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

2004－NL－159　(2)
2004／1／13

サポートベクトルマシンを用いたプロフィール情報の抽出

吉谷　仁志†　　黄瀬　浩一†　　松本　啓之亮†

固有表現抽出および情報統合の技術を用いて，電子文書から表形式のプロフィール情報を作成する手法を提案する．固有表現抽出で得られる情報は単語が主であり，単独で有用な情報となることがそれほど多くない．また，従来の情報統合では人手によりあらかじめ作成された表を対象としているものが多く，電子文書の持つ膨大な情報を十分に生かしきれているとはいえない．本研究では，以下に示す手順で電子文書から表形式の情報を作成する手法を紹介する．まず，サポートベクトルマシンを用いて固有表現抽出及びプロフィール情報を表形式化する．次に，それらを氏名の一致により統合する．手法の有効性を確認するために新聞記事 1ヶ月分を対象として実験を行った結果，F 値で 0.56 程度の結果が得られた．

Extraction of Profile Information Using Support Vector Machines

Hitoshi Yoshitani,† Koichi Kise† and Keinosuke Matsumoto†

This report presents a method for extracting profile information in tabular formats based on existing technologies called named entity extraction and information integration. Named entity extraction enables us to provide elements of tables for profile information. Information integration allows us to unify tables for making the profile information fruitful, though it requires predetermined initial tables. In this report, we propose a whole system of extracting profile information by bridging the gap between the two technologies. For this purpose we employ a method of grouping named entities for making initial tables. For the extraction and grouping of named entities we utilize support vector machines. Initial tables are then unified using if these are with the same name. From the experimental results on newspaper articles for one month, we obtained the results of $F = 0.56$.

1. はじめに

インターネットの急速な普及に伴い，膨大な数の電子文書が世の中に存在するようになっている．このような背景のもと，電子文書から目的とする情報のみを自動的に取り出すための様々な手法が提案されている．情報抽出 (Information Extraction) とよばれる研究もその一つである．

従来の情報抽出では，「人名」，「組織名」などの与えられた種類に対する固有表現を抜き出す固有表現抽出 (Named Entity Extraction) [3)~6)] と，何らかの形で作成された表の中から同じものに対する情報をみつけてまとめる情報統合 (Information Integration)[7),8)] の 2 つが盛んに研究されている．しかし，これら 2 つの研究の間には溝があり，双方を統括的に扱う手法がまだ確立されているとはいえないのが現状である．

また従来の情報統合では，あらかじめ人手により作られた表を対象としているものが多い．これは，膨大な電子文書の中で表の情報しか利用しないことを意味している．このような観点から，表以外の部分に対しても情報統合のような枠組みを実現することが望まれている．

そこで本研究では，固有表現抽出および情報統合に固有表現抽出からの表生成を加えた統括的な情報抽出手法を提案する．具体的には，表生成の手順を固有表現抽出，局所的統合，大域的統合の 3 つに分け，前 2 つの部分に関してはサポートベクトルマシン，残りの部分に関しては単純な文字列の一致により処理を行う．

このような手法の有効性を確かめる上で，人物に関する情報 (プロフィール情報) を抽出の題材として取り上げる．これは，人物に関する情報に対して高い関心が持たれていること，また抽出すべき固有表現の種類が多彩であることなどから，抽出の対象として適当であると考えられるためである．

以下，**2.** で従来の情報抽出手法について概観し，**3.** で本研究において使用するサポートベクトルマシンについて説明する．**4.** では本研究の提案手法について説

† 大阪府立大学大学院工学研究科情報工学分野
Dept. of Computer and Systems Sciences, Graduate School of Eng., Osaka Prefecture Univ.

明する．また，提案手法に対する実験およびその考察を **5.** で報告し，**6.** でまとめと今後の課題を述べる．

2. 情報抽出

情報抽出は，電子文書から目的とする情報を取り出す処理のことを指す．その中でも，固有表現抽出に関する研究と情報統合に関する研究が盛んに行われている．ここでは，これら 2 つの現状について概観する．

2.1 固有表現抽出

固有表現抽出とは,「人名」,「組織名」など特定の種類に対応する表現の書かれている部分 (固有表現) を抜き出す処理のことである．この分野の研究は古くから行われており，1987 年から開催されている MUC[1] や，99 年に日本で開催された IREX[2] において固有表現抽出のコンテストも行われている．

固有表現抽出の主たる処理は，パターンマッチングとよばれる部分である．パターンマッチングとは，特定の語の存在から固有表現となる部分を決定する処理のことである．この特定の語，すなわちパターンを決める方法としては，辞書などを用いてパターンを人手で与える手法[3] と，機械学習器によりパターンを自動的に学習させる手法[4),5] が提案されている．

人手によりパターンを与える手法は，辞書などのパターンを与えさえすればよいので，実装が容易である．反面，辞書に登録されていない未知の単語などに対応できないことなどから，大量の文書に対して有効になる手法を確立するのが困難である．一方，機械学習による方法は与えられた学習データから一般性の高い抽出規則を学習するため，人手によるものより高い精度が得られることが多い．中には F 値 (再現率と精度の調和平均) が 9 割を超えている事例も報告されている[6]．しかし，機械学習による手法は事前に大量の学習データを用意する必要があり，少量の文書集合に対しては適切な学習が行いにくいという問題点がある．

2.2 情報統合

情報統合とは，何らかの形で与えられた表形式の情報の中から，同じものに対する情報を探し出し 1 つにまとめる処理を指す．この情報統合に関しては佐藤らの研究[7] が代表的なものとしてあげられる．また，この研究に関しては，データベース上の表を統合させることなどと類似性が高いことから，様々な分野からのアプローチが試みられている．基本的な処理手順として，一致する項目の数や種類ごとに重みをつけたり，人手で規則を与えたりする方法がある．最新の研究では，機械学習を取り入れたもの[8] もある．しかし，従来の情報統合は統合すべき表が人手によりあらかじめ作成されていることを前提としているものが多く，そのような形式になっていない情報を扱えない場合が多い．このことは，大量に存在する電子文書の多くを最初から対象として扱わないことになるため，得られる情報もそれだけ少なくなることを示している．このような観点から，一般の文書形式に対しても情報統合の技術を利用できるような手法の提案が望まれている．

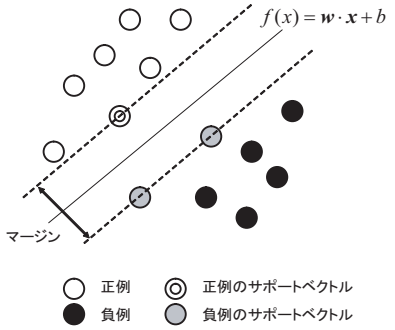

図 1 サポートベクトルマシンによる分離超平面

3. サポートベクトルマシン

2.1 で述べた固有表現抽出において現在最も優秀な結果を示しているのは，サポートベクトルマシンを用いた手法である．そこで本研究でもサポートベクトルマシンを用いて固有表現抽出などの処理を行う．ここでは，その準備段階としてサポートベクトルマシンの概要について説明する．

サポートベクトルマシン (Support Vector Machine ; SVM) は n 次元素性ベクトル　　に対し，正・負のラベル y_t を付与した組 (　,y_t) で表現される l 個の訓練データ ($0 \leq t \leq l$) を用いて，正例と負例を正しく分離する超平面　 ・ $+b = 0$ (　,　 ∈　　) を求める二値線形分類器の一種である．

概要を図 1 に示す．この図において，破線は求める超平面から等距離にある平行な超平面である．この破線間の距離をマージンとよぶ．SVM の学習は，このマージンが最大となる超平面を求めるアルゴリズムである．マージンの最大化は，|| || を最小化することに相当し，これは制約条件 (1) の元で式 (2) を最大化する双対問題と等価であることが知られている．

$$\sum_{i=1}^{l} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \tag{1}$$

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j}^{l} y_i y_j \alpha_i \alpha_j K(\quad , \quad) \tag{2}$$

$$K(\quad , \quad) = \boldsymbol{\Phi}(\quad) \cdot \boldsymbol{\Phi}(\quad) \tag{3}$$

ここで，式 (2) の $K(\quad , \quad)$ を Kernel 関数とよ

ぶ．Kernel 関数は，式 (3) のように 2 つのベクトル　，　∈　　を関数 $\mathbf{\Phi}$ で写像し，それらの内積をとったものである．この Kernel 関数を適切に選択することで，SVM を非線形の分離問題に適用することも可能になる．

自然言語処理の分野では，Kernel 関数として d 次の多項式関数を採用することが多い．d 次の多項式関数を用いると，d 個までの素性の組み合わせを考慮することになる．例として，2 次元の素性ベクトル　$=(a_1, a_2)$，　$=(b_1, b_2)$ に対し 2 次の多項式 Kernel 関数を適用する場合を考える．この場合，2 次の多項式 Kernel は式 (4) で表される．式 (4) を変形していくと，式 (5) に示す$\mathbf{\Phi}(\ )$ で表せることがわかる．この$\mathbf{\Phi}(\ )$ で写像した空間の成分は，元の 2 次元空間における各成分の 2 次までの項で表されている．これは，元のベクトルにおける素性に関して 2 つまでの組み合わせを考慮した空間といえる．

$$
\begin{aligned}
K(\ ,\ ) &= (\ \cdot\ + 1)^2 \qquad (4)\\
&= (a_1 b_1 + a_2 b_2 + 1)^2\\
&= a_1^2 b_1^2 + 2a_1 a_2 b_1 b_2 + 2a_1 b_1\\
&\quad + 2a_2 b_2 + a_2^2 b_2^2 + 1\\
&= \mathbf{\Phi}(\ ) \cdot \mathbf{\Phi}(\ )\\
\mathbf{\Phi}(\ ) &= (x_1^2, \sqrt{2}x_1 x_2, \sqrt{2}x_1, \sqrt{2}x_2, x_2^2, 1) \qquad (5)\\
&= (x_1, x_2)
\end{aligned}
$$

SVM は正例，負例を分類するための二値分類器である．そのため，固有表現抽出などの 3 つ以上のクラスを分類する問題に適用するときは，多値分類器に拡張する必要がある．この拡張方法として pairwise 法がよく用いられる．例えばあるデータを A,B,C のいずれかに分類する場合，pairwise 法では A と B，A と C，B と C の 3 つの二値分類器を用意し，その中で最も多く識別されたクラスを識別結果として用いる．仮に A と B の分類器で B，A と C の分類器で C，B と C の分類器で B という判定が出たならば，識別結果を B とするのが pairwise 法の考え方である．

4. サポートベクトルマシンによる情報抽出

人物に関する情報は多くの人が興味を持つものである．例えば毎日新聞 94 年 1 月分の記事のうち約 7 割は何らかの人物が関与している記事である．このように人物に関する情報，とりわけプロフィール情報に対する要求は非常に高いものであると考えられるので，本研究ではこれを対象とする．

本研究では，電子文書からプロフィール情報を得るための手順を固有表現抽出，局所的統合，大域的統合の 3 つに分ける．固有表現抽出の部分では，表 1 に示すような固有表現を抽出する．局所的統合の部分では，連続するいくつかの固有表現が同じものに対する情報を構成しているという前提の下で，固有表現をグループ化する．大域的統合では，局所的統合で得られた固有表現のグループ同士を比較し，同じものに対する情報があればそれらを 1 つにまとめる．すなわち，本稿で「大域的統合」とよぶ部分が従来の情報統合とよばれている部分と対応する．以下で，各手順の具体的な内容について順に説明する．

表 1　使用する固有表現の種類

固有表現の種類	例
NAME(氏名)	橋田寿賀子
BORN(生年)	１９２５年
FROM(出身)	東京
GRAD(学歴)	大阪府立大学卒
JOB(職歴)	日本フィルコン入社
PRIZE(受賞歴)	芥川賞
AGE(年齢)	５２歳
HEIGHT(身長)	１７２センチ
WEIGHT(体重)	６０キロ
OTHER(その他)	妻孝子

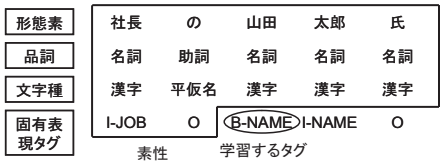

図 2　固有表現抽出の学習に用いる素性

4.1 固有表現抽出

固有表現抽出に関しては，山田ら[5] の手法を適用する．まず，対象文書に対して形態素解析を行う．次に，図 2 に示すように，得られた形態素そのもの，品詞，文字種，固有表現タグを SVM の学習に用いる情報として収集する．ここで，固有表現タグは IOB2 フォーマットに基づくタグである．IOB2 フォーマットでは，固有表現の開始単語には B，それ以外の固有表現である単語には I，固有表現でない単語には O を付与する．これに，表 1 に示す固有表現の種類を組み合わせた B-NAME のような表記を固有表現タグとする．

SVM の学習の際には図 2 の太線で囲まれた部分，すなわち $i-2$ から $i+2$ 番目までの形態素そのもの，品詞，文字種と $i-2$ 番目および $i-1$ 番目の固有表現タグを素性として用いる．実際にタグを推定する場合は，学習時と同様の素性を用いる．このようにして，対象文書から固有表現を抽出する．

4.2 局所的統合

局所的統合では，連続するいくつかの固有表現が同

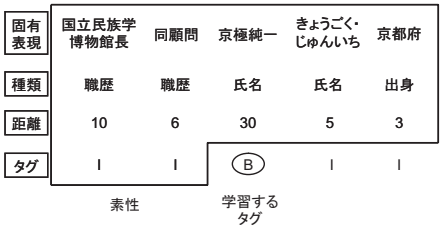

図 3 局所的統合の学習に用いる素性

じものに対する情報を構成しているという前提の下でプロフィール情報の境界を推定する．境界の推定には，先ほどの固有表現抽出と同様の手法を適用する．

まず SVM で学習を行うにあたって，図 3 に示すように，固有表現そのもの，固有表現の種類，固有表現間の距離，学習用のタグの情報を用いる．ここで，学習用のタグは，IOB2 フォーマットに基づくタグである．ただし，ここでの B は連続する固有表現グループにおいて最初に位置する固有表現を表し，I は固有表現グループ中の固有表現，O はそれ以外の固有表現を表す．例えばプロフィール情報において抽出誤りを起こし，プロフィール情報とは全く関係のない部分を固有表現として抽出してしまった場合などは O のタグを付与する．また固有表現間の距離は，直前の固有表現との間に含まれる形態素の数で定義する．

学習の方法としては，図 3 の太線部分，すなわち $i-2$ から $i+2$ 番目までの固有表現そのもの，属性，固有表現間の距離と $i-2$ 番目および $i-1$ 番目の学習用タグを素性として SVM を学習させる．タグの推定時にも学習時と同様の素性を用いる．このようにして得られた固有表現のグループを，以下では固有表現組とよぶことにする．

4.3 大域的統合

大域的統合では，得られた固有表現組どうしを互いに比較し，同じものに対する情報があれば統合する．ここで重要となるのは，どのようにして同じものに対する情報であると判断するかである．本研究では，最も単純な方法である氏名の一致によって統合を行う．

はじめに，各固有表現組の氏名部分を比較し，文字列が完全に一致したものどうしを同じ人物に対する情報と判断する．統合すると判断された固有表現組同士は，互いの固有表現の和集合をとることで統合を行う．これに，統合すると判断されなかったプロフィール情報を合わせて最終結果とする．

5. 実　　験

提案手法の有効性を検討するために，新聞記事を対象としたプロフィール情報の抽出実験をした．まず，固有表現抽出に関する実験を行い，その結果を用いて統合の実験を行った．

表 2 固有表現抽出の結果

テストデータ	R_t	P_t	F_t
1	0.747	0.918	0.824
2	0.684	0.898	0.777
3	0.757	0.902	0.823
4	0.814	0.961	0.881
5	0.703	0.903	0.790
平均	0.741	0.916	0.819

5.1 固有表現抽出に関する実験

固有表現抽出に関する実験では，奈良先端大の TinySVM および yamcha[9] を使用し，表 1 に示す種類の固有表現を抽出した．SVM の仕様としては，山田らの手法で最も結果の良かった 2 次の多項式 Kernel を用いた．対象記事集合を 5 つに分割し，そのうちの 1 つをテスト，残りの 4 つを学習に用いる 5 分割交差検定 (5-fold cross validation) を行い，その平均を最終結果とした．

実験の評価は，正解の固有表現タグと出力された固有表現タグを比較する方法をとった．評価尺度としては式 (6) に示す F 値を用いた．

$$F_t = \frac{2R_tP_t}{R_t + P_t} \tag{6}$$

ここで，$R_t = |C_{tag}|/|A_{tag}|$ は再現率，$P_t = |C_{tag}|/|B_{tag}|$ は精度であり，$|A_{tag}|$ は正解の固有表現タグ数，$|B_{tag}|$ は結果として得られた固有表現タグ数，$|C_{tag}|$ は $|B_{tag}|$ 内の正解数である．

実験結果を表 2 に示す．山田らの手法は固有表現の種類が変わっても，8 割程度の F 値が得られることがわかった．

本来固有表現であるものを正しく抽出できなかったものとしては,「１９５８年入社」などのように職歴に関する部分が 46%と最も多く，次いで「放送文化賞」などの受賞歴に関する部分が 22%,「芸術院会員」などの肩書に関する部分が 11%を占めた．これは，職業や賞の名前，肩書には様々な形態が存在するためそれらを網羅的に抽出できるような規則を学習することが困難であったためと考えられる．

固有表現でないものを誤って抽出したものとしては，数字や「・」,「、」などの記号を含む部分が全体の 47%を占め，次いで「年」や「会」などの形態素を含む部分を抽出したものが 23%を占めた．これは，氏名として「えんどう・しゅうさく」などのようなひらがな表現を学習させたことや,「１９５８年，大阪府立大学卒」などのように，学歴や職歴などの多くに「、」や

表 3 局所的統合の結果 (タグ単位の評価)

テストデータ	R_t	P_t	F_t
1	0.946	0.924	0.935
2	0.906	0.862	0.883
3	0.936	0.950	0.943
4	0.974	0.957	0.965
5	0.908	0.902	0.905
平均	0.934	0.919	0.926

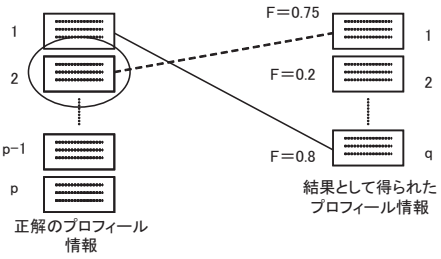

図 4 情報統合の評価方法

「年」が含まれていたことによると考えられる.

5.2 局所的統合に関する実験

5.1 の結果を用いて，局所的統合を行った．局所的統合においても 5 分割交差検定を行い，学習およびテストをした．SVM では同様に 2 次の多項式 Kernel を使用した.

まず，局所的統合のタグがどの程度正しく判定されているかを確かめるため，タグ単位での評価をした．評価尺度としては，先ほどの固有表現抽出で用いた R_t, P_t, F_t の $|A_{tag}|, |B_{tag}|, |C_{tag}|$ をそれぞれ局所的統合タグの数で置き換えたものを用いた．その結果を表 3 に示す．平均の F 値が 0.926 と高く，ほぼ正しい精度で切れ目を判断することができていた.

誤っていたものの 69%は，固有表現の抽出誤り部分 (本来なら O になる部分) をプロフィール情報の一部と認識したものであった．また，これに伴ってその直後の部分を誤ってしまったものが 17%を占め，これが誤りの主原因であることがわかった.

次に，局所的統合だけでどの程度プロフィール情報を正しく得られているのかを調べた．評価基準としては図 4 に示すような方法を用いた．まず，正解データと結果をプロフィール情報単位で比較し，式 (7) で示す F 値を求める.

$$F_p = \frac{2}{\frac{1}{R_p} + \frac{1}{P_p}} = \frac{2R_pP_p}{R_p + P_p} \tag{7}$$

ここで，$R_p = |C_{nee}|/|A_{nee}|$ は再現率，$P_p = |C_{nee}|/|B_{nee}|$ は精度であり，$|A_{nee}|$ は正解の固有表現数，$|B_{nee}|$ は結果として得られた固有表現数，$|C_{nee}|$ は $|B_{nee}|$ 内の正解数である．今，F_p が最大となる組合せが正解データの i 番目と結果の j 番目のプロフィール情報の組であるとすると，この組を対応付けとして採用する．次に，この i と j を含まない組合せの中で F_p が最大となる組合せを対応付けに加える．以下同様の手順で対応する組合せを greedy 法で定める．図 4 の例の場合，正解データの 2 番目に対応する結果のプロフィール情報のうち，q 番目が最も F 値が高い対応である．しかし，これは既に正解データの 1 番目と対応付けられているので，その次に F 値の高い結果データの 1 番目を対応付けとする．この上で，正解数を数え F_p を求めた.

表 4 局所的統合の結果 (プロフィール情報単位での評価)

テストデータ	R_p	P_p	F_p
1	0.608	0.418	0.496
2	0.461	0.288	0.354
3	0.642	0.468	0.541
4	0.671	0.559	0.610
5	0.521	0.315	0.393
平均	0.581	0.410	0.479

表 5 大域的統合の結果

テストデータ	R_p	P_p	F_p
1	0.615	0.651	0.632
2	0.473	0.412	0.440
3	0.650	0.580	0.613
4	0.671	0.608	0.638
5	0.524	0.439	0.478
平均	0.587	0.538	0.560

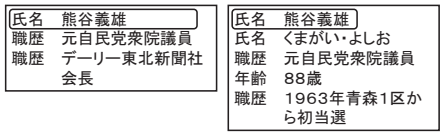

図 5 大域的統合の成功例

実験結果を表 4 に示す．この結果より，局所的統合のみの場合でもある程度の再現率が得られる半面，重複しているものがまとまっていないため精度が低くなることがわかった.

5.3 大域的統合に関する実験

5.2 の結果に対し，大域的統合を試みた．その結果を表 5 に示す．局所的統合のみの状態と比べ，再現率を維持したまま精度を大幅に上昇させることができた．このことより，提案手法による統合の有効性が示せたと考えられる.

大域的統合によって，図 5 のようにいくつかの氏名が一致したことにより同じものに対する情報と認定され，統合が行われているものが多く見受けられた.

氏名　毎日新聞社特別顧問4 氏名　有馬朗人 職歴　京極純一氏 職歴　丸谷才一氏	氏名　有馬朗人 氏名　ありま･あきと 出身　大阪市 職歴　1953年東大理学部卒 職歴　75年同教授 職歴　93年から理化学研究所理事長 学歴　日本学術会議会員 年齢　63歳

図 6　統合の失敗例

氏名　根岸重治 職歴　判事	氏名　ねぎし･しげはる 学歴　1953年 職歴　官房長 職歴　東京 職歴　長 職歴　91年 職歴　退官 年齢　65歳

図 7　文字種の違いにより統合されなかった例

氏名　中野一氏 職歴　前群馬テレビ社長	氏名　中野一 氏名　なかの･はじめ 職歴　前群馬テレビ社長 年齢　79

図 8　固有表現の抽出誤りにより統合されなかった例

しかしその弊害として，図 6 のように局所的統合の誤りによって発生した，本来プロフィール情報ではないものと統合されるという例がいくつか見られた．このような誤りは全体の 57%にのぼり，誤りの主原因となっていることがわかった．また，次に多かったケースは固有表現の抽出ミスにより偶然氏名が一致してしまったものであり，これが全体の 29%を占めた．今回のような統合手法は，前の処理の結果に誤りが多いほど誤った統合を行う可能性が高くなることを示している．また，単なる氏名の一致では同姓同名の別人の区別ができないため，この部分に関しては更なる検討が必要である．

また，図 7 のように，片方の氏名が漢字，もう片方の氏名がひらがなによって書かれていたため，文字列的に一致せず統合されなかったものがあった．これは本来統合すべきであるが統合できなかった誤りの 50%を占めた．残りのうち 47%は図 8 のように，固有表現の抽出誤りによって文字列的に氏名が一致せず，同じ人物であるにもかかわらず統合されなかったものであった．このような問題点を解決するためには，表記ゆれに対処するか，あるいは他の固有表現の一致なども考慮する必要があると考えられる．

6. おわりに

本稿では，固有表現抽出と情報統合の手法を組み合わせて電子文書から表形式のプロフィール情報を得る手法を提案した．本手法の特徴は，固有表現抽出と情報統合 (大域的統合) をつなぐ処理として，局所的統合を導入している点にある．新聞記事を対象とした実験の結果，$F_p = 0.56$ を得た．

従来の情報統合技術には，単純な文字列の一致だけでなく様々な要素を考慮して精度の高い統合を行っている手法がある．そのため，これらの手法を適用すれば更なる精度の向上が見込めると思われる．また，現在の電子文書の主流は Web ページであり，これらからいかに的確に目的とする情報を収集できるかに大きな関心が寄せられている．このような観点から，今後様々な情報統合手法を適用し精度の向上を図ることと，Web に対してこの手法を適用し，どれほどの成果が得られるかを検証することを今後の課題としたい．

参 考 文 献

1) *Seventh Message Understanding Conference* (MUC-7), DARPA, 1998
2) 関根　聡，伊佐原　均：“IREX:情報検索，情報抽出コンテスト”, 情処研報, NL-127-15, pp. 109-116, 1998
3) 竹元義美，福島俊一，山田洋志：“辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出”, 情報処理学会論文誌, Vol.42, No.6, pp.1580-1591, 2001
4) 内元清貴，馬青，村田真樹，小作浩美，内山将夫，井佐原均:“最大エントロピーモデルと書き換え規則に基づく固有表現抽出”, 自然言語処理, Vol.7, No.2, pp.63-90, 2000
5) 山田寛康，工藤　拓，松本裕治：“Support Vector Machine を用いた日本語固有表現抽出” , 情報処理学会論文誌, Vol.43, No.1, pp.44-53, 2002
6) 磯崎秀樹，賀沢秀人：“SVM に基づく固有表現抽出の高速化”, 情処研報, NL-149-1, pp.1-7, 2002
7) 佐藤理史：“情報の自動編集と WIT プロジェクト”, 電子図書館–デジタル情報の流通と図書館の未来, 日本図書館情報学会研究委員会編, pp.131-149, 勉誠出版, 2001
8) AnHai Doan, Ying Lu, Yoonkyong Lee and Jiawei Han：“Profile-Based Object Matching for Information Integration”, *IEEE Intelligent systems*, September/October, pp.54-59, 2003
9) ⟨URL : http://cl.aist-nara.ac.jp/~taku-ku/software/⟩