

# 日本語固有表現抽出の難易度を示す指標の提案と評価

野 畑 周<sup>†</sup> 関 根 聡<sup>††</sup> 辻 井 潤一<sup>†††</sup>

本論文では、固有表現抽出の難易度をテストコーパスから評価する指標を提案する。固有表現抽出システムの性能は客観的な指標によって評価される。しかし、システムの出力に対する評価だけでは、あるコーパスに対する固有表現抽出がどのように難しいのか、どのような情報がそのコーパスに対して固有表現抽出を行なう際に有効なのかを知ることは難しい。本論文で提案する指標は、個々のシステムの出力に依存することなく、複数のコーパスについて統一的に適用できる。指標の有効性は固有表現抽出システムの性能評価と比較することで検証される。さらに固有表現のクラス間における難易度の比較や、有用な情報の違いについても議論する。

キーワード: 固有表現, 情報抽出, コーパス比較

## Analysis on Difficulty Indices for Japanese Named Entity Task

CHIKASHI NOBATA <sup>†</sup>, SATOSHI SEKINE <sup>††</sup> and JUN'ICHI TSUJII <sup>†††</sup>

We propose indices to measure the difficulty of the named entity (NE) task by looking at test corpora, based on expressions inside and outside the NEs. These indices are intended to estimate the difficulty of each task without actually using an NE system and to be unbiased towards a specific system. The values of the indices are compared with the systems' performance in Japanese documents. We also discuss the difference between NE classes with the indices and show useful clues which will make it easier to recognize NEs.

**KeyWords:** *Named entity, Information extraction, Corpus comparison*

## 1 はじめに

本研究の目的は、情報抽出のサブタスクである固有表現抽出 (Named Entity Task) の難易度の指標を定義することである。情報抽出とは、与えられた文章の集合から、「人事異動」や「会社合併」など、特定の出来事に関する情報を抜き出し、データベースなど予め定められた形式に変換して格納することであり、米国のワークショップ Message Understanding Conference (MUC) でタスクの定義・評価が行われてきた。固有表現 (Named Entity) とは、情報抽出の要素となる表現のことである。固有表現抽出 (Named Entity Task) は MUC-6(DARPA 1995) において初め

<sup>†</sup> 独立行政法人通信総合研究所けいはんな情報融合研究センター自然言語グループ, Computational Linguistic Group, Keihanna Human Info-Communication Research Center, Communications Research Laboratory

<sup>††</sup> ニューヨーク大学コンピュータサイエンス学科, Computer Science Department, New York University

<sup>†††</sup> 東京大学大学院情報理工学系研究科コンピュータ科学専攻, Graduate school of Information Science and Technology, University of Tokyo

て定義され、組織名 (Organization)、人名 (Person)、地名 (Location)、日付表現 (Date)、時間表現 (Time)、金額表現 (Money)、割合表現 (Percent) という 7 種類の表現が抽出すべき対象とされた。これらは、三つに分類されており、前の三つが entity names (ENAMEX)、日付表現・時間表現が temporal expressions (TIMEX)、金額表現・割合表現が number expressions (NUMEX) となっている。1999 年に開かれた IREX ワークショップ (IREX 1999) では、MUC-6 で定義された 7 つに加えて製品名や法律名などを含む固有物名 (Artifact) というクラスが抽出対象として加えられた。

固有表現抽出システムの性能は、再現率 (Recall) や適合率 (Precision)、そしてその両者の調和平均である F-measure といった客観的な指標<sup>1</sup>によって評価されてきた。しかし、単一システムの出力に対する評価だけでは、あるコーパスに対する固有表現抽出がどのように難しいのか、どのような情報がそのコーパスに対して固有表現抽出を行なう際に有効なのかを知ることが難しい。例えば、あるコーパスについて、あるシステムが固有表現抽出を行い、それらの結果をある指標で評価したとする。得られた評価結果が良いときに、そのシステムが良いシステムなのか、あるいはコーパスが易しいのかを判断することはできない。

評価コンテストを行い、単一のシステムでなく複数のシステムが同一のコーパスについて固有表現抽出を行い、それらの結果を同一の指標で評価することで、システムを評価する基準を作成することはできる。しかしながら、異なるコーパスについて、複数の固有表現抽出システムの評価結果を蓄積していくことは大きなコストがかかる。また、継続して評価を行なっていくとしても、評価に参加するシステムは同一であるとは限らない。異なるコーパスについて、個別のシステムとは独立に固有表現抽出の難易度を測る指標があれば、コーパス間の評価、また固有表現抽出システム間の評価がより容易になると考えられる。本研究は、このような指標を定義することを目指すものである。

## 1.1 固有表現抽出の難易度における前提

異なる分野における情報抽出タスクの難易度を比較することは、複数の分野に適用可能な情報抽出システムを作成するためにも有用であり、実際複数のコーパスに対して情報抽出タスクの難易度を推定する研究が行われてきている。Bagga et. al (Bagga and Biremann 1997) は、MUC で用いられたテストコーパスから意味ネットワークを作成し、それをを用いて MUC に参加した情報抽出システムの性能を評価している。固有表現抽出タスクに関しては、Palmer et.

<sup>1</sup> 再現率は、正解データ中の固有表現の数 (G) のうち、正しく認識された固有表現の数 (C) がどれだけであったかを示す。適合率は、固有表現とみなされたものの数 (S) のうち、正しく認識された固有表現の数 (C) がどれだけであったかを示す。F-measure は、両者の調和平均である。それぞれの評価基準を式で示せば以下ようになる。

$$\begin{aligned} \text{再現率 (R)} &= C / G \\ \text{適合率 (P)} &= C / S \\ \text{F-measure} &= 2PR / (P + R) \end{aligned}$$

表 1 IREX 固有表現抽出のテストコーパス

	予備試験	本試験	
		総合課題	限定課題
記事数	36	72	20
単語数	11173	21321	4892
文字数	20712	39205	8990

al (Palmer and Day 1997) が Multilingual Entity Task (DARPA 1998) で用いられた 6 カ国語のテストコーパスから，各言語における固有表現抽出技術の性能の下限を推定している．

本研究では，固有表現抽出の難易度を，テストコーパス内に現れる固有表現，またはその周囲の表現に基づいて推定する指標を提案する．指標の定義は，「表現の多様性が抽出を難しくする」という考えに基づいている．文章中の固有表現を正しく認識するために必要な知識の量に着目すると，あるクラスに含まれる固有表現の種類が多ければ多いほど，また固有表現の前後の表現の多様性が大きいほど，固有表現を認識するために要求される知識の量は大きくなると考えられる．

あらゆるコーパスを統一的に評価できるような，固有表現抽出の真の難易度は，現在存在しないので，今回提案した難易度の指標がどれほど真の難易度に近いのかを評価することはできない．本論文では，先に述べた「複数のシステムが同一のコーパスについて固有表現抽出を行った結果の評価」を真の難易度の近似と見なし，これと提案した指標とを比較することによって，指標の評価を行うことにする．具体的には，1999 年に開かれた IREX ワークショップ (IREX 1999) で行われた固有表現抽出課題のテストコーパスについて提案した指標の値を求め，それらと IREX ワークショップに参加した全システムの結果の平均値との相関を調べ，指標の結果の有効性を検証する．

このような指標の評価方法を行うためには，できるだけ性質の異なる数多くのシステムによる結果を得る必要がある．IREX ワークショップでは，15 システムが参加しており，システムの種類も，明示的なパターンを用いたものやパターンを用いず機械学習を行ったもの，またパターンと機械学習をともに用いたものなどがあり，機械学習の手法も最大エントロピーや HMM，決定木，判別分析などいくつかバラエティがあるので，これらのシステムの結果を難易度を示す指標の評価に用いることには一定の妥当性があると考えている．

## 1.2 IREX ワークショップの固有表現抽出課題

IREX ワークショップの固有表現抽出課題では，予備試験を含め，3 種類のテストコーパスが評価に用いられた．表 1 に各々の記事数，単語数，文字数を示す．単語の切り分けには JUMAN3.3 (松本，黒橋，山地，妙木，長尾 1997) を用い，単語の切り分けが固有表現の開始位置・終了位置と異なる場合には，その位置でさらに単語を分割した．

表 2 IREX 固有表現抽出の性能評価

クラス	予備試験	本試験	
		総合課題	限定課題
組織名	55.6	57.3	55.2
人名	71.3	67.8	68.8
地名	65.7	69.8	68.1
固有物名	18.8	25.5	57.9
日付表現	83.6	86.5	89.4
時間表現	69.4	83.0	89.8
金額表現	90.9	86.4	91.4
割合表現	100.0	86.4	—
全表現	66.5	69.5	71.7

IREX ワークショップに参加した固有表現抽出システムの性能評価は F-measure で示されている。表 2 に各課題における F-measure の値を示す。本試験の評価値は、IREX ワークショップに参加した全 15 システムの平均値である。一方、予備試験においては、全システムの評価は利用できなかったため、一つのシステム(野畑 1999) の出力結果を評価した値を用いている。このシステムは、決定木を生成するプログラム(Quinlan 1993) を用いた固有表現抽出システム(Sekine, Grishman and Shinnou 1998) を IREX ワークショップに向けて拡張したものである。

IREX では、8 つの固有表現クラスが定義された。表 2 から、最初の 4 つの固有表現クラス(組織名、人名、地名、固有物名) は残り 4 つの固有表現クラス(日付表現、時間表現、金額表現、割合表現) よりも難しかったことが分かる。以下では、両者を区別して議論したいときには、MUC での用語に基づき前者の 4 クラスを「ENAMEX グループ」と呼び、後者の 4 クラスを「TIMEX-NUMEX グループ」と呼ぶことにする。

### 1.3 指標の概要

以下、本稿では、まず固有表現内の文字列に基いて、固有表現抽出の難易度を示す指標を提案する。ここで提案する指標は 2 種類ある。

- Frequency of tokens: 各固有表現クラスの頻度と異なり数を用いた指標 (2 節)
- Token index: 固有表現内の個々の表現について、その表現のクラス内における頻度とコーパス全体における頻度を用いた指標 (3 節)

これらの指標の値を示し、それらと実際のシステムの評価結果との相関を調べた結果について述べる。

次に、固有表現の周囲の文字列に基いた指標についても、固有表現内の文字列に基いた指標と同様に 2 種類の指標を定義し、それらの値とシステムの評価結果との相関の度合を示す (4 節)。

表 3 各クラスの固有表現の異なり数

クラス	予備試験	本試験	
		総合課題	限定課題
組織名	131	187	48
人名	113	217	71
地名	89	191	78
固有物名	31	39	9
日付表現	71	126	49
時間表現	16	32	15
金額表現	28	13	7
割合表現	6	16	-
全表現	482(485)	818(821)	277

## 2 Frequency of tokens

本節では，固有表現クラスに含まれる文字列の頻度と異なり数とを用いて，固有表現抽出の難易度を示す指標について述べる．このような指標は，ある固有表現クラス内において，異なる文字列が数多く現れるならば，そのクラスの固有表現を認識することは難しくなる，という仮定に基づいている．頻度や異なり数を考慮する文字列の単位には，固有表現そのもの，単語，また単一の文字をとることができる．

### 2.1 固有表現を単位とする指標

まず，固有表現そのものを単位として分析を行なう．表 3 に，各クラスがもつ固有表現の異なり数を示す．予備試験と本試験の総合課題では，全表現の異なり数が各クラスの異なり数の合計よりも少ない．これは，複数のクラスに分類される固有表現がそれぞれ 3 個ずつあったからである．また，限定課題には割合表現が現われなかったので，数値が入っていない．

異なり数を指標として用いるには，コーパスサイズの影響を除く必要がある．最初に定義する指標は，各クラスについて固有表現の異なり数を出現頻度で正規化したものである．以下これを FE(Frequency of Entities) と呼ぶ．FE の定義を式で示せば式 1 となる． $D_E$  は各クラスに含まれる固有表現の異なり数， $N_E$  は各クラス内の固有表現の総出現数である．FE は，あるクラス内の固有表現を抽出することが難しいときに，指標の値が大きくなることを意図して定義されている．

$$FE = \frac{D_E}{N_E} \quad (1)$$

FE の値を求める際には，文章中に現れる数字を全て文字 “#” で置き換えた．これは，各数字を異なる表現とみなすよりも同一の表現とみなす方が固有表現の多様性を捉える際にはよ

表 4 各クラスの FE の値

クラス	予備試験	本試験	
		総合課題	限定課題
組織名	0.61 (=131/214)	0.48 (=187/389)	0.65 (=48/ 74)
人名	0.67 (=113/169)	0.61 (=217/355)	0.73 (=71/ 97)
地名	0.46 (=89/192)	0.46 (=190/416)	0.75 (=78/106)
固有物名	0.71 (=30/ 42)	0.80 (=39/ 49)	0.69 (=9/ 13)
日付表現	0.33 (=36/110)	0.18 (=51/277)	0.24 (=17/ 72)
時間表現	0.46 (=11/ 24)	0.27 (=16/ 59)	0.53 (=10/ 19)
金額表現	0.09 (= 3/ 33)	0.13 (= 2/ 15)	0.13 (= 1/ 8)
割合表現	0.50 (= 3/ 6)	0.29 (= 6/ 21)	— —
全表現	0.53 (=415/790)	0.45 (=706/1581)	0.60 (=235/389)

り適切であるという判断による．数字を同一の文字とみなすことによって TIMEX-NUMEX グループに含まれる固有表現クラスの FE の値は小さくなるが，これは TIMEX-NUMEX グループの認識精度が非常に高いという結果に合致する．FE の値を表 4 に示す．

## 2.2 単語，文字単位の指標

前節では固有表現そのものを指標を計算する単位として用いたが，単語や文字を単位としても同様に指標を定義することができる．固有表現よりも短かく頻度の大きい単語や文字を単位にすることで，よりコーパスサイズの影響を受けにくい指標が得られると期待される．以下，単語単位の指標を  $FW$ ，文字単位の指標を  $FC$  と呼ぶ． $FW$ ， $FC$  の定義は  $FE$  と同様に，それぞれ式 2，式 3 によって表わせる．

$$FW = \frac{D_W}{N_W} \tag{2}$$

但し:

$D_W$  各固有表現クラスに含まれる単語の異なり数

$N_W$  各固有表現クラスに含まれる単語の総出現数

$$FC = \frac{D_C}{N_C} \tag{3}$$

但し:

$D_C$  各固有表現クラスに含まれる文字の異なり数

表 5 各クラスの FC の値

クラス	予備試験	本試験	
		総合課題	限定課題
組織名	0.29 (=258/883)	0.20 (=365/1792)	0.38 (=139/365)
人名	0.39 (=222/575)	0.26 (=319/1228)	0.48 (=148/311)
地名	0.30 (=186/618)	0.19 (=284/1491)	0.34 (=155/462)
固有物名	0.53 (=131/245)	0.50 (=175/ 347)	0.58 (= 34/ 59)
日付表現	0.16 (= 44/282)	0.07 (= 54/ 737)	0.07 (= 15/226)
時間表現	0.18 (= 12/ 66)	0.09 (= 16/ 182)	0.14 (= 10/ 71)
金額表現	0.06 (= 4 / 72)	0.09 (= 3/ 34)	0.13 (= 2/ 16)
割合表現	0.38 (= 5 / 13)	0.10 (= 7/ 58)	— —
全表現	0.20 (=555/2754)	0.12 (=717/5869)	0.24 (=355/1510)

$N_C$  各固有表現クラスに含まれる文字の総出現数

FE と同様に，FW・FC においても数字は同一の文字とみなして値を求めた．FW と FC の値の傾向は似通っているので，ここでは FC の値のみを示す (表 5)．FC ではクラス間の差が FE よりも際だっており，特に TIMEX-NUMEX グループ内のクラスに対する FC の値はきわめて小さい．

### 2.3 指標の有効性

指標の有効性を確かめるために，各指標がシステムの評価結果とどの程度相関しているかを調べる．まず，各固有表現クラスに対する FE・FW・FC の値と F-measure との散布図を予備試験 (図 1)，本試験の総合課題 (図 2)，限定課題 (図 3) それぞれについて示す．どの図においても，縦軸に指標の値，横軸に F-measure の値をとっている．各クラスに対応する F-measure の値には，以下のような，クラス名を示す英字 3 文字のラベルを付した：

組織名 (ORG)，人名 (PRS)，地名 (LOC)，固有物名 (ART)，

日付表現 (DAT)，時間表現 (TIM)，金額表現 (MON)，割合表現 (PRC)

予備試験と総合課題においては，TIMEX-NUMEX グループが右下にまとめられ，固有物名を除いた ENAMEX グループがややその左上に位置する．その左上に位置しているのは，固有物名である．限定課題においては，固有物名は他の ENAMEX グループに属するクラスと同様の位置にある．F-measure の値においても，指標の値においても，値の傾向としてはほぼ同様で

表 6 指標と F-measure との相関

課題	FE	FW	FC
予備試験	-0.66	-0.63	-0.61
本試験 (総合)	-0.91	-0.92	-0.97
本試験 (限定)	-0.80	-0.87	-0.89

あるといえる。

各クラスごとに FE・FW・FC の値と F-measure との相関係数を求めた結果を表 6 に示す。これらの指標は、固有表現の抽出が難しいときに値が大きくなることを意図して定義されたものである。即ち、F-measure の値との負の相関が高くなることを意図して作成された指標である。

表 6 から、FW・FC は予備試験のコーパスにおいては FE よりも相関が弱いですが、本試験のコーパスにおいては総合・限定課題どちらにおいても FE より相関が強いことが分かる。予備試験に対するシステムの評価結果は 1 つのシステムによるものであることを考慮すると、本試験の二つの課題に対して相関が強いほうが指標としてより信頼できる。本試験のコーパスに対する結果から、固有表現よりも単語の方が、単語よりも文字の方が指標の値を求める単位としては安定しているといえる。

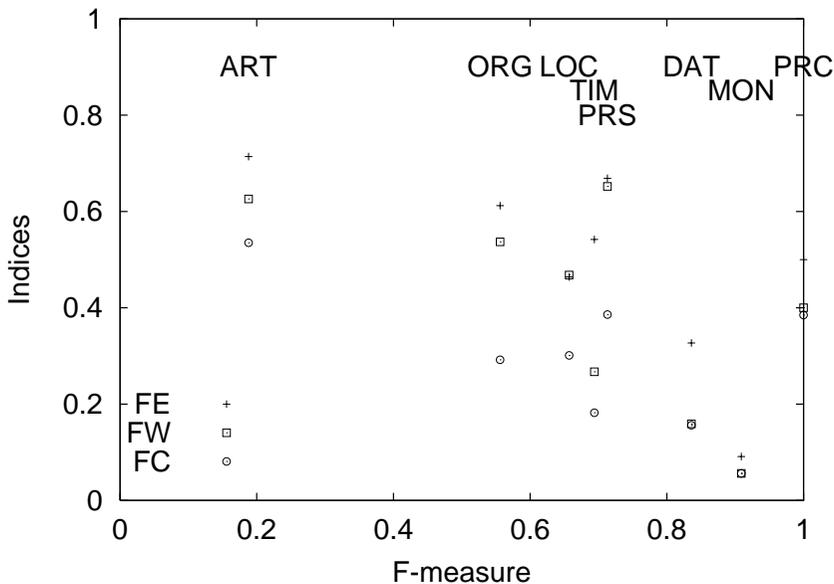


図 1 指標と F-measure との散布図 (予備試験)

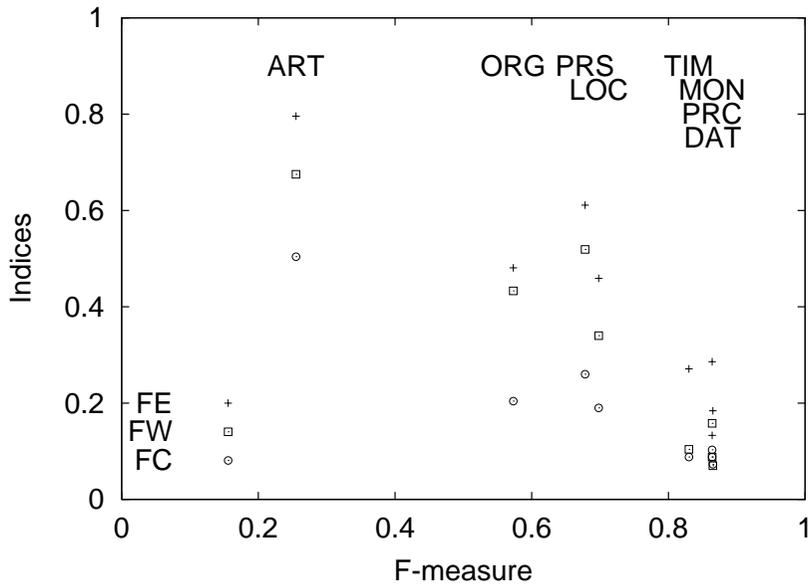


図 2 指標と F-measure との散布図 (総合課題)

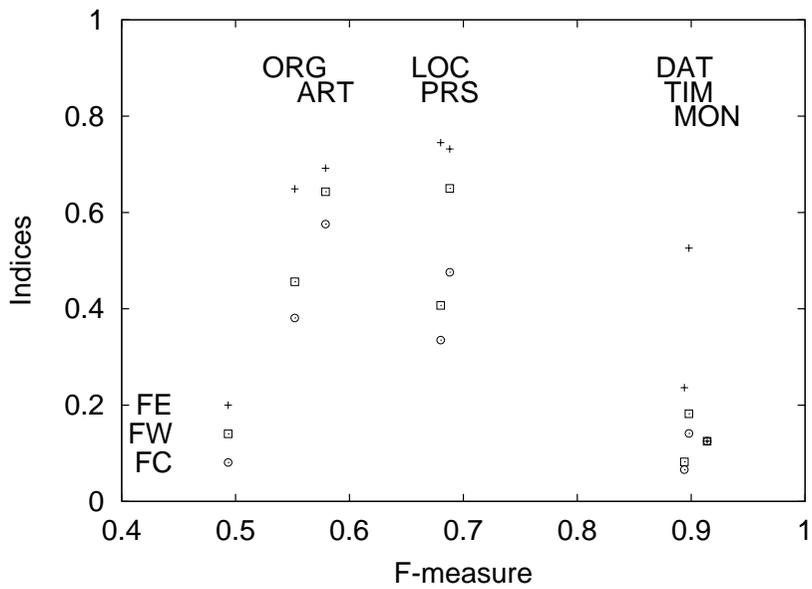


図 3 指標と F-measure との散布図 (限定課題)

### 3 Token index

本節では、固有表現内の個々の表現について、その表現のクラス内における頻度とコーパス全体における頻度との関係に基いて固有表現抽出の難易度を示すことを考える。これは、あるクラスに相対的に関連の強い文字列が多いほど、そのクラスの固有表現を抽出することはより易くなるという仮定に基づいている。先に定義した指標では、クラス内の頻度のみを用いており、個々の固有表現内の文字列については考慮していなかった。本節で考える指標では、ある文字列の固有表現クラスとの関連の強さを、その文字列のクラス内での頻度とコーパス全体の頻度の双方を用いて定義する。つまりある文字列の頻度が高く、かつそのほとんどが特定の固有表現クラス内に限られるならば、その文字列はそのクラスと関連が強くなり、そのような文字列が多いほどそのクラスにおける固有表現の抽出は易くなるという仮定に基づく。

#### 3.1 文字単位の指標

以下では、文字を単位として指標を定義する。文字を単位として選んだのは、先に定義された指標の中では、文字を単位とした指標が最もシステムの評価結果との相関が強かったためである。

まず、各文字ごとの指標  $CI_c$  を定義する。文字  $c$  のクラス  $L$  に対する  $CI_c$  の値は、式 4 によって与えられる。 $n_L(c)$  は文字  $c$  のクラス  $L$  における頻度、 $n_{(c)}$  はコーパス全体での頻度を現わす。 $N_{CL}$  はクラス  $L$  内の総文字数である。つまり、右辺第 1 項  $\frac{n_L(c)}{N_{CL}}$  はクラス  $L$  での文字  $c$  の相対頻度を示し、第 2 項  $\frac{n_L(c)}{n_{(c)}}$  は文字  $c$  がクラス  $L$  にどれだけ偏って現れるかを示している。なので、 $CI_c$  は文字  $c$  のクラス  $L$  における偏りを相対頻度で正規化したものとなる。

$$CI_c = \frac{n_L(c)}{N_{CL}} \frac{n_L(c)}{n_{(c)}} \tag{4}$$

各固有表現クラスに現れる全文字の  $CI_c$  の値を合計した値を、新たな指標 CI(Character Index) として用いることにする。

$$CI = \sum_{c \in C^L} CI_c \tag{5}$$

この指標は、固有表現の抽出が易いときに値が大きくなることを意図して定義されたものである。従って、システムの評価結果との正の相関が強ければ、指標として優れていることになる。

$CI_c$  は、クラス  $L$  の表現に文字  $c$  が生じる条件付き確率  $p(c|L)$  と、文字  $c$  があったときにそれがクラス  $L$  の表現の一部である条件付き確率  $p(L|c)$  との積を推定する式となっている。

$$CI_c = p(c|L) \cdot p(L|c)$$

$CI_c$  は、文字  $c$  の出現確率  $p(c)$ 、クラス  $L$  内の文字が出現する確率  $p(L)$ 、文字  $c$  とクラス  $L$

表 7 各クラスの CI の値

クラス	予備試験	本試験	
		総合課題	限定課題
組織名	0.34	0.31	0.45
人名	0.51	0.45	0.59
地名	0.38	0.40	0.56
固有物名	0.21	0.15	0.27
日付表現	0.39	0.48	0.60
時間表現	0.36	0.40	0.47
金額表現	0.47	0.51	0.51
割合表現	0.33	0.27	—
全表現	0.57	0.58	0.71

の同時確率  $p(c, L)$  を用いて次のように変形できる．

$$CI_c = \frac{p(c, L)^2}{p(c) \cdot p(L)}$$

これは，文字  $c$  とクラス  $L$  に対する相互情報量に基づく尺度 (式 6) に類似する．

$$MI_c = \log_2 \left( \frac{p(c, L)}{p(c) \cdot p(L)} \right) \quad (6)$$

異なる点は， $\log$  を取っていないことと，同時確率  $p(c, L)$  が二乗になっていることである．この違いによって，文字  $c$  がクラス  $L$  にのみ出現する場合，相互情報量に基づく尺度では，その文字の頻度に関わらず一定値になるのに対し， $CI_c$  の値では，さらにその文字がクラス  $L$  の全表現のうちどのくらいの割合を占めるかを指標として含むことができる．また， $CI_c$  の定義は， $CI_c$  の総和として  $CI$  を求める際に必要な正規化となっており，クラス  $L$  内の全ての文字が  $L$  にのみ現れるならば， $CI$  は最大値 1 をとる．これに対し，相互情報量に基づく尺度では，そのクラス内での文字の分布により最大値は一定でない．

### 3.2 CI の有効性

表 8 に， $CI$  とシステムの評価結果との相関係数の値を示す． $CI$  とシステムの評価との相関は先に定義した指標のそれと比べると低く，指標としては十分でないことを示している．相関が低い理由の一つとしては， $CI$  の値が，各固有表現クラスに含まれる全文字の  $CI_c$  の値を合計した値であることが考えられる． $CI_c$  の値が低い文字はそのクラスに含まれる固有表現を抽出するのに有用であるとはいえないので，そのような文字は  $CI$  を求める際に取り除く必要がある． $CI_c$  の値に対する閾値を設け，閾値以上の値についてのみ  $CI$  の値に加えることで， $CI$  の値をより指標として優れたものにできると考えられる．

図 4 は， $CI_c$  に対する閾値と相関係数との関係を示すグラフである． $CI_c$  に対する閾値を示す

表 8 CI と F-measure との相関

CI に対する条件	予備試験	本試験 (総合)	本試験 (限定)
閾値なし	0.62	0.75	0.49
最大値 ( $CI_c$ への閾値)	0.86 (0.005)	0.88 (0.004)	0.96 (0.009)
予備試験の閾値に対する値	-	0.86 (0.005)	0.95 (0.005)

軸には対数軸を取っている。グラフから、3種類のテストコーパス全てについて相関係数の値は一旦上昇し、その後低下していることが分かる。各々の相関係数の最大値と、それに対応する閾値は表 8 に示してある。これらの値は前節で提示した指標の相関係数と同程度になっている。

もっとも、相関係数の最大値を与える閾値は、システムの評価結果を用いて初めて明らかになるので、新しいタスクのテストコーパスにおいては、事前に閾値を何らかの方法で決定する必要がある。新しいタスクにおいて閾値を求める方法の一つとしては、予め閾値を求めるために本当に評価したいコーパスと同じ種類のデータを用意し、同じ固有表現クラスの定義を用いて複数の参加システムについて実験をしておき、そこで得られた閾値を、本当に評価したいコーパスについて用いることが考えられる。例えば、性質の似た 2 種類のコーパスを用いて予備試験と本試験を行い、それぞれについて複数システムの評価結果を得ることができれば、予備試験の結果から閾値を得て本試験に用いることができる。今回の実験においては、予備試験に対して 1 システムの結果のみを用いているが、それでもその結果から得られた閾値を本試験のコーパスに対して用いるならば、表 8 の最後の行に示すように、最大値に近い相関係数の値が得られるので、この方法によって妥当な閾値が得られたといえる。

CI の値の振舞いをより詳しく調べるために、固有表現クラスを ENAMEX グループと TIMEX-NUMEX グループの二つに分け、各々について  $CI_c$  の値が大きい順に文字を並べて  $CI_c$  の値の変化を示したのが図 4 である。TIMEX-NUMEX グループにおいては  $CI_c$  の値が他に比べて極立って大きい文字がいくつか存在するのに対し、ENAMEX グループにはそのような文字は存在せず、なだらかに  $CI_c$  の値が変化していくことがグラフから見てとれる。このことは、ENAMEX グループの固有表現には多くの文字がほぼ同程度に関連しているが、極立って強い関連を持つものはなく、固有表現を抽出する際にはほぼ全ての文字を考慮する必要があること、一方 NUMEX-TIMEX グループの固有表現には、少数の文字が非常に強く関連していることを示唆している。

### 3.3 $CI_c$ による文字の重要度

本節では、 $CI_c$  の値に基づいて、固有表現を抽出する際に有用と思われる文字を具体的に挙げて述べる。

表 9 に TIMEX-NUMEX グループにおいて  $CI_c$  の値が大きい文字を示す。対象課題は本試

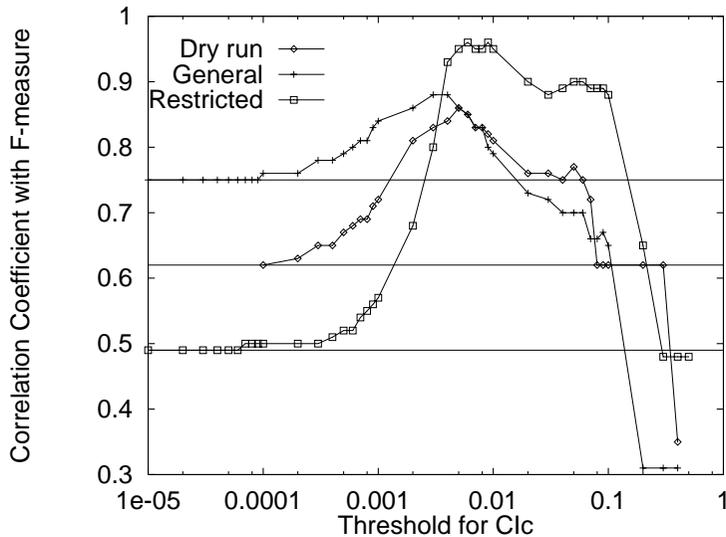


図 4 CI と F-measure との相関の変化

験の総合課題である。「#」は数字全体を示している。前節で見たように、TIMEX-NUMEX グループには、 $Cl_c$  の値が非常に大きい文字がいくつか存在する。これらの文字がそのクラスに属する表現と強く結びついていることは人間の直観から見ても妥当だといえる。実際、金額表現クラスにおける「円」、割合表現クラスにおける「%」の  $Cl_c$  の値は非常に大きく、各クラスに対する CI の値の半分以上を占めている。一方、コーパス中の数字の頻度は非常に大きい、TIMEX-NUMEX グループ内の各クラスに同様に現れるため、日付表現以外では  $Cl_c$  の値は小さい。

次に、ENAMEX グループにおける各文字の  $Cl_c$  の値の傾向を調べる。表 10 に、ENAMEX グループにおいて  $Cl_c$  の値が大きい文字を示す。対象課題は同様に本試験の総合課題である。これを見ると、人名以外の 3 つのクラスにおいては、接尾辞として用いられる文字において  $Cl_c$  の値が比較的大きいことが分かる。これをより明確に示すために、ENAMEX グループにおいて  $Cl_c$  を文字 bi-gram について求めた結果を表 11 に示す。[BOE] は固有表現の開始、[EOE] は終了を示す。文字 bi-gram に対する結果からは、組織名クラスにおける「党」や「銀」、固有物名における「法」、地名における「市」や「国」など、いくつかの接尾辞に対して高い  $Cl_c$  の値が得られた。これらの接尾辞が、特定の固有表現クラスに属する表現と強く結びついていることは人間の直観から見て妥当だといえる。今回の実験では、固有表現中の先頭にあるか末尾にあるかといった位置の情報は用いなかったが、このような位置情報を取り入れることで、指標の値から固有表現抽出に必要な知識の一部をより効率良く得ることができると考えられる。

表 9 TIMEX-NUMEX グループ内で  $CI_c$  の値が大きい文字

クラス	$CI_c$	$n_L(c)$	文字
日付表現	0.1113	277	#
	0.1071	143	日
	0.0931	75	月
	0.0893	98	年
	0.0421	31	昨
時間表現	0.1868	34	午
	0.0586	32	時
	0.0368	23	後
	0.0352	8	夜
	0.0330	6	夕
金額表現	0.4412	15	円
	0.0588	2	銭
	0.0091	17	#
割合表現	0.1379	8	%
	0.0616	5	倍
	0.0276	4	半
	0.0212	4	割
	0.0134	27	#

## 4 固有表現周囲の文字列に基づく指標

固有表現内の文字列に関する分析だけでは、難易度を調べるのに十分ではない。あるクラス内の固有表現が多様であったとしても、その周囲の表現が定まっているならば、そのクラスの固有表現抽出に関する難易度は小さくなると考えられる。本節では、固有表現の周囲の表現に着目して新たな指標を定義し、その有効性を先に定義した指標と同様に検証する。以下では指標を求める際の文字列の単位としては全て単語を用いている。

### 4.1 Frequency of context words

まず、FE・FW・FCと同様に、固有表現の周囲の単語について、その頻度と異なり数に基づいた指標 FCW(Frequency of context words) を定義する。FCW は固有表現クラスの周囲  $m$  語以内の単語を対象とする指標であり、式 7 のように定義される。

$$FCW = \frac{DCW_m}{NCW_m} \tag{7}$$

但し:

$DCW_m$  各固有表現クラスの周囲  $m$  語以内に現れる単語の異なり数

$NCW_m$  各固有表現クラスの周囲  $m$  語以内に現れる単語の総出現数

周囲の単語とみなす範囲  $m$  を、固有表現の直前または直後 1 単語から最大 4 単語まで変化

表 10 ENAMEX グループ内で  $CI_c$  の値が大きい文字

クラス	$CI_c$	$n_L(c)$	文字
組織名	0.0177	41	銀
	0.0159	43	党
	0.0108	22	庁
	0.0106	19	衆
	0.0087	22	A
人名	0.0200	34	原
	0.0172	35	田
	0.0155	19	郎
	0.0126	18	藤
	0.0109	21	山
地名	0.0323	51	米
	0.0161	36	市
	0.0151	30	京
	0.0125	24	ボ
	0.0124	33	東
固有物名	0.0206	20	法
	0.0080	6	商
	0.0058	2	仙
	0.0043	3	賞
	0.0038	2	鳳

表 11 ENAMEX グループ内で  $CI_c$  の値が大きい文字 bi-gram

クラス	$CI_c$	$n_L(c)$	文字 bi-gram
組織名	0.0125	39	党 [EOE]
	0.0119	27	長銀
	0.0111	26	銀 [EOE]
	0.0110	24	自民
	0.0101	22	民党
人名	0.0126	20	上原
	0.0120	19	郎 [EOE]
	0.0107	17	原 [EOE]
	0.0074	24	[BOE] 上
	0.0069	11	佐藤
固有物名	0.0146	14	法 [EOE]
	0.0130	6	商法
	0.0057	5	[BOE] 商
	0.0057	3	ドラ
	0.0051	2	鳳仙
地名	0.0252	49	[BOE] 米
	0.0163	31	米 [EOE]
	0.0139	46	[BOE] 日
	0.0136	36	本 [EOE]
	0.0121	36	日本
	0.0110	21	京都
	0.0104	26	市 [EOE]
	0.0100	44	国 [EOE]

表 12 FCW と F-measure との相関

課題	FCW <sub>pre</sub> : 直前の単語			
	1 語	2 語	3 語	4 語
予備試験	0.50	0.22	0.20	0.18
本試験 (総合)	0.16	-0.05	0.01	0.01
本試験 (限定)	-0.56	-0.36	0.00	0.16
課題	FCW <sub>fol</sub> : 直後の単語			
	1 語	2 語	3 語	4 語
予備試験	0.58	0.50	0.49	0.46
本試験 (総合)	0.34	0.43	0.23	0.26
本試験 (限定)	0.06	0.13	0.34	0.49

させ、指標の値を求めた。また、固有表現の直前に現われる単語に関する指標  $FCW_{pre}$  と、固有表現の直後に現れる単語に関する指標  $FCW_{fol}$  とをそれぞれ求めた。

$$FCW_{pre} = \frac{DCW_{pre_m}}{NCW_{pre_m}} \quad (8)$$

但し:

$DCW_{pre_m}$  各固有表現クラスの直前  $m$  語以内に現れる単語の異なり数  
 $NCW_{pre_m}$  各固有表現クラスに直前  $m$  語以内に現れる単語の総出現数

$$FCW_{fol} = \frac{DCW_{fol_m}}{NCW_{fol_m}} \quad (9)$$

但し:

$DCW_{fol_m}$  各固有表現クラスの直後  $m$  語以内に現れる単語の異なり数  
 $NCW_{fol_m}$  各固有表現クラスに直後  $m$  語以内に現れる単語の総出現数

指標とシステムの評価結果との相関を表 12に示す。負の相関が強いほどこの指標の値がシステムの結果とよく合致していることになるが、相関係数の値から、FCW は指標として適切であるとはいえない。

## 4.2 Context word index

固有表現の周囲の単語を用いた新たな指標として、CI と同様に CWI(Context Word Index) を定義する。CWI の定義は式 10で与えられる。

$$CWI_w = \frac{1}{m} \frac{n_L(w)}{N_{WL}} \frac{n_L(w)}{n(w)}$$

$$CWI = \sum_{w \in W_m^L} CWI_w \quad (10)$$

表 13 各クラスの CWI の値 ( $m = 1$ )

クラス	予備試験		総合課題		限定課題	
	CWIp <sub>pre</sub>	CWIf <sub>ol</sub>	CWIp <sub>pre</sub>	CWIf <sub>ol</sub>	CWIp <sub>pre</sub>	CWIf <sub>ol</sub>
組織名	0.23	0.30	0.16	0.22	0.15	0.20
人名	0.18	<b>0.47</b>	0.17	<b>0.53</b>	0.16	<b>0.58</b>
地名	0.22	0.35	0.20	0.21	0.29	0.27
固有物名	0.09	0.10	0.05	0.18	0.02	0.56
日付表現	0.13	0.25	0.15	0.22	0.14	0.33
時間表現	<b>0.29</b>	0.07	<b>0.29</b>	0.20	<b>0.44</b>	0.40
金額表現	0.14	0.20	0.25	0.28	0.37	0.45
割合表現	0.07	0.04	0.12	0.27	—	—
全表現	0.32	0.41	0.30	0.36	0.34	0.43

$m$  は固有表現の周囲の単語とみなされる語の範囲を示し，右辺第 1 項  $\frac{1}{m}$  は，範囲  $m$  を大きくしたときに頻度を補正するための項である． $W_m^L$  は，固有表現クラス  $L$  の周囲  $m$  語以内に現れる単語の集合を示す． $n_L(w)$  は，単語  $w$  がクラス  $L$  の固有表現の周囲  $m$  語以内に現れる頻度， $n_{(w)}$  はコーパス全体での頻度を現わす． $N_{WL}$  はクラス  $L$  の固有表現周囲に現れる単語の総数である．すなわち，右辺第 2 項  $\frac{n_L(w)}{N_{WL}}$  はクラス  $L$  に対する単語  $w$  の相対頻度を示し，第 3 項  $\frac{n_L(w)}{n_{(w)}}$  は単語  $w$  がクラス  $L$  に属する表現の周囲  $m$  語以内にどれだけ偏って現れるかを示す．表 13 に， $m = 1$  のときの各クラスにおける CWI の値を示す．FCW と同様に，直前の単語に関する指標 CWIp<sub>pre</sub> と直後の単語に関する指標 CWIf<sub>ol</sub> とを別々に求めた．各課題における指標の値のうち，クラス間で最も大きいものを太字で示した．

### 4.3 CWI<sub>w</sub> による単語の重要度

固有表現の周囲の単語とみなす範囲を，固有表現の直前または直後 1 単語から最大 4 単語まで変化させ，システムの評価結果との相関を調べた．結果を表 14 に示す．ここでは正の相関が強いほどシステムの結果とよく合致していることを表わす．CWI の指標としての有効性は FCW より高いが，その他の指標と比べると低い．

CWI は固有表現の周囲の表現がもつ情報を十分に利用しているとはいえないが，しかし課題や固有表現クラスによっては，人間の直観に沿うような結果が得られている． $m = 1$  のときの結果から，具体的な単語の例を表 15，表 16，また表 17 に示す．これらの表において単語に添えられている値は，単語ごとの指標 CWI<sub>w</sub> と，あるクラスに属する固有表現の前または後に現われた頻度  $n_L(w)$  である．

表 13 から 3 種類の課題全てにおいて時間表現クラスは他のクラスより CWIp<sub>pre</sub> の値が大きいことが分かるが，これは表 15 に示すように，時間表現の直前には日付表現がよく現われていることによる．この逆が成り立たないことは，日付表現クラスの CWIf<sub>ol</sub> の値が時間表現の

表 14 CWI と F-measure との相関

課題	CWIpre: 直前の単語			
	$m = 1$	$m = 2$	$m = 3$	$m = 4$
予備試験	-0.07	-0.34	-0.53	-0.49
本試験 (総合)	0.66	-0.01	-0.01	-0.04
本試験 (限定)	0.67	0.39	0.46	0.20
課題	CWIfol: 直後の単語			
	$m = 1$	$m = 2$	$m = 3$	$m = 4$
予備試験	-0.01	-0.24	-0.07	-0.09
本試験 (総合)	0.14	0.29	0.00	0.02
本試験 (限定)	0.06	0.46	0.36	0.10

表 15 総合課題の時間表現に対して CWIpre<sub>w</sub> の値が大きい単語

CWIpre <sub>w</sub>	$n_L(w)$	単語
0.1805	35	#日
0.0920	8	同日
0.0086	1	#年#月#日
0.0067	5	同
0.0057	1	昨年#月#日

表 16 人名に対して CWIfol<sub>w</sub> の値が大きい単語

課題	CWIfol <sub>w</sub>	$n_L(w)$	単語
本試験 (限定)	0.0471	30	容疑者
本試験 (限定)	0.0407	33	(
予備試験	0.0406	28	氏
本試験 (総合)	0.0370	54	さん
本試験 (限定)	0.0340	13	さん
予備試験	0.0228	17	さん
本試験 (総合)	0.0214	29	氏
本試験 (総合)	0.0170	28	】
本試験 (総合)	0.0164	25	被告

表 17 限定課題において CWIfol<sub>w</sub> の値が大きい単語

クラス	CWIfol	CWIfol <sub>w</sub>	単語
固有物名	0.5640	0.5470	違反
時間表現	0.4015	0.3876	ごろ
金額表現	0.4460	0.3750	相当

CWIpre の値ほど高くないことから分かる．日付表現クラスは時間表現クラスとともに現れることも多いが，単独で現れることも多いからである．人名クラスについても同様に，どの課題でも他のクラスより CWIfol の値が大きいことが表 13 から分かる．表 16 に CWIfol<sub>w</sub> の値が大きい単語を示す．どの課題においても敬称や呼称が人名の直後によく現れており，これらの単語は人名を抽出する際に有用であることが分かる．

固有物名，金額表現，時間表現クラスはそれぞれ本試験の限定課題において CWIfol の値が

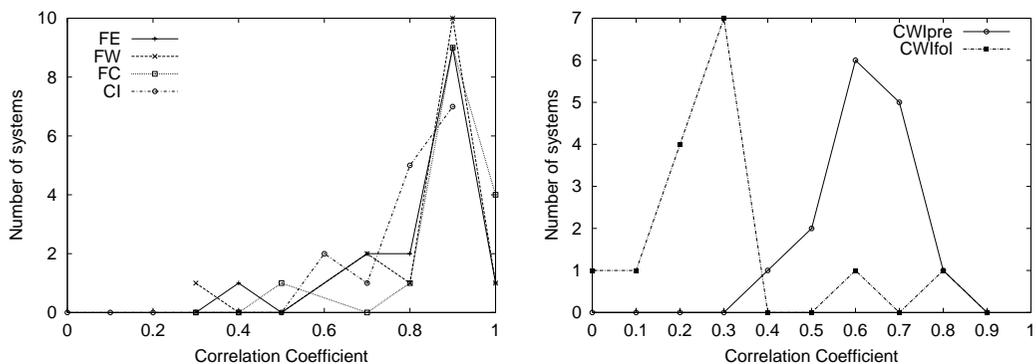


図 5 指標とシステムの評価結果との相関係数 (総合課題)

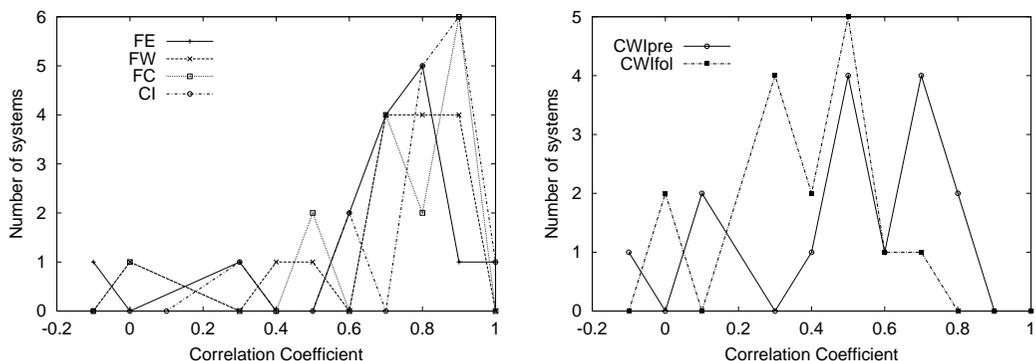


図 6 指標とシステムの評価結果との相関係数 (限定課題)

大きい．表 17に示すように，そのほとんどが特定の一単語がもつ  $CWIfol_w$  の値によるものである．これは，限定課題におけるコーパスが逮捕に関する記事のみから成っており，単語の用いられ方が他の種類の記事に比べてより固定されていることが理由であると考えられる．

## 5 個々のシステムとの相関

IREX ワークショップに参加した全システムの性能について，その平均値との相関を調べることによって指標の結果の有効性を検討してきた．本節では，各システムの性能と指標との相関について示す．

IREX で行なわれた本試験の総合課題・限定課題それぞれについて，定義した指標と参加した各システムの評価結果との相関係数を調べた結果を図 5, 6に示す．指標の種類は FE, FW, FC, CI, CWIpre, CWIfol の全 6 種類である．CI については，図 4 から得られたしきい値を用

表 18 システム E の各クラスごとの評価結果

クラス	全平均	システム E	値の差
組織名	55.2	74.2	(+19.0)
人名	68.8	68.9	(+0.1)
地名	68.1	61.2	(-6.8)
固有物名	57.9	91.7	(+33.8)
日付表現	89.4	91.2	(+1.8)
時間表現	89.8	89.5	(-0.3)
金額表現	91.4	100.0	(+8.6)
全表現	71.7	74.6	(+2.9)

いた結果を示した。固有表現内の文字列を用いた指標 (FE, FW, FC, CI) との相関は、どちらの課題においても、ほとんどのシステムで高い。固有表現の周囲の単語を用いた指標 (CWIpre, CWIfol) との相関はそれに比べると低いことが分かる。特に限定課題では、固有表現の周囲の単語を用いた指標との相関はシステムによって大きくばらつきがある。

表 19 に、個々のシステムについて、その F-measure の値と手法の特徴をまとめた。括弧内に示した値は、各々のシステムの F-measure について、それ以外の全システムの F-measure の平均との相関係数をとったものである。また、表 20, 表 21 に、IREX で行なわれた本試験の総合課題・限定課題それぞれについて、定義した指標と参加した各システムの評価結果との相関係数を調べた結果を示す。これは、図 5, 図 6 を記述するのに用いたものである。固有表現の周囲の単語を用いた指標は、どのシステムにおいても相関係数の値が低く、また、バラツキが大きいので以下の考察ではふれない。

総合課題においては、各システムの評価と全体との相関は、システム O を除いて非常に高い。限定課題においては、システム E, F, O において、全体との相関が低くなっている。固有表現内の文字列を用いた指標 (FE, FW, FC, CI) との相関は、両課題においてほぼシステム全体との相関に類似した結果になっている。システム E, F, O はそれぞれ異なる機械学習手法を用いており、また評価結果も互いに近い値ではないので、手法の特徴が指標との相関に影響しているとはいえないが、システム F と O については、評価時に用いたプログラムやデータに不備があり、本来の性能が発揮されていなかったことがワークショップにて発表された。このことが、F-measure の値や指標の値の双方において、他のシステムとの相関が低い原因になっていると考えられる。システム E については、総合課題においては相関が高いが、限定課題では相関が低くなっている。システムの各クラスごとの評価を見ると、とくに固有物名・組織名での結果が平均に比べて高く、この差が相関が低くなった原因と考えられる (表 18)。システム E は、限定課題用にチューニングは行っていないが、手作業および自動生成によって得られた N グラムパターンを用いており、これらのパターンが、限定課題の固有物名としてよく現れる法律名などに対応していたと考えられる。

表 19 各システムの性能評価・手法の特徴:

システムの評価は F-measure の値．括弧内の数字は，各システムの F-measure と，それ以外の全システムの F-measure の平均との相関係数．

システム	各システムの評価				手法の特徴	
	総合課題		限定課題		パタンの使用	機械学習の手法
A	57.69	(0.956)	54.17	(0.972)	Y	-
B	80.05	(0.989)	78.08	(0.901)	Y	有限状態変換器
C	66.60	(0.969)	59.87	(0.756)	Y	-
D	70.34	(0.973)	80.37	(0.927)	N	決定木
E	66.74	(0.975)	74.56	(0.520)	Y	N グラムパタン
F	72.18	(0.876)	74.90	(0.493)	N	最大エントロピー
G	75.30	(0.967)	77.61	(0.901)	Y	-
H	77.37	(0.990)	85.02	(0.905)	N	最大エントロピー
I	57.63	(0.901)	64.81	(0.908)	Y	-
J	74.82	(0.961)	81.94	(0.820)	Y	-
K	71.96	(0.975)	72.77	(0.923)	Y	決定木
L	60.96	(0.984)	58.46	(0.882)	N	隠れマルコフモデル
M	83.86	(0.892)	87.43	(0.933)	Y	-
N	69.82	(0.932)	70.12	(0.779)	Y	-
O	57.76	(0.424)	55.24	(0.229)	Y	パタン学習と判別分析

表 20 指標とシステムの評価結果との相関係数 (総合課題)

システム	FE	FW	FC	CI	CWlpre	CWlfol
A	-0.927	-0.935	-0.906	0.894	0.570	0.156
B	-0.944	-0.943	-0.984	0.877	0.699	0.223
C	-0.923	-0.931	-0.979	0.806	0.625	0.122
D	-0.870	-0.897	-0.914	0.821	0.572	0.205
E	-0.922	-0.938	-0.942	0.925	0.661	0.270
F	-0.676	-0.704	-0.821	0.629	0.384	0.343
G	-0.836	-0.881	-0.905	0.832	0.645	0.275
H	-0.900	-0.908	-0.967	0.883	0.737	0.344
I	-0.899	-0.854	-0.904	0.770	0.471	0.150
J	-0.832	-0.825	-0.922	0.755	0.504	0.318
K	-0.913	-0.902	-0.920	0.906	0.616	0.316
L	-0.896	-0.920	-0.965	0.865	0.704	0.274
M	-0.733	-0.704	-0.884	0.725	0.630	0.579
N	-0.966	-0.979	-0.942	0.894	0.681	0.038
O	-0.369	-0.342	-0.494	0.556	0.767	0.751

総じて，固有表現内の文字列に基づいた指標と各システムの性能との相関は，ほぼ全システムの平均との相関と同じ傾向を示しているが，固有表現の周囲の単語を用いた指標は改善の必要があるといえる．

表 21 指標とシステムの評価結果との相関係数 (限定課題)

システム	FE	FW	FC	CI	CWlpre	CWlfol
A	-0.753	-0.855	-0.894	0.923	0.726	0.483
B	-0.756	-0.687	-0.684	0.886	0.444	0.547
C	-0.721	-0.884	-0.929	0.787	0.744	0.096
D	-0.771	-0.792	-0.770	0.870	0.646	0.344
E	-0.767	-0.535	-0.451	0.622	0.133	0.058
F	-0.267	-0.355	-0.484	0.582	0.110	0.682
G	-0.729	-0.684	-0.679	0.869	0.482	0.547
H	-0.754	-0.841	-0.906	0.886	0.708	0.345
I	-0.904	-0.852	-0.818	0.926	0.509	0.353
J	-0.587	-0.776	-0.858	0.775	0.802	0.322
K	-0.959	-0.886	-0.886	0.958	0.519	0.383
L	-0.575	-0.791	-0.868	0.838	0.758	0.452
M	-0.672	-0.709	-0.725	0.832	0.704	0.503
N	-0.671	-0.701	-0.646	0.770	0.492	0.315
O	0.135	0.035	-0.023	0.260	-0.085	0.570

## 6 結論

本論文では、固有表現抽出の難易度を示す指標を定義し、IREX ワークショップで行なわれた課題についてそれらの指標を適用し、参加したシステムの評価結果と相関を調べることで、その有効性を検証した。指標を定義するために、固有表現内の文字列、あるいは固有表現周囲の文字列に対して、固有表現クラスごとの頻度・異なり数や、個々の表現のクラス内における頻度とコーパス全体における頻度を用いた。定義された指標のうち、固有表現内の文字列に基いた指標に対しては非常に高い相関が得られた。また、個々の表現に対する指標の値と固有表現抽出における有効性との関係を具体例から考察した。

今後の課題としては、まず固有表現の周囲の表現に基づいた指標を改良して指標としての有効性を高めることが挙げられる。また、固有表現内の文字列に基づいた指標に位置情報を加え、接頭辞や接尾辞などの有効性を測れるようにすることも考えられる。最終的には、指標による分析を通して、与えられた分野の固有表現抽出に有用な情報を自動的に獲得したいと考えている。

## 参考文献

- Bagga, A. and Biremann, A. W. (1997). "Analyzing the Complexity of a Domain With Respect To An Information Extraction Task." In *The Tenth International Conference on Research on Computational Linguistics(ROCLING X)*, pp. 175-184.
- DARPA (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, MD, USA. Morgan Kaufmann.

- DARPA (1998). *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA, USA.
- IREX 実行委員会 (編) (1999). IREX ワークショップ予稿集. IREX 実行委員会.
- 松本裕治，黒橋禎夫，山地治，妙木裕，長尾真 (1997). 日本語形態素解析システム JUMAN (version 3.3). 京都大学工学部, 奈良先端科学技術大学院大学.
- 野畑周 (1999). “決定木を用いた学習に基づく固有表現抽出システム.” IREX ワークショップ予稿集, pp. 201–206.
- 野畑周，関根聡，辻井潤一 (2000). “固有表現抽出技術の難易度に関する分析.” 言語処理学会第6回年次大会併設ワークショップ.
- Nobata, C., Sekine, S. and Tsujii, J. (2000). “Difficulty Indices for the Named Entity task in Japanese.” In *Proceedings of the 38th Annual Meeting of Association for Computational Linguistics (ACL2000)*, pp. 344–351.
- Palmer, D. D. and Day, D. S. (1997). “A Statistical Profile of the Named Entity Task.” In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97)*, pp. 190–193.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc., San Mateo, California.
- Sekine, S., Grishman, R. and Shinnou, H. (1998). “A Decision Tree Method for Finding and Classifying Names in Japanese Texts.” In *Proceedings of the Sixth Workshop on Very Large Corpora*, pp. 171–178 Montreal, Canada.

## 略歴

野畑 周: 1995年東京大学理学部情報科学科卒業。2000年東京大学大学院理学系研究科博士課程修了。博士(理学)。同年通信総合研究所関西先端研究センター知的機能研究室非常勤研究員。2001年より、同けいはんな情報通信融合研究センター自然言語グループ専攻研究員。言語処理学会、情報処理学会、ACL各会員。

関根 聡: Assistant Research Professor, New York University。1987年東京工業大学応用物理学科卒業。同年松下電器東京研究所に入社。1990年～1992年 UMIST 客員研究員。1992年 UMIST 計算言語学科修士。1994年から NYU, Computer Science Department, Assistant Research Scientist。1998年 Ph.D.。同年から現職。自然言語処理の研究に従事。コーパスベース、パーザ、分野依存性、情報抽出、情報検索等に興味を持つ。言語処理学会、人工知能学会、ACL各会員。

辻井 潤一: 京都大学大学院工学博士。1971年京都大学工学部電気工学科卒業、

1973年同大学大学院修士課程修了。同年4月より、同大学電気工学第2教室助手、助教授を経て、1988年から英国UMIST(University of Manchester Institute of Science and Technology)の教授。同大学の計算言語学センター所長などを経て、1995年より東京大学大学院理学系研究科情報科学専攻・教授。組織変更により、現在は、同大学院情報理工学系研究科・コンピュータ科学専攻教授。また、1981年～1982年、フランスCNRS(グルノーブル)の招聘研究員。言語処理学会、情報処理学会、ACL各会員。

(2001年10月25日受付)

(2002年7月24日再受付)

(2002年10月4日採録)