企業内情報共有のための専門用語抽出方式の提案

A Terminology Extraction Method for Knowledge Sharing System on Enterprise

立石 健二♥ 久寿居 大♠
Kenji TATEISHI Dai KUSUI

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。本項では、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。

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. The proposed method extracts terms that are strongly-correlated with few categories of authors under the assumption that a terminology like 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 attached on a document, and improves it to address plural categories on a document. The experimental result shows that it enables high accuracy of terminology extraction for documents in the company by combining it with other previous methods.

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上で必要となる。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。

従来の専門用語抽出方式は大きく(従来1)抽出パタ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つ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る。本手法は、前述の(従来3)の手法を1文書に複数の作成者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ものである。評価実験では、提案手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を向上でき、特に高頻度の不要語を削除できることを示す。

2. 専門用語抽出システム

2.1 企業内文書のモデル

企業内文書の特徴として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。例えば、図のような社外向けの広報記事や、研究部門の報告記事がある。広報記事では新製品発売や製品の導入事例、新技術開発等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。研究部門の報告記事では、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など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おり、複数の人が記事に関係する場合がある。このような作成者カテゴリ付文書から製品名/技術名/顧客名等の専門用語を抽出できれば、部門/人の専門領域をデータベース化した情報共有システム開発も可能となる。本稿では、作成者カテゴリ付き文書を企業内文書とみなし、専門用語抽出の対象とする。

2.2 要件

製品名等の専門用語は、社内において初期は少数の人/部門が関連する小型プロジェクトとして開始し、徐々に関連する人/部門を増やしながら、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。情報共有システムを目的とした専門用語抽出では、情報共有の効果の観点からこれらの早期の段階で抽出することが必要である。

♥ 正会員 NECインターネットシステム研究所
k-tateishi@bq.jp.nec.com
♠ 非会員 NECインターネットシステム研究所
kusui@ct.jp.nec.com

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、放送事業者などを中心に幅広く提案していく考えであります。 …

図 1 社外向けの広報記事の例
Fig. 1 Example of a press release

作成者 カテゴリ (人物名)	○○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 2 研究部門の報告記事の例
Fig. 2 Example of a report on research section

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。上記の初期の段階では製品名、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存在するため、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 節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 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、及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。本節では、エントロピーを用いる方法を代表として説明する。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 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 が偏って出現していることになる。

$$Entropy(NP_i) = \sum_j - p(C_j \mid NP_i) \log_2 p(C_j \mid NP_i)$$

$$p(C_j \mid NP_i) = \frac{f(C_j \cap NP_i)}{\sum_k f(C_k \cap NP_i)}$$

ここで、$p(C_j \mid NP_i)$はNP_iのカテゴリC_jにおける出現確率であり、$f(C_j \cap NP_i)$はNP_iのカテゴリC_jにおける出現頻度である。

図 1 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わかる。この図の(1-a)の NP のエントロピーは NP が C 1 から C8 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 Entropy(NP)=3 となる。また、図 1 の(1-b)の NP のエントロピーは NP が C1 で 32 回、C2-C9 で各 0 回出現していることから(1-b)の Entropy(NP)=0 となる。

3.2 従来方式の課題

従来方式の課題は、1 文書に複数のカテゴリが付与されている企業内文書から専門用語を抽出する際に、その文書を複数のカテゴリの頻度計算で使用するため、少数のカテゴリに関連が深く専門用語となるべき用語であっても、スコアが高く見積もられ、結果として専門用語とならない場合があることである。この点について以下具体的に説明する。

カテゴリ付文書から専門用語を抽出する従来の方法は、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、その文書を複数のカテゴリの頻度計算で使用するため、その文書に付与されたカテゴリの数が多いほど NP は多くのカテゴリに偏りなく出現するとして扱う。これは、ある文書に C1 から Cn 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 C1 から Cn 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。

例えば図 1 の(1-c)は、1 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 C1 に対して 32 回、C2 から C9 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 Entropy(NP)=2.972 となる。このスコアは(1-b)よりも高く、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 1 文書に複数のカテゴリが付与されている場合に多くのカテゴリに偏りなく出現する扱われる用語は、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 C1 に所属しており NP はカテゴリ C1 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。

3.3 改良方式

そこで、改良方式では、ある用語の専門用語のスコアを計算する際に、各文書に付与されたカテゴリの内でその用語の出現頻度が最も高い一つを選択して出現頻度の計算に用いる。これは、企業内文書において、概念的には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。例えば、図 2 の(2-a)の NP1 の分布の場合は、出現頻度が C1→C2→C5→C4→C3 の順序で高い。このとき、各文書のカテゴリはこの優先順位で○のついた一つのみを選択する。その結果、NP1 のカテゴリ毎の出現頻度は、(C1,C2,C3,C4,C5)=(15,1,0,3,8)となる。この提案方式に従うと、カテゴリ毎の出現頻度は、図 1 の(1-b)と(1-c)は同一となり（C1＝32、C2－C9＝0）、その結果エントロピーも(1-b) と(1-c)は同一の Entropy(NP)=0,となり、(1-b)の NP が専門用

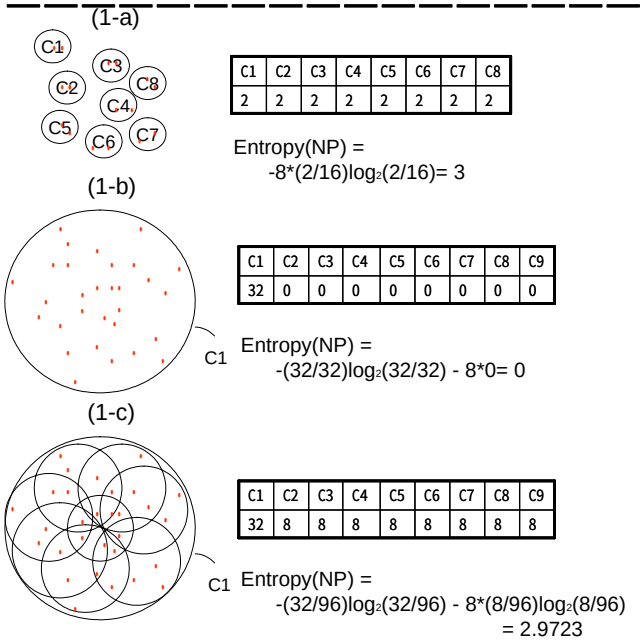

図 3 エントロピーの計算例
Fig.3 Example of calculation of Entropy.

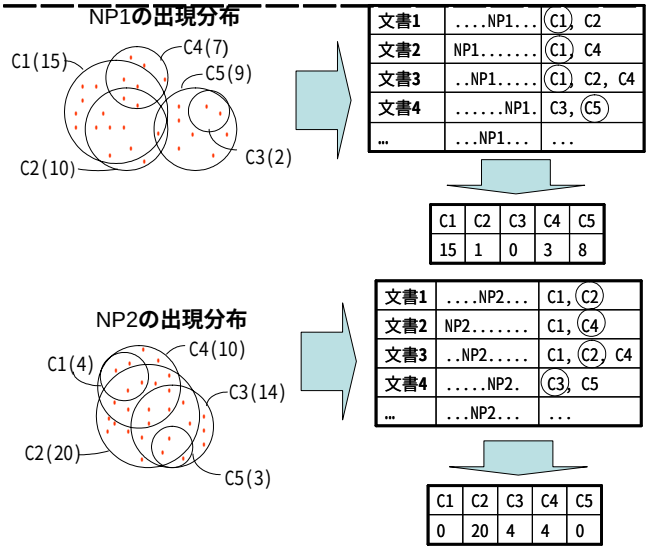

図 4 提案手法
Fig.4 Proposed method

語の場合は、(1-c)も専門用語とすることが可能である。

提案方式では、着目する用語に応じてカテゴリを動的に選択する。例えば、図 2 の(2-b)の NP2 の場合は、(2-a)と同じ文書からあっても、NP2 の出現分布に応じて NP1 と異なるカテゴリを選択する。これは、専門用語ごとにリーダとなる人/部門が異なる場合に対応できることを意味し、企業内における実際の製品名等の担当/管理範囲に柔軟に対応できる。

このような提案方式を導入することにより、人/部門といった作成者カテゴリが 1 文書に複数付与された文書において、少数のリーダ的な作成者カテゴリに管理/担当される専門用語も抽出することができる。例えば、図 2 の(2-a)は、C1 と C5 という 2 つのリーダが率いて管理する専門用語であり、(2-b)は C2 という 1 つのリーダが率いて管理する専門用語と解釈する。これらの専門用語は従来手法では C1 から C5 の全てのカテゴリに均等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見積もられ、専門用語とできない可能性が高い。

4. 評価実験

本節では、従来手法とそれに 3 節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 実験方法

評価に用いた企業内文書は、2.1 節で述べた社外向け広報記事と、研究部門の報告記事である。広報記事は 2004 年度 1 年分 441 記事を用い、1 記事当たりの平均文字数は 1663 文字、平均カテゴリ数は 1.36 である。同様に、報告記事は、2003 年度 1 年分 1390 記事を用い、1 記事当たりの平均文字数は 589 文字で、平均カテゴリ数は 3.23 である。

専門用語の候補となる用語は次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、過去の研究において専門用語の多くは 2 形態素以上から構成されるという報告がある[5]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。

評価式として、下記の(方式 a)から(方式 d)の 4 つの式を用いる。(方式 a)から(方式 c)は従来方式の組み合わせた方式であり、(方式 d)は従来方式と提案方式を組み合わせた方式である。

tf_mは、1 節の(従来 1）の抽出パタンを用いた手法のスコア

$$tf_m \quad \text{(方式 a)}$$

$$tf_m \times \log(\frac{N}{df}) \quad \text{(方式 b)}$$

$$tf_m \times \log(\frac{N}{df}) \times 2^{-Entropy(NP_i)} \quad \text{(方式 c)}$$

$$tf_m \times \log(\frac{N}{df}) \times 2^{-Entropy2(NP_i)} \quad \text{(方式 d)}$$

であり、専門用語候補の内部及び前後の文字列が抽出パタンを満足する回数を示す。抽出パタンは、正規表現で記述し、”EXP_E.{0,5}(を | の)(開発 | 受注 | 発売 | 発表 | 開始 | 活用) ” や“「EXP_S.*?EXP_E」”等 6 個を用いる。なお、EXP_S は専門用語候補の開始位置を、EXP_E は終了位置を示す。次に、log(N/df)は、(従来 2)の tf/idf を用いた手法のスコアである。N は文書数、df は専門用語候補を含む文書数を意味する。tf を省略しているのは、先の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である。$2^{-Entropy(NPi)}$は、(従来 3)のカテゴリに対する出現頻度の偏りを用いた手法の専門用語候補のスコアであり、エントロピーの値を用いる。エントロピー値は専門用語らしさが大きいほど値が低り、他の要素と逆の傾向を持つため正規化を行っている。$2^{-Entropy2(NPi)}$は、(従来 3)の手法を複数カテゴリが付与された文書に対応できるよう改良した提案方式である。

専門用語であるか否かの判断を明確にするため、本実験では専門用語を製品名, 開発物成果名、技術名、機能名、顧客企業名、提携先企業名とする。これらは、企業内での情報共

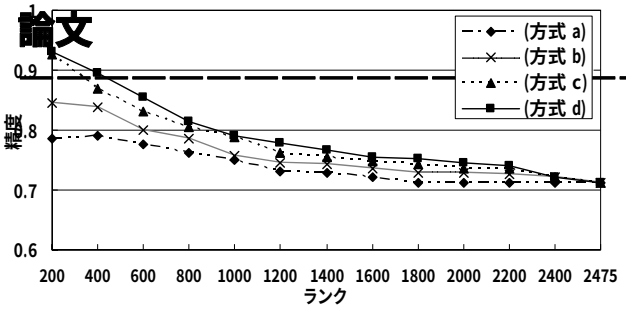

図 5 実験結果(社外向け広報記事)
Fig.5 Experimental result (press releases)

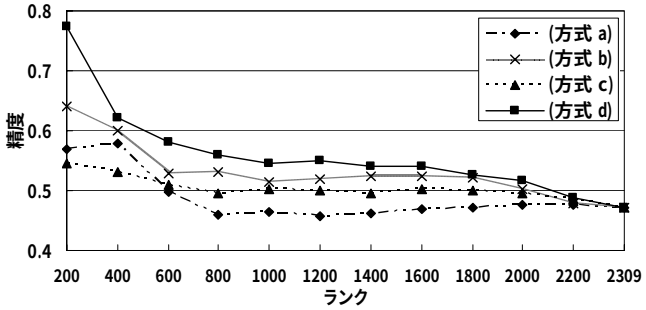

図 6 実験結果(研究部門の報告記事)
Fig.6 Experimental result (reports on research section)

有に必要という観点から挙げたものである。つまり、(方式 a)-(方式 d)によって抽出された候補がこれらの範囲の専門用語ならば正解としそれ以外の場合は不正解と判定する。

4.2 実験結果

図 5 に社外向け広報記事の、図 6 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。図 5 において 2475 位、図 6 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ンに当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わかる。

5. 考察

4.2 節の結果より、企業内文書において提案手法を抽出パタンや 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できることが分かる。(方式 d)は、カテゴリを用いた専門用語抽出において 1 文書に複数カテゴリが付与される場合に対応した提案手法を含むが、対応していない(方式 c)では、1 文書当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する。したがって提案手法は 1 文書当りのカテゴリ数が多い記事で特に有効である。

表 4 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 10 件の結果を示す(製品名等は伏字で表記)。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わかる。一方、図 6 の(6- b)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。詳細を見ると図 6 の 2309 の用語の内、297 位からの 1319 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 Entropy2 の項では常に最大値である 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。以上のことから、提案手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がわかる。

表 1　研究部門の報告記事のスコア上位 10 件の結果
Table. 1 Top 10 results of reports on research section.

順位	用語(方式 d) 提案方式	頻度	判定	用語(方式 b) 従来方式	頻度	判定
1	**H システム**	19	○	本システム	56	×
2	**A** 社	8	○	本技術	29	×
3	3 次システム	10	×	**H システム**	19	○
4	**D システム**	9	○	要素技術	18	×
5	**E システム**	8	○	基礎技術	16	×
6	**R 技術**	9	○	一システム	13	×
7	**R サービス**	12	○	デモシステム	13	×
8	一システム	13	×	**M技術**	42	○
9	**T システム**	8	○	重要技術	13	×
10	**I システム**	8	○	**A 社**	8	○

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一般化した。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ることを示した。

[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, 第 42 巻, 第 6 号別冊, pp.1580-1591, 2001.
[2] Sparck-Jones, K., “A Statistical Interpretation of Term Specificity and Its Application in Retrieval”, Journal of Documentation, 28(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, 17(2), pp.110-117, 1976.
[4] 形態素解析ツール茶筅, http://chasen.naist.jp/hiki/Chasen/
[5] 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.

立石 健二　Kenji TATEISHI
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士課程修了。同年日本電気(株)入社。現在、NEC インターネットシステム研究所主任。情報抽出, 情報検索に関する研究に従事。情報処理学会会員、日本データベース学会正会員。

久寿居 大　Dai KUSUI
1992 年京都大学大学院工学研究科修正課程修了。現在、NEC インターネットシステム研究所主任研究員。情報分析・知識管理システムの研究・開発に従事。情報処理学会会員。