4 テキストからの情報抽出
―文書から特定の情報を抜き出す―

関根　聡
ニューヨーク大学コンピュータサイエンス学科

情報抽出とは何か

最近，テキストからの情報抽出という技術が注目されている．本稿ではその中心的な存在である米国の Message Understanding Conference (MUC)[1] における情報抽出を中心に解説する．その他の広義の情報抽出の解説には文献2）などを参照のこと．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，その情報をデータベースに入力する，という技術である．まずは具体例を見てみよう．

<新聞記事>
ABC株式会社は十三日，臨時取締役会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．鈴木二郎会長は代表取締役にとどまる．二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，若返りをはかる．…

<異動イベント情報>
人名：　田中一郎
会社名：　ABC株式会社
異動前役職名：　社長
異動後役職名：　会長
異動理由：　昇格
異動発生日：　3月25日

人名：　山田次郎
会社名：　ABC株式会社
異動前役職名：　副社長
異動後役職名：　社長
異動理由：　昇格
異動発生日：　3月25日

人名：　鈴木二郎
会社名：　ABC株式会社
異動前役職名：　会長
異動後役職名：　代表取締役
異動理由：　降格
異動発生日：　3月25日

図-1　情報抽出の例

図-1に人事異動に関する新聞記事を元にした情報抽出結果を載せる．ここでは，抽出したい情報は，企業の重役の異動（昇進，降格，退任等）に関する情報であり，抽出したい情報の内容としては，該当者の人名，会社，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，面倒な文章で描かれた人事異動の情報がすっきりとしたデータベース形式で抽出されている．

このような技術は，特定の情報を簡単に調べたいときなどに非常に役に立つ．たとえば，過去10年の新聞記事から，企業の重役の異動に関する情報を得たいという場面を想定してみよう．現在よくある情報検索の技術を利用すると，適当な検索式を作成し，異動に関連した記事を引っぱり出すことはできる．しかし，多くの読者が簡単に想像できるように，その結果は膨大であり，実際に必要な情報は1つ1つの記事を読んでみなければ得ることができない．また，自動要約という技術を利用して，その手間を減らすことは可能であるが，それでも文章を読まなければいけないことには変わりない．また，今の技術では，特定の視点に注目した要約はまだ開発途上であると言わざるを得ない．これに対し情報抽出の技術を利用した場合には，図-2のような表形式で，複数の異動の情報を視覚的に一度に見ることができる．表形式になっているので，特に注目したい欄を対象にフィルタリングすることなども可能である．つまり，情報検索結果の記事を1記事あたり数分かけて読みながら調べるというのに比べて，非常に高速に必要な情報を得ることができる．

ここでは一例として，重役の人事異動を取り上げたが，対象となる情報は，もちろん，それだけに限らない．一般的に前もって，「抽出したい情報の

型」が決められるものなら特にこだわらない．

たとえば，合弁事業の情報，新製品の情報，新事業の情報など産業界に役に立ちそうな内容や，研究者向けには科学技術論文における技術内容の情報抽出や，医療カルテ，ゲノムといった特定分野のテキストにおける注目される情報（たとえばタンパク質の役割の情報抽出），またアイドル歌手の活動情報抽出，スポーツなどの特定イベントの情報など個人的な利用も含め広く考えることができる．

この技術のポイントは，一般的にテキストを理解するという技術に比べて，あらかじめ抽出したい情報の型が与えられているということにより，比較的に実現容易だという点にある．すでに，いくつかのデモシステムは作成されており，実際に商用に使われ始めているシステムもある．

利用されている技術

前章で，この技術がテキストを理解するよりも容易であると述べたが，実際にどのような点で容易なのか，また，逆にシステムを開発したり使用する際に制約が生じていないかというようなことについて述べる．

パターンマッチングによる情報抽出

一般にテキスト解析というと，自然言語処理における構文解析や意味解析を行うと考えられがちであるが，これらの技術はまだ困難な問題が残り，なんら制約のない環境で広く利用可能であるとは言い難い．それに対して，情報抽出の実用が広く可能になっているのは，これらの難しい技術を使用して深い理解を試みることなく情報抽出を実現できる技術が生まれたという点に集約できる．それがパターンマッチングによる情報抽出という技術である．

ちなみに，初期のMUCでは構文解析などの技術を用いる方式が主流であったが，パターンマッチングの方式が性能的に優れていたため淘汰されてしまった．現在は一部で統計的構文解析等を利用した新しい試みはあるものの，広くは利用されていない．

パターンマッチングは，その情報抽出の対象に関係する文や文の一部にマッチするパターンを用意しておいて，それを決まった順に適応し，決定的に情報をつかんでいくという技術である．図-3に例を載せる．

たとえば，図-3の最初のパターンを言葉に直すと，「固有名詞かカタカナ列かアルファベット列の後に，株式会社という文字列があった場合にはそれらをまとめて企業と判断する．」ということである．@で始まる要素は，なんらかのカテゴリを示す．ここでは，プロセスや定義の詳細には立ち入らないが，図-4では，入力文が次第にパターンマッチングにより@の付いたカテゴリに変換され，最終的に昇進イベントを得るという様子が描かれている．まず，「XYZ株式会社」が最初のパターンにより「@企業」に変換され，その他，日時，人名，役職等が随時変換されていく．そして，最後のパターンを利用して，昇進イベントを得るという流れになる．最後の昇進イベントを得る際には，それぞれの項の内容を記憶し，それぞれの役割の抽出も同時に行う．これによって図-1に示したような情報が抽出されることになる．

ここでは，説明のために簡略化してパターンマッチングの流れを示したが，実際にはこれだけでは十分ではない．人名等の抽出はなかなかパターンのみでは困難であり，辞書的な情報との組合せが必要になってくる．また，パターンは注意深く設計しないと，意図しない文字列とマッチングしてしまうことも考えられる．そして，正確に情報を把握するためにはパターンマッチングの順序も重要である．また，現実的にはこのような1文に対

人名	会社名	異動	異動理由	発生日
○○○	●●●	社長→会長	会長専任	1999年1月17日
△△△△	(株)▲▲▲	専務→社長	昇進	1999年1月20日
××××	※※※※(株)	社長→取締役	引責	1999年2月22日
◇◇◇◇	◆◆◆◆◆	会長→なし	引退	1999年2月30日
田中一郎	ABC株式会社	社長→会長	昇進	1999年3月25日
山田次郎	ABC株式会社	副社長→社長	昇進	1999年3月25日
鈴木三郎	ABC株式会社	会長→取締役	降格	1999年3月25日

図-2 情報抽出結果の例

[固有名詞 | カタカナ列 | アルファベット列] "株式会社" →@企業 数字 "月" 数字 "日" →@日時
["社長" | "副社長" | "会長" | ...] →@役職 @日時 ， "@企業名 "の" @人名 @役職 "は" @役職 "に" "昇進した." →@昇進イベント

図-3 パターンの例

入力文：2月7日，XYZ株式会社の高橋四郎副社長は社長に昇進した．
↓
2月7日，(@企業) の高橋四郎副社長は社長に昇進した．
↓
(@日時)，(@企業) の (@人名) (@役職) は (@役職) に昇進した．
↓
(@昇進イベント)

図-4 パターンマッチングの例

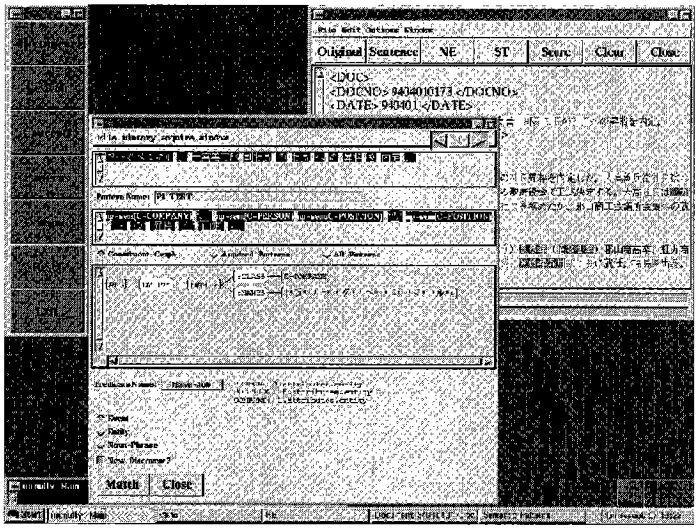

図-5　パターン作成ツールの例

するパターンマッチングだけですべての情報を得られるとは限らない．1つのイベントが複数の文にまたがって書かれていることもある．このような問題を解決するためには照応解析の技術が必要になってくる．これはたとえば，図-1の例で，2つ目の文にある鈴木三郎がABC株式会社の会長であるということを理解するためには，その前の文の情報が必要であるといったことである．このような照応を厳密に行うには，難しい問題があるが，ある程度までの照応解析は，シンプルなルールで解決することができる．

情報抽出は，これまで述べたパターンマッチングの技術を利用して，実用化に向かって大きく成長してきた．しかし，このパターンマッチングの枠組みから，システムの開発や利用においてある種の制約を生じさせてしまっている．次に，その制約を技術的な課題として紹介し，その解決に向けての取り組みを紹介する．

技術的な課題

情報抽出においてパターンマッチング技術を利用したことによって生じた最大の問題は，パターンを情報抽出の課題ごとに作成しなければいけないという点である．つまり，人事異動のパターンは人事異動にしか使用できず，新たに企業合併に関する情報抽出を行いたいとしたときにはそのためのパターンを新たに作らなければいけないということである．課題によって必要なパターンの数は異なるが，ある程度複雑な課題の場合には，少なくとも500から1000近いパターンが必要になってくる．これらのパターンを課題が与えられるごとに，1から手作業で作成していくのでは，そのコストが膨大になり，あまり実用的であるとは言い難い．システムがある特定の情報抽出の要望だけを処理するのならばいいが，望むべくは幅広い課題に対処できるようにしたい．この問題に対しては大きく3つの方向で研究が進められている．

1つは，課題の種類に依存しないパターンと，課題に依存したパターンを切り分け，前者をライブラリとして用意しておくという方法である．たとえば，図-3の例では最初の3つのパターンは重役の昇進という課題以外でも十分に有効であり，課題に依存しないパターンとして利用できる．また，課題に依存するパターンも課題の種類によっては流用できるものもあると思われる．そのような分野依存性を的確に判断しパターン作成の労力を少しでも減らそうというのが1つ目の方針である．

2つ目は，パターンを作成するためのツールを用意し，言語処理や計算機システムなどに詳しくない人でも簡単にパターンを作成できるような環境を用意するという方法である．たとえば，ニューヨーク大学では例文を元に，簡単にパターンを作成するツールを開発している．パターンの要素の一般化や構文的なバリエーションの自動的な作成などをサポートし短時間で特定の課題に対するパターンを作成できるようになっている．図-5にニューヨーク大学のシステムの動作画面の例を載せる．後ろの画面にある記事の例文に対して，手前の画面でパターンを作成する．簡単なマウス操作でこの例文に対するパターンを生成できるようになっている．

最後には，パターンを自動的に作成するという方法である．基本的には大量の文章を元に，動詞の使われ方や固有名詞の出現などの情報からパターンを自動的に作成するという取り組みが行われている．この分野はまだ研究が開始されたという段階であり，今後が楽しみな分野である．

要素技術の分割

最初に紹介したMUCでは，システムのモジュラリティやポータビリティを向上させることを目的に，情報抽出における要素技術を分割し，それぞれの技術をそれのみで評価しようという試みが行われた[3]．この分割は，単に情報抽出の開発に役に立っているだけではなく，それぞれの要素技術の問題点を整理したり，また単独で応用システムに組み込まれたりするのに役に立っている．

まずは，図-3の例からも明らかであるが，人名，組織名，時間などの表現の抽出は1つの要素技術として分割できる．この課題は，テキストの中に

ある固有表現を抽出し，種類を認定することが課題とされており，すでに情報検索の分野で応用されている．一例としては，地名とも人名ともとれる表現（たとえば，成田やワシントン）を地名として検索したいか，人名として検索したいかをあらかじめ指定しておくことによって，検索ミスを減らそうといった応用である．

また，組織名や人名には課題によらない決まった型があると考えられる．たとえば，組織名なら，所在地，社長名，事業分野などがある．そのような情報の型（テンプレート）を分野によらず決めておいて，その情報を抽出することを目的とした課題もある．この技術はテンプレート抽出技術と呼ばれている．

最後に，パターンマッチングのところで触れたが，照応解析というのは情報抽出によって重要な技術である．これも独立した課題として扱うことができるし，この技術は他の自然言語処理応用システムに利用することができる．

これらの要素技術をまとめて模式的に描くと図-6のようになる．

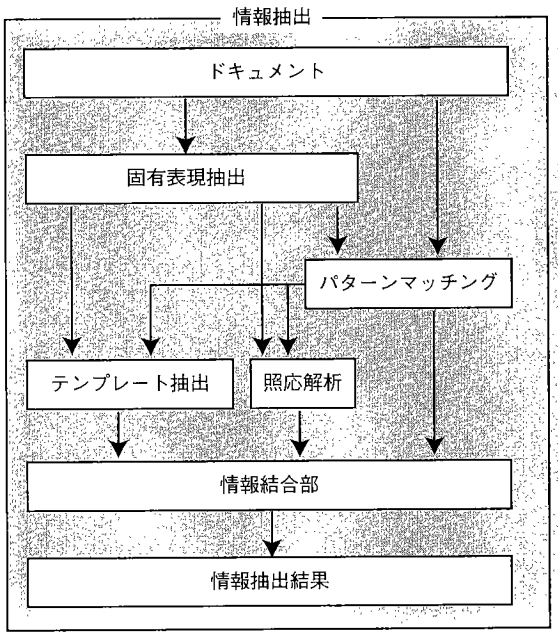

図-6　情報抽出の要素技術

応用分野

情報抽出は技術的には成熟期に手が届いているという状態であり，米国はもとより，日本でも製品化やデモシステムの実例がある．ここでは，それらを大雑把にまとめ，応用分野として紹介したい．

高度情報検索としての応用

本稿の最初にも述べたように，現在の情報検索が持つ問題点を解決していこうという目的が，今一番ホットな対象である．インターネットを始めとする電子上のテキストが膨大になるにつれ，非常に有望な応用課題となってきている．新聞記事や特定の情報リソースからの情報抽出はもとより，電子化図書館での応用なども考えられる．

データベースの（半）自動構築

世界のデータベースコンテンツのシェアは米国が断突であり，日本でもその作成が急務であると伝えられている．このようなデータベースの作成は非常に手間のかかる仕事であるが，それを推進するツールとしての利用価値が考えられる．

情報抽出のツール

米国でMUCが盛んであった背景には，米国の軍や情報中央局で，実際に情報抽出の作業を人手で行っているということが挙げられる．著者の聞くところによると，実際に100人単位の人間が，世界の治安，政治，麻薬等に関する情報を，世界中の新聞や通信社から集め整理し，政策立案に役立てているということである．このような作業を自動化することがMUCのスポンサーである米軍や情報中央局の意図であり，実際に補助ツールとして使用されているそうである．

日本の場合には，先日の北朝鮮のミサイル疑惑の際に，情報収集の重要性が話題になったが，企業戦略における情報収集といった応用も考えられる．

研究グループ

米国のMUCを継承し，日本でも情報抽出の技術を確立し，その向上を目的としたグループが形成されている[4),5)]．興味ある読者はこれらのホームページを見てみることをお薦めする．

参考文献
1) MUC Homepage: http://www.muc.saic.com/
2) Pazienza, M. T. (ed.): Information Extraction, Springer-Verlag, Lecture Notes in Artificial Intelligence, Rome (1997).
3) Grishman, R. and Sundheim, B.: Message Understanding Conference-6: A Brief History, The 16th International Conference on Computational Linguistics (COLING-96), pp.466-471 (1996).
4) IREX Homepage: http://cs.nyu.edu/cs/projects/proteus/irex/
5) NACSIS Workshop Homepage: http://www.rd.nacsis.ac.jp/~ntcadm/

（平成11年3月3日受付）