複合語・文節解析誤り個所の検出

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

1　はじめに

文節解析結果の誤っている可能性のある個所を指摘することができれば，文節解析を応用するいろいろなシステムで有効に利用出来る．例えば，OCRによる文字認識の後編集で，指摘された個所だけを点検することで正確な文書を作成することが出来る．我々は自動点訳システム (IBUKI-TEN) を開発しているが，自動点訳システムにおける後処理で，指摘された個所だけを点検することで正確な点字文書を作成することが出来る．また，入力文書そのものに最初からあるタイプミスや変換誤りなどの一部をこれによって発見することが出来る．あるいは，解析システムで用いている辞書や諸規則の改善に役立てることが出来る．これらの応用のためには，もちろん誤り指摘の再現率・適合率が高いシステムを追求することが目標となる．

本稿では，我々が現在開発中の日本語解析システムIBUKIを対象として [1]，文節区切り誤りおよび複合語内の区切り誤りを検出する手法について述べる．

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には，統計的手法による試みがいくつか報告されている [3]．本報告では，文節内の表記，品詞，機能語の有無といった表層的な情報により文節区切り誤りおよび複合語内の単語区切り誤り個所を指摘する試みについて述べる．

2　文節解析システムIBUKI

IBUKIは，形態素解析だけでなく文節まとめ上げまでを行うシステムで，文節単位のコスト最小法で解析を行っている．形態素解析で一般的に用いられているコスト最小法では，個々の単語に与える単語コストと，隣接する単語の接続に対する連接コストを用いて，総コストの少ない単語列を優先解として出力する (単語単位の方法)．これに対し，我々の手法では，全ての単語および単語間にコストを与えるのではなく，文節自身および隣接する文節間にコストを与えている．

文節コストとしては，自立語コストおよび，文節タイプコストからなる．自立語コストは，単語の品詞 (品詞コスト) および単語の表記 (表記コスト) により決定する．辞書としては，EDRの日本語単語辞書をベースとしているが，表記コストは、EDR単語辞書中の単語見出しにはない語で，その仮名表記を辞書登録する場合に高いコストを与えている．文節タイプコストは，例えば，体言文節内に機能語がない場合，高いコストを与えている．

IBUKIの文節区切りの精度を調べるために，京大コーパス38,383文を用いて [4]，以下のように適合率，再現率を計算した．

$$\text{適合率} = \frac{R\_Count}{I\_Count} * 100$$

$$\text{再現率} = \frac{R\_Count}{K\_Count} * 100$$

ここで，I_Count をIBUKIの文節区切り数，K_Count を京大コーパスの文節区切り数，R_Count を文節の区切りが一致する数とする．結果を表1に示す．

表1: IBUKIの文節解析精度

I_Count	324,112
K_Count	333,653
R_Count	322,338
文節解析適合率 (%)	99.45
文節解析再現率 (%)	96.61
文節解析誤り個所	1,774

表1より，IBUKIの文節区切り数は，京大コーパスに比べると約9500箇所，少ないことが分かった．この多くは，京大コーパスでは，以下のように数量を表す単語の前で文節区切りが入れられているが，IBUKIでは，1つの文節として扱っていることによるものであった．

- 戦車 | 五十両を
- 保守 | 二党論は
- 二月 | 十一日を | めどに

3 文節区切り誤りの検出

今回は，IBUKIの文節区切りが誤っている可能性がある箇所を検出するための指摘条件について，京大コーパスとの比較により抽出された誤り1,774箇所(*L_Count-R_Count*)を分析し，以下の規則を作成した．

3.1 誤り指摘条件

(1) 独立文節，未登録語文節での誤り指摘

以下のような「不服」「逃げ切り」のように，体言，用言文節内に活用語尾を除く機能語が存在しない文節を独立文節と呼ぶことにする．独立文節が出現する場合は，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．また，「律せ，(律：名詞)(せ：未登録語)」のように，ひらがな未登録語を含む文節を未登録語文節とする．独立文節，未登録語文節の直前，直後の文節との区切りを誤り個所として指摘する．

・少し┃不服┃そうな┃口調に┃なったが
・大差を┃つけて┃逃げ切り┃態勢に┃入る。
・政治という┃ことだけでは┃律せ┃なくなった。

(2) 命令形文節，口語体文節の誤り指摘

用言命令形文節や，「って」などの口語体機能語が含まれる文節は，その文節で文が終わる可能性が高く，句読点がなく文中に存在している場合は，その文節の周辺で文節解析が誤っている可能性が高い．そこで，このような句読点を持たない文節とその直後の文節との区切りを誤り個所として指摘する．

・民間経済交流の┃進め┃方などを┃協議するもので
・対価を┃求めたわ┃いろとばかりは┃言えないとの
・考えよ┃うでは┃議院内閣制の┃根幹に

(3) 表記コストの高い文節の直前での誤り指摘

2節で述べたように，IBUKIでは，自立語の単語毎に表記コストを付与している．これは，文中に出現しにくい単語にはコストを高く設定することで，解析結果で，この単語を選ばれにくくしている．そこで，表記コストが高い単語が文節の先頭単語となる文節とその直前の文節との区切りを誤り個所として指摘する．

・怖い┃もの┃知らずの┃一年生の
・就職は┃いず┃こも┃同じ
・党大会で┃採択する┃かどう┃かなどで

(4) 漢字連続部分での指摘

漢字連続部分で2文節に分割されている個所では，漢字連続部分で1つの単語または複合語となる個所が過分割されている場合が多い．そこで，終端文字が漢字である文節で，その直後の文節の先頭文字が漢字となる場合，2文節の区切りを誤り個所として指摘する．ただし，「二百年┃待った」のように，前方の文節が数量を表す文節の場合は，2文節に分割される場合が多いので，誤り指摘を行わない．

・いつの間にか┃当然┃視され
・極端に┃低いという際┃立った┃傾向が
・今年度┃当┃初並みの┃六十三万戸と┃決まった。

(5) 終端がひらがな小文字文節での指摘

文節の終端文字が「っ」などのひらがな小文字の場合は，その直後の文節と過分割になっている可能性が高い．そこで，終端文字がひらがな小文字の文節とその直後の文節との区切りを誤り個所として指摘する．

・外出は┃おっ┃くう
・ヤツとは┃カモカの┃おっ┃ちゃん。
・勇気を┃出して┃出ちゃ┃いな┃さいよ。

(6)1文字自立語文節での指摘

1文字の自立語のみで構成される文節は，基本的に並列表現以外には出現しにくく，本来，1つの文節が過分割されて生成される可能性が高い．そこで，並列表現以外で，1文字の自立語のみで構成される文節とその直前と直後の文節との区切りを誤り個所として指摘する．

・「 | 今 | さらながら、 | 異常な | 時代だった。
・今 | シーズンが | 終わる | ころには
・ほっと | 一 | 息ついているような

(7) 文節コストの差による指摘

該当する文節と直後の文節との分割点を前後に 2 文字分の範囲でずらした時，この時の 2 文節が文節候補として存在し，かつ，その 2 文節の文節コストの合計値と元の 2 文節の文節コストの合計値との差が小さいならば，該当する文節とその直後の文節との区切りを誤り個所として指摘する．

・事情が | あるならとも | かく
・全体の | 三分の | 一前後を | 占めている。
・スタイルさえ | もとられなかったことには、

3.2 評価実験

京大コーパスを用いて，IBUKI で文節解析を行った解析結果データに対して，文節区切り誤り個所の指摘を行った．このとき，適合する誤り指摘条件が複数ある場合は，最初に適合した誤り指摘条件を選択することで，誤り指摘条件の重複を避けることとする．結果を表 2 に示す．区切り誤り個所指摘全体の適合率・再現率はそれぞれ 15.47%，53.79% であった．

また、誤り個所のうち、図 1 のように，実際には文節解析誤りとは考えなくてもよい箇所があった．このような誤り個所を数えてみたところ，全部で 515 個所存在し，誤り個所全体の約 30%近くを占めていた．そこで，このような誤り個所を正解とした時の誤り指摘条件毎の結果を表 3 に示す．この時の誤り個所指摘全体の適合率，再現率はそれぞれ，14.39%(835/5804)，69.64%(835/1199) となった．

図 1: 文節解析誤りとは考えなくてもよい個所

・大会で | 一躍有名に | なったのは、
・ディクソンの | ち密で | かつ | 大胆な
・能力の | なさは | 歴然と | してくる。

表 2: 実験結果 1

条件	誤り指摘数	実際に誤っている数	適合率	再現率
(1)	3466	591	17.05	34.48
(2)	80	23	28.75	1.34
(3)	264	45	17.05	2.63
(4)	412	62	15.05	3.62
(5)	23	11	47.83	0.64
(6)	264	38	14.39	2.22
(7)	1295	147	11.35	8.58
計	5804	917	15.80	53.50

表 3: 実験結果 2

条件	誤り指摘数	実際に誤っている数	適合率	再現率
(1)	3466	567	16.35	47.29
(2)	80	22	27.5	1.83
(3)	264	40	15.15	3.34
(4)	412	58	14.08	4.84
(5)	23	11	47.83	0.91
(6)	264	30	11.36	2.50
(7)	1295	107	8.26	8.92
計	5804	835	14.39	69.64

4 複合語内の単語区切り誤りの検出

4.1 誤り個所の抽出方法

次に．複合語内の単語区切り誤り個所の検出方法について検討した．今回は，IBUKI が抽出した文節に対して，2 つ以上の「名詞接頭語，名詞接尾語，名詞」が含まれる文節を複合語文節とし，「書き/続ける」といった動詞連続は複合語の対象としていない．以下に複合語とした箇所を示す．

村山富市首相 は | 年頭にあたり | 首相官邸で |
内閣記者会 と | 二十八日会見 し、|

京大コーパスに対して，複合語の抽出を行ったところ，延べ 92,315 語の複合語が抽出された．抽出された複合語を京大コーパスの単語分割を正解データとして，比較を行った．その結果，単語の区切りが異なる複合語は 13,999 語あった．13,999 個所を調査すると，表 4 のように，IBUKI の解析誤りとは考えなくてもよいものが多く含まれていた．

表 4: 解析誤り例

IBUKI	京大コーパス
関西空港/着	関西/空港/着
明治/維新	明治維新
ウェディング/ドレス	ウェディングドレス

そこで，今回は以下の条件にあてはまる時は誤りとしないことにした．

- IBUKI で 1 単語と解析され，京大コーパスで分割されている

- 京大コーパスで 1 単語として登録され，IBUKI では分割されている．しかし，IBUKI で分割された場合，1 文字単語が含まれる場合は誤りとする．

これに該当する複合語を誤りから除外すると，IBUKI の誤り個所は 6,858 箇所であった．

表 5: 複合語区切り誤り指摘条件

	誤り指摘条件
1	1 文字の普通名詞が連続する
2	1 文字普通名詞と 1 文字名詞接尾語が連続する
3	ひらがなが連続する
4	接尾語が連続する
5	漢字未登録語が出現する
6	「漢字+ひらがな」で構成される単語が分割されている
7	接頭語と接尾語が連続する
8	人名 (姓) と 1 文字単語が連続する
9	人名 (姓) と名詞接尾語が連続する
10	1 文字単語と人名 (名) が連続する
11	1 文字単語と人名 (姓) が連続する
12	「人名 (姓)/人名 (名) の間に 1 文字単語がある
13	地名と人名 (名) の間に 1 文字単語がある
14	人名詞と人名詞の間に他の品詞が入る
15	人名後接語の前に人名詞が出現しない
16	地名と 1 文字単語が連続する

4.2 誤り指摘条件

IBUKI の自立語辞書は，EDR 日本語単語辞書 [5] をベースに作成している．EDR 日本語単語辞書には人名などの固有名詞があまり登録されていないため，固有名詞が出現した場合，1 文字普通名詞に過分割されることが多かった．このように，誤り箇所を分析し，表 5 に示す 16 個の誤り指摘条件を作成した．

4.3 評価実験

表 5 の誤り指摘条件を用いて，京大コーパスに対する複合語解析誤り個所の指摘実験を行った．結果を表 4.3 に示す．今回の実験では，複合語区切り誤り個所指摘の全体の適合率・再現率はそれぞれ 54.52%，38.31%であった．

表 6: 実験結果

	検出数	A:誤り数	適合率	再現率 (A/6858)
1	1046	720	68.83%	10.50%
2	1123	600	53.43%	8.75%
3	83	56	67.47%	0.82%
4	116	101	87.07%	1.47%
5	975	672	68.92%	9.80%
6	562	277	49.29%	4.04%
7	133	38	28.57%	0.55%
8	113	104	92.04%	1.52%
9	32	15	46.88%	0.22%
10	185	184	99.46%	2.68%
11	118	80	67.80%	1.12%
12	18	17	94.44%	0.25%
13	10	10	100%	0.15%
14	17	16	94.12%	0.23%
15	1027	464	45.18%	6.77%
16	302	170	56.29%	2.48%
合計	4818	2627	54.52%	38.31%

5 おわりに

本論文では，我々が現在開発中の日本語解析システム IBUKI に対して，文節区切り誤りおよび複合語内の区切り誤りを検出する試みについて述べた．具体的には，京大コーパスを利用し，IBUKI の区切り誤り箇所を抽出して分析し，誤り箇所を指摘するための条件を作成した．京大コーパスを正解データとして，文節区切り誤りおよび複合語内の区切り誤り指摘に関する再現率は，それぞれ約 70%，31%であった．今後は単語の意味属性を含めた誤り指摘条件を作成し，適合率・再現率の向上を追求したい．

参考文献

[1] 兵藤, 池田，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] 横平，兵藤，早川，生川，村上，太田，池田，自動点字翻訳システム IBUKI-TEN の校正支援機能，電子情報通信学会技術研究報告，WIT00-18，pp37–42，2000.

[3] 内山, 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，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.

[5] 日本電子化辞書研究所，EDR 電子化辞書仕様説明書，1995.