分かち書きの効率化について

教育方法開発センター　河原正治
情報処理学科　夏目　武

要　旨：点訳のための分かち書きを効率化することを目的として，EDR電子化辞書の可能性について検討した。EDR電子化辞書は，次世代の自然言語処理，知識処理システムの研究と応用開発に広く利用されると期待されている。このような，大規模かつ意味レベルでの解析までを考慮したデータベースは，より完全な分かち書きシステムの構築には必要不可欠の要素となるだろう。本稿では，簡単な評価プログラムを使うことにより，実用化のために解決しなければならない問題点を指摘した。EDR辞書を利用した研究開発の成果も発表され始めており，今後は，それらとの連携も検討する必要がある。
キーワード：自動点訳システム，自然言語処理，形態素解析，EDR電子化辞書

1. はじめに

近年，パーソナルコンピュータで動作する日本語自動点訳ソフトウェアが，いくつか開発され広く利用されるようになった。ここで「日本語自動点訳」とは，かな漢字混じりのべた書き文を6ビット記号体系の触知文字(点字)に変換することを意味している。この処理の最も困難な部分は，かな漢字混じり文を点字のルール[1]に 基づき，かなの分かち書きにすることにある。このような機能を持つプログラムの精度を比較した報告によると，以下のような改善すべき点が指摘 されており[2]，より効率的な分かち書きシステムの開発が待たれている。

・原文１万字あたり300～600個の誤変換がある。
・特に，ひらがなの並びの解析に不備がみられる。
・より精度を向上させるためには文意に踏み込んだ解析が必要である。

本論文では，分かち書き効率化の試みにおけるEDR電子化辞書の可能性について考察する。

2. EDR電子化辞書の概要

本節では，EDR電子化辞書仕様説明書[3]をもとに，その概要を説明する。

EDR電子化辞書は，基盤技術研究促進センターとコンピュータメーカ８社との共同出資のもとに，９年間のプロジェクト(1986年度～1994年度)により得られた成果であり，単語辞書，対訳辞書，概念辞書，共起辞書，専門用語辞書とEDRコーパスから構成されている。

単語辞書は，25万語の語彙を持つ日本語単語辞書と19万語の語彙を持つ英語単語辞書に分けられる。

概念辞書は，単語辞書に語義として導入された40万の概念についての知識が記述され，概念体系辞書と概念記述辞書に分けられる。概念体系辞書は40万の概念に対して，それらの間の上位下位関係を記述したものである。

共起辞書は，90万句の日本語共起辞書と46万句の英語共起辞書から構成され，言葉の言い回しに関する情報を２項関係で記述したものである。

EDRコーパスは，22万センテンスの日本語コーパスと16万センテンスの英語コーパスから構成されおり，大量の用例を収集し，意味レベルまで解析して得られる言語データである。

このようにEDR電子化辞書は，大規模かつ意味レベルでの解析までを考慮したデータベースであり，次世代の自然言語処理，知識処理システムの研究と応用開発に広く利用されると期待されている。

3. 評価実験

ここでは，EDR電子化辞書の分かち書きへの応用の可能性を評価するための簡単な実験について報告し，実用化に対する問題点を指摘する。

3.1 形態素解析器について

現在，容易に利用可能な形態素解析器としては，京都大学および奈良先端科学技術大学院大学において開発されたJUMAN[4]と奈良先端科学技術大学院大学において開発された茶筌[5]がある。これらは，フリーソフトウェアとして公開されている。

茶筌(ChaSen)は，JUMAN version 2.0をベースに改良されたものであり，コンパイル後の辞書サイズが4分の1(8MB 程度)になり，また解析速度も8～10倍になっており十分実用的である。茶筌の辞書は約11万5千語である。

また，JUMANの最新バージョンはversion 3.11であ

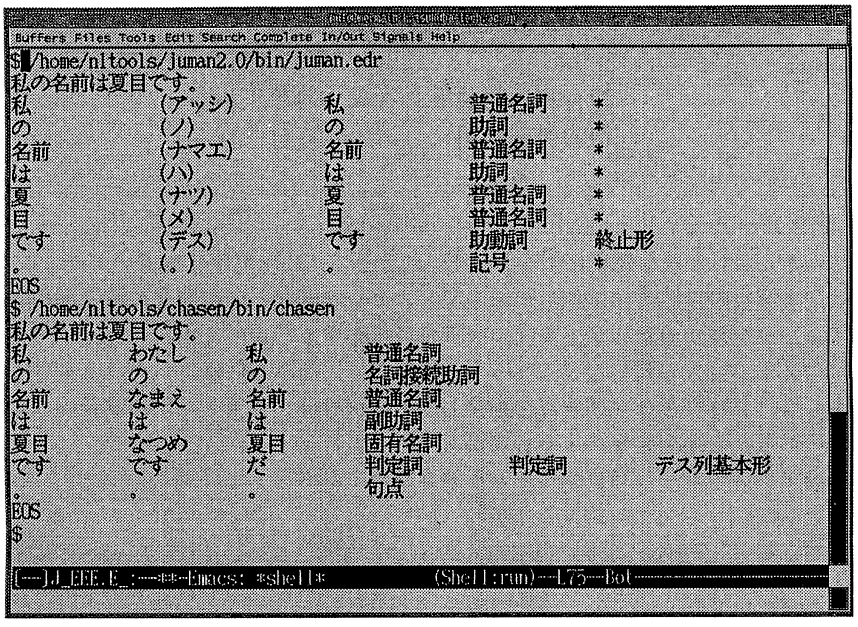

図１：形態素解析プログラムの実行例

り，種々の改良が続けられている。

3.2 実用化のための課題

JUMAN version 2.0をEDR電子化辞書の日本語基本単語辞書を検索するように拡張し簡単な実験を行った。その結果，つぎのような問題点が判明した。

(1)多数の複合語が登録されていること。

例えば，電子工業連盟［デンシコウギョウレンメイ］，女子栄養大学［ジョシエイヨウダイガク］などが登録されており，直接，点字の分かち書きを生成することができない。辞書の読みデータを修正することで回避できる。

(2)多数の読みが登録されていること。

例えば，「私」という単語については，私［アッシ］，私[ワタシ]などが登録されている。頻度情報などを参照することによって回避できると思われる。

(3)人名が登録されていないこと。

フリーソフトウェアとして公開されているかな漢字変換システムに付属する人名辞書などを利用することができるだろう。

(4)辞書の容量が大きいこと。

日本語単語基本辞書だけで90MB程度ある。不要なエントリを削除することによって，辞書サイズを縮小することを検討するべきである。

図1にプログラムの実行画面を示す。この実験は，基本単語辞書の品詞情報と接続情報のみを使っており，EDR電子化辞書の能力を引き出すためには，概念辞書，共起辞書を参照するような解析プログラムの開発が必要である。

4. 終わりに

3.で述べたように，EDR電子化辞書を利用した分かち書きシステムを実用化するには，いくつかのハードルがあることがわかった。EDR辞書は，大規模かつ意味レベルでの解析までを考慮したデータベースであり，より完全な分かち書きシステムの構築には必要不可欠な要素となるだろう。EDR辞書を利用した研究開発の成果も発表され始めており，それらとの連携も考慮しつつ研究を発展させていきたい。

参考文献

［１］日本点字委員会：“日本点字表記法”（1990）

［２］福井哲也：“日本語自動点訳ソフト４種の精度の比較”，第２回視覚障害リハビリテーション研究発表大会論文集，pp.114-117（1993）

［３］“EDR電子化辞書仕様説明書”日本電子化辞書研究所（1993）

［４］松本裕治他：“日本語形態素解析システムJUMAN Version 3.11使用説明書”，京都大学工学部 長尾研究室，奈良先端科学技術大学院大学 松本研究室(1996)

［５］松本裕治他：“日本語形態素解析システム『茶筌』varsion 1.0b5使用説明書”，奈良先端科学技術大学院大学 松本研究室(1996)

［６］研究代表者田中穂積：“EDR辞書を用いた日本語解析 ツールに 関する研究”，http://www.icot.or.jp/AITEC/ACTIVITY/itaku95/T18/main-s.html（1996）