日本語文節解析システムの自動点訳への応用

兵藤安昭　横平貫志　池田尚志
岐阜大学工学部
{hyodo,kanji,ikeda}@ikd.info.gifu-u.ac.jp

1 はじめに

我々は現在開発中の日本語解析システム (IBUKI) を，自動点訳に応用することを検討している．点字翻訳では，一般に，点字規則に従って分かち書きを行い，漢字をカナに変換する必要があるが，今回は分かち書きに関してシステムの作成を行った．

通常の形態素解析は，文を基本的に単語単位に分割するが，これは点字の分かち書き単位と異なっており、そのまま点訳に適用することは難しい．文献1では，形態素解析を行わず，知識ベース化した分かち書き規則に基づくシステムを開発している．

解析システムIBUKIでは文節単位の切り出しを行っており，意味的なまとまりに従って比較的長い単位で切り出している [2]。例えば「進まざるを得なくなる」を1つの文節として切り出す．これを点訳規則に従った短い単位に分解することは容易である．これは、IBUKIが使用するRDB上の辞書に，点訳用の分かち書き規則を記述することで比較的簡単に実現できる．

2 点字分かち書き規則

点字では，基本的には，文節の単位で分かち書きを行うが，点字用の規則に従って分かち書きを行わなければならない [4][5]．例えば，形式名詞などは，自立語とし，前で区切りを入れる必要がある．

- 「家に帰ったところ」… 家に / 帰った / ところ
- 「悲しみのあまり」 … 悲しみの / あまり
- 「思ったまま」 … 思った / まま
- 「忘れずに読むこと」 … 忘れずに / 読む / こと

また，補助動詞についても自立語とし，前で区切る必要がある．

- 「持っていく」… 持って / いく
- 「現代における」 … 現代に / おける

3 解析システム (IBUKI) 概要

日本語解析システム (IBUKI) の構成図を図1に示す．IBUKIは，辞書引き，文節候補の作成，最良な文節の選択，係り受け解析，の4つの部分からなる．

IBUKIの解析用自立語辞書は，EDR日本語単語辞書 [6] をベースに平仮名以外の同一文字種からなる普通名詞・サ変名詞を削除したもの (約13万語) を用いた．なお，IBUKIでは，漢字連続文字を名詞として解析し，複合語解析は係り受け解析の前に行う．また，機能語辞書は次節で述べるように独自に作成した．

パトリシア構造によるメモリ上の辞書を参照して，入力文中の各位置から始まる単語を切り出し，解析表に書き込む．そして，連接可能性をチェックし，文節内の末尾の単語が未然型ではないなどの規則により文節候補を作成する．その後，文節候補に対して，文節と文節間のコストを参照し，Viterbiアルゴリズムを用いてコスト最小解を求める．

係り受け解析は，文節に付与した文節カテゴリに基づいて係り可能な文節を求め，文頭から文末に向かってブロック化処理を適用することで，係り先を決定する [3]．

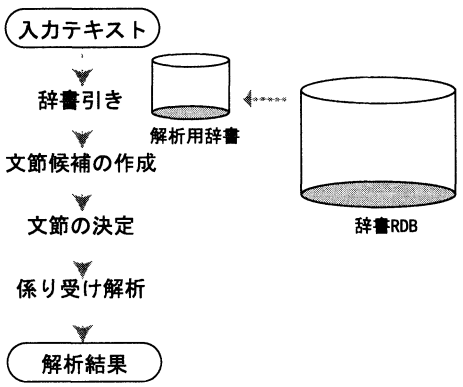

図 1: 解析システム構成図

4 点字分かち書き手順

4.1 辞書ＲＤＢ

図 1 に示すように，すべての辞書情報は RDB 上で一括管理している．文節，構文解析で直接参照する辞書は，この RDB から，表記，連接属性など解析に必要最低限の情報を取り出し，メモリ上に読み込む．

辞書 RDB 上のテーブル例を図 2 に示す．自立語辞書テーブル [図 2-(1)] は，EDR 単語辞書に含まれる辞書情報に加えて，点訳規則情報を登録した．これは，点字では「宇宙 (うちゅう)」などのウ列の長音は「ウチュー」のように表記するからである．

(1) 自立語辞書テーブル

ID	表記	左連接	右連接	カナ表記	点訳規則
100	美し	JLA1	JRA1	うつくし	うつくし
101	宇宙	JLN1	JRN1	うちゅう	うちゅー
	……	……	……	……	……

(2) 機能語辞書テーブル (スロット表現)

スロット表記	@1	@2	左連接		右連接1	右連接2	右連接3	右連接4
ざるを / @@	え，得/え	ゆ，ぬ，ず，まい	AM		RAX1	ZZ	ZZ	ZZ
……	……	……	……	……	……	……	……	……

(3) 機能語辞書テーブル

ID	表記	左連接	右連接	カナ表記	点訳規則
600	ざるをえ	AM	RAX1	ざるをえ	ざるを/え
601	ざるをえぬ	AM	ZZ	ざるをえぬ	ざるを/えぬ
602	ざるをえず	AM	ZZ	ざるをえず	ざるを/えず
603	ざるをえまい	AM	ZZ	ざるをえまい	ざるを/えまい
604	ざるを得	AM	RAX1	ざるをえ	ざるを/え
605	ざるを得ぬ	AM	ZZ	ざるをえぬ	ざるを/えぬ
606	ざるを得ず	AM	ZZ	ざるをえず	ざるを/えず
607	ざるを得まい	AM	ZZ	ざるをえまい	ざるを/えまい

図 2: 辞書テーブル

IBUKI では，文節を意味的なまとまりに従って切り出すために，できるだけ長い単位で機能語を登録している．そのため，点訳の際には，点字規則に従って機能語を分割する必要がある．例えば，「ざるを得ない」の場合は「ざるを/得ない ('/' は区切りを示す)」となる．我々は，機能語辞書を，図 2-(2) に示すようなスロットモデルを用いて整理，分類している．これにより，例えば「ている，てはいる，てもいる」「かもしれない，かも知れない」のように，表記の違い，活用，助詞の付加による意味の添加などにより派生的な多数の機能語を系統的に整理分類することが可能となった．図 2-(2) は，機能語「ざるをえない」とその派生語に対する辞書表現である．[@] をここではスロットと呼ぶ．スロット表記に区切り記号 ('/') を挿入することで，点字用の分かち書き規則を表現している．この場合は，「ざるを」の後で区切りを入れることになる．また，@1,@2 式がスロットに入り得る要素を示す．このテーブルから，図 2-(3) に示す機能語辞書テーブルが自動的に作成される．

4.2 処理手順

解析された各単語には，辞書 RDB 上への ID が付与されている．分かち書き処理は，この ID をもとに，対応する点訳規則を辞書 RDB から検索することで行われる．例えば，図 3 に示すように，文節解析で「これについて」,「書いてある」と 2 文節に分割される．この後，辞書 RDB を参照して，機能語「について」は「に＋ついて」に「てある」は「て＋ある」に分割する．

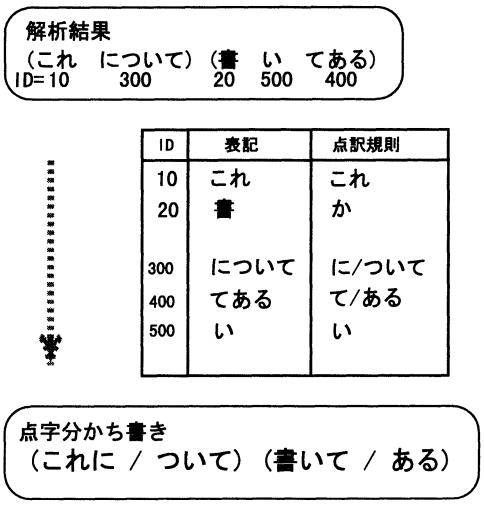

図 3: 点字分かち書き処理

5 評価実験

5.1 分かち書き実験

朝日新聞記事社説100文(4537文字)について，本システムを用いて点字分かち書き処理の評価実験を行った．結果を表1に示す.「切りすぎ」とは，必要のない箇所を区切ってしまった誤り,「切り忘れ」とは，区切り忘れの誤りである．それぞれの正解率は以下のように計算した.

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

$$\text{正解率 (切り忘れ)} = (1 - \frac{\text{切り忘れ箇所}}{\text{正解の区切り数}}) \times 100$$

表 1: 分かち書き正解率

	切りすぎ	切り忘れ
全体	97.22%	94.37%
複合語を除く	97.76%	98.25%

5.2 誤りに対する考察

分かち書きが正しく行われなかった箇所として，自立語に関する誤りが多かった．以下に誤り例を示す.

1. 敵方といっしょになって
   - (正解) 敵方と / いっしょに / なって/
   - (誤り) 敵方と / いっしょになって/
2. よく知っている
   - (正解) よく / 知って / い る
   - (誤り) よく / 知っている/
3. このへんでなんとしても
   - (正解) このへんで / なんと / しても/
   - (誤り) この / へんで / なんと / しても/

1,2については，EDR辞書に「いっしょになる」「知っている」が1語として登録してあるためによる誤り，3は「このへん」等，前の語と複合して1語で意味をなす場合，点字では区切ってはいけないという規則による誤りである.

本システムでは，現在のところ，複合語に対する点字用の分かち書き規則を導入していないため，複合語の解析精度が低かった．特に漢字，カタカナ混じりの単語で，誤って区切ってしまう例が多い．これは，複合語を字種によって分割しているからである．複合語で，正しく分かち書きされなかった(切りすぎによる誤り)箇所としては以下の例がある.

- (誤り) ダルシー / 賞
- (誤り) ワールド / 紙
- (誤り) U / ターンする

6 おわりに

本論文では，現在開発中の日本語解析システム(IBUKI)を，点訳用の分かち書きシステムとして適用することについて述べた．今後は，複合語に関する分かち書き規則を導入し，システムの精度向上を図りたい.

参考文献

[1] 鈴木恵美子, 小野智司, 狩野均: 点字翻訳ボランティアのための対話型分かち書き支援システム, 自然言語処理,Vol5,No.4,pp95-110(1998)

[2] 兵藤安昭, 池田尚志：文節単位のコストに基づく日本語文節解析システム, 言語処理学会 第5回年次大会, (1999)

[3] 兵藤安昭, 若田光敏, 池田尚志：文節ブロック化規則による浅い係り受け解析と精度評価, 信学技報, NLC98-30, pp. 33–39, 1998

[4] 点訳のてびき第2版:全国視覚障害者情報提供施設協議会, (1991)

[5] 最新点字表記辞典増補改訂版:視覚障害者支援総合センター, (1998)

[6] 日本電子化辞書研究所: EDR 電子化辞書仕様説明書, (1995)