基応般

2. 自然言語処理技術の概要

乾 健太郎（東北大学）　関根 聡（ニューヨーク大学）

自然言語処理の現状と期待

自然言語処理は自然言語の機械処理，平たく言えば「言葉が分かる」計算機システムの構築を目指す研究領域である．もちろん，言語の理解というのは遠大な目標であり，我々はまだ途についたばかりだ．しかし，今世紀に入って機械学習に基づく統計的アプローチが大きな成果を上げ，ビッグデータ時代の到来とともに応用領域も急速に広がりつつある．

自然言語処理への期待は大きく２つある．

１つは膨大な言語データに対する情報アクセスである．Web や SNS を駆け巡り蓄積される言語情報，コールセンターや EC サイトに寄せられる質問や意見，研究者も追い切れない速さで急増する科学技術論文など，言語データはどこにでも溢れている．大量の文書を高速に読みこなし，内容を解析する技術があれば，膨大な情報を分析し，有効に活用することができる．情報検索，情報抽出，質問応答，テキストマイニング，文書要約など，さまざまな応用技術が研究され，商用化されている．また，１つの言語に閉じず複数の言語間での課題である機械翻訳の研究にも多くの研究者が取り組み，実用化されつつある．

もう１つは人間と機械のコミュニケーションである．音声認識の精度が飛躍的に向上し，アップルの Siri のような音声対話アプリが多くのユーザを獲得するまでになった．今後，モバイル端末やウェアラブルデバイスの普及によってこの流れはさらに加速すると想像される．

本特集の導入として，本稿では自然言語処理の基本事項を紹介する．なお，以下ではもっぱら日本語の例を使って話を進めるが，本質的な問題の多くはどの言語にも共通であり，ここでの考え方は他の言語にも適用できる．

自然言語処理の基本タスク

情報抽出や質問応答，機械翻訳など，自然言語処理のさまざまな応用タスクの核となる部分は，**図 -1** に示すような基本タスクの組合せに分解することができる．以下，それぞれについて概観する．

形態素解析と固有表現抽出　形態素解析は，入力文を単語に分割し，各単語の品詞を特定するもので，自然言語処理の最も基本的なタスクである．「母 / とそ / ば / の / 実」のような間違った分割でなく，「母 / と / そば / の / 実」のように正しい分割を選択するにはどうすれば良いかが問われる．固有表現抽出は，「そばの実」のような組織名のほか，人名や商品名など，入力文章に出現する多種多様な名前（固有表現）を同定する処理である．固有表現は日々新しく生産され，ロングテールがきわめて長いため，すべての表現をあらかじめ辞書に登録しておくことは難しい．膨大な数の固有表現をどのように収集するか，未知の固有表現をいかに文脈から推定するかが大きな課題である．

構文解析　文は単なる単語の並びではなく，階層的な構文構造を持っている．構文構造の表現方法には主として，名詞句・動詞句・名詞節といった句や節からなる**句構造**と，図の例のような単語や基本節（日本語では文節）の間の修飾関係からなる**係り受け構造**がある．係り受け構造を特定する「係り受け解析」では，単語列を文節にまとめ上げ，文節間の係り受け関係を特定する．「母と」は「母と子」のように並列構造になる可能性があるが，図の例では「母」と「そばの実」の並列ではなく，「母と」は「行きました」に係る．この関係が正しく

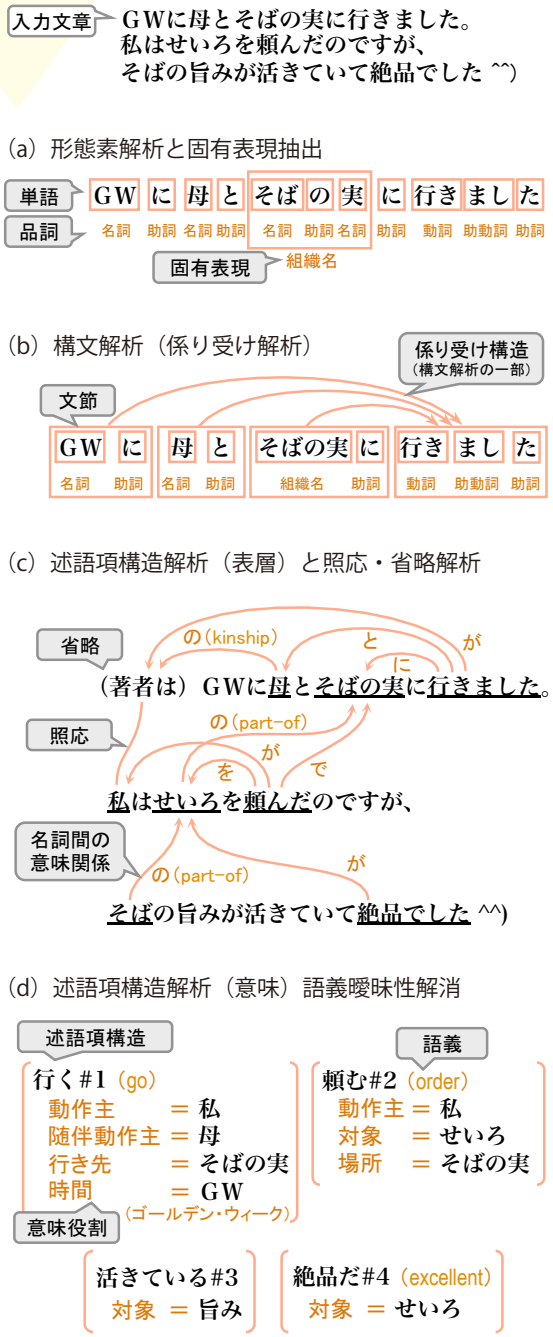

図 -1　自然言語処理の基本タスク

認識できないと，「母と」を“with mom”と翻訳することもおぼつかない．

述語項構造解析と照応・省略解析　自然言語では，文脈から容易に特定できる要素はしばしば省略される．図の例では，「そばの実に行った」のが誰か，「絶品だった」のは何かといった情報が明示的には書かれていない．つまり，構文解析だけでは「誰が何をどうした」のような情報を十分に抽出できない．「どうした」に相当する要素を**述語**，「誰が」や「何を」に相当する情報を**項**といい，全体を**述語項構造**という．図の例では，述語「行く」が「私が」，「母と」，「そばの実に」などの項を持っている．述語と項の意味的な関係を区別したい場合は，図 -1 (d) のように〈動作主〉，〈行き先〉などの**意味役割**の特定までをタスクに含める．述語項構造の解析は文の意味の解析への入り口である．また，図の例のように，省略（第 1 文の「著者」や第 2 文の「せいろ」と「絶品だ」の関係）や照応関係（第 1 文の「著者」と第 2 文の「私」など）の解析と重なりがある．

語義曖昧性解消　図の例の「GW」は「ゲートウェイ」や「ギガワット」ではなく，「ゴールデンウィーク」のことである．「頼む」は「注文する（order）」の意味で，「懇願する（beg）」とは少し違う．このように自然言語では，まったく違う単語が偶然に同じ表記を持っていたり，同じ単語がいくつかの語義を持っていたりするため，テキストから意味的な情報を抽出したり，翻訳したりする場合に，それぞれの単語がどの意味を指しているかを文脈に照らして特定する必要がある．

同義・含意関係認識　自然言語には同じ情報を表現する（あるいは含意する）言い回しがいくつも用意されており，そのことが言語処理のさまざまな応用で障害になる．たとえば，「ざるそばが美味しい」と言っている口コミを検索したいときに，単純な文字列照合だけでは図 -1 の記述を見つけることはできない．「ざるそばが美味しい」と「せいろが絶品だ」の同義関係を認識する何らかの仕組みが必要である．

自然言語処理の応用タスク

前章で紹介した基本タスクの実現はさまざまな応用タスクの実現・高度化につながる．本稿では，紙面の制約から個別の応用タスクに言及することは控えるが，1 つの例として情報抽出の一種である意見マイニングを取り上げ，個々の応用タスクの中核的問題が前章の基本タスクの組合せに分解できること

を例示する．

意見マイニングとは，ソーシャルメディア上の言語データや社会調査の自由回答アンケートのような文書集合から個人が発信する意見を抽出し，構造化情報として蓄積することにより，ユーザの関心に合わせて検索したり，要約・分析したりすることを可能にする情報加工サービスである（本特集「3.11 レビュー解析」を参照）．図 -1 の例文からは，そこに含まれる著者の意見をたとえば**図 -2** のような構造化された情報として抽出する．単なる文字の羅列でしかなかった元の文章に比べて，図 -2 の情報は明示的な構造を持っており，情報の検索や分析に都合がよい．

図 -1 の入力文章から図 -2 の構造化情報を抽出する過程で図 -1 の (a) から (d) のような基本問題を解決する必要があることは比較的容易に想像できるだろう．述語「絶品だ」が〈評価〉を述べる表現であることが分かっている（辞書に書かれている）とすると，最初の目標は「何が絶品だ」と言っているのかを解析することである．これは「絶品だ」の述語項構造の項〈対象〉を特定することに対応する．また，「何が」の省略を解消することでもある．ここで「絶品だ」の〈対象〉，すなわち「せいろ」は，より正確には「そばの実」で出された「せいろ」であるので，その関係が分かれば，「絶品だ」の〈評価の対象〉と〈評価の着眼点〉が特定できる．「そばの実」と「せいろ」の関係を認識するには，「そばの実」を飲食店名と認識する固有表現解析も必要であろう．さらに，抽出された多数の情報を類似意見ごとにまとめて俯瞰したい場合は，同義・含意関係認識（「せいろが絶品だ」と「ざるそばが美味しい」）も入ってくる．

経験的手法による言語処理

自然言語処理の基本タスクはいずれも解釈の選択肢の中から正しい解釈を選択することであるので，原理的にはすべて分類問題あるいはラベル付け問題に帰着させることができる．

たとえば，形態素解析の単語分割は，入力文字列

意見のタイプ	＝〈評価〉
評価者	＝〈著者〉
評価の対象	＝ そばの実
評価の対象のクラス	＝〈飲食店〉
評価の着目点	＝ せいろ
評価の着目点のクラス	＝〈料理〉
評価の値	＝ 絶品だ
評価の値のクラス	＝〈ポジティブ〉

図 -2　構造化された意見情報

のそれぞれの文字と文字の間が単語の切れ目になるか否かの 2 値分類問題が 1 列に並んだ系列ラベル付け問題と考えることができる．各単語に品詞を付与する問題も同様に系列ラベル付け問題である．同様に，係り受け解析は入力の各文節について係り先の文節を 1 つ選ぶ問題，述語項構造解析も入力の各述語の各項のフィラーをしかるべき候補の中から選ぶ問題と見なせる．つまり，ラベル付け問題の組合せとして定式化できる．機械学習，情報抽出，文書要約といった応用タスクについても，中核的な問題は基本タスクの組合せと見なせるので，やはりラベル付け問題の組合せに帰着できる．

このように言語処理の問題を一旦ラベル付け問題として定式化できれば，種々の統計的機械学習アルゴリズムが適用できるようになる．基本タスク，応用タスクを問わず，幅広い問題で統計的手法が活発に研究され，電子的な言語データの急増，機械学習理論の発展，計算機の能力の飛躍的な向上と相まって，自然言語処理はこの 20 年の間に飛躍的な発展を遂げた．

人間が正解ラベルを付与した訓練事例集合から学習する教師あり学習の場合，解きたい問題とよく似た訓練事例を十分に用意できれば，良い性能が期待できる．たとえば，新聞記事については日本語・英語ともに比較的大規模な正解付き訓練事例が早くから整備されたため，現在の形態素解析器や構文解析器は新聞記事に対しては実用レベルの精度で解析できる．しかし，技術論文や小説，あるいはメールやソーシャルメディアといった，大規模な訓練事例が存在しないジャンルでは，同様の性能を得ることは

難しい．そこで，半教師あり学習，転移学習など，少量の訓練事例から，あるいは別のドメインの訓練事例からでも効果的な学習ができる手法がひろく研究され，一定の成果を収めるに至っている．

訓練事例の不足に対するもう１つの方策は知識の整備である．幅広いタスクで有用な言語知識・世界知識をあらかじめ整備し，これを効果的に使うことによって，訓練事例の不足を補うことができる．知識の整備と共有化に関する研究は言語処理全般にわたる重要な課題である．

たとえば，図 -1 の係り受け解析では，(i) 述語「行く」が「○○と」で表示される項〈随伴動作主〉をとり，その要素には「人間」が入ること（格フレーム），(ii)「母」は「人間」であること（概念階層）を知識として持っていれば，「母と」が「行きました」に係る解釈をうまく選択できると考えられる．こうした知識は照応・省略解析でも有用である．「そばの実」という名前の蕎麦屋が知識ベースに入っていれば，固有表現抽出も容易になる．同義・含意関係認識では，「せいろ」と「ざるそば」，「(料理が）絶品だ」と「(料理が）美味しい」のような，語彙的な同義・含意関係の知識が必要である．

大規模知識獲得の可能性

こうした語彙レベルの言語知識や世界知識は一般にきわめて長いテールを持ち（きわめて広範な低頻度表現の知識が必要），またドメインによる違いもあるため，人手で網羅的に収集するのはほとんど不可能である．そこで，こうした広範な知識を大量の言語データ（生データ）から自動的あるいは半自動的に獲得する知識獲得の研究が精力的に進められ，近年その成果が顕在化してきた（本特集「3.7 知識獲得」を参照）．

知識獲得の高度化・大規模化は，現在はまだほとんど解けていない省略解析や文章の「行間」を読む深い文脈解析の研究に大きな変革をもたらすと期待できる．計算機システム自身が大量の文章から知識を集め，集めた知識を使って文章の行間を読む．そんな可能性が少し見えてきた．

近年劇的な復活をとげ，日進月歩の発展を見せている深層学習（ディープ・ラーニング）もこの流れの中で重要な役割をはたすだろう．深層学習の強みである特徴学習は知識ベースを支える意味表現・知識表現を一新する可能性もあり，今後の動向に注目したい．

(2015 年 11 月 11 日受付)

乾 健太郎（正会員）inui@ecei.tohoku.ac.jp
東北大学大学院情報科学研究科教授．1995 年東京工業大学博士課程修了，博士（工学）．同大学助手，九州工業大学助教授，奈良先端科学技術大学院大学助教授を経て，2010 年より現職．2014 年度本会論文誌編集委員長，同年度より本会自然言語処理研究会主査．

関根 聡（正会員）sekine@cs.nyu.edu
New York University Associate Research Professor. 1998 年 NYU Ph.D．松下電器産業，University of Manchester，ソニー CSL，MSR，楽天技術研究所ニューヨークなどでの研究職を歴任．ランゲージ・クラフト代表．専門は自然言語処理，特に情報抽出，固有表現抽出，質問応答の研究に従事．