Special 特集 feature

ここまできた
自然言語処理

－例文の収集とその利用－

Now Natural Language Processing
Has Come to This Stage
- Collecting and Using of Example Sentences -

編集にあたって

冨浦 洋一 tom@is.kyushu-u.ac.jp
九州大学大学院システム情報科学研究院

渡辺 日出雄 hiwat@jp.ibm.com
日本IBM（株）東京基礎研究所

本特集では，近年の言語コーパスを用いた自然言語処理技術の進展と今後の可能性を読者に紹介する．

急速な計算機ハードウェアの進歩と言語コーパス（例文集）に基づいた手法の導入により，1990年代に入って自然言語処理技術が大きな進展を遂げている．自然言語処理技術を用いたアプリケーションには，機械翻訳，テキスト要約，検索など，インターネット上の大量の情報の中から必要な情報を取得するために有用なものが多い．これらのアプリケーションの基盤技術は構文解析技術であり，翻訳や要約の質，検索のヒット率などのアプリケーションの精度は，構文解析システムの精度に大きく依存する．

構文解析とは，簡単にいえば，述部は何か，その目的語は何か，といった文の構造を解析するものである．一般に1つの文に対して複数の構文構造の候補が得られ，しかも，意味的におかしな構文構造も多数含まれる．たとえば，「赤いイタリアの車」や「赤い帽子の子供」のような＜形容詞＞＜名詞＞「の」＜名詞＞の形をした名詞句を考えよう．『連体形の形容詞は名詞に係る（名詞を修飾する）』という単純な文法規則に従うと，この名詞句の先頭の形容詞はどちらの名詞にも係り得る．しかし意味を考えれば，「赤いイタリアの車」では，「赤い」は「車」に係り，「赤い帽子の子供」では「赤い」は「帽子」に係るとするのが正しい．したがって，いかにして正しい構文構造を選択するか（構文構造の曖昧さを解消するか）が構文解析における大きな問題である．

従来，構文解析システムは，経験則などに基づいて，曖昧さを解消する規則を人手で用意することにより構築されていたため，規則作成者の言語的センスに強く依存するものとなり，保守・拡張性の観点からは問題があった．それに対して，近年，言語コーパスを用いて学習した統計的言語モデルに基づいた曖昧さ解消法や，言語コーパス中の類似用例に基づいた曖昧さ解消法などのコーパスベースの手法が盛んに研究されるようになった．このコーパスベースの手法は，従来の方式で開発されたものに精度の点で比肩し得るほどになりつつある．

第1編では，構文解析の重要性，構文構造の曖昧さを解消するための古典的な手法とその限界について述べ，言語コーパスを用いた構文解析の概要を解説する．

第2編では，良質なコーパスを作成するという問題に関して，京大コーパスの作成を通して得られた知見を披露してもらう．コーパスに付与する構文構造は正確で一貫性のあるものでなければならない．これをすべて人手で行うのは困難であり，構文解析器（パーサー）による支援が有効である．一方，高精度のパーサーの開発に当たっては，大規模な構文構造付コーパスが必要となる．この「鶏が先か卵が先か」のような問題の1つの解決法を与える．

第3編では，言語資源の共有化の動向について解説する．言語コーパスの開発には多大な労力を必要とするため，大規模かつ良質の言語コーパスの構築法として，さまざまな研究機関で開発した言語コーパスを共有することも考えられている．

第4編では，自然言語処理と密接な関係を持ち，かつ，コーパスの利用ということでは長い歴史を持つ音声認識の分野での状況を紹介する．十分な性能のシステムを開発するために必要な，音声コーパス，テキストコーパスの量と質について考察し，また，そのようなコーパスを整備するための取組み方について述べる．

第5編では，第1編の内容を掘り下げて，有限のコーパスを有効利用するための手法（モデルに取り入れる情報の選択手法やパラメタの推定手法など）について解説する．

第6編では，本特集のまとめとして，コーパスベースの言語処理の背景，適用範囲，現状について概観し，コーパスベースの言語処理の得失と評価における問題点について述べ，今後の展望について論じる．

最後に，ご多忙にもかかわらず，快く執筆をお受けいただいた筆者の皆様，ならびに，閲読をいただいた皆様に感謝申し上げます．

（平成12年5月23日受付）