辞書データ主導型の自動点字翻訳システム

3N－8

横平貫志　早川哲史　兵藤安昭　池田尚志
岐阜大学工学部

1　はじめに

通常，点字は6ビットの点で文字を表すため，日本語の場合，すべてをひらがな表記し，点字に翻字する必要がある．その際，点字の規則に従った文節単位に分かち書きをしなければならない．この作業の多くは，点訳ボランティアによって手作業で行われている．これを自動的に行う自動点訳ソフトも市販されているが，精度が十分でないため，まだ広くは使われていない．それは，点訳ボランティアが，全体を見直して，間違った箇所を見つけ出し修正するという校正作業に労力がかかりすぎるからである．

我々は，現在開発している日本語解析システムIBUKIによる文節解析を応用し，点訳ボランティア支援のための自動点訳システムの開発を試みている[1]．目標は，しっかりした言語解析に裏付けられた精度の高い分かち書きが行えることと，校正作業がスムーズに行えるシステムである．

本報告では，辞書データ主導による点訳手法の概要と，分かち書きや読みの校正作業を行うための点訳後編集について述べる．

2　システムの概要

本システムの構成を図1に示す．入力した日本語文は，IBUKIによる文節解析によって，文節単位の切り出しを行う．漢字連続文字は，複合語解析によって名詞，接辞等に分割する．文節解析の結果を基に，点訳処理では，点訳規則を記述したRDB上の辞書情報を参照しながら，点字の規則に従った分かち書き，点字の表記法に従ったひらがな表記への変換を行う．ユーザは，点訳後編集インターフェースにおいて間違った分かち書きや読みを修正し，最後に点字ディスプレイ，点字プリンタ等に出力する．

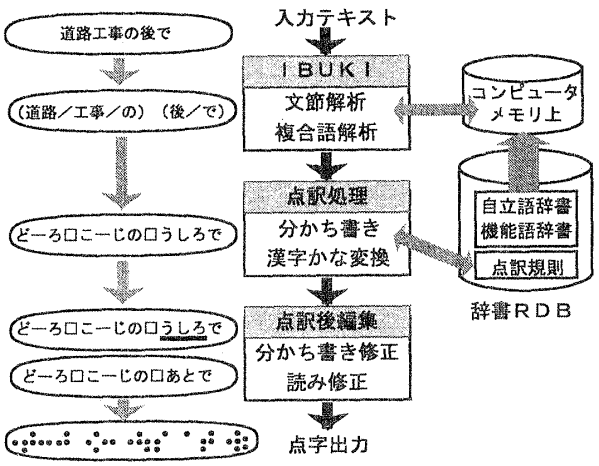

図 1: システムの構成図

3　辞書RDBと点訳手順

3.1　辞書RDB

点字の規則[2]では，表記法については「ウ列の長音は長音符を用いる」といった規則，複合語については「3拍以上の意味のまとまりが2つ以上あれば，その境目で区切るが，連濁を生じる場合は続ける」などの規則がある．例えば「栄養満点」は「えいよー口まんてん」（口：区切り），「建設会社」は「けんせつがいしゃ」となる．しかし，個々の単語に依存した規則や，例外的事項も多く，特に複合語に関しては，拍数とは別に，自立性の強さなどによって区切り方が違ってくる．例えば，「都市国家」の「都市」は2拍であるが，「とし口こっか」となる．このような例外的な規則にも対応できるように，我々のシステムは，辞書データ主導型の方式を取っている．これらの点訳規則をIBUKIが使用するRDB上の各単語に対して，個々に記述することで木目細かに対応できる．

RDB上の辞書情報を図2に示す．辞書は大きく分けて自立語辞書テーブルと機能語辞書テーブルに分けられる．点訳規則には，各単語ごとに点字の表記法に従ったひらがな表記で登録し，濁音には，単語の濁音の有無を登録した．また，読みコストには，4節で述べる読みの優先順位を表すコストを登録した．

Japanese Braille transcription system
based on dictionary driven method
Kanji Yokohira, Yasuaki Hyodo, Takashi Ikeda
Faculty of Engineering, Gifu University
Gifu-shi,501-11,Japan

自立語辞書テーブル

ＩＤ	表記	左連接	右連接	点訳規則	連濁	読みコスト
１００	急に	JLD2	JRD2	きゅーに		０
１０１	会社	JLN1	JRN1	かいしゃ	○	０
１０２	今日	JLN7	JRN9	きょー		１
１０２	今日	JLN7	JRN9	こんにち		２
１０２	今日	JLN7	JRN9	こんじつ		２
１０２	今日	JLN7	JRN9	こんち		３
１０３	都市国家	JLN1	JRN1	とし□こっか		０
１０４	生年月日	JLN1	JRN1	せいねんがっぴ		０
・・・	・・・	・・・	・・・	・・・	・・・	・・・

機能語辞書テーブル

ＩＤ	表記	左連接	右連接	点訳規則
１０００	にもかかわらず	D	ZY	にも□かかわらず
１００１	にも関わらず	D	ZY	にも□かかわらず
１００２	ておかねばならぬ	ABT	ZZ	て□おかねば□ならぬ
１００３	たろう	DDA	ZZ	だろー
１００４	でさえ	D	ZY	でさえ
・・・	・・・	・・・	・・・	・・・

図 2: 辞書 RDB

3.2 点訳手順

IBUKI による文節解析の結果得られた文節の区切りを点字の基本的な分かち書きの区切りとする．各単語には，RDB 上の単語へ対する ID が付けられていて，その ID から RDB 上の点訳規則を検索することで点訳を行う．

辞書に登録されていない複合語については，IBUKI による複合語解析の結果を基に，RDB 上から得た連濁の有無や単語の拍数などをを考慮して分かち書きの箇所をプログラム処理で判断する．

表 1 には，分かち書きの精度を，朝日新聞社説 300 文（13730 文字）を用いて評価した結果を示す．

表 1: 分かち書き正解率

	切り過ぎ	切り忘れ
複合語除く	98.8%	99.5%
複合語含む	97.4%	98.6%

$$\text{正解率} = (1 - \frac{\text{切り過ぎ (or 切り忘れ)}}{\text{正解の分かち書き数}}) \times 100[\%]$$

4 点訳後編集

自立語辞書は，EDR 単語辞書 [3] を基に作成している．その中に登録されている漢字の読みには，通常では使わない読みも含まれているため，次のようなコストを付けることで読み候補の絞り込みを行った．

コスト０ 複数の読みを持たない単語
コスト１ 複数ある場合に通常選択する読み
コスト２ その他に使う可能性がある読み候補
コスト３ 通常では使わない読み

複数の読み候補がある漢字は，コスト１の読みを優先的に選択する．その他にコスト２の読み候補がある場合は，点訳後編集インターフェース（図 3）において曖昧な読みであるとして表示色を変えて表示した．その箇所でマウスをクリックすると，コスト２の読みが全て表示され，該当する読みが選択できる．該当する読み候補がない場合は，直接任意の読みを書き込んで編集を行う．また，文節解析結果に未登録語が含まれる場合や，５文字以上の長い複合語が出現した場合を，分かち書きが曖昧であるとして表示色を変えて表示した．その箇所でマウスをクリックすることで容易に区切りの追加や削除ができる．

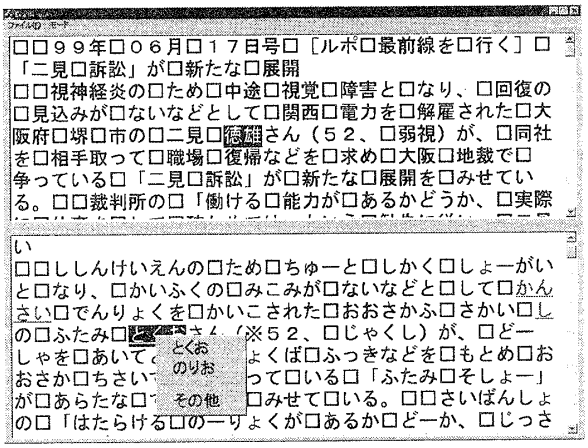

図 3: 点訳後編集インターフェース

5 おわりに

日本語解析システム IBUKI を応用して自動点訳システムを開発した．分かち書き精度は現在のところ，概ね 98%程度である．本システムは，辞書データ主導型であり，分野別に辞書データを整備することで，より高精度の正解率が期待できる．

今後は，意味解析などを行い，読みの選択の正解率を向上させ，分かち書きや読みが曖昧である箇所をより明確にし，ユーザによる校正作業が容易に行えるシステムの開発を目指したいと考えている．

参考文献

[1] 兵藤, 横平, 池田：長単位文節解析を利用した点字分かち書きシステム, 電子情報通信学会技術研究報告 NLC99-4,Vol.99,No.88,(1999)

[2] 全国視覚障害者情報提供施設協議会：点訳のてびき第２版, (1991)

[3] 株式会社日本電子化辞書研究所：EDR 電子化辞書仕様説明書, (1995)