機械学習を用いた日本語機能表現のチャンキング

高木 俊宏 [†1]　注連 隆夫 [†2]　土屋雅稔 [†3]
内元 清貴 [†4]　松吉 俊 [†2]　宇津呂 武仁 [†2]　佐藤 理史 [†5]

[†1] 京都大学 工学部 電気電子工学科　[†2] 京都大学 情報学研究科
[†3] 豊橋技術科学大学 情報メディア基盤センター
[†4] 情報通信研究機構　[†5] 名古屋大学大学院 工学研究科

1. はじめに

機能表現とは、「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る内容表現が存在する。例えば、「にあたって」という表現は、「出発するにあたって、荷物をチェックした」という文では、「機会が来たのに当面して」という意味に相当する機能表現であるが [1]、「ボールが壁にあたって跳ね返った」という文では内容表現である。このような表現においては、機能表現としての非構成的用法と、内容語としての構成的用法とを識別する必要がある。

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(ver.5.1) と構文解析器 KNP(ver.2.0) 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、機能表現である可能性を考慮した解析を行っている。本稿の対象とする 52 種類の表現の内、JUMAN/KNP によって機能表現を考慮した解析が行われる可能性がある表現は 31 種類あるが、この 31 種類の表現の全ての用例において、機能的用法と内容的用法の識別が正しく行えているわけではない。また、ChaSen(ver.2.3.3) と構文解析器 CaboCha(ver.0.52) の組合わせ☆☆を考える。その際、形態素解析器は IPA 品詞体系（THiMCO97）の形態素解析用辞書 (ver.2.6.1) を用い、構文解析器は、京都テキストコーパス [2] から機械学習したモデルを用いるとする。この場合、形態素解析用辞書に「助詞・格助詞・連語」と登録されている機能表現は、形態素解析時に検出される。また、「ざるを得ない」などの表現は直前の文節の一部としてまとめられ、機能的な表現として解析される。本稿で対象としている 52 種類の表現の内、ChaSen/CaboCha によって機能表現を考慮した解析が行われる可能性がある表現は 26 種類あるが、ChaSen/CaboCha の場合も、この 26 種類の表現の全ての用例において、機能的用法と内容的用法の識別が正しく行えているわけではない。

このような現状を改善するには、機能表現である可能性がある形態素列 (機能表現候補) の用法を正しく識別する検出器が必要である。これまで、形態素解析結果に基づいて人手で作成した規則によって機能表現を検出する手法も提案されてきた [3]。しかし、これらの手法では検出規則を人手で作成するのに多大なコストが必要であり、検出対象とする機能表現集合の規模の拡大に対して追従が困難である。そこで、本稿では、機能表現検出をチャンク同定問題として定式化し、SVM[4] を用いたチャンカー YamCha(ver.0.32) ☆☆☆を利用した日本語機能表現の検出器を提案する。日本語複合辞用例データベース [5] を訓練データとして学習した日本語機能表現検出器によって、既存の解析系、および、土屋らが提案した手法 [3] と比べ、機能表現を高精度に検出できることを示す。

2. 日本語機能表現の検出

2.1 日本語複合辞用例データベース

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表現とは区別している。現代語複合辞用例集 [1](以下、複合辞用例集と呼ぶ) は、主要な 125 種類の複合辞について、用例を集成し、説明を加えたものである。

日本語複合辞用例データベース [5](以下、用例データベースと呼ぶ) は、機能表現の機械処理を研究するための基礎データを提供することを目的として設計・編纂されたデータベースである。用例データベースは、複合辞用例集で扱っている 125 種類の複合辞およびその異形 (合計 337 種類の機能表現) を対象として、機能表現候補と一致する文字列のリストと、個々の機能表現候補に対して最大 50 個の用例を収録している。そして、各機能表現候補が文中において果たしている働きを、表 1 に示す 6 種類の判定ラベルのうちから人手で判定し、付与している。ラベル F は、複合辞用例集で説明されている用法の表現 (すなわち、複合辞) に付与される。また、機能表現は、ラベル F、A、M に相当する。

☆ http://www.kc.t.u-tokyo.ac.jp/nl-resource/{juman-e.html, knp-e.html}

☆☆ http://chasen.naist.jp/hiki/ChaSen/, http://www.chasen.org/~taku/software/cabocha/

☆☆☆ http://www.chasen.org/~taku/software/yamcha/

表 1 判定ラベル体系

F	用例集 [1] で説明されている用法
A	接続詞的用法
M	その他の機能的用法
C	内容的用法
Y	読み不一致
B	判定単位が不適切

2.2 チャンキングによる定式化

機械学習を用いて機能表現を検出する場合、機能表現検出をクラス判別として定式化するアプローチと、チャンキングとして定式化するアプローチが考えられる。クラス判別として定式化する場合は、機能表現検出は、機能表現となる可能性がある候補部分を単位として、その候補部分の用法を分類するという手順になる。しかし、このアプローチでは、以下のように、一部分が重複して出現している複数の機能表現に対して矛盾した検出を行ってしまう可能性がある。

(1) 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。

(2) それが試合 という ものの 難しさだ。

例文の下線部は、機能表現となる可能性がある候補部分である。文（1）では、機能表現として検出される可能性がある候補部分は「にあって」と「ても」の 2 つであるが、これらは一部が重なっている。そのため、クラス判別によるアプローチでは、2 つの候補部分が同時に機能表現であるという矛盾した判定が行われる可能性がある。文（2）の場合、2 つの候補部分「という」と「というものの」は、包含関係にあり、文（1）と同様に、2 つの候補部分が同時に機能表現であると矛盾した判定が行われる可能性がある。このような問題が発生した場合、検出された複数の機能表現を、なんらかの指標に従って 1 つに絞り込まなければならない。それに対して、機能表現検出をチャンキングとして定式化した場合は、形態素を単位として、どのような機能表現であるか否かの判断を行うため、このような問題は発生しない。

そのため、本研究では、機能表現検出をチャンキングとして定式化する。

3. SVM を用いたチャンキングによる機能表現検出

3.1 チャンクタグの表現法

機能表現の検出時に付与するチャンクタグは、2 つの要素を用いて表記されるものを使用する。ひとつは、チャンクの範囲を示す要素であり、もうひとつは、チャンクの用法を示す要素である。

チャンクの範囲を示す要素の表現法としては、以下で示すような IOB2 フォーマット [7] が広く利用されている。本研究でも、この IOB2 フォーマットを使用する。

- **I** チャンクに含まれる形態素 (先頭以外)
- **O** チャンクに含まれない形態素
- **B** チャンクの先頭の形態素

チャンクの用法を示す要素の表現法としては、様々なものが考えられるが、予備実験の結果、いずれの表現法を用いても大きな性能の差は見られなかったため、本論文では、その中で最も性能が良かった下記の表現法を用いる。

F	A M	C Y B

これは、6 種類の判定ラベル F、A、M、C、Y、B のうち、ラベル A、M とラベル C、Y、B をそれぞれ区別せずに 1 つの分類とみなすものである。そして、各機能表現候補は、チャンクであることを表す要素 (B/I) と、用法を示す要素 (F/AM/CYB) を組み合わせた次の 6 種類のチャンクタグによって表現される。

B-F　B-AM　B-CYB

I-F　I-AM　I-CYB

本研究では、用例データベースで設定されている判定ラベルのうち、ラベル F が付与される表現（複合辞）を検出する検出器 (これを、検出器 F と呼ぶ) と、ラベル F、A、M のいずれかが付与される表現（機能表現）を検出する検出器 (これを、検出器 FAM と呼ぶ) を作成する。検出器 FAM においては、評価時に、判定ラベル F と AM の区別を行わない。

SVM は二値分類器であるため、そのままでは、2 クラスの分類しか扱えない。本研究のようにクラス数が 3 以上の場合には、複数の二値分類器を組み合わせて拡張する必要がある。本研究では、拡張手法としては、広く利用されているペアワイズ法を用いる。

3.2 素　　性

学習・解析に用いる素性について説明する。文頭から i 番目の形態素 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。

$$F_i =< MF(m_i), CF(i), OF(i) >$$

形態素素性 $MF(m_i)$ は、形態素解析器によって形態素 m_i に付与される 10 種類の情報（表層形、品詞、品詞細分類 1 ～ 3、活用型、活用形、原形、読み、発音）である。

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候補に基づいて定まる素性である。今、下図のような形態素列 $m_j \ldots m_i \ldots m_k$ からなる機能表現候補 E が存在したとする。

$$m_{j-2} \quad m_{j-1} \quad \boxed{m_j \ldots m_i \ldots m_k} \quad m_{k+1} \quad m_{k+2}$$

機能表現候補 E

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候補 E を構成している形態素の数（機能表現候補の長さ）と、機能表現候補中における形態素 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候補の直前 2 形態素および直後 2 形態素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。

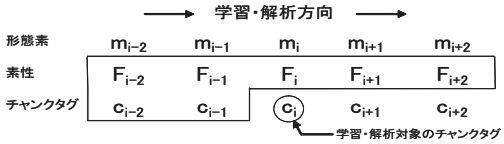

図 1 YamCha の学習・解析

表 2 データセットの各統計量

	各判定ラベル数						全形態素数
	F	A	M	C	Y	B	
全データセット	1974	55	453	523	9	169	92899
部分データセット 1	1478	52	342	465	8	155	16229
部分データセット 2	1478	52	342	465	8	155	90813

$$CF(i) = \langle k-j+1,\ i-j+1 \rangle$$
$$OF(i) = \langle MF(m_{j-2}),\quad CF(m_{j-2}),$$
$$MF(m_{j-1}),\quad CF(m_{j-1}),$$
$$MF(m_{k+1}),\quad CF(m_{k+1}),$$
$$MF(m_{k+2}),\quad CF(m_{k+2}) \rangle$$

複数の機能表現候補が、共通の形態素を構成要素に含む場合、チャンク素性とチャンク文脈素性は、次の優先順位に従って付与する。

1 先頭の形態素が最も左側の機能表現候補を用いて素性を付与する。

2 1 を満たす候補が複数存在する場合は、その中で最も形態素数が多い候補を用いて素性を付与する。

例えば、文（1）と文（2）に対してチャンク素性とチャンク文脈素性を付与する場合、文（1）は「にあって」を、文（2）は「というものの」をそれぞれ用いて、素性を付与する。

以上の素性を用いて、学習・解析を行う。i 番目のチャンクタグの学習・解析を行う場合に用いる素性は、F_{i-2}、F_{i-1}、F_i、F_{i+1}、F_{i+2}、c_{i-2}、c_{i-1} である (図 1)。解析時に素性として用いるチャンクタグは、解析によって得られたチャンクタグを順に利用する。

4. 実験と考察

検出器 F と検出器 FAM に対して、性能の評価を行った。

4.1 データセット

実験には、用例データベースにおいて、判定ラベル F とそれ以外の用法とがバランスよく収録されている 52 表現に対する 2600 例文 (1 つの表現につき 50 例文) について、全ての機能表現候補に判定ラベルを付与したものを使用した。以下、判定ラベルが付与されたこの 2600 例文のことを、全データセットと呼ぶ。全データセットに含まれる各ラベル数と、全形態素数を、表 2 に示す。

4.2 評価尺度

実験を評価する際の尺度には、以下の式で表される精度、再現率、F 値、および判別率を用いた。

$$\text{精度} = \frac{\text{検出に成功したチャンク数}}{\text{解析によって検出されたチャンク数}}$$
$$\text{再現率} = \frac{\text{検出に成功したチャンク数}}{\text{評価データに存在するチャンク数}}$$

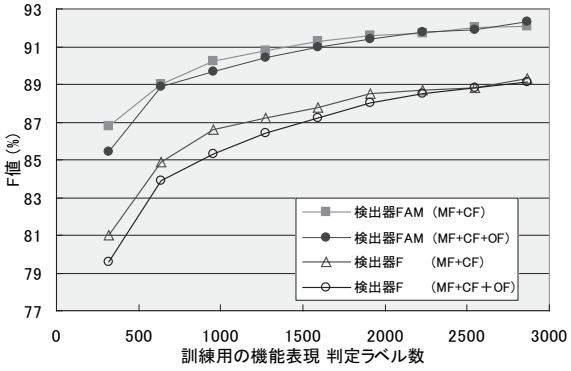

図 2 訓練データサイズと学習性能の関係

$$\text{F 値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}}$$
$$\text{判別率} = \frac{\text{正解した判定ラベル数}}{\text{全判定ラベル数}}$$

また、実験は、10 分割交差検定を用いて行った。

4.3 評価結果

検出器 F および検出器 FAM と、各ベースラインの検出性能を表 3 に示す。 表 3 において、「頻度最大の判定ラベル」は、解析時に常に頻度最大のもの (ラベル F) を出力した場合である。また、「人手により作成した規則」は、土屋らが提案した手法 [3] である。「JUMAN/KNP」および「ChaSen/CaboCha」といった既存の解析系は、機能表現の用法の区別を意識した検出は行わないため、ラベル F、A、M を正解とする評価のみを行った。

また、「MF」は形態素素性のみを用いた場合を表し、「MF+CF」は、形態素素性とチャンク素性を用いた場合、「MF+CF+OF」は、形態素素性とチャンク素性とチャンク文脈素性の全てを素性として用いた場合を表す。

表 3 からわかるように、SVM を用いたチャンキング手法は、どのベースラインよりも高い F 値を示した。また、用いる素性の違いによる性能の違いに着目すると、形態素素性のみを用いた場合に比べて、形態素素性とチャンク素性を併用した場合 (MF+CF) のほうが、F 値で 2 ポイント以上上回った。このことから、チャンク素性は、機能表現を検出するための素性として有効であったと言える。一方、MF+CF と MF+CF+OF の間に、性能の差はみられなかった。

4.4 訓練データサイズの違いによる比較

本稿では、用例データベースに基づき、1 表現につき 50 例文を用いて実験を行ったが、このデータサイズが、機能表現のチャンクの学習に十分であるという保証はない。そこで、訓練データにおける機能表現の判定ラベル数を減少させたとき、検出性能がどのように変化するかを調査した。結果を図 2 に示す。

図 2 より、学習に用いる判定ラベル数が全データセットの約 10 分の 1 のときは、検出性能が大きく低下しているが、判定ラベル数の増加に伴って検出性能も向上していき、全データセットに相当する判定ラベル数付近で

表 3 各検出器の検出性能 (%)

		F				FAM			
		精度	再現率	F 値	判別率	精度	再現率	F 値	判別率
ベースライン	頻度最大の判定ラベル	72.4	100	76.6	62.0	78.0	100	87.6	78.0
	JUMAN/KNP	—	—	—	—	89.2	49.3	63.5	55.8
	ChaSen/CaboCha	—	—	—	—	89.0	45.6	60.3	53.2
	人手により作成した規則	86.8	83.7	85.2	82.0	90.7	81.6	85.9	79.1
SVM を用いた検出器	MF	85.1	89.2	87.1	85.5	88.0	91.0	89.4	86.5
	MF+CF	87.6	91.1	89.3	87.9	91.0	93.2	92.1	89.0
	MF+CF+OF	87.1	91.3	89.1	87.5	91.1	93.6	92.3	89.2

MF は、形態素素性を、**CF** はチャンク素性を、**OF** は、チャンク文脈素性を表す。

は、検出性能がほぼ飽和していることがわかる。これより、機能表現のチャンクの学習に用いるデータサイズは、用例データベースで収集されている用例量で十分であるといえる。

4.5 訓練データの作成コストの削減

用例データベースを、SVM を用いた機能表現の検出器における訓練データとして用いるときには、判定ラベルを付与すべき箇所が 1 例文につき 1 箇所以上となる場合がある。機能表現候補「ばかりだ」の用例 (3) のように、1 つの用例中に、複数の別の機能表現候補が現れることがあるからである (“/” は形態素区切りを表す)。

(3) /セミナー/開催/に/あたり/、/最初/は/戸惑う/こと/ばかり/だっ/ た/と/いう/。/

これには、次のような問題が考えられる。

- 「という」などの出現頻度の高い機能表現と、出現頻度の低い機能表現の収集数に差ができ、学習に偏りが生じる。
- 検出対象とする機能表現の数を増やしていった場合、判定ラベルを付与すべき箇所が膨大になる。

この問題を解決する手法として、1 例文につき、判定ラベルを付与する箇所をその例文が対象としている機能表現候補の部分のみとし、その例文に出現したその他の機能表現候補は機械的に削除することが考えられる。しかし、この処理によって、学習性能が劣化することも懸念される。そこで、次の 2 つの部分データセットを用いて、この手法が有用であるかを調査した。

- 部分データセット 1
  - 各用例において、機能表現候補と、その前後 2 形態素ずつを切り出したデータ。ただし、前後 2 形態素内にその他の機能表現候補が出現した場合はそれを含み、さらに前後 2 形態素を含める、という操作を繰り返す。

  (例) /戸惑う/こと/ばかり/だっ/た/と/いう/。/
- 部分データセット 2
  - 部分データセット 1 +その他の機能表現候補以外の形態素。ただし、対象以外の機能表現を削除することによって文が分断された場合は、それぞれを 1 文として用いる。

  (例) /セミナー/開催/ /、/最初/は/戸惑う/こと/ばかり/だっ/た/と/いう/。/

なお、部分データセット 1 と部分データセット 2 では、作成する際に人手を必要とする作業量は、どちらも同じである。部分データセット 1、2 に含まれる各判定ラベル数と形態素数を表 2 に、実験結果を表 4 に示す。 部分データセット 1 で学習を行った場合は検出性能が大きく低下したが、部分データセット 2 の場合は、検出性能の低下を、検出器 F において約 1.0 ポイント、検出器 FAM において約 0.6 ポイントに抑えることができた。したがって、上で述べた方法によって訓練データの作成コストの削減ができているといえる。

表 4 訓練データの違いによる性能比較 (%)

データセット	F			FAM		
	精度	再現率	F 値	精度	再現率	F 値
全データセット	87.6	91.1	89.3	91.1	93.6	92.3
部分データセット 1	86.4	39.1	53.7	90.3	47.4	62.1
部分データセット 2	87.1	89.8	88.4	90.7	92.4	91.5

5. まとめと今後の課題

本稿は、SVM を用いたチャンカー YamCha を利用した、日本語機能表現の検出器を提案し、その性能評価を行った。そして、機械学習による機能表現の検出が、人手による規則を用いた機能表現の検出よりも高い性能を示すことを報告した。検出対象の機能表現の種類を増やし、その性能を評価することが今後の課題である。

参 考 文 献

1) 国立国語研究所: 現代語複合辞用例集 (2001).
2) 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).
3) 土屋雅稔, 宇津呂武仁, 佐藤理史, 中川聖一: 形態素情報を用いた日本語機能表現の検出, 言語処理学会第 11 回年次大会発表論文集, pp. 584–587 (2005).
4) Vapnik, V. N.: *Statistical Learning Theory*, Wiley-Interscience (1998).
5) 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, 言語処理学会第 12 回年次大会発表論文集 (2006).
6) 森田良行, 松木正恵: 日本語表現文型, NAFL 選書 5, アルク (1989).
7) Tjong Kim Sang, E.: Noun Phrase Recognition by System Combination, *Proc. 1st NAACL*, pp. 50–55 (2000).