予備実験により確認している．また，エントロピー値 $H(NP)$，$H_{cat}(NP)$ は用語の専門用語らしさが大きいほど値が低くなるため組み合わせの際には正規化を行った．

$$tf_m \quad \text{（方式 a）}$$
$$tf_m \times \log \frac{N}{df} \quad \text{（方式 b）}$$
$$tf_m \times \log \frac{N}{df} \times 2^{-H(NP)} \quad \text{（方式 c）}$$
$$tf_m \times \log \frac{N}{df} \times 2^{-H_{cat}(NP)} \quad \text{（方式 d）}$$

4.4 正 解 判 定

専門用語であるか否かの基準は 2 章の専門用語の範囲に従う．すなわち，上記の（方式 a）–（方式 d）によって抽出された用語が 2 章で定義した専門用語であるならば正解とし，それ以外の場合は不正解と判定する．

4.5 実 験 結 果

図 7 の (1-a) に社外向け広報記事の，(2-a) に社内向け報告記事の実験結果のグラフを示す．横軸は評価式（方式 a）–（方式 d）のスコアの順位を示し縦軸はある順位以上の精度を示す．(1-a) において 2475 位，(2-a) において 2309 位までをグラフに表示しているのは抽出パターンにあてはまる用語がこの個数であったためである．したがって，これらの順位ではすべての評価式で精度が同一となる．4 つの評価式（方式 a）–（方式 d）のうち提案方式を含むものは（方式 d）であるがどの

表 2 使用した抽出パターン
Table 2 Extraction patterns.

ID	抽出パタン
1	EXP_E.{0,5}（を｜の）（開発｜受注｜発売｜発表｜開始｜活用）
2	EXP_E.{0,5}（を｜の）（共同｜提携｜連携｜協業）
3	EXP_S.*?（殿｜[^（同）（仕）（多）] 様）EXP_E
4	EXP_S.*?システム EXP_E
5	EXP_S.*?技術 EXP_E
6	「EXP_S.*?EXP_E」

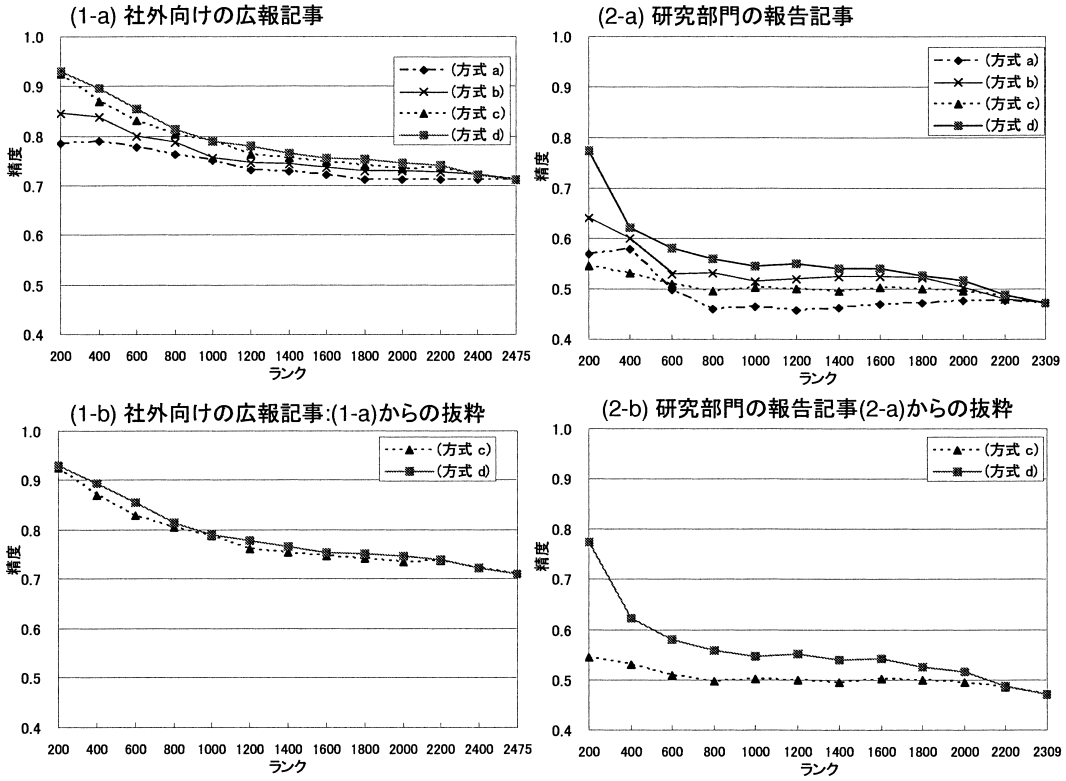

図 7 実験結果
Fig. 7 Experimental result.

順位を閾値としても，どちらの記事においても最も高い抽出精度となっており，従来手法と比較して精度/再現率ともに高いことが分かる．

4.6 考　　察

4.5 節の結果より，企業内文書において提案手法を抽出パターンや tf/idf といった従来方式と組み合わせることよって専門用語の抽出精度を向上させることができることが分かる．**表 3** に精度向上がより顕著であった社外向け広報記事における（方式 b）と（方式 d）のスコア上位 20 件の結果を示す．この結果を見ると（方式 b）では削除しきれなかった高頻度の不要語を（方式 d）により削除できていることが分かる．このような高頻度の不要語は，検索時に多くの入力キーワードと結び付くため削除の効果は大きい．一方，図 7 の (2-a) では，（方式 d）のスコア 400 件以降の効果が小さくなっている．詳細を見ると (2-a) の 2,309 の用語のうち，297 位からの 1,319 個が出現頻度 1 の用語であった．これらの用語は，提案手法（方式 d）の $H_{cat}(NP)$ の項ではつねに最大値である 1 になってしまい，専門用語とそうでないものを区別することができなかったことが原因である．

図 7 の (1-a) と (2-a) から（方式 c）と（方式 d）を抜き出したグラフを (1-b) と (2-b) に示す．（方式 c）と（方式 d）を比較することにより，カテゴリを用いた専門用語抽出におけるカテゴリ選択処理の効果を確認できる．カテゴリ選択処理を含まない（方式 c）では，1 文書あたりの平均カテゴリ数が多い社内向け報告記事では大幅に精度が低下し，提案手法は 1 文書あたりのカテゴリ数が多い記事で特に有効であることが分かる．これは，1 文書あたりのカテゴリ数が多い記事ほど，従属的なカテゴリが多いため，その用語に最も中心的にかかわるリーダ的な部門や人を各文書から優先的に選択する（従属的なカテゴリを用いないようにする）提案手法の効果が大きいと考えられる．

製品名等の専門用語は，初期は比較的少数の人/部門が関連する小型プロジェクトとして開始され，その後徐々に関連する人/部門を増やしながら大きなプロジェクトに成長し，最終的には企業内で広く一般的に認知される流れをたどる場合が多い．提案手法は，この 1 段階および 2 段階で専門用語を抽出することができる．カテゴリ選択処理を用いない従来手法でも第 1 段階の専門用語を抽出できる場合もあるが，一般的

表 3 社外向け広報記事における提案方式（方式 d）：左と従来方式（方式 b）：右のスコア上位 20 件の結果（実際の製品名等は伏字に変換し表記）
Table 3 Top 20 results from reports on research section.

順位	用語	スコア	頻度	判定	順位	用語	スコア	頻度	判定
1	**H システム**	135.3	19	○	1	本システム	284.7	56	×
2	**A 社**	83.5	8	○	2	本技術	168.1	29	×
3	3 次システム	78.6	10	×	3	**H システム**	135.3	19	○
4	**D システム**	73.1	9	○	4	要素技術	117.6	18	×
5	**E システム**	67.5	8	○	5	基礎技術	107.8	16	×
6	**R 技術**	65.4	9	○	6	一システム	102.1	13	×
7	**R サービス**	61.8	12	○	7	デモシステム	90.8	13	×
8	一システム	59.8	13	×	8	**M 技術**	88.8	42	○
9	**T システム**	59.5	8	○	9	重要技術	87.6	13	×
10	**I システム**	59.5	8	○	10	**A 社**	83.5	8	○
11	**K 技術**	56.8	7	○	11	**R サービス**	82.3	12	○
12	**J 技術**	53.4	7	○	12	3 次システム	78.6	10	×
13	**B 社**	52.2	5	○	13	**D システム**	73.1	9	○
14	**C 社**	52.2	5	○	14	試作システム	72.7	10	×
15	**S 技術**	51.2	48	○	15	**T ソフトウェア**	71.2	12	○
16	**D 社**	47.2	5	○	16	**I 技術**	71.2	12	○
17	**L 技術**	47.2	5	○	17	**E システム**	67.5	8	○
18	**J システム**	47.2	5	○	18	**R 技術**	65.4	9	○
19	**M 技術**	46.1	42	○	19	**A システム**	65.4	9	○
20	**S システム**	42.2	5	○	20	プロトタイプ	65.4	9	×

には小型プロジェクトであっても 1 人でない限りチームリーダとその部下という形態で進められる場合が多く，また，1 段階の状態では統計的に信頼できる出現数に達せずに抽出漏れになる可能性があることを考慮すると，提案手法の効果は大きいと考える．一方，第 3 段階の専門用語は従来方式/提案方式どちらであっても，一般的な用語との区別がつかないため抽出が難しい．以上のことから，運用上は比較的短いスパンで新規登録文書を加えた専門用語抽出を繰り返し，専門用語が第 2 段階を過ぎる前に専門用語リストに追加することが重要である．

5. 従来研究

1 章で述べたように，従来の専門用語抽出方式の代表的な方式は大きく 3 つに分類することができる．本章では，従来方式とそれらに対する提案方式の貢献について述べる．

（従来 1）の抽出パターンを用いる方式は，新聞記事等の文書から「組織名」「人名」「地名」等の固有表現を抽出する方式としてよく用いられる[1]．最も基本的な方式は，固有表現を特徴づける表現内部や周囲の文字列を手がかりとして抽出する方式である．これらの手がかりは抽出パターンとして人手または機械学習を用いて作成する．新聞記事を利用した評価実験では，「日付表現」「人名」「地名」といった比較的語彙や周囲の文脈の種類が限定される場合には有効であるが，限定されにくい製品名等の「固有物名」は抽出が難しいことが報告されている[6]．本方式では，4 章の実験から，作成者カテゴリ付き文書からの製品名等の専門用語抽出として，抽出パターンを用いる方式と組み合わせることで精度向上が実現できることを示した．

（従来 2）の頻度と文書に対する用語の出現頻度の偏りを用いる方法は，最も代表的な方式として，tf/idf 値がある[2]．この方式と提案方式は，2 章の専門用語を抽出する場合において，多くの文書に出現する一般的すぎる不要語を排除する効果がある点で共通する．しかしながら，4 章の実験から，提案方式と組み合わせることで tf/idf では削除できなかった高頻度の不要語を削除できることを示した．

（従来 3）の文書に付与されたカテゴリに対する用語の出現頻度の偏りを用いる方式は，3.1 節で述べたように，エントロピーやカイ二乗値[3] を用いて用語が特定のカテゴリに偏って出現する場合に専門用語とする方式である．従来方式は 1 文書に 1 カテゴリが付与される場合を前提としているが，本方式は，3.3 章で述べたように，カテゴリ選択処理を加えることによって，1 文書に複数カテゴリが付与される場合にも対応できるよう改良し，4 章の実験から，改良方式の有効性を示した．なお，3.1 節で紹介したエントロピーやカイ二乗値以外にも，用語が関連するカテゴリ数が少ないほど専門用語らしいとする方式がある[5]．しかしながら，単純にカテゴリ数を数えただけではカテゴリ数は多くてもその中の一部のカテゴリに用語が大きく偏って出現する場合にスコアが低くなるという課題が

あるため，エントロピーやカイ二乗値の方が優れていると判断した．

上記の（従来 1）–（従来 3）以外にも専門用語を抽出する方式として，用語の周囲に出現する単語分布の異なりの度合いに着目する手法[8] や，複合名詞を構成する単名詞の統計的分布を利用する手法[7] が提案されている．これらの方式と本方式組み合わせた評価は今後の課題であるが，本方式とは異なる手がかりを利用して専門用語を抽出する補完的な関係にあると考えている．

6. おわりに

本稿では，人や部門といった作成者情報が 1 文書に複数付与された文書から専門用語を抽出する手法を提案した．本方式は，カテゴリを利用して専門用語を抽出する従来方式を利用するが，さらに，1 文書に複数のカテゴリ情報が付与された文書にも対応できるように改良した．評価実験により，企業内文書において提案手法を従来方式と組み合わせることよって，専門用語の抽出精度を向上でき，特に高頻度の不要語を削除できる利点があることを示した．

今後の課題として，まず，今回の 4 章の実験の規模は比較的小規模であり，今後はテストコレクションを構築してより大規模な他の評価式を含めた評価を行いたい．また，本稿では，作成者情報がカテゴリとして事前付与された文書を前提としていたが，現状の固有表現抽出技術は人名の抽出精度は比較的高いため提案方式と組み合わせることでカテゴリ情報が付与されていない文書にも適用することが可能になると期待できる．さらに，本方式を企業内文書以外，たとえば論文や特許文献，e-mail 等へ適用し，有効性を検証する予定である．

参考文献

1) 竹元義美，福島俊一，山田洋志：辞書およびパターンマッチングルールの増強と品質強化に基づく日本語固有表現抽出，情報処理学会論文誌，Vol.42, No.6 別冊，pp.1580–1591 (2001).
2) Sparck-Jones, K.: A Statistical Interpretation of Term Specificity and Its Application in Retrieval, *Journal of Documentation*, Vol.28, No.1, pp.11–21 (1972).
3) 長尾　真，水谷幹男，池田浩之：日本語文献における専門用語の自動抽出，情報処理学会論文誌，Vol.17, No.2, pp.110–117 (1976).
4) 形態素解析ツール茶筅．http://chasen.naist.jp/hiki/Chasen/
5) 内元清貴，関根　聡，村田真樹，小作浩美，井佐原均：異分野コーパスを用いた用語抽出，*Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp.444–450 (1999).
6) 野畑　周，関根　聡，辻井潤一：日本語固有表現抽出の難易度を示す指標の提案と評価，自然言語処理，Vol.10, No.1, pp.3–26 (2003).
7) 中川裕志，湯本紘彰，森　辰則：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.27–46 (2003).
8) 久光　徹，丹羽芳樹，辻井潤一：タームの representativeness を測る，情報処理学会研究報告，NL-133-16, pp.115–122 (1999).

(平成 17 年 9 月 19 日受付)
(平成 18 年 3 月 13 日採録)

(担当編集委員　石川 博，有次 正義，片山 薫，
木俵 豊，土田 正士)

立石　健二（正会員）

1975 年生．1997 年東京理科大学理学部応用物理学科卒業．1999 年九州大学大学院システム情報科学研究科知能システム学専攻修了．同年日本電気（株）入社．以来，情報検索，情報抽出システムに関する研究に従事．現在，同社インターネットシステム研究所勤務．日本データベース学会会員．

久寿居　大（正会員）

1967 年生．1990 年京都大学工学部数理工学科卒業．1992 年京都大学大学院工学研究科修士課程修了．同年日本電気（株）入社．以来，情報分析・知識管理システムの研究・開発に従事．現在，同社インターネットシステム研究所勤務．