言語処理の歴史と展望

長尾　眞

概要

１９４５年前後にコンピュータが誕生した直後から歴史が始まった。コンピュータは論理操作をする機械であるから、文字列、単語列を変換して他言語の文字列、単語列に直すことによって機械による翻訳ができるだろうと考えられ、１９５０年ころから研究が始まった。一方、コンピュータは大量のテキストやデータを記憶することができるので、そこから必要な情報を取り出す情報検索のシステムが作られるようになってきた。

これらの処理をするために、単語抽出技術、文字列照合技術、テキスト並べ替え技術、構文解析技術、辞書構造の在り方など、多くの研究がなされた。初期のころはデータ量がそれほど多くなかったこともあって、アルゴリズム(理論)中心で種々の研究が行われたが、１９９０年代に入るとコンピュータネットワークが社会に浸透し始め、ネット上に膨大なテキスト情報が存在するようになり、研究はデータ中心に移行してきている。今日では数十億の文をスーパーコンピュータで処理できるようになり、事例ベースのシステムが作られるようになってきている。

これまでの言語学は人間の集めうる範囲の言語データをもとに研究が行われ、人間の思考能力の範囲で文法などの理論が作られてきた。しかしこれらの理論は複雑多様な言語現象に潜んでいるであろう言語の骨格を推定した、いわば第一近似のものであって、それでは説明できない言語現象は膨大に存在する。

コンピュータネットワーク上には、子供から大人まで、男女、あらゆる場面での発話など、生きた言語が膨大に存在するので、これらをくまなく集め、それらの言語データを利用目的に従ってどのように処理するかが問われる時代になってきている。すなわち言語学研究者としては何十億文という言語データ(テキストデータ、音声データなど)をどのような観点からどのような処理をすれば何が分かるかに注力しなければならないし、言語工学者はこの膨大なデータをどのように利用して機械翻訳、情報検索、事実検索、自動要約、人と機械との対話システムなどを作り上げてゆくことができるかを考えねばならない。我々に課された課題は大きく、また深刻であるといわねばならない。

本論ではこのような言語処理技術の歴史的変遷について述べ、今後の方向性について議論を行う。その内容は以下のとおりである。

Ⅰ　草創期（１９４５－１９５５）

Ⅱ　機械翻訳発展の時代（１９５５－１９６５）

Ⅲ　計算言語学の出発（１９６５－１９７５）

Ⅳ　実用システム開発の時代（１９７５－１９８５）

Ⅴ　新しい展開の時代（１９８５－１９９５）

Ⅵ　巨大言語データ利用の時代(１９９５－　　)

Ⅶ　これからの課題