特集　自然言語による情報アクセス技術

1 情報抽出

ー情報を整理して提示するー

関根 聡
ニューヨーク大学
sekine@cs.nyu.edu

情報抽出の技術の解説を行い，いくつかの応用システムを紹介する．情報抽出とは特定のテーマの情報を非構造的な文書の中から抽出し，構造化されたかたちで提示する技術である．具体的には，図-1にあるように新聞記事の中から，「人事異動」に関する情報を，「人名，会社名，異動前役職名」などをコラムとした表形式に整理したかたちで提示する．この技術の起源は1980年代に米国のMUC（Message Understanding Conference）で始まった情報抽出プロジェクトにある．MUCでは新聞記事などの文章からテロ活動，製品開発，人事異動，企業合併，ロケットの発射情報などの情報を整理して抽出するというタスクが行われた．このMUCではシステムを作成する参加団体がスコアを競うという形式で行われたが，1990年代半ばに終了した．その後，米国では情報抽出のサブタスクがACE（Automatic Content Extraction）などのプロジェクトで続けられている．また，情報抽出のまとまったかたちでのプロジェクトはTIDES（Translingual Information Detection Extraction and Summarization）の一部にあるが，評価型のプロジェクトではない．それ以外での現在の情報抽出の研究は，MUCで行われてきたようなあらかじめ決められた特定の情報だけを新聞記事のような決まった情報源から抽出するというものから，対象をWeb文書などに広げ需要の高い特定の情報について大量に抽出する，または，さまざまな情報について自由に抽出を行う，というような方向に広がっている．特に前者の技術はテキストマイニングを源流としたWebラッパーと呼ばれる技術である．本稿では，まずMUC型の情報抽出を紹介し，現在の情報抽出で用いられている代表的な要素技術，そして，新しい情報抽出の展開を具体例を挙げながら紹介する．

これまでの情報抽出技術

MUCを源流とする情報抽出技術については，筆者が本学会の会誌40巻4号（1999年4月）の「特集：フィールドを広げる自然言語処理」の「4．テキストからの情報抽出」で紹介させていただいた．本稿はこの続編という性格も持つが，この章ではまず，これまでの情報抽出技術というかたちでMUCを源流とした情報抽出技術について簡単にまとめてみる．

MUCにおける情報抽出とは，新聞記事のようなテキストからあらかじめ指定されたイベントや事柄に関する情報を抽出し，その情報を表形式のデータベースに自動的に入力するという技術である．図-1に人事異動に関する新聞記事を基にした情報抽出結果を載せる．ここで抽出したい情報は，企業の重役の異動（昇進，降格，退任など）に関する情報であり，抽出したい情報の内容と

＜新聞記事＞
　ABC 株式会社は十二日，臨時取締役会義で田中一郎社長が代表権のある会長に就任し，山田次郎副社長が社長に昇格する人事を内定したと発表した．鈴木三郎会長は代表取締役にとどまる．三月二十五日に開く株主総会後の取締役会で正式決定する．田中社長は五期十年社長を務め，年齢も七十一歳と高齢になったため，若返りを図る．…

＜異動イベントデータベース＞

人名	田中一郎	山田次郎	鈴木三郎
会社名	ABC 株式会社	ABC 株式会社	ABC 株式会社
異動前役職名	社長	副社長	会長
異動後役職名	会長	社長	代表取締役
異動理由	昇格	昇格	降格
異動発生日	3 月 25 日	3 月 25 日	3 月 25 日

図 -1　情報抽出例

パターン：（日時），（企業）は，（人名）（役職名）の（役職名）への昇格を発表した．

入力分：　2 月 7 日，XYZ 株式会社は，高橋四郎副社長の取締役社長への昇格を発表した．

抽出情報：日時＝ 2 月 7 日
企業名＝ XYZ 株式会社
人名＝高橋四郎
異動前役職名＝副社長
異動後役職名＝取締役社長

図 -2　パターンマッチングによる情報抽出

しては，該当者の人名，会社名，異動前役職名，異動後役職名，異動理由，異動発生日というように与えられている．ご覧のように，面倒な文章で書かれた人事異動の情報が整理され，人事異動に関する重要な情報がすっきりとした表形式で提示されている．

　現在の情報抽出の応用との比較として重要なのは，この時点の情報抽出では，抽出する対象がきちんとした自然言語で書かれた新聞記事であり，その中には抽出しようと思っている情報が比較的高い確率で，綺麗なかたちで含まれていることと，抽出したい情報の種類や詳細がタスクというかたちで規定されているという点がある．したがって，実際にシステムを作成する際には，そのタスクに限定された言語解析のための知識を作るということが主なるテーマであった．つまり，汎用な構文解析や意味解析などを行う必要はなく，求めたい情報を抽出するに十分な解析知識のみが重要であった．実際には重要な情報が表現される部分を局所的なパターン（言語的言い回し）として用意し，それを用いて対象文章から情報を抽出するという手法が主流であった．パターンとはたとえば**図 -2** に示すようなものであり，そこにある入力文から，パターンマッチングによって，日時，企業，人名，役職が抽出される．

　このようなパターンマッチングは，基礎となるいくつかの技術によって支えられている．次の章で大きな 3 つの技術について紹介する．これらの基礎技術は MUC において認識され研究が開始されて，現在の情報抽出の新しい展開に沿って，発展してきている．

情報抽出のための要素技術

　情報抽出システムを実現するためには，いくつかの要素技術が必要である．もちろん，日本語では一般的な自然言語解析技術である形態素解析なども必要であるが，ここでは特に情報抽出に重要なものとして，固有表現抽出，照応解析，パターンの自動作成の 3 つについて解説する．固有表現抽出は文中にある名前，数値表現，重要な表現を特定する技術である．照応解析は，名前などの表現が繰り返し示される場合に別の表記で表されていたり，代名詞が使われていたりしても，その関係を特定する技術である．また，前述したパターンマッチングのためのパターンをなるべく自動的に作成するという研究も盛んに行われている．

名前や重要な表現を特定する固有表現抽出

　MUC において情報抽出のサブタスクとして，情報抽出の対象になりやすい人名，組織名，地名や時間表現，金額，割合表現を文中で特定するというタスクが設定され，これらの表現を「固有表現（Named entity）」，固有表現を文中で特定する技術を「固有表現抽出」と呼んでいる．このタスクは最近広く研究されている質問応答においても重要な要素技術となっており，情報抽出のためだけではなく，幅広い自然言語処理の基礎技術として認識されるようになってきている（質問応答については本特集の解説記事を参照のこと）．MUC で定義された 7 種類の固有表現は，当時の MUC における情報抽出タスクである企業活動の情報抽出を意識したものであり，それだけでは不十分であることが知られている．その後，日本で行われた IREX（Information Retrieval and Extraction Exercise）では 7 種類の固有表現に固有物名が追加され 8 種類となり，米国の ACE でも，施設名と GPE（Geographical and Political Entity: 政府組織を持つ地名）という固有表現が追加された．しかしながら，質問応答や幅広い情報抽出の応用を考えた際にはそれでも不十分であり，もっと広い固有表現の定義というものが提

案されてきている．また，生物学情報抽出などの分野ではたんぱく質やDNAの名前などの新しいタイプの固有表現も必要となってきており，これまでの固有表現といったものだけで問題が解決するものではない．2004年の言語処理学会併設ワークショップの「固有表現と専門用語」ではその辺りの問題意識が1つの課題となり，筆者の論文[2]では固有表現の研究の歴史と現在抱える課題や解決の方向性を含めた広範なレビューが述べられている．固有表現を特定する技術は固有表現が誕生して以来いろいろな方法が試されてきたが，7～8種類の固有表現を特定する固有表現抽出システムは，教師付き学習で実現する方法が主流であり，いろいろな言語で高い精度を得ている．教師付き学習というのは，ある程度の規模のテキストに正解の固有表現をタグ付けしておき，それを機械が学習するという方法である．学習の手法としては，決定木，決定リスト，HMM（隠れマルコフモデル），ME（最大エントロピー法），SVM（サポートベクターマシン），CRFs（コンディショナルランダムフィールド）などが試され，さながら機械学習の実験場といった様相であった．しかしながら，固有表現の種類が拡張され，たとえば筆者の提案する200種類もの固有表現になると教師付き学習をそのまま実現できるか疑問である．この問題を解決する可能性や関連技術として以下の方法が考えられており，これは今後の課題となっている．（詳しくは文献2））

- 人手による辞書，ルールの作成
- 部分的教師付き学習（Weakly Supervised Learning）
- 能動学習（Active Learning）
- 言語的手がかりを使った教師なし学習
- 統計的用語抽出（専門用語）
- 語義分類・クラスタリング
- 未知語処理

照応関係を特定する

ある情報抽出の対象となっているイベントが2つ以上の文で表現され，2文目以降では最初の文にあった名前などが，代名詞，省略形，ゼロ代名詞などで表されていることがある．たとえば，1文目に「日本銀行」とあったものが，その後では「同行」「日銀」と書かれたり，または「25日，政府に対して金融政策の説明を行った．」というように，「日本銀行」のことを言っているにもかかわらず，完全に省略されたりするような場合（これがゼロ代名詞と呼ばれている）がある．情報抽出をする場合には，このように代名詞，省略形などで表された表現でも，それが本来何を指しているのかを特定できなければ，適切に情報を抽出できない場合がある．この問題を解決するのが照応解析のシステムである．これまでの，日本語の照応解析では，日本語では頻繁に起こるゼロ代名詞の認定とその先行詞の同定の研究が主であったが，情報抽出の場合には，名詞句照応，特に固有表現の照応関係の同定も重要になってくる．この分野では米国のACEが評価型のプロジェクトを行っており，現在は英語だけではなく，中国語，アラビア語でも同様のタスクが設定されている．照応解析を学習するための教科書としては，たとえば文献4）がある．また，関連する技術として時間表現の正規化という課題もある．これは，同じ日でも文章中では「2004年4月2日」「2日」「昨日」「1カ月前の今日」といったさまざまな表現がされているものを同一のものであることを認識するタスクである．

パターンを自動的に学習する

これまでの情報抽出技術の紹介で，パターンマッチングが重要な技術であることを紹介した．その際に用いられるパターンをどのように作成するか，という点が，現在，大きな研究テーマになっている．MUCの時代にはパターンは主に人間が例を見ながら作成していた．しかし，それでは効率が悪い．そこで半自動または完全自動でパターンを作成しようという方向の研究が盛んである．基本的なアイディアは，情報抽出のトピックに関連するドキュメントを集めると，その中にはそのトピックで重要と思われる言い回しが他のドキュメントに比較してより多く現れるという仮説に基づいている．これは情報検索でキーワードの重要性を評価するTF/IDFという指標の考え方と同じである．パターンをどのような形式で表現するかという問題があるが，これについても，「動詞と主格」や「ofでつながれる名詞句説」などよく現れる形式を固定しておく方法や，「動詞と主格，目的格」に限る，係り受け関係に限る，任意の部分構文木構造とするなどさまざまな方法が提案されている[3]．このような方法によって情報抽出のパターンはある程度効率的に作成できることが分かってきている．

特定の情報について大量に抽出する

新しい情報抽出の応用の1つである，需要の高い特定の情報の抽出について解説する．需要の高い特定の情報というのは，その情報が大量にあることにより価値がある情報となり，多くの人が使いたがるような応用アプリケーションになり得るようなものをいう．具体的には，1）Webに公開されているショッピング情報や製品販売の情報，2）人事募集の情報，3）論文の題，著者名などの情報に関する応用について紹介する．製品情報は，ショッピングサイトを始めいろいろなサイトに載って

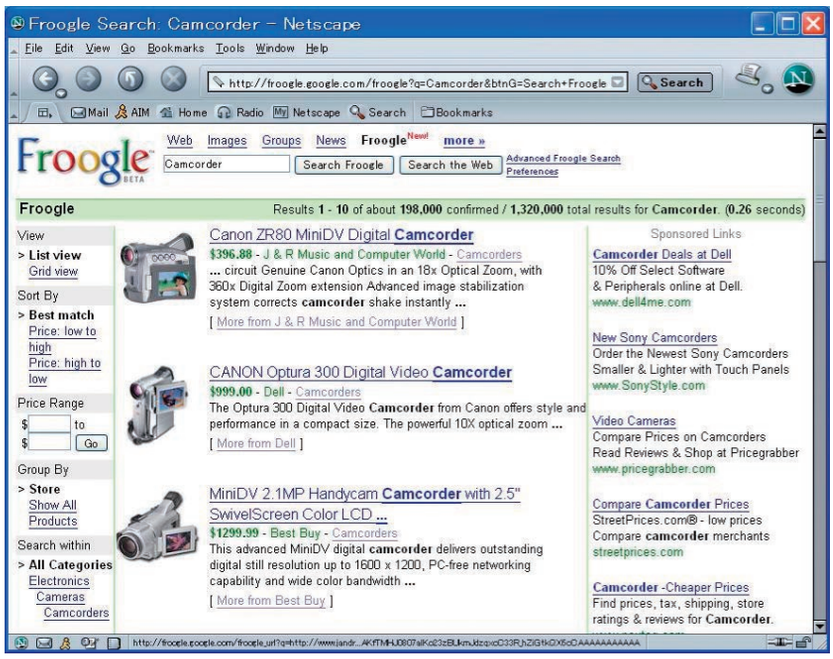

図 -3　自動的に集められたショッピング情報のサイト（http://froogle.google.com/）

いる．それらの製品について，製品の名前，値段，カテゴリー，写真などの情報が何万と集められれば，それだけで重宝されるショッピングサイトが形成できる．人事募集についても同様で，Webページで企業などが公開している人事情報を集めれば，ハローワークにも匹敵するような人事募集のポータルサイトが形成できる．また，論文の情報も大量に集めれば，検索の容易さも手伝って，図書館などで検索するよりもずっと便利な論文データベースが構築できる．このような技術は自然言語処理とは別に，テキストマイニングの研究分野からも形成されており，特にWebデータを対象にしたものはWebラッパーと呼ばれる技術として形成されてきている．Webラッパーについての技術的なサーベイはたとえば文献5）に詳しい．すでに 米国では実際にこの技術を使って自動的に情報を収集したポータルサイトが存在し，大きな注目を集めている．以下に実例を紹介する．

製品情報を抽出する

製品情報を収集しポータルサイトとして公開されている有名なサイトの1つに，googleがやっているfroogleがある．ここでは，googleによって集められた大量のページの中から製品の販売に関するページを見つけ出し，そこから製品の名前，値段，写真，概要などの情報を自動的に収集する．そして，集められた情報を適切な形式で表示することによって，製品に関する巨大なポータルサイトができ上がることになる．ただし，froogleのhelpにもあるように，このサイトにある情報は自動的に集められたものだけではなく，販売小売店などが特定の形式でfroogleに商品の情報を送っている場合もある（これはdata feedによって集めたもの，と呼ばれている）．自動的に集められたものと，直接集められた情報の比率がどのくらいであるかは明確にされていない．しかしながら，たとえば「Camcorder」で検索すると約130万検索されたもののうち，確認されたものが約20万あるとされている（2004年4月2日現在：図-3参照）．これらすべてが直接集められた情報ではなく，自動的に収集された情報が大部分を占めるであろうことは容易に想像される．ショッピングサイトをこのように自動収集したもので作られたポータルサイトはfroogleだけではなく，いくつかのショッピングサイトでも開発されているようである（実際，筆者の所にもコンサルタントの依頼がきている）．googleで使われている技術の詳細は不明であるが，基本的にはこれまでに紹介したようなパターンマッチングによる情報抽出が使われていると考えられる．ただし，パターンの作成は人手で行うことは不可能であり，htmlの特性や製品紹介サイトの特異性を活用，駆使したものであると考えられる．また，ブートストラピングを使用することによって効率化を図るという方法は，現実的に使用されているかどうかは不明であるが，研究的に面白い課題であると考えられる．

リクルート情報を抽出する

製品情報と同様にリクルート情報を広範なWebページから抽出するという応用が行われている．代表的なものはCMUの教授らによるWizBangというベンチャー企業が作ったシステムであり，現在はFlipDogという名前のページに受け継がれている．図-4にあるように2004年4月2日現在，全世界から約20万のリクルート情報を収集しており，地域，カテゴリー，職種などの情報が抽出される．職を探しているユーザは分類された地域やカテゴリーを指定し，自分の希望する職を検索することができる．自動的にリクルート情報を抽出する技術の詳細は明らかにされていないが，図-5の左側にあるようなリクルート情報のあるWebページを検索し，そこから右に示してある"Last Update"，"Location"，"Category"，"Function"などの情報を自動的に抽出している．基本的には，製品情報と同様にリクルート情報の書かれ方のヒューリスティックを利用したパターンマッチングによる情報抽出の技術が試用されていると考えられる．

論文情報を抽出する

英文の論文を検索している人にはご存知の方が多いと思うが，NECの米国プリンストンにある研究所が開発したCiteseerというシステムがある（図-6）．これは，Webに公開されている論文を一堂に集め，それらに対してキーワード検索，引用のリンクをたどった検索，内容の似た論文の検索ができるシステムである．このシステムの実現には，Webにある論文のファイルから，その論文のタイトル，著者，アブストラクトの情報や，引用論文のタイトル，著者，雑誌名，ページ番号，論文中で引用している場所の同定などを行う必要がある．また，抽出した引用論文の同一性や関連論文の検索など情報検索や他の技術も駆使されたシステムになっている．論文の中からのタイトルなどの同定や引用論文の解析には，情報抽出の技術が使われている．文献1）によると，使われている技術は基本的に人手で作成したパターンのようなもので，それを利用して，それぞれの情報を同定しているとのことである．この論文情報の抽出は，前に紹介した自然言語文やWebページからの情報抽出に比べれば，もともとの情報がある程度構造化されたフォーマットで書かれているために，比較的容易であることが予想されるが，それでも，引用論文の書き方は人によってバリエーションがあり，正解率は80%程度にとどまっているようである．しかし，この程度の正解率でも十分に有用な応用であるということ，無料で公開したことなどにより，このシステムがこれだけポピュラーになったものである．

図-4　自動的に集められたリクルート情報のサイト（http://www.FlipDog.com/）

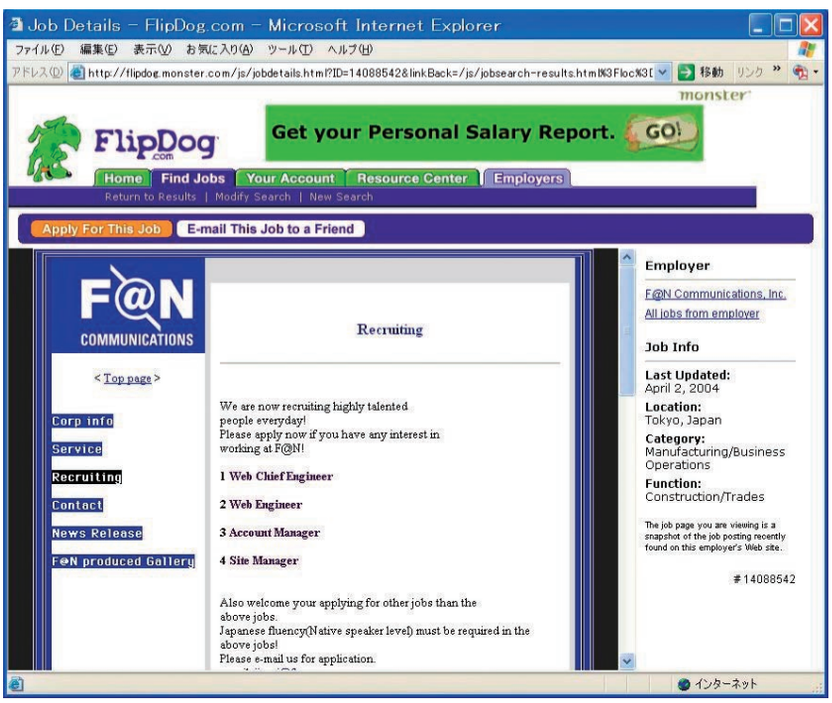

図-5　リクルート情報の自動抽出

幅広い情報を抽出する

ここまでは，製品情報，リクルート情報，論文情報のような特定の情報に限って大規模に情報を集めることによって有用なアプリケーションを形成してきた例を見てきた．ここでは逆に，幅広い情報を抽出するという方向で筆者らが取り組んでいる「オンデマンド情報抽出（On-demand Information Extraction）＝ODIE」の研究を紹介する．幅広い情報抽出は，ユーザがその場その場の個人的な興味やビジネスのニーズから，あるテーマについての情報をまとめてみたいというときに有効なアプリケーションである．たとえば，あるユーザが，ふと「家で作れるフランス料理」について興味を持ったときに，システムにそのように入力するとWebなどの大量なテキストから関連情報を抽出し，そこにある重要な要素，たとえば「料理名，材料，調理時間」などが整

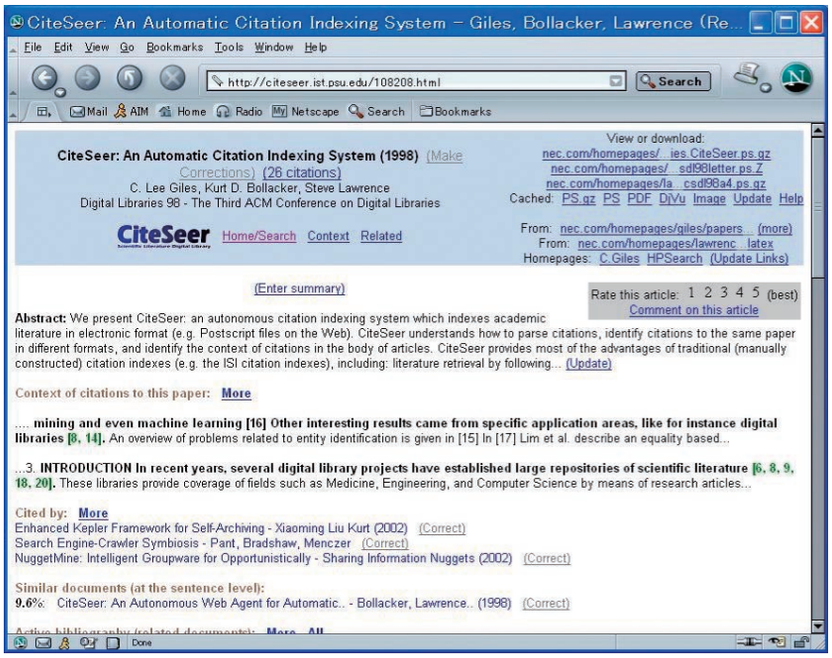

図 -6　自動的に集められた論文情報のサイト（http://citeseer.ist.psu.edu/）

理されたかたちで提示されるというアプリケーションイメージである．本稿の最初の方で紹介したように，特にMUCを中心としたこれまでの情報抽出では，抽出したい情報を規定するタスク（たとえば，「人事情報」や「企業の合併情報」）をシステムの開発前に知らされ，それからそのタスクに関するパターンや辞書などの知識を開発し，システムを構築していた．しかし，各タスクに対して1カ月といった開発期間が必要であり，利便性にかけていた．この1カ月という期間をなくそうというのがODIEの発想である．具体的には4つのコンポーネントから構成される．そのうちの3つの技術は本稿の「要素技術」で紹介した，固有表現抽出，照応解析，パターンの自動学習である．固有表現では，幅広いタスクに対応できるように数種類の固有表現ではなく，200種類にも渡る固有表現を設定した．もちろん，これでもあらゆるタスクに対応できるわけではないが，新聞記事を調査し設定したこの200種類の固有表現は多くのタスクの役に立つものと考えられる．また，照応解析はこれまでのゼロ照応だけではなく，固有表現を中心とした名詞句照応をも含め，情報抽出のためのものとしている．パターンの自動学習は筆者らの論文[3]にもあるように，ある程度の精度で学習できることが確認されており，情報抽出のためのパターンは自動的に抽出できるようになってきている．あと1つ必要な技術は，以下のものである．現在，パターンを自動的に学習し，文章からどの情報がそのタスクにとって重要な情報であるかが分かるようになったが，その意味付けはできない．言い換えると，いろいろなパターンによって抽出した情報がばらばらであり，たとえば「企業合併における合併前の企業名」と「企業合併における合併後の企業名」の区別がつけられなかったり，違ったパターンで抽出した情報の役割の同一性が認定できなかったりするという問題がある．この問題に対しては，パターンの意味的な類似性，つまり，パターン同士の言い換え（パラフレーズ）を同定することによって解決できると考えている．言い換えについては自然言語処理の分野で最近注目されている技術分野であり，言語処理学会第7回年次大会併設ワークショップや2003年のACLのワークショップでも取り上げられ研究が盛んになっているところである．これらの4つの技術を組み合わせ，幅広い情報を抽出するオンデマンド情報抽出の技術の方も実現に向けて歩みを進めているところである．

まとめ

情報抽出の技術と具体的な応用例をみてきた．1980年代に始まった情報抽出の研究も，MUCでやられたようなかたちの「前もって指定した特定の情報に関する抽出」から，製品情報，リクルート情報，論文情報といった人々が興味ある特定の情報について大量に抽出するという方向で応用が盛んになっていたり，幅広い情報をオンデマンドに抽出するという研究方向が模索されるなど，新しい方向へ発展してきている．そのための要素技術として，固有表現抽出，照応解析，パターン抽出，言い換えの発見などが必要であるが，それらは他の自然言語処理でも有用な技術であり盛んに研究が行われている．比較的アプリケーションに近い場所に位置する情報抽出技術の今後が楽しみである．

参考文献

1）Giles, C. L., Bollacker, K. D. and Lawrence, S.: CiteSeer: An Automatic Citation Indexing System, Digital Libraries 98 - Third ACM Conference on Digital Libraries, I, ACM Press, New York, pp.89-98 (1998).
2）関根　聡：固有表現から専門用語，言語処理学会第10回年次大会併設ワークショップ「固有表現と専門用語」発表論文集，pp1-4.
3）Sudo, K., Sekine, S. and Grishman, R.: An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition, Annual Meeting of Association of Computational Linguistics（ACL 03）2003; Sapporo Japan.
4）Mitkov, R. : Anaphora Resolution, Longman（2002）.
5）山田泰寛，池田大輔，坂本比呂志，有村博紀：WWWからの情報抽出？ウェブラッパーの自動構築？，特集「WWW上の情報の知的アクセスのためのテキスト処理」，奥村（編），人工知能学会誌，Vol.19, No.3 (May 2004).

（平成16年4月16日受付）