日本語文解析システム ibukiC と 文節解析の曖昧さ解消および日常語テキストの解析

脇田 貴之，安藤 健二郎，太田 哲也，池田 尚志
岐阜大学工学部

1　はじめに

我々は文節の単位や構造に着目した解析器として ibukiC を開発してきた．開発途上版としてであるが公開している．[1]

ibukiC についてはこれまでも報告してきたが ([2][3])，本稿では ibukiC における係り受け解析の現状，また，”通って（とおって/かよって）”などの文節構造解析における曖昧さの解消法，ibukiC における日常語彙の収集と解析法について述べる．

2　ibukiC の概要

日本語解析システム ibukiC は形態素・文節解析システム、文節構造解析システム、構文解析システムの 3 つから構成される．(図 1)

形態素・文節解析システムでは入力文を形態素および文節に分割する．入力文から辞書引きを行い，内容語や機能語との接続などにコストを与え，Viterbi アルゴリズムで最適解を導出している．

文節構造解析システムでは各文節に文節構造（図 2）を与えている．また機能文節や複合語をサブ文節として通常の文節から分割して解析している．

構文解析システムでは文節構造解析の結果から係り受け規則を適用して係る可能性がある文節を決定していく．係り受け規則には“ 文節カテゴリ”と“ 係り先情報”だけから決定される基本的な 59 規則と，内容語や文節内要素を用いた詳細規則 183 からなる．その後，近接 3 ブロック以内の係り受けに限定するブロック化処理を行い，係り先を決定する．

3　係り受け解析における曖昧さの解消

係り受け解析では，ブロック化を行っても係り先が１つに限定されない場合もあり, 曖昧さを含んだ解析となっている.(図 3）そこで係り先が複数となる文節の半数を占める係り先情報連用の文節について表 1 の後処理を追加して係り先を絞り込んだ．京都テキストコーパス [4] を正解として, 処理追加前後の係り受け解析の精度比較結果を表 2 に示す．なお,ibukiC と文節の区切りが等しい文のみを比較対照としている．

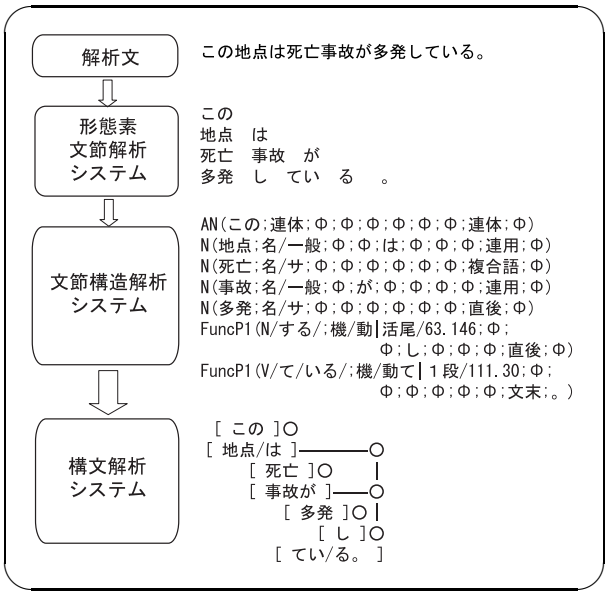

図 1: ibukiC の流れ

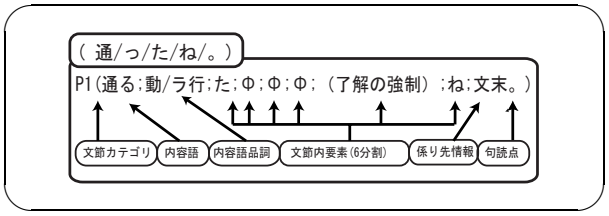

図 2: 文節構造

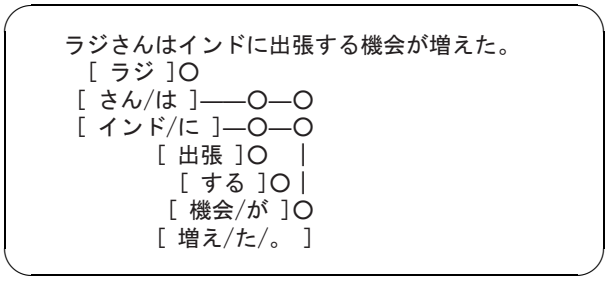

図 3: 曖昧を含んだ係り受け解析

表 1: 曖昧解消処理

- 文節カテゴリが AV（副詞文節）
  - 一番近い係り先に限定
- 「～は」という形
  - 読点を超えない最も遠い係り先に限定
- それ以外
  - 係り先情報が連体でない最初の係り先に限定
  - 全ての係り先が係り先情報連体なら
    最初の係り先に限定

表 2: 係り受け解析の精度

	曖昧解消処理 追加前	曖昧解消処理 追加後
文節区切りが 同じ文数（A）	14846 文	14846 文
A中の正解文数	8180 文 (55%)	7031 文 (47%)
A中の文節数	89850 文節	89850 文節
A中の正解文節数	77164 (86%)	74179 (83%)
曖昧文節数 (正解文節の中で)	15167 文節	7662 文節

この処理によって半数以上の曖昧な係り先を解消できたが, 中には正しい係り先を消してしまった場合もあった．また, 残っている他の係り先が複数ある文節に関しても, 今後の課題である.

4 局所的な文節 Bigram 規則による文脈依存文節の解析

4.1 はじめに

「会社に通って（かよって）」,「トンネルを通って（とおって）」などの，文節が文脈に依存している場合を正しく解析するために, 現在の文節と直前の文節の特徴の組み合わせによって文節間コストを調節し, 正しい解析を得ようと試みた.

4.2 文節間 Bigram 規則

使用する文節の特徴は, 文節のカテゴリ, 内容語の字面, 内容語の意味属性, 機能語の字面, 句読点の 5 つ. 意味属性については日本語語彙体系 [5] と ibukiC の辞書の品詞を基に付与した.

表 3: 文節 Bigram 規則の記述表

直前の文節					現在の文節					コスト
(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(6)

(1) 文節カテゴリ　(2) 内容語の字面
(3) 内容語の意味属性　(4) 機能語の字面
(5) 句読点　(6) 調節するコストの値

形態素・文節解析において,Viterbi アルゴリズムで最適解を求める際に, 文節間 Bigram 規則に記述されている組み合わせがあればコストを調節する．このように，全ての文節間のコストを考慮するのではなく，必要な場所でのみ文節間コストを操作している（局所的な文節 Bigram）ので，無駄がなく効率的である。

4.3 作成した規則

上記のように，文節の解析が文脈に依存する事例について，毎日新聞 1995 年から抜き出して分析し，規則を作成した．以下に作成した規則の例を 3 つ示す.

- **「通る/通う」の解析**

基本的には「通る」と解析し，表 4 の場合はコストを調整し「通う」と解析するようにした.

表 4: 「通る/通う」の文節 Bigram 規則

直前の文節				現在の文節					コスト
*	*	*	へ	*	*	通	*	*	-15
*	血、心	*	の	*	*	通	*	*	-15
*	*	※ 1	に	*	*	通	*	*	-15

※ 1
265(医師)400(機関)425(場所)2753(固有名詞)

- **「行く/行う」の解析**

基本的には「行う」と解析し，表 5 の場合はコストを調整し「行く」と解析するようにした.

表 5: 「行く/行う」の文節 Bigram 規則

直前の文節				現在の文節					コスト
*	*	*	へ	*	*	行	*	*	-15
*	*	※ 1	まで	*	*	行	*	*	-15
*	*	※ 2	に	*	*	行	*	*	-15

※ 1
403(立法機関)425(場所)900(建造物)2754(地名)
2837(組織名)932(家具)2647(場)2740(先・後)
※ 2
425(場所)900(建造物)2754(地名)2837(組織名)400(機関)
1260(式・行事等)1459(調査・研究)1521(読み・書き)
1532(言動)1624(食)1629(住)1637(寝起き)1690(行為)
1694(娯楽)1731(出会い・別れ)1738(招致)1743(送迎)
1746(仲介)1751(挨拶)1752(交渉・約束)1789(迎合)
1815(支配)1904(取引)1959(勤労・徒労等)1963(従業)
1973(仕事)2067(行ない)2181(流動・滑り・飛翔)2647(場)
2698(遠近)

- **「ある（動詞/連体詞）」の解析**

基本的には連体詞の「ある」と解析し，表 6 の場合はコストを調整し動詞の「ある」として解析するようにした.

表 6: 「ある（動詞/連体詞）」の文節 Bigram 規則

直前の文節					現在の文節					コスト
*	*	*	に	*	*	あ	*	る	*	-15
*	*	*	Φ	*	*	あ	*	る	*	-15

4.4 解析結果

毎日新聞 2000 年からそれぞれの事例が含まれている 300 文を抜き出してテストを行った．結果を表 7,8,9 に示す．以前はどちらか一方の解析しか出来ていなかったが，文節 Bigram 規則の記述後は全体的に解析精度が大きく向上した.

問題点として、直前の文節だけではどちらか判断できない場合，また，全ての単語に意味属性が付与されているわけではない，付与されていても意味属性が正しくないことがある，などがある.

他の形態素解析機である茶筌 [6] とこれらの文で比較すると,「ある」に関しては ibukiC の方が 37%ほど低く，それ以外では ibukiC の方が 10%〜30%ほど精度が高い結果となった.

表 7: 「通る/通う」の解析結果

正解	正しく解析できた数/総数
「通る」	106/107 (99.1%)
「〜へ通う」	6/6 (100%)
「〜の通う」	5/8 (62.5%)
「〜に通う」	80/96 (83.3%)
その他の「通う」	0/83 (0%)
計	197/300 (65.7%)

表 8: 「行く/行う」の解析結果

正解	正しく解析できた数/総数
「行う」	142/154 (92.2%)
「〜に行く」	73/85 (85.9%)
「〜へ行く」	13/13 (100%)
「〜まで行く」	1/1 (100%)
その他の「行く」	22/47 (46.8%)
計	251/300 (83.7%)

表 9: 「ある (動詞/連体詞)」の解析結果

正解	正しく解析できた数/総数
連体詞の「ある」	31/36 (86.1%)
「〜にある」	67/74 (90.5%)
「ある」	38/44 (86.4%)
その他の「ある」	48/146 (32.9%)
計	184/300 (61.3%)

5 解析誤り箇所を推定する機能と日常語彙の収集

5.1 誤り箇所推定機能

ibukiC には, 形態素・文節解析において, ヒューリスティックな規則によって誤っている可能性がある箇所を指摘する機能がある．これは, 指摘した誤り箇所を参考にして解析に使われる辞書や接続規則等の整備を行い, ibukiC の精度を向上させること，また，ibukiC の応用の一つとして我々が開発している日本語点訳システム ibukiTenC で，点訳語の後編集の際，誤りの可能性がある箇所を指摘し，後編集の効率を上げることなどを目的としている.

以下の形態素・文節解析例において, #が出力されている部分が誤り指摘箇所である.

```
（例）風変わりな客
0; #;0; 風;608388; 名/一般; かぜ
1; #;0; 変わ;610563; 動/ラ行; かわ
1; #;1; り;70338; 機/動|活尾/60.111; り
1; #;2; な;88215; 機/動|命令/117.169; な
2; Φ;0; 客;544995; 名/一般; きゃく
```

この例では, 辞書に「風変わり」という名詞が登録されておらず, 解析誤りとなっている.

現在, 新聞記事を対象とした誤り指摘機能の再現率は 70%, 精度は 80%ほどである．また, 構文解析においてブロック化で係り先が存在しなくなった場合, システムで係り先を与えているが，その場合も誤っている可能性があると出力している.

5.2 日常語彙の誤解析から誤りの分析

ibukiC では新聞などのいわゆる硬い文章に比べ, 日常会話などの表現（日常語テキスト）では解析誤りが多かった．そこで, 日常語テキストに対する誤り箇所推定機能の精度・再現率が上がるよう改良し, また, 日常語彙を辞書登録し，日常語のテキストにも十分対応できるように ibukiC の精度向上を試みた.

日常語の文章として物語文中の会話文を正しく解析することを目標に, 青空文庫 [7] の『赤毛連盟』から会話文を抜き出したものを対象として，解析誤りの分析を行い, 表 10 のように分類した.

日常会話では新聞などとは違い感動詞が多く, その中で辞書に登録されていない語彙が見られた．また, 従来の ibukiC では音韻的な変化, 敬語表現などに対する語彙が少なく, 誤解析が多くなってしまっていた. その他に, 複合語の問題などコストや接続規則による誤りも見られた.

この分析から誤り箇所推定機能を改良すると共に, ibukiC の辞書に足りない語彙を見つけ, 辞書登録を行い, 日常語に対する解析精度を向上させた.

5.3 誤り箇所推定の改良

前節の分析を元に誤り箇所推定機能を表 11 のように改良し, 日常語テキストにおける誤解析の幾つかを ibukiC で指摘できるようにした.

表 11: 誤り指摘の改良点

- 会話文の文頭（「の後）且つ、ひらがな又はカタカナ＋句読点で構成された文節が感動詞又は接続詞でないときに指摘
- 「で」の後に続く文節の先頭がひらがなで、文節間に切れ目（句読点）がないときに指摘
- ひらがな未知語、漢字未知語を含む複合語のときに指摘
- ひらがな 1 文字動詞の連用中止形＋読点のときに指摘

これらの条件を満たしていても, 誤解析が少ない場合があり, その場合は細かい条件を付けることによって指摘しないようにしている.

この改良で前節の誤解析の幾つかをシステムで指摘できるようにしたが, 複合語の区切り違いなど, 指摘することが難しいものが残っている。今後はさらに多くの日常語テキストを解析し, 誤解析を分析して残りの誤解析を指摘できるような案を考えていく予定である.

表 10: 解析誤りの分析

- **感動詞の誤り**

  ふむ, では〜
  誤：動詞「ふむ」 → 正：感動詞「ふむ」
  いや, 実に〜
  誤：形容動詞「いや」 → 正：感動詞「いや」
  え えと, それは〜
  誤：動詞「える」, 名詞「えと」
  → 正：感動詞「ええと」
  ほら, 私が〜
  誤：名詞「ほら」 → 正：感動詞「ほら」

- **会話独特の表現**

  することなんて ありゃ しませんし.
  誤：感動詞「ありゃ」
  → 正：動詞「ありゃしません」
  来てくれ とるんです.
  誤：動詞「来てくれ」, 動詞「とる」
  → 正：動詞「来てくれとる」
  思っ とった.
  誤：名詞「思」＋未知語「っ」, 動詞「とる」
  → 正：動詞「思っとった」
  役に立つかも しれんので,
  誤：名詞「しれん」
  → 正：機能語「かもしれん」

- **敬語による誤り**

  半額だったわけで すね.
  誤：名詞「すね」
  → 正：機能語「ですね」
  仕事を していら したんですから,
  誤：名詞「していら」, 動詞「した」
  → 正：動詞「していらした」
  ご存じ ないんですか？
  誤：形容詞「ない」 → 正：機能語「ない」

- **複合語と誤解析**

  何千もの事件 の例を
  誤：名詞「何千」＋名詞「もの」＋名詞「事件」の複合語
  → 正：名詞「何千」＋機能語「もの」, 名詞「事件」

- **複合語の区切り違い**

  右手首 のすぐ上に
  誤：名詞「右手」＋名詞「首」の複合語
  → 正：名詞「右」＋名詞「手首」の複合語

5.4 日常語彙の辞書登録と解析

以上のような分析を基に，内容語 216 単語，機能語 99 単語を新たに登録して ibukiC の辞書を日常語用に拡張した．

以下に，例文を通常の ibukiC と日常語用の ibukiC で解析した結果をそれぞれ示す．

例文：扉が閉まって開かんのですわ。

- 通常の ibukiC の文節構造解析結果

1;1;2;0; 開/かん/の/です/わ/。;N; 開; 名/個/地名/駅名; Φ; Φ; Φ; Φ; Φ; Φ; 複合語; Φ
1;1;2;1; 機能文節;N; かん; 名/一般; Φ; Φ; Φ; Φ; Φ; Φ; 直後; Φ
1;1;2;2; 機能文節;FuncSN;N/の/; 機/名—ノ (中間)/104.181; Φ; Φ; Φ; Φ; Φ; Φ; 直後; Φ
1;1;2;3; 機能文節;FuncP1;N/です/; 機/名—ダ 文末/102.164; Φ; Φ; Φ; Φ;（宣言）; わ; 文末;。

- 日常語用の ibukiC の文節構造解析結果

1;1;0;0;開/か/ん/の/です/わ;P1;開く;動/カ行;ない; Φ; Φ; Φ; Φ; Φ; 直後; Φ
1;1;0;1; 機能文節;FuncP1;Y のだ; 機/用ノダ—文末/140.164; Φ; Φ; Φ; Φ;（断定）/（宣言）; わ; 文末; Φ

青空文庫『阿亀（おかめ）』でテストした結果，精度は文節単位で 91%と以前と比べて 5%向上した．今のところ，これらの追加単語などによって新聞などでの解析への影響は出ていないが，通常の辞書とは別に日常語用の専用辞書として分離することも検討している．

6 おわりに

ibukiC の係り受け解析の現状，文脈依存文節や日常語彙への対応について述べた．ibukiC は開発途上版としてではあるが公開している．今後も ibukiC 発展を目指し，開発・整備を続けていく．

参考文献

[1] 岐阜大学 池田研究室：http://ikd.info.gifu-u.ac.jp/
[2] 池田, 脇田, 大口，機能文節を導入した文節構造解析システム ibukiC(v0.20) について，言語処理学会 第 14 回年次大会, 2008
[3] 山田，脇田，大口，池田，文節構造解析システム ibukiC の解析仕様および精度の比較と評価，言語処理学会 第 13 回年次大会, 2007
[4] 京都テキストコーパス Version 4.0：http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html
[5] NTT コミュニケーション科学基礎研究所 監修, 池原 悟, 宮崎 正弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良彦 編集：日本語語彙体系：岩波書店 (1997, 1999)
[6] 形態素解析器 茶筌 Version 2.4.2：http://chasen-legacy.sourceforge.jp/
[7] 青空文庫：http://www.aozora.gr.jp/