日本語解析システム ibukiC の誤解析の分析と改良の試み

原 裕樹　　松本 忠博
岐阜大学大学院 工学研究科
{hara, tad}@mat.info.gifu-u.ac.jp

1. はじめに

ibukiC は岐阜大学池田研究室で開発された日本語解析システムである [4]．このシステムは，長単位の機能語，機能文節 [2] などの導入などの特徴を持っている．自動点訳システム ibukiTenC[5] や日中機械翻訳システム jaw/Chinese[3] などの入力日本語解析部としても利用されており，ibukiC の解析精度はこれら応用システムの翻訳精度にも直接影響する．ibukiC は日本語文に対する形態素・文節解析，文節構造解析，係り受け解析の機能を持つ．本研究では，京都大学テキストコーパス [6] を正解データとして，係り受け解析における誤解析を収集・分析することで，誤解析の原因を調査した．また，その調査を元に ibukiC の改良を行い，ibukiC の解析精度を向上させることを試みた．

2. ibukiC による解析の流れ

ibukiC による解析の概要を簡単に説明する．入力となる日本語文に対し，まず形態素・文節解析が施され，形態素・文節単位での分割とそれらに関する言語情報の付加がされる．次に，それらの情報を基に文節構造解析がされる．これにより，文節の機能語部が意味的・機能的な観点から 6 つの機能語要素に分割される．さらに，文節構造解析の解析結果を基に，文節間の修飾関係である係り受けの解析がされる．

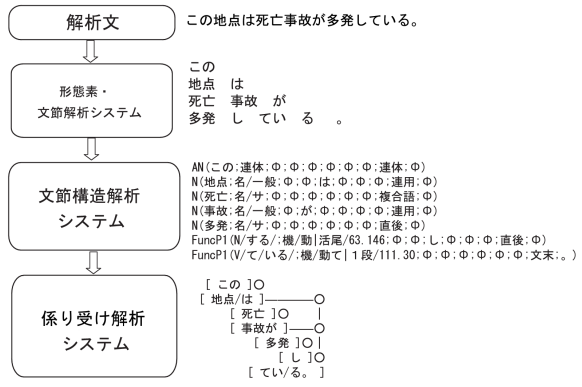

図 1　ibukiC の概要

3. 係り受け解析における誤解析の分析

ibukiC の係り受け解析における誤解析を検出するためのソフトウェアを作成し，機能語などの要素や，処理過程における精度について分析することで誤解析の原因を調査した．この時，ibukiC の誤解析を検出するための正解データとして京都大学テキストコーパスを用いた．京都大学テキストコーパスに含まれる文章と同じ文章を ibukiC で解析し，比較した．京都大学テキストコーパスは，京都大学によって開発された，文節分割，品詞，係り受けなどの言語情報が付加されたテキストコーパスである．自動解析により解析された文章が，人手により修正されたものである．なお，ibukiC と京都大学テキストコーパスには形態素・文節・係り受けの扱いで以下のような違いがある．

1. 辞書の違いによる形態素単位の分割
2. 機能語，複合語などによる文節分割
3. 鍵括弧内の読点による文の分割
4. 鍵括弧による文節分割
5. 機能文節などの特殊な文節による分割
6. 係り先の文節を複数持つか

１～３について，これらの相違点を含む場合係り先を正しく比較することができなくなるため，これらを含む文節は除外することとした．４の相違点は，鍵括弧を１つの文節として扱うかどうかによる違いである．ibukiC では鍵括弧を独立した文節としているが，これを京都大学テキストコーパスと同じ形になるよう前処理をすることで比較ができるようにした．５の相違点は，ibukiC 側一部の機能語や複合語が独立した文節として扱われることにより発生する．そのため，比較の際に ibukiC 側のそれらの文節を再び結合することで比較ができるようにした．６の相違点は，ibukiC において係り先が複数与えられる性質によるものである．これについては，複数ある係り先の内１つでも正解の係り先と一致すれば，係り先が正解であると判定した．

3.1　誤解析検出の流れ

誤解析の検出における処理の流れを図 2 に示す．図中の①～⑤の手順により誤解析を収集する．この処理において比較を行えるのは，ibukiC と京都大学テキストコーパスで文節分割が等しい文中にある文節だけとなる．また，対象となる文節は 16020 文中の 113353 文節となっている．

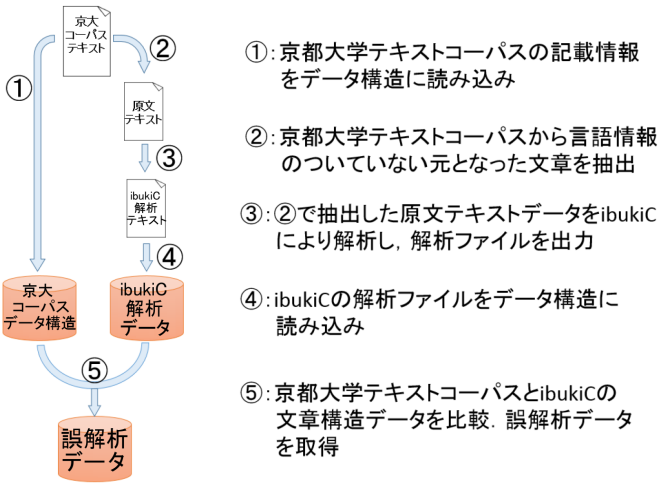

図 2　誤解析検出の流れ

3.2　係る文節の機能語による精度検証

3.1 項で収集した誤解析データを使い，機能語の種類ごとの正解率を調べた．頻繁に現れる機能語の一部と，その解析精度を表 1 に示す．機能語ごとの解析精度については，際立って低いというものは見当たらなかった．

表 1　機能語ごとの解析精度の一部

機能語	出現文節数	係り先正解文節	正解率 (%)
の	16440	15249	92.75
を	9859	8927	90.54
は	9076	7210	79.44
に	8011	7057	88.09
が	7552	6594	87.31

3.3　係る文節の機能文節による精度検証

ibukiC は，特定の機能語を機能文節という独立した文節として扱う．前節と同様に，この機能文節についても種類ごとの精度を調査した．頻繁に現れる機能文節の種類とその解析精度は表 2 のようになった．機能文節においては，全体的に高い精度が出ているという結果が得られた．

表 2　機能文節ごとの解析精度の一部

機能文節	出現文節数	係り先正解文節	正解率 (%)
V ている	2001	1911	95.50
N だ	1251	1192	95.28
N する	1023	965	94.33
Q と	974	849	87.16
Q」と	798	726	90.97

3.4　係り受け解析の処理過程における精度検証

ibukiC の係り受け解析における処理を大まかに分けると，図 3 のようになっている．この①～④それぞれの段階について，係り受けの精度を調査することで，どの処理が精度に影響を与えているのかを検証した．各処理過程における精度は表 3 のようになった．なお，この調査で利用した文節数は 113353 文節である．

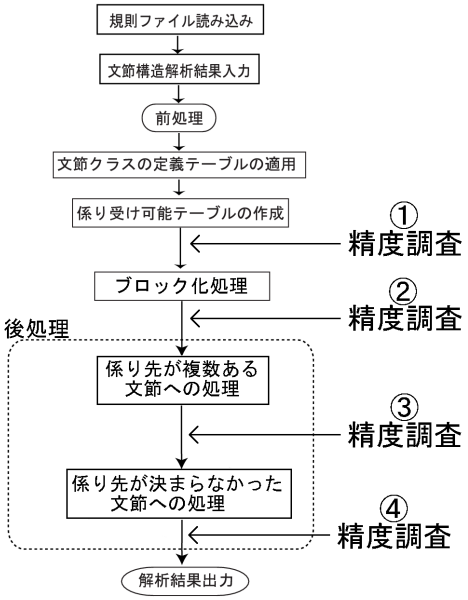

図 3　係り受け解析処理の精度調査

表 3　係り受け解析処理の段階ごとの精度

段階	係り先のある文節数	係り先正解文節数	正解率 (%)
①	37519	34169	91.07
②	86128	77678	90.18
③	90773	78247	86.20
④	97333	82270	84.52

4. ibukiC の改良と評価

前節の分析結果を参考に，ibukiC の改良を試みた．またそれによる精度の推移についての評価を行った．

4.1 「係り先のない文節に対する処理」における係り先情報が「連体」の文節に対する処理

図 3 の調査結果から,「係り先のない文節への処理」の精度が低くなっていることが分かった．また，この中でも特に，文節構造解析で付加されるデータの 1 つである「係り先情報」が「連体」となっている文節の精度が低いことがわかった．正解データ側の傾向では，直後の文節にかかることが多かったため，そのようにルールを書き換えたところ，精度が表 4 のように向上した．また，システム全体としては精度が表 5 のように向上した．

表 4　係り先が決まらず，係り先情報「連体」の文節の精度変化

	対象の文節	係り先正解文節数	正解率 (%)
改良前	949	356	37.51
改良後	949	774	81.55

表 5　係り受け解析全体の精度変化

	評価対象の文節	係り先正解文節数	正解率 (%)
改良前	113353	98290	86.71
改良後	113353	98719	87.08

4.2 「係り先のない文節に対する処理」における係り先情報が「独体」の文節の処理

前項と同様に,「係り先のない文節への処理」における「係り先情報」が「独体」となっている文節の精度が低くなっていることが分かった．こちらも正解データ側の傾向では，直後の文節にかかることが多かったため，そのようにルールを書き換えた．それにより，精度が表 6 のように向上した．また，システム全体としては精度が表 7 のように向上した．

表 6　係り先が決まらず，係り先情報「独体」の文節の精度変化

	対象の文節	係り先正解文節数	正解率 (%)
改良前	225	63	28.00
改良後	225	163	72.44

表 7　係り受け解析全体の精度変化

	評価対象の文節	係り先正解文節数	正解率 (%)
改良前	113353	98719	87.08
改良後	113353	98821	87.17

4.3 京都大学テキストコーパスを利用したルールの書き換え

ibukiC はルールベースの解析システムである．文節のカテゴリ，自立語，品詞，係り先情報，機能語要素，句読点の組を文節定義とし，文節ごとにいくつかの文節定義を与える．その文節定義ごとで係るか係らないかを判別している (図 4 参照)．また，このルールは主に図 3 における「係り受け可能テーブルの作成」において利用されている．

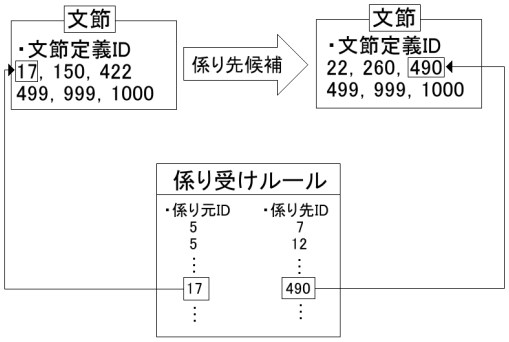

図 4　文節定義の役割

しかし，表 3 の結果から，ibukiC の「係り受け可能テーブルの作成」では，それほど多くの係り先を決定できてはいないようであった．そこで，この文節定義ごとの係りやすさを京都大学テキストコーパスの正解データを基に調べ，係りやすい文節定義同士のルールを加える事で,「係り受けテーブルの作成」処理における係り先の付加率と精度を向上されることができるのではないかと考えた．また，この調査の手法については文献 [1] を参考にしている．係りやすい分類を調査する上でのデータの取得には，京都大学テキストコーパスに含まれる比較が可能な 16020 文，それらに含まれる 113353 文節を用いた．この中から，文節定義ごとの係り元と係り先の組で，文章中の出現回数と正解率がある程度あるものを調査した．表 8 にその一例を示す．なお，この表は文節定義の要素の一部を省いており，実際にはこの項の始めに述べただけの要素がある．要素が＊となっているものはその要素が何でも該当することを表している．

表 8　係り受けルール (一部省略) に追記した文節定義の例

	分類 ID	カテゴリ	品詞	係り先情報	機能語要素	句読点
係り元	17	N	、/引用	*	*	Φ
係り先	990	*	、/引用/49.141	*	*	、

こういった文節定義でみた係り元と係り先の組が，京都大学テキストコーパス全体で 500 組以上出現しており，かつ係り受けの正解率 94 %以上あるものをルールに追加した．元々のルールの数は 18321 組あり，今回の追記により新たに 12 組のルールを追記した．その結果，精度が表 10 のようになった．

表 9　係り受けルール追記前の精度

	係り先がある文節	係り先正解文節	正解率
図 3①の段階	37519	34169	91.07
システム全体	97333	82801	85.06

表 10　係り受けルール追記後の精度

	係り先がある文節	係り先正解文節	正解率
図 3①の段階	38527	35437	91.97
システム全体	97333	82952	85.22

これにより，「係り受けテーブルの作成」の段階で 1008 文節が新たに係り受けを得られるようになり，その段階での精度も 0.9 %とわずかではあるが向上した．また，係り受け解析全体においては精度が 0.16 %向上する結果となった．このルールの改変では，わずかながらに精度を向上させることができたが，京都大学テキストコーパスに合わせてルールを変えたため，京都大学テキストコーパスを用いて精度を検証すれば高くなりやすくなるものと推測される．より正確に精度が向上しているかを検証するには，他の正解データを基準にして精度を測定する必要があると思われる．

5.　おわりに

ibukiC の係り受け解析に関して，機能語ごとの解析精度や解析プロセスごとの解析精度を調査した．機能語における調査からは大きな誤解析の原因を見つけることは出来なかったが，解析プロセスにおいては「係り先のない文節への処理」において解析精度が低くなっていることが分かった．これより「係り先のない文節への処理」の改良を試みた結果，その処理における精度が向上した．係り受け解析全体においても，解析精度をわずかながらに向上させることができた．また，京都大学テキストコーパスの，文節定義ごとの統計データを利用することで「係り受けテーブル作成」段階での精度を向上させることができた．本研究による ibukiC の精度は，着手前と後で表 11 のように向上した．

表 11　研究による精度変化

	評価対象の文節	係り先正解文節数	正解率 (%)
着手前	113353	98290	86.71
着手後	113353	98972	87.31

参考文献

[1] 阿辺川 武 , 奥村学 : 大規模統計情報を用いた日本語係り受け解析の精度向上 , 言語処理学会第 11 回年次大会, pp.919–922 , 2005

[2] 池田 尚志 , 脇田 貴之 , 大口 智也 : 機能文節を導入した文節構造解析システム (ibukiC_v0.20) , LACE2007（第 12 回「言語認識表現」研究会）, pp.1〜14，2007

[3] 池田 尚志 :『日本語からアジア諸言語への機械翻訳システムの構築奮闘記』日本語学，28(12), pp.62–71, 2009

[4] 池田　尚志 : 日本語解析システム　ibukiC, http://www.ikd.info.gifu-u.ac.jp/ibukiC/about_ibukiC.html

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

[6] 黒橋禎夫, 長尾眞 : 京都大学テキストコーパス・プロジェクト, 言語処理学会 第 3 回年次大会 , pp.115–118, 1997