# 人手作成パタンとコーパスにおける統計情報の融合による固有表現抽出システム

宮本昌幸　松尾衛　森辰則
横浜国立大学工学部電子情報工学科, E-mail: {laforge,mamoru,mori}@forest.dnj.ynu.ac.jp

## 1　はじめに

膨大な文書情報が氾濫している昨今，文書から特定の情報を抜き出す情報抽出が注目されている [MUC96, MUC98, IREX99]．文書中の組織名，人名，地名，商品名のような固有名や，時間表現，数値表現，割合表現などを同定する課題である，固有表現抽出タスク (Named Entity タスク, NE タスク) はその基礎技術として重要である．昨年開催された IREX においては，NE タスクに参加したシステムの約半数がパタン駆動型であり，好成績をおさめているものが多い．しかし，パタン駆動型システムの問題点はパタンの大部分が人手で書かれており，作成や調整のための労力が大きいことである．

そこで本稿では，この労力を軽減するために，人手や機械的に作成した各々のパタンに対し，正解タグ付きコーパスから得られた固有表現に関する統計情報を付与することにより，パタンが同定する固有表現の推定を行なう枠組を提案する．さらに，この統計情報を用いることにより，パタンの性能に関する評価やそれにもとづく選別を行なえることを示す．

## 2　パタンと統計量

パタン駆動型 NE システムにおいて，パタンのなす役割とは，形態素解析等により得られた形態素列 (ならびに品詞などの付加情報) において，ある特別な部分列を発見することである．そして同時に，その部分形態素列の全部又は一部が，ある固有表現であるという情報を付加する役目もなす．これは，図 1に示すとおり，該当する各形態素の先頭と末尾に対して，ある固有表現の開始 (例えば組織名の開始 ORG-ST)，中間 (例えば組織名の中間 ORG-CN)，終了 (例えば組織名の終了 ORG-ED) というクラスを付与することである．図 1は三形態素により一つの組織名が構成されている場合に相当する．このクラスを以下では「NE クラス」あるいは単に「クラス」と呼ぶ．

図 1: NE クラスと形態素の関係

図 2に示す人名ならびに組織名を同定するパタンを例にとってみよう．このパタンでは左辺が特定の形態素列を発見するための条件部であり，右辺が照合した形態素の先頭と末尾に付与される NE クラスを表す[1]．

ここで，図 2の右辺に見られるような NE クラス情報付与について注目する．通常，この情報はパタンにより記述された特定の形態素列に対して，人間が割当を行ない，固有表現同定に利用する．これは，人間が NE の正解を与えていることになる．

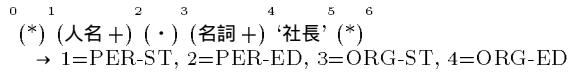


図 2: NE 用パタンの例

しかし，もしも，固有表現に関するタグを付与したコーパスがあるならば，それに対して各パタンを照合させた時の結果を蓄積することにより，NE クラスに関する情報が収集できるはずである．さらに，この方法では，パタンに対して人間が NE クラスの情報を付与した箇所以外 (図 2においては，区切り 0,5,6) についても情報を収集できるので，その箇所にも NE が有意に出現した場合にはそれを自動的に獲得できる可能性がある．

すなわち，パタンの持つ機能のうち「特定の形態素列を発見する機能」のみを用い，そのパタンに現れる全ての部分パタンについてのコーパスにより NE に関する統計量を求めれば，人間が記述したパタンを包摂する情報が獲得されると期待される．

## 3　パタンに付与された統計情報に基づく NE 抽出

前節で述べた考え方に従うと，未知の文書に対する NE 同定は以下のステップからなる．

1. NE タグ付コーパスからの統計量の収集
   (a) NE のためのパタンを用意する．NE のクラスに関する情報はあってもなくてもよい．
   (b) 正解タグ付コーパスよりパタンに照合する形態素に現れる NE クラスの頻度を調べる．
   (c) NE クラスの頻度情報から，そのパタンが適用された時の NE の生起確率を求める．
2. 未知の文書における NE の同定
   (a) 未知の文書に対して，各々のパタンの適用を試みる．パタンが適用された場合，そのパタンが持っている NE クラスの生起確率を対応する形態素に割り当てる．
   (b) 付与された NE クラスの生起確率に基づき，スムージングを行ないつつ，NE クラスのの連接可能性を満足しつつ文全体の生起確率が最も高くなるような NE クラスの選択をし，NE クラスの推定値とする．

[1] このパタンは説明のためのものであり，実際の記述とは異なる．

### 3.1 正解タグ付きコーパスによるパタンへの統計情報の付与

ステップ 2 においては，パタンの各部がどのような確率である特定の NE クラスになるかを推定する．これには，正解タグ付コーパスに対して各パタンの照合を繰り返うことにより，実際に照合した NE クラスの頻度情報を蓄積する．具体的には，各パタン中の各々の部分パタンについて，START,END,CENTER-ST, CENTER-ED という名前の頻度リスト (NE 頻度リスト) を持たせ，どのような NE のクラスが何回現われたかという頻度情報を保存する．NE 頻度リスト START にはその部分パタンの先頭に現れる NE のクラスと頻度を，END には末尾に現れる NE のクラスと頻度を保存する．パタンにおいて繰り返し表現が現れる場合には，その繰り返しの途中の形態素に付与される NE のクラスについても集計する．これには NE 頻度リスト CENTER-ST と CENTER-ED が用いられ，それぞれ，繰り返しの内側に現れる形態素の先頭，末尾に対応する．

例として，図 2のパタンが「PER-ST ‘山田’ PER-CN ‘太郎’ PER-ED ‘・’ ORG-ST ‘日本’ ORG-CN ‘電子’ ORG-CN ‘機器’ ORG-ED‘社長’」という NE 情報付きの文字列に照合した場合を考える[2]． 区切番号 3,4 の間にある繰り返しパタンは 3 形態素 (‘日本’, ‘電子’, ‘機器’) に照合するが，この時にはその繰り返しパタンが以下のように展開されたと考え，文書中の NE のクラスの頻度情報を対応するリストに保存する．

| START CENTER-ED | CENTER-ST CENTER-ED | CENTER-ST END |
|---|---|---|

この 1 回の照合の結果，このパタンに付随する NE 頻度リストはそれぞれ次のようになる．ただし，各要素は「NE クラス * 頻度」の形式である．

START: (ORG-ST*1), END: (ORG-ED*1), CENTER-ST: (ORG-CN*2), CENTER-ED: (ORG-CN*2)

### 3.2 パタンにおける NE クラスの生起確率の推定

あるパタンが文書に照合した時に NE クラスが生起する確率は，各 NE 頻度リストにおける NE クラスの相対頻度で推定される．例えば，ある部分パタンの持つ NE 頻度リストが次のようになっていたとしよう[3].

START: (ORG-ST*1 NON*1), END: (ORG-ED*1 PER-ED*1), CENTER-ST: (ORG-CN*2 PER-ST*1), CENTER-ED: (ORG-CN*2 NON*1)

これより，各部分の生起確率は以下のように推定される．

START: (ORG-ST=0.50 NON=0.50), END: (ORG-ED=0.50 PER-ED=0.50), CENTER-ST: (ORG-CN=0.66 PER-ST=0.33), CENTER-ED: (ORG-CN=0.66 PER-ST=0.33),

しかし，このようなナイーブな方法では統計量をとるためのコーパスの規模によるデータスパースネスの問題が生じる．つまり，正解付コーパスにおいて照応できる事例が少なかったパタンについては，ほとんどの NE クラスで推定確率が 0 になってしまう．ここでは，以下の m 推定 (m-estimate) を用いて，平滑化を行なう [Mit97]．ただし，$n$ はそのパタンが適用された回数，$n(c)$ はその中での NE クラス c の頻度，$p(c)$ は NE クラス $c$ の事前確率 (正解付コーパス全体での出現確率)，$M$ は定数[4]である．これらを各クラスについて求めている．以下では，M=m*n，m=0.2 としている．

$$\frac{n(c) + M * p(c)}{n + M} \tag{1}$$

### 3.3 未知の文書に対する NE クラスの生起確率の推定

正解タグ付コーパスから NE クラスの生起確率に関する情報を付与されたパタンがあれば，未知の文書に対しても各形態素における NE クラスの生起確率を推定可能である．すなわち，文書全体に対して全てのパタンを順次照合可能性を調べる．照合した場合には，対応する形態素に対し，パタンの持つ NE クラスの生起確率情報をそのまま付与する．このとき文書のある部分に複数のパタンが照合することがある．この場合，厳密には条件付確率の重ね合わせを計算すべきであるが，ここでは簡単のために各パタンの NE クラスの推定性能が等しいと仮定し，各パタンのもつ生起確率の平均値を採用している．

### 3.4 NE クラスの生起確率に基づく固有表現の抽出

前記の方法により各形態素に NE クラスの生起確率を求められるが，最も生起確率が高い NE クラスを採用すると，その前後にどのような NE クラスが採用されたかとは無関係にクラスが推定されてしまう．このため，各形態素に付与された生起確率と NE クラスの連接可能性の両者を考慮しつつ，一文全体として NE クラスの生起確率が最大となる候補を選択する．探索方法としてはビタビアルゴリズムを用いた．

## 4 人手で付与した NE クラス情報に基づく NE 抽出との比較

本稿で提案する枠組による NE 抽出手法の基本的な特性を調べるために，人間がパタンに付与した NE クラスによる NE 抽出と比較する実験を行なった．使用するパタンは IREX NE タスクに参加するために作成したものを用いた．従来のパタンに基づく NE 抽出では，これをそのまま用いる．本手法の枠組においては，同じパタンについて，付与されている NE クラスの情報を使わず，正解付コーパスから得た NE クラスの生起確率を用いる．なお，形態素解析器として両システムとも JUMAN3.6 を用いた．正解タグ付コーパスとしては郵政省通信総合研究所により作成，公開された CRL 固有表現データ (毎日新聞 1174 記事) を用いた．また，評価対象としては，IREX 本大会の NE

[2] ここでは，見やすさのため，品詞情報を省略している．

[3] これは，前節の例において注目していた部分パタンが，さらに，「| NON NON | | PER-ST PER-ED |」という形態素列に 1 回照合した状況である．

[4] 等価サンプル数 (equivalent sample size) と呼ばれる．

タスクで使用された文書の自由トピック記事 (72 記事) により行なった．結果を表 1,2に示す．

表 1: 従来のパタン型システムによる抽出結果

| | 再現率 | 適合率 | F 値 |
|---|---|---|---|
| ORGANIZATION | 44.99 | 76.75 | 56.73 |
| PERSON | 49.86 | 56.91 | 53.15 |
| LOCATION | 61.06 | 72.36 | 66.23 |
| ARTIFACT | 4.08 | 15.38 | 6.45 |
| DATE | 88.04 | 83.79 | 85.86 |
| TIME | 96.55 | 76.71 | 85.49 |
| MONEY | 86.67 | 86.67 | 86.67 |
| PERCENT | 80.95 | 100 | 89.47 |
| ALL | 59.34 | 72.19 | 65.14 |

表 2: 統計情報に基づくシステムによる抽出結果

| | 再現率 | 適合率 | F 値 |
|---|---|---|---|
| ORGANIZATION | 41.39 | 74.19 | 53.14 |
| PERSON | 52.68 | 57.72 | 55.08 |
| LOCATION | 65.38 | 71.77 | 68.43 |
| ARTIFACT | 4.08 | 33.33 | 7.27 |
| DATE | 87.68 | 78.57 | 82.88 |
| TIME | 87.93 | 96.23 | 91.89 |
| MONEY | 86.67 | 86.67 | 86.67 |
| PERCENT | 80.95 | 100 | 89.47 |
| ALL | 59.85 | 71.65 | 65.22 |

これらの表に示される通り，パタンのもつ特定の形態素を同定する機能だけに注目し，人間の付与した NE クラス情報を用いずとも，同等以上の精度で抽出できることがわかる．個別の固有表現クラスについて見ると，統計を用いたことにより PERSON，LOCATION，ARTIFACT，TIME の表現について改善が見られた．一方，ORGANIZATION,DATE については精度が落ちている．

## 5 機械的に作成したパタンでの評価

前節ではパタンに対して NE クラスの確率情報を付与することにより，人間が NE クラスを付与した場合と同程度の精度で NE を抽出できることを示した．そのことはまた，正解コーパスから機械的に半網羅的につくりだしたパタンであっても，統計情報を用いることによりある程度の精度で NE を抽出できるものと考えられる．ここでは，人名抽出の場合に注目して，その評価を行なう．

### 5.1 評価に用いる人名抽出パタン

パタンを機械的に作成するにあたって，抽出の戦略として以下のものを採用した．

- 人名を構成し得る形態素として，名詞，片仮名列と‘・’，平仮名列のみを考える．これらの集合を Mid-PER とする．
- 人名に連接し，なおかつ，人名にならない名詞を集める．そのうち，前方に現れる表現の集合を Pre-PER，後方に現れる表現の集合を Post-PER と表す．
- 人名の一部となる形態素を発見する．
- 人名の一部がわかっている時その前 (後) の形態素について，Mid-PER でありなおかつ，Pre-PER (Post-PER) に属していないものであれば，そこも人名と認定する．

なお Pre-PER ならびに Post-PER については，CRL 固有表現データから集めた．

上記戦略に基づき人名抽出パタンを作成する．一つめのパタンを図 3に示す．これは，形態素解析の結果，人名と判定された箇所から人名候補を探索するものである． 次に考えるべきは，人名であることが形態素解

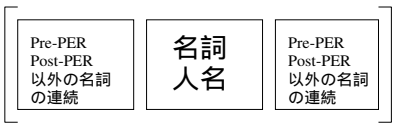


図 3: 人名抽出パタン 1

析の結果からはわからなかった場合に対処するパタンのグループである．これは，図 4に示すように，人名の直後 (直前) に現れ，その直前の表現が人名であると特定できるような表現に基づくパタンである[5]．人名の直後 (直前) に来る表現としては，Post-PER(Pre-PER) の他にも表現がある．ここでは，その多数を占める名詞性接尾辞 (接頭辞相当のもの) を追加して考える．実際のパタンはその表現を埋め込むことにより生成されるので，多数のパタンのグループからなる．また，表層表現をそのまま展開しただけでは正解付コーパスに特化し過ぎているものも数多くみうけられる．そこで，人名の直後の表層表現をその品詞などに置き換えて一般化したパタンもも含めている． これらのパタ

図 4: 人名抽出パタングループ 2

ンを用いて，節 4と同じ評価実験を行なった．結果を表 3,4 に示す．

これらの表をみると，まず，従来のパタン型システムでは，網羅的に書かれたパタンによるため，再現率は高いものの適合率が非常に低くなっていることがわかる．一方，統計情報に基づくシステムでは，パタ

[5]図 4は，人名の直後の表現に基づくパタンのグループである．これと同様に人名の直前の表現に基づくパタンのグループもある．

ンごとに付与されたNEクラスの統計情報が有効に働き，適合率が大幅に向上している．最終的にF値では約10ポイントの改善が見られた．

表3: 従来のパタン型システムによる抽出結果

| 再現率 | 適合率 | F値 |
|---|---|---|
| 71.27 | 45.67 | 55.67 |

表4: 統計情報に基づくシステムによる抽出結果

| 再現率 | 適合率 | F値 |
|---|---|---|
| 61.41 | 69.21 | 65.08 |

## 6 パタンの選別

前節では，統計情報に基づくシステムを用いると，例え，網羅的に展開されたパタンを用いても適合率の低下を招くことなく，十分な精度が保つことができることを示した．しかし，不要なパタンを数多く残しておくと，パタンの照合に非常に時間がかかり有用なシステムとならない．例えば，図3に示すパタンが1つのパタンで記述できるのとは対象的に，図4のパタングループは表層表現の組合せにより非常に多数のパタンのグループとなる．さらに，人名の直後の表層表現をその品詞などに置き換えて一般化したパタンもあるので，これらも含めるとその数は増大するので問題となる．

そこで，パタンに付与されている統計情報を用い，確率的に有用でないパタンを排除することを考える．その尺度としては，エントロピーを用いる．この選別ができれば，パタン数を減少させることが可能であり，統計情報に基づくシステムでは処理時間の短縮が見込まれる．更に，選別されたルールを使えば，パタン型でも十分な精度が得られることが期待される．

### 6.1 パタンのエントロピー

各パタンの各々の部分パタンに対して，正解付コーパスから獲得されたNEクラスの生起確率が付与されているとすれば，頻度リストSTART,END,CENTER-ST,CENTER-ED ごとにエントロピー次式で求めることができる．ただし，$c$ はNEクラスを表す．

$$-\sum_{c} P(c)\log P(c) \qquad (2)$$

さて，1つのパタンは複数の部分パタンを持ち，また，一つの部分パタンは4つの頻度リストを持つためパタン全体のエントロピーはそれらから求める必要がある．ここでは，それらの平均値を採用した．

ここでパタンが持つエントロピーの意味を考える．エントロピーが大きい場合には，NEクラス分布が均一化されているということを表すので，ある形態素列に対してそのパタンが照合したとしても，ある特定のNEクラスを指すことがなく，有用な情報となり得ない．逆に，エントロピーが小さいパタンは，照合した部分について，ある特定のクラスになる可能性が高い．

### 6.2 エントロピーによる選別の効果

エントロピーによるパタン削減の効果を示すために，大量に生成された図4に示される型のパタンについて，エントロピーを計算し，ある閾値以上のパタンを削除する実験を行なった．結果を表5に示す．

表5: エントロピーによる選別後の従来のパタン型システムによる抽出結果

| 閾値 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 |
|---|---|---|---|---|---|
| 再現率 | 64.23 | 64.51 | 67.89 | 67.89 | 67.89 |
| 適合率 | 65.71 | 65.43 | 64.96 | 60.86 | 58.50 |
| F値 | 64.96 | 64.97 | 66.39 | 64.18 | 62.85 |

閾値が0.20のときF値は66.39となり，パタンを削減する前から約11ポイントの改善がみられた．

## 7 まとめ

本稿では人手で作成された固有表現抽出パタンに対して，正解タグ付コーパスから取得された統計情報を付与することにより，確率に基づいてに固有表現を特定する手法を提案し，その有効性を確認した．特に，正解タグ付コーパスから半網羅的に作成した人名抽出パタンにおいて，統計情報を用いた手法で固有表現を特定することにより，パタン単独による抽出と比べて，F値において約10ポイントの改善が見られ，適合率の低下を軽減した．また，パタンに付与された統計情報から求めたエントロピーを用い閾値をもうけて精度の低いパタンを削除することで，パタン型システムにおいても約11ポイントのF値の改善がみられた．

今後の課題としては他の学習型システムとの融合などがある．

## 参考文献


[IREX99] IREX実行委員会 (編). IREXワークショップ予稿集. IREX実行委員会, 1999.

[Mit97] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[MUC96] MUC6, editor. *Proceedings of the sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann Publishers Inc., 1996.

[MUC98] MUC7, editor. *Message Understanding Conference Proceedings*. http://www.muc.saic.com/, 1998.