

英語の点字表記から墨字表記への自動変換のための言語処理

藤崎博也 窪敏喜 伊藤博保
東京理科大学 基礎工学部

4Q-4

1 はじめに

重度の視覚障害者は、知能が優れていても、その障害のゆえに情報の入出力が困難である。特に現行の情報入出力のための補助機器は、高度な内容を持つ学術情報の入出力に適していない。従って、重度視覚障害者が大学・大学院レベルで高度な学術情報を検索・利用したり表現したりするには大きな支障がある。

本稿では、高度な学術情報について点字を墨字（我々が通常用いる文字）に自動変換するシステムを考え、その中でも英語の点字表記に着目し、変換のための言語処理について述べる。

2 英文点字の略字表記

英語の点字表記法にはアルファベットをそのまま表記する一級表記と、略字表記を用いる二級表記とがある[2][3]。二級表記は一級表記に比べてマス数が少なく、速読が可能になる、といった利点をもつ反面、点字表記に特有の規則や英語に関する知識がないと正しく解読することができない。

略字表記は7種類に分類され、それぞれ表記内容、使用可能箇所さまざまな規則がある[3]。

表1. 二級表記の種類と機能

種類	表記内容
1マス略語	1マスで1語
1マス略字	1マスで1語、または語の1部分
低下略字	1マスで語の1部分
低下略語	1または2マスで1語
頭字略字	2マスで1語、または語の1部分
末字略字	2マスで語の1部分
縮語	類出語や単語の部分を短縮

3 点字解析木による解析

3.1 点字番号

点字の各点には、左上から右下へ順次1番から6番までの番号が付けられており、その凹凸情報は6ビットの2進数で表現できる。以下では、これを10進数に変換したものを点字番号とする。なお、点字番号0は空白を表す。

3.2 点字解析木

自然言語処理の手段として木構造による解析があるが[4]、ここでは点字解析木を定義する。これは点字番号列を墨字に変換するための解析木で、各節点の成分は辞書アクセスする点字番号列であり、候補の数だけの枝を持つ。探索は最長一致的に枝を作りながら再帰的に行う。図1は、点字番号列(1,27,21)に対して、墨字単語候補“ago”が生成される例を示すものである。

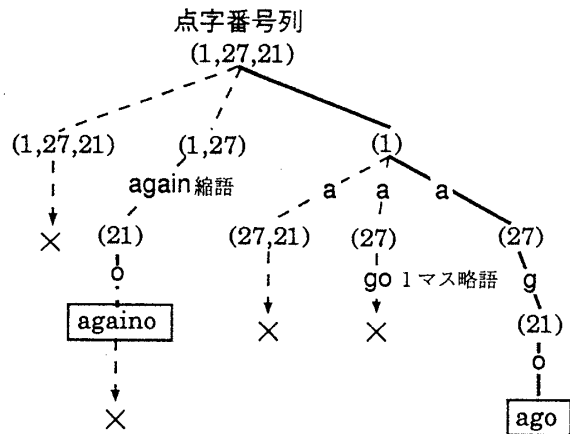


図1. 点字解析木による解析例

図中の破線の枝は候補として存在せず切り捨てられたものを表す。点字番号列(1,27,21)および(27,21)については対応する点字単語または一部分が存在しない。また、(27)についてはアルファベットの‘g’と1マス略語“go”が考えられるが、1マス略語は語の一部として使用できないという規則から捨てられる。

略字使用には発音との関連など使用可能条件にさま

Linguistic Processing in the Automatic Conversion of Braille to Roman Characters for English Sentences
Hiroya Fujisaki, Toshiki Kubo, Hiroyasu Itoh
Department of Applied Electronics
Science University of Tokyo
2641 Yamazaki, Noda, 278, Japan

ざまな規則があり、生成された2つの候補のうちどちらを選択するかは、点字番号列の情報からだけでは決定することはできない。これは、縮語なのか否か木の生成過程では判断できないことが原因である。この例では“againo”に含まれる縮語(“again”)は、縮語を語の一部とする場合の規則に反する。しかし、略字表記に関する知識がなくても“againo”という単語が存在しないことから“ago”が正解であろうと判断できる。

複数候補の中からどれを選択するかは、まず、それぞれの候補について英単語として存在するかどうか辞書アクセスする(a)方式で行なう。しかしながら縮語を語の一部として使用できる英単語の数は、複数候補が生じる英単語の数に比べてはるかに少ない。従って点字解析木を作る段階で、縮語が語の一部として使われる候補は一時的に除外する。生成単語が英単語として存在しない場合には、縮語を含むものも候補として辞書アクセスする(b)方式の方が処理効率が高いと思われる、以下、両方法による比較検討を行なった。

4 評価および検討

4.1 評価

点字使用者の略字表記に関する知識の有無によって、表記内容は一級表記であったり二級表記であったりし、場合によっては2つが混在する。従ってここでは、

- 1) すべて一級表記による点字英文。
- 2) すべて二級表記による点字英文。
- 3) 一級表記と二級表記とが混在する点字英文。
略字が使用できる部分について、使用するかないかを乱数によって決定した点字英文。

の3通りの方法で点訳した高等学校レベルの100英文(延単語数1215)、および日本国憲法の条文の英訳点字(207文, 延単語数4658)を対象としてシステムの評価を行なった。複数候補の中からどれを選択するかは、先に述べた(a), (b)それぞれの方法で行ない、正解率、辞書アクセス回数を比較した(表2)。

すべての単語は候補内に正解があり、候補が1つのものはすべてその候補が正解であった。

4.2 検討

一級表記文、二級表記文、混在文ともに、問題となるのは縮語なのかそうでないのかの区別である。結果をみるといずれも正解率は100%であるが、辞書アクセス回数に大きな差がある。(a)方式は単純な辞書ア

表2. 処理結果

文 種類	複数候補 単語数	正解率(%)		アクセス回数		
		(a)	(b)	(a)	(b)	
高 校	1)	110	100	100	241	110
	2)	61	100	100	135	65
	3)	89	100	100	197	92
憲 法	1)	748	100	100	1752	748
	2)	422	100	100	913	427
	3)	585	100	100	1326	588

クセスの繰り返しであるが、アクセス回数は膨大となる。一方、(b)方式の処理は(a)方式に比べて複雑になるが、アクセス回数は少ない。

ここで処理の対象とした単語の中で点字解析木の枝分かれの数が最も多かった(候補数が最も多かった)のは、点字番号列(39,10,7,7,1,27,17)、(1,7,15,19,1,3,17,30)などで、4通りの候補が生成された。対応する墨字はそれぞれ“village”と“alphabet”である。2つの単語はいずれも略字表記を使わない単語であるが、縮語使用の可能性のある候補が生成される。(a), (b)いずれの方法でも正解が得られているが、辞書アクセス回数が異なり、(a)では4回であるのに対し、(b)ではわずか1回であった。

5 おわりに

点字解析木を用いて英語の点字表記を墨字表記に変換する方式について提案し、変換候補が複数ある場合でも正しい墨字表記に変換されることを確認した。(a)方式と(b)方式の2つを組み合わせることで、より効率の高い処理が期待できる。

今回対象にした点字英文はいずれも表記に誤りのない英文ばかりであった。点字を表記するのが人間である以上、誤りは発生するおそれがあり、これを訂正しながら墨字に変換する処理も必要となろう。

参考文献

- [1] 木塚泰弘：“視覚障害者のための情報入出力技術の課題,” 「文字言語・音声言語の知能的処理」第152委員会 第21回研究会資料(1991.7).
- [2] 文部省：点字学習指導の手引, 東山書房(1978).
- [3] 福井哲夫：初歩から学ぶ英語点訳 改定版, 日本点字図書館(1991).
- [4] 岡田直之：自然言語処理入門, 共立出版(1989).