テキストからの知識抽出の基盤となる日本語基本解析技術

インターネット上に氾濫しているテキストから有用な情報・知識を取り出すためには，まずテキストを解析しなければなりません．本稿では，テキスト基本解析技術のうち，日本語を対象とした形態素解析，固有表現抽出，係り受け解析を中心に，その仕組みを紹介します．

今村 賢治（いまむら けんじ） /齋藤 邦子（さいとう くにこ）
浅野 久子（あさの ひさこ）
NTTサイバースペース研究所

3つのテキスト基本解析技術

インターネット上で流通している情報には，画像や音声など，さまざまな種類がありますが，中心となるのは，HTMLやワープロ文書に代表されるテキストデータです．これらテキストデータに書かれている情報を有効活用するには，まず，そのテキストに何が書かれているか，解析する必要があります．本稿で紹介する技術は，インターネット上の文書だけでなく，一般的なテキストデータを解析し，そこに書かれている情報を有効活用するための基本技術です．

各技術の流れを**図1**に示します．

■形態素解析

テキストはコンピュータにとっては単なる文字列であるので，まず，文字列のどこが単語であるか，知らなければなりません．形態素解析は，入力文を単語に分割し，各単語に品詞などの情報を付与します．

■固有表現抽出

テキストが単語に分けられたとしても，複数の単語が集まって特別な意味を持つ場合が多くあります．固有表現抽出は，人名，地名，組織名等（固有表現）を単語列から抽出します．固有表現はテキストの特徴を表すキーワードになりやすいため，情報検索などの場面で有用です．

■係り受け解析

「誰がどうした」など，文の意味を情報として把握するためには，まず文の構造を知らなければなりません．係り受け解析は，日本語の文の構造を，どの文節がどの文節を修飾するか（係るか）という観点で解析します．

次に，これら技術の仕組みと，私たちの取り組みを順に紹介します．

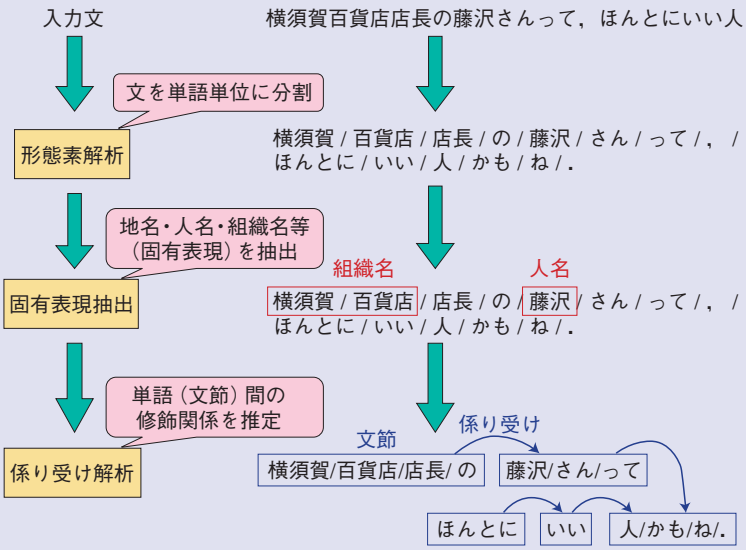

図1　テキスト基本解析技術の概要

形態素解析エンジンJTAG

日本語のように，スペースなど用いずに記述する言語では，まず「どこからどこまでが何という単語か」を知る必要があります．形態素解析では，あらかじめ「知っている」単語を辞書に格納しておき，これを引くことで，入力文を一番もっともらしい単語列に分割します．辞書に品詞（名詞，動詞など，単語の種類）や読みなどを付与しておくことで，単語にさまざまな情報を付けることができます．

一般的には，形態素解析は，辞書

から単語の候補を取得し，単語どうしが接続できるかどうか，できるなら，それはどの程度のコストかを評価しながら，入力文を単語に分割します．**図2**に例を示します．接続可能性をルールで記述するタイプと，確率モデル（後述）で記述するタイプがあります．

私たちが開発した形態素解析エンジンJTAG [(1)] は，ルールベースの解析器です．辞書に品詞，読みのほかに日本語語彙大系に基づく意味カテゴリを含んでおり，格納語数も約90万語と豊富であることと，大量文書を処理できるよう，高速であることが特徴です．

JTAGは日本語専用の形態素解析ですが，単語に区切らずに記述する言語には，日本語のほかにも中国語や韓国語があります．NTTサイバースペース研究所で開発した多言語形態素解析エンジンは，中国語，韓国語，英語を対象に，確率モデルを用いて，精度の高い解析を実現しています [(2)]．

固有表現抽出エンジン NameLister

たとえ文が単語に分割されたとしても，複数の単語が集まって特別な意味を持つことは多くあります．例えば，「日本」「電信」「電話」という3つの単語が連続して現れると，「日本電信電話」という会社名になるなどです．単語の組合せにより無限のパターンがあるため，形態素解析とは別の処理として実行されます．

私たちの開発した固有表現抽出エンジンNameListerでは，**表**に示す8種類の固有表現を抽出します [(3)]．これらはテキストの特徴を表すキーワードになりやすく，情報検索には必須の情報となります．

■系列タギング

単語列から固有表現を抽出するため，NameListerでは，「タグ」という概念を導入しています．固有表現の開始単語に開始マーク付きのタグ（B-種別名），2単語目以降には後続マーク付きのタグ（I-種別名），固有表現ではない単語にはその他タグ（O）を付与します．タグという概念を導入することにより，各単語にもっとも適したタグを付与すれば，固有表現が抽出できるようになります．これを系列タギング（または系列ラベリング）と呼びます．

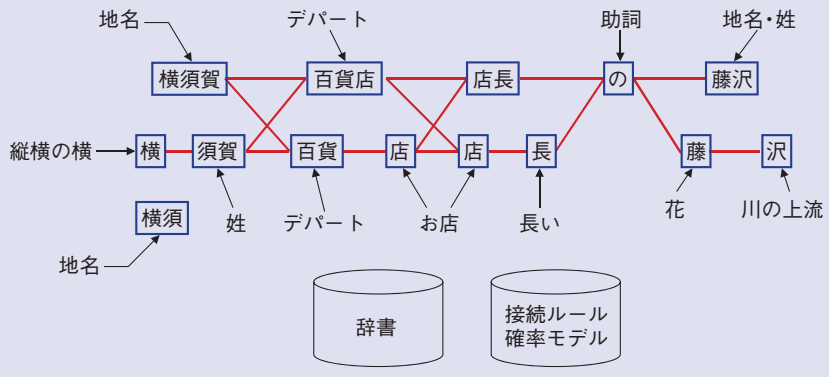

図2　形態素解析の例

NameListerの系列タギングは，条件付確率場（CRFs：Conditional Random Fields）という確率モデルに基づいて [(4)]，最適なタグ列を探索するようにつくられています．確率モデルは，ある程度の量の正解タグが付けられた文（コーパス）から，固有表現の出現パターンを自動的に学習します．簡単な例では，「○○さん」というパターンであれば，「○○」は人名である可能性が高いとか，「鈴木」であれば，前後にどのような単語があっても人名らしいとかです．確率モデルは，このような「らしさ」を数値として表現したものです．私たちが使う言語は，このようなパターンを無限に近く持っており，人間がすべて列挙するのは困難です．NameListerは実例から自動学習することにより，人間が気付かないような微妙な「らしさ」を数値化しています．

実際の固有表現抽出は，この確率モデルを参照しながら，すべてのタグの組合せの中から，もっとも「らしい」タグ列を探し出す（探索する）ことにより実現されています．探索の例を**図3**に示します．図3は「藤沢」という単語にタグを付与する場合ですが，当

表　固有表現の種別

種　別	意味，例
人　名	姓名
地　名	国名，県名や住所，建物など
組織名	会社などの組織，グループ
固有物名	商品名や映画・書籍タイトルなど
日　付	2008年，1月1日，去年など
時　間	8時30分，8時半など
金　額	200円，$3 000など
割　合	30%，五割など

該単語以外にも，前後数単語の情報（文脈）を用いて，すべてのタグの「らしさ」を計算し，もっとも「らしい」タグを付与します．この例の場合，直後に「さん」が出現しており，直前に助詞「の」が出現しているため，人名の開始単語（B-人名）が一番もっともらしいタグと判断されます．文脈も考慮して最適なタグ列を決定しているため，もし入力文が，「横須賀から藤沢まで電車で行きました」なら，電車で行ける「藤沢」は地名であると，正しく解析されます．

NameListerで採用されている系列タギングは，日本語以外にも同様に適用できるため，現在は中国語，韓国語，英語の固有表現抽出も可能になっています．

係り受け解析エンジン Jdep

形態素解析および固有表現抽出は，単語レベルの解析機能でした．それに対して係り受け解析は，文レベルの解析機能です．最終的には，解析対象であるテキストの意味を理解してさまざまな処理を行うのが理想ですが，意味を理解するためには，まず文の構造を把握しなければなりません．係り受け解析は，このような文の構造解析機能です．

■係り受け解析とは

日本語の場合，文の構造は，通常文節と呼ばれる基本的な句と，どの文節がどの文節を修飾しているか（係るか），という2つの要素で表現されます．例えば，「望遠鏡でカゴを持った少女を見た」という文があると，文節としては「望遠鏡で／カゴを／持った／少女を／見た」と分割されます．「望遠鏡で」が係るのは「見た」ですので，「望遠鏡で少女を見る」という文に解釈されます．もし誤って，「持った」に係ると解析してしまうと，「望遠鏡でカゴを持つ」という，通常は考えにくい意味になってしまいます．このように，文の構造を知ることは，「誰がどうした」という，文の内容を知る重要な手掛かりとなります．

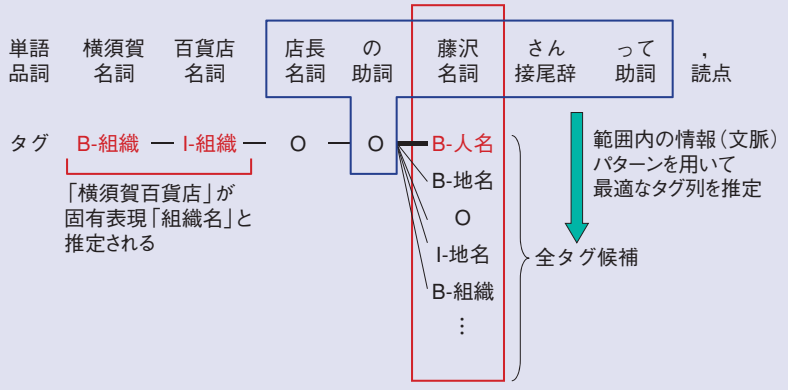

図3　固有表現タグ列の探索

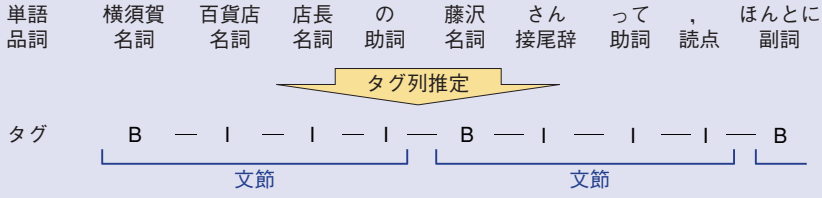

図4　系列タギングを用いた文節分割

■Jdepの仕組み

係り受け解析エンジンJdepの文節分割も，NameListerと同様に，系列タギングを用いて実行されます．**図4**にその例を示します．文節分割の場合，文節開始単語（B）か，後続する単語（I）か，という2種類のタグを各単語に付与します．B-I-I-の連続を取り出せば，文節列が得られます．

文節が決定すると，次にどの文節がどの文節を修飾しているか，決定します．これを狭義の係り受け解析と呼んでいますが，ここでも，系列タギングが使われています．**図5**に例を示します．係り受け解析の場合，各文節に対して，修飾先の文節の相対位置をタグとして付与します．図5の例ですと，「藤沢さんって，」という文節は，「人かもね．」を修飾しているため，タグとして3D（3つ先の文節を修飾）を付与しています．

■CGM文書の解析

インターネット上には，ブログなどユーザが直接発信した文書（CGM：Consumer Generated Media）が多数存在しています．これらには話し言葉的な要素も多く，「えーっと」

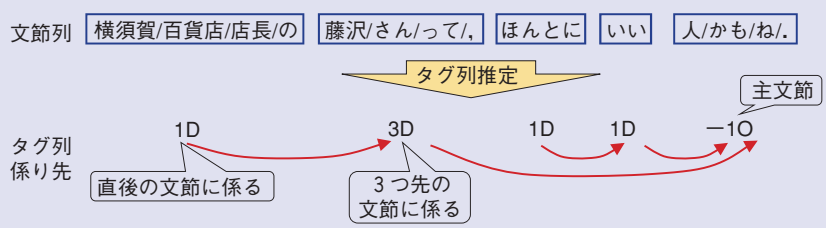

図5　系列タギングを用いた係り受け解析

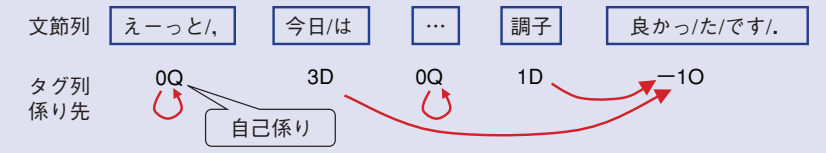

図6　独立した文節を含む文の係り受け解析

「……」のように間を持たせるためだけに書かれているものや，顔文字など，修飾先がない独立した文節が含まれます．Jdepでは，系列タギングのタグに「自己係り」を導入することにより，どこにも修飾先がない独立した文節を，「独立している」と明示して出力します．**図6**は，独立した文節を含む文の解析例です．この機能を導入することで，CGM文書の解析時にも，意味の理解に直接必要な文構造だけを取り出せるようになりました[(5)]．

技術の適用分野

前述の3つの技術は，さまざまなインターネット上のテキスト処理に適用可能です．例えば，形態素解析を検索エンジンに適用すると，単なる文字列一致ではなく，単語を考慮した検索ができるようになるため，「京都」で検索するときに「東京都」も検索されてしまうというミスを防ぐことができます．固有表現抽出は，キーワードになりやすい表現を抽出できるので，検索結果のランキングや質問応答に有効です．係り受け解析は，「どうした」に係る文節が分かるので，これを集計すると，「最近急に動かなくなったもの」を大量のテキストから発見したりするなど，マイニング分野に応用することができます．

今後の取り組み

インターネット上のテキストであっても，書かれている意味を理解したうえで処理をするのが理想です．本稿で紹介した3つの基本解析技術は，意味理解のための重要な要素ですが，これだけではまだ「誰がいつどこで何をどうした」までは解析できません．また，一般的なテキストは読み手が知っている要素などは省略されることも多く，コンピュータがテキストを理解して処理をするのはまだまだ困難です．今後はまず，日本語を対象に，文の意味理解や，省略要素の同定などの解析技術を検討していきます．

■参考文献

(1) T. Fuchi and S. Takagi：“Japanese Morphological Analyzer using Word Co-occurrence -JTAG,”Proc. of COLING-ACL, pp.409-413, 1998.
(2) K. Saito and M. Nagata：“Multi-Language Named-Entity Recognition System based on HMM,”ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, 2003.
(3) http://nlp.cs.nyu.edu/irex/
(4) 鈴木・磯崎：“学習誤り最小化に基づく条件付き確率場の学習：言語解析への適用，”言語処理学会第12回年次大会発表論文集，pp.548-551，2006.
(5) 今村：“系列ラベリングによる準話し言葉の日本語係り受け解析，”言語処理学会第13回年次大会発表論文集，pp.518-521，2007.

（左から）今村　賢治/ 齋藤　邦子/
浅野　久子

人にやさしい自然言語処理技術の研究開発に取り組んでいきます．

◆問い合わせ先
NTTサイバースペース研究所
TEL　046-859-2687
FAX　046-855-1054
E-mail　imamura.kenji@lab.ntt.co.jp