8N-5

大規模複合語辞書による形態素解析

相川 勇之　　高山 泰博　　鈴木 克志

三菱電機株式会社　情報技術総合研究所

1．はじめに

インターネット上のWWW文書数が億ページ規模を越え、日々数百万ページの勢いで増加している中、検索ポータルサービスの重要性が高まっている。われわれは上記サービスの核となる検索エンジンにおいて、索引キーワードを検索対象文書から抽出するための日本語形態素解析ソフトウェアを開発している。

検索用のキーワード抽出に用いる形態素解析では、検索漏れを防ぐために複合語をできるだけ細かく単語分割した結果を出力することが望ましいが、日本語の複合語分割には曖昧性がある。特に検索漏れを防ぐために複合語相当の長い見出しを取り除き、短かい見出しだけからなる辞書を用いて解析する場合には、漢字１文字からなる接辞などで多くの曖昧性を生じて解析誤りが増加する。複合語の構造解析に関する先行研究は多いが、いずれも係り受け解析や意味解析を要するため、短時間で大量の文書を処理しなくてはならない大規模検索エンジンには適さない。われわれは大量のWWW文書から大規模な複合語情報を獲得することにより、速度性能を維持しつつ複合語の解析精度向上を狙う。本稿では、WWW文書をコーパスとして利用する方法を提案し、大規模複合語辞書獲得のための予備実験について述べる。

2．複合語処理の課題

全文検索の索引生成におけるキーワード抽出では、検索時に多用される自立語を正確に抽出することが要求される。特に、検索漏れを防ぐために複合語をできるだけ細かく正確に分割する必要がある。たとえば「三菱電機」を１形態素として索引キーワードを生成すると、「三菱」では検索できなくなってしまう。従って形態素解析結果としては「三菱/電機」のような単語区切りが望ましいが、日本語の複合語解析には曖昧性があることが良く知られている[横尾 97] 。

複合語解析処理については[宮崎 84]をはじめとして多くの研究がなされているが、大部分は複合語の構造解析に関する研究であり、係り受け解析や高度な意味解析を必要とするため、高速性が要求される検索エンジンのキーワード抽出処理には適していない。

一方、最近ではコーパスベースの形態素解析に関する研究が盛んである。統計情報を用いた接続確率の学習[浅原 00]や、コーパスそのものを用いた解析に関する研究[伊東 99]などがある。しかし、これらの先行研究で用いている既存のコーパスでは複合語の分割が十分ではないため、本研究の目的には適さない。たとえばEDR日本語コーパスでは、「日本国憲法」や「情報処理学会」が分割されていない。また、新聞記事データをもとにしたものが多く、WWW文書には多数出現する専門用語が少ないという問題もある。

3．WWW文書からの大規模複合語辞書獲得

3．1．既存コーパスとWWW文書の比較

われわれは複合語の解析精度向上のため、大規模な複合語辞書を構築することとした。複合語辞書構築にあたって、既存コーパスのかわりに大量のWWW文書を利用する。既存コーパスとWWW文書それぞれの利点および欠点を表１に示す。

表 1：既存コーパスとWWW文書の比較

既存コーパス（ＥＤＲ日本語コーパスなど）	
利点	(a)タグつきである。 (b)定型的、模範的な文が多い。
欠点	(c)正確なタグ付きデータは量が少ない。 (d)口語や新語が少ない。
WWW文書	
利点	(e)大量のデータがある。 (f)口語や新語が多い。
欠点	(g)一種の生コーパス（タグなし）である。 (h)信頼性の低いデータが混在。

A Japanese Morphological Analyszer with Large Compound Words Dictionary
Takeyuki AIKAWA, Yasuhiro TAKAYAMA, Katsushi SUZUKI
Mitsubishi Electric Corporation.
5-1-1 Ofuna, Kamakura, Kanagawa 247-8501, JAPAN

表１における WWW 文書の欠点(g)については、既存の形態素解析プログラムによる解析結果を用いることとした。解析結果には誤りも含まれるが、人手で修正しながらコーパスを拡充していくことにした。また、欠点(h)については、利点(a)により問題にならないと考えている。大量の WWW 文書中には誤植などの誤りデータも含まれていて当然だが、正しい単語のほうが桁違いに多いので頻度統計などを利用することにより機械的に誤りデータを排除することができる。

３．２．WWW文書からの複合語辞書獲得

図１に複合語辞書獲得手順を示す。

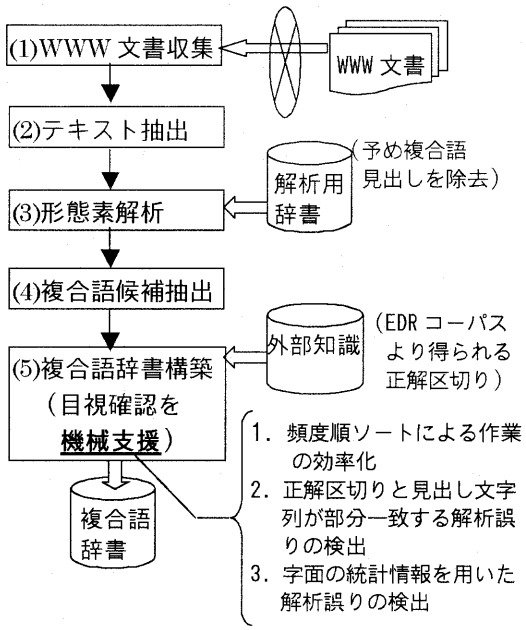

図１　複合語辞書獲得手順

（１）WWW 文書の収集

ロボットプログラムで WWW 文書を収集する。

（２）テキスト抽出

収集文書から日本語テキストを抽出する。このとき日本語コード変換も同時に行ない、日本語 EUC コードに統一する。

（３）形態素解析

コスト最小法を用いて形態素解析を行なう。解析用辞書から複合語見出しは削除しておく。

（４）複合語候補抽出

形態素解析結果の品詞情報を参照し、「接頭語」「名詞」「接尾語」からなる形態素列を抽出する。

（５）複合語辞書構築

形態素解析結果には誤りも含まれているので、抽出した複合語候補を人手で検査する必要がある。その際に、大量のデータに対する目視確認作業の負荷を軽減するために、出現頻度順にソートしたり、既存のコーパスから得られる結果と比較するといった機械的な支援を行なう。

たとえば、EDR 日本語コーパスを正解データとして利用することで作業を効率化することができる。同コーパスの単語区切りにはゆれがあり、「日本国憲法」が１語とされている一方で、「小/規模」「表現/力」などのように細かく区切られている例もある。EDR コーパス中の「地区/別」という正解区切りを参照することにより「県営/住宅/地/区別/一覧」という解析誤りを検出できる。

また、元の形態素解析結果とは独立に、字面に着目した統計情報を用いた単語分割を行ない、分割結果の差分をとることによって解析誤りを検出する方法も検討中である。

４．複合語辞書獲得の試行

まずは試行のために約 12 万ページを収集した。このうち約 8.7 万文書が日本語テキストを含んでおり、テキスト抽出の結果、約 890 万文、約 7700 万文字からなるデータを得た。上記データは EDR 日本語コーパス（約 21 万文、約 800 万文字）と比較すると、かなり大きなデータである。これを形態素解析し、のべ約 508 万語、異なり数で約 107 万語の複合語データを得た。現在、収集したデータをもとに効率よく複合語辞書を作成するための機械支援環境を構築中である。

５．まとめ

WWW 文書から大規模な複合語辞書を構築する手法を提案した。今後は、複合語辞書構築のための機械支援方法を確立した上で、解析精度および速度性能の評価をしていく。さらに、収集した WWW 文書をもとに、新語の獲得や、口語表現の収集なども検討していく予定である。

参考文献

[浅原 00]　浅原,松本「統計的日本語形態素解析に対する拡張ＨＭＭモデル」，情報処理学会研究報告 NL137-6，2000.
[伊東 99]　伊東「Suffix array を用いた日本語単語分割」，情報処理学会研究報告 NL131-7，1999.
[宮崎 84]　宮崎「係り受け解析を用いた複合語の自動分割法」，情報処理学会論文誌 Vol.25 No.6，1984.
[横尾 97]　横尾他「日本語形態素解析の誤りの回復について」，言語処理学会第３回年次大会，pp.429-432，1997.