

日本語点字翻訳のための漢字仮名変換

1W-2

鈴木和洋 浅川智恵子

日本アイ・ビー・エム(株) 東京基礎研究所

1. はじめに

視覚障害者にとって、点訳された図書や文書は非常に重要な情報源である。しかし、点字の表記法を完全にマスターすることは、非常に難しく時間を要するため、そうした点訳ボランティアは少なく、点訳される図書は十分とはいえない。

そのような点訳者の不足を補うため、著者らは、これまで点字翻訳の自動化について研究開発を行ってきたが、今回漢字仮名交じり文書を点字にするまでの一連の処理を実現した。ここでは、その中で特に、入力文章を、分かち書きされた仮名の並びに変換する漢字仮名変換部について報告する。

2. システム構成

点字翻訳システムの全体の構成を図1に示す。このシステムは、入力の漢字仮名交じりの日本語テキストを分かち書きされた仮名テキストに変換する漢字仮名変換部と仮名及び英語のテキストを点字に変換する点字変換部[1]の2つの処理から構成される。なお、システムで用いた点字翻訳の規則は[2]に準拠した。

3. 漢字仮名変換部

点字文書は、日本語文書をただ単に読みに置き換えたものではなく、単語境界や文節境界においては、分かち書きがなされている。従って、漢字仮名文字列を点字のための仮名の並びに変換するには、読みへの置き換えとともに、分かち書きのための解析処理、情報抽出が必要となってくる。

今回作成した漢字仮名変換部は、テキスト音声合成の言語処理部[3]をもとに構築した。テキスト音声合成とは、任意の日本語文書を音声に変換するシステムであるが、この言語処理部を点字翻訳に利用する利点として、

1) テキスト音声合成、点字翻訳のどちらも基本的に仮名体系を生成するものであるため、テキスト音声合成で正しい読みを得るための情報(連濁変化、助数詞変化、アクセント情報など)を用いることができる。

Kanji/Kana Conversion for Text-to-Braille System
Kazuhiro Suzuki, Chieko Asakawa
Tokyo Research Laboratory, IBM Japan Ltd.

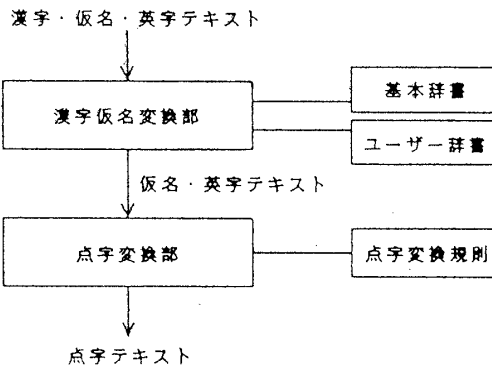


図1. 日本語点字翻訳システムの構成

2) テキスト音声合成では、ポーズの位置やイントネーションを決定するために、文節の係り受け解析などの文法解析を行う必要があり、その情報も点訳の分かち書きの情報に応用できる。

一などがあげられる。

漢字仮名変換部では、単語の読み仮名、品詞や文節境界及び簡単な係り受け関係を得るためのテキスト解析と得られた情報をもとにした分かち書き処理の2段階の処理を行っている。以下、この具体的な処理について説明する。

3.1 テキスト解析

3.1.1 辞書

テキスト解析処理ではテキストを仮名に変換すると同時に、分かち書きに必要な単語の品詞、文節境界及び係り受けの情報を抽出する。解析のための辞書として、約77000語の基本単語辞書及び5000語まで登録できるユーザー辞書、及び121語の拡張付属語辞書を用いている。基本単語辞書には、単語の読み仮名のほか、品詞や分かち書きのためのフラグが設定されている。このフラグは、品詞情報だけでは判定できない単語固有の分かち書きについて対処するためのもので、テキスト音声合成の言語処理に用いている辞書の情報から抽出した。このフラグとして、自立語については1種類、接辞については4種類のフラグを設定した。

[一般語のフラグ]

- ・前接非分かち書きフラグ(連濁語)
(株式会社→かぶしきがいしゃ)

[接辞のフラグ]

- ・後端分かち書きフラグ (独立アクセント接頭)
(反社会的→はん□しゃかいてき)
- ・前端分かち書きフラグ (固有名詞につく「さん」等)
(鈴木さん→すずき□さん)
- ・前端非分かち書きフラグ (助数詞接尾等)
(一人当り→ひとりあたり)
- ・後端非分かち書きフラグ (接頭)
(うすら寒い→うすらさむい)

拡張付属語辞書については、品詞や活用に応じた付属語の接続条件を、分かち書きの観点から、さらに細かく分類し、2種類のフラグを設定した。

- ・前端非分かち書きフラグ
 - 一 接続によって分かち書きしない付属語
用言連用形+ない (起きない→おきない)
サ変名詞+する (維持する→いじする)
- ・前端分かち書きフラグ
 - 一 形式名詞, 補助動詞 (読むことを忘れてしまう→
よむ□ことを□わすれて□しまう)
 - 一 接続によって分かち書きする付属語
形容詞連用形+ない (良くない→よく□ない)
普通名詞+する (お茶する→おちゃ□する)

また、規則で対応できない複合語などについては、すでに分かち書きされた形で辞書にしておくことにより対処している。

3.1.2 文節確定

上記の辞書を用いて得られた単語列より文節候補を構成する。この文節候補の中から、文節長、付属語の数、語の接続情報をもとに、適切なものを選択する。次に、求められた文節列に対して、隣接する文節間の係り受け関係を調べる。この係り受け関係は、サ変名詞や複合動詞の分かち書きの判定に用いている。

3.2 分かち書き規則

テキスト解析で得られた読み仮名、単語の品詞、文節境界、係り受けの情報をもとに、次の1)~5)に示す分かち書き規則によって、分かち書きを行う。

- 1) 文節境界で分かち書きする。
- 2) 単語境界での分かち書きは、フラグがある場合は、そのフラグの指定に従う。フラグが指定されていない場合、構成される複合語が3拍以上であれば、その後で分かち書きする。
- 3) 「サ変名詞+する」は、前後の語の並びまたは文節の係り受けを考慮して分かち書きを指定する。

- ・副詞+サ変名詞 非分かち書き
(突然(副詞)停止する→とつぜん□ていしする)

表1. 評価テストの結果 (学術論文の解析)

総文数	153文
総文字数	8660字
解析誤り	53個

表2. 解析誤りの内訳

内訳	個数
辞書の未登録語による誤り	27
英数字記号の扱いの誤り	12
文脈に起因する誤り	8
その他	6

- ・複合名詞+する 分かち書き
(非常停止する→ひじょー□ていし□する)
 - ・連体修飾関係 分かち書き
(ハードな練習する→ハードな□れんしゅー□する)
- 4) 括弧類は、前端につく括弧("(", "[")と後端につく括弧(")", "]")などを判定し、分かち書きする。
(彼は「今日いくの」と言った→
かれわ□「きょー□いくの」と□いった)
- 5) 用言連用形が用言に係るといった係り受け関係の場合は、複合動詞であるから分かち書きしない。
(歩き続ける→あるきつづける)

4. 評価

変換の性能を調べるために、実際の文章の解析を行った。文章として学術論文を使用した。表1に変換結果を、表2に誤りの内訳を示す。解析誤りの多くは単語が辞書に含まれていないことによるもので、今後変換精度を向上するためには、辞書の整備や英数字記号への対応が課題である。

5. おわりに

従来の仮名及び2級英語の点字翻訳システムに漢字仮名変換部を追加し、日本語点字翻訳システムを試作した。その中で漢字仮名変換部について、その概要と評価について述べた。今後は、様々な文書に対してテストを行い、精度の向上を図るとともに、実際の翻訳作業における使い易さの検討を行う予定である。

なお、システムの評価実験に協力していただいた点訳ボランティアの方に深く感謝いたします。

文献

- [1] 平山：情報全大，4T-8, pp.1683-1684, (1986)
- [2] 日本盲人社会福祉施設協議会 点字図書部会：点訳の手引き(第2版)，(1991)。
- [3] 大河内他：音響講論，1-P-12, (1992.3)