

知識ベースに基づく対話型点字翻訳システム

4 B-9

小野 智司[†] 西森 雄一[†] 平岡 大樹[†] 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

情報処理分野への視覚障害者の職域を拡大するためには、コンピュータのマニュアルや教科書等を点字に翻訳(点訳)する必要がある。現状ではボランティアが一冊ずつ手作業で行っており効率が悪い。また、製品化された点訳システムもあるが、誤訳された箇所についての見直しの手間がかかることが指摘されている [1]。

本稿では、ユーザとシステムが協調して精度の高い点訳を行う方式を提案する。このため、分かち書きを自動化するための知識を知識ベース化し、区切りが一意に決定できない場合には、ユーザに問い合わせることにした。以下では、本システムの処理手順、構成と機能ならびに分かち書きの実用例について述べる。

2 点字翻訳の概要と問題点

2.1 点字翻訳の概要

欧米諸言語のように分かち書きされ、単語と単語の間に空白のある文章と異なり、日本語の文章は字種も豊富である上に単語と単語が分かれていない [2]。このため、点訳するには漢字をかなに変換するとともに、分かち書きを行い単語を認識する必要がある。分かち書きの知識は表1のようなものがあげられる。本稿では点訳の問題の中でも特にこの分かち書きの問題について考察する。

表 1: 分かち書きの知識の例

知識 1	自立語は前を区切って書く
知識 2	付属語は自立語または他の付属語に続けて書く
知識 3 (1)	形容詞の「ない」は前を区切って書く
知識 3 (2)	助動詞の「ない」は前に続けて書く

2.2 現状技術の問題点と本システムの特徴

従来の点訳システムには以下のような特徴と問題点がある。

- (1) 辞書引きを基本とする形態素解析を行っており、文法情報が必要となるため辞書が膨大である。
- (2) 一括処理を行っているため、誤訳された箇所を修正するために全文を見直す必要がある。
- (3) 正解率の改良方法がユーザ辞書への単語登録に限られている。

Interactive System for Translating Japanese into Braille using Knowledge Base

Satoshi Ono[†], Yuuichi Nishimori[†], Taiki Hiraoka[†], Emiko Suzuki[‡], Hitoshi Kanoh[†], Seiichi Nishihara[†]

[†]Institute of Information Sciences and Electronics, University of Tsukuba

[‡]Tokyo Kasei Gakuin Tsukuba Junior College

表 2: 表層情報に基づく分かち書きの知識の例

知識 A-1	句読点の後ろで切る
知識 A-2	ひらがな→漢字間で切る
知識 A-3	助詞の後ろを切る

表 3: 曖昧な区切り箇所を特定するための知識の例

知識 B-1	サ変動詞は複合語になると曖昧である
知識 B-2	「ない」は助動詞である場合と形容詞である場合があり曖昧である
知識 B-3	漢字テーブルに登録されていない漢字連続の区切りは誤っている可能性がある

これに対し本システムは以下のような特徴を持つ。

- ① 表層解析を基本とする知識を用いるので、辞書を大幅に縮小することができる。
- ② システムによる分かち書き(自動分割)の結果、誤りを含んでいると考えられる箇所をユーザに問い合わせることにより、ユーザが修正を容易に行うことができる。
- ③ 分かち書きの知識を知識ベース化することにより、知識を追加・削除・更新することができる。

3 システムの開発

3.1 設計方針

本システムの設計方針を以下に述べる。

- a) 文法情報の付加されていない見出し語のみからなる助詞、自立語および漢字などの7種類のテーブルを用いる。
- b) 分かち書きの処理を自動分割と対話処理の2段階に分ける。
- c) 分かち書きの知識(表1)を表層解析に基づく知識(表2)に書き換え、知識ベース化する(以下、知識ベース A と称す)。
- d) 分かち書きが疑わしい箇所をユーザに示すための知識(表3)を知識ベース化する(以下、知識ベース B と称す)。

3.2 知識ベース B の作成

知識ベース B における知識は次の2種類の方法で導出した。

- (1) 表層情報のみでは曖昧な分かち書きの知識から導く。例えば、表1の知識3(1),(2)は表3の知識B-2のように書き換える。

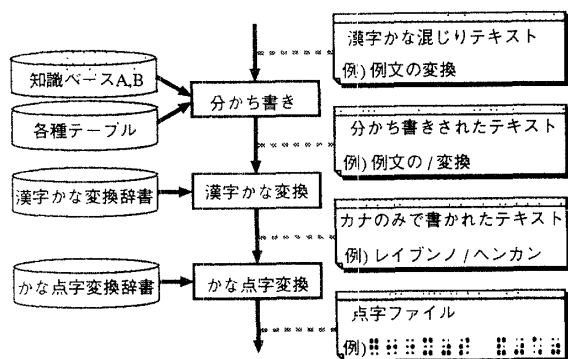


図 1: データと処理の流れ

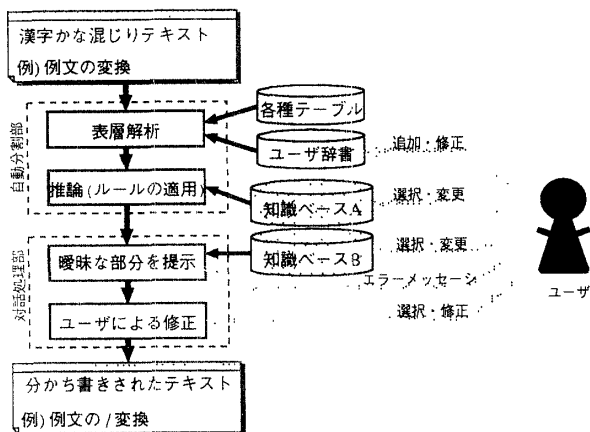


図 2: 分かち書き処理部におけるシステム構成

(2) 自動分割で誤っている箇所の原因を調べ、同様な傾向のKWIC(Key Word In Context) リストを出力し、誤りを含む割合が高ければ知識として採用する。例えば、情報処理の標準的なテキスト [4] について調べると、3文字以上からなる漢字連続において誤りが多かった。そこで、漢字テーブルに登録されていない漢字熟語のKWICリストを調べると、全文中 1048 回出現しており、このうち 32% の漢字連続が誤りであった。よって、このような漢字連続は誤りを含んでいる可能性が高いので、表 3 の知識 B-3 を導いた。

3.3 構成と機能

点訳システム全体の処理の流れを図 1 に示す。特に本システムの分かち書き処理部は 3.1 節の方針を実現するため、図 2 のように自動分割部と対話処理部の 2 段階で構成した。

自動分割部は、漢字かな混じりテキストを入力し、知識ベース A を用いて表層情報に基づく分かち書き処理を行う [3]。ここでは、ユーザがユーザ辞書と知識ベース A の知識を選択・追加・修正することにより、自動分割の精度を上げることができる。

また、対話処理部は自動分割されたテキストに対して、知識ベース B を用い正しく分割されないと考えられる箇所をユーザに提示する。ユーザは提示される箇所から誤りを探してその修正を行うことで、正しい分かち書きテキストを得ることができる。

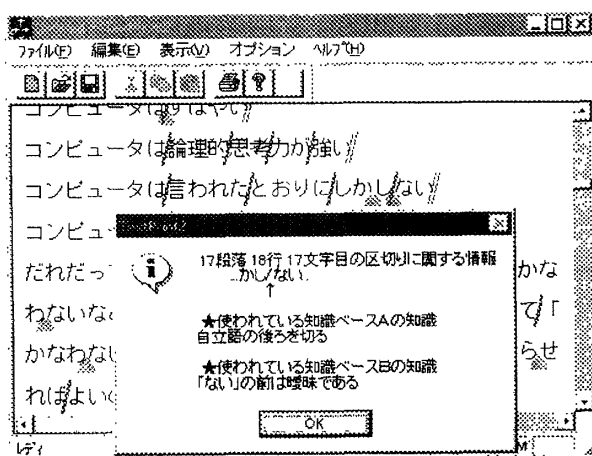


図 3: システムの実行例

4 実行例

本システムの実行例を図 4 に示す。本システムは 1 段落ごとに処理を行う。図 4 は、自動分割処理の後、対話処理で修正を行っているところである。ここでシステムは知識ベース A をもとに区切った箇所を '/' で示し、知識ベース B の知識をもとに疑わしい箇所を三角と網かけで示している。ユーザは上記のような記号で示される箇所について、図のようなダイアログボックスにより、どのような知識によって区切られている (いない) か、疑わしいとされているかを知り、修正を行う。

なお、自動分割の精度については文献 [3] に示す。

5 おわりに

本稿では、従来の点訳システムより精度が高く、曖昧な区切り箇所をユーザに示すことによりユーザと対話しながら点訳を進めていくシステムについて述べた。

今後の課題としては、疑わしいところを過不足なく指摘すること、漢字かな変換、かな点字変換の処理部を作成することなどがあげられる。

謝辞

本研究にご協力いただいた日本 IBM 東京基礎研究所の西野氏、点訳ボランティアの辰巳氏、筑波女子大学の鈴木ゼミの皆様、筑波大学 工学研究科の川崎氏に感謝の意を表します。

参考文献

- [1] 河原: 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] 鈴木: 統計調査に基づく文字列パターンを用いた日本語自動分割, 電子情報通信学会論文誌 Vol.J77-D-II, No.7, pp.1236-1243 (1996).
- [3] 平岡, 水野, 小野, 鈴木, 狩野, 西原: 知識ベースに基づく点字翻訳のための日本語文分かち書き手法, 第 54 回情報処理学会 全国大会 4B-10 (1997).
- [4] 宇都宮: コンピュータ入門, 共立出版 (1990).