自動点字翻訳編集システム ibukiTenC
－点字毎日との比較による精度評価－

高田 和典，江本 倫基 ，脇田 貴之，山田 佳裕，池田 尚志
岐阜大学工学部

1　はじめに

我々の研究室では視覚障害者の読書支援･点訳ボランティアの活動支援となる自動点訳システム ibukiTen を開発し，公開してきた [1]．今回，これまで EDR の辞書をベースにしていた ibukiTen の辞書を我々独自の辞書に置き換え，整備を加えた新しいシステム ibukiTenC を開発し公開することとした．

本論文では，ibukiTenC について述べるとともに，ibukiTenC の点訳精度を評価するために，毎日新聞社より発行される点字による週刊新聞「点字毎日」を対象として行った評価実験の結果を示す．比較には，我々の研究室で開発された比較プログラムを用いた．実験から得られた精度評価，そこから考えられる問題点について述べる．

2　自動点訳システム ibukiTenC

点訳を行うエンジン部は，我々の研究室で開発している日本語解析システム IBUKI の文節解析をベースに開発している．点訳エンジンの処理の流れ，点訳辞書，単語条件・点訳規則の構築などについて述べる．

2.1　システム概要

システム構成を図 1 に示す．

点訳の流れとしては，日本語文が入力されると，日本語解析システム IBUKI によって文節解析と複合語解析を行う．そのあと ibukiTenC の点訳処理（分かち書き処理，点字表記変換），そして校正作業ができる ibukiTenCEdit では後編集処理（点訳誤り検出，かな表記修正，分かち書き修正）を行い，点字を出力する．

2.2　点訳用辞書

辞書には，解析用の情報のほかに，点訳用の点訳表記の項目がある．点字で分かち書きを必要とする個所には，”/ ”を記述してある．自立語は，現在のところ

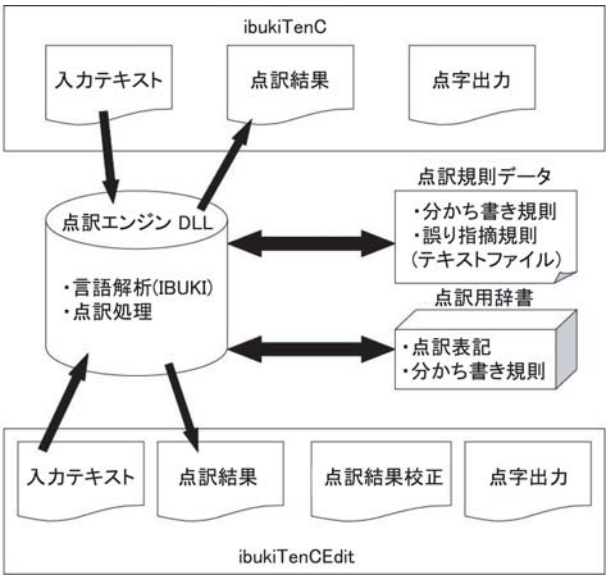

図 1: システム構成

約 22 万語登録されている．機能語は，IBUKI の特徴である長単位での登録となっており，約 2 万語登録されている．必要な個所を ”/ ”で区切った点訳表記が記述されている．辞書の例を図 2 に示す．複数の読みを持つ単語には読みコストを与え，点訳の際に用いる読みの優先順位としている．

見出し語	品詞説明	点訳表記	コスト
会社	名/一般	かいしゃ	0
今日	名/時	きょー	1
今日	名/時	こんにち	2
作物	名/一般	さくもつ	1
作物	名/一般	さくぶつ	2
作物	名/一般	つくりもの	2
民主主義	名/一般	みんしゅ/しゅぎ	0
日	尾/名/時	にち	0
にちがいない	機/ダ｜文末	に/ちがい/ない	0
に違いない	機/ダ｜文末	に/ちがい/ない	0

図 2: 点訳用辞書

3　評価実験

今回，毎日新聞社が毎月発行している「点字毎日」の点字データを正解データとして評価を行った．点字毎日の記事は点字コード BASE（拡張子：bse）で保存されている．それに対応した原文記事は毎日新聞から，点字毎日付録「テキスト版」として発行されている．評価に用いた点字毎日の記事は，2006 年 10 月号 1ヶ月のデータ（2,338 文，196KB）である．精度評価を行うにあたって，BASE で保存されているデータを一度ひらがな表記へ変換し，これと原文とを対応させた．

評価は ibukiTenC と，現在 WWW にて公開中の ibukiTen Ver0.56，市販の自動点訳システム E を用いて，それぞれの点訳結果の分かち書き正解率（切り過ぎ，切り忘れ），漢字の読み正解率の評価を行った．

3.1　ibukiTenC 精度評価プログラム

比較作業を自動化するために精度評価プログラムを用いた．比較するための前処理と，精度評価プログラムの概要について述べる．

3.1.1　前処理

今回評価の対象としたテキストは点字毎日の記事である．点字毎日の点字データとその原文では，記事の位置がばらばらであったり，また，点字データでは原文を簡略化していたりして，点字データと原文は必ずしもそのまま一致しない．そのため，その対応関係を取る必要があった．また，点字毎日のテキスト中には，短歌や俳句が含まれる記事，方言が多い口語的な記事なども含まれているが，それらは評価対象から除外した．前処理の手順を以下に示す．

1. 点字毎日から提供されている月々の点訳データ（フロッピーディスク）を ibukiTenC によって点字かな変換を行い，テキスト（正解テキスト）に保存する．
2. 点字毎日付録「テキスト版」の点字毎日の原文から，正解テキストに対応する文章を抜き出す．
3. 原文と正解テキストの対応をとるため，あらかじめ原文と正解テキストは対応する文ごとに改行を行う．「文ごと」とは，句点までを一文としている．
4. 題名は，原文と正解テキストがあまりにも違っているので削除する．
5. 更に細かく正解テキストと原文を比較し，文章の表現が違っていたり，または簡略化や削除されている部分を修正する．
6. 点字毎日のデータは日本点字表記法 2001 年版 [2] によって変更があった「〜する」の表記に対応していないため，その部分の修正を手作業で行う．

点字毎日の原文データを表 1 に，点字データを ibukiTenC によって点字かな表記へ変換したテキストを表 2 にそれぞれ示す．

表 1: 点字毎日原文

「高松でヘレン・ケラー写真展」
― 映画上映会場に点毎パネル　―

　視力と聴力を失った山口県在住の 70 代の女性をモチーフにした映画「ヘレン・ケラーを知っていますか」（小林綾子主演）の高松での上映会場で１日、ヘレン・ケラー（1880〜1968）の写真展が開かれた．毎日新聞高松支局が「来場者に三つの障害を乗り越え活躍したヘレン・ケラー本人の写真も見てもらえれば」と企画、点字毎日が写真パネル 12 点を貸し出した．
　上映会は地元の映画サークルなどで結成した同映画上映実行委員会が主催した．会場の香川県県民ホールのロビーには、ヘレンの生い立ちの説明文と共に、10 代のヘレンと家庭教師、アン・サリバンさんとのツーショットや、戦前の１９３７（昭和 12）年と、戦後の 48（同 23）年、55（同 30）年の３回にわたり来日した際の写真など、ヘレンの素顔を伝えるパネルが並んだ．３回目の来日で、大阪市の点字毎日を訪問した写真もある．上映会には約１０００人が入場．映画の後、じっくりと写真展に見入る家族連れもあった.【毎日新聞高松支局・久門たつお、矢島弓枝】

表 2: 点字毎日点字データ

　＿たかまつで　へれん　けらー　しゃしんてん＿
――　えいが　じょーえいかいに　てんまいの　ばねる　――

　しりょくと　ちょーりょくを　うしなった　やまぐちけん
ざいじゅーの　数７０だいの　じょせいを　もちーふに　した
えいが　＿へれん　けらーを　しって　いますか＿（こばやし
あやこ　しゅえん（の　たかまつでの　じょーえい　かいじょーで
ついたち　へれん　けらー（数１８８０〜〜数１９６８の
しゃしんてんが　ひらかれた。　まいにち　しんぶん　たかまつ
しきょくが　＿らいじょーしゃに　みっつの　しょーがいを
のりこえ　かつやくした　へれん　けらー　ほんにんの　すがたも
みて　もらえれば＿と　きかく　てんじ　まいにちが
しゃしん　ばねる　数１２てんを　かしだした。
　じょーえいかいわ　じもとの　えいが　さーくるなどで
けっせいした　どー　えいが　じょーえいかい　じっこー
いいんかいが　しゅさいした。　かいじょーの　かがわけん
けんみん　ほーるの　ろびーにわ　へれんの　おいたちの
せつめいぶんと　ともに　数１０だいの　へれんと　かてい
きょーし　あん　さりばん　さんとの　つーしょっとや
せんぜんの　数１９３７しょーわ　数１２ねん　せんごの
数４８どー　数２３ねん　数５５どー　数３０ねんの
数3かいに　わたり　らいにちの　さいの　しゃしんなど　へれんの
すがおを　つたえる　ばねるが　ならんだ。　数3かいめの
らいにちで　おおさかしの　てんじ　まいにちを　ほーもんした
しゃしんも　ある。　じょーえいかいにわ　やく　数1せんにんが
にゅーじょー。　えいがの　あと　じっくりと　しゃしんてんに
みいる　かぞくづれも　あった。　（まいにち　しんぶん
たかまつ　しきょく　くもん　たつお　やじま　ゆみえ（

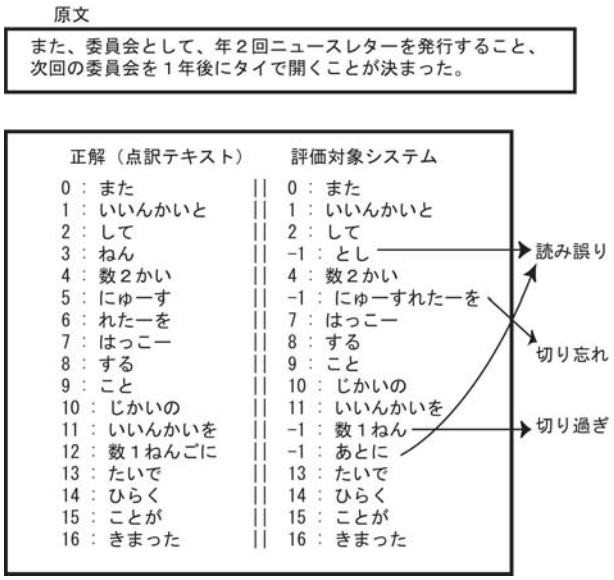

図 3: 点訳評価プログラムの点訳結果比較

3.1.2 評価プログラム概要

プログラムでは，評価対象のシステムが出力した点字かな表記の点訳結果と，正解の点字かな表記の点訳テキストと，その原文の 3 つのファイルを読み込んでシステムの評価を行う．

正解の点訳テキストからは全ての分かち書きの合計をカウントし，原文からは，漢字連続文字列を 1 と数えた漢単語数をカウントする．よって,「岐阜大学」などの複合語でもカウント数は 1 となる．

そして，システムの点訳結果と，正解の点訳テキストを比較することによって，分かち書きの誤り，漢字の読みの誤りを抽出する．

図 3 を用いて，比較手法を説明する．

1. 正解の点訳テキスト（以下，テキスト A）と評価対象システムの点訳結果（以下，テキスト B）から，それぞれ 1 文ずつ切り出す．
2. テキスト A から切り出した 1 文を分かち書きの単位に分け，先頭から順に番号を付ける．
3. テキスト B から切り出した 1 文を，分かち書きの単位でテキスト A と比較する．
4. 分かち書きが一致した場合は，テキスト A の分かち書きに振られた番号をテキスト B の分かち書きに付ける．分かち書きが一致しなかった場合は，テキスト B の分かち書きに「-1」を付ける．
5. 「-1」が付いたテキスト B の分かち書きと，それに対応するテキスト A の分かち書きを比較し，分かち書きの誤り，漢字の読み誤りを抽出する．

比較が終了すると，その結果から点訳精度を計算する．分かち書き正解率は，区切ってはいけないところを区切ってしまった誤りに対する切り過ぎ正解率，及び区切らなければならないところを区切らなかった誤りに対する切り忘れ正解率として，それぞれ以下の (1) 式，(2) 式により求める．漢字かな変換における読みの正解率は (3) 式によって求める．

$$\text{正解率（切りすぎ）} = (1 - \frac{\text{切りすぎ個所}}{\text{正解の区切り数}}) \times 100 \quad (1)$$

$$\text{正解率（切り忘れ）} = (1 - \frac{\text{切り忘れ個所}}{\text{正解の区切り数}}) \times 100 \quad (2)$$

$$\text{正解率（読み）} = (1 - \frac{\text{漢字読み誤り個所}}{\text{漢単語数}}) \times 100 \quad (3)$$

3.2 評価実験

3.2.1 精度評価

精度評価実験に用いた点字毎日の記事データを表 3 に示す．漢単語数は漢字連続文字列を 1 と数えたときの合計数である．

表 3: 点字毎日の記事データ

正解区切り数	28,665
漢単語数	17,635

ibukiTenC，ibukiTen Ver0.56，市販の自動点訳システム E，それぞれの点訳結果を用いた分かち書き，及び読み正解率の評価結果を表 4 に示す．

表 4: ibukiTenC による自動点訳正解率

	ibukiTenC	ibukiTen	システム E
正解率 (切り過ぎ)	96.9301%	98.1197%	96.5986%
正解率 (切り忘れ)	96.8777%	97.565%	97.1429%
正解率 (読み)	92.5035%	95.8265%	97.2498%

3.2.2 誤り個所の考察

分かち書き・漢字読み取りが誤った個所について，その例文を示し考察を行った．

入力文　シャッフルボード
文節解析　シャッフルボード（名/未知語/カタカナ）
点訳結果　しゃっふるぼーど
正しい点訳結果　しゃっふる　ぼーど

正しい点訳では「シャッフル」と「ボード」の複合語として区切って点訳するが,「シャッフルボード」が未知語として解析され，区切らずに点訳されている.「ボード」という単語は既に辞書に登録されており,「シャッフル」という単語を辞書に登録することで，複合語の点訳規則を適用でき，正しい点訳を行うことができる.

入力文　バスターミナル
文節解析　バス（名/一般）／ターミナル（名/一般）
点訳結果　ばすたーみなる
正しい点訳結果　ばす　たーみなる

ibukiTenC では，複合語内の 2 拍以下のカタカナは続けて書くとしているため,「ばす」と「たーみなる」を続けて点訳している．日本点字表記法には「意味のまとまりが 2 拍以下であっても，自立性が強く，意味の理解を助ける場合には区切って書いてよい」とあり，点字毎日では「バス」の意味を強調するために「ばす□たーみなる」と区切って書いている．しかしこの表記に従うには，単語の意味やその扱われ方を考慮する必要がある.

入力文　藤縄佑樹君
文節解析　藤縄（名/一般）／佑（名/一般）／樹（名/一般）／君（尾/名/人名 | 名/人名）
点訳結果　ふじなわじょーきくん
正しい点訳結果　ふじなわ　ゆーき　くん

「藤縄」「佑樹」ともに人名として単語登録されておらず，誤った解析が行われている．また ibukiTenC では,「君」は直前の単語が人名であった場合に前を区切るとしているため，ここでは続けて点訳されている．これらを正しく点訳するには,「藤縄」「祐樹」といった単語を人名として辞書に登録する必要がある.

入力文　ｃｒｏｗ
文節解析　ｃｒｏｗ（名/未知語/ローマ字）
点訳結果　外ｃｒｏｗ
正しい点訳結果　≪ｃｒｏｗ≫

アルファベットや，アルファベットによる略称などは，外字符を前置して書き，ibukiTenC もその規則に従い点訳を行っている．ただし，アルファベットで書かれた語句や文は，外国語引用符（ ”≪ ”〜 ”≫ ”）で囲んで書かなくてはならない．しかし外字符と外字引用符のどちらを用いるかという判断には，アルファベットが表しているのが略称なのか英単語なのか，といった情報が必要であり，これを正しく点訳するのは難しいと考えられる.

4　公開状況

我々の研究室では 2000 年 9 月より，ibukiTen を WWW 上に公開している [1]．今年度（2006.4〜2007.1）のダウンロード延べ数はおよそ 1,500 件（公開開始からの延べ数はおよそ 10,000 件）異なり数でおよそ 1,100 件（公開開始からの異なり数はおよそ 5,500 件）である．今回，本論文で述べた ibukiTenC を新たに WWW 上に公開することとした.

5　おわりに

精度評価プログラムを用いて，ibukiTenC，ibukiTen-Ver0.56，市販の自動点訳システム E，それぞれの点訳結果の評価を行った．精度の比較では，ibukiTen や市販のシステムに比べ，現在開発中の ibukiTenC の精度はやや低かった．特に読みの精度が低く，解析誤り，分かち書き誤りに加え，辞書の語彙不足も精度低下の大きな原因と考えられる．辞書の整備を進め，ibukiTenC の精度向上を目指したい.

参考文献

[1] 自動点字翻訳編集システム IBUKI-TEN, http://www.ikd.info.gifu-u.ac.jp/IBUKI-TEN/

[2] 日本点字委員会，日本点字表記法 2001 年版，2001.

[3] 高松大地，岸井謙一，伊佐治和哉，松本忠博，池田尚志 (岐阜大)，自動点訳システム IBUKI-TEN と新点字規則への対応，言語処理学会 第 10 回 年次大会 (NLP2004)，2004.