社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－NL－180 (12)
2007／7／24

拡張固有表現獲得の精度向上

塩入寛之† 関根聡†† 梅村恭司†

† 豊橋技術科学大学
〒 441-8580 愛知県豊橋市天伯町字雲雀ヶ丘 1-1
†† ニューヨーク大学
E-mail: †shio@ss.ics.tut.ac.jp, ††sekine@cs.nyu.edu, †††umemura@tutics.tut.ac.jp

あらまし 固有表現辞書の整備は固有表現抽出ツールだけではなく，言語知識や世界知識の把握のために有用である．人手での整備は高コストであり，関根らは WEB 検索エンジンの英語の検索ログを用いて固有表現辞書を整備する手法を提案した（関根，鈴木 2007）．この方法はブートストラッピングを利用し同義の単語を見つける方法の一種であり，その特徴は大規模な固有表現辞書と固有表現のコンテキストを利用することである．しかしながら，検索ログの入手は一般には困難であるため，新聞記事などで同様のことができると便利である．本論文では対象コーパスを新聞記事とし，獲得精度の向上を得るためのいくつかの手法を提案する．

キーワード 固有表現，ブートストラッピング，語彙獲得

Precision Improvement of Extended Named Entity Acquisition

Hiroyuki SHIOIRI†, Satoshi SEKINE††, and Kyoji UMEMURA†

† Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi, Aichi, 441-8580 Japan
†† New York University
715 Broadway, 7th floor, New York, NY 10003 USA
E-mail: †shio@ss.ics.tut.ac.jp, ††sekine@cs.nyu.edu, †††umemura@tutics.tut.ac.jp

Abstract Maintenance of a NE(Named Entity) dictionary is important for grasp of not only a NE extraction tool but also language knowledge and world knowledge. Maintenance by human is expensive, and Sekine suggested technique to get a Named Entity dictionary ready with English search log of WEB search engine. (Sekine, Suzuki 2007). It is a kind of a method to find a synonymous word in using bootstrapping, and the characteristic is to use a large-scale NE dictionary and context of NE. However, acquisition of search log is usually difficult. It is convenient if we can use newspaper articles. We conducted experiments to improve the acquisition precision using newspaper articles.

Key words Named Entity, Bootstrapping, Vocabulary Acquisition

1. はじめに

固有表現は文書中で重要になることが多く，情報としての単位が明確なものである．例としては人名，組織名，地名，時間，日時，金額表現，割合表現，固有物名が固有表現の種類として挙げられる．文書中でこれらの固有表現の種類をタグ付けすることを固有表現抽出という．図 1 に示すように，地名：フランス，人名：ルイ 16 世のように固有表現抽出が行われる．固有表現は文章中で重要な意味を表すことが多いため，固有表現の場所を特定できれば，情報抽出に役立つ．質問応答についても同様に重要な意味を持つことの多い固有表現は，質問応答の答えとして要求されることが多い．これら情報抽出，質問応答等の自然言語処理の分野で固有表現抽出は活用される．

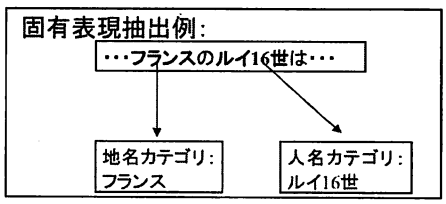

図 1 固有表現抽出

固有表現抽出を実現する手法がいくつか存在する．HMM(Hidden Markov Model) [1]，SVM(Support Vector Machine) [2]，CRF(Conditional Random Fields) [3]，ルールベース [4]，辞書の利用等がある．ここでは辞書の利用について考える．辞書を利用した固有表現抽出の利点は高い抽出精度である．辞書利用の欠点は辞書に無い固有表現は抽出できず，辞書に依存した再現率となることである．この欠点については辞書の語彙を増加することで再現率の向上を期待できる．この辞書の語彙増加を自動的に行う手法として，大規模な固有表現辞書と固有表現のコンテキストを用いる固有表現獲得手法が関根らによって提案されている．[5] この手法は検索エンジンから得られた英語の検索ログを用いて固有表現の獲得を行う．

本論文では，関根らの固有表現獲得手法について，対象コーパスを日本語新聞記事に変更して固有表現獲得を行う．検索ログは入手が困難であり，それに比べて新聞記事は入手しやすい．また，この条件下で固有表現獲得について獲得の精度を向上させる試みも行う．これによって固有表現辞書の語彙増加の効率が上がり，辞書を用いた固有表現抽出の性能が向上することを期待する．以上の 2 点を本論文の目的とする．

2. 日本語新聞記事からの固有表現獲得手法

この章では日本語新聞記事を対象とした固有表現の獲得手法について説明する．本手法は 3 ステップで固有表現を獲得する．まず，既存の固有表現辞書を使ってコンテキストを大量に取得する．そのコンテキストの中から各カテゴリにとって特徴的と考えられるものを計算する．最後に特徴的なコンテキストを使って固有表現の獲得を行う．なお，辞書としては幅広い分野への応用が期待される拡張固有表現の辞書を利用する．

2.1 拡張固有表現

固有表現は一般的に 7〜10 種類のカテゴリである．例えば 1990 年前後の MUC(Message Understanding Conference) では人名，組織名，地名，時間，日時，金額表現，割合表現の 7 種類の固有表現を扱っていた．その後，1998-1999 年の IREX(Information Retrieval and Extraction Exercise) では上記の 7 種類に固有物名を加えた 8 種類を固有表現として設定した．

以上のような固有表現のカテゴリが存在しているが，情報抽出の応用の広がりと質問応答という新しいタスクの出現によってより多くのカテゴリ数への要求がある．情報抽出については「伝染病の発生」という情報からは「病気の名前」，「ロケットの発射」という情報からは「ロケットの名前」というような固有表現のカテゴリが要求される．質問応答は，「島津製作所の田中さんが受賞した国際的な賞は何ですか」という質問に対して「ノーベル賞です」と返す技術である．この技術は特定の知識源（ここでは賞の名前）から答えを探す方法があり，そのためには「賞の名前」という固有表現のカテゴリが要求される．そこで固有表現のカテゴリ数を約 200 種類まで拡張した拡張固有表現が存在する [6]．この拡張固有表現は次の図 2 にあるようなカテゴリ等を固有表現として設定している．

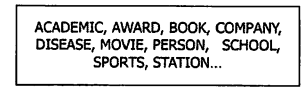

図 2 拡張固有表現カテゴリの例

2.2 固有表現獲得

固有表現の獲得は次の 3 ステップで行う．

（1） 固有表現のコンテキスト獲得

（2） コンテキストの特徴量計算

（3） 未知の固有表現獲得（コンテキストを利用）

固有表現のコンテキストはコーパス中で出現する固有表現周辺の文字列を指す．このコンテキストはコーパス中に大量に存在するため，コンテキストの選別を行う．選別の方法は各コンテキストについて固有表現カテゴリの特徴らしさを計算する．ここで選別されたコンテキストを用いて固有表現を獲得する．

この 3 ステップを図 3 に示した．

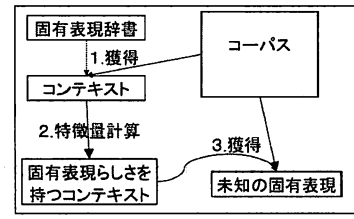

図 3 固有表現獲得の流れ

次に，各ステップについて詳細に説明する．

2.3 固有表現のコンテキスト獲得

本手法においてコンテキストは重要な役割を持つ．本節ではコンテキストについて詳細に説明し，コーパス中からのコンテキスト獲得の具体例を示す．

2.3.1 固有表現とコンテキスト

固有表現とコンテキストとの関係について説明する．図 4 のように文章中に「アカデミー賞」という AWARD カテゴリの固有表現が出現した場合，その前後の文字列をコンテキストとして獲得する．この場合のコンテキストは「今年の#AWARD#を受賞」である（#AWARD# は AWARD カテゴリの固有表現であることを示すタグ）．

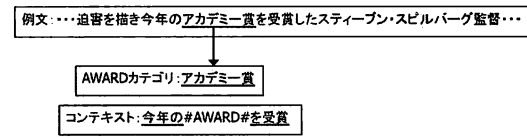

図 4 コンテキストの例

本手法では他の AWARD カテゴリの固有表現も「今年の#AWARD#を受賞」というコンテキストによって同様に出現するという仮説に従う．この固有表現獲得手法ではコーパス（文書集合）中から大量のコンテキストを獲得，選出する必要がある．そのため，コンテキストを獲得するための固有表現が

大量に必要である．本手法では約 10 万語が記載された拡張固有表現を利用することで，コーパスから大量のコンテキストの獲得を実現する．

本研究の手法ではコンテキストを前後の両方を利用しており，あえて係り受けではなくて，表層上の情報をコンテキストに使っている．これは先行研究と異なる [7] [8] [9].

2.3.2 コンテキストの獲得

固有表現辞書に記載されている既知の固有表現と共起するコンテキストを獲得する．また，各コンテキストについて固有表現の全カテゴリとの共起頻度も求める．この共起頻度を用いてコンテキストの特徴量を計算する．コンテキスト獲得の例を図 5 に示す．

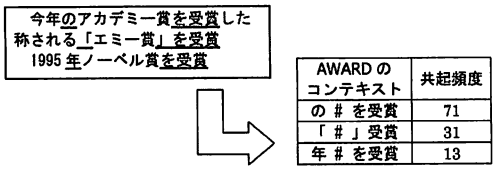

図 5 コンテキストの獲得

また，コンテキストはコーパス中の固有表現の前後の文字列であり，この文字列の長さをいかに設定するかが問題となる．前後の文字列が短すぎると意味の無いコンテキストとなってしまうし，長すぎると計算負荷が高くなる．ここでは固有表現前後の文字列を獲得した際，2 パターンの文字列をコンテキストとして得る．1 つは固有表現の直前にある 2 要素と直後の 1 要素をコンテキストとし，もう 1 つは固有表現の直前にある 1 要素と直後の 2 要素をコンテキストとする．ここでいう'要素'とは文を構成する情報単位を指す．英語ならば 1 単語が 1 要素，日本語ならば文を形態素解析し，1 形態素を 1 要素とすることが考えられる．図 6 は日本語文からコンテキストを獲得する例であり，形態素解析された文から複数の形態素をコンテキストとして獲得している．

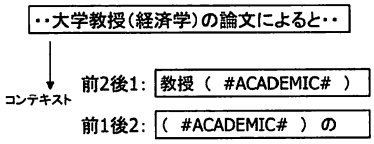

図 6 コンテキスト獲得パターン

2.4 コンテキストの特徴量計算

コーパスから得られる大量のコンテキストを選別するための，コンテキストの特徴量とその計算方法について説明する．

2.4.1 コンテキストの特徴量

上記のように既存の固有表現周辺のコンテキストを獲得することで，未知の固有表現を獲得することが期待できる．しかし，獲得したコンテキストの全てを利用すると不要な文字列も大量に獲得されてしまう．そこで各コンテキストについてカテゴリらしさを定量的に表す特徴量を求め，この数値から利用するコンテキストを選出する．次に示す図 7 で AWARD カテゴリを例に特徴量について説明する．「の # を受賞」は他の AWARD カテゴリが出現しそうであるので高い特徴量が望ましい．もう一方は AWARD カテゴリが出現しそうではあるが，「受賞」のような特徴的な文字列が無いので他のカテゴリも大量に出現するだろうと予測される．つまり，AWARD カテゴリらしさを持たないので，低い特徴量が適している．

例:AWARDカテゴリ（#は固有表現）
の # を受賞 → 高い特徴量
「 # 」の → 低い特徴量

図 7 カテゴリらしさを表す特徴量

2.4.2 特徴量計算によるコンテキストの選別

各コンテキストの特徴量計算のために 2 つの情報 f_{type}, F_{inst} を利用する．

f : 特定のカテゴリと共起する場合

F : 全てのカテゴリで共起する場合

$type$: 共起する固有表現の種類数

$inst$: 固有表現との共起頻度

以上の情報を引用した特徴量算出関数が次の関数 $Score(c)$ である．

$$Score(c) = f_{type}(c) \cdot \log\left(\frac{f_{type}(c)}{F_{inst}(c)} \cdot \frac{1}{R}\right)$$

$$R = \left(\frac{f_{type}(c_{top1000})}{F_{inst}(c_{top1000})}\right)$$

c : 特徴量計算の対象コンテキスト

$c_{top1000}$: f_{type}の上位 1000 件のコンテキスト

この関数は TF-IDF と似た計算を行う関数である．また，元々は検索ログを対象コーパスとして考案された関数である．. f_{type} が高いコンテキストの特徴量が高くなり，F_{inst} が高い場合は特徴量が低くなる．また，log 内では f_{type} が高いコンテキストの上位 1000 件の f_{type} と F_{inst} で正規化を行っている．この特徴量算出関数を AWARD カテゴリに適用した例を表 1 に示す．なお，#EOS$は文章の終わりを意味する．

表 1 コンテキストの特徴量

順位	特徴量	コンテキスト
1	475.18	で #AWARD# を 受賞
2	371.48	、 #AWARD# を 受賞
3	369.99	回 #AWARD# を 受賞
4	319.54	の #AWARD# を 受賞
5	251.67	が #AWARD# を 受賞
6	251.07	で #AWARD# 。 #EOS#
7	240.55	に #AWARD# を 受賞
8	217.28	で #AWARD# 受賞 。
9	214.03	で #AWARD# 、 「
10	207.77	回 #AWARD# #EOS# #EOS#

2.5 未知の固有表現獲得

特徴量の高いコンテキストを使ってコーパス中から未知の固有表現を獲得する．今回は使用するコンテキストの下図は特徴量の高いものの上位100件を用いた．100という件数はおおよその数であり，使用するコンテキストの数として最適かどうか不明であるため，議論の余地がある．これらのコンテキストと1回でも共起する文字列はコーパス中で大量に出現し，その全てが固有表現ではなく，不適当なノイズも混じっている．そこで獲得した文字列について，「固有表現らしさ」を評価する．評価の方法は共起したコンテキストの種類数が多いものが「固有表現らしい」としてランキングする．表2はAWARDカテゴリについての獲得した未知の固有表現である．上位の10件について著者の主観に基づいて正解判定を行ったものである．

表2 獲得した未知の固有表現

順位	AWARDカテゴリ	正解判定	種類数
1	ノーベル平和賞	○	53
2	毎日芸術賞	○	46
3	芸術選奨文部大臣賞	○	44
4	グランプリ	△	44
5	菊池寛賞	○	38
6	大賞	△	36
7	日	×	34
8	特別賞	△	27
9	大会	×	26
10	新人賞	△	23

正解である固有表現がいくつか獲得できており，日本語新聞記事を対象として固有表現獲得手法が十分に動作することが確認できる．表2の結果では正解のものが多いが，不正解のものとして「目」や「大会」などAWARDカテゴリの固有表現周辺に出現する文字列が取れてしまっている．また，△は固有表現ではなく，固有表現を表すことのできる普通名詞である．照応解析の際に有用使える．検索ログの場合には取れておらず，新聞記事を使った方法の特徴である．

3. 精度改善手法の提案

2章で説明した手法をベースラインとし，固有表現獲得の精度を改善するため本論文は次の3つのアプローチを試行した．

（1） コンテキストの特徴量計算関数の提案
（2） コンテキストから助詞を除去
（3） コンテキストの重み付け

以上の3つのアプローチについて述べる．

3.1 コンテキストの特徴量計算関数の提案

検索ログ向けの特徴量計算関数は f_{type}，F_{inst} という2つの情報を利用していた．提案する関数ではさらに f_{inst}，F_{type} の情報を利用する．これによって f_{inst} が低い場合（特定カテゴリでの出現回数が低く，カテゴリらしさが小さいと判断できる場合）に特徴量が低くなる．F_{type} が高い場合（全体のカテゴリで共起するコンテキストの種類数が多いため，特定のカテゴリらしさが小さいと判断できる場合）に特徴量が低くなる．この2つの情報を用いることで，不要と思われるコンテキストを少なくし，多くの特徴的なコンテキストを利用できることを期待する．

f_{type},f_{inst},F_{type},F_{inst} の4つの情報を利用したスコアは次の数式を扱う．提案する関数は検索ログ向け関数と比べて，特定カテゴリ/全体カテゴリの比を重視するものとなっている．c は特徴量計算の対象となるコンテキストである．

$$Score(c) = f_{type}(c) \cdot \frac{f_{type}(c) \cdot f_{inst}(c)}{F_{type}(c) \cdot F_{inst}(c)}$$

3.2 コンテキストからの助詞の除去

コンテキストの獲得手法について，2章で説明した手法では固有表現の前2後1，前1後2の形態素をコンテキストとして獲得している．ここで提案するのは，少ない形態素で構成されるコンテキストから助詞を除去することによって，特徴的なコンテキストをより多く獲得しようという試みである．ACADEMICカテゴリのコンテキストを例にとって説明する．

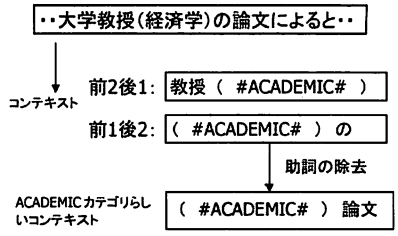

図8 コンテキストの獲得

図8のようなコンテキストの獲得において，前1後2というパターンでは「（ #ACADEMIC# ）の」というコンテキストが得られる．このコンテキストはACADEMICカテゴリらしいコンテキストではない．他の多くのカテゴリで出現するコンテキストである．このコンテキストについては，図8のように助詞の除去を施すことによってACADEMICカテゴリらしいコンテキストとなりうる．助詞の除去によって，この例のように特徴的なコンテキストが増えることを期待する．

3.3 コンテキストの重み付け

2章で説明した手法では獲得した未知の固有表現は共起するコンテキストの種類数でランキングしている．コンテキストの出現頻度等の情報は利用していない．そこでコンテキストのスコアを重み付けすることにより，より特徴的なコンテキストと共起する未知の固有表現がランキングの上位となることを期待する．これを数式で表すと次のようになる．c_i は共起したコンテキスト，all_c は共起したコンテキストの種類数である．

ベースライン：$Rank\_score = all\_c$

提案手法：$Rank\_score = \sum_{i}^{all\_c} Score(c_i)$

AWARDカテゴリについて獲得した未知の固有表現の上位10件を例として示す．

表3と表4を比較すると．表3で存在する「目」「大会」な

表 3　コンテキスト重み付け無し未知固有表現ランキング

順位	AWARD カテゴリ	正解判定	種類数
1	ノーベル平和賞	○	53
2	毎日芸術賞	○	46
3	芸術選奨文部大臣賞	○	44
4	グランプリ	△	44
5	菊池寛賞	○	38
6	大賞	△	36
7	日	×	34
8	特別賞	△	27
9	大会	×	26
10	新人賞	△	23

表 4　コンテキスト重み付けによる未知固有表現ランキング

順位	AWARD カテゴリ	正解判定	スコア
1	芸術選奨文部大臣賞	○	2627
2	毎日芸術賞	○	2578
3	グランプリ	△	2471
4	菊池寛賞	○	2405
5	ノーベル平和賞	○	2202
6	大賞	△	1887
7	紀伊国屋演劇賞	○	1755
8	芸術祭大賞	△	1752
9	芸術祭優秀賞	△	1563
10	特別賞	△	1533

どの誤りが，提案する手法の上位 10 件からは消えていることがわかる．このような不要と思われる未知固有表現の削減のために，提案する手法が有効に働くことを期待する．

4. 評価実験

2 章で説明した手法，検索ログを対象とした固有表現獲得手法をナイーブに日本語新聞記事コーパスへ対応させた固有表現獲得手法をベースラインとして，提案する 3 つの獲得精度向上の手法を実験により評価した．

4.1 実験方法

獲得した固有表現について，高スコアの上位 100 個での正解精度で評価した．用いた固有表現辞書とコーパスである．固有表現辞書は関根らにより作成された日本語の拡張固有表現辞書．約 10 万語の拡張誇張固有表現を持つ．コーパスは毎日新聞の日本語記事を形態素解析システム juman で形態素解析したもの．1994 年〜1999 年の 6 年分の記事を用いた．

未知の固有表現の獲得に用いるコンテキストは特徴量の高いコンテキストの上位 100 件を用いている．コンテキストは前 2 後 1，前 1 後 2 の 2 パターンについて特徴量の高い上位の 100 件を利用している．獲得する固有表現のカテゴリは AWARD, ACADEMIC, DISEASE, SPORTS, PERSON の 5 種類である．求める精度はその 5 つのカテゴリの精度と 5 つの平均精度である．100 以上ある拡張固有表現の種類の中で 5 つのカテゴリに限定した理由は計算時間と正解集合の作成時間に制約があったためである．AWARD, ACADEMIC, DISEASE, SPORTS, PERSON の 5 種類を選んだ理由は新聞記事中で多く出現することが期待できるからである．新聞記事コーパス中で出現頻度が高くない固有表現カテゴリにおいて，本手法が有効に働くことが期待できないと考えられる．

4.2 実験結果

ベースライン手法と提案する 3 手法についての精度比較と，それらを組み合わせた場合の精度比較を示す．

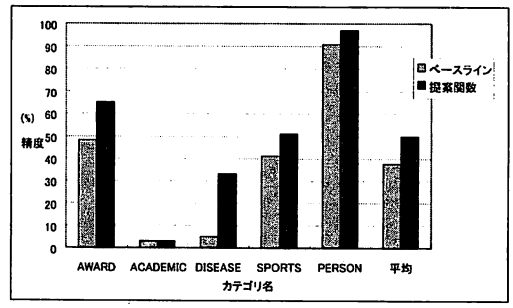

図 9　ベースラインの関数と提案関数

図 9 では ACADEMIC 以外ではベースラインの関数よりも提案関数が良い精度である．特に DISEASE で精度の向上幅が大きい．提案関数で新たに利用した情報 f_{inst}, F_{type} が DISEASE では効果的に働き，ACADEMIC では効果がなかったと考えられる．

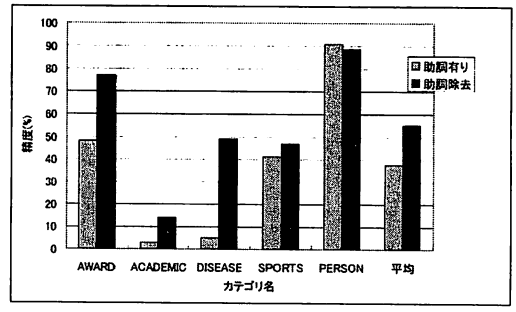

図 10　助詞有り (ベースライン手法) と助詞除去 (提案手法)

図 10 では PERSON 以外では提案手法のほうが良い精度である．特に AWARD と DISEASE の精度向上が大きい．コンテキスト中の助詞情報が AWARD, DISEASE カテゴリにおいては特徴の小さい情報であり，それらが無いほうが特徴的なコンテキストになりやすいと考えられる．一方で，PERSON ではベースライン手法のほうが優れている．PERSON カテゴリは AWARD, DISEASE カテゴリに比べて汎用性の高いカテゴリであり，助詞のような出現頻度の高い情報が有効なのかもしれない．

図 11 では 5 つのカテゴリ全てにおいて提案手法がベースラインを上回っている．重み付けの手法は獲得した固有表現の評価についてのアプローチであり，獲得した固有表現数はベースラインと同じである．この点が他の 2 手法とは異なる点である．「たくさんのコンテキストと共起する文字列は固有表現」とす

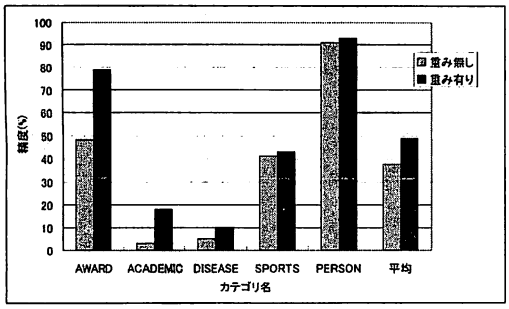

図 11　重み無し (ベースライン) と重み有り (提案手法)

るベースライン手法に対して，重み付け手法は「優れた (特徴的な) コンテキストと共起する文字列は固有表現」という考え方に基づいており，後者が効果的であると考えられる．

提案する 3 つの手法は 5 つのカテゴリの平均正解精度については，いずれもベースラインを上回っている．このことから 3 つの提案手法が固有表現獲得精度の向上に効果があると考えられる．次に助詞除去と助詞除去に重み付けを組み合わせた場合の精度比較結果を示す．

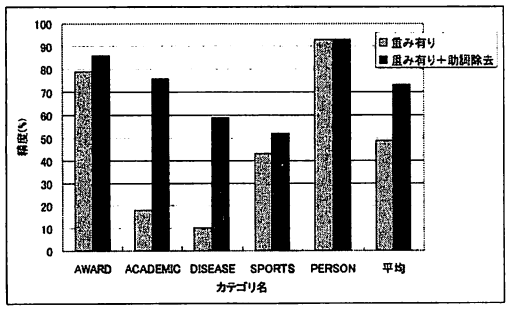

図 12　重み有りと助詞除去の組み合わせ

図 12 では重み有りに助詞除去を組み合わせることで PERSON 以外の正解精度が向上した．特に ACADEMIC と DISEASE について精度向上が大きくなっている．これは助詞除去による特徴的なコンテキスト増加と重み付けの手法のランキング手法が組み合わさって効果的に働いていることが考えられる．その具体的な例として，ACADEMIC の固有表現「環境衛生学」について考える．「環境衛生学」はベースライン手法での獲得固有表現ランキングは 1448 位であった．助詞除去の手法を用いると 342 位に向上する．重み付けの手法を用いた場合には 267 位となった．2 つの手法はそれぞれ「環境衛生学」の獲得固有表現ランキングを向上させるが，本実験での評価範囲である 100 位以内から外れたままである．一方で助詞除去手法に重み付け手法を組み合わせると，「環境衛生学」が 100 位まで上がり，本実験の評価範囲に入る．このようなパターンが他にも存在していることを確認した．これにより，ACADEMIC と DISEASE の獲得固有表現ランキング上位 100 件での精度が向上したと考えられる．

5. 考　　察

今回の実験は AWARD, ACADEMIC, DISEASE, SPORTS, PERSON の 5 つの固有表現カテゴリのみを実験している．拡張固有表現のカテゴリは 100 以上存在する．これらに対応し，より多くの固有表現カテゴリでの固有表現獲得の精度向上を確認することが考えられる．

実験で用いたコーパスは 6 年分の新聞記事を利用している．固有表現獲得手法はコーパスの量が多ければ獲得できる固有表現は増加し，その精度向上も期待できる．今後はより多くのコーパスに対応したシステムを構築することが考えられる．

固有表現の獲得に利用するコンテキスト数について，今回は特徴量の高い上位 100 件のコンテキストを用いた．実際には固有表現獲得に利用すべきコンテキスト数は 50 件かもしれないし，200 件かもしれない．この件数を動的に決定する方法があれば不要なコンテキストを減らすことができ，固有表現獲得の精度向上が期待できる．同様に獲得した固有表現についての評価はスコアの高い上位 100 件で行っている．獲得した固有表現の中で正しい固有表現と判定する閾値を決定できれば固有表現獲得の精度が向上を期待できる．

6. ま と め

固有表現獲得手法を日本語新聞記事に対応させることができた．提案関数，助詞の除去，重み付け，という 3 つの手法についてはそれぞれ固有表現獲得の精度を高めることができた．ベースライン手法の 5 つのカテゴリの平均精度 40%未満であったが，3 つの手法はそれぞれ 50〜60%にまで固有表現獲得精度を高めた．さらに助詞除去と重み付け手法を組み合わせた場合には，70%以上に精度が向上させることができた．

文　　献

[1] D. Bikel, S. Miller, Richard Schwartz and Ralph Weischedel: "Nymble: a High-Performance Learning Name Finder", ANLP 1997.

[2] M. Asahara and Y. Matsumoto: "Japanese Named Entity Extraction with Redundant Morphological Analysis", HLT-NAACL 2003.

[3] A. McCallum and W. Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Fetures Induction and Web-Enhanced Lexicons", CoNLL 2003.

[4] L. F. Rau. "Extracting Company Names from Text", Proceedings of the Seventh Conference on Artificial Intelligence Applications, 1991.

[5] 関根聡，鈴木久美.「検索ログによる拡張固有表現辞書の整備」. 言語処理学会，2007

[6] S. Sekine, C. Nobata, "Definition, Dictionary and Tagger for Extended Named Entities". LREC 2004

[7] Patrick Pantel and Dekang Lin , "Discovering Word Senses from Text", In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2002

[8] 山本和英.「テキストからの語彙的換言知識の獲得」. 言語処理学会 2002.

[9] 土野友司，森辰則，木下冬子，中川裕志.「係り受けの 2 部グラフと共起関係を利用した同義表現抽出」. 言語処理学会 2004

[10] 関根聡.「固有表現から専門用語」. 固有表現と専門用語ワークショップ，2004