# 助動詞型機能表現の形態・接続情報と自動検出

中 塚 裕 之　　佐 藤 理 史　　宇 津 呂 武 仁

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

## 1. はじめに

機能表現とは、「〜に関して」、「〜に違いない」のように、2形態素以上に分解されうるが、文中ではひとまとまりとなって、機能的な意味をもつ表現である。機能表現は、日本語の中に数多く存在し、日常的によく使われる表現であるが、自然言語処理において、それらを高精度で検出するようなツールは、まだ存在しない。

これらの機能表現の多くは、非構成的な意味を持つ。例えば、「彼は気休めを言ったに過ぎない」における「〜に過ぎない」という表現は、「過ぎる」の否定形の意味とは異なる意味を持つ。このため、この文における「〜に過ぎない」は、ひとまとまりの表現として認識しなければならない。例えば、機械翻訳においては、このような機能表現を直訳すると、正しい翻訳が得られないので、翻訳しようとする表現が、機能表現であるかどうかを判定した上で、機能表現であれば、機能表現として、そうでなければ、内容語として訳し分けをする必要がある。

本論文は、機能表現を検出するツールを作成することを目的として、まず、助動詞型機能表現についての情報を整理し、判定指標を導入する。ここで、機能表現の判定には、前後の形態素の接続情報にもとづいた方法を考える。次に、検出システムを作成して、実際に検出実験を行った結果について述べる。これらを通じて、どの程度の助動詞型機能表現が、接続情報を用いることで、曖昧性解消を行えるのかを明らかにする。

## 2. 機能表現

どのようなものを「機能表現」と呼ぶかについて、統一された明確な定義はないが、松吉 [1] は、「機能的」とは、「助詞的」、「助動詞的」、「接続詞的」のいずれかであることととらえ、それぞれの機能をもった表現を、助詞型機能表現、助動詞型機能表現、接続詞型機能表現と定義した。本論文ではこの分類のうち、助動詞型機能表現を研究対象とする。

### 2.1 助動詞型機能表現

助動詞型機能表現とは、例えば、「〜に違いない」、「〜ざるを得ない」といったものである。松吉 [1] は、助動詞型機能表現は、「前件の述語に付加的なニュアンスを与える」という機能をもつと定義した。これは主に、助動詞によるムードの表現に相当する。本論文では、よく使われる代表的な助動詞型機能表現として、国立国語研究所「現代語複合辞用例集」(以下では、「用例集」と略記する) [2] に掲載されている「助動詞的複合辞」42 種を扱う。これを表 1 に示す。

### 2.2 エントリの確定

機能表現を同定するためには、どの範囲を1つのエントリと考えるかという問題を解決する必要がある。本研究では、佐藤 [3] の同語異語判定についての基本的な考え方を参考にして、以下の2つの指標を採用した。

(1) 意味が明らかに異なるものは同一エントリとは考えない
(2) 構成形態素が明らかに異なるものは同一エントリとは考えない

まず、(1) については、「用例集」[2] に従い、42 種は別エントリとして扱うことにした。次に、(2) については、以下の (a)、(b) を別エントリとして、(c)-(g) を同一エントリの別表記として扱うことにした。

(a) 助詞の挿入・脱落・変化 → 別エントリ
例:「〜わけがない」≠「〜わけもない」
(b) 内容語の変化 → 別エントリ
例:「〜に違いない」≠「〜に相違ない」
(c) 異表記同語 → 同一エントリ
例:「〜う (得) る」=「〜え (得) る」
(d) 丁寧形 → 同一エントリ
例:「〜に違いない」=「〜に違いありません」
(e) 口語形 → 同一エントリ
例:「〜ものだ」=「〜もんだ」
(f) 否定派生形 → 同一エントリ
例:「〜ざるを得ない」=「〜ざるを得ぬ」
(g) 省略形 → 同一エントリ
例:「〜ところだ」=「〜とこだ」

この結果、本研究で扱う助動詞型機能表現の総数は、42 種 129 エントリとなった。

## 3. 助動詞型機能表現の整理

前述のように、本研究では、助動詞型機能表現 42 種 129 エントリを扱う。それらの表記は常に機能表現となるとは限らない。例えば次の2文を考えてみよう。

(1) このことで礼を言う**には及ばない**。
(2) A 社の製品は、B 社の製品**には及ばない**。

どちらの文も「〜には及ばない」という表記を含むが、(1)

**表 1** 42 種より代表的な 42 エントリ

| | |
|---|---|
| (1)「〜ものだ」 | (22)「〜に足りない」 |
| (2)「〜はずだ」 | (23)「〜に違いない」 |
| (3)「〜つもりだ」 | (24)「〜にほかならない」 |
| (4)「〜ところだ」 | (25)「〜に過ぎない」 |
| (5)「〜一方だ」 | (26)「〜には及ばない」 |
| (6)「〜どころではない」 | (27)「〜ほうがいい」 |
| (7)「〜ほかない」 | (28)「〜たらいい」 |
| (8)「〜わけだ」 | (29)「〜たらいけない」 |
| (9)「〜わけがない」 | (30)「〜てもいい」 |
| (10)「〜わけにはいかない」 | (31)「〜てもしょうがない」 |
| (11)「〜ことだ」 | (32)「〜ないではいられない」 |
| (12)「〜ことがある」 | (33)「〜てならない」 |
| (13)「〜ことができる」 | (34)「〜といったらない」 |
| (14)「〜ことになる」 | (35)「〜つつある」 |
| (15)「〜ことにする」 | (36)「〜うとする」 |
| (16)「〜までだ」 | (37)「〜かもしれない」 |
| (17)「〜までもない」 | (38)「〜とは限らない」 |
| (18)「〜ばかりだ」 | (39)「〜得る」 |
| (19)「〜に決まっている」 | (40)「〜ざるを得ない」 |
| (20)「〜に限る」 | (41)「〜べきだ」 |
| (21)「〜にとどまらない」 | (42)「〜なければならない」 |

は機能表現であり、(2) は自立的 (内容的) な表現である。このように、一般に同表記であっても、文中で、機能的に働く場合と自立的に働く場合がある。我々は、機能的に働く場合のみ、機能表現 (の表記) と考える。なお、機能的に働く場合は、構成的な意味を持つ場合と、非構成的な意味を持つ場合がありうるが、今回の研究では，それらは区別しないことにする。

このことから、機能表現を計算機によって検出しようと考えた場合、その課題は以下の２つに分けられる。

(a) 文中より「機能表現の表記」を検出する。

(b) (a) で検出した表記が「機能的に働いている」かどうかを判定する。

本章では、まず、表記を検出するために機能表現の形態情報について述べ、機能的に働いているかどうかを判定するための接続情報について述べる。次に、意味分類について述べる。この意味分類は、言い換えを想定したものである。最後に、判定指標について述べる。この判定指標は、検出システムを想定したものである。なお、「機能的に働いている」か否かの区別は、本研究では、基本的に「用例集」[2] の意味・用法欄、用例欄に準拠することとする。なお、これらの情報の整理において、「用例集」[2]、森田・松木「日本語表現文型」[4]、グループ・ジャマシィ「日本語文型辞典」[5]、沖森ら編「ベネッセ表現読解国語辞典」[6] を参考にした。

### 3.1 形態情報

表記を検出することを考えた場合、次の例のように、偶然表記が一致しただけのものを検出する可能性がある。

例：言葉を大**事にする**。(「〜こと (事) にする」)

このような誤検出は、形態素解析を用いて避けることができる。そこで、助動詞型機能表現 42 種 129 エントリが、JUMAN4.0[7] によって、どのような形態素に分割されるかを調査し、この結果を整理した。

**表 2** 形態情報の整理 (例)

| 機能表現 | 構成形態素 | | |
|---|---|---|---|
| 〜に限る | | 助詞「に」 | 動詞「限る」 |
| 〜ても仕方ない | 「て」 | 助詞「も」 | 形容詞「仕方ない」 |

ところで、機能表現は、独立した形態素のみから構成されるとは限らない。例えば「言っても仕方ない」の場合、「〜ても仕方ない」を機能表現と考えるのが自然だが、形態素解析の結果は、「て」は「言う」の活用語尾として出力される。このように、機能表現は、独立した形態素のみから構成されるとは限らない。「て」は非形態素として扱った。例として、整理した表の一部を表 2 に示す。

### 3.2 接続情報

本章の最初に挙げた、２つの例文にある「〜には及ばない」は、JUMAN4.0[7] においては、どちらも『接続助詞「に」＋副助詞「は」＋動詞「及ぶ」未然形＋接尾辞「ない」基本形』と形態素解析される。そこで、前後の接続にもとづく判定方法が必要となる。例えば、該当表記の直前の品詞に注目すると、(1) は普通名詞「製品」であり、(2) は動詞「言う」である。ここで、機能表現「〜には及ばない」は、動詞には接続するが、普通名詞には接続しないということが成り立つのならば、これらを正しく判定することができる。

#### 3.2.1 左接続

助動詞型機能表現 42 種 129 エントリの、直前の品詞を大きく動詞、イ形容詞、ナ形容詞、体言、その他の５つに分け、各エントリがそれらの語に接続した場合に、機能的な用法が存在するか否かを調査した。その一部を表 3 に示す。

#### 3.2.2 右接続

助動詞型機能表現 42 種 129 エントリの直後の形として、「ない」(否定形)、「だろう」(推量形)、「、(読点)」(連用中止形)、「体言」(連体形)、「た」(過去形) の５つを考え、各エントリにそれらが接続した場合に、機能的な用法が存在しうるか否かを調査した。なお、例えば「〜かもしれない」のような、否定形の表現に対しての、「ない」用法は、「〜なくはない」の形の二重否定として扱うこととした。この調査結果の一部を表 4 に示す。

### 3.3 意味分類

機能表現では、複数の異なる表現が、ほぼ同じ意味を表すことがある。例えば、次の２文はほぼ同じ意味を持つ。

(1) あいつがやった**に決まっている**。

(2) あいつがやった**にちがいない**。

そこで、機能表現の意味を、大きく「当為・判断系」、「可能・可能性・帰結系」、「意思系」、「程度系」、「自発系」、「その他」の６つに分け、助動詞型機能表現 42 種を分類した。ただし、この分類においては、１種の機能表現が、必ずしもひとつの意味カテゴリに属するとは限らない。例えば、次の２文は全く異なった意味を持っている。

(1) 病人はおとなしくしている**ものだ**。

(2) 昔はよく虫取りをした**ものだ**。

**表 3** 左接続情報の記述例

| 機能表現 | 動詞 | イ形容詞 | ナ形容詞 | 体言 | その他 |
|---|---|---|---|---|---|
| ～に限る | 基本形<br>タ系連用テ形+接尾辞「いる」基本形 | 基本形 | × | 名詞 | × |
| ～にほかならない | × | × | × | 名詞 | 接続助詞「から」 |

**表 4** 右接続情報の記述例

| 機能表現 | ない | だろう | 読点 | 体言 | た |
|---|---|---|---|---|---|
| ～に限る | × | に限るだろう | に限り、 | × | × |
| ～うとする | うとしない | うとするだろう | うとし (て)、 | うとする～ | うとした |

注：表中の接続情報における () は省略可能を表し、～は体言を表す。

**表 5** 意味分類ごとの種数

| 意味カテゴリ | 種数 |
|---|---|
| 当為・判断系 | 17 |
| 可能・可能性・帰結系 | 17 |
| 意思系 | 8 |
| 程度系 | 6 |
| 自発系 | 2 |
| その他 | 10 |
| 計 | 60 |

このような場合は、「ものだ $_1$」、「ものだ $_2$」のように、意味 ID を付けて区別した。それぞれの意味カテゴリに含まれる種数を表 5 に示す。

### 3.4 機能表現の判定指標

判定指標とは、ある表記が機能的に働いているか、自立的に働いているかを判定するにあたり、どのように判定すればよいかを示したものである。本研究では、松吉 [1)] の判定指標を参考に、５つの判定指標を導入した。この判定指標と、助動詞型機能表現 42 種の代表的な 42 エントリにおけるその割合を表 6 に示す。

ここで、“c” を含む判定指標は、表現の中に「自立的用法を持っている形態素」が含まれていることを表している。また、機能表現として使われることが多いと考えられる表現には、cf を付与し、自立的な表現として使われることが多いと考えられる表現には、cg を付与している。

## 4. 自動検出実験

３章の整理に基づいて、機能表現を検出するシステムを作成した。その構成を図 1 に示す。

### 4.1 京大コーパスに対する検出実験

京大コーパス [8)] より 1,129 文を選択し、その中から対象となる 42 エントリの表記を人手で抽出した。ここで抽出した表現を、抽出データと呼ぶ。抽出データに含まれる機能表現は 19 エントリ 83 件であった。次に、この抽出データに、人手で自立的な表現、機能表現の判定を与えた。

この 1,129 文を対象に、検出システムが、正しく機能表現を検出できるかどうかを調べた。その結果、判定誤りはわずか 1 件のみであった。

この実験における、判定指標と抽出データ、システムの判定タイミングの関係を表 7 に示す。ここで、表中のシステム判定タイミングの欄は、システムの各判定部で、

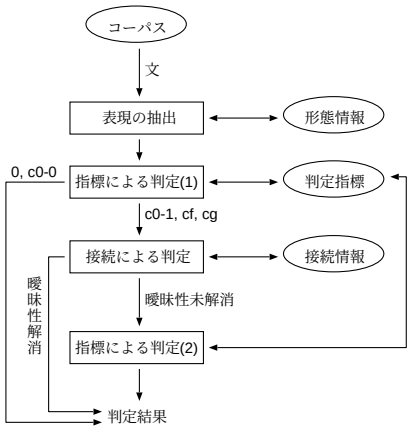


**図 1** システム構成

判定が行われた件数を表している。この欄を見ると、指標による判定 (1) と接続による判定で判定できた件数は、累積で全体の 71 %である。残り 29 %は、指標による判定 (2)—すなわち頻度—で判定されている。この頻度による判定がうまくいったことにより、全体として、非常に高い正解率が得られている。なお、この実験での唯一の判定誤りは、指標による判定 (1) の cg に含まれる 22 件のうち、１件が機能表現であったことによる。

抽出データとシステム判定タイミングを比較すると、抽出データでは cf、cg に含まれる 33 件のうち 9 件 (27 %) が、接続による判定により曖昧性が解消されている。これらは、左右の接続が、機能表現の可能性がない接続であったため、自立的な表現と判定されたものである。

### 4.2 機能表現用例コーパスに対する検出実験

「機能表現用例コーパス」[9)] は、1995 年の毎日新聞の記事データから、「用例集」[2)] に記載されている複合辞の表記を含む例文を抽出したものである。すべての例文には、そこに含まれる表記が機能的に働いているのか、それとも自立的に働いているのかを示すマークが付与されている。このコーパスより、判定指標 c0-1、cf、cg が付与されている機能表現を 8 エントリ選択し、これらに対して、上記のシステムを用いて検出実験を行った。その結果を表 8 に示す。この表において、「～に決まっている」を除いた 7 エントリでは、非常に高い正解率を示している。「～に決まっている」は、接続条件だけでは判定ができず、かつ、正例・負例の頻度に大きな偏りがなかった

**表 6** 判定指標とその割合

| 判定指標 | 内容 | エントリ数 | 割合 | 累積 |
|---|---|---|---|---|
| 0 | 表現が機能的にのみ働くもの | 4 | 10% | 10% |
| c0-0 | 表現が機能的にのみ働くもの | 20 | 48% | 57% |
| c0-1 | 左右の形態素の接続関係によって判定可能であるもの | 5 | 12% | 69% |
| cf | 左右の形態素の接続関係による判定で曖昧性が残る場合に機能表現と判定するもの | 7 | 17% | 86% |
| cg | 左右の形態素の接続関係による判定で曖昧性が残る場合に機能表現ではないと判定するもの | 6 | 14% | 100% |
| 計 | | 42 | 100% | - |

**表 7** 判定指標と京大コーパスに対する検出実験の抽出データ・システム判定タイミング

| 判定 | 抽出データ | | | システム判定タイミング | | | |
|---|---|---|---|---|---|---|---|
| 指標 | 件数 | 割合 | 累積 | 指標による判定 (1) | 接続による判定 | 指標による判定 (2) | 計 |
| 0 | 3 | 4% | 4% | 3 | - | - | 3 |
| c0-0 | 29 | 35% | 39% | 29 | - | - | 29 |
| c0-1 | 18 | 22% | 60% | - | 18 | - | 18 |
| cf | 5 | 6% | 66% | - | 3 | 2 | 5 |
| cg | 28 | 34% | 100% | - | 6 | 22 | 28 |
| 計 | 83 | 100% | - | 33(39%) | 27(33%) | 22(29%) | 83 |

**表 8** 機能表現用例コーパスに排する検出実験結果

| エントリ | 判定指標 | 例数 | 正例数 | 判定 | | 負例数 | 判定 | | 正解率 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | 正 | 誤 | | 正 | 誤 | |
| ～一方だ | c0-1 | 47 | 47 | 47 | 0 | 0 | - | - | 100% |
| ～てならない | c0-1 | 50 | 49 | 49 | 0 | 1 | 1 | 0 | 100% |
| ～なければならない | c0-1 | 50 | 50 | 50 | 0 | 0 | - | - | 100% |
| ～わけだ | cf | 50 | 50 | 50 | 0 | 0 | - | - | 100% |
| ～わけがない | cf | 46 | 46 | 46 | 0 | 0 | - | - | 100% |
| ～に過ぎない | cf | 49 | 49 | 48 | 1 | 0 | - | - | 98% |
| ～かもしれない | cf | 52 | 52 | 52 | 0 | 0 | - | - | 100% |
| ～に決まっている | cg | 37 | 16 | 0 | 16 | 21 | 21 | 0 | 57% |
| 計 | | 381 | 359 | 339 | 20 | 22 | 22 | 0 | 96% |

ため、正解率が低くなっている。

## 5. おわりに

本研究では、助動詞型機能表現の形態と接続情報を整理し、判定指標を導入して、助動詞型機能表現を自動検出するシステムを作成した。本システムを用いて、実際に助詞型機能表現の検出実験を行ったところ、69 %の表現 (0、c0-0、c0-1) について、形態と接続情報を用いることで、自立的な表現か機能表現かを判定することができた。また、残りの表現も、形態・接続情報と頻度を用いることにより、機能表現か否かを、高い精度で判定できる可能性がある。

本研究で行った実験は、新聞コーパスを対象としており、そこに含まれている機能表現は偏っている可能性がある。そのため、新聞以外のコーパスに対しても、同様の結果が得られるかどうか、今後、調査する必要がある。

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) 松吉俊: 機能・意味・形態にもとづく日本語機能表現の分類と自動検出, 修士論文, 京都大学 (2005).
2) 国立国語研究所: 現代語複合辞用例集 (2001).
3) 佐藤理史: 異表記同語認定のための辞書編纂, 情報処理学会研究報告 2004-NL-161, pp. 97–104 (2004).
4) 森田良行, 松木正恵: 日本語表現文型, アルク (1989).
5) グループ・ジャマシィ編: 日本語文型辞典, くろしお出版 (1998).
6) 沖森卓也, 中村幸弘編: ベネッセ表現読解国語辞典, ベネッセ (2003).
7) 黒橋禎夫, 河原大輔: 日本語形態素解析システム JUMAN version 4.0 (2003).
8) 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 回年次大会発表論文集, pp. 115–118 (1997).
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 回年次大会発表論文集 (2005).