日本語処理における頑健な文節区切り手法とその応用

第７章

結論

本論文では、自然言語における曖昧さの取り扱いを大きく、２つの立場から行なったものである。

これは、筆者が１９８３年に日本アイ・ビー・エム株式会社サイエンス・インスティテュートに入社してから１９９０年まで所属していた日本アイ・ビー・エム株式会社東京基礎研究所に至るまでの７年間と、東京家政学院筑波短期大学 (現在東京家政学院筑波女子大学短期大学部) 情報処理科において筑波大学電子・情報工学系非数値処理アルゴリズム研究室との共同研究を行なってきた研究活動の歴史でもある。

この歴史は、１９８３年にサイエンス・インスティテュートにおいて、知識ベースを用いた自然言語処理の検討を始め、約１年間の調査・検討ののちに「ＣＲＩＴＡＣ」プロジェクトとして、確率的手法を用いた漢字短単位分割を行なうことから始めた。それまでは文節単位かな漢字変換や文節内部構造の決定の仕方等、日本語処理の基礎とも言える形態素解析に関わる処理は、多くの人手をかけて細かく一つ一つの辞書項目について情報を蓄えた大規模な辞書を用いて行なうのが一般的であった。その結果、長期的に多くのマン・パワーをかけて作成された辞書程、形態素解析における文節区切り、品詞付与等、全般の精度が高くなり、ある程度以上の精度を保つシステムにおいては、どれだけの「例外」を扱うことので切る「例外辞書」を用意しているか、という点が問題になるにいたった。この従来手法の問題点は、例えば、ある特定の目的のために辞書を整備した場合、目的が変化することによって、辞書を新たに造るところから始めなければならない、ということにある。しかし、従来のこの方法を踏襲するには、サイエンス・インスティテュートにおける人員、および許される期間が短すぎて他のシステムに追随するのが困難であることが明らかである。そこで、必要最小限の人員とデータを基に、効率の良い形態素解析を行なう方法として既に機械可読な形で用意されていたＪＩＣＳＴの文献抄録を用いてまず文節内の漢字複合語の解析を行なう手法を実現する方法について検討を始めた。

当時は、あくまでも従来の方法に依存した、大規模辞書を用いた形態素解析を行う手法が主流

で、統計的な手法を用いてどれだけの成果がでるか疑う人も多かったが、大学等で効率よく仕事を進めるために、筆者らと同様のマルコフモデルを用いて漢字複合語の解析を行なうシステムが造られるようになり、その手法には期待がもたれている。

第3章では、統計情報を用いた文字列パターンを用いた日本語文自動分割について述べた。この方法では従来の形態素解析のように接続関係をチェックしないため、ワードプロセッサによって作成された誤変換や、未変換を含むような、ノイズのある文書に対しても精度の高い文節区切りを行うことが可能である。

第4章に述べたように、実際に構造化文書に第3章で提案する手法を用いて文節区切りを行ない、ワードプロセッサで作成された文書に対して誤った箇所を提示するＣＲＩＴＡＣという日本語文書校正支援システムを試作したところ、日本語の前処理(形態素解析、辞書引き)に関する部分については十分な精度が得られた。文書校正部分に関してはまだルールの数が少なく、現段階で実用化するには不安もあるが、本方式の基本部分を用いて、新聞社の校正支援システムが稼働中である。今後、ワードプロセッサで作成された文書の誤りについてさらに調査を行ない、系統的な校正ルールを整備することにより、さらに効率良く文書校正が支援可能であると考えられる。

第5章に述べた統計的な手法を用いた日本語の形態素解析の曖昧さ解消方式については、対象を限定することなく調査を重ねることによって、例えば「と」の使われ方の確率分布等が得られ、かつその使用状況に関する条件が特定できれば、この方式によって形態素解析の結果の曖昧性を減らせるとともに、その後の構文解析における係り受け解析可能性の組合せ数の爆発を防ぐことが可能である。

第6章で述べた点字翻訳ボランティアのための分かち書き支援手法については現在も調査・研究中である。現段階ではまだ点字翻訳のための「分かち書き処理」のための手法までしか確立されていないが、今後、さらに検討をかさね、ユーザの修正を学習したりすることでさらに精度の高い、使い勝手の良いシステムのための点字翻訳手法を構築する計画中である。

以上見てきたように、筆者が行なってきた確率的、あるいは統計的日本語処理のための基本アルゴリズムは、より大量のデータに関して調査し、統計量や確率の値を求めることにより十分実用的である。

また、従来の形態素解析では文書中に未知語(誤り)が出現した際、未知語の切り出し処理と未知語の前後の語の接続関係のチェック、それぞれのレベルで処理に誤りが含まれる可能性があり、未知語が解析に与える影響は大きい。一方、本論文で提案する手法では、統計的な文字列情報とヒューリスティックなルールベースを用いて文章を文節に区切るため、誤りが局所的に留まり、周囲に与える影響が少ないという点で、形態素解析を行うよりも未知語に対して頑健であるといえる。

今後も日本語処理の曖昧さをひきおこしている各種条件の検討・分離をはかり、本手法の精緻

化を行うとともに、より実用的なシステムのための手法の確立を目指すことを考えている。