ibukiC の機能文節解析と機能語/内容語の曖昧さ解消

高田 和典，脇田 貴之，池田尚志
岐阜大学工学部

1　はじめに

日本語には「にあたって」や「に応じて」のように，複数の語から成りながら，一つの機能的な意味を持つ複合辞が存在している。これを機能表現という。また，機能表現と同一の表記でありながら，内容的な意味を持つ表現が存在する。機能表現を正しく検出するためには，そうした表現が機能的に用いられているか，内容的に用いられているかの曖昧さを解消する必要がある。

本研究は，我々の研究室で開発している日本語解析システム ibukiC[1] において，機能表現の曖昧性を解消することによって，解析精度を向上させることを目的とし，ibukiC の辞書 (以下 ibukiDic) に登録された曖昧な機能表現の抽出，また局所的文節 Bigram を用いた機能表現の曖昧性解消に関する考察を行った。

2　曖昧な機能表現の抽出

機能表現の曖昧性解消に取り組むにあたり，対象とする機能表現を，ibukiDic に登録されている機能語 4539 語の中から抽出する作業を行った。まず，日本語複合辞用例データベース v1.0[2](以下 MUST1) にもリストされている語で，その用例中に機能的用法と内容的用法が混在している語 (128 単語) を曖昧な機能表現として抽出した。さらに，MUST1 にリストされていない語についても，独自に判断して，94 単語を曖昧な機能表現とした。

3　局所的文節 Bigram による曖昧性解消

機能表現の曖昧性を解消する手法として，局所的文節 Bigram による処理を用いた。ここでは，ibukiC の局所的文節 Bigram に関して説明し，局所的文節 Bigram を用いた機能表現の曖昧性解消について述べる。

3.1　局所的文節 Bigram による文節・形態素解析

ibukiC では，文節解析において，文節間に設定されたコスト (以下 文節間コスト) を判断材料の一つとして文節を決定するメカニズムを備えている。しかし，これまで文節間コストは全てに一定値を与えており，実際には用いていなかった。今回はこのメカニズムを利用して，ある単語に対してコストの調整を行うかどうかの条件を定義する規則 (以下 局所的文節 Bigram 規則) を作成し，文節間のコスト調整を行った。

局所的文節 Bigram 規則では，まず処理の対象とする単語と，その単語が現れた場合に適用する規則を定義する。規則の内容は，現在の文節およびその直前の文節の情報 (文節カテゴリ，字面，意味属性，句読点) を条件として，文節間コストを定義するものである。処理の例を図 1 に示す。

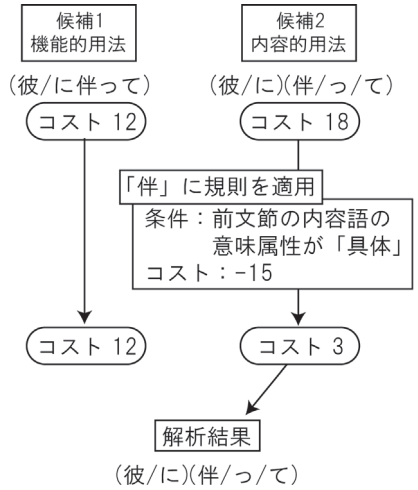

図 1: 局所的文節 Bigram を用いた解析

3.2　意味属性を用いた局所的文節 Bigram 規則

機能表現の曖昧性解消のために，局所的文節 Bigram の条件として意味属性を用いた。そのために行った辞書の整備と，局所的文節 Bigram の条件の記述について述べる。

3.2.1　ibukiDic への意味属性付与

局所的文節 Bigram の条件として意味属性を用いるため，ibukiDic に意味属性の付与を行った。意味属性を付与する対象は、ibukiDic に登録されている名詞および接尾辞とし、意味属性は日本語語彙大系 [3] において定義されたものを実験的に用いた。意味属性の

付与は，次に示す方法で行った。

(1) ibukiDic の内容語の見出しと日本語語彙大系に登録されている単語の見出しを比較し，一致した単語の意味属性を付与する。

(2) 接尾辞として意味属性が定義されている単語は，ibukiDic の内容語の見出しの最後の文字から比較し，一致した単語の意味属性を付与する。

(3) ibukiDic の品詞から意味属性を特定できる場合，該当する品詞の単語に，適切な意味属性を付与する。

上記の (3) において用いた、ibukiDic に登録された品詞と意味属性の対応の一部を表 1 に示す。なお「名/一般」「名/サ」といった品詞に対しては、品詞から意味属性を特定できないため、(3) における意味属性付与は行われていない。

表 1: 品詞と意味属性の対応

品詞名	意味属性
名/一般	(該当属性無し)
名/人	人
名/個/人名/姓	姓
名/個/人名/名	名
名/個/施設名	施設名
名/個/自然名	自然名
名/個/地名/一般	地域名
名/個/年号	年号
名/サ	(該当属性無し)
名/代/人	人間（人称）

3.2.2 用例による意味属性の傾向の分析

MUST1 には，一つの機能表現に対して最大で 50 の用例が登録されている。また各用例は，その用例において機能表現がどのような用法で用いられているかを意味するラベルを持っている。それらの情報を用いて，機能表現の直前の単語の意味属性と，機能表現の用法との間の関係の分析を行った。

まず，各用例の機能表現の直前の単語と，その単語が持つ意味属性を取得する。例として、機能表現「に応じて」の用例を対象とした処理によって得られた意味属性と，その意味属性を持つ単語数を表 2 に示す。なお、「に応じて」の用例は、機能的用法が 35 用例、内容的用法が 11 用例である。また、単語が複数の意味属性を持つ場合がある。

表 2: 「に応じて」直前の単語の意味属性

機能的用法		内容的用法	
意味属性	単語数	意味属性	単語数
事情	8	希望	4
程度	7	求め	4
等級	5	買い	3
立場	5	情報	2
状況	5	大字 (その他)	2
限度	4	打合せ	2
是	4	会談	2
類型	3	会議	2
値・額	2	案内	1
願い	2	願い	1
⋮	⋮	⋮	⋮

次に，意味属性の親子関係を考慮して、意味属性に属する単語数を集計する。各意味属性は、その先祖にあたる全ての意味属性に属しているとする。上で挙げた「に応じて」の用例を集計した結果の一部を、図 2 に示す。括弧内の数字は、その意味属性に属している単語数である。

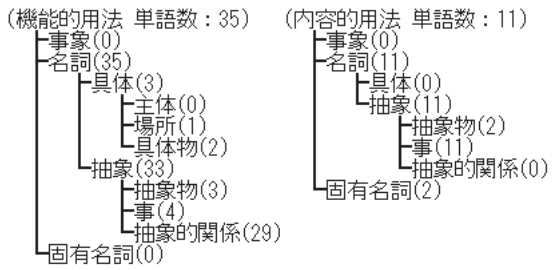

図 2: 「に応じて」親子関係を考慮した意味属性

3.2.3 MUST1 に登録されていない表現の用例作成

ibukiDic より抽出した曖昧な機能表現のうち，94 単語は MUST1 に登録されていない機能表現である。これらの機能表現についても分析を行うために，新聞記事を元に用例の作成を行った。用例作成の手順を以下に示す。

(1) 毎日新聞の記事 1 年分から，対象となる機能表現を含む文を抽出する。

(2) 各機能表現に対して，最大で 50 文を無作為に抽出し用例とする。

(3) 各用例における機能表現の用法を人手で判断し、ラベルを付与する。

(1) の処理により，新聞記事約 1,500,000 文から 18,761 文が抽出され，(2) の処理により，用例とする文は 1,698 文となった。

ここで作成した用例に対して，3.2.2 の処理を行い，同様に意味属性についての分析を行った。

3.2.4 用例の分析をもとにした局所的文節 Bigram 規則

3.2.2 および 3.2.3 で得られた情報を用いて，局所的文節 Bigram 規則の条件を記述した。その一例として，機能表現「に応じて」に関する規則を表 3 および表 4 に示す。

表 3: 「に応じて」文節候補

例	候補	用法
に応じて	(に応じて)	機能的用法
	(に)(応じ/て)	内容的用法

表 4: 「に応じて」局所的文節 Bigram 規則

前文節		自文節	コスト
内容語	機能語	内容語	
意味属性	字面	字面	
行為	に	応じ	-30

表 3 に示すように，機能表現「に応じて」には，同一の表記でありながら内容的な意味を持つ場合がある。「金額に応じて」「増加に応じて」などが機能的用法であるのに対して,「呼びかけに応じて」「要求に応じて」などは内容的用法となる。ここで,「に応じて」の持つ曖昧性を解消するために，その直前の単語に着目する。用例を用いた分析結果において,「に応じて」が内容的用法で用いられる場合，直前の単語が高い確率で意味属性「行為」に属していた。また，機能的用法で用いられる場合，直前の単語が意味属性「行為」に含まれる確率は低いということが分かった。これらの傾向から,「に応じて」の直前の単語の意味属性が「行為」に属していれば内容的用法として，属していなければ機能的用法として解析することによって,「に応じて」の曖昧性の解消につながると考えられる。表 4 では，直前の単語の意味属性が「行為」に含まれる場合，内容的用法としての解析のコストを下げるという規則を記述している。

3.2.5 局所的文節 Bigram 規則の記述

前項の例をはじめとし，用例の分析結果をもとにして，25 種類の機能表現に対して局所的文節 Bigram 規則の記述を行った。記述した規則の一部を表 5 に挙げる。

表 5: 局所的文節 Bigram 規則

前文節		自文節		コスト
内容語	機能語	内容語	機能語	
意味属性	字面	字面	字面	
具体 | 交際	に	当た	って | り	-30
行為	に	応じ	*	-30
具体 | 事 | 抽象物	に	かけ	*	-30
具体	を	も	って | ちまして	-30
*	に	従	い |って	-30
変動	に	従	い |って	60
*	に	よ	らず*	-30
主体 | 人名	から	言	*	-30
主体 | 人名 | 組織名 | 場所 | 地名	から	見	*	-30
時間	の	末		-30
時間 | 要点	を	お	いて*	-30

4 評価実験

4.1 評価の方法

(1) 3.2.3 と同様の方法で，新聞記事から用例集を作成する。(毎日新聞の記事 3 年分から機能表現を含む文を抽出し，各用例に対して最大 50 文にラベルを付与し，用例とする。)

(2) 用例を ibukiC で解析し，機能表現の解析結果がラベルの用法と同じであれば正解とする。

(3) 局所的文節 Bigram 規則適用前と適用後で，正解率を各機能表現ごとにまとめ，評価する。

4.2 評価実験

評価結果の一部を表 6 に示す。

4.3 誤りに関する考察

4.3.1 機能的用法の用例の割合が多い場合

ibukiC では局所的文節 Bigram 規則を適用しなかった場合，基本的には機能的用法のコストが低くなる。そのため，機能的用法の用例の方が多く登録されている機能表現では，わずかな解析誤りでも，局所的文節 Bigram 規則適用後の方が正解率が低くなる場合がある。

表 6: 評価結果 (局所的 Bigram 規則適用前後比較)

見出し	正解数 (正解率)					
	局所的文節 Bigram 規則適用前			局所的文節 Bigram 規則適用後		
	機能的用法	内容的用法	全体	機能的用法	内容的用法	全体
にあたって	20/20(100%)	2/18(11%)	22/38(58%)	18/20(90%)	16/18(89%)	34/38(89%)
に応じて	36/37(97%)	3/12(25%)	39/49(80%)	33/37(89%)	11/12(92%)	44/49(90%)
にかけて	43/43(100%)	1/3(33%)	44/46(96%)	41/43(95%)	3/3(100%)	44/46(96%)
にしたがって	7/7(100%)	1/43(2%)	8/50(16%)	6/7(86%)	43/43(100%)	49/50(98%)
によらず	13/14(93%)	4/24(17%)	17/38(45%)	0/14(0%)	24/24(100%)	24/38(63%)
をもって	17/17(100%)	1/26(4%)	18/43(42%)	13/17(76%)	6/26(23%)	19/43(44%)
からいうと	31/31(100%)	0/19(0%)	31/50(62%)	29/31(94%)	3/19(16%)	32/50(64%)
からみれば	12/12(100%)	0/38(0%)	12/50(24%)	9/12(75%)	31/38(82%)	40/50(80%)
の末に	45/45(100%)	0/2(0%)	45/47(96%)	44/45(98%)	0/2(0%)	44/47(94%)
をおいて	10/16(63%)	3/32(9%)	13/48(27%)	6/16(38%)	21/32(66%)	27/48(56%)

表 7: 評価結果 (茶筅，Ko-BaKo/J)

見出し	正解数 (正解率)					
	茶筅			Ko-BaKo/J		
	機能的用法	内容的用法	全体	機能的用法	内容的用法	全体
にあたって	20/20(100%)	16/18(89%)	36/38(95%)	0/20(0%)	18/18(100%)	18/38(47%)
に応じて	0/37(0%)	12/12(100%)	12/49(24%)	3/37(8%)	12/12(100%)	15/49(31%)
にかけて	42/43(98%)	2/3(67%)	44/46(96%)	0/43(0%)	3/3(100%)	3/46(7%)
にしたがって	5/7(71%)	5/43(12%)	10/50(20%)	0/7(0%)	43/43(100%)	43/50(86%)
によらず	0/14(0%)	24/24(100%)	24/38(63%)	14/14(100%)	0/24(0%)	14/38(37%)
をもって	9/17(53%)	16/26(62%)	25/43(58%)	0/17(0%)	26/26(100%)	26/43(60%)

4.3.2 複数の意味属性を持つ単語による誤り

単語が複数の意味属性を持っていても，用例中でどの意味として単語が用いられているかまでは，解析することができない。

4.3.3 ibukiDic への意味属性付与の誤り

ibukiDic に誤った意味属性を付与している場合がある。例えば,「〜高」という接尾辞をもつ単語に「学校」という意味属性を付与する際，単語を字面で比較しているため,「輸出高」といった単語にも「学校」という意味属性が付与されてしまっている。

4.4 その他のシステムとの比較

茶筅 [4]，Ko-BaKo/J[5] によって，4.2 で使用した用例集を同様に解析し，評価を行った。なお．評価対象の機能表現は，MUST1 に登録されている表現のみとし，ibukiDic から独自に抽出した曖昧な機能表現は評価対象から除外した。評価結果を表 7 に示す。

5 おわりに

局所的文節 Bigram を用いた機能表現の曖昧性解消のために，意味属性の導入と，局所的文節 Bigram 規則の記述を行った。評価実験の結果，多くの機能表現において，精度の向上が見られた。しかし他システムとの比較においては，機能表現によってバラつきはあるが，局所的文節 Bigram 規則適用後も，ibukiC の正解率の低さが目立つものがあった。

4.3.2 で挙げた問題に配慮した局所的文節 Bigram 規則の記述や，4.3.3 の問題に対処するためにより正確な意味属性の付与を行うことで，さらなる精度向上を目指したい。

参考文献

[1] 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) について：言語処理学会 第 14 回年次大会 (2008)

[2] 松吉俊，宇津呂武仁，佐藤理史，土屋雅稔：日本語複合辞用例データベース v1.0

[3] 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)

[4] 茶筌 v2.1 for Windows
http://chasen.aist-nara.ac.jp/index.html.ja

[5] Ko-BaKo/J
http://www.jsa.co.jp/index.html