『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

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　(筑波大学システム情報系)*
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)

Detection of Compound Functional Expressions in "Balanced Corpus of Contemporary Written Japanese" and its Evaluation

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．

我々は，このような日本語機能表現の解析の課題に対して，これまでに，国立国語研「現代語複合辞用例集」 [国研 01] に収録されている 125 機能表現の異表記を展開した 300 表現について，新聞記事中の用例に対して機能的用法・内容的用法を判別した用例データベース [土屋 06] を作成・公開した．また，機能的用法・内容的用法の自動判別ツールを作成し，係り受け解析ツールとの統合により，複合辞としての機能的用法を考慮した係り受け解析を実現した [注連 07]．また，日本語機能表現の全表記を網羅した辞書として，日本語機能表現の全表記約 17,000 を網羅的に収録した「つつじ」 [松吉 07, 松吉 08][2] が公開されたのを受けて，日本語機能表現の全表記約 17,000 を網羅的に収録した辞書「つつじ」の階層的構造および言語学的特性を活用して，全 17,000 表記を対象とした網

*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．本論文において対象とする機能表現は，いずれも複数形態素から構成される複合辞に相当するため，本論文においては，複合辞と同等の意味で機能表現という用語を用いる．また，本論文では，特に，機能表現を構成する表記が，複合辞として用いられる機能的用法となる場合と，複合辞を構成する形態素が本来の意味で用いられている内容的用法となる場合の曖昧性を持つ場合に注目する．

[2]http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

表 1: 「(1) 複合辞として一長単位を構成する」または「(2) 複数の長単位から構成される」という曖昧性を持つ短単位列の例

	表記	短単位列	長単位	意味	例文
(1)	に当たって	に (助詞) +当たっ(動詞) +て (助詞)	に当たって (助詞)	状況	証券化 **に当たって**、より有利な 商品設計が可能である点など でメリットがあると考えられる。
(2)			に (助詞) +当たっ(動詞) +て (助詞)	-	… 紫外線 **に　当たっ　て** も 分解されず、においの 成分が長もちするんです」
(1)	ことがあっ	こと (名詞) +が (助詞) +あっ(動詞)	ことがあっ (助動詞)	経験	誰かに似ているな、と 彼はこれまでにも時折 思った **ことがあっ** たが、…
(2)			こと (名詞) +が (助詞) +あっ(動詞)	-	どんな人にも … 、得意なことと 苦手な **こと　が　あっ** て、…
(1)	ところが	ところ (名詞) +が (助詞)	ところが (接続詞)	逆接	**ところが**、それは意外に素早く 簡単に済んだのだった。
(2)			ところ (名詞) +が (助詞)	-	… 部屋の北側に一段高い **ところ　が** あって、…

羅的な日本語機能表現表記の用法判定 [鈴木 12]，および，日本語機能表現の集約的翻訳の枠組み [島内 10, Nagasaka10, 阿部 12] を提案した．

ここで，これらの研究のうち，日本語機能表現表記の機能的用法・内容的用法の分析および自動判定手法の研究 [土屋 06, 注連 07, 鈴木 12] は，いずれも，新聞記事という限定されたジャンルのコーパスを対象としたものであった．そこで，本研究では，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』において，上述した機能的用法・内容的用法の曖昧性を持つ機能表現表記を対象として，機械学習により用法判定を行う手法を適用し，その性能を評価した結果を報告する．具体的には,『現代日本語書き言葉均衡コーパス』コアデータ [BCCWJ 総括班 09] を対象として，複合辞となり得る表記 (機能表現表記) を構成する短単位列が,「全体として 1 つの機能的長単位」となるのか，それとも,「複数の長単位から構成される列」となるのかという曖昧性を解消することを目的とする．適用する手法としては，条件付き確率場 (Conditional Random Fields, CRF) [Lafferty01] を利用したチャンキングを用い，ツールとしては CRF++[3]を用いた．評価実験の結果，機能的な長単位の検出において 97%近い F 値を達成した．

2. 『現代日本語書き言葉均衡コーパス』における複合辞

2.1 複合辞の短単位列・長単位の分析

本研究ではまず,『現代日本語書き言葉均衡コーパス』(以下，BCCWJ) コアデータ [BCCWJ 総括班 09] を用いて，複合辞となり得る表記 (機能表現表記) を構成する短単位列について，以下の曖昧性の有無を調査した．

1. 短単位列が全体として一つの機能的長単位 (=複合辞) を構成する．

2. 一つの短単位が一つの長単位を構成する．

[3] http://crfpp.sourceforge.net/

表 1 にこれらの曖昧性の例を示す．例えば，表中の「ところが」が，機能的な長単位となる場合は，非構成的に「逆接」の意味で用いられており，その品詞は接続詞である．一方，複数の長単位から構成される長単位列となる場合は,「ところ」は場所を表す名詞,「が」は格助詞として用いられている．

2.2 検出対象の複合辞の選定

前節を踏まえて，本節では，検出対象となる機能表現表記 (複合辞となり得る短単位列) の選定手順およびその結果について述べる．

BCCWJ においては，[小椋 11] において，助詞相当 75 語，助動詞相当 55 語の複合辞が収録されているが，これらは，

1. [国研 01, グループ・ジャマシイ 98, 森田 89] における収録状況に基づき，重要度を判断．
2. BCCWJ において一定以上の頻度 (50 程度) がある．
3. BCCWJ において，機能的用法の割合が 80%程度である．
4. 複合辞前後の短単位から，機能的用法であると判定できる．

という条件のもと選定されたものである．

一方，本論文では，前節で述べたように,「複合辞として一長単位を構成する」，もしくは,「複数の長単位から構成される」の曖昧性を持つ短単位列に対して，その曖昧性の解消を目的とする．そこで，まず，以下の手順により，検出対象とする短単位列 (機能表現表記) を選定した．

1. BCCWJ コアデータを対象として，品詞が助詞，助動詞，接続詞[4]となる長単位を列挙することにより，1,010 種類の長単位が得られた．
2. 文字長が一文字である長単位，78 種類を除外し，932 種類となった．
3. 口語調の崩れた日本語や誤字，古語，方言など，合計 213 種類を人手で除外し，719 種類となった．
4. 719 種類の長単位に対して，短単位列の種類数は 727 種類となった．これらの短単位列のうち,「複合辞として一長単位を構成する」，もしくは,「複数の長単位から構成される」の曖昧性を持つ短単位列は，201 種類 (助詞 63 種類，助動詞 112 種類，接続詞 26 種類) となった．なお，これらの短単位列のうち，[小椋 11] において選定された複合辞，および，その他連語と重複する短単位列の数は，助詞 60 種類，助動詞 70 種類，接続詞 18 種類の合計 148 種類 ([小椋 11] での単位に従い，表記の揺れ，活用形の違いをまとめた場合は，助詞 43 種類，助動詞 25 種類，接続詞 16 種類の合計 84 種類) である．

上記の手順により除外した長単位の一例を表 2 に示す．

3. 条件付き確率場を用いたチャンキングによる複合辞の検出

3.1 条件付き確率場

本論文では,「複合辞として一長単位を構成する」，もしくは,「複数の長単位から構成される」という曖昧性を持つ短単位列を対象として，複合辞としての長単位を検出する手法として，条件付き確率場 (Conditional Random Fields, CRF) [Lafferty01] を適用する．CRF は正しいラベル系列を他の

[4]「つつじ」においては接続詞型の機能表現が収録されていることをふまえて，本論文でも，接続詞の長単位を検出の対象とする．なお，これらは，[小椋 11] においては、その他連語 86 語として選定されている．

表 2: 除外した長単位 (短単位列) の例

分類	例
長単位としての表記長 1 字	ぬ (助動詞)，た (助動詞)
古語	てゐる (助動詞)，といふ (助詞)
誤字	て ”くる (助動詞)，な　かっ(助動詞)
口語調の崩れた日本語	てくださぁ〜ィ(助動詞)，でぇえええええええす (助動詞)
方言	でっしゃろ (助動詞)，のお (助詞)
長単位としての 曖昧性を持たない	にもかかわらず (助詞)，かも知れない (助動詞)， あるいは (接続詞)

全ラベル系列の候補と弁別する学習を行う．本論文では，CRF による学習・解析用のツールとして CRF++[5]を利用する．正規化項としては，L1 正則化，L2 正則化，MIRA の 3 通りを評価し，最も性能のよかった L1 正則化を採用した．

3.2 チャンキングタグの表現法

本論文では，短単位を最小単位として，検出対象とする機能表現表記を構成する短単位列に対して，共通のチャンクタグを付与するという手順で，機能的な長単位の検出を行う．チャンクタグは，そのチャンクタグが付与された短単位が，検出対象とする機能的な長単位のいずれかに含まれるか否かを表し，チャンクの範囲を示す要素によって表現される．チャンクタグの範囲を示す要素の表現法としては，以下で示す IOB2 フォーマット [Tjong Kim Sang00] を使用する．

I 機能的な長単位 (=複合辞) を表すチャンクに含まれる短単位 (先頭以外)

O 機能的な長単位 (=複合辞) を表すチャンクに含まれない短単位

B 機能的な長単位 (=複合辞) を表すチャンクの先頭の短単位

3.3 素性

学習・解析に用いる素性は，[土屋 07,注連 07] で用いられているものに従う．また，本論文では，[土屋 07,注連 07] で述べている形態素の情報を，全て短単位における該当情報に置き換えるものとする．

文頭から i 番目の短単位 m_i に対して与えられる素性 F_i は，形態素素性 $MF(m_i)$，チャンク素性 $CF(i)$，チャンク文脈素性 $OF(i)$ の 3 つ組として，次式によって定義される．

$$F_i \quad = \quad \langle MF(m_i), CF(i), OF(i) \rangle$$

[土屋 07, 注連 07] では，形態素素性 $MF(m_i)$ は，形態素解析によって形態素 m_i に付与される情報である．IPA 品詞体系の形態素解析用辞書[6]に基づいて動作する形態素解析器 ChaSen[7]による形態素解析結果を入力としているため，以下の 10 種類の情報 (表層系，品詞，品詞細分類 1〜3，活用型，活用形，原形，読み，発音) を形態素素性として用いていた．

[5] http://crfpp.sourceforge.net/
[6] http://sourceforge.jp/projects/ipadic/
[7] http://chasen-legacy.sourceforge.jp/

一方，本論文では，これに対応するものとして，UniDic[8]の品詞体系に従い付与された，BCCWJのコアデータ中の短単位における以下の 10 種類 (書字体，品詞，品詞細分類 1~3，活用型，活用形，語彙素，仮名形，発音形) を利用する．

チャンク素性 $CF(i)$ とチャンク文脈素性 $OF(i)$ は，i 番目の位置に出現している機能表現表記に基づいて定まる素性である．下図の短単位列 $m_j \cdots m_i \cdots m_k$ からなる機能表現候補 E が存在したとする．

$$m_{j-2} \quad m_{j-1} \quad \underbrace{\boxed{m_j \ldots m_i \ldots m_k}}_{\text{機能表現表記 } E} \quad m_{k+1} \quad m_{k+2}$$

チャンク素性 $CF(i)$ は，i 番目の位置に出現している機能表現表記 E を構成している短単位の数 (機能表現表記の長さ) と，機能表現表記中における短単位 m_i の相対的位置の情報の 2 つ組である．チャンク文脈素性 $OF(i)$ は，i 番目の位置に出現している機能表現表記の直前の 2 つの短単位および直後の 2 つの短単位の形態素素性とチャンク素性の組である．すなわち，i 番目の位置に対する $CF(i)$ および $OF(i)$ は次式で表される．

$$\begin{aligned} CF(i) &= \langle k-j+1,\ i-j+1 \rangle \\ OF(i) &= \langle\ MF(m_{j-2}), CF(m_{j-2}), MF(m_{j-1}), CF(m_{j-1}), \\ &\qquad MF(m_{k+1}), CF(m_{k+1}), MF(m_{k+2}), CF(m_{k+2})\ \rangle \end{aligned}$$

素性の詳細な定義については，[土屋 07, 注連 07] を参照されたい．

4. 評価

4.1 評価手順

本節では，評価手法，及び，データセットに関して述べる．本論文では，評価にあたり，BCCWJコアデータ中の 50,693 文のうち，201 種類の機能表現表記 (短単位列) を含む 37,231 文を利用して 10 分割交差検定を行った．表 3 に示すように，評価対象となる個所は，2.2 節で選定した 201 種類の短単位列が出現する個所で，それらの個所に対応する長単位の総数は，48,178 個である．本論文では，これらの長単位の個所を基本単位として，以下で定義する適合率，再現率，F 値を測定し，評価尺度として用いる．

$$\begin{aligned} \text{適合率} &= \frac{\text{検出に成功した長単位数}}{\text{システムによって検出された長単位数}} \\ \text{再現率} &= \frac{\text{検出に成功した長単位数}}{\text{評価データに存在する評価箇所の長単位数}} \\ \text{F 値} &= \frac{2 * \text{再現率} * \text{適合率}}{\text{再現率} + \text{適合率}} \end{aligned}$$

4.2 評価結果

評価結果を表 4 (1) に示す．短単位列全体が一つの長単位となる個所に対しては，97%近い F 値を達成した．また，ベースラインとして，全ての評価対象個所に対して，「短単位列が全体として一つの長単位となる」として判定した場合の評価結果を表 4 (2) に示す．結果として，全ての評価対象個所に対しては，75%程度の F 値にとどまり，提案手法はベースラインに対して，F 値を 20%以上改善した．

[8] http://www.tokuteicorpus.jp/dist/

表 3: 評価箇所に関する統計情報

単位	内訳		総計
	短単位列を構成する 複数の短単位が それぞれ一つの長単位となる	短単位列を構成する 複数の短単位が 1 つの長単位となる	
短単位の数を集計	19,274 (23.3%)	63,349 (76.7%)	82,596 (100%)
長単位の数を集計	19,274 (40.0%)	28,904 (60.0%)	48,178 (100%)

表 4: 評価結果

(1)　CRF によって出力された長単位

類型	適合率 (%)	再現率 (%)	F 値 (%)
タスク 1: 短単位列が全体として 一つの長単位となる個所を検出	96.1	97.5	96.8
タスク 2: 短単位列を構成する短単位それぞれが 一つの長単位となる個所を検出	97.9	90.5	94.0
合計 (タスク 1 + タスク 2)	96.8	94.7	95.7

(2)　ベースライン：評価対象の短単位列を全て「一つの長単位」となると判定

類型	適合率 (%)	再現率 (%)	F 値 (%)
タスク 1: 短単位列が全体として 一つの長単位となる個所を検出	60.0	100	75.0
タスク 2: 短単位列を構成する短単位それぞれが 一つの長単位となる個所を検出	0	0	0
合計 (タスク 1 + タスク 2)	60.0	60.0	60.0

5. 関連研究

[首藤 88, 首藤 98, Shudo04] は，機能表現や慣用表現を含む複数の形態素からなる定型的表現をできるだけ網羅的に収集し，機能表現間に類似度を定義して，機能表現の言い換えや機械翻訳に利用することを提案している．特に，文献 [Shudo04] では，機能表現を検出することを目的として，機能的用法と内容的用法を識別するための規則を人手で作成している．しかし，人手で規則を作成するにはコストがかかるため，網羅できる機能表現の規模には限界がある点が課題であると言える．

一方，我々は，[鈴木 12] において,「つつじ」 [松吉 07] の全 16,801 表現を対象とした方式を提案している．この方式においては,「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っている．

また，[松吉 08] においては,「つつじ」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，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]，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方式 [山本 02] 等が知られている．同様に,「つつじ」 [松吉 07] の機能表現を

対象として，代表的表現への言い換えを介した機械翻訳を行う手法の研究事例として，日本語文型辞典 [グループ・ジャマシイ 98] 中の例文を対象とした集約的英訳 [坂本 09]，特許文を対象とした集約的英訳 [島内 10, Nagasaka10, 阿部 12]，及び集約的中国語訳 [劉 10] についての手法が提案されている．

6. おわりに

本論文では，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』において，機能的用法・内容的用法の曖昧性を持つ機能表現表記を対象として，機械学習により用法判定を行う手法を適用し，その性能を評価した結果を報告した．具体的には，『現代日本語書き言葉均衡コーパス』コアデータ [BCCWJ 総括班 09] を対象として，複合辞となり得る表記 (機能表現表記) を構成する短単位列が，「全体として 1 つの機能的長単位」となるのか，それとも，「複数の長単位から構成される列」となるのかという曖昧性を解消することを目的とした．適用する手法としては，条件付き確率場 (Conditional Random Fields, CRF) [Lafferty01] を利用したチャンキングを用い，評価実験の結果，機能的な長単位の検出において 97%近い F 値を達成した．

参考文献

[阿部 12] 阿部佑亮, 鈴木敬文, 宇津呂武仁, 山本幹雄, 松吉俊, 河田容英：対訳用例および意味的等価クラスを用いた機能表現の日英翻訳, 言語処理学会第 18 回年次大会論文集 (2012).

[グループ・ジャマシイ 98] 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).

[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).

[国研 01] 国立国語研究所: 現代語複合辞用例集 (2001).

[Lafferty01] Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. 18th ICML*, pp. 282–289 (2001).

[劉 10] 劉颯, 長坂泰治, 宇津呂武仁, 松吉俊：意味的等価クラスを用いた日本語機能表現の集約的日中翻訳規則の作成と分析, 言語処理学会第 16 回年次大会論文集, pp. 194–197 (2010).

[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).

[森田 89] 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).

[Nagasaka10] Nagasaka, T., Shimanouchi, R., Sakamoto, A., Suzuki, T., Morishita, Y., Utsuro, T. and Matsuyoshi, S.: Utilizing Semantic Equivalence Classes of Japanese Functional Expressions in Translation Rule Acquisition from Parallel Patent Sentences, *Proc. 7th LREC*, pp. 1778–1785 (2010).

[小椋 11] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 小西光, 原裕：『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版 (下)』, 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書 (2011).

[坂本 09] 坂本明子, 宇津呂武仁, 松吉俊：日本語機能表現の集約的英訳, 言語処理学会第 15 回年次大会論文集, pp. 654–657 (2009).

[島内 10] 島内蘭, 長坂泰治, 坂本明子, 宇津呂武仁, 松吉俊：日英特許翻訳における日本語機能表現の集約的英訳可能性の調査, 言語処理学会第 16 回年次大会論文集, pp. 611–614 (2010).

[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).

[首藤 88] 首藤公昭, 吉村賢治, 武内美津乃, 津田健蔵：日本語の慣用的表現について—語の非標準的用法からのアプローチ—, 情報処理学会研究報告, 第 1988-NL-66 巻, pp. 1–7 (1988).

[首藤 98] 首藤公昭, 小山泰男, 高橋雅仁, 吉村賢治：依存構造に基づく言語表現の意味的類似度, 電子情報通信学会研究報告, 第 NLC98-30 巻, pp. 33–40 (1998).

[Shudo04] Shudo, K., et al.: MWEs as Non-propositional Content Indicators, *Proc. 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pp. 32–39 (2004).

[鈴木 12] 鈴木敬文, 阿部佑亮, 宇津呂武仁, 松吉俊, 土屋雅稔：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, 言語処理学会第 18 回年次大会論文集 (2012).

[Tjong Kim Sang00] Tjong Kim Sang, E.: Noun Phrase Recognition by System Combination, *Proc. 1st NAACL*, pp. 50–55 (2000).

[土屋 06] 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).

[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).