代表・派生関係を利用した日本語機能表現の解析方式の評価*

鈴木 敬文† 阿部 佑亮† 宇津呂 武仁‡ 松吉 俊§ 土屋 雅稔¶
筑波大学大学院 システム情報工学研究科† 筑波大学 システム情報系‡
山梨大学大学院 医学工学総合研究部§ 豊橋技術科学大学 情報メディア基盤センター¶

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 1 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．

(i) 出発する にあたって，荷物をチェックした．

(ii) ボールは壁 にあたって，跳ね返った．

文 (i) 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 (ii) 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．

これまでの研究で，現代語複合辞用例集 [3](以下，用例集) 中の代表的複合辞一覧に基づいて，それらの派生形である 337 種類の機能表現について，その用例データベース (日本語複合辞用例データベース [11]，以下，用例データベース) が作成された．さらに，それらの用例データベースを訓練事例として，機械学習により機能表現の用法判定・係り受け解析を行う方式が提案された [10, 8]．また，機能表現の異形の語構成パターンを網羅することにより，日本語機能表現一覧 [6](以下,「つつじ」[2]) が作成された．これらを受けて，本論文では,「つつじ」に収録されている，16,801 種類の機能表現を対象として，機能表現表記の用法判定を行うことを目的とする．ここで，[10, 8] の機械学習による機能表現表記の用法判定においては，一つの表現あたり 50 例程度の訓練用例に対して，人手で機能的・内容的等の用法判定を行う必要がある．しかし,「つつじ」の全機能表現 16,801 種類に対して，それだけの規模の作業を行うことは容易ではない．

そこで，本論文では,「つつじ」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の用法判定に際して，用法が類似するより上位の代表的表現の用例を参照することで用法判定を行う手法について述べる．本論文では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，という特徴を利用し，派生的表現の用法判定の際に前後の形態素品詞が一致する代表的表現の用法判定済み用例を参照することでその用法判定を行う．この方式に基づいて，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，85%程度の F 値で正しく判定できることが分かった．さらに,「つつじ」に記述されている左・右接続規則を一部改変した規則を併用することにより，用法判定の精度が改善できることを示す．

2 階層的機能表現辞書

「機能表現一覧」 [6] は，9 つの階層構造をなしており，各階層は,「見出し語、意味、派生、機能後の交替、音韻的変化、とりたて詞の挿入、活用、「です/ます」の有無、表記のゆれ」の観点によって分類されている．

*Evaluating the Method of Analyzing Japanese Functional Expressions using Canonical/Derivational Relation

†Takafumi Suzuki, Yusuke Abe, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Takehito Utsuro, Faculty of Engineering, Information and Systems, University of Tsukuba
§Suguru Matsuyoshi, Dpt. Comp. Sci. and Media Eng., Faculty of Engineering, University of Yamanashi
¶Masatoshi Tsuchiya, Information and Media Center, Toyohashi University of Technology

[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，本論文では，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/

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 基本的な考え方

本節では，機能表現の解析における本論文の基本的な考え方について述べる．本論文では，2 節で述べた「つつじ」の持つ階層構造を利用することにより，階層構造の上位にある機能表現を代表的表現，その代表的表現から下位に派生している表現を派生的表現として扱う．そして，解析に用いる知識源としては代表的表現の用法を判定した用例を用意し，派生的表現はそれらを参照することにより解析を行う．

ここで,「つつじ」には約 17,000 もの機能表現が収録されており，その全てに対して，解析のための用法判定済み用例を用意するのは，容易ではない．しかし，L^1〜L^4 の比較的上位の階層の表現種類数は 1,000 未満であり，これらのうちのいずれかの階層までの表現を代表的表現として，それらの代表的表現の用法判定済用例を参照して，残りの派生的表現の用法を解析することができるならば，派生的表現の用法判定済用例を作成する労力を省略することができる．

例えば，代表的表現を L^4 階層の表現とした場合は，代表的表現の表現数は 800 程度であり,「つつじ」の全表現に対して，用法判定済用例の作成作業に必要な労力は約 1 / 20 になる．また，より上位の階層の表現を代表的表現とする場合は，用法判定済用例の作成作業に必要な労力は更に少なくて済む．

この方式の具体例を図 1 に示す．この例においては,「なくてもいい」，および,「てもいい」という二つの派生的表現の表記に対して，用法判定済用例データベース中の代表的表現の表記の用例が照合し，結果的に，表記の文字長の大きい「なくてもいい」の代表的表現「なくていい」の機能的用法が採用されている．

3.2 代表的表現の選定

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 [7]．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．

3.3 用例に基づく解析の方式

本節では,「用例に基づく解析の方式」の概要を説明する．この方式は以下の 2 つのステップから構成される．

1. 個々の機能表現表記に対する仮説の生成
2. 構成形態素を共有する複数の機能表現表記に対する解析

ステップ 1 では，文中に出現する機能表現表記に対して，機能的用法，および，内容的用法の両方の用法を仮説として設定し，出現文脈の類似する代表的表現の用法判定済み用例集合を参照して，それぞれの仮説を検証する．ステップ 2 では，ステップ 1 で処理した個々の機能表現表記のうち，構成形態素を共有する (出現箇所が重複する) 複数の機能表現表記に対して，主に機能表現表記の文字長に基づいて，最も適切な仮説を絞り込む．

4 評価

4.1 代表的機能表現表記の用例データベース及び評価対象文集合

本節では，用例に基づく解析方式における用例データベース，および，評価対象文集合について述べる．まず，代表的機能表現表記の用法判定済み用例データベースとして，毎日新聞 1995 年の 1 年分から収集した文に対して，人手で機能表現表記の用法判定を行った約 38,000 用例を用いる．また，評価対象文集合としては，同じく毎日新聞 1995 年の 1 年分から，上記の代表的表現の用法判定済み用例と重複しない 2,832 用例 (248 表現) を選定し，評価対象文集合とする．

4.2 評価手順

評価においては，評価対象文集合の用例に対し，システムが出力を行った個所に対して，人手で正解の用法判定ラベルを付与し，その成否を評価するものとする．また，評価尺度として，以下で定義する適合率，再現率，F 値を用いる．

$$\text{適合率} = \frac{\text{システムが出力した用法判定結果のうち正解した箇所数}}{\text{システムが出力した用法判定結果の総箇所数}}$$

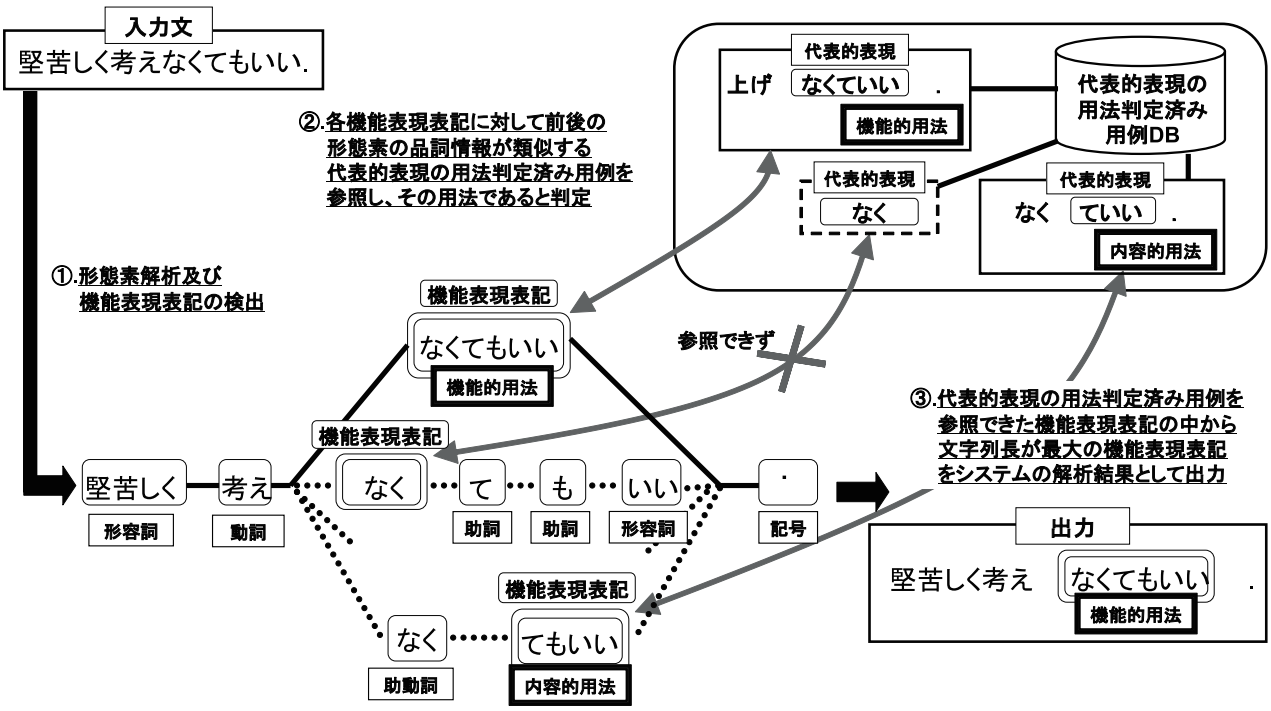

図 1: 模式図: 「代表的表現の表記の用例」を参照して「派生的表現の表記の用例」の用法を判定

$$\text{再現率} = \frac{\text{システム出力した用法判定結果のうち正解した箇所数}}{\text{人手で付与した正解の総箇所数}}$$

$$\text{F 値} = \frac{2 * \text{再現率} * \text{適合率}}{\text{再現率} + \text{適合率}}$$

4.3 評価結果

評価結果を表 1(a) の最上段に示す．また，3.3 節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例して用法判定済用例集合に追加した場合に，正解可能か否かの分析を行った結果も併せて表 1(a) の中段に示す．この結果からわかるように，適切な作例を行わない場合の F 値は約 85%，作例を許す場合の F 値は約 95%である．また，これらに対するベースラインとして，評価対象の機能表現表記を全て機能的用法と判定した場合 (ただし，複数の機能表現表記の出現箇所が重複する場合は無条件に最長の表記を選択) の評価結果を表 1(a) の最下段に示す．

また，表 1(b) 上半分には，用法判定済用例集合に対して，用法判定済用例の一つとして，左・右接続情報 [6, 5] を一部改変して追加した場合の評価結果を示す．同様に，左・右接続情報に加えて代表的表現の適切な用例を作例して用法判定済用例集合に追加した場合に，正解可能か否かの分析を行った結果も併せて表 1(b) の下半分に示す．ここで，左・右接続情報とは，機能表現表記の用法が機能的用法となる場合の前後の形態素についての情報である．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[3]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」 [6] 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，F 値は約 88%に改善する．更に作例を許す場合には，F 値は約 91%となる．

左・右接続情報を追加し作例を許す場合，左・右接続情報を追加しないで作例を許す場合に比べ，F 値が下がっているが，これは，左・右接続情報が複数の表記にまたがって作成された規則であり，個々の機能表現表記に対しては条件の緩い規則となっているためである．つまり，代表的表現の用例を参照する場合と比べると，前後の形態素に課す条件が緩くなる傾向にあるため，代表的表現の用例を参照する場合に比べて不正解となる個所が増える．

[3]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．

表 1: 評価結果

(a) 左・右接続情報を参照しない場合

類型	適合率 (%)	再現率 (%)	F 値 (%)
3.3 節の手順に従い，代表的表現の用例を参照する手法	85.9	83.8	84.8
作例した用例を用法判定済み用例データベースに追加して，3.3 節の手順に従い，代表的表現の用例を参照する手法	96.4	93.9	**95.1**
ベースライン	78.4	77.3	77.8

(b) 左・右接続情報を参照する場合

類型	適合率 (%)	再現率 (%)	F 値 (%)
「左・右接続情報」を利用する場合	89.4	86.4	**87.9**
「左・右接続情報+作例」を利用する場合	93.4	90.3	91.2

5 関連研究

文献 [5] においては,「機能表現一覧」 [6] 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 [12]，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 [10, 8, 1] がある．また，文献 [9] では，大規模な均衡コーパスである「現代日本語書き言葉均衡コーパス」(2009 年度板) [2] において，機械学習による複合辞の検出を行った．

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．また，提案手法により 88%程度の F 値で機能表現表記を正しく用法判定できることを示した．

また，提案手法との比較として，本論文で述べた代表・派生関係を利用し，条件付き確率場 (CRF：Conditional Random Fields) [4] を用いた機械学習による機能表現のチャンキングを行う方式の評価を行った．機械学習による機能表現のチャンキングの研究事例 [10, 8] において，その対象は 59 種類の機能表現表記のみであり，1 表現に必要な訓練事例は 50 例程度とされていた．このことから，約 17,000 の表現に対して必要とされる訓練事例数は，850,000 例となるが，これを人手で作成するためには膨大なコストを要する．そこで，このタスクにおいて，代表・派生関係を利用することにより，人手コストの問題を解消する．具体的には，訓練事例中の代表的表現を人工的に派生的表現に置換した訓練事例を作成し，派生的表現のチャンキングモデルの学習を行う方式を導入する．この方式においては，実際に人手により訓練事例を作成するためのコストは，代表的表現を対象としたものだけとなる．この方式に従い，提案手法と同テストデータで評価した結果，評価結果は F 値で 75%程度であった．

参考文献

[1] 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.

[2] 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.

[3] 国立国語研究所: 現代語複合辞用例集. 2001.

[4] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th ICML*, pp. 282–289, 2001.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 鈴木敬文, 阿部佑亮, 宇津呂武仁, 松吉俊, 土屋雅稔. 『現代日本語書き言葉均衡コーパス』における複合辞の検出と評価. 『コーパス日本語学ワークショップ』予稿集, 2012.

[10] 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.

[11] 土屋雅稔, 宇津呂武仁, 松吉俊, 佐藤理史, 中川聖一. 日本語複合辞用例データベースの作成と分析. 情報処理学会論文誌, Vol. 47, No. 6, pp. 1728–1741, 2006.

[12] 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.