

共起関係データの今後の研究について

5F-1

田 中 康 仁

姫路短期大学

吉 田 将

九州工業大学

1) はじめに

機械翻訳システムが実用化されはじめてきた。しかし、色々な問題がまだ多くあり、その問題の解決がはかられている。一つの方法としてはポスト・エディット、プレ・エディットといった解決方法である。これは一つの解決方法ではあるが、機械翻訳システムの本質的解決方法ではない。

機械翻訳システムをもう一度みなおす時期に来ている。このためには文法と辞書の改良が必要である。ここでは辞書、特に共起関係について考えてみる。

2) 共起関係データの有効性

共起関係データの有効性については仮名漢字変換システムの研究の中で又機械翻訳システムの研究の中で述べられてきた。しかし、ハードウェアのメモリスぺースと処理能力、共起関係データが充分でなかったため実用化されなかった。筆者等の研究により、本として資料を公開することにより、実用化への動きが活発になってきた。例えば

- 1) 市販のワードプロセッサではA I 辞書、関連語辞書の名前で実用化されてきた。
 - 2) 音声認識システムでも、A T Rの研究の中で共起データが使われてきた。A T Rではデータ収集を独自に行って研究に使っている。
 - 3) パーザーの中に共起データを組み込み処理能力の向上と曖昧さの減少のために使われている。これはK D Dの研究所で英文の解析に使われている。
 - 4) 機械翻訳のモデルの研究では語と語の関係の共起関係データが使われている。日韓機械翻訳システムモデルが東工大安居院研究室で作成されている。
- このように有効性が確認され、実績ができあがってきた。

3) 共起関係データの収集状況

- I) 'が'について朝日新聞記事データの分析で2種類(84日分、一年分)を作成した。
- II) 'を'についてJ I C S T抄録データと朝日新聞記事データ(84日分)で2種類を作成した。

III) 'に'について朝日新聞記事データ(84日分)で1種類を作成した。

IV) 四文字漢字列データをJ I C S T記事データを分析することにより作成した。

これらにより20冊の資料集を作成し関係者に配布した。

近い将来'で'について(朝日新聞記事データ、84日分と一年分の統合)を作成する。

'に'について(朝日新聞記事データ一年分)を作成する。

'が'について(J I C S T抄録データ)を作成する。

'の'について(朝日新聞記事データ84日分)を作成する。等の計画がある。

知識データの収集計画

Data \ 格助詞等	を	が	に	で	の	四文字漢字列	三文字漢字列
朝日新聞84日分	○	○	○		I	/	/
朝日新聞一年分	P	○	P	I	/	/	/
日本科学技術情報センター(JICST)	○	I	/	/	/	○	○

I: Input中: 1990年夏頃(Data収集終了、Input作業)

P: 計画中: 1991年夏頃(KWIC作成済、Data収集計画中)

/: 計画していない

4) 今後の収集について

機械翻訳システムは長文(例えば40~50文字を超えるもの)になると機械処理の曖昧さと処理能力が急激に増大すると言われている。

これは文法の面での複文、重文についての解決が充分でないことにもよるし、共起データその他の辞書の不十分な点が考えられる。ここでは共起関係データについて考えてみる。

(1) 連体修飾データ

例えば 壊れた 橋、保守的な 人
美しい 姫路城

このように動詞の連体形と名詞、形容詞と名詞の関係等についても共起データを集める必要がある。

(2) 連用修飾

例えば 頻繁に 行う
自動的に 動く

このように副詞や連用修飾語と用言の関係を集める

必要がある。

(3) 並列関係

並列関係の研究は多くなされてきたが、これらの研究は構造を明らかにすることに役立ち、文法等の中には組込まれている。しかし、並列関係の判別には具体的データが必要であるし、さらに詳細な研究をするためには具体的なデータが必要である。並列データを翻訳する場合の語の順序等についても研究する必要がある。これについては一部研究が行われている。

（九工大村田）

(4) 'の'の分析 'の'は名詞と名詞を結合するために文章の中でしばしば用いられる。この構造については研究が行われているが、具体的にデータを集め対訳語の研究等についての本格的研究はなされていない。

これら4つの研究が行われ、データの収集が必要である。今後ますます格助詞以外の共起関係データを研究・収集しなければならない。

このほか長い文を解析するために必要なものとして辞書データとしては次のものがある。

- I) 専門用語（大量に集める）
- II) 複合語（複合名詞、複合動詞）
- III) 慣用表現

また、共起関係データは分野別（医学、法律、経済…）で語集が異なるため各分野別に調べる必要がある。筆者の研究対象としては朝日新聞とJICST抄録データを対象としている。

5) データの評価

どの程度の量の対象文を基準にすればよいかということが問題となる。

また、どのような分野を対象とすればよいかを考えなければならない。

しかし、ここでは朝日新聞記事データとJICST抄録データを対象とする。データ量が多いほどよいが、最低でも一年分程度の記事データを分析する必要があると考えている。

さらに、データが収集されるにつれ抽出したデータがどの程度に適用されるものか評価しなければならない。同時に抽出方法もKWICからの手作業で行うのではなく一度抽出したデータとKWICと照合し既に集めたものの以外のデータを集めるようにしなければならない。

また、共起データの翻訳、日本文-英文の対になった文からの自動収集・半自動収集も検討しなければならない。共起データのカテゴリ化の方法としてはシソーラスとの照合を提案しているが、シソーラスの完備したものがなく実現していない。今後の検討課題である。

6) おわりに

共起関係データについて研究するようになり約9年になるが、やっとこのごろになり研究成果が認識されるまでになってきた。今後、急速にハードウェアが安くなり処理スピードが向上すれば、それに対応するアルゴリズム（文法）と大量で高品質の辞書が必要になるであろう。我々は先を読んで研究をしておかなければならない。

