情 報 学 基 礎 57−14
自 然 言 語 処 理 136−14
（2000. 3. 22）

機械学習と人手作成のパタンを組み合わせた固有表現抽出

松尾 衛　　宮本 昌幸　　森 辰則
横浜国立大学 工学部

テキストから固有表現を抽出する手法は，大別すると局所的なパタンマッチによる抽出手法とコーパスを用いた統計的な抽出手法の２種類が用いられている．本稿ではこれらの２つの手法を組み合わせた固有表現抽出手法についての提案を行なう．我々は，これらの固有表現抽出手法を組み合わせるために，複数の固有表現抽出手法を組み合わせて固有表現抽出を行なう枠組を提案する．この枠組に従い，パタンとコーパスからの統計を用いた抽出手法と機械学習に基づく抽出手法の２つの抽出手法を作成することにより，それぞれを統合して固有表現抽出を行なうシステムを実現した．さらに，このシステムを用いて人名の抽出を行なった結果，抽出精度が改善されることが確認された．

NE Tagging System based on a Combination of Machine Learning and Handmade Patterns

Mamoru Matsuo , Masayuki Miyamoto and **Tatsunori Mori**
Faculty of Engineering,
Yokohama National University
{mamoru,laforge,mori}@forest.dnj.ynu.ac.jp

Techniques of named entities extraction may be classified into pattern driven method and statistical method. This paper describes named entity tagging system which employs a combination of these two methods. To combine these methods, we propose a framework to incorporate multistrategies for named entity spotting method. With this framework, we developed an experimental system, in which a pattern driven system and a system based on machine learning. Our experimental result of person names shows that the combination improves F-measure.

1 はじめに

近年の計算機と計算機ネットワークの発展にともない，現在さまざまな文書が電子的な媒体を介して利用されるようになっている．しかし，そのような文書群は計算機ネットワーク上に大量に散在しているため，すでに利用者が管理できるレベルを越えており，その中から必要な情報を取り出してくることが非常に困難な状況となっている．そのため，膨大な量の文書群の中から利用者に必要な情報を獲得するための技術に対する要求が非常に高まっており，固有表現抽出などの情報抽出に関するさまざまな研究が盛んに行なわれている．

テキストから固有表現を抽出する手法としては，局所的なパタンマッチによる抽出手法とコーパスを用いた統計的な抽出手法の大きく分けて２つの手法が用いられている．局所的なパタンマッチによる手法の利点は処理が軽く高速であり，パタンに照合しさえすれば非常に強力であることが挙げられる．しかし，パタンの作成には多大な労力を要するという欠点を持っている．そのため抽出する情報を変更したりする場合にはパタンの作

成をやり直す必要がある．それに対し，コーパスを用いた統計的な抽出手法ではトレーニング用のコーパスを作成すれば，固有表現抽出ルールを自動的に作成することができる．残念ながらトレーニングコーパスの作成にかかる労力までは軽減できないもののパタン作成の手間については軽減することができる．

我々はこれらの抽出手法について，それぞれのアプローチで研究を行なっており，人手によるパタンを利用した固有表現抽出システムと，統計処理を利用した固有表現抽出システムの両方について独立にシステムを作成し研究を進めてきた．そして，我々はこれらの手法を組み合わせた固有表現抽出システムについて検討を行なっている．特に統計による抽出手法とパタンによる抽出手法を組み合わせるために，我々は複数の固有表現抽出手法を組み合わせて表現の抽出を行なうための枠組について研究を進めている．固有表現を抽出する際に，抽出する手法や着目点が異なれば，当然のように推定結果も異なってくる．ある手法では抽出できない固有表現でも他の手法を用いることにより抽出できる可能性がある．つまり，より多くの固有表現推定手法を用いて固有表現の抽出を行なうことにより，より高い精度で抽出を行なうことが期待できる．

本稿では，統計的な推定手法とパタンによる推定手法の2つの固有表現推定手法を組み合わせるために，複数の固有表現推定結果を組み合わせる枠組を提案している．ここで注目するべきことは，我々の枠組が単にパタン型と統計型の固有表現推定手法を組み合わせるためだけの枠組ではなく，2つ以上の複数の固有表現推定手法による推定結果を組み合わせるための枠組であるということである．つまり，本稿で示す手法以外の固有表現推定手法を，我々の枠組に取り込むことにより，更に固有表現の推定精度を向上させることも期待することができる．本稿で示す各々の固有表現推定手法も統計やパタンを用いた固有表現抽出のための多くの手法のうちの一例にすぎないことに注意されたい．

複数のシステムの推定結果を組み合わせる手法は，現在，人工知能の分野で複数の学習アルゴリズムを利用する手法が多く報告されており (元田，1999)，情報抽出の分野でも Freitag(Freitag, 1998) により，3種類の学習手法により推定された結果を用いて情報抽出を行なう手法が提案されている．しかし，Freitag の手法では抽出する表現のクラスは1種類であるという前提の元に枠組が組み立てられている．例えば，地名を抽出するのであれば，表現を抽出する際に推定するのは地名であるか地名でないかのどちらかである．そのため，複数のクラスを抽出するためにはそれぞれのクラスで独立して抽出する表現であるかないかを推定する必要がある．我々の枠組では，複数のクラスの抽出が一度の抽出で可能である．つまり，組織名，人名，地名等の複数のクラスの表現が推定された結果を組合わせて，固有表現の抽出を行なうことが可能である．

本稿では，複数の固有表現推定結果を組み合わせる手法についての説明を行ない，固有表現の推定手法となる統計を用いた固有表現推定手法と人間が作成した固有表現抽出パタンを用いた推定手法について，それぞれの説明をする．そして，これらの手法を統合した抽出結果についての報告を行なう．

2 抽出する固有表現

本稿で抽出する固有表現は IREX の NE task で抽出対象となった表2に示す8種類の表現である．

ORGANIZATION 組織名，政府組織名

PERSON 人名

LOCATION 地名

ARTIFACT 固有物名

DATE 日付表現

TIME 時間表現

MONEY 金額表現

PERCENT 割合表現

IREX の NE task では以上の固有表現に対し，元の文書に SGML のタグを挿入することにより固有表現の抽出を行なっている．なお，SGML のタグづけを行なう際は重複や入れ子のない形でタグづけを行なう．もし，表現が重なっている場合は原則として長い単位の表現を抽出する．例えば「日本銀行」の場合「日本」を地名として抽出するのではなく「<ORGANIZATION> 日本銀行 </ORGANIZATION>」として日本銀行全体を組織名として抽出する．なお抽出する固有表現の詳細は (IREX 実行委員会，1999) を参照されたい．

3 推定手法の統合

ここで，複数の推定結果を統合する手法についての説明を行なう．

3.1 複数の推定結果の組合せ方法

固有表現の抽出は，固有表現内に使われている表現，固有表現の周りの表現，文脈等さまざまな情報を利用した抽出が行なわれている．これらのような情報を用いて，その表現が組織名なのか人名なのかという判断が下される．しかし，単語ごとの分かち書きがされない日本語の場合，表現がどの種類の固有表現かという判定の他にも，どこからどこまでが固有表現を形成しているのかという判定を下す必要がある．つまり，日本語文書から固有表現の抽出を行なう場合固有表現の範囲を同定するセグメンテーションの問題も同時にとかなければならない．

この問題に対し関根らは，文書を形態素解析しその形態素が含まれる固有表現の種類と，その形態素が固有表現の開始位置か中間か終了位置かをまとめあげたクラスを作成し，学習アルゴリズムを用いることによりその形態素がどのクラスかを推定する手法を用いて固有表現の抽出を行なった(Sekine, Grishman, & Shinnou, 1998)．固有表現の抽出を，注目している形態素が固有表現の始まりであるか，終わりであるか，あるいは途中であるかを判別するタスクとしてとらえたのである．

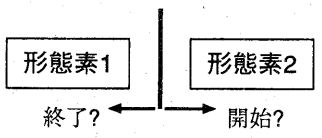

図 1: 固有表現の抽出手法

固有表現を求める際には図1の1形態素目と2形態素目の間が固有名詞の開始か終了か中間かそれとも固有名詞ではないかをということ求めれば良い．つまりこれらが抽出システムにより求めるクラスということになる．また，この縦棒の位置では，1番目の形態素が固有表現の終了位置であるかに注目した場合と，2番目の形態素が固有表現の開始位置であるかに注目した場合の2種類の注目の仕方がある．表現の始まりに注目した場合，求めるのは

- ある固有表現の開始位置である．
- ある固有表現の中間である．
- 固有表現ではない．

ということである．これらを固有表現の際に求めるクラスとして定義する．表現の開始と表現の中間を表すクラスについては，抽出する固有表現それぞれに対応するクラスが用意されている．一方，固有表現の終了に注目した場合では，固有表現の終了位置，中間，固有表現ではない，というクラスを求めることになる．以上をまとめると具体的には求めるクラスは表1のようになる．

表 1: 求めるクラス

開始		終了	
組織名の開始	ORG_ST	組織名の終了	ORG_ED
組織名の中間	ORG_CN	組織名の中間	ORG_CN
人名の開始	PSN_ST	人名の終了	PSN_ED
人名の中間	PSN_CN	人名の中間	PSN_CN
地名の開始	LOC_ST	地名の終了	LOC_ED
地名の中間	LOC_CN	地名の中間	LOC_CN
固有物名の開始	ART_ST	固有物名の終了	ART_ED
固有物名の中間	ART_CN	固有物名の中間	ART_CN
日付の開始	DAT_ST	日付の終了	DAT_ED
日付の中間	DAT_CN	日付の中間	DAT_CN
時間の開始	TIM_ST	時間の終了	TIM_ED
時間の中間	TIM_CN	時間の中間	TIM_CN
金額の開始	MNY_ST	金額の終了	MNY_ED
金額の中間	MNY_CN	金額の中間	MNY_CN
割合の開始	PCT_ST	割合の終了	PCT_ED
割合の中間	PCT_CN	割合の中間	PCT_CN
固有表現ではない	NONE	固有表現ではない	NONE

たとえば，「ここは横浜国立大学です」という文を形態素解析してクラスを付与すると図2のようになる．このようにして，固有表現の抽出をこれらのクラスを求めるタスクと置き換えることになる．

図 2: クラスの適用例

3.2 複数手法の統合

固有表現の開始位置か終了位置か中間かということが，どの程度の精度であるかという確率を求めれば，複数の推定法で求めた結果を確率を用いて組み合わせることが可能である．例えば，図3の位置にいくつかの固有表現推定規則が適用された場合を考える．この場合に，クラスを求める手法として，それぞれのシステムから出力されるクラスを重みづけするモデルが考えられる．これらの規則の中から規則 j の重みを m_j とし，規則 j によるクラス C_i の推定率が $\mathrm{P}_j(C_i)$ とするとク

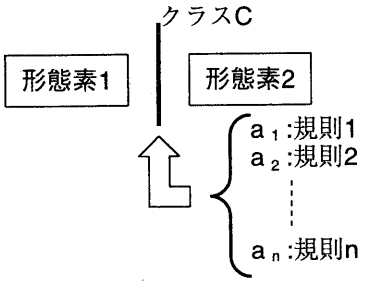

図 3: 固有表現の抽出手法

ラス C_i になる確率 $\mathrm{P}(C_i)$ は

$$\mathrm{P}(C_i) = \sum_{j=1}^{n} m_j \mathrm{P}_j(C_i) \qquad (1)$$

となる．ただし，

$$\sum_{j=1}^{n} m_j = 1;$$

とする．現在は m_j は全て同じ値として扱っている．

3.3 確率データからの固有表現の抽出

上述の手法によりクラスの付与確率を求めることが可能である．しかし，それぞれの付与確率は，その前後にどのようなクラスが採用されたかとは無関係に付与されるため，各々の形態素の切れ目で付与確率が最も高いクラスを採用した場合，クラスのつながりに不整合を生じる場合がある．そのため，付与された確率から固有表現を抽出するための手法が必要となる．付与確率から固有表現を抽出するための手法としては，確率データからモデルを作成し，ビタビアルゴリズムを用いて表現を抽出する手法が広く用いられている．クラスの付与確率を利用し，文頭から文末までを，クラスが連接可能であるものに限って接続した時の一文全体の確率を求めることにより，一文全体でのクラスの付与確率を求めることが可能である．この一文全体の付与確率の内，最大のものを採用することにより，最適であると考えられる固有表現の抽出を行なうことが可能である．

しかし，適用できる事例が少なかった規則については，ほとんどのクラスで付与確率が 0 になってしまう．この場合，例えばORG_ST が 100% の形態素と PSN_ED が 100% の形態素が隣接するということがおこり，連接できるものがなくなってしまう．このように結果が求まらなくなる場合が出現するため，ビタビアルゴリズムを適用する際には，以下の式によりクラスの付与確率の平滑化を行なう．

$$\frac{n_c + mp}{n + m} \qquad (2)$$

ただし，n はその規則が適用されたデータの事例数，n_c はその規則が適用されたトレーニングデータの事例のうち，あるクラス c に属している事例の数，p はクラス c のトレーニングコーパス全体での出現確率，m は重みである．現在 m は経験により n の 15–20% の値に定めている．

4 決定木学習による固有表現抽出システム

ここでは，コーパスから学習アルゴリズムを用いて統計的に固有表現抽出ルールを求める手法について述べる．先に述べたように統計型の利点としては学習アルゴリズムにより抽出ルールの生成を行なうため，抽出ルールの作成の手間を省くことができる点にある．もちろん計算機による学習で求めた固有表現抽出規則はトレーニングコーパスで多く現れた事例を基に導いた規則であるため，手作業で求めた細かな規則や大規模な辞書を用いたシステムほどの精度は期待できないだろう．しかし，1つのシステムで，さまざまな分野の固有表現に対応できる汎用性の高さを考慮すれば機械学習手法を利用して抽出規則を自動的に獲得する手法に対する研究も重要であると考えられる．また，人間により作成されたパタンと組み合わせて利用する場合は，人間が記述しきれなかったパタンを補強するために利用することも可能である．

我々の学習型の固有表現抽出手法は関根 (Sekine et al., 1998) の手法に基づくシステムであり，この手法にいくつかの拡張を行なっている．我々の手法では固有表現の開始位置と終了位置を求めるために，図4に示すように，表現の開始に注目した学習と終了に注目した2つの決定木学習を行なっている． 1つめの拡張は複合名詞からなる固有名詞に対する拡張である．日本語の場合，いくつかの形態素からなる複合名詞が頻繁に登場する．日本語は head final であるため，名詞句の一番最後の形態素が，その複合名詞が表す内容を強く表している．そこで，我々は注目している名詞句の一番最後の形態素の情報を利用することにより抽出精度の改善を試みている．2つめの拡張として，言語の持つ概念に注目した拡張を行なっ

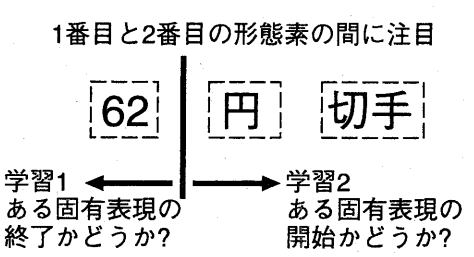

図 4: 注目する形態素の境界

ている．言語の表層表現をそのまま用いて学習を行なった場合，トレーニングコーパスにほとんど現れなかった未知の表現を含んだ文書に対して精度の低い抽出ルールが生成されることが予想される．この問題に対し，我々はその言葉の持つ意味(概念)に注目して，同様の意味を持つ言葉をまとめてからルールを求めることにより，改善を試みている．

4.1 学習型システムの構成

システムはトレーニングコーパスから固有表現抽出ルールを生成する学習システムとそのルールを他の文書に適用して固有表現を抽出する固有表現抽出システムの2つのシステムから構成される．このシステムの概要を図5に示す．

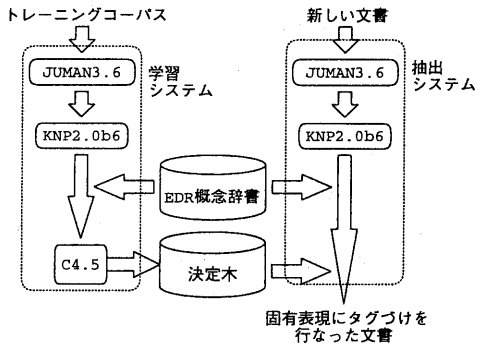

図 5: システム概要

学習アルゴリズムにはC4.5(Quinlan, 1995)を用いており，形態素解析にはJUMAN version 3.6(黒橋 長尾, 1998)を用いている．また，我々の手法ではいくつか文節に関する情報を利用するので，文節認識のためにKNP version 2.0 b6(黒橋, 1998)を用いている．

なお，このシステムの詳細については文献(松尾 森, 1999)を参照されたい．

5 人手による抽出パタンとコーパスデータを用いた固有表現抽出システム

この章では，パタンを用いて固有表現の抽出を行なう手法について述べる．この手法では固有表現自身または固有表現の周りに現れる特徴的な表現に注目し，その表現を手がかりにしたパタンを作成することにより固有表現の抽出を行なう手法である．しかし，パタンを用いた手法の場合，求めているのはその表現が固有表現であるかどうかの真偽値である．そのため，確率の統合により固有表現を推定する我々の枠組には利用できない．そこで，抽出パタンとコーパスを用いて固有表現の開始，終了確率を求める手法について述べることにする．

5.1 パタンと統計量

パタン駆動型固有表現抽出システムにおいて，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6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，中間，終了というクラスを付与することである．図6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．

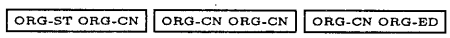

図 6: NE クラスと形態素の関係

図7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]．

```
0  1    2  3    4    5  6
(*) (人名+) (・) (名詞+) '社長' (*)
  → 1=PER-ST, 2=PER-ED, 3=ORG-ST, 4=ORG-ED
```

図 7: NE 用パタンの例

[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．

ここで，図7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．

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7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．

5.2 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．

5.3 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すことにより，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(CN-ST)，CENTER-ED(CN-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7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 CN-ED | CN-ST CN-ED | CN-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)

5.4 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．

[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．

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),

このように，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することが可能となる．具体的には，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，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とにより，固有表現の推定確率を与えている．

6 システムの評価

以上で述べた固有表現抽出システムの抽出結果について述べる．まず，機械学習に基づく抽出システムとパタンに基づく抽出システムのそれぞれのシステムの抽出精度を示し，それらの抽出精度と推定結果を統合したシステムの抽出精度の比較を行なう．ただし，パタン型の抽出システムでは，人名抽出のみで実験を行なっているため，パタン型抽出システムと統合システムについては人名の抽出のみで評価を行なうことにする．

6.1 評価用のデータについて

評価はIREX-NEの本大会で使用された文書に対して行なうことにする．IREXの本大会で用いられた文書は毎日新聞の記事である．この毎日新聞の記事を対象として，トピックを限定した記事(逮捕に関する記事．ARREST)20記事と，自由トピックの記事(GENERAL)72記事から固有表現の抽出が行なわれた．ここでは，紙面の都合により自由トピックの72記事を評価対象として固有表現の抽出を行なうことにする．自由トピック記事内に含まれていた固有表現の数は企業名361，人名338，地名413，固有物名48，日付表現260，時間表現54，金額表現15，割合表現21である．

この試験を行なう際に学習用データとして利用した文書は以下の2つの文書である．

1. CRL固有表現データ
2. IREX-NE本試験逮捕トレーニングデータ

1.のデータはIREXの本大会の前に通信総合研究所により作成，公開されたデータであり，毎日新聞の1174記事を対象にIREXのNE taskの定義にしたがって固有表現の抽出を行なった文書である．また，2.のデータはIREXの本大会のトピック限定課題のトピックが発表された際に同時に公開されたデータで，23記事に対して固有表現の抽出が行なわれているトピック限定トレーニングデータである．

6.2 機械学習に基づくシステムの抽出結果

機械学習に基づくシステムによりIREX NE taskの自由トピック記事から固有表現の抽出を行なった結果を表2に示す．人名の抽出精度に注目すると，再現率70.12%，適合率71.17%となり，F値は70.64となっている．

表2: 機械学習に基づくシステムによる抽出結果

	再現率	適合率	F-measure
組織名	50.69	54.79	52.66
人名	70.12	71.17	70.64
地名	65.86	72.15	68.86
固有物名	18.75	29.03	22.78
日付表現	90.77	85.51	88.06
時間表現	87.04	82.46	84.69
金額表現	93.33	93.33	93.33
割合表現	61.90	92.86	74.28
全体	66.95	70.35	68.61

6.3 パタンに基づくシステムの抽出結果

5章で述べた，パタンとコーパスからの統計量に基づくシステムによりIREX NE taskの自由トピック記事から人名の抽出を行なった結果を表3に示す．

表3: パタンに基づくシステムによる抽出結果

	再現率	適合率	F-measure
人名	61.41	69.21	65.08

6.4 2つの手法を組み合わせたシステムの評価

学習型システムとパタン型システムの出力を組み合わせたシステムにより，IREX NE taskの自由トピック記事から固有表現の抽出を行なった結果を人名について表4に示す．

表 4: 2つの抽出手法を統合したシステムの結果

	再現率	適合率	F-measure
人名	77.81	67.96	72.25

まず，適合率と再現率について考察する．適合率は，機械学習に基づくシステムでは71.17%であり，パタンに基づくシステムでは69.21%である．そのため，それぞれのシステムよりも低い値となっている．これに対し再現率は，機械学習に基づくシステムでは70.12%であり，パタンに基づくシステムでは61.41%である．再現率についてはそれぞれのシステムからの改善が見られている．この結果は統合する手法の一方が局所的なパタンを用いたシステムであるためと考えられる．パタン型の場合，パタンは固有表現(この場合は人名)であると想定される部分に対して，抽出パタンが作成される．これに対し，固有表現が出現しない部分では，適用されるパタンは作成されない．そのため，固有表現ではないという確率が高いパタンは作成されないことになる．つまり，我々のパタンとコーパスからの統計による固有表現の推定手法では，固有表現であるという情報が付与されやすくなることを示しており，固有表現ではないという情報は付与されにくくなる．学習型との統合を行なった場合も，学習型のシステムの推定結果に，固有表現であるという情報をさらに追加することを示している．そのため，抽出の傾向としては再現率重視のシステムとなっている．

次に，F値についての比較を行なうことにする．機械学習に基づくシステムのF値が70.64であり，パタンに基づくシステムのF値が65.08であるため，機械学習型のシステムとパタンに基づくシステムのどちらに対してもF値の向上が見られている．適合率がやや低下したものの，それよりも大幅に再現率が改善したということがいえるだろう．このように，機械学習型とパタン型のそれぞれのシステムの推定率を組み合わせるで，双方のF値以上の結果が得られることが示されている．

複数のシステムから得られる結果を元にすれば，より抽出精度の高い固有表現抽出システムが作成できることが期待できるだろう．ここで，示した統合結果では再現率重視の抽出システムとなったが，統合する固有表現抽出システムの正確や，それぞれのシステムの推定結果を統合する際の手法を検討することにより，より精度の高い固有表現抽出システムを作成することが可能であると考えられる．

7 おわりに

複数の固有表現抽出手法を統合するための枠組を提案し，この枠組に基づいた機械学習に基づく手法ととパタンに基づく手法のそれぞれの固有表現抽出手法について述べた．そして，これらの抽出手法を統合して固有表現抽出を行なった結果，統合前の抽出システムよりも再現率を重視したシステムとなり，F値についても改善されることを示した．

今後の課題としては，それぞれの抽出システムの精度向上や他の抽出手法の統合，また，統合手法のさらなる検討などがあげられる．

参考文献

Freitag, D. (1998). Multistrategy Learning for Information Extraction. In *proceedings of The Fifteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers.

IREX 実行委員会 (編) (1999). IREX ワークショップ予稿集. IREX Committee.

Quinlan, J. R. (1995). AI によるデータ解析. トッパン.

Sekine, S., Grishman, R., & Shinnou, H. (1998). A Decision Tree Method for Finding and Classifying Names in Japanese Texts.. WVCL-98.

元田浩 (1999). "データマイニング - 機械学習と知識獲得 -." 1999 年度 人工知能学会全国大会 (13 回) チュートリアル講演テキスト 第 I トラック: データマイニング –3 つの側面–. 人工知能学会.

黒橋禎夫 (1998). 日本語構文解析システム KNP version 2.0 b6 使用説明書.

黒橋禎夫 長尾真 (1998). 日本語形態素解析システム JUMAN version 3.6 使用説明書.

松尾衛 森辰則 (1998). 教師あり学習と EDR 概念辞書に基づく固有表現抽出システム. 「知識発見のための自然言語処理」シンポジウム, 1999.11 http://www.pluto.ai.kyutech.ac.jp/plt/inui-lab/pub/NLP_Sympo99/