企業名抽出のための特徴量の検討

中野 翔平† 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, yoshida@cs.tut.ac.jp, umemura@tut.jp}, ‡okabe@imc.tut.ac.jp

あらまし 先行研究において，辞書に登録されていない名前も抽出可能な企業名の抽出方法が提案されている．この先行研究を改良することで，より正確に名前抽出が行えるようになると考えた．本研究ではこの先行研究を基にして，企業名の先頭及び末尾の文字列の情報を用いる新たな特徴量を提案した．先行研究と提案手法に対して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先行研究を上回り，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．以上より，提案手法を用いることで先行研究に比べ，近似された適合率及び近似された再現率を向上できることを明らかにした．

キーワード 情報抽出，企業名，N-gram

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など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，人手のコストをかけず，辞書に含まれない名前にも対応可能な特定の種類の名前の抽出法が提案されている[4].

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を基にした，新たに特徴量を検討する．特徴量として，名前の直前直後の文字列にはその種類を特定する有用な情報が多く含まれると考え，それを反映させた．さらに，先行研究で最も適合率及び再現率の高かった特徴量と本稿で提案する特徴量の比較実験を行い，本稿で提案する特徴量の方が適合率及び再現率を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，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．

未知語に対しても適用可能な特定の種類の名前の抽出に関連する研究として，次のようなものがある．小山内[3]は，隠れた正例を含む教師データを前提とした Passive Aggressive を利用し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，形態素解析で得られる企業名の前後の形態素の品詞を学習に用いて抽出を行なっている．また企業名抽出においては文字列の末尾 2 文字の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．菅野[4]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．

この中で菅野の手法は，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られる．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り，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．

以上より本研究では，菅野の手法を基にした企業名の抽出について新たな特徴量の検討を行う．

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 4 つの概念，N-gram，分布仮説，尤度比とスムージング，辞書と文書について述べる．ここで述べることは菅野[4]と同じものである．

3.2. *N*-gram

N-gram[5]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．本研究では形態素解析を行わないため，文章を文字単位で区切った *N*-gram（文字 *N*-gram）を用いる．さらに，菅野は文字 *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．

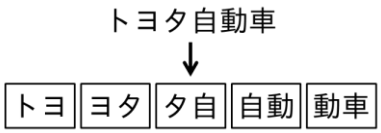

図 1 文字 Bigram の例

3.3. 分布仮説

Harris の分布仮説[6]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．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. 尤度比とスムージング

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」（企業名自身またはその直前直後の文字 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」（企業名自身またはその直前直後の文字 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．ただし，尤度をそのまま用いると文字 Bigram の出現頻度が 0 のときにゼロ頻度問題が発生するため，スムージングを用いて確率の補正を行う．菅野はスムージング別の比較実験を行い，Good-Turing 推定法[7]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．

$$\frac{\text{企業名自身またはその直前直後の文字 Bigram から求めた尤度}}{\text{文書全体の文字 Bigram から求めた尤度}} \quad (1)$$

3.5. 辞書と文書

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 2 のような既存の企業名のリストである辞書と図 3 のような文章中に企業名が含まれており直前及び直後の文字 Bigram を得ることのできる文書を使用する．

```
ＡＣＣＥＳＳ
ＡＤＥＫＡ
ＡＯＣホールディングス
ＡＯＫＩホールディングス
ＡＱインタラクティブ
ＣＦＳコーポレーション
ＣＳＫホールディングス
ＤＣＭ　Ｊａｐａｎホールディングス
ＤＮＡチップ研究所
・・・
```

図 2 辞書（既存の企業名のリスト）の例

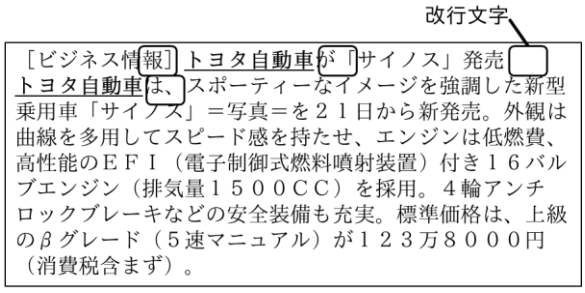

図 3 文書と利用する直前直後の文字 Bigram の例

4. 企業名抽出のための特徴量

4.1. 概要

ここでは，管野が提案した分布仮説に基づく特徴量と企業名自身を用いる特徴量，及び本稿で提案する企業名の先頭及び末尾を用いる特徴量について述べる．

4.2. 分布仮説に基づく特徴量

3.3 節の分布仮説を企業名抽出に適用した場合，企業名直前直後の文字 Bigram から,それらの間にある文字列が企業名らしいかの評価を行うこととなる．図 4 の例では，前の「月に」という文字列と後の「が新」という文字列から「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なっている．

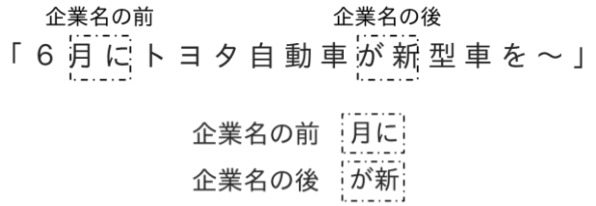

図 4 分布仮説に基づく特徴量の例

4.3. 企業名自身を用いた特徴量

管野は企業名の抽出においては，企業名自身の文字列にも，類似する意味をもつ語を特定できる情報が出現することが多く，抽出に有用であると考え，企業名直前直後の文字 Bigram に加えて企業名自身の文字 Bigram も使用する特徴量を提案している．また，4.2 節のように企業名直前直後の文字 Bigram のみを使用した場合より本節の企業名自身の文字 Bigram も使用した場合の方が適合率及び再現率が高かったことを報告している．図 5 の例では，前後の文字 Bigram に加えて，「トヨ」「ヨタ」「タ自」「自動」「動車」という文字 Bigram も用いて「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なっている．

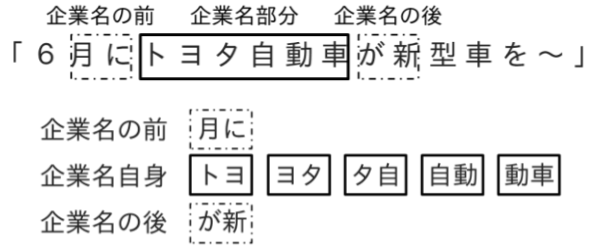

図 5 企業名自身を用いた特徴量の例

4.4. 企業名自身の先頭及び末尾を用いた特徴量

本稿では 4.3 節の特徴量を基に企業名自身を図 6 のように細かく分類した．企業名自身の前には「住友」などのグループ名や「東京」などの地域名，企業名自身の後には「工業」などの業種名といった単語が出現するように，企業名自身の前及び企業名自身の後の文字列も特徴として有用なのではないかと考えたためである．またこれ以降，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 6 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．

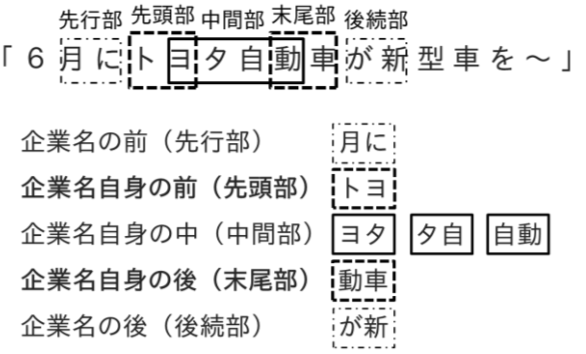

図 6 企業名自身の先頭及び末尾を用いた特徴量の例

5. 出現頻度の学習

尤度の計算には先行部から末尾部までの各部の文字 Bigram の出現頻度を使用するため，3.5 節の文書を用いて頻度を集計した学習データを用いる．

例として，図 7 上部のような複数の企業名を含む文書の各部の頻度を計算すると図 7 下部のようになる．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頻度が高くなり企業名らしい文字列を得ることができる．

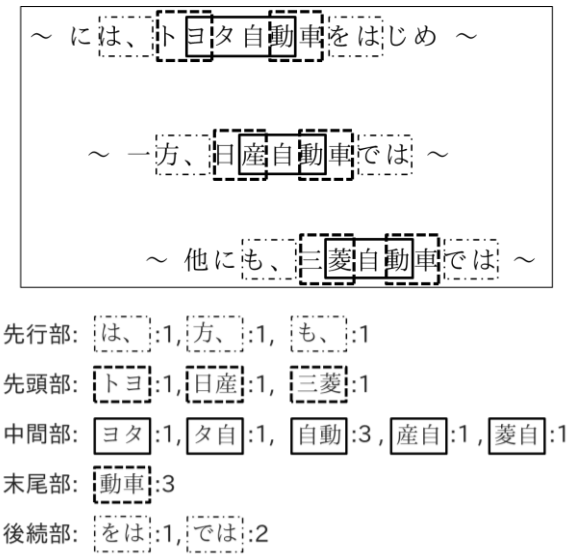

図 7 複数の企業名を含む文書及び各部の頻度の集計の例

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．

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.

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

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

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

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

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

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

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）

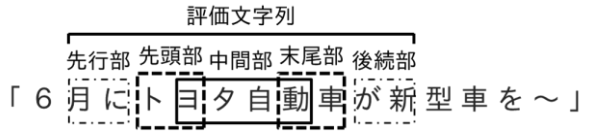

図 8 評価文字列の例

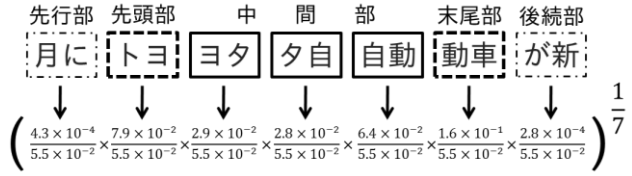

図 9 評価値の計算例

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，特徴量を変更したことによる影響を確認するため，今回提案した企業名自身の先頭及び末尾を用いた特徴量による抽出法（以下提案手法）と管野が提案したものの中で適合率及び再現率が高かった企業名自身を用いた特徴量による抽出法(以下菅野の手法)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，菅野の手法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[8]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．辞書（既知の企業名のリスト）は，東京証券上場企業一覧（2011 年）から 5 文字以上の企業名のリスト[9]を使用する．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．また，スムージング法には 3.4 節で述べたように菅野の手法で最も適合率及び再現率が高かった Good-Turing 推定法を使用した．

今回使用した Good-Turing 推定法の推定値 $P_{SGT}(w_1^n|S_X)$ を式(3)に示す．Gale ら[10]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 Simple Good-Turing を使用した．

$$P_{SGT}(w_i^j|S_X) = \begin{cases} P_{GT}(w_i^j|S_X) & (\sigma \times 1.65 < |P_{GT}(w_i^j|S_X) - P_{LGT}(w_i^j|S_X)|) \\ P_{LGT}(w_i^j|S_X) & (\sigma \times 1.65 \geqq |P_{GT}(w_i^j|S_X) - P_{LGT}(w_i^j|S_X)|) \\ \frac{N_1}{N_0 N} & (r = 0) \end{cases} \quad (3)$$

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

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

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

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 推定法の推定値
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

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（頻度取得用）	使用文書中からテスト用の 1 文書を除いた 13 文書
辞書（既存の企業名のリスト）	東京証券上場企業一覧（2011 年）から 5 文字以上の企業名（計 1441 社）
***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 推定法

7.3. 部分正解による評価

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．

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

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

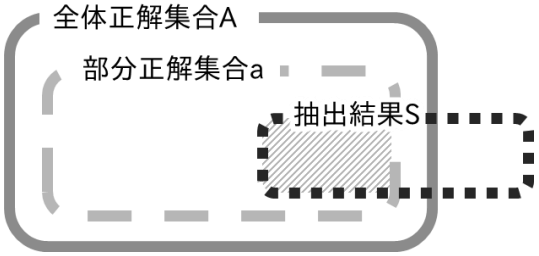

図 10 正解集合と抽出結果

7.4. 実験結果・考察

抽出の結果、提案手法を用いた場合に新たに正解又は誤りと判定された企業名の例（同じ企業名は除く）を表 3 に示す．正解例では，先頭部又は後続部に「ＮＴ（Ｔ)」や「野村」などのグループ名，「山陰」や「富士」などの地域名，「銀行」や「工業」などの業種名といった文字 Bigram が含まれている．これは，提案手法の特徴量が有効に働いているためと考えられる．一方，同じように地域名や業種名を含む「東京コスモス電機」が誤りになったのは，学習ファイル作成において既知の企業名を単純な部分文字列によって一致させており，学習文書中に「東京日産自動車販売」というような他の企業名（日産自動車）を部分文字列に含んだ企業名がある場合に「東京」が先頭部ではなく先行部の学習ファイルに誤って集計されてしまったためと考えられる．企業名の位置を正しく指定することでこの問題は改善可能であると考えている．

また，「**"コー"**プケミカル」や「ユ**"アサ"**商事」など

片仮名を含む企業名は，「アート"コー"ポレーション」や「"アサ"ヒ飲料」のように別の部分に同様の文字 Bigram が出現する例も多く，今回の抽出では誤ってしまったと考えられる．

部分適合率と部分再現率を計算した結果，全ての対象文書のそれぞれにおいて提案手法が菅野の手法を上回った．表 4 に部分適合率及び部分再現率を示す．また，抽出の正誤を基に符号検定を行なった結果，全ての対象文書において有意水準 1%で提案手法と菅野の手法との有意差が認められた．表 4 で有意差が認められた項目を下線で示す．

表 3 新たに正解又は誤りと判定された企業名の例

新たな正解例	新たな誤り例
山陰合同銀行	東京コスモス電機
ＮＴＴデータ	コープケミカル
小田急電鉄	ユアサ商事
グローリー工業	
野村総合研究所	
富士火災海上保険	
川崎重工業	
オリンパス	
岩崎通信機	
・・・	

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	菅野の方法	提案手法	菅野の方法
'91(1)	**0.274**	0.248	**0.932**	0.847
'91(2)	**0.256**	0.232	**0.945**	0.854
'92(1)	**0.300**	0.268	**0.937**	0.839
'92(2)	**0.296**	0.269	**0.940**	0.854
'93(1)	**0.415**	0.354	**0.940**	0.803
'93(2)	**0.457**	0.404	**0.940**	0.832
'94(1)	**0.398**	0.358	**0.938**	0.841
'94(2)	**0.438**	0.383	**0.933**	0.817
'95(1)	**0.430**	0.362	**0.932**	0.785
'95(2)	**0.356**	0.306	**0.936**	0.804
'96(1)	**0.479**	0.408	**0.912**	0.776
'96(2)	**0.476**	0.408	**0.917**	0.788
'97(1)	**0.568**	0.462	**0.917**	0.746
'97(2)	**0.553**	0.452	**0.913**	0.746
平均	0.407	0.351	0.931	0.809
分散	0.0092	0.0052	0.0001	0.0012

8. おわりに

本稿では菅野の手法を基にした，企業名の先頭及び末尾の文字 Bigram を新たな評価文字列として追加する特徴量の提案を行った．そして，新聞記事を対象とした提案手法と菅野の方法の比較実験を行い，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．

今後の課題としては，抽出精度の向上のために新たな評価式を検討すること，今回の評価に用いた正解以外の企業名も含めて評価を行うことが挙げられる．

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] 小山内 一由. 隠れた正例を含む教師データに対する機械学習法とその学習法による名前抽出. 豊橋技術科学大学, 2014, 53p. 修士論文.

[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.

[5] 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.

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

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

[8] 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).

[9] ADVFN PLC.“東京証券取引所 : 上場企業一覧”. ADVFN. http://jp.advfn.com/tse/tokyostockexchange.asp, (参照 2011-10-28).

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