公募研究：2000年度

自然言語処理の応用によるゲノム文献の高度検索システムの構築

●黒橋 禎夫

京都大学大学院情報学研究科

〈研究の目的〉

自然言語の文章では，人間にとって理解可能な範囲で頻繁に省略や代名詞化がおこる．この問題は，文章を単語集合として扱っている現在の情報検索でさほど表面化しないが，今後，情報検索を高度化していくためには，省略・代名詞に対する照応詞の同定が必須の要素技術となる．

省略・代名詞解析には，用言(動詞，形容詞，名詞＋判定詞)ごとに，どのような名詞が主語，目的語(格要素)になるかという情報をまとめた格フレーム辞書が必要となる．しかし，数千から数万の用言について，専門分野における特殊な用法までカバーする大規模で実用的な格フレーム辞書はこれまでのところ存在しなかった．

本研究では，格フレーム辞書を大規模テキストから自動構築する方法を考案し，さらに，構築した格フレーム辞書を用いて実際に省略の解析を行い，その有効性を検証した．

〈研究成果〉

格フレーム自動構築における最大の問題は，用言の意味の多義性である．たとえば「(友達に)なる」と「(病気に)なる」，「(塩，調味料などを)加える」と「(砲撃を)加える」では，同じ動詞でも格フレームのパターンがまったく異なる．この多義性を解消しなければ，格フレームは自動的には構築できない．

ここでのポイントは，用言の意味を決定づける重要な名詞は用言の直前にあり，かつそれは文章中で省略されることは比較的少ない，という点である．そこで，用言単独ではなく，用言とその直前の名詞のペア，すなわち「友達になる」や「病気になる」を格フレームの単位とし，そのまわりに他にどのような格要素が存在するかを大量のテキストから学習するという手法を考案した．

新聞記事を対象とし，約370万文から格フレームを学習したところ，9,900用言について平均6.0個の格フレームが学習された．さらに，この格フレーム辞書を用いて文章中の省略要素を同定する実験を行ったところ，70%程度の正解率が得られた．この手法は言語独立，分野独立であり，必要となるのはある分野の大量のテキストだけである．

〈研究の達成度〉

自然言語処理の高度化のための最も基本となる知識源，格フレームについて，意味的曖昧性の問題を解消し，テキストからの完全自動による学習方法を考案した点は大きな成果である．なお，計算機による自然言語処理の高度化は非常に広い文脈で今後の社会にとって重要であり，その中の基盤技術となる格フレーム学習，照応省略解析にじっくり取り込むことを想定していたが，ゲノム情報科学ではより即戦力的な文献検索の改善が切望されていたため，公募研究への応募は2年目以降見合わせた．

〈研究期間の全成果公表リスト〉

論文／プロシーディング

・Kawahara, D. and Kurohashi, S.: Japanese Case Frame Construction by Coupling the Verb and its Closest Case Component, Proc. of Human Language Technology Conference (HLT 2001), San Diego (2001.3)

・河原大輔,黒橋禎夫: 用言と直前の格要素の組を単位とする格フレームの自動構築, 自然言語処理, Vol.9, No.1, pp.3-19 (2002.1)