形態素解析におけるデータベース解析方式の提案

035108　三枝　優一

（ 指導教員　速水　治夫　教授 ）

1. はじめに

近年，音声対話システムや一問一答システムなどにおいて，人が発する言葉をコンピュータが処理できるシステムへの期待が高まっている．また，個人による情報発信が一般化されていることから，企業などがそうした発信源から世間の反応や評判を得たいという要求が高まっている．

このような要求に応えるシステムを実現する為には自然言語処理の技術が重要である．本研究では，その第一段階である形態素解析システムについて採り上げた．

業界標準の形態素解析システムである茶筌において用いられている辞書は，データ構造が木構造(以下，木構造方式)で，高速化の為にバイナリ化されて構成されている．このことから，追加・更新といった処理の度に辞書の再コンパイルを行う必要があり煩雑である．

2. 本研究の目的

上記のような要求に応えるには，日常的な会話で用いられているような，必ずしも辞書に記載されていない話し言葉や方言，くだけた表現による未知語の頻出に対応する必要がある．

本研究では，追加・更新のような管理の容易さから，辞書のデータ構造をデータベースとし，データベースを用いた解析方式(以下, DB 方式)を提案する．

3. 比較システムの実装

提案方式を従来方式(茶筌)と比較するための比較システムを実装した．

比較システムは，DB 方式と茶筌を比較するため，DB 方式の解析部，辞書データを追加・更新・削除・SQL 文の発行などを実現した DB 管理機能部，また DB 方式と同一文書を比較対象とする為の茶筌起動部の三構成からなる．

4. 評価

評価では，DB 方式と茶筌について，解析速度と解析精度の比較を行った．

解析速度の比較では，同時に音声対話システムへの適応を考慮して，読み上げ速度との比較を行った．実験結果を図 1 に示す．

解析精度の比較では，それぞれの解析方式における誤解析率を算出した．実験結果を図 2 に示す．

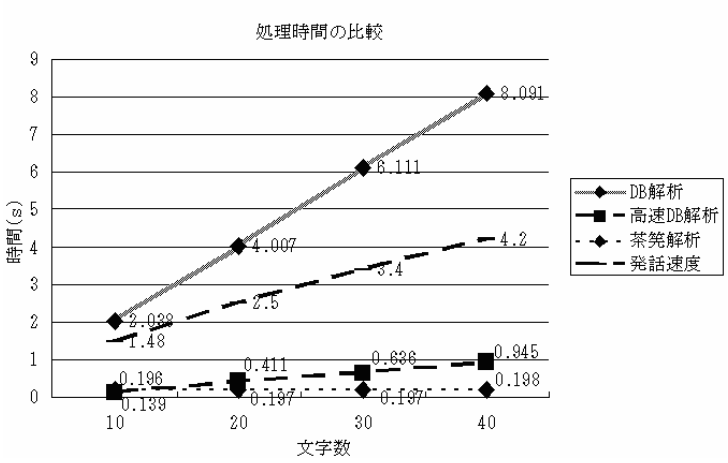

図 1. 処理時間の遷移

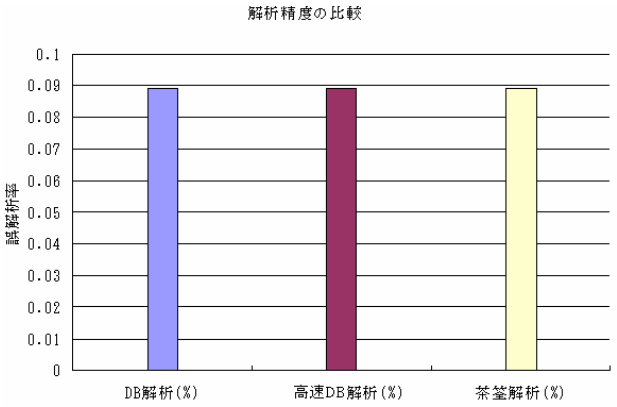

図 2. 解析精度の比較

5. まとめ

木構造方式と，DB 方式の比較実験の結果，DB 方式は茶筌に近い処理効率を実現できた．また，音声対話システムを想定した場合，処理効率が発話速度内におさまる必要があるが，茶筌と DB 方式は双方とも充分余裕をもった効率で解析可能である．また，精度では顕著な差はなかった．

このことから，DB 方式は，辞書データの頻繁な追加・更新を伴う場合，充分有用である．

今後の課題として，DB 方式で用いた解析方式を改善し精度の向上を図ること，未知語を自動検出する機能を追加することがあげられる．