知識ベースに基づく点字翻訳のための日本語文書分かち書き手法

4B－10

平岡 大樹+　水野 一徳+　小野 智司+　鈴木 恵美子‡　狩野 均+　西原 清一+
+筑波大学　電子情報工学系　‡東京家政学院筑波女子大学　短期大学部

1．はじめに

近年、視覚障害者のために、コンピュータのマニュアルや勉強に必要な教科書等を点字に翻訳（点訳）するシステムの開発が望まれている。しかし従来の点訳システムでは、最長一致法を基本とする形態素解析を用いて単語を認定しており、辞書の整備に非常に多くの時間や労力がかかる[1][2]。本稿では、従来のような大規模な辞書や文法ルールを用いずに簡便な表層解析のみを行うことによって点訳のための分かち書きを行う手法について提案する。

2．研究分野の概要

2.1　点訳のための分かち書き

一般の漢字かな混じり文は漢字とかなの使い分けで文を構成しているのに対し、点字は表音文字体系であり、分かち書きを行わないと、読みにくいばかりでなく正確に意味も伝わらない。よって、日本語文書を点訳する際には分かち書きが必要とされるが、ルールの複雑さや日本語特有の曖昧性により計算機で正確な分かち書きを行うのは、非常に困難なものとなっている。

2.2　従来手法の問題点と対策

日本語文書の分かち書きは従来、形態素解析によって行われてきた[3]。この手法の問題点として、必要とされる辞書の構築に多大な時間と労力を要する上、辞書引きにも時間がかかってしまうという点がある。また、改良方法はユーザ辞書の追加のみに限られるため正解率の向上が難しいという点もある[5]。

これに対して本手法は、次の基本戦略により、これらの問題点を解決し、従来方法より正解率の高い分かち書きを行うというものである。

(1) 文法情報を含む大規模な辞書の代わりに単語のみの小規模なテーブルを用いる。

(2) 知識の追加、削除、更新を容易とさせるため、各文字間に区切りを入れるかどうかという知識を知識ベース化する。

(3) 各ルールの曖昧性に対処するため、各ルールに優先点数を導入する。

Sentence Segmentation Algorithms to Translate into Braille using Knowledge Base
Taiki Hiraoka+, Kazunori Mizuno+, Satoshi Ono+,
Emiko Suzuki ‡, Hitoshi Kanoh+, Seiichi Nishihara+
+Institute of Information Sciences and Electronics, University of Tsukuba
‡Tokyo Kasei Gakuin Tsukuba Junior College

3．提案する手法

3.1　知識ベースの構築

分かち書きに必要な知識の例を表1の(1)に示す。本研究では13個の指標（表1(2)）を用いて、これらの知識から36個のルールを構築した（表1(3)）。

表１：知識ベースの例

(1)知識の例
知識１　促音の前では切らない
知識２　助詞の後ろでは切る
知識３　ひらがな書きの自立語の内部では切らない

(2)指標の例
・前の文字は（促音, 拗音, …）
・前の文字は助詞の（１文字目, ２文字目, 最後, …）
・区切りは（切る, 切らない）

(3)知識ベースの例
ルール（1，（優先点数６）
　[if，　[前の文字は , [促音]]]
　[then , [区切りは , [切る]]])
ルール（2，（優先点数２）
　[if，　[前の文字は助詞の , [最後]]]
　[then , [区切りは , [切る]]])

3.2　テーブルの作成

本手法では、表２の１列目に示した７項目からなる小規模なテーブルを作成し、これと字種のみの表層解析を行うことで、大規模な辞書を引く手間を省き、時間の短縮を図った。

表２：各テーブルの概要

テーブル	書式	例	語数
ひらがな書きの自立語	単語,区切り方	しかし, 前後 はっきり, 前	250
助詞	単語	と は	25
混ぜ書き語	単語	引き算 繰り返	11000
漢字２字熟語	単語	圧力 計算	62500
漢字３字熟語	単語	亜熱帯 委員長	24000
接頭語	単語	不 副	20
接尾語	単語	秒 機	60

3.3　アルゴリズム

図１の手順で分かち書きを行う。

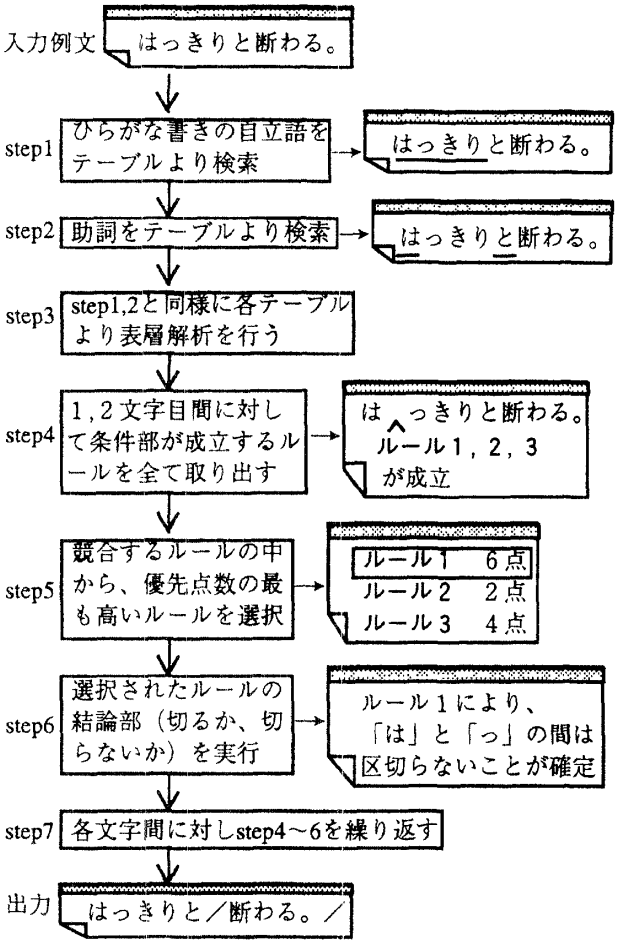

図1：本手法の分かち書きを行う手順

4. 評価実験

4.1 評価方法

情報処理の標準的なテキスト[4]を用いて、正解率を計算した。このテキストは文数が3463文、一文の平均文字数が32.8文字である。

正解率は次式によって計算した。なお、空振りは区切る必要のない部分を区切ってしまった間違い、見逃しは区切らなくてはならない部分を区切らなかった間違いを示す。ただし、以下の4.2節の評価にはテキストの全章(1〜4章)、4.3節では1〜2章までを用いた。

$$\text{正解率(空振り無し)}=\left(1-\frac{\text{空振りの回数}}{\text{正解の区切り数}}\right)\times 100$$

$$\text{正解率(見逃し無し)}=\left(1-\frac{\text{見逃しの回数}}{\text{正解の区切り数}}\right)\times 100$$

4.2 本手法の正解率

図2に本手法の句あたりの文字数と正解率の関係を示す。これより、本手法は句の長さが長くなっても、正解率は低下しないといえる。

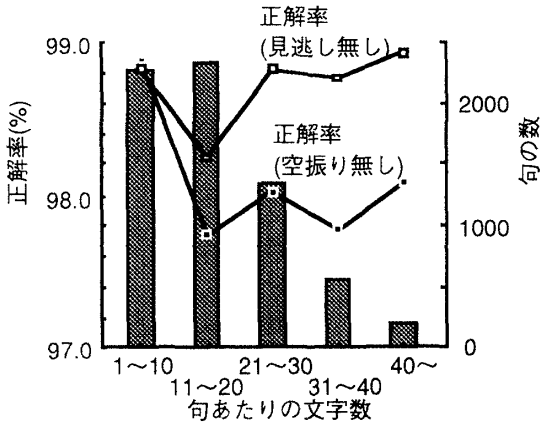

図2：句あたりの文字数と正解率の関係

4.3 他のシステムとの比較

本手法と他のシステム(EXTRA Ver3.0)との正解率の比較結果を表3に示す。これより、本手法の方が優れており、有効な手法であるといえる。

表3：本手法とEXTRAとの正解率の比較結果

	本手法	EXTRA
正解率(空振り無し)	98.50%	96.08%
正解率(見逃し無し)	99.05%	96.00%

5. おわりに

本稿では、従来より小規模なテーブルを使った表層解析のみで点訳のための分かち書きを行う手法について提案した。また、分かち書きの知識を知識ベース化し、ルールに優先度をつけることで競合を解消し、文の長さに依存しない高い正解率を得ることができた。今後の課題として、更に正解率を上げるためにテーブルやルールを改良すること、また分かち書きされた文を漢字かな変換して、最終的に点字の出力を得ることがあげられる。

謝辞

本研究に御協力頂いた筑波技術短期大学視覚部 長岡英司先生、辰巳公子氏、筑波女子大学の鈴木ゼミの皆様、またデータを提供して頂いた筑波大学 宇都宮公訓先生に深く感謝致します。

参考文献

[1]河原正治：日本語自動点訳ソフトウェアの開発について，信学技報 HC94-49(1994).
[2]吉田将：辞書構築における諸問題，情報処理，Vol.27, No.8, pp.933-939 (1986).
[3]長尾真：日本語情報処理，電子通信学会編，コロナ社，(1984).
[4]宇都宮公訓：コンピュータ入門，共立出版，(1990).
[5]小野智司ほか：知識ベースに基づく対話型点字翻訳システム，第54回 全国大会 講演論文集 4B-09 (1997).