形態素情報を用いた日本語機能表現の検出

Detecting Japanese Functional Expressions based on Morphological Clues

土屋 雅稔 * 宇津呂 武仁 † 佐藤 理史 † 中川 聖一 ‡

*豊橋技術科学大学 情報処理センター

† 京都大学 情報学研究科

‡豊橋技術科学大学 工学部情報工学系

1. はじめに

日本語には，複数の語が定型的・複合的に使われ，ひとかたまりの表現として機能的な関係を表す表現が多数存在する．このような表現は機能表現と呼ばれ，日本語文の構造を理解するために非常に重要である．機械翻訳をはじめとする多様な応用において機能表現を適切に取り扱うには，多数の機能表現を網羅的かつ明示的に検出するような統一的な解析系が必要である．しかも，機能表現と同一の形態素列が内容的な用法で用いられている例も多く存在するため，機能表現を正しく検出するには，単に機能表現と同一の形態素列を検出するだけでなく，その形態素列の用法を判定しなければならない．

しかし，既存の解析系における機能表現の取り扱いは不十分である．例えば，形態素解析器 JUMAN[1] と構文解析器 KNP[2] の組合わせは，形態素解析時には機能表現を検出していない．構文解析時に，解析規則に記述された特定の形態素列が現れると，直前の文節の一部として処理したり，直前の文節からの係り受けのみを受けるように制約を加えて解析を行っている．この時，その形態素列が内容的な用法で用いられている可能性は考慮されていない．また，形態素解析器 ChaSen[3] と構文解析器 CaboCha[4] の組合わせは，IPA 辞書に「助詞・格助詞・連語」と登録されている 88 種類の機能表現以外の機能表現を明示的に検出していない．

このような背景を踏まえ，本稿では形態素情報を用いて機能表現を検出する手法を提案し，現代語複合辞用例集 [5] に収録されている機能表現を対象として評価を行った結果について述べる．そして，形態素情報による機能表現の検出がどの程度可能であるかを示す．

2. 機能表現の検出

2.1 対象とする機能表現

機能表現とは，幾つかの語が複合してひとまとまりの形で辞的な機能を果たす表現である．森田ら [6] は，機能表現の中でも特に「単なる語の連接ではなく，表現形式全体として，個々の構成要素のプラス以上の独自の意味が生じている」表現を**複合辞**と呼び，個々の構成要素の意味から構成的に表現形式全体の意味を説明できるような機能表現とは区別している．

現代語複合辞用例集 (以下，用例集と呼ぶ) は，森田らが複合辞として取り上げた表現を基本とし，その中でも 1 つの複合形式として熟合度が高く，また一般性も高いと判断される表現を，さしあたりの研究対象として列挙したものである．本稿では，この用例集に収録されている機能表現を検出対象とする．収録されている表現と同一の文字列または形態素列が含まれる文が見つかった場合，この表現の用法として，3 つの可能性が存在する．

- 機能的用法である
  - 用例集の用法である …(1)
  - それ以外の機能的用法である …(2)
- 内容的用法である …(3)

例として,「A となると B」の形で「A という事態になった場合には B」という関係を示す「〜となると (A6–1000)」の各用法に対応する例文を以下に示す．

(1) しかし風邪という特定の感染症に特効があるとなると事は重大だ．

(2) となると「たつ」とはなにか？

(3) 脳死移植の適応基準は心臓，肝臓に次ぐもので，国会で継続審議中の臓器移植法案成立へ向けての条件整備の一つとなるとみられる．

本研究では，本稿執筆の段階での評価用コーパスの状態を考慮して,「用例集に収録されている表現が，用例集の用法で用いられていることを検出する」検出器を作成する．ただし，検出規則を修正すれば，機能的用法 (用例集の用法を含む) で用いられている表現に対する検出器も同様に作成可能である．

2.2 機能表現の検出に用いる情報

本研究では，(1) 表層文字列，(2) 形態素境界，(3) 形態素列パターンおよび (4) 接続制約という 4 つの情報を用いて機能表現を検出する．

■表層文字列

各表現の表層文字列を与える．入力文において機能表現が使われている可能性がある候補部分を発見するために用いる．単純な文字列一致と，基本形を考慮した検出手法を併用する．

■形態素境界

文字列によって発見された候補部分に対して，候補部分の先頭と末尾が形態素境界と一致する場合に限定する．

■形態素列パターン

各表現に対応する形態素列パターンを用意し，その形態素列パターンに一致する候補部分に限定する．それぞれの形態素は，基本形・品詞・活用形の 3 つ組で指定す

ID: A42-1011	ID: A62-1000	ID: A77-1000
表層文字列: に応じた	表層文字列: として	表層文字列: からして
パターン: に:助詞:* / 応じる:動詞:連用形 / た:助詞:*	パターン: と:助詞:* / する:動詞:連用形 / て:助詞:*	パターン: から:助詞:* / する:動詞:連用形 / て:助詞:*
制約: 左体言 / !右文末 / !右補助動詞	制約: 左体言 / !右補助動詞	制約: 左体言 / !右補助動詞

図 1 機能表現検出規則の例

る．実際は，形態素解析誤りに対処するため，解析誤りに対応する形態素列パターンも用意した．

■接続制約

用例集の各項目では「接続」などの説明文の形で，各機能表現の直前に現れ得る表現に対する条件を述べている．その条件としては，以下のようなものが見られる．

- 体言に限る
  - 名詞に限る (または，動作的名詞に限る)
  - 「〜から」などの理由節に限る
  - 「〜時」などの時間節に限る
  - 名詞または文 (または文相当) に限る
- 用言に限る
  - 動詞に限る (活用形指定あり/なし)
  - 形容詞に限る (活用形指定あり/なし)
  - 形容動詞に限る (活用形指定あり/なし)
- 上記の組み合わせ

用例集には記述がほとんどないが，機能表現の直後に現れ得る表現にも制限がある．典型的な制限としては，以下の 2 つがある．

- 直後に補助動詞が現れてはならない．
- 直後が文末であってはならない．

これらの接続制約を表現するため，以下の素性とその組み合わせを接続制約として指定できるようにした．

左体言，左用言，左基本形，左過去形，左形態素，右形態素，右補助動詞，右文末

例えば,「左体言」は体言に接続することを意味する．

2.3 検 出 手 順

本研究の検出器は，用例集の用法で用いられている機能表現を以下の手順で検出する．

I. 入力文を形態素解析する．
II. 表現毎に以下の手順 (i)〜(iv) で，用例集の用法で用いられている候補部分を検出する．
   i) 表層文字列による候補部分の検出
   ii) 形態素境界との一致
   iii) 形態素列パターンとの照合
   iv) 接続制約の検査
III. 表現毎に個別に求められた候補部分から，長い候補部分から順に選択する．

以下，詳細を述べる．

最初に，入力文を MeCab[7] を用いて形態素解析する．解析用辞書として IPA 辞書を使っているため,「〜について」などは「助詞, 格助詞, 連語」という 1 形態素として解析される．しかし，後段の検出処理との整合を取るため，これらの連語は全て構成要素の形態素に分解する．

i) 表層文字列による候補部分の検出

次に，表層文字列を用いて，以下の 2 つの方法のいずれかで検出された個所を,「用法を判定しなければいけない候補部分」として取り出す．

(a) 文字列一致による検出 表層文字列を含む個所を無条件に検出する．例えば,「〜として (A62–1000)」に対しては，下線部分のような候補部分が検出される．

助手<u>として</u>働く
彼はきちん<u>とし</u>ている
財布を落<u>とし</u>て困っている

(b) 基本形を考慮した検出 機能表現の末尾形態素が活用して用いられている場合を検出する．以下に,「台風は本土を北上しつつあった」という入力文から「〜つつある (B35–1000)」を検出する手順 (1) 〜 (2) を示す．

(1) 文中の活用している語の 1 つだけを基本形に置き換えた文を生成．
台風/は/本土/を/北上/<u>する</u>/つつ/あっ/た
台風/は/本土/を/北上/し/つつ/<u>ある</u>/た

(2) 表層文字列「つつある」と一致している部分を検出．
台風/は/本土/を/北上/し/<u>つつ/ある</u>/た

ii) 形態素境界との一致

検出された候補部分について，その候補部分の先頭と末尾が形態素境界と一致しているか検査する．一致していない場合は，その候補部分を棄却する．例として,「〜として (A62–1000)」の候補部分が検出された 3 つの文について考える．

○助手<u>として</u>働く
○彼はきちん<u>とし</u>ている
×財布を落<u>とし</u>て困っている

最初の 2 文の候補部分の先頭と末尾は，いずれも形態素境界となっているのに対して，最後の文に現れている候補部分「として」の先頭は，動詞「落とす」の内部に含まれており，形態素境界とは一致していない．そこで，最後の文を棄却する．

iii) 形態素列パターンとの照合

続いて，候補部分の形態素列を，表現毎に登録されている形態素列パターンと照合する．例として,「A からして」の形で「A に基づいて考えれば」という意味を表わす「〜からして (A77–1000)」について考える．

○鮮やかな手口<u>からして</u>プロの仕業に違いない．

表 1 判定性能 (全 127 表現の平均値)

正例の割合	ベースライン			形態素境界			形態素列パターン			接続制約		
	判別率	精　度	再現率	判別率	精　度	再現率	判別率	精　度	再現率	判別率	精　度	再現率
100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
90%以上	96%	96%	100%	97%	98%	99%	97%	98%	99%	98%	99%	100%
50%以上	69%	69%	100%	74%	73%	99%	76%	75%	99%	85%	83%	99%
50%未満	74%	0%	0%	52%	35%	99%	55%	36%	98%	84%	63%	94%
全体	86%	89%	92%	82%	81%	99%	83%	82%	99%	92%	91%	99%

表現毎の判別率 $= \frac{\text{用法を正しく判定できた文の数}}{\text{文数}}$, 表現毎の精度 $= \frac{\text{正しく判定できた正例の数}}{\text{用例集の用法と判定した文の数}}$, 表現毎の再現率 $= \frac{\text{正しく判定できた正例の数}}{\text{正例の数}}$

表 2 判定性能 (全 127 表現の重み付き平均値)

正例の割合	ベースライン			形態素境界			形態素列パターン			接続制約		
	判別率	精　度	再現率	判別率	精　度	再現率	判別率	精　度	再現率	判別率	精　度	再現率
100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
90%以上	96%	96%	100%	98%	98%	100%	98%	98%	100%	99%	99%	100%
50%以上	66%	66%	100%	71%	70%	100%	74%	72%	100%	84%	81%	100%
50%未満	75%	0%	0%	69%	44%	100%	71%	46%	99%	84%	61%	98%
全体	77%	79%	83%	77%	71%	100%	79%	73%	99%	87%	83%	99%

判別率 $= \sum \frac{\text{その表現のために収集された文の数}}{\text{収集された全ての文の数}}$ · 表現毎の判別率, とする．精度および再現率も同様．

×声をからして叫ぶ．

後の文に含まれている候補部分は,「〜からして (A77–1000)」の形態素列パターン (図 1) と一致しないため，後の文を棄却する．

iv) 接続制約の検査

次に，その候補部分の前後が，接続制約を満たしているか検査する．満たしていない場合は，その候補部分を棄却する．例として,「〜として (A62–1000)」の候補部分が検出されている 2 つの文について考える．

○助手として働く

×彼はきちんとしている

「〜として (A62–1000)」の接続制約 (図 1) には，直後に補助動詞が現れることを禁止する制約「!右補助動詞」が含まれている☆．前者の文は制約を満たしているが，後者の文は制約を満たしていないため，後者の文を棄却する．

ここまでの手順 (i)〜(iv) を表現毎に独立に行うと，表現毎に「この部分が用例集の用法で用いられている表現である」という検出結果が得られる．検出された表現を，形態素列長によって降順に並べ替え，長い表現から順に採用し，既に採用された表現と一部でも重なっている表現は棄却する．

現実装では，個々の表現に対する検出手順 (i)〜(iv) を表現毎に全く独立に実行しているが，これは動作速度的に非常に不利な設計なので，いずれ改良する予定である．

3. 実　　験

機能表現用例コーパス [8] に収録されている文を対象として評価実験を行う．ただし，収録予定の 337 表現の内，本稿執筆の段階で検証済の 127 表現のみを対象とする．このコーパスは，用例集に収録されている機能表現について，表層文字列によって検出された候補部分を対象として，人手で用法を判定して作成している．評価対象の文の内，用例集の用法で用いられている文を正例，用例集の用法以外の用法で用いられている文を負例と呼ぶ．

表現毎の判別率・精度・再現率の平均値を表 1，各表現のコーパス作成時に新聞記事から収集された文数を重みとした重み付き平均値を表 2 に示す．ベースラインとは，表層文字列によって検出された全ての候補部分に対して，評価対象の文集合において過半数を占める用法であるという判定を行った場合の結果である．

表 3 形態素境界を用いた場合の判定性能 (表現数)

正例の割合	判別率				計
	100%	90%以上	50%以上	50%未満	
100%	42	2	0	0	44
90%以上	8	14	1	0	23
50%以上	0	4	25	1	30
50%未満	1	3	11	15	30
計	51	23	37	16	

2.3 節 (i)〜(ii) の検出手順の判定性能を，表 1,2 の「形態素境界」欄および表 3 に示す.「〜にもかかわらず (A32–1000)」「〜なければならない (B42–1000)」などの 51 表現は，この段階で 100%判定可能である．これらの表現は，用例集の用法でしか用いられないため，判定のための追加情報は必要ない．なお,「〜ほかない (B7–1000)」「〜ばいい (B28–3000)」は，コーパスに負例が含まれていないにも関わらず判別率が 100%に満たないが，これは形態素解析誤りが原因である．

表 4 形態素列パターンを用いた場合の判定性能 (表現数)

正例の割合	判別率				計
	100%	90%以上	50%以上	50%未満	
100%	42 (±0)	2 (±0)	0 (±0)	0 (±0)	44
90%以上	8 (±0)	14 (±0)	1 (±0)	0 (±0)	23
50%以上	1 (+1)	5 (+1)	23 (−2)	1 (±0)	30
50%未満	1 (±0)	5 (+2)	9 (−2)	15 (±0)	30
計	52 (+1)	26 (+3)	33 (−4)	16 (±0)	

括弧内は，形態素境界を用いた場合の判定性能 (表 3) との差を表す

2.3 節 (i)〜(iii) の検出手順の判定性能を，表 1, 2 の「形態素列パターン」欄および表 4 に示す．用意した形態素

☆ !は素性の否定を表す．

列パターンは 206 個 (その内，形態素解析誤りに対処するためのパターンは 69 個) である．形態素境界のみを用いた場合に比べて判定性能が改善した表現は，「〜とはいえ (A2–1000)」「〜からして (A77–1000)」などの 4 表現である．これらの表現は，2.3 節 (iii) で例を示したように，候補部分の先頭と末尾が形態素境界と一致している場合であっても，候補部分の形態素列が機能表現としての形態素列とは異なっている可能性がある．そのため，形態素列パターンとの照合によって判定性能が改善される．

表 5 接続制約を用いた場合の判定性能 (表現数)

正例の割合	判別率 100%	90%以上	50%以上	50%未満	計
100%	42 (±0)	2 (±0)	0 (±0)	0 (±0)	44
90%以上	14 (+6)	9 (−5)	0 (−1)	0 (±0)	23
50%以上	4 (+3)	9 (+4)	17 (−6)	0 (−1)	30
50%未満	8 (+7)	10 (+5)	9 (±0)	3 (−12)	30
計	68 (+16)	30 (+4)	26 (−7)	3 (−13)	

括弧内は，形態素列パターンを用いた場合の判定性能 (表 4) との差を表す

2.3 節の検出手順全体の判定性能を，表 1,2 の「接続制約」欄および表 5 に示す．接続制約を追加することによって判定性能が改善した表現は，「〜につれて (A55–1000)」や「〜なければならず (B42–1020)」などの 20 表現である．これらの表現が「直後に補助動詞が現れない」や「用言のみに接続する」などの接続制約を満たしている場合には，用例集の用法で用いられていることが多いため，判定性能が改善した．

表 6 接続制約を用いても 100%判定できなかった表現

用例集の接続制約	表現数	例
名詞に接続する	12	〜として (A62–1000)
文 (または文相当) に接続する	17	〜といって (A1–1000)
用言 (動詞・形容詞) に接続する	18	〜だけに (A35–1000)
用言または名詞に接続する	3	〜と思ったら (A9–6000)
動詞に接続する	7	〜ところだ (B4–1000)

接続制約を用いても 100%判定できなかった 57 表現を，用例集に説明されている接続制約で分類すると表 6 が得られる．「文 (または文相当) に接続する」という接続制約を，2.2 節で説明した素性で表現すると，以下のようになる．

左用言 ··· 通常の文に接続する
左体言 ··· 体言止の文に接続する

しかし，この制約はあらゆる表現に接続可能なので，用法判定には役立たない．つまり，「文 (または文相当) を受ける」に分類されている 17 表現は，本研究の検出器が検出に用いている素性だけでは判定不可能である．「用言または名詞を受ける」に分類されている 3 表現についても同様の問題があるが，直前直後の形態素に対して詳細な語彙的制約を加えると判定できるようになる可能性は残されている．残る 37 表現には，(1) 用例集の接続制約が不十分な表現と，(2) 用例集の接続制約が厳しすぎる表現が含まれている．例えば，ある状況の中にいることを表す「〜にあって (A38–1000)」は，用例集では「名詞に付く」と説明されているが，以下のように接続制約を満たす負例が存在しているため，用例集の接続制約では不十分である．

○どんな逆境にあっても全力を尽くす．
×温泉と聞けば，どんな場所にあっても心が弾むものである．

また，前件にも関わらず後件が成り立つことを表す「〜にしても (A34–1000)」は，用例集では「動詞のスル形・シタ形 (シテイル形・シテイタ形を含む) に付く」と説明されているが，実際には名詞も受けることができるので，用例集の接続制約は厳しすぎる．

○例えば，ベースランニングにしても，監督の言われた通りに走るだけでなく，楽しさも取り入れる．

これらの表現については，直前直後の形態素に関する詳細な語彙的制約を加えると判定できる可能性がある．

4. おわりに

本稿では，形態素情報を用いて現代語複合辞用例集に収録されている機能表現を検出する方法を提案し，機能表現用例コーパスを対象として評価を行った．ただし，今回の評価実験では，127 表現中 30 表現程度について評価対象のコーパスを参照しながら，検出規則の開発を行った．今後は，検出規則の開発時には参照しなかった用例に対する評価実験を追加して行う予定である．

謝辞: 本研究の一部は，次の研究費による: 文部科学省 科学研究費 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)，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 使用説明書, 7 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, 6 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 ChaSen. http://chasen.aist-nara.ac.jp/hiki/ChaSen/.
4) 工藤拓, 松本裕治. チャンキングの段階適用による係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.
5) 国立国語研究所. 現代語複合辞用例集, 2001.
6) 森田良行, 松木正恵. 日本語表現文型, NAFL 選書, 第 5 巻. アルク, 1989.
7) 工藤拓. 形態素解析器 MeCab. http://chasen.org/~taku/software/mecab/.
8) 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.