大規模階層辞書と用例を用いた日本語機能表現の解析

鈴木敬文　(筑波大学大学院システム情報工学研究科)
阿部佑亮　(筑波大学大学院システム情報工学研究科)
宇津呂武仁　(筑波大学大学院システム情報工学研究科)*
松吉俊　(山梨大学大学院 医学工学総合研究部)
土屋雅稔　(豊橋技術科学大学情報メディア基盤センター)

Analysis of Japanese Functional Expressions Using a Large Scale Hierarchical Lexicon and Examples

Takafumi Suzuki　(University of Tsukuba)
Yusuke Abe　(University of Tsukuba)
Takehito Utsuro　(University of Tsukuba)
Suguru Matsuyoshi　(University of Yamanashi)
Masatoshi Tsuchiya　(Toyohashi University of Technology)

1. はじめに

機能表現[1]とは,「にあたって」や「をめぐって」のように，2 つ以上の語から構成され，全体として 1 つの機能的な意味をもつ表現である．一方，この機能表現に対して，それと同一表記をとり，内容的な意味をもつ表現が存在することがある．例えば，文 (1) と文 (2) には「にあたって」という表記の表現が共通して現れている．

(1) 出発する にあたって，荷物をチェックした．

(2) ボールは壁 にあたって，跳ね返った．

文 (1) では，下線部はひとかたまりとなって,「機会が来たのに当面して」という機能的な意味で用いられている．それに対して，文 (2) では，下線部に含まれている動詞「あたる」は，動詞「あたる」本来の内容的な意味で用いられている．このような表現においては，機能的な意味で用いられている場合と，内容的な意味で用いられている場合とを識別する必要がある．

本稿では，16,801 表現を収録する日本語機能表現一覧 [松吉 07](以下,「機能表現一覧」[2]) の階層性を利用し，階層において下位に位置する派生的表現について，用法が類似するより上位の代表的表現の用例を参照して，用法判定を行う手法について述べる．特に，前後の形態素の品詞が代表・派生間において不変の場合には，代表的表現と派生的表現の間で用法の傾向に相関がある，という特徴を利用する方式を提案する．さらに，前後の形態素品詞に加え，代表的表現と派生的表現の間で，機能表現の表記を構成する形態素列の品詞パターンの間に派生関係があるという特性を利用する方式を提案する．提案方式に基づいて，派生的表現の用法の分析を行った結果，代表的表現の表記の用法判定済み用例集合 (約 38,000 例) を参照して，派生的表現の表記の用法判定を行うことにより，80%以上の用例の用法を正しく判定できることが分かった．しかし，この手法に関して， [鈴木 11] では処理全体を全自動で行うには至っておらず，人手によって手法の性能を評価する手順が残されており，

*utsuro @ iit.tsukuba.ac.jp
[1]機能表現は，複数形態素からなる複合辞と一つの形態素からなる機能語から構成されるが，本稿では，複合辞と同等の意味で機能表現という用語を用いる．
[2]http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

表 1: 機能表現辞書の 9 つの階層

階層		分類数	表現数 合計 (L^9 表現数)	助動詞型以外	助動詞型
L^1	見出し語	—	341 (488)	281	207
L^2	意味	45/128/199	435 (488)	281	207
L^3	派生（格助詞型，接続助詞型，連体助詞型，接続詞型，助動詞型，形式名詞型，とりたて詞型，提題助詞型）	8	555	348	207
L^4	機能語の交替	—	774	492	282
L^5	音韻的変化	38	1,187	633	554
L^6	とりたて詞の挿入	18	1,810	659	1151
L^7	活用	—	6,870	659	6211
L^8	「です/ます」の有無	2	9,722	895	8827
L^9	表記のゆれ	—	16,801	1360	15411

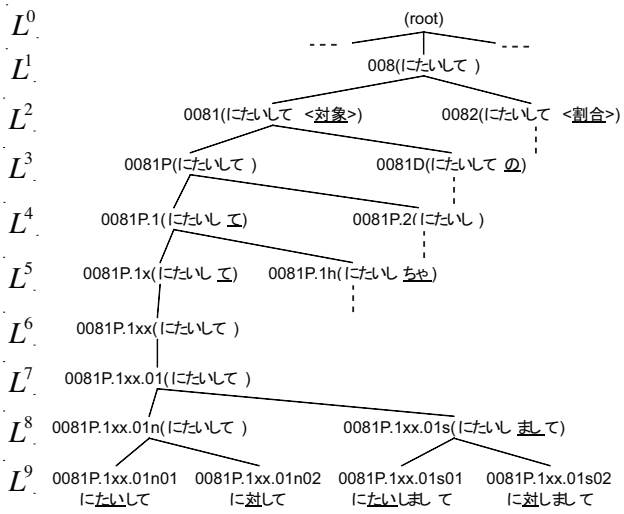

図 1: 機能表現辞書階層構造の一部

その用法判定精度は約 87%であった．一方，本稿では，処理全体を全自動化した結果，その用法判定精度は約 85%となった．

2. 階層的機能表現辞書

「機能表現一覧」 [松吉 07] は，9 つの階層構造をなしており，各階層は，表 1 に示されるような観点によって分類されている．同表に，各階層における機能表現数が示されており，図 1 に階層構造の一部をそれぞれ示す．

3. 派生関係及び用例を利用した日本語機能表現の解析

3.1 代表的な表現の選定

階層の上位に位置する代表的表現は，L^4 階層相当の 1,000 表現程度の規模とする [長坂 08]．そして,「機能表現一覧」において，代表的表現を除く表現を派生的表現と定義する．ただし，代表的表現を選定する際には，以下の制約を課す．

- 機能表現の語頭の無声・有声の制約により前接する活用語の活用型が制限される場合は，この制限を保持する．

- 機能表現の仮名表記・漢字表記の違いを保持する．
- 助動詞型の機能表現の場合には，活用形を保持する．

3.2 派生的な表現の解析方式

本節では，代表的表現の表記の用法判定済用例集合 S_c^{tr} を参照して，派生的表現の表記の用法判定を行う方式について述べる．

3.3 機能表現表記照合個所の表現形式

まず，一文中で，機能表現表記と文字列照合する個所を $e = \langle f, l, r \rangle$ (ただし，f は機能表現表記，l は機能表現表記の先頭の文字位置，r は末尾の文字位置) によって表現する[3]．このとき，評価用の文において機能表現表記 f_{ts} と照合した個所を $e_{ts} = \langle f_{ts}, l_{ts}, r_{ts} \rangle$ とし，e_{ts} に前接する形態素を m_{-1}^{ts}，後接する形態素を m_{+1}^{ts} とする．一般には，f_{ts} の可能性としては，派生的表現の表記 f_d の場合，および，代表的表現の表記 f_c の場合の二通りが考えられる．ここで，f_{ts} が派生的表現 f_d の場合には，f_d の代表的表現 f'_c の用例が，用法判定済用例集合 S_c^{tr} 中の機能表現表記照合個所の一つ $e_{tr} = \langle f'_c, l_{tr}, r_{tr} \rangle$ となる．一方，f_{ts} が代表的表現 f_c の場合には，f_c 自身の用例が，用法判定済用例集合 S_c^{tr} 中の機能表現表記照合個所の一つ $e_{tr} = \langle f_c, l_{tr}, r_{tr} \rangle$ となる．いずれの場合も，e_{tr} に前接する形態素を m_{-1}^{tr}，後接する形態素を m_{+1}^{tr} とする．

ここで，次節の解析手順においては，評価用の文における用法判定対象個所の単位として，相互に重複して連続する複数の機能表現表記から構成される列をひとまとめとして，機能表現表記列の用法判定を一括して行う．具体的には，評価用の文において，連続する 2 個の機能表現表記の文字列のうちの少なくとも一部が重複するような機能表現表記列 $E = e_i, \ldots, e_k$ (すなわち，機能表現表記列 $E = e_i, \ldots, e_k$ 中における連続する任意の 2 個の機能表現表記の組 e_j，e_{j+1} において表記の文字列の少なくとも一部が重複する: $l_j < l_{j+1} < r_j < r_{j+1}$) をひとまとめとする．

3.4 解析手順

まず，評価用の文における用法判定の単位である機能表現表記列 $E = e_i, \ldots, e_k$ に対して，以下の条件**「前後形態素が類似する用法判定済用例の存在」**の成否を判定する．

> **「前後形態素が類似する用法判定済用例の存在」**
> $E = e_i, \ldots, e_k$ 中で，少なくとも一つの機能表現表記照合個所 $e_{ts} = \langle f_{ts}, l_{ts}, r_{ts} \rangle$ に対して，機能表現表記 f_{ts} に対応する機能表現表記照合個所 e_{tr} が用法判定済用例集合 S_c^{tr} 中に存在する．さらに，前接形態素 m_{-1}^{ts} と m_{-1}^{tr}，および，後接形態素 m_{+1}^{ts} と m_{+1}^{tr} の間で，それぞれ，品詞大分類 (IPAdic を用いる) が一致する．

そして，この成否に応じて，下記の手順 (I) もしくは (II) を行う．

(I) **「前後形態素が類似する用法判定済用例の存在」**が成り立たない場合，機能表現表記列 $E = e_i, \ldots, e_k$ 中の全ての機能表現表記が内容的用法であると判定して終了する．

(II) **「前後形態素が類似する用法判定済用例の存在」**が成り立つ場合，以下を行う．

[3]ただし，機能表現表記 f としては,「機能表現一覧」 [松吉 07] における一文字表記の機能語は除外する．

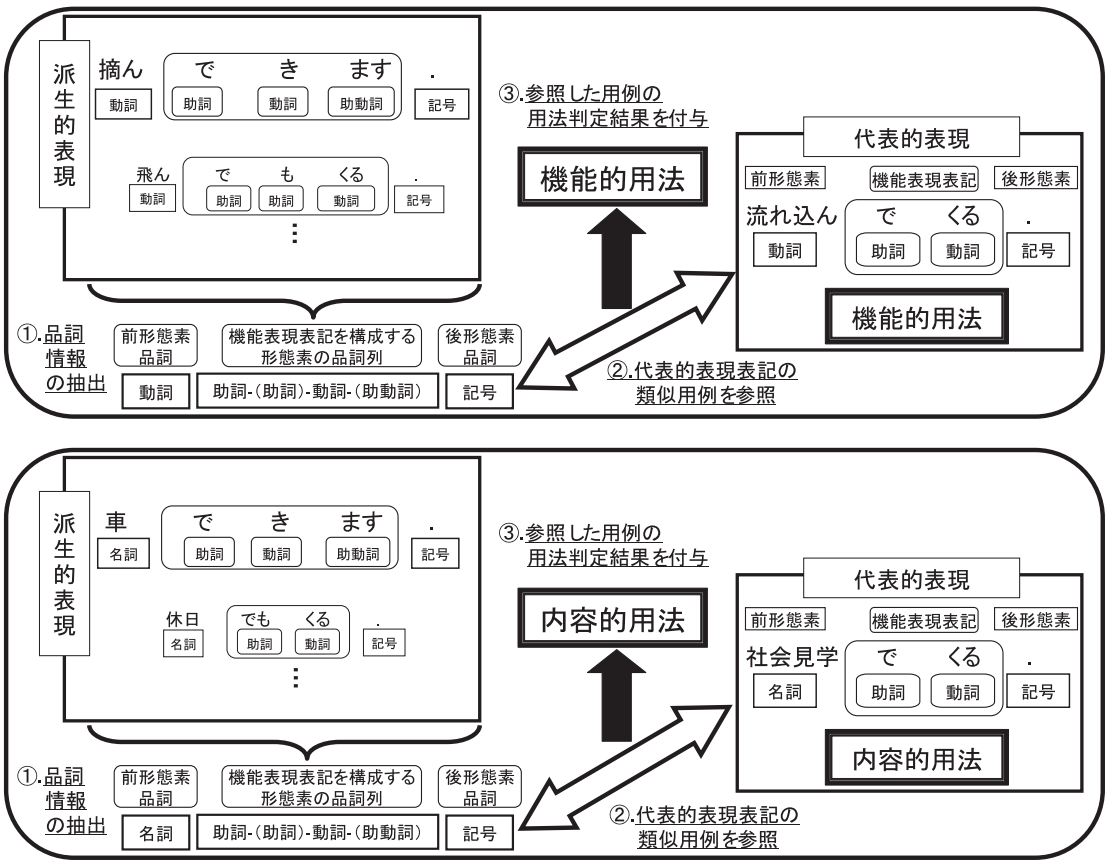

図 2: 模式図: 「代表的表現の表記の用例」を参照して「派生的表現の表記の用例」の用法を判定

(II-i) 条件「機能表現表記列 E において，最長の表記となる照合個所 e_{ts} がただ一つである．さらに，e_{ts} に対して，用法判定済用例集合 S_c^{tr} 中の対応する機能表現表記照合個所 e_{tr} (複数個所の場合もあり得る) を参照することにより，e_{tr} に対する用法判定結果 l_{tr} が一意に決まる.」が成り立つならば，機能表現表記列 E に対して,「e_{ts} の用法は l_{tr}，E 中のその他の機能表現表記の用法は内容的用法」を採用して終了する．その他の場合は，(II-ii) を行う.

(II-ii) 条件**「前後形態素が類似する用法判定済用例の存在」**において,「前後形態素の品詞大分類の一致」の代わりに「前後形態素の品詞細分類が一定以上の基準で類似する」を課し，(II-i) と同様の手順を行う．機能表現表記列 E に対する用法判定結果が一意に決まらない場合には，(II-iii) を行う.

(II-iii) 条件**「前後形態素が類似する用法判定済用例の存在」**において,「前後形態素の品詞大分類の一致」の代わりに「機能表現表記を構成する形態素の品詞列が一定以上の基準で類似する」を課し，(II-i) と同様の手順を行う．機能表現表記列 E に対する用法判定結果が一意に決まらない場合には,「不正解」と判定し終了する.

以上の手順にしたがって，派生的表現の表記の用法が機能的用法であると判定した例の模式図を図 2 上半分に，内容的用法であると判定した例の模式図を図 2 下半分に，それぞれ示す[4].

[4]この解析手順に関する詳細な分析結果は文献 [鈴木 11] に示す.

表 2: 評価結果

(a) 代表的表現の用例を参照する手法

類型		割合 (%)	
「節の手順 (II)」前後の形態素の品詞もしくは機能表現表記を構成する形態素の品詞列の条件を満たす代表的表現の用法判定結果を採用し正解		73.7	82.0
「節の手順 (I)」前後の形態素の品詞が一致する代表的表現が存在しないため，内容的用法と判定し正解		8.3	
不正解	適切な作例をすることにより正解可能	12.7	18.0
	作例しても正解不可能	5.3	
合計		100	

(b)「代表的表現の用例+左・右接続接続情報」を参照する手法

類型		割合 (%)	
「節の手順 (II)」において用法判定済用例の一つとして左・右接続情報を追加して正解		77.4	85.3
「節の手順 (I)」により正解		7.9	
不正解	適切な作例をすることにより正解可能	6.2	14.5
	作例しても正解不可能	8.5	
合計		100	

4. 評価

代表的表現の表記の用法判定済用例としては，毎日新聞 1995 年の 1 年分から収集して人手で機能表現表記の用法判定を行った約 38,000 用例を参照することとする．評価対象としては，同じく毎日新聞 1995 年の 1 年分のうち，機能的用法と内容的用法として適度な割合で新聞記事内に出現する代表的表現に対して，用例数が 10 例以上となる派生的表現を中心に収集した 1,916 用例，及び，機能的用法に偏って新聞記事内に出現する代表的表現に対して，用例数が 50 例未満となる派生的表現を中心に収集した 916 用例の計 2,832 用例 (248 表現) を評価対象とする．

評価結果を表 2(a) に示す．また， 節のいずれかの手順における判定結果が「不正解」となる場合について，代表的表現の適切な用例を作成して用法判定済用例集合 S_c^{tr} に追加した場合に，正解可能か否かの分析を行った結果も併せて示す．この結果から分かるように,「適切な用例の作例なしで正解」となる割合は約 82%，作例を許す場合は約 95%である．

また，表 2(b) には，用法判定済用例集合 S_c^{tr} に対して，用法判定済用例の一つとして，左・右接続情報 [松吉 07, 松吉 08] を追加した場合の評価結果を示す．左・右接続情報とは，機能表現表記の用法が機能的用法である場合の情報である．左接続情報は，直前に接続可能な形態素の情報を示しており，右接続情報[5]は，機能表現表記を構成する末尾の形態素の情報を示したものである．これらは「機能表現一覧」 [松吉 07] において，各機能表現ごとに定義されており，53 種類の左接続情報，および，51 種類の右接続情報が掲載されている．これらの左・右接続情報を追加した場合,「適切な用例の作例なしで正解」となる割合は，約 85%に改善する．

更に，作例をすることで正解となる個所に関し，BCCWJ 2009 年度版 [BCCWJ 総括班 09] 及び，

[5]右接続情報に加えて，IPAdic を用いて形態素解析を行った場合の形態素列の情報を参照することにより，機能表現表記の直後に接続可能な形態素の情報が得られる．

コアデータ [BCCWJ 総括班 11] 中において，作例の代用となる文の有無を調査した．作例をすることで正解となる個所は，評価用例中に 176 用例存在し，その代表的表現数は 43 表現であった．本手法では代表的表現の表記の用法判定済用例として，約 38,000 用例を用いたが，機能表現表記及び前後の形態素の品詞が一致する用例は存在しなかった．一方，BCCWJ 2009 年度版及び, コアデータ（8,356 万語=8,226 万語+130 万語）においては，176 用例中の 82 用例に対して，機能表現表記とその用法，及び前後の形態素の品詞が一致する用例が存在した．以上の結果，用法判定の精度は最大で 2.9%(82 用例分に相当) 改善し，約 88%の精度となる可能性がある．

5. 関連研究

文献 [松吉 08] においては,「機能表現一覧」 [松吉 07] 中の機能表現を対象として，意味を保存する言い換えが可能な機能表現の分類を規定している．その他，内容語と口語的な機能表現を対象として，代表的表現への言い換えを介した機械翻訳の研究 [山本 02]，機能表現の検出・係り受け解析等の解析を対象とした研究 [土屋 07, 注連 07, 小早川 09] がある．

6. おわりに

本稿では,「機能表現一覧」の階層性を利用し，階層において下位に位置する派生的表現について，用法が類似するより上位の代表的表現の用例を参照して，用法判定を行う手法について述べた．

参考文献

[小早川 09] 小早川健, 関場治朗, 木下明徳, 熊野正, 加藤直人, 田中英輝：単語格子とマルコフモデルによる日本語機能表現の解析 — 日本語機能表現辞書「つつじ」を用いて —, 電子情報通信学会技術研究報告, NLC2009-1, pp. 15–20 (2009).

[BCCWJ 総括班 09] 文部科学省科学研究費特定領域研究「日本語コーパス」総括班: BCCWJ 領域内公開データ（2009 年度版） (2009).

[BCCWJ 総括班 11] 文部科学省科学研究費特定領域研究「日本語コーパス」総括班: 特定領域研究「日本語コーパス」研究成果報告 (2011).

[松吉 07] 松吉俊, 佐藤理史, 宇津呂武仁：日本語機能表現辞書の編纂, 自然言語処理, Vol. 14, No. 5, pp. 123–146 (2007).

[松吉 08] 松吉俊, 佐藤理史：文体と難易度を制御可能な日本語機能表現の言い換え, 自然言語処理, Vol. 15, No. 2, pp. 75–99 (2008).

[長坂 08] 長坂泰治, 宇津呂武仁, 土屋雅稔：大規模日本語機能表現辞書の階層性を利用した機能表現検出, 言語処理学会第 14 回年次大会論文集, pp. 837–840 (2008).

[注連 07] 注連隆夫, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史：日本語機能表現の自動検出と統計的係り受け解析への応用, 自然言語処理, Vol. 14, No. 5, pp. 167–197 (2007).

[鈴木 11] 鈴木敬文, 宇津呂武仁, 松吉俊, 土屋雅稔：代表・派生関係および用例を利用した日本語機能表現の解析, 言語処理学会第 17 回年次大会論文集, pp. 155–158 (2011).

[土屋 07] 土屋雅稔, 注連隆夫, 高木俊宏, 内元清貴, 松吉俊, 宇津呂武仁, 佐藤理史, 中川聖一：機械学習を用いた日本語機能表現のチャンキング, 自然言語処理, Vol. 14, No. 1, pp. 111–138 (2007).

[山本 02] 山本和英：換言と言語変換の協調による機械翻訳モデル, 言語処理学会第 8 回年次大会発表論文集, pp. 307–310 (2002).