2008－NL－186（3）
2008/7/17

固有表現自動獲得に向けての固有表現とコンテキストの関連度

塩入寛之† 岡部正幸† 阿部洋丈† 梅村恭司†

† 豊橋技術科学大学 情報工学系
〒441-8580 愛知県豊橋市天伯町字雲雀ヶ丘1-1
E-mail: †shio@ss.ics.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．本報告では固有表現辞書の語彙自動獲得手法についての性能向上を目的とし，文書中の固有表現と周辺文字列との関連度をより尤もらしく求めるための方法を提案する．関連度を求めるために固有表現と周辺文字列との共起頻度情報を利用し，固有表現獲得に有用な関数を報告，考察する．

キーワード 固有表現，語彙獲得

A Degree of Relation between Named Entity and Context for Named Entity Automatic Acquisition

Hiroyuki SHIOIRI†, Masayuki OKABE†, Hirotake ABE†, and Kyoji UMEMURA†

† Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi, Aichi, 441-8580 Japan
E-mail: †shio@ss.ics.tut.ac.jp

Abstract Maintenance of a Named Entity dictionary is important for not only a Named Entity extraction but also language knowledge and world knowledge. This report aims to improve a method which acquires vocabularies for Named Entity dictionary. We propose a convincingly approach to calculate a degree of relation between Named Entity and Context on a corpus. The approach leverages the co-occurrence information between Named Entity and Context. We report useful functions leveraging the clue for Named Entity acquisition.

Key words Named Entity, Vocabulary Acuisition

1. はじめに

固有表現は文章中で重要になることが多く，文章の意味理解において役立つ概念である．例えば「…フランスのルイ16世は…」という文章においては地名「フランス」と人名「ルイ16世」が固有表現となる．地名や人名の他にも組織名，時間，日時，金額表現，割合表現，固有物名らも固有表現であり，文章中でその位置を特定することができれば意味情報の抽出に利用できる．このような，文章中の人名，地名といった情報の位置を文章中から特定することを固有表現抽出と呼ぶ．固有表現の応用例としては質問応答，要約文抽出，情報検索が挙げられる．

固有表現抽出を実現する手法としては，HMM(Hidden Markov Model)[1]，SVM(Support Vector Machine)[2]，CRF(Conditional Random Fields)[3]，ルールベース[4]，辞書の利用等が存在している．ここで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，辞書に存在する固有表現に関しては抽出できる可能性が高いことである．一方，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．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．また，関根らの手法について対象を日本語の新聞記事として，その性能向上のために関数の提案や固有表現の重み付けを用いた「拡張固有表現獲得の精度向上」[6]を著者らが提案した．これらの手法[5][6]では固有表現と共起する文字列の出現頻度情報等を利用して固有表現の獲得を行う．

本論文の目的は「拡張固有表現獲得の精度向上」[6]についてさらに適合率を向上させることである．その方針として，固有

表現とその共起文字列との関係性について，新しい統計情報を用いるアプローチを取る．既存手法においては固有表現と一つの共起文字列との関係にのみ注目している処理について，提案手法では固有表現と複数の共起文字列との関係についての統計情報を利用する．

2. 固有表現獲得手法

固有表現獲得の既存手法 [6] について説明する．まず，固有表現獲得手法で利用する拡張固有表現について述べ，次に既存手法の詳細な手順について説明する．

2.1 拡張固有表現

一般的な固有表現について解説し，拡張固有表現の必要性について述べる．

固有表現は一般的に7〜10種類のカテゴリに分類される．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 [7]．その後，1998-1999年のIREX(Information Retrieval and ExtractionExercise)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 [8]．しかし，この固有表現カテゴリでは分類できない固有表現への要求が発生した．

自然言語処理の分野においては，情報抽出の応用の広がりや質問応答という新しいタスクが出現し，上記の固有表現カテゴリには無いカテゴリが必要となった．この情報抽出と質問応答において新しいカテゴリが要求される例について述べる．情報抽出の例としては，「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．もう1つのタスク，質問応答の例としては「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．これらのカテゴリは10種類のカテゴリを持つ固有表現では対応できない．

この問題に対して，より多くの固有表現カテゴリを保持する拡張固有表現 [9] が存在する．拡張固有表現は固有表現のカテゴリ数を約200種類まで拡張したものであり，そのカテゴリ例としては，ACADEMIC, AWARD, BOOK, COMPANY, DISEASE, MOVIE, PERSON, SCHOOL, SPORTS, STATION...等が存在している．固有表現獲得手法では拡張固有表現を用いて固有表現獲得を行う．

2.2 手法の概要

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） コンテキストの関連度計算

（3） 固有表現獲得

本手法では，コーパス中に存在する既知の固有表現の周辺に出現する文字列が未知の固有表現の周辺においても出現するという仮説を軸とする．そこで，(1) で既存の固有表現辞書を用いて，コーパス中から固有表現周辺に出現する文字列を収集する．これ以降，固有表現周辺に出現する文字列をコンテキストと呼ぶ．コンテキストはコーパス中に大量に存在し，固有表現の判別に役立つもの，その逆の雑音となるものが混在している．その中から役に立つコンテキストとそうでないものを分別するために，(2) では各コンテキストと固有表現カテゴリとの関連の強さ（関連度）という概念を利用する．最後の (3) では役に立つコンテキストを利用することで固有表現をコーパス中から獲得する．

2.3 固有表現自動獲得

各ステップについて詳細に説明する．

2.3.1 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．本節では固有表現とコンテキストとの関係について説明する．図1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，その前後の文字列から助詞を除去したものをコンテキストとして獲得する．図1で示す例文のコンテキストは「今年#AWARD#受賞した」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とする）．

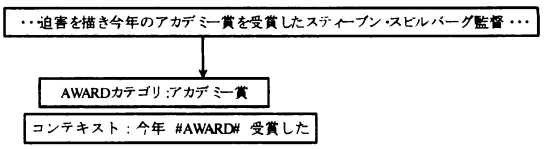

図1 コンテキストの例

本手法ではAWARDカテゴリに含まれる他の固有表現も「今年#AWARD#受賞した」というコンテキストによって同様に出現することを期待する．しかし，単一のコンテキストだけでは多様な文脈で出現するであろう固有表現に対応することはできないため，様々なコンテキストを獲得することが望ましい．このために，あらかじめ十分な量の固有表現辞書を用意する必要がある．本手法では約10万語が記載された拡張固有表現を利用している．

コンテキストの獲得と同時に，コンテキストとその固有表現カテゴリとの共起度数情報も取得する．これを用いて関連度を求める方法を次に述べる．

2.3.2 コンテキストの関連度計算

コンテキストとある固有表現カテゴリとの関わりの強さを表わす数値として関連度を考える．この値が高いコンテキストほどそのカテゴリに属する固有表現と共起しやすいと見なす．逆に，不要なコンテキストについては関連度は低くなるのが望ましい．先行研究 [6] の関連度を求める関数を示す．関連度関数を説明するための変数として，評価対象のコンテキスト CO，ある固有表現カテゴリ CA，CA 以外の固有表現カテゴリ群 $\overline{CA}$，CO と CA との共起出現頻度 a_i，CO と CA との共起出現種類数 a_t，CO と $\overline{CA}$ との共起出現頻度 b_i，CO と $\overline{CA}$ との共起出現種類数 b_t を定義する．

$$\text{関連度}：a_t \cdot \frac{a_t}{a_t + b_t} \cdot \frac{a_i}{a_i + b_i} \tag{1}$$

この関数は a_t を重視している．それに加えて，コンテキストがある単一の固有表現カテゴリとだけ共起しているかを表わす情報を利用する．これによって他のカテゴリにも出現するようなコンテキストの関連度を低くさせる．

この関連度を AWARD カテゴリに適用し，ランキングしたのが表 1 である．この表から，「受賞」を含むコンテキストに高い関連度が与えられていることがわかる．なお，表 1 は後述する評価実験と同じ実験環境において得られたコンテキストである．

表 1　コンテキストと AWARD カテゴリとの関連度

順位	関連度	コンテキスト
1	27	」 #AWARD# 受賞 。
2	23	」 #AWARD# 受賞 した
3	18	回 #AWARD# 受賞 。
4	18	、 #AWARD# 受賞 した
5	16	、 #AWARD# 受賞 。

2.3.3 固有表現獲得

先行研究 [6] と同様に本研究では関連度の高いコンテキストと共起する文字列を固有表現候補としてコーパス中から収集する．この際，得られた固有表現候補がどのコンテキストと共起したかという情報を保持しておく．これは得られた固有表現候補をランキングするための数値の算出に用いる．この数値は固有表現候補および共起した各コンテキストの関連度の総和とする．つまり，ある固有表現カテゴリとの関連が強いコンテキスト群と共起するほどにその固有表現候補がより固有表現らしいと判断する．表 2 は先行研究で AWARD カテゴリの固有表現候補として獲得された文字列と，そのスコアランキングの上位 5 件である．

表 2　AWARD カテゴリでの固有表現ランキング

順位	スコア	固有表現
1	239	直木賞
2	229	毎日芸術賞
3	178	菊池寛賞
4	165	ノーベル平和賞
5	156	アカデミー賞

3. 関連度関数の提案

固有表現カテゴリとコンテキストとの関連度を求める関数について，性能向上のための新しい情報と関連度関数を提案する．

3.1 利用する情報

関連度は 1 つの固有表現と 1 つのコンテキストとの関連の強さについての数値であり，あるカテゴリに属する固有表現とそのコンテキストとの共起度数から求める．共起度数情報の説明のため，AWARD カテゴリでの例を図 2 に示す．a は AWARD の固有表現と 1 つのコンテキスト「受賞した」との共起度数である．このコンテキスト「受賞した」と AWARD の固有表現との関連の強さをより尤もらしく求めることが固有表現獲得性能の向上にとって重要である．b は AWARD 以外の全ての固有表現と「受賞した」との共起度数である．c は AWARD の固有表現と「受賞した」以外のコンテキストとの共起度数，d は AWARD 以外の全ての固有表現と「受賞した」以外のコンテキストとの共起度数である．

ベースラインとなる関数はあるカテゴリの固有表現とコンテキストとの共起度数 a，b を利用している．これに対して提案する手法では a，b に加えて c，d といった他のコンテキストとの共起度数情報も利用している．この新しい情報 c，d の利用によってより尤もらしい関連度を求め，固有表現獲得性能を向上させることを検討する．

共起度数情報の利用については 2 通りの方法がある．1 つは固有表現とコンテキストが共起する全ての事例をカウントする延べ数であり，もう 1 つはコンテキストと共起する固有表現の種類数をカウントする異なり数である．ベースライン関数は延べ数と異なり数の両方を利用しているが，提案する手法では異なり数のみを利用する．

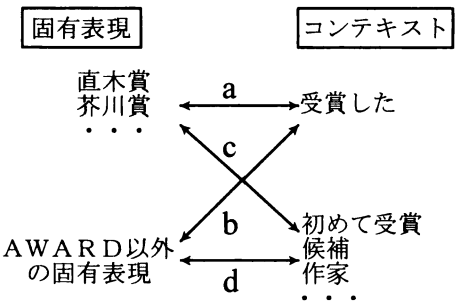

図 2　共起度数情報 a,b,c,d の例

3.2 関連度関数

関連度を求める 3 つの関数について説明する．既存の関数 (1) 式をベースライン関数とし，固有表現獲得性能向上のための 2 つの関数，ϕ 相関式，自己相互情報量を示す．次の 2 つの関数においては a, b, c, d は異なり数を利用している．また，自己相互情報量の n は a, b, c, d の総和である．

$$\phi \text{相関式} : \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \tag{2}$$

$$\text{自己相互情報量} : \log \frac{an}{(a+b)(a+c)} \tag{3}$$

4. 評価実験

3 つの関連度関数について，固有表現獲得の評価方法と実験結果について述べる．なお，関連度関数以外の手順は既存手法と同じ内容で実験した．

4.1 評価方法

固有表現獲得手法を実行して得られた固有表現について，適合率，再現率，F 値による評価を行う．正解判定のためには既存の辞書を利用する．このため，既存の固有表現辞書はコンテキスト獲得のための固有表現群と評価セットの 2 つに分割する．

分割は無作為に行う．既存の固有表現辞書を利用して正解判定をする評価方法では，辞書に含まれない固有表現が不正解と判定されることに注意が必要であるが，性能を相対的に比較することが容易に行えるという利点がある．この評価方法による適合率と再現率の式を示す．なお，$|A|$ は A 中に存在する固有表現の個数を意味する．

$$\text{適合率} = \frac{|\text{辞書の評価セット} \cap \text{出力結果}|}{|\text{出力結果}|} \quad (4)$$

$$\text{再現率} = \frac{|\text{辞書の評価セット} \cap \text{出力結果}|}{|\text{辞書の評価セット} \cap \text{コーパス}|} \quad (5)$$

本実験では適合率による評価を重視する．固有表現獲得手法の目的とする辞書の語彙増加のためには，より正確に固有表現を求めることが重要であると判断するからである

各カテゴリで獲得した固有表現について，よりスコアの高い固有表現を上位から正解判定をする．そして全てのカテゴリでの適合率の平均を求め，獲得した固有表現全体の評価とする．本実験では 89 種類のカテゴリで固有表現が得られた．

4.2 実験の準備

固有表現獲得手法を実行するためのデータとしては，固有表現辞書と固有表現を獲得するためのコーパスが必要である．固有表現辞書は関根らにより構築された日本語の拡張固有表現辞書で，約 10 万語の固有表現を有する．コーパスは毎日新聞の 1994 年〜1999 年の 6 年分の日本語記事を用い，あらかじめ形態素解析システム juman [10] で形態素解析をしておく．

4.3 実験結果

ベースライン関数と ϕ 相関式，自己相互情報量の適合率比較実験の結果を図 3 に示す．

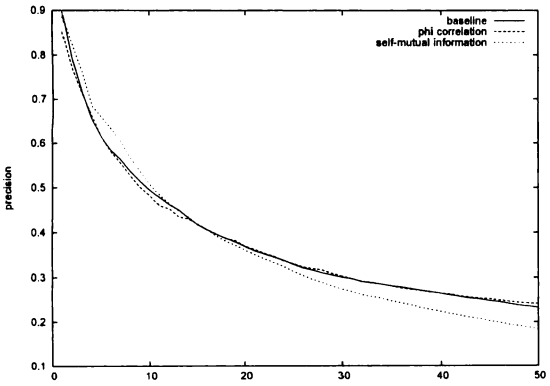

図 3 上位 50 件までの適合率

ϕ 相関式は全体的にベースライン関数に近い適合率であった．自己相互情報量は上位 10 件程度まではベースライン関数と ϕ 相関式をおおむね上回り，それ以降は下回っている．

固有表現辞書の語彙を増加するという本手法の目的からすれば，得られた固有表現の上位で高い適合率となる自己相互情報量がより適した関数であると考えられる．

参考として，再現率と F 値を図 4 と図 5 に示す．

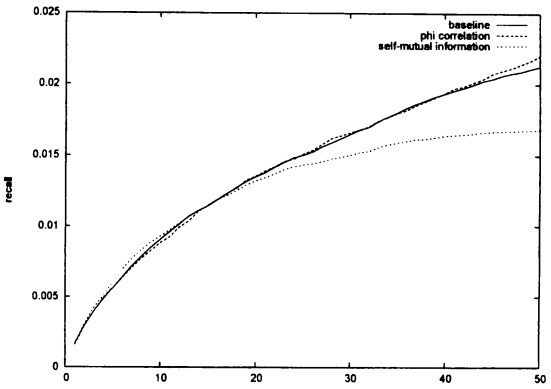

図 4 上位 50 件までの再現率

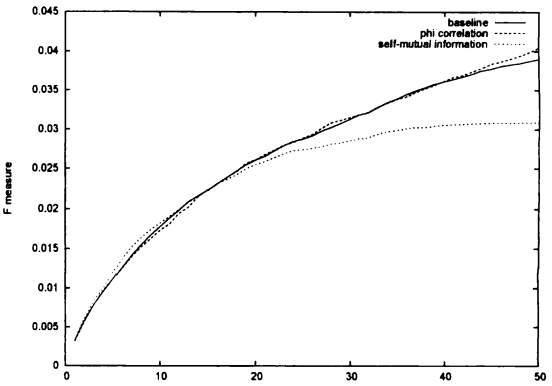

図 5 上位 50 件までの F 値

5. 考　　察

自己相互情報量を用いた結果が上位で優れており，下位ではベースラインと ϕ 相関式よりも劣る理由について考える．表 3 は ϕ 相関式で関連度が高いとされたコンテキストで，表 4 は自己相互情報量についての同様の表である．なお，いずれも AWARD カテゴリについて扱っている．ϕ 相関式では，コンテキスト毎に関連度に差がついており，関連度が高いと見なされるコンテキストがベースライン関数と同じである．一方の自己相互情報量では関連度が一定の値となっており，かつ関連度が高いコンテキストは ϕ 相関式やベースライン関数と異なったものを上位としている．自己相互情報量の関連度が一定の値を取るのは関連度が高い上位のコンテキストで見られる傾向である．表 4 の自己相互情報量のコンテキストを見ると [、#AWARD# 『 二] のような不要なコンテキストが混じっている．不要なコンテキストと他の良いコンテキストが同じ評価となっているために，自己相互情報量は不要なコンテキストの影響を受ける．このため，自己相互情報量では 1 つのコンテキストだけでなく，より多くのコンテキストと共起するような固有表現が有利だろう．そうでないと誤りが混入しやすい．ϕ 相関式については，多くのコンテキストと共起する固有表現候補な

らば信頼して良いと考えられる．上位の固有表現候補は下位よりも共起するコンテキストが多いので説明がつく．

このようなコンテキストに対して，得られた固有表現を表5と6で比較するとφ相関式よりも自己相互情報量の方が1位〜5位までのスコアの相対的な差が大きくなっている．φ相関式のスコアが1.56〜1.11に対して，自己相互情報量は48.34〜20.95である．コンテキストの場合とは逆の傾向である．この傾向は5位よりも下位の固有表現でも同様であり，自己相互情報量では1位のスコアに対する他の固有表現のスコアが低くなっていた．このことから，自己相互情報量は上位の固有表現に比べて下位のそれは正しい固有表現である期待が低いと考えられる．自己相互情報量を用いた場合の結果が上位に比べて，下位で低くなるという傾向は図3の結果と一致する．

表3　φ相関式でのコンテキスト

順位	関連度	コンテキスト
1	0.107	」 #AWARD# 受賞 。
2	0.099	」 #AWARD# 受賞 した
3	0.087	回 #AWARD# 受賞 。
4	0.082	、 #AWARD# 受賞 した
5	0.076	、 #AWARD# 受賞 。

表4　自己相互情報量でのコンテキスト

順位	関連度	コンテキスト
1	2.66	等 #AWARD# 授与 する
1	2.66	、 #AWARD# 終身 年金
1	2.66	緩 #AWARD# 関係 地方
1	2.66	、 #AWARD# 『
1	2.66	、 #AWARD# 該当 作

表5　φ相関式で得られた固有表現 (AWARD カテゴリ)

順位	スコア	固有表現
1	1.57	直木賞
2	1.38	毎日芸術賞
3	1.23	ノーベル平和賞
4	1.14	ノーベル文学賞
5	1.11	アカデミー賞

表6　自己相互情報量で得られた固有表現 (AWARD カテゴリ)

順位	スコア	固有表現
1	48.34	文化勲章
2	47.57	ノーベル文学賞
3	39.58	直木賞
4	37.30	褒章
5	20.95	芸術選奨

6. ま と め

固有表現獲得の性能を向上するために，新たな統計情報を用い，関連度を求めるための関数を提案した．その評価方法として既存の固有表現辞書を分割し，固有表現獲得と評価のための2セットを利用する方法を取った．固有表現語彙増加という目的のために，評価は適合率に注目している．結果としては，自己相互情報量は獲得した固有表現の上位でベースラインやφ相関式を上回り，下位ではベースラインを下回った．また，φ相関式はベースラインは同程度の性能であった．固有表現辞書の語彙増加のためには上位で優れた適合率をもたらす自己相互情報量が適していると考えられる．

謝　辞

この研究は，住友電工情報システムとの共同研究の成果である．この成果を分析するときに使用したシステムには，平成19年度科学研究費課題 (課題番号 19500120) の研究成果を使用した．また，本研究で利用した拡張固有表現辞書はニューヨーク大学の関根らによって構築されたものである．

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] 塩入寛之，関根聡，梅村恭司.「拡張固有表現獲得の精度向上」，情報処理学会自然言語処理研究会 (NL-180-12)，2007.

[7] R. Grishman, B. Sundheim. "Message Understanding Conference - 6: A Brief History", COLING-96.

[8] S. Sekine, H. Isahara. "IREX: IR and project in Japanese", LREC 2000.

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

[10] 黒橋禎夫，河原大輔. 日本語形態素解析システム JUMAN version 5.1，2005. http://www.kc.t.u-tokyo.ac.jp/nl-resource/.juman/juman-5.1.tar.gz.