信頼区間の下限値による確率推定を用いた企業名抽出

中野 翔平† 菊地 真人† 吉田 光男† 岡部 正幸‡ 梅村 恭司†

†豊橋技術科学大学 情報・知能工学系 〒441-8580 愛知県豊橋市天白町雲雀ヶ丘 1-1
‡豊橋技術科学大学 情報メディア基盤センター 〒441-8580 愛知県豊橋市天伯町雲雀ケ丘 1-1
E-mail: †{nakano14@ss.cs.tut.ac.jp, kikuchi14@ss.cs.tut.ac.jp, yoshida@cs.tut.ac.jp, umemura@tut.jp},
‡okabe@imc.tut.ac.jp

あらまし Good-Turing のスムージングとナイーブベイズを用いた先行研究において，名前の周辺と名前を構成する文字列を特徴量としている抽出方法がある．本研究では Good-Turing の代わりに信頼区間の下限値による確率推定を用いた企業名抽出を提案する．先行研究と同様の条件で新聞記事から企業名の抽出を行う比較実験を行なった結果，近似された適合率及び近似された再現率のそれぞれにおいて提案手法が Good-Turing を用いた方法を上回り，有意水準 1%で提案手法と先行研究の有意差が認められた．

キーワード 情報抽出，企業名，*N*-gram，ナイーブベイズ，確率推定

1. はじめに

文書の分類をするために，同じ種類の名前のリストが用いられることがある．例えば，野球やサッカーのチーム名や選手の名前が含まれる文書はスポーツに，パソコンの OS 名や携帯電話の機種名が含まれる文書は IT に分類することができる．このように，特定の種類の名前のリストがあることで人手によらない分類が可能になる．

特定の種類の名前のリストを作成する方法として，既存の辞書から名前を取り出し利用する方法や手作業でリストに名前を追加していく方法，形態素解析又は構文解析で名前を取り出し利用する方法が挙げられる．しかし，既存の辞書から名前を取り出す方法は新たな語が含まれないという問題がある．手作業で追加する方法は一から作成した場合，コストが膨大となる，最初だけ既存の辞書を用いたとしても新たな語が出続けるたびに追加していくのは同様にコストが大きい，また人為的なミスも発生しやすいという問題がある．形態素解析又は構文解析を利用する方法は固有名詞が抽出できたとしても，そこからは特定の種類の名前だけを人手で選別しなければならない，また辞書に含まれない名前が出現した場合に漏れが生じるという問題もある．これらの問題を解決するために先行研究において，ナイーブベイズを基にした特定の種類の名前の抽出法が提案されている[6,7].

ナイーブベイズを基にした抽出法では，未知語が現れた場合，本来の確率が 0 で無いにも関わらず全体の尤度が 0 になるという問題がある．この問題を解決するために確率推定が用いられる．未知の名前を抽出したい場合，確率推定法の選択が重要となる．

本研究では，先行研究で提案された特定の種類の名前の抽出法に，新たに信頼区間の下限値による確率推定を検討する．これは先行研究の課題として挙げられた，片仮名を含む名前に対して誤りが発生しやすいという問題に対して対処したものである．さらに，先行研究で最も適合率及び再現率の高かった確率推定法と本稿で提案する確率推定法の比較実験を行い，本稿で提案する確率推定法の方が適合率及び再現率を有意に向上できることを示す．

2. 関連研究

ここでは，本研究に関連する未知語の抽出，日本語からの特定の種類の名前の抽出に関連する研究，ナイーブベイズを基にした特定の種類の名前の抽出に関する研究について述べる．

未知語の抽出が可能な研究として，次のようなものがある．森ら[1]は，*N*-gram 統計値を用いた単語の抽出と品詞の推定を同時に行う手法を提案している．この研究では形態素解析済みのコーパスに対し，名詞の前後の *N*-gram の分布を用いることで未知語を含む名詞の抽出を行なっている．梅村[2]は，出現頻度と出現集中を表す統計量を用いることで辞書を用いず文書中の特有の語を抽出する手法を提案している．この研究ではある文字列を含む文書の数を用いて文書中の特有の語を抽出している．以上の研究は未知語を抽出できるものであるが，特定の種類の名前の抽出は行なっていない．

日本語からの特定の種類の名前の抽出に関連する研究として，固有名詞の分類の 1 つである組織名を抽出する研究[3,4,5]がある．これらの手法は単語ごとに分割済みの文書を用意，又は先に単語ごとに分割を行なっている．

ナイーブベイズを基にした特定の種類の名前の抽出に関する研究として，名前の周辺と名前を構成する文字列を特徴量としている次のような研究がある．菅野[6]は， *N*-gram の統計値を用いて語の抽出を行う手

法を提案している．この研究では企業名を適用例として，企業名の前後の文字 *N*-gram の出現頻度を用いて抽出を行なっている．また，企業名抽出においては企業名自身の文字 *N*-gram の出現頻度も特徴量として用いることが有用であることを報告している．中野ら[7]は，菅野の手法を基にした新たな特徴量を提案している.この研究では企業名自身を新たに企業名自身の前，企業名自身の中及び企業名自身の後に分け，それぞれの文字 *N*-gram の出現頻度を特徴量として用いることが有用であることを報告している．

この手法は，形態素解析を利用せずに抽出を行うため 1 章に挙げた漏れが生じるという問題を回避できると考える．さらにこの方法は，既存の辞書の増強として用いることもでき，抽出した未知語をリストに追加することでより内容を充実させられるという点も有用である．この方法を改良することでより正確に特定の種類の名前の抽出が行えるようになると考える．

中野ら[7]の手法は Good-Turing のスムージングを用いている．これは観測されなかった語に対して一定の頻度を分配し，観測された語に対しても頻度の補正を行なっている．一方，提案手法はベイズ統計の枠組みで，観測によって計算できる事後分布を扱う．通常は,この事後分布における確率の期待値をとる方法(ラプラススムージング)で確率を推定する場合が多いが，本研究では事後確率における信頼区間を構成し，その下限値をとるアプローチをとった．

提案手法と中野ら[7]において，特徴とするもの，尤度比の計算方法，評価の行い方は同一であるが，確率推定の方法だけが異なる本研究では，確率推定の方法を取り換えることで適合率及び再現率が有意に向上することを示す．

3. 使用する概念

3.1. 概要

ここでは,本研究で使用している 5 つの概念,*N*-gram,分布仮説,評価文字列,尤度比,文書について述べる．これらの概念は中野ら[7]と同じものである．

3.2. *N*-gram

N-gram[8]とは，文字，単語又は品詞などの連続した組み合わせである．単語を空白で区切る英語などの言語では単語単位で区切った *N*-gram（単語 *N*-gram）が使用される．しかし，日本語は空白で区切られていないため，直接単語 *N*-gram を用いることは出来ない．この問題の解決として,文字単位で分割を行う方法[1,9]がある．今回はこの文字単位で区切った *N*-gram（文字 *N*-gram）を用いる．また，菅野[6]は企業名抽出に対して文字 *N*-gram の大きさ別の比較実験を行い，図 1 のような 2 文字区切りの *N*-gram（文字 Bigram）を用いた場合に最も適合率及び再現率が高かったことを報告している．中野ら[7]も文字 Bigram を用いている．

図 1 文字 Bigram の例

3.3. 分布仮説

Harris の分布仮説[10]とは，「同じ文脈で使われる言葉は，類似する意味をもつ傾向がある」という仮説である．中野ら[7]と同様に，本研究ではこの分布仮説における文脈を企業名の直前及び直後の文字 Bigram と考える．

3.4. 評価文字列

中野ら[7]は分布仮説を基づいて企業名周辺の文字列を，企業名の前，企業名自身の前，企業名自身の中，企業名自身の後及び企業名の後の 5 つに分類し，特徴量として用いることが有用であると報告している．

これ以降，企業名の前，企業名自身の前，企業名自身の中，企業名自身の後及び企業名の後という各部分を図 2 のように先行部，先頭部，中間部，末尾部及び後続部として表す．

また，図 3 のように文書全体の評価文字列を集めた集合を評価文字列集合とする．

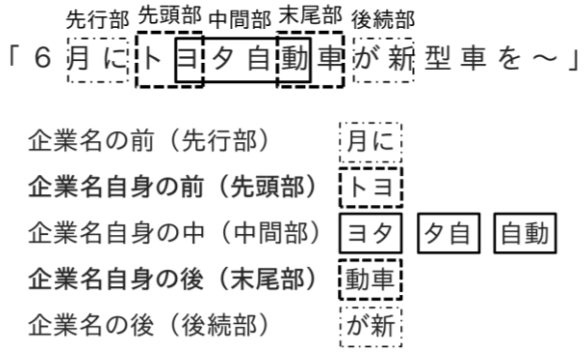

図 2 評価文字列の各部の例

	先行部	先頭部	中間部	末尾部	後続部
1	月に	トヨ	ヨタ,タ自,自動	動車	が新
2	たに	日産	産自,自動	動車	の社
3	は、	トヨ	ヨタ,タ自,自動	動車	をは
4	方、	日産	産自,自動	動車	では
⋮	⋮	⋮	⋮	⋮	⋮
n	も、	三菱	菱自,自動	動車	では

図 3 評価文字列集合の例

3.5. 尤度比

文字列の企業名らしさを評価する値として尤度比を用いる．尤度比とは，帰無仮説の尤度$L(H_0)$と対立仮説の尤度$L(H_1)$の比を取り，どちらが尤もらしいかを比較する指標である．対立仮説H_1より帰無仮説H_0の方が尤もらしいときに尤度比は小さくなり，帰無仮説H_0より対立仮説H_1の方が尤もらしいときに尤度比は大きくなる．どちらも同じくらい尤もらしいときには尤度比は 1 となる．

本研究では，帰無仮説H_0を「与えられた文字 Bigram が文書中から任意に取り出したものである」(評価文字列集合の文字 Bigram ではない)，対立仮説H_1を「与えられた文字 Bigram が企業名自身またはその直前直後から取り出したものである」(評価文字列集合の文字 Bigram である）とする．

以上の帰無仮説H_0と対立仮説H_1を尤度比として表すと式(1)となる．

$$\frac{\text{評価文字列集合の文字 Bigram から求めた尤度}}{\text{文書全体の文字 Bigram から求めた尤度}} \tag{1}$$

3.6. 文書

3.3 節の分布仮説により，尤度の計算には企業名の直前直後の文字 Bigram の出現頻度が必要となるため，図 4 のような文章中に企業名が含まれており直前直後の文字 Bigram を得ることのできる文書を使用する．

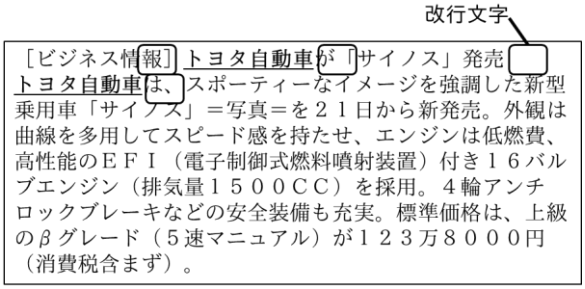

図 4 文書と利用する直前直後の文字 Bigram の例

4. スムージング

4.1. 概要

尤度の計算において，未知の文字 Bigram が現れた場合，本来の尤度が 0 で無いにも関わらず文字列全体の尤度が 0 になるという問題(ゼロ頻度問題)がある．このため,スムージングによる確率推定が用いられる．

ここでは,中野ら[7]の手法で用いられた Good-Turing,及び本稿で新たに用いる信頼区間の下限値によるスムージングについて述べる．

4.2. Good-Turing

Good-Turing[11]は頻度rの語の種類数N_rを用いて出現頻度に対して補正を行い，出現しなかった語の確率を推定する．また頻度が高い語の場合，N_rの値が不安定になるため，ジップの法則を用いることでさらに補正を行う．ジップの法則とは「頻度順位が n 位の単語は 1 位の単語の 1/n の確率であらわれる」という法則である．この法則により，$\log N_r$及び$\log r$が線形の関係で表される．

中野ら[7]は Gale ら[12]の方法に基づく，通常の Good-Turing と線形回帰を用いた Good-Turing を頻度が低いものと高いもので切り替える Good-Turing を用いて確率推定を行なっている．中野らで使用されている Good-Turing を式(2)に示す．

$$P_{SGT}(w_i^j|S_X) = \begin{cases} P_{GT}(w_i^j|S_X) & (\sigma \times 1.65 < |P_{GT}(w_i^j|S_X) - P_{LGT}(w_i^j|S_X)|) \\ P_{LGT}(w_i^j|S_X) & (\sigma \times 1.65 \geqq |P_{GT}(w_i^j|S_X) - P_{LGT}(w_i^j|S_X)|) \\ \dfrac{N_1}{N_0 N} & (r = 0) \end{cases} \tag{2}$$

$$P_{GT}(w_i^j|S_X) = \frac{(r+1)\cdot\frac{N_{r+1}}{N_r}}{N}$$

$$P_{LGT}(w_i^j|S_X) = \frac{r(1+\frac{1}{r})^{b+1}}{N}$$

$$\sigma = \sqrt{(r+1)^2 \cdot \frac{N_{r+1}}{N_r}\left(1+\frac{N_{r+1}}{N_r}\right)}$$

n	評価文字列の文字数
w_i^j	評価文字列中のi文字目からj文字目までの部分文字列
P_{SGT}	使用する Good-Turing の推定値
S_X	任意の文字 Bigram 集合
r	S_X内のw_i^jの頻度
N	文字 Bigram の総頻度
N_r	S_X内の頻度rの文字 Bigram の種類数

4.3. 信頼区間の下限値によるスムージング

一般的なスムージングの手法の一つとして，コーパスに出現した全ての語に対して，それぞれの頻度に 1 を加えるラプラススムージングがある．ラプラススムージングは，観測における確率の事後分布を計算したあと，確率の期待値をとることに等しい．

実際にはラプラススムージングを使用すると，まれな事象への確率を過大に推定するという問題が生じる．このとき，事後分布が計算できるならば，その分布の信頼区間を構成することはできる．

例として工業製品の製造不良の確率を推定するときには，観測から計算される事後分布から求められる信頼区間の上限値を用いる．これは，不良品をあやまって良品と判断とするリスクが，良品を不良品と判断するリスクよりも大きいからである．

我々は，名前の抽出のタスクに置いては，確率を過

大に評価するリスクが確率を過小とするリスクよりも大きいと考えた．この理由は，抽出の精度は五分五分よりも高くするのが自然ということにある．

信頼区間を構成する近似公式は多くあるが，頻度が極めて低い場合には近似が使えないという制限がある．そこで，本研究では菊地ら[13]の方法を使って信頼区間を構成し，尤度比の計算に用いる確率には，その信頼区間の下限値を使用した．

二項分布の尤度関数が式(3)で表されるときの信頼区間を式(4)に示す．この信頼区間の下限値p_{lb}をスムージング値として用いる．下限値p_{lb}は代数的に求めることができないため，式(4)の下限値側を整理して二分法によって求める．整理した式を式(5)に示す．

$$L(p;n,x) = {}_nC_x\, p^x(1-p)^{n-x} \tag{3}$$

$$\frac{\alpha}{2}\int_0^1 p^x(1-p)^{n-x}dp = \int_0^{p_{lb}} p^x(1-p)^{n-x}dp = \int_{p_{ub}}^1 p^x(1-p)^{n-x}dp \tag{4}$$

$$(1-p_{lb})^{n-x+1}\cdot\sum_{i=0}^{x}\left(\frac{(n-x+i)!}{(n-x)!}\cdot(p_{lb})^i\cdot\frac{1}{i!}\right)-\left(1-\frac{\alpha}{2}\right)=0 \tag{5}$$

p	成功確率
n	試行回数
x	成功数
α	有意水準
p_{lb}	信頼区間の下限値
p_{ub}	信頼区間の上限値

5. 出現頻度の学習

5.1. 概要

ここでは，評価値の計算に用いる頻度の集計及び尤度の計算方法について述べる．この集計方法及び計算方法は中野ら[7]と同じものである．

5.2. 学習方法

尤度の計算には先行部から末尾部までの各部の文字 Bigram の出現頻度及び全体の文字 Bigram の出現頻度を使用するため，3.6 節の文書を用いて頻度を集計した学習データを用いる．また，尤度に対しては 4 章のスムージングを適用する．

例として，図 5 のような複数の企業名を含む文書から，各部の頻度及び全体の頻度を集計し尤度を計算すると図 6 のようになる.図 6 上部は各文字 Bigram の頻度，図 6 下部は各文字 Bigram の尤度を表す．なお，尤度にはスムージングされた値を用いている．

これにより，今回の例の「自動」や「動車」のような，企業名によく使われる文字列の尤度が高くなり企業名らしい文字列を得ることができる．

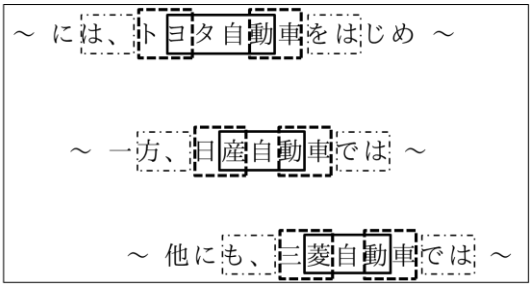

図 5 複数の企業名を含む文書の例

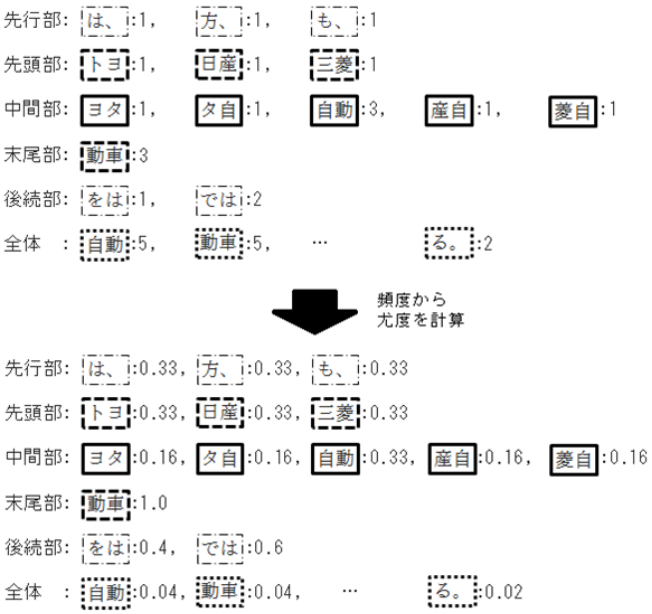

図 6 各部の頻度の集計及び尤度の計算例

6. 企業名の評価及び抽出

6.1. 概要

ここでは，企業名らしさの評価方法及び企業名の抽出方法について述べる．この評価方法及び抽出方法は中野ら[7]と同じものである．中野ら[7]から引用したものについて横線で表す．

6.2. 評価方法

抽出の段階では，対象となる文書の先頭から順に部分文字列が企業名らしいかの評価を行う．この評価は抽出したい文字長内に含まれる全ての部分文字列が対象となる．この部分文字列を評価文字列と呼ぶことにする．評価時は評価文字列を企業名とその直前直後の文字列と仮定して，先行部から後続部の各部に対して 3.4 節の尤度比を計算する．この値が企業名らしさを表すものとなる．図 8 の例の「月にトヨタ自動車が新」を評価したいとすると，この評価文字列に対する尤度比を計算し，「トヨタ自動車」という文字列が企業名らしいかの評価を行うこととなる．

本研究では，評価文字列に対する尤度比を先行部

から末尾部までの各部の尤度比の相乗平均と仮定して，この値を評価値と定義する．これは図 9 のように表される．

評価値を求めるための評価式$LR(w_1^n)$を式(2)に示す．文字数nの評価文字列wにおけるi文字目からj文字目までの部分文字列をw_i^jとする．この時，各部の尤度比LR_{Pre}, LR_{Head}, LR_{Mid}, LR_{Tail}, LR_{Post}は，先行部，先頭部，中間部，末尾部，後続部の文字 Bigram 集合S_{Pre}, S_{Head}, S_{Mid}, S_{Tail}, S_{Post}内の文字 Bigram の推定値$P^*(w_i^j|S_X)$（S_Xは各部の文字 Bigram 集合）と抽出用文書の文字 Bigram 集合S_{doc}内の文字 Bigram の推定値$P^*(w_{n-1}^n|S_{doc})$の比で表される．

今回は引用内の式(2)の$P^*(w_i^j|S_X)$について，Good-Turing 及び信頼区間の下限値によるスムージングを使用している．

$$LR(w_1^n)=\left(LR_{Pre}\times LR_{Head}\times\prod_{i=4}^{n-4}LR_{Mid}\times LR_{Tail}\times LR_{Post}\right)^{\frac{1}{n-3}} \quad (2)$$

$$LR_{Pre}=\frac{P^*(w_1^2|S_{Pre})}{P^*(w_1^2|S_{doc})}$$

$$LR_{Head}=\frac{P^*(w_3^4|S_{Head})}{P^*(w_3^4|S_{doc})}$$

$$LR_{Mid}=\frac{P^*(w_i^{i+1}|S_{Mid})}{P^*(w_i^{i+1}|S_{doc})}$$

$$LR_{Tail}=\frac{P^*(w_{n-3}^{n-2}|S_{Tail})}{P^*(w_{n-3}^{n-m}|S_{doc})}$$

$$LR_{Post}=\frac{P^*(w_{n-1}^{n}|S_{Post})}{P^*(w_{n-1}^{n}|S_{doc})}$$

n	評価文字列の文字数
w_i^j	評価文字列中のi文字目からj文字目までの部分文字列
LR	評価文字列の尤度比（＝評価値）
LR_X	各部の尤度比
S_{Pre}	先行部の文字 Bigram 集合
S_{Head}	先頭部の文字 Bigram 集合
S_{Mid}	中間部の文字 Bigram 集合
S_{Tail}	末尾部の文字 Bigram 集合
S_{Post}	後続部の文字 Bigram 集合
S_{doc}	抽出用文書の文字 Bigram 集合
$P^*(w_i^j\|S_X)$	S_X中のw_i^jの出現確率の推定値のスムージング値（今回は Good-Turing 推定法を使用）

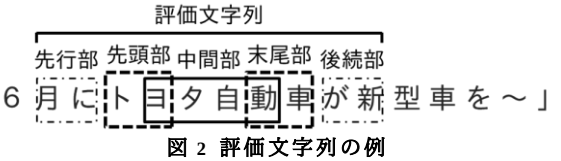

図 2 評価文字列の例

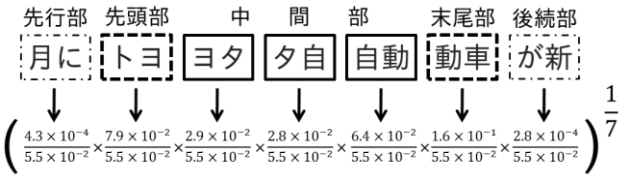

図 3 評価値の計算例

6.3. 抽出方法

抽出したい最小文字数から最大文字数までの評価文字列について 6.2 節の評価値を計算し，その値の高い順から一定数の企業名を抽出する．

例として「６月にトヨタ自動車が新型車を発売した.」という文章に対し評価値を計算して，値が高い順に並べ替えると表 1 のようになる．この例では企業名が 1 つしか含まれていないが，実際の文書では多くの企業名が含まれるため上位一定数を抽出する．

表 1 評価文字列と評価値の例

評価文字列	評価値（尤度比）
月に トヨタ自動車 が新	**0.2447**
月に トヨタ自動車が 新型	0.0572
にト ヨタ自動車 が新	0.0510
６月 にトヨタ自動車 が新	0.0461
月に トヨタ自動 車が	0.0424
にト ヨタ自動車が 新型	0.0121
にト ヨタ自動車が新 型車	0.0082
６月 にトヨタ自動 車が	0.0081
トヨ タ自動車が 新型	0.0065
・・・	・・・

7. 比較実験

7.1. 概要

ここでは，確率推定を変更したことによる影響を確認するため，今回提案した信頼区間の下限値による確率推定を用いた抽出法（以下提案手法）と Good-Turing を用いた抽出法（以下ベースライン）との比較実験を行う．

7.2. 実験条件

実験の各条件は表 2 に示す，ベースラインにおいて最も適合率及び再現率の高かった条件を使用する．文書は，毎日新聞コーパス 91-97 年[14]の年始から 2 万記事を 1 万記事ごとに分割したものを 1 つの文書として計 14 文書を作成する．また，K-分割交差検証で 14 文書中の 13 文書を学習用,残りの 1 文書をテスト用とする．既知の企業名はテスト用文書から形態素解析で組織名を抽出後，パターンマッチにより企業名以外を除去したものを用いた．5 文字から 30 文字までの企業名を対象に評価値の計算を行い，評価値の高い順に上位

2000 件を企業名として抽出した．

表 2 実験条件

使用文書	毎日新聞コーパス 91-97 年の年始から 2 万記事（1 万記事ごとに分割）の計 14 文書
テスト用文書	使用文書中の 1 文書
学習用文書	使用文書中からテスト用の 1 文書を除いた 13 文書
既知の企業名リストの作成方法	形態素解析で組織名を抽出後,パターンマッチにより企業名以外を除去
***N*-gram**	文字 Bigram
企業名抽出の文字数の範囲	5 - 30 [文字]
抽出件数	評価値の上位 2000 [件]
スムージング法	・Good-Turing ・信頼区間の下限値を用いた確率推定 (α=0.9999997)

7.3. 部分正解による評価

正解の評価方法は中野ら[7]と同じく部分適合率及び部分再現率を使用している．それらの説明について中野ら[7]から引用したものを横線で表す．

人が企業名だと認識できる全ての文字列の集合を全体正解集合 A としてこの外に正解は無いものとする．この時，既知の企業名のリストを全体正解集合 A に包含される部分正解集合 a とする．図 10 に全体正解集合 A，部分正解集合 a 及び抽出結果 S の関係図を示す．

以下の評価は菅野[4]を踏襲したものである．本来ならば抽出の正誤の判定には全体正解集合 A を用いるべきであるが，全体正解集合 A は実際には得られない，もしくは得るために大きなコストがかかるため，部分正解集合 a を用いる．この際，抽出結果 S に対して部分正解集合 a から得られる精度及び再現率を全体正解集合 A から得られる精度及び再現率とは区別して部分適合率と部分再現率と表現する．

部分適合率と部分再現率を式(4.1)と式(4.2)に示す．

$$\text{部分適合率} = \frac{\text{部分正解に含まれる抽出文字列の数}}{\text{抽出文字列の数}} \quad (4.1)$$

$$\text{部分再現率} = \frac{\text{部分正解に含まれる抽出文字列の数}}{\text{文書に存在する部分正解に含まれる企業名の数}} \quad (4.2)$$

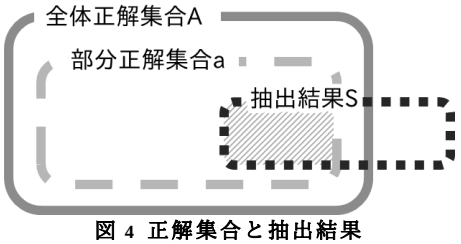

図 4 正解集合と抽出結果

（中略）

[4] 菅野 弘太. n-gram の統計値による企業名の抽出. 豊橋技術科学大学, 2014, 43p. 修士論文.

7.4. 実験結果・考察

抽出の結果，提案手法とベースラインを比較して片方のみに現れた企業名の例（同じ企業名は除く）を表 3 に示す．提案手法のみに現れた例ではベースラインのみに現れた例に比べて，企業名に「**"ソロモン・ブラザーズ"**」や「山崎製**"パン"**」などの片仮名を含む企業名が多く含まれている．これは，Good-Turing の確率推定がジップの法則を基にしているためと考えられる．漢字の文字 Bigram は異なる単語同士で同じ文字 Bigram が現れることが少なく，文字 Bigram と単語の出現頻度がおおよそ等しくなりジップの法則に従っているが,片仮名や平仮名の文字 Bigram は異なる単語同士でも同じ文字 Bigram が現れることが多いためジップの法則に従わない分布となる．このため，信頼区間の下限値による確率推定の方が有効に働いたと考えられる．

表 4 に部分適合率及び部分再現率を示す．また，各項目で上回っている手法を下線で示す．

部分適合率と部分再現率のいずれも'94(1)以外の全ての対象文書において提案手法がベースラインを上回っており，符号検定を行なった結果，有意水準 1%で提案手法とベースラインとの有意差が認められた．

表 3 片方のみに現れた企業名の例

提案手法のみに現れた例	ベースラインのみに現れた例
ソロモン・ブラザーズ	帝国ホテル
山崎製パン	日立製作所
サンケン電気	日本輸出入銀行
アエロフロート	毎日新聞社
ワシントン・ポスト	全日本空輸
ニューヨーク・タイムズ	富士重工業
全日本空輸	石川島播磨
積水ハウス	東洋信託銀行
ホテルオークラ	テレビ東京
みすず書房	中央公論社
⋮	⋮

表 4 部分適合率及び部分再現率

	部分適合率		部分再現率	
	提案手法	ベースライン	提案手法	ベースライン
'91(1)	**0.724**	0.708	**0.735**	0.718
'91(2)	**0.751**	0.712	**0.734**	0.696
'92(1)	**0.788**	0.753	**0.751**	0.717
'92(2)	**0.741**	0.714	**0.702**	0.676
'93(1)	**0.842**	0.798	**0.724**	0.687
'93(2)	**0.888**	0.867	**0.670**	0.654
'94(1)	0.860	**0.862**	0.704	**0.706**
'94(2)	**0.866**	0.857	**0.652**	0.645
'95(1)	**0.842**	0.833	**0.665**	0.657
'95(2)	**0.813**	0.783	**0.706**	0.680
'96(1)	**0.890**	0.862	**0.637**	0.617
'96(2)	**0.878**	0.874	**0.627**	0.624
'97(1)	**0.891**	0.864	**0.594**	0.576
'97(2)	**0.895**	0.878	**0.585**	0.574
平均	0.834	0.812	0.678	0.659
分散	0.0034	0.0040	0.0026	0.0021

8. おわりに

本稿では，Good-Turing の代わりに信頼区間の下限値によるスムージングを用いた企業名抽出の提案を行った．そして，新聞記事を対象とした提案手法とベースラインの比較実験を行い，部分適合率及び部分再現率が向上できることを明らかにした．

謝辞

本研究は，住友電工情報システム株式会社との共同研究の成果です．ここに感謝の意を表します．

参 考 文 献

[1] 森 信介, 長尾 眞. n グラム統計によるコーパスからの未知語抽出. 情報処理学会論文誌. 1998, 39(7), p. 2093-2100.

[2] 梅村 恭司. 未踏テキスト情報中のキーワードの抽出システム開発. 未踏ソフトウェア創造事業, 2000.

[3] 山田 寛康ほか. Support Vector Machine を用いた日本語固有表現抽出. 情報処理学会論文誌. 2002, 43(1), p. 44-53.

[4] 宇津呂 武仁, 颯々野 学. ブートストラップによる低人手でコスト日本語固有表現抽出. 情報処理学会研究報告. 2000, 2000(86), p. 9-16.

[5] 齋藤 邦子ほか. CRF を用いたブログからの固有表現抽出. 言語処理学会第 13 回年次大会, 2007, p. 1-4.

[6] 菅野 弘太. n-gram の統計値による企業名の抽出. 豊橋技術科学大学, 2014, 43p. 修士論文.

[7] 中野 翔平ほか. 企業名抽出のための特徴量の検討, 第 7 回データ工学と情報マネジメントに関するフォーラム(DEIM 2015), E8-5, 2015.

[8] 長尾 眞, 森 信介. 大規模日本語テキストの n グラム統計の作り方と語句の自動抽出. 情報処理学会研究報告. 1993, 93(61), p. 1-8.

[9] 浅原 正幸,松本 裕治. 日本語固有表現抽出におけるわかち書き問題の解決. 情報処理学会論文誌. 2002, 45(5), p.1442-1450.

[10] Zellig S. Harris. Distributional structure. Word. 1954, 10(23), p. 146-162.

[11] 北 研二. 確率的言語モデル. 東京大学出版会, 1999, 239p.

[12] W. A. Gale, G. Sampson. Good-Turing Frequency Estimation without Tears. Journal of Quantitative Linguistics. 1995, 2(3), p.217-237.

[13] Masato Kikuchi et al. “Confidence Interval of Probability Estimator of Laplace Smoothing”. ICAICTA2015. The Tide Resort, Bang Saen Beach, Chonburi, Thailand, 2015-08-19/22.

[14] 毎日新聞社. CD-毎日新聞データ集'91-97 年版. 日外アソシエーツ, 1991-1997. (CD-ROM).