接続情報にもとづく助詞型機能表現の自動検出

松吉　俊　佐藤理史　宇津呂武仁

京都大学大学院 情報学研究科
matuyosi@pine.kuee.kyoto-u.ac.jp, {sato, utsuro}@i.kyoto-u.ac.jp

1. はじめに

日本語には、例えば、「に関して」、「からには」のように、2 形態素以上から構成され、全体として 1 つの機能語のように働く**機能表現**が存在する。かなりの数の機能表現に対して、それと同一表記をとる内容表現が存在する。例えば、「をめぐり」という表現は、「貿易をめぐり対立する」という文では「について」に相当する機能表現であるが、「四国をめぐり旅をする」という文では内容表現である。このため、このような機能表現は、内容表現と区別して認識する必要がある。

しかしながら、現在の自然言語処理においては、機能表現の認識は不十分である。JUMAN[1)]・KNP[2)] は、KNP の係り受け規則の一部に機能表現の情報を使用しているが、機能表現と認識した場合も、内容表現と認識した場合と全く同じ係り受け構造を出力する。一方、ChaSen[3)]・CaboCha[4)] は、ChaSen の辞書に「によって」、「に当たって」など 90 語を「助詞-格助詞-連語」として登録することにより、機能表現の認識を行なっているが、その認識結果はそれほど信頼できるものではない。例えば、文 (A) を ChaSen・CaboCha で解析した結果を図 1 に示す。

(A)　下からの光が、天井に当たって反射している。

この文の「に当たって」は内容表現であるが、機能表現であると誤認識されている。このような出力は、後続の応用システムに致命的な誤りをもたらす可能性がある。

機能表現を正しく認識するためには、次の 2 つのことが必要である。

(1) 認識すべき機能表現の網羅的リストを作成すること。

(2) それぞれの機能表現の表記に対して、機能表現として認識すべき場合と、そうでない場合を区別する能力をもった、機能表現の検出システムを実現すること。

本論文では、(2) の機能表現を検出するシステムについて報告する。まず、第 2 章で、助詞型機能表現について述べ、機能表現の適切な判定時期を示す判定指標について説明する。次に、第 3 章において、接続情報にもとづく、助詞型機能表現の検出システムについて説明する。そして、第 4 章で、このシステムの評価を行なう。最後に、第 5 章でまとめを述べる。

```
* 0 1D 0/2 1.80152089
下 シタ 下 名詞-一般 O
から カラ から 助詞-格助詞-一般 O
の ノ の 助詞-連体化 O
* 1 3D 0/1 4.30264170
光 ヒカリ 光 名詞-一般 O
が ガ が 助詞-格助詞-一般 O
、 、 、 記号-読点 O
* 2 3D 0/1 0.00000000
天井 テンジョウ 天井 名詞-一般 O
に当たって ニアタッテ に当たって 助詞-格助詞-連語 O
* 3 -1D 1/3 0.00000000
反射 ハンシャ 反射 名詞-サ変接続 O
し シ する 動詞-自立 サ変・スル 連用形 O
て テ て 助詞-接続助詞 O
いる イル いる 動詞-非自立 一段 基本形 O
。 。 。 記号-句点 O
EOS
```

図 1　ChaSen・CaboCha による、「下からの光が、天井に当たって反射している。」の解析結果

2. 助詞型機能表現とその検出

2.1 助詞型機能表現

本研究では、機能表現のうち、前稿 [5)] で助詞型機能表現に分類したもの 326 個を研究対象とする。これらは、森田・松木 [6)] でとり上げられている、271 個の助詞相当の複合辞を独自に整理したものに、いくつかの助詞を追加したものである。下位分類における助詞型機能表現の例と数を表 1 に示す。

2.2 助詞型機能表現の検出と判定指標

機能表現を検出する最も単純な方法は、「機能表現を構成する形態素列が文中に見つかった場合、その形態素列を機能表現であるとして検出する」というものである。当然のことながら、この方法は上手くいかない。なぜならば、その形態素列中のある形態素が、例文 (A) の「当たって」のように、内容語として用いられていることがあるからである。それゆえ、機能表現を検出するためには、機能表現の候補となりうる形態素列 (以下、候補表現) が機能表現であるかどうかの判定を行なわなければならない。

この判定には、3 つの段階があると考える。

(1) 表現全体が機能的か、内容的か

(2) 一つの機能表現か、機能表現の列か
　例) < 定義 > の「とは」か、「と」+「は」か

(3) どの意味か (同表記異義語の曖昧性解消)

(2) と (3) の判定は、(1) の判定後に、併せて行なうのが望ましいと考え、本研究では、(1) の判定問題を扱う。

(1) の判定問題の見通しを良くするために、「内容的」、「機能的」という属性を、語に付いた属性ではなく、左接続・右接続を伴った語・表現 (例えば、『用言-「ものを」-

表 1 助詞型機能表現

機能	下位分類	例	表現数
前件の体言を後件の用言に関係付ける	格助詞型機能表現	に対して、によって、に至るまで	57
前件の体言を後件の体言に関係付ける	連体助詞型機能表現	に関しての、といった、における	42
前件を話題化する	係助詞型機能表現	といったら、としては、としても	34
前件に付加的なニュアンスを与える	副助詞型機能表現	に限らず、ならでは、をよそに	27
用言で終わる節を後件の用言に関係付ける	接続助詞型機能表現	からには、おかげで、かと思ったら	123
文末に付加的なニュアンスを与える	終助詞型機能表現	ものだ、ことだ、ないかい	43

表 2 判定指標の割合

	0	c0	c1	c2	c2-	計
格助詞型	3(5%)	4(7%)	31(54%)	18(32%)	1(2%)	57(100%)
連体助詞型	4(10%)	8(19%)	27(64%)	3(7%)	0(0%)	42(100%)
係助詞型	4(11%)	7(21%)	7(21%)	13(38%)	3(9%)	34(100%)
副助詞型	8(30%)	4(15%)	15(55%)	0(0%)	0(0%)	27(100%)
接続助詞型	41(33%)	51(42%)	15(12%)	10(8%)	6(5%)	123(100%)
終助詞型	28(65%)	6(14%)	9(21%)	-(–%)	-(–%)	43(100%)
計	88(27%)	80(25%)	104(32%)	44(13%)	10(3%)	326(100%)
累計	88(27%)	168(52%)	272(**83%**)	316(97%)	326(100%)	

文末』は、終助詞という機能をとる) の用法に付いた属性であると考える。このような立場をとるとき、すべての語・表現を、次の 3 つのカテゴリのいずれかに分類できる。

A 機能的用法のみを持つ
例) 「が」、「けれども」、「からには」、「に関して」

B 内容的用法と機能的用法を持つ
例) 「もの」、「ところ」、「をめぐり」、「に当たって」

C 内容的用法のみを持つ
例) 「本」、「食べる」、「日本語能力」、「手を焼く」

このうち、A と B に分類されるものが、機能表現の候補表現であるが、これらは、機能表現かどうかの判定の難易度が異なる。そこで、本研究では、候補表現が A に属する機能表現に、形態素解析中に判定可能であるとして、判定指標 0 もしくは c0 を、候補表現が B に属する機能表現に、形態素解析後に判定すべきであるとして、判定指標 c1, c2, c2-のいずれかを指定した。

それぞれの判定指標について説明する。

判定指標 0 「からには」のように、「その中に内容的用法を持つ要素を含まない」機能表現に指定する。

判定指標 c0 「に関して」のように、「その中に内容的用法を持つ要素を含んでいるが、この形態素列 (と左接続・右接続) においては全体として機能的用法でのみ用いられる」機能表現に指定する。

判定指標 c1 局所的文脈から検出が可能であると考えられる機能表現に指定する。本研究では、「局所的文脈」を候補表現の前後一語の範囲とし、この文脈情報を、それぞれの機能表現ごとに、以下の 4 種類の集合として機能表現辞書 [5] に記述した。

- 前集合: 前に来ることができる語の集合
  例) < 状態 > の意味を持つ「をもって $_1$」に対して、{ 自信, 確信, 余裕, 心, ··· }
- 非前集合: 前に来ることができない語の集合
  例) < 状況 > の意味を持つ「に当たって」に対して、{ 壁, 天井, 顔, 腕, ··· }
- 後集合: 後に来ることができる語の集合
  例) < 逆-確定 > の意味を持つ「ものの」に対して、{“、”, “,”}
- 非後集合: 後に来ることができない語の集合
  例) < 逆-想外 > の意味を持つ「くせに」に対して、{ こだわる, 悩む }

これらの集合は、「日本語文型辞典」[7] を参考にして一つの機能表現に対して、2 から 100 程度記述した。なお、判定指標 c1 の機能表現であっても、文脈によっては、上の集合を用いても候補表現が機能表現かどうかの判定ができないことがある。そのような場合には、後の処理で広い文脈を考慮して判定を行なうことを想定している。

判定指標 c2, c2- 検出に広い文脈を考慮する必要がある機能表現に指定する。機能表現かどうかの判定ができずに曖昧性が残った場合に、機能表現であるとみなすべきものに c2 を、機能表現ではないとみなすものに c2-を指定する。

助詞型の下位分類の機能型における判定指標の割合を表 2 に示す。判定指標 0, c0, c1 の機能表現で、全体の 83%(272/326) を占める。

3. 接続情報にもとづく助詞型機能表現の検出システム

本研究では、接続情報にもとづいて、判定指標が 0, c0, c1 の機能表現を検出するシステム (以下、検出システム) を実装した。この検出システムは、機能表現辞書 [5] から必要な情報を取り出して、機能表現の検出を行なう。検出システムの概要を図 2 に示す。

検出は次の手順で行なう。まず、入力文を形態素解析器 JUMAN で解析する。このとき、同時に、判定指標 0,

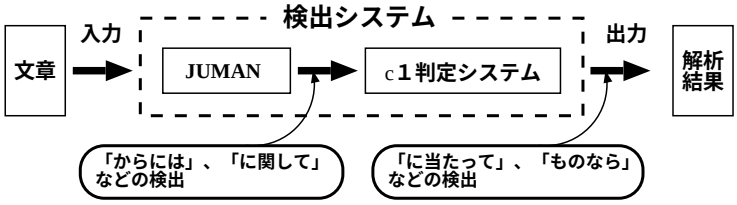

図 2 検出システムの概要

表 3 JUMAN の辞書に登録したエントリ数

	0	c0	小計
格助詞型	0	10	10
連体助詞型	4	9	13
係助詞型	3	13	16
副助詞型	9	6	15
接続助詞型	19	84	103
終助詞型	25	6	31
計	60	128	188

c0 の機能表現を検出する。続いて、c1 判定システムが、この出力を受けて、判定指標 c1 の機能表現を検出する。

以下の節で、それぞれの検出について詳しく述べる。

3.1 形態素解析器による機能表現の検出

機能表現を形態素解析中に検出するためには、それらを形態素解析器の辞書にエントリとして追加すればよい。

本研究では、すでに辞書に登録されているものを除いて、判定指標 0, c0 の機能表現をすべて JUMAN の辞書に登録した。「に際し (て)」のように、「に際し」、「に際して」と複数の表記があるものはすべて異なるエントリとして辞書に登録した。登録したエントリ数を表 3 に示す。

3.2 直前・直後の文脈による機能表現の検出

検出システム内の c1 判定システムは、JUMAN の出力に存在する判定指標 c1 の機能表現の候補表現に対して、前集合、非前集合、後集合、非後集合の順に条件を確かめていき、条件が一致した場合、その集合の種類に応じて候補表現が機能表現であるかどうかの判定を行なう。なお、判定ができずに曖昧性が残った場合は、候補表現を機能表現ではないと判定する。

4. 検出システムの評価実験

前章で述べた検出システムの評価実験を行なった。

4.1 現代語複合辞用例集を用いた評価実験

4.1.1 実 験 対 象

「現代語複合辞用例集」[8] は、助詞的複合辞 83 個および助動詞的複合辞 42 個に対して、その意味・用法、用例、文法などを記述したものである。本研究では、助詞的複合辞に参考表現 4 個を合わせたものの中から、76 個の表現に対する用例 (すべて正例) を評価に用いた。この評価実験は、クローズドテストである。

4.1.2 結果と検討

システムの出力を人手で評価した結果を表 4 に示す。システム全体として、正例から機能表現を検出できたものは、96.3%(847/880) であった。

表 4 現代語複合辞用例集による検出システムの評価

判定指標	表現数	正例	正	誤
0	12	163	163	0
c0	30	306	300	6
c1	34	411	384	27
計	76	880	847	33

形態素解析器における検出失敗 (6 個) の原因は、その辞書にエントリがないことであった。これは、辞書にエントリを追加することにより対処できる。

c1 判定システムの検出失敗の原因は、次のものであった。

(a) 並列の「と」の挿入 (1 個):「血と労苦 と をもって」
(b) 局所的判定の助けにならない語が前後に接続 (15 個)
(c) JUMAN の解析誤り (11 個)

(b) の問題のうち、「にかけて」、「にわたって」など用言を中心として構成された機能表現に関するもの (9 個) は、後の処理で係り受け関係などの大域的な情報を用いることにより、解消することが可能であると思われる。

4.2 機能表現用例コーパスを用いた評価実験

4.2.1 実 験 対 象

「機能表現用例コーパス」[9] は、「現代語複合辞用例集」に記載されている複合辞に対して、1995 年の毎日新聞の記事データからその候補表現を含む例文を抽出したものである。すべての例文には、そこに含まれる候補表現が機能的用法で使用されているのか、それとも内容的用法で使用されているのかを示すマークが付与されている。

評価には、このコーパスのうち、現在利用可能な、25 個の表現に対する用例 (正例と負例) を用いた。この評価実験は、オープンテストである。

4.2.2 結果と検討

システムの出力を人手で評価した結果を表 5 に示す。システム全体として、正しく判定ができたものは、90.7%(1064/1173) であり、提案した手法が有効であることがわかった。

形態素解析器による誤りは、(i) 接続条件の不備 (14 個) と (ii) 不適切な判定指標 (17 個) によるものである。これらの誤りは、適切な条件等を指定することにより、容易に除去することができる。

c1 判定システムの誤りの原因は、次のものであった。

(I) 判定条件の不備 (65 個)、辞書にエントリがない (1 個)
(II) 局所的判定の助けにならない語が前後に接続 (9 個)

その他は、JUMAN の解析誤り (3 個) である。(I) の問題は容易に解決でき、(II) の問題のうち、「にかけて」と

表 5 機能表現用例コーパスによる検出システムの評価

機能表現	正例	正	誤	負例	正	誤	全体	正	誤	検討
のみならず	51	51	0	0	0	0	51	51	0	
からには	50	50	0	0	0	0	50	50	0	
との	47	47	0	0	0	0	47	47	0	
0 (3 個)	148	148	0	0	0	0	148	148	0	
にもかかわらず	51	51	0	0	0	0	51	51	0	
に限らず	51	51	0	0	0	0	51	51	0	
に関する	50	50	0	0	0	0	50	50	0	
に際し	50	50	0	0	0	0	50	50	0	
に至っては	50	50	0	0	0	0	50	50	0	
につれ	50	50	0	0	0	0	50	50	0	
といえども	50	50	0	0	0	0	50	50	0	
に関して	50	49	**1**	0	0	0	50	49	1	(i)
とはいえ	33	33	0	18	5	**13**	51	38	13	(i)
に応じて	38	38	0	12	0	**12**	50	38	12	(ii)
にしても	20	20	0	5	1	**4**	25	21	4	(ii)
につれて	49	49	0	1	0	**1**	50	49	1	(ii)
c0 (12 個)	542	541	1	36	6	30	578	547	31	
を問わず	50	50	0	0	0	0	50	50	0	
をめぐって	50	50	0	0	0	0	50	50	0	
をよそに	50	50	0	0	0	0	50	50	0	
にかけては	9	0	**9**	1	1	0	10	1	9	(I)
にかけて	40	2	**38**	0	0	0	40	2	38	(I),(II)
にわたって	49	37	**12**	0	0	0	49	37	12	(I)
くせに	38	37	**1**	13	10	**3**	51	47	4	(I)
に限り	40	39	**1**	8	3	**5**	48	42	6	(I)
において	48	47	**1**	1	0	**1**	49	47	2	(I),(II)
ものの	46	39	**7**	4	4	0	50	43	7	(II)
c1 (10 個)	420	351	69	27	18	9	447	369	78	
計 (25 個)	1110	1040	70	63	24	39	1173	1064	109	

「において」に関するもの (2 個) は、後の処理で大域的な情報を用いれば、解消することができると考えられる。

5. おわりに

本研究では、判定指標を導入し、それぞれの機能表現に対して、その機能表現の検出を文解析処理のどの時期に行なうべきかを指定した。そして、このうち、形態素解析中に検出を行なう手法と、形態素解析後に局所的文脈から検出を行なうシステムを実装した。

現在、判定条件は人手で記述しているが、今後は、コーパスなどを用いて判定条件を自動的に獲得する必要がある。また、係り受け解析器を利用して大域的な情報にもとづく判定システムを実装し、検出システムの判定精度を上げるとともに、検出対象の機能表現を増やしていく必要がある。

本研究の一部は、次の研究費による；基盤研究 (A)「円滑な情報伝達を支援する言語規格と言語変換技術」(課題番号 16200009)、21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」、京都大学–NTT コミュニケーション科学基礎研究所共同研究「グローバルコミュニケーションを支える言語処理技術」。

参 考 文 献

1) 黒橋禎夫, 河原大輔: 日本語形態素解析システム JUMAN version 4.0, 東京大学大学院情報工学系研究科 (2003).
2) 黒橋禎夫: 日本語構文解析システム KNP version 2.0 b6 使用説明書 (1998).
3) 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸: 形態素解析システム「茶筌」version 2.3.3 使用説明書 (2003). http://chasen.naist.jp/.
4) 工藤拓, 松本裕治: チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842 (2002).
5) 松吉俊, 佐藤理史, 宇津呂武仁: 機能・意味・形態にもとづく助詞型機能表現の分類, 言語処理学会第 11 回年次大会 B2-2 (2005).
6) 森田良行, 松木正恵: 日本語表現文型 用例中心・複合辞の意味と用法, アルク (1989).
7) グループ・ジャマシイ編: 教師と学習者のための日本語文型辞典, くろしお出版 (1998).
8) 山崎誠, 藤田保幸: 現代語複合辞用例集, 国立国語研究所 (2001).
9) 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史, 中川聖一: 日本語機能表現用例コーパスの作成, 言語処理学会第 11 回年次大会 A5-4 (2005).