文節解析における解析誤り箇所の検出について

村上裕　　兵藤安昭　　池田尚志
岐阜大学工学部
{yutaka,hyodo,ikeda}@ikd.info.gifu-u.ac.jp

1　はじめに

我々は，日本語解析システム (IBUKI) を開発し，いろいろの応用を試みている [1]．IBUKI はまず文節解析を行い，ついで係り受け解析を行う．本論文では，IBUKI の文節解析結果の誤り個所を検出することを目的として行った検討結果について述べる．

文節解析結果におかしな個所があるということは，入力文がおかしな個所を含んでいるのが原因である場合と，解析システムがおかしな解析をしたのが原因である場合とがある．誤り個所を指摘することは，前者の場合には OCR 後処理や文書校正への応用において，後者の場合には機械翻訳や点訳など人間との対話を通して仕事を遂行する応用において，基本的に重要である．我々は点字翻訳システム IBUKI-TEN を開発しているが [2]，点訳ボランティアに有効に活用してもらうためには，点訳誤り (解析誤り) の可能性のある場所を (100 %に近い再現率で) 提示して，その場所を点検し手直ししてもらうといったことが出来ることが望まれる．

通常の形態素解析（単語分割と品詞付与）結果に対する誤り個所の検出法にはいくつかの試みが報告されているが [3]，本報告では文節の形態に注目して，誤り個所を調査・検討した．

2　文節解析における誤り箇所の調査

IBUKI は，単語分割だけでなく文節まとめ上げまでを含めて文節単位のコスト最小法で解析を行っており，文節には，文節カテゴリ (75 種類) を付与している．文節カテゴリとは，構文的な観点から文節を分類したもので，係り受け解析を行う際に，係り受け可能な文節を求める時に利用する．

文節カテゴリは 3 文字で示されており，左から文節の種類，文節の種類の細分類，係り得る文節の種類を示している。例えば,「体の体」は，体言に係る体言文節で，細分類として助詞の「の」を含んでいる文節のことを示している．

今回の実験では，文節解析が誤っている文節を抽出し，特定の文節カテゴリによって文節解析誤りを判別することができるかについて調べた．

具体的には，

①毎日新聞記事約 2 万文に対する京都大学コーパスの解析結果（正解）と，同記事に対する IBUKI の解析結果を比較することで文節解析が誤っている箇所を調べる．

②特定の文節カテゴリで文節解析が誤っているものを数え上げる．

ということを行った．比較の方法は，図 1 のように，IBUKI 解析結果の内，ある文節の直後の文節内の自立語部分が京都大学コーパスでは文節の自立語先頭部分でないとき，その文節において文節解析が誤るとした．例えば，IBUKI 解析結果では文節「髭も」の直後の文節「じゃで,」の自立語部分「じゃ」が，京都大学コーパスでは，文節「髭もじゃで」の自立語部分の先頭でないので，IBUKI の解析結果「髭も」は文節解析が誤っているとする．

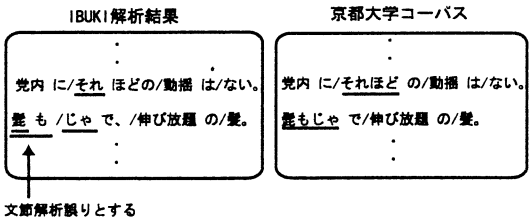

図 1: 比較方法

3　はだかの体言文節，用言文節に関連する解析誤り

3.1　はだかの体言文節，用言文節における文節分割誤り調査

文節カテゴリの内,「体＊＊」,「ＶΦ用」は，文節内に機能語部分が存在しないもの（はだかの体言文節と

はだかの用言文節）であり，本来一つの文節部分が過分割されて生成されているなど，解析誤りである可能性が高い（図 2）と考え調査した．

IBUKI 解析結果の内，「体＊＊」，「ＶΦ用」，また未定義語を含む文節である「未＊独」を含むものを抽出し，実際に文節解析が誤っているかについて調査した．結果を表 1 に示す．ただし，読点を含んだ「体＊＊」，「ＶΦ用」や，直後に括弧が続く「体＊＊」，「ＶΦ用」はすべて文節解析が正しいものであったので，ここでは考慮しない．

ガス会社を ／ **グループ分** ／ **けして、**
体を用　　体＊＊　　Ｖ連用

延べ ／ **床面積は** ／ **五千三百平方メートル。**
ＶΦ用　　体は用　　だ＊末

図 2: 文節解析誤り例

IBUKI の文節解析の精度は，誤り率約 0.5%(937/181,927)，正解率約 99.5%である．
「体＊＊」，「ＶΦ用」，「未＊独」における文節解析誤りの適合率はそれぞれ，30.8%，20.2%，100%となる．また，「体＊＊」，「ＶΦ用」，「未＊独」全体としての再現率は 33.4%となる．

表 1: 調査結果

文節カテゴリ	全個数	文節解析が正しい個数	文節解析が誤っている個数
体＊＊	517	358	159
ＶΦ用	242	193	49
未＊独	105	0	105
それ以外	181,063	180,439	624
合計	181,927	180,990	937

3.2 はだかの体言文節，用言文節における誤り箇所の分析

「体＊＊」，「ＶΦ用」における誤りの傾向として，次の 5 つに分類してみた．

1. 仮名異表記単語をもつ文節部分での誤り

・ ざん　新な　出来事だった。
体＊＊　Ａ＊体　だ＊末
・ 実力者だけに、　ひと　暴れしそうだ。
体に用　体＊＊　Ｖ＊末
・ 若者の　アブ　ない　揺れと　痛みを　描く。
体の体　体＊＊　形＊体　体＊並　体を用　Ｖ＊末

これは，通常漢字で表記される単語を異表記として仮名で表現されている場合に文節が 2 つ以上の文節に分割したり，その直前・直後の文節と結び付いてしまう場合である．

2. 固有名詞部分での誤り

・ さくらもも　この　原作で,･･･
体＊＊　副＊体　体で用
・ 右ＣＫを　けった　鈴木　いどむは･･･
体を用　Ｖ＊体　体＊＊　Ｖ終用

これは，1 つの固有名詞である部分が複数の文節に分割してしまう場合である．

3. 直前または直後が「ＶΦ用」，「体＊＊」，「未＊独」のいずれかの場合

・ ぎっしりと　重なり合　ひ　･･･」
副＊用　体＊＊　ＶΦ用　括＊末
・ いたずらっぽ　く　笑った。
体＊＊　未＊独　Ｖ＊末
・「　心身ともに　佳　い　女の子に･･･
括＊独　体＊用　体＊＊　ＶΦ用　体に用

これは，該当する文節の直前・直後の文節が，「ＶΦ用」，「体＊＊」，「未＊独」のいずれかであるものある．

4. 体言文節部分での過分割

・ 現実感覚と　成熟　さも　物語っている。
体＊並　体＊＊　副＊用　Ｖ＊末
・ 延べ　床面積は　五千三百平方メートル。
ＶΦ用　体は用　だ＊末
・ ムチ打ち　刑を　受けた･･･
ＶΦ用　体を用　Ｖ＊体

これは，1 つの体言文節が複数の文節に分割してしまう場合である．

5. その他

それぞれの場合において，文節解析が誤っている「体＊＊」,「ＶΦ用」の個数を調べた.(表 2)
直前・直後の文節が「体＊＊」,「ＶΦ用」,「未＊独」のいずれかである場合の「体＊＊」,「ＶΦ用」の個数はそれぞれ全体の 42.76%, 40.8%と高い割合を示しており,「体＊＊」,「ＶΦ用」が連続している場合は文節解析が誤っている可能性が高いといえる.

表 2: 「体＊＊」,「ＶΦ用」の誤り分類

分類	体＊＊	ＶΦ用
分類 1	11	0
分類 2	6	2
分類 3	68	20
分類 4	2	21
分類 5	72	6
合計	159	49

4 その他の部分での解析誤りの調査

はだかの体言文節，用言文節と未知語文節以外の箇所での解析誤りについて検討し，いくつかのパターンについて調査した．結果を表 3 に示す.

- パターン 1:該当する文節が，またはその直後の文節がひらがな 1 文字からなる用言文節である場合

・ 少し [のす] きも 見せてはならない。
副＊用 Ｖ＊体 体も用 Ｖ＊末
・ 灯を 再びと [もす] 必要性を･･･
体を用 副＊用 Ｖ＊体 体を用

- パターン 2 : 直前または直後が「ＶΦ用」,「体＊＊」,「未＊独」のいずれかの場合

・ お正月を [大切に] しな くちゃ､･･･
体を用 ＡΦ用 体＊＊ Ｖ引用
・ [少し] だけ 気が 休まる･･･
副＊用 ＶΦ用 体が用 Ｖ＊体

- パターン 3 : 解析誤りとはみなさなくてもよいと考えられる場合

・ 自分の 国という [感じは] しない。
体＊体 体＊体 体は用 Ｖ＊末
・ [この] 手が 指される 前の 加藤は､･･･
副＊体 体が用 Ｖ＊体 体の体 体は用

- パターン 4 : その他

表 3: その他の部分の解析誤りの分類

誤りパターン	個数
パターン 1	19
パターン 2	55
パターン 3	242
パターン 4	296
合計	624

はだかの体言文節，用言文節と未知語文節以外の個所の解析誤りの内，1/3 強 (パターン 3) が実際には解析誤りとはいえないものであった．これらを解析誤りから除き，前節の結果にパターン 1，パターン 2 を加えれば再現率は，55.7%となる.

5 おわりに

解析システム IBIKI の分析解析結果における誤り箇所について調査，検討した．はだかの体言文節，用言文節，未登録語文節で誤り箇所を検出した時の再現率は 32%，適合率は 35%であった。パターン 1，2 についても調べれば，再現率は 50%を超えるという結果が得られた．さらに高い再現率・適合率が得られるよう検討を進めたい.

参考文献

[1] 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, pp.502–504 (1999)

[2] 辞書データ主導型の自動点字翻訳システム IBUKI-TEN, 信学技報 WIT99-22, pp.131–136, (1999)

[3] 内山将夫, 形態素解析結果から過分割を検出する統計的尺度，自然言語処理 Vol.6 No.7, pp3–28(1999)

[4] 黒橋禎夫：構文情報付きテキストコーパスの作成と構文解析システムの改良，言語処理学会 第 5 回年次大会 ワークショップ論文集, pp.57–62 (1999)