鍵語に着目した字句分解法の試み

1N－11

山口 政道　佐藤 匡正
(島根大学大学院理学研究科)

1.序論

自然語の文を、構文解析や意味解釈のために、その構成要素である単語に分解する事（形態素解析）が行われている[2]。しかし、日本語の場合、英語文のように分かち書きがされておらず、単語や字句に分解するには大掛かりな単語辞書を用いて形態素に分解することが一般的である。

現在、日本語形態素解析システム（ALTJAWS、JUMAN、Breakfast 等）は、膨大な単語辞書を用いて形態素解析を行っている[3]。

こうした解析システムでは、辞書が鍵となるが、辞書作成、維持、手直しなどに費用や権利など、利用上の制限が生ずる。この様な事情から、解析精度は多少劣っていても、費用や制約の少ない簡便な解析方法の創案が望まれる。

そこで、本格的な単語辞書を用いず、文中に存在する鍵語に着目して字句に分解する簡便な方法を考案し、本方法の実現性について実験を試みたので、報告する。

2．鍵語に着目した字句分解方法

（1）考え方

本提案の方法は、文に含まれている特定の文字列からなる語句に着目した鍵語によって字句を分解する方法である。ここでの字句とは、単語に付属語、または他の単語が付属したものが、鍵語によって機械的に分解されたものとする。

論文など文章に書かれた日本語文は通常、格助詞、副詞、句読点などの特定の語句によって分けられている。

日本語文を字句に分解するとき、これらを鍵語として字句に句切ると、単語辞書を用いずに簡便なシステムで実現できる。

（2）実現上の問題点

本提案の字句分解法は、鍵語によって字句を分解するため、方式上に次の様な問題が生ずる。

①鍵語の部分衝突

複数の鍵語において、共通する部分文字列を持つことがある。これを部分衝突ということにする。

複数の鍵語を逐次に検索するため、部分衝突の鍵語が含まれていると、不都合が生ずる事がある。

例えば、「に」と「にて」のように、先頭の部分列に衝突を持つ鍵語の場合、先に検索対象となる鍵語が、鍵語として短い場合は、次に検索対象となる鍵語は認知されない。これを防ぐ検策方法を考案する必要がある。

②字句分け誤り

字句の部分列に鍵語が存在している場合は、この語は部分文字列で分けられる。これは、鍵語が認知されてしまうためである。

この誤りを完全に防ぐには、原理上からは単語辞書が必要であるが、この発生事例の少ない場合は、字句分解法として意義がある。この誤りを防ぐ方法を、処理系として実現する場合、この誤りについては、バックトラッキングが必要である。

③バックトラッキング

処理系としては、字句分け誤りの生じた字句を正当な字句として取扱うための仕掛けが必要である。

簡便に構成するため、字句分け誤りが起こる語句を簡単なファイルにまとめ、分解を行う前に、字句分解禁止処理を行った後、字句分解を行うのである。

また、この時の字句分解法は、訂正部分を分解しないような仕組みを持つ。

3．実験

（1）実験方法

本提案に基づく字句解析方法をもつシステムを試作して、提案の妥当性を確める。字句分解法の対象とする資料としては、きちんと書かれている文として、商業的技術雑誌[3]（以下資料1とする）の記事と学会論文誌[4]（以下資料2とする）を対象とする。

（2）鍵語

鍵語は、活用のない品詞を選定した。格助詞 12種、助詞5種、副詞5種、代名詞8種、連語2種、接続詞3種、句読点である。本試作で用いた鍵語は、37種である。

①鍵語の出現頻度と誤り率

字句分解における鍵語の出現頻度と、その誤り率を表1、2に示す。

表1：鍵語の出現頻度

	解析率	
鍵語	資料1	資料2
が	8.27%	7.45%
の	13.6%	17.6%
に	10.9%	11.9%
その他	67.2%	63.1%

表2：鍵語の誤り率

	誤り率	
鍵語	資料1	資料2
が	9.2%	19.9%
の	8.0%	17.8%
に	10.4%	3.7%
その他	72.4%	58.6%

②字句の長さ

字句分解により、資料毎の一文における句切りの数のグラフを図1、2に示す。

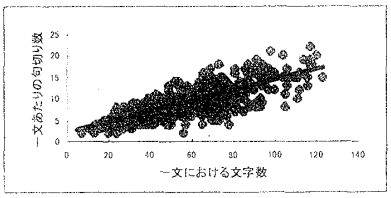

図1：一文の句切りの数（資料1）

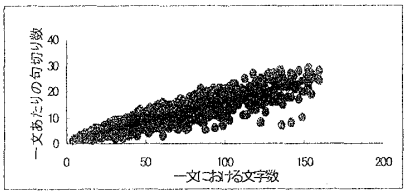

図2：一文の句切りの数（資料2）

特徴として、資料1は、字句長が9文字句であるが、切り数に幅がある。資料2は、句切り数が文の長さと比例し、字句長が7文字と一定である。

（3）字句の分解

字句分解によって得られた字句の正確さを字句解析率ということにする。これを（正当な字句数／全字句数）で現わす。計測の結果、次が得られた。

①格助詞のみの鍵語

格助詞、句読点の14種を鍵語とした時の字句解析率を（正当な字句数／全字句数）で示すときの字句解析率は資料1で92.3%、資料2で90.9%である。

②副詞、代名詞などの追加

格助詞に、副詞、代名詞等の37種を鍵語としたときの字句解析率は、資料1で96.9%、資料2で97.9%である。

4．評価

①鍵語を格助詞のみで構成

鍵語に格助詞だけを用いて字句分解を行っても9割以上の字句分けの正確さがある。

②副詞、接続詞、代名詞等を追加

副詞、接続詞などを鍵語に含めると、95%以上の正確さが得られる。

③文章の性質と解析率

解析率を出した2つの資料は、正当な漢字、仮名混じりの文である。ひらがな遣いの多い場合は、解析率が低下する。

④誤りの性質

鍵語を追加して生じた字句の誤りの性質としては、活用語句、固有名詞に鍵語を含んでいる字句である。動詞、助動詞などは語尾に活用があるため、鍵語に選定できない。また、固有名詞は、出現の予測できない語である。

5．結論

本論文で、鍵語に着目した字句分解法を提案し実験を行った。その結果、大規模な単語辞書を用いずとも、鍵語に着目するという簡便な方法により、日本語文を字句に分解することができ、本提案の有効性を把握することができた。

参考文献

1）佐藤匡正：流れ図文の性質－文記述の違いに着目した分析，情報処理学会論文誌 Vol.32、No.5、(1995) pp1260 - pp1270

2）池原悟：自然言語処理の基本問題と意味辞書の役割 池原研究室・年度報告,平成10年度,pp134-pp144

3）日経バイト：11月号、(1998)

4）診療録管理：日本診療録管理学会　刊　Vol.9、(1997)