# 日本語文型同定システムにおける曖昧性解消とその評価

ミロシニク ロマン　加藤 恒昭
東京大学 大学院総合文化研究科
miroshnik.roma@gmail.com; kato@boz.c.u-tokyo.ac.jp

## 1 はじめに

日本語学習者を支援するための総合的な日本語文型同定システムを開発している。初心者の日本語学習者が、自分で日本語文を作文する時や書かれた日本語文を読み解く時に使えるような語彙辞書、漢字辞書、文型辞書を組み込んだ総合的な読解支援システムを目指している。

語彙的な支援、漢字に関する支援に対して、文型に関する支援を行うシステムは多くない [5]。ここで文型とは、文における語順を定めたルールのことである。つまり、言語習得に必要な文法の知識のパターンのことをいう。 例えば、「だけで (は) なく 〜(も)」、「〜ざるを得ない」、「〜はいうまでもなく」、「〜と (は) いって」などが文型である。これらに関する支援を行うために、与えられた文からこれらの文型を同定することが必要になる。総合的に文型を取り扱うこのような読解支援システムの核となる文型同定システムが求められている。

文型同定システムで問題となるのは、複数種類の文型が同じ構成を持っていたり、ある文型が、文型として同定すべきでない内容表現と同じ構成を持っていたりして、その部分だけではその判定ができないという曖昧性の存在である。前者を文型と文型の間の曖昧性、後者を文型と内容表現の曖昧性と呼ぶ。

本稿では、開発中のシステムにおけるこれらの曖昧性の解消について論じる。まずシステムの概要を紹介し、次に曖昧性の手法を述べる。実現のための規則構築を報告した後、評価結果を示す。

## 2 システムの構成

開発しているシステムでは、形態素解析器 MeCab の解析結果を利用して、XML 形式で記述されている規則に基づいて、3 段階の処理を行うことで文型要素を同定する。ルールベースのシステムであるが、全ての処理のための規則を XML パターンで記述し、見通しのよいものとすることで、多数の文型を扱うことを可能としている。本システムで扱う文型は、文献 [2] に挙げられているもので、現在、618 件の文型の同定を実装している。

各段階の処理は以下の通りである。

1. **形態素解析器 MeCab によるテキストの解析**分析対象となるテキストを MeCab によって解析し、結果を形態素オブジェクトの配列とする。
2. **XML パターンによる文型構成素候補の抽出**文中で同じような役割をなしている形態素オブジェクトをまとめて、より抽象度の高い文型構成素候補に変換する。
3. **XML パターンによる文型要素の同定**文型構成素候補の配列にパターンを適用し、文型を含んでいるかを分析し、文型要素を同定する。
4. **XML パターンによる曖昧性解消**文型として検出された表現が文型と文型の間の曖昧性や文型と内容表現の曖昧性を持つ場合、実際はどちらに属するかの判断を行ったり、内容表現になっているものを取り除いたりなど、曖昧性の解消を行う。

形態素を直接扱わず、文型のパターンが記述しやすい文型構成素候補への変換を行うこと、文型の同定を主に内部の構造に注目した抽出部分と周囲の文脈を利用した曖昧性解消に分離したことが特徴で，これらによりモジュール性の高い構造となっている。

本稿では、この第 4 の段階について説明する。第 1, 2, 3 の段階については文献 [5] に述べられている。

## 3 曖昧性解消の手法

文献 [4] では、形態素解析結果に対して機械学習を用いることで、機能表現と内容表現の曖昧性を持つ表現の曖昧性解消を行う方式が報告されている。そこでは表現の前後二つ形態素を素性としている。本稿の曖昧性解消処理では、表現の前後二つ形態素に加えて、表現内の先頭と末尾、各一つずつの形態素を条件として用いて判定を行う。

曖昧性解消は、XML で記述された曖昧性解消パターンに基づいて行われる。曖昧性解消パターンのテンプレート

を図 1 に示す。name 属性は、そのパターンを適用する文型を指定する。second_head、 first_head、 inner_head、inner_back、first_back、second_back は二つ前、一つ前、表現内先頭、表現内末尾、一つ後、二つ後の形態素についての条件である。パターン内のすべての条件がマッチした場合、パターンが同定され、パターンの種類によってそれぞれの処理が行われる。pattern_TYPE は pattern_yes もしくは pattern_no のいずれかで、pattern_yes タグであれは、同定された場合に対象とする文型が同定結果として残され、同定できなければ同定結果から削除される。pattern_no タグであれば、同定された場合に同定結果から削除され、同定できなければ、対象とする文型が同定結果に残される。このような 2 種類の操作を設けることで、パターンの記述を容易にしている。

```
<bunkei name="">
<pattern_TYPE>
        <second_head>
        <cand_name></cand_name>
        </second_head>
        <first_head>
        <cand_name></cand_name>
        </first_head>
        <inner_head>
        <cand_name></cand_name>
        </inner_head>
        <inner_back>
        <cand_name></cand_name>
        </inner_back>
        <first_back>
        <cand_name></cand_name>
        </first_back>
        <second_back>
        <cand_name></cand_name>
        </second_back>
</pattern_TYPE>
</bunkei>
```

図 1: 曖昧性解消の XML パターンのテンプレート

以下、曖昧性解消パターンの例を挙げる。

**例 1 複数種類の文型と文型の間の曖昧性解消**　「から $^1$」“ from ”(パーティーは八時から始まる。) と「から $^2$ 」”after doing s.t. ”(ご飯を食べてから映画に行った。) の曖昧性
格助詞「から」は「から $^1$」と「から $^2$」の文型要素として同定される。これに対して図 2 に示したパターンを用いて、曖昧性を解消し、一方の文型を選択する。ひとつ前の形態素が条件として参照される。これが接続助詞「て/で」の場合、文型「から $^2$」が選ばれ、「から $^1$」は削除される。そうでなければ、「から $^1$」が選ばれ、「から $^2$」は削除される。

**例 2 文型と内容表現の曖昧性**　文型「ものを」(彼は日本語を続けて勉強すればいいものを、1 年間やっただけでやめてしまった。) と内容表現「ものを」(自分で描きたいものを描けばよい。) の曖昧性
図 3 のパターンによって、曖昧性解消を行う。同定した

```
<bunkei name="kara1">
<pattern_no>
<first_head>
<cand_name>te</cand_name>
<cand_name>de_setsuzokujyoshi</cand_name>
</first_head>
</pattern_no>
</bunkei>

<bunkei name="kara2">
<pattern_yes>
<first_head>
<cand_name>te</cand_name>
<cand_name>de_setsuzokujyoshi</cand_name>
</first_head>
</pattern_yes>
</bunkei>
```

図 2: 「から $^1$」「から $^2$」の曖昧性解消パターン

文型「ものを」に対して、ひとつ後の形態素を参照し、その形態素の文型要素候補ラベルが読点であれば、文型「ものを」として同定する。読点が見つからなければ、文型「ものを」が削除され、形態素オブジェクト「もの」と「を」はともに内容表現として処理される。

**例 3 文型と内容表現の曖昧性**　文型「とあって」“ because; as expected; ”(夏休みが始まったとあって、子供たちはみんな嬉しそうだ。) と助詞「と」と動詞「ある」からなる内容表現 (ゴルフ場造成の仕事なども次々とあってしばらくはうまくいった) の曖昧性
同定した文型「とあって」に図 4 のパターンを適用する。文型のひとつ後の形態素が「非自立動詞」、「非自立形容詞」、「非自立名詞」、「助動詞」である場合、同定した文型要素が削除され、内容表現として処理される。それ以外の場合、同定した文型要素は残される。

```
<bunkei name="mono_o">
<pattern_yes>
<first_back>
<cand_name>touten</cand_name>
</first_back>
</pattern_yes>
</bunkei>
```

図 3: 「ものを」の曖昧性解消パターン

```
<bunkei name="to_atte">
<pattern_no>
<first_back>
<cand_name>doushi_hijiritsu</cand_name>
<cand_name>keiyoushi_hijiritsu</cand_name>
<cand_name>hijiritsuteki_meishi</cand_name>
<cand_name>jyodoushi</cand_name>
</first_back>
</pattern_no>
</bunkei>
```

図 4: 「とあって」の曖昧性解消パターン

# 4 曖昧性例文収集と分析

各文型は様々な異形を持つ。例えば，文型「によって」は「によりまして」や「により」のような異形を持つ。

「機能表現辞書つつじ」[3] を利用して、各文型の異形を作成し、それらに対して「現代日本語書き言葉均衡コーパス（BCCWJ）」[1] から 20 件程度の例文を抽出し、それらの例文を曖昧性の有無を分析した。

618 文型のうちに文型と文型の間の曖昧性があり得る文型を 53 件発見し、文型と内容表現との曖昧性があり得る文型を 38 件発見した。

曖昧性があり得る各文型に対して BCCWJ からさらに 100 件程度の例文を収集した。各例文を分析し、それぞれの例文を文型であるか内容表現であるかによって分別した。

文型ごとこれらの例文に対して MeCab 形態素解析と文型構成素候補ラベル検出を行った。各文型の前後の 3 つ形態素の文型構成素候補ラベル（表現の前後の二つ形態素と表現内の前後の一つ形態素）を抽出した。抽出した情報を参照して 92 曖昧性解消パターンを作成した。

なお、文献 [2] であげられている文型の一部については、本システムで用いている情報と枠組みでは曖昧性解消パターンを記述できなかった。これについては、次節で詳しく述べる。

# 5　評価

評価は二つの方法で行った。第一は、日本語能力試験に出題する三つのテキストを対象とした評価である。第二は、曖昧性を生じる可能性があり、曖昧性解消のパターンを記述できた 38 文型それぞれに対しての評価である。

評価の指標として正解率、適合率、再現率と F 値を計測した。文型として同定されるべきであるという期待と実際にシステムによって同定されたかという事実に対して以下のように判断する：対象形態素オブジェクト列が文型として期待され、実際に文型として同定された場合、True Positive(TP) とされる。文型として期待されていたが、実際には文型として同定されなかった場合、False Positive(FP) とされる。対象形態素オブジェクト列が文型と期待されないにもかかわらず、実際には文型として同定された場合、False Negative(FN) とされる。文型と期待されず、実際も文型として同定されなかった場合、True Negative(TN) とされる。

## 5.1　総合的評価

評価のためのテキストとして日本語能力試験 N1 に出題された 3 つのテキストを選んだ。評価結果を表 1 に示す。この評価では期待される文型を正しく同定したか、誤って文型をしたかに着目したので、TN は常に 0 である。

総合的評価の結果として 96-99%の再現率と 98,1-99,5%の F 値を達成している。3 つのテキストの合わせたものについて、再現率は 97%であり、98,7%の F 値を達成している。ただし、助詞、副詞など同定が比較的簡単なものの出現率が高く、それらが評価に強く影響を与えていると考えられる。そのため、次に、問題となり得る文型（特に曖昧性が生じ得る文型）に対して文型別の評価を行なった。以下でその結果について述べる。

## 5.2　文型別の評価

この評価では評価対象を曖昧性があり得る文型 38 件とした。対象文型に対して BCCWJ から各異形の 20 例文ずつを収集し、文型ごと評価を行った。618 文型についての異形の数は 16884 であるが、曖昧性のある 38 文型では、多い場合は 710、少ない場合で 1 の異形を持ち、異形の平均数は 34 であった。これらの異形ごとに 20 例文の文例をランダムに集めた。文型と文型の間の曖昧性におけるそれぞれの文型の数、文型と内容表現の曖昧性における内容表現の数は、バランスさせていない。結果は表 2 の通りである。F 値の高い文型 3 件、低い 3 件、合計を示した[1]。

文型別評価の結果として以下が分かった。曖昧性の生じ得る文型全体で 93%の再現率で同定できる。F 値も 94,4%という高い値で曖昧性を解消できている。F 値が 100%になっている文型は 12 件であった。90%以下の F 値を持つ文型は 6 件あり、そこに「くわえて」、「につけ」、「だけに」、「そのうえ」、「に向けて」が含まれる。特に F 値が悪かった文型は「から言って」で、その F 値は 77,4%であった。

総合的に見れば、曖昧性解消がかなり高い正確さで行われていると言える。しかし、特定の文型の評価が悪く、それらについては曖昧性解消パターンの改善が必要である。

## 5.3　曖昧性解消できなかった文型

現在の提案アルゴリズムでは、いくつかの文型の曖昧性を解消できない。内容表現と文型の曖昧性では以下の文型に対して解消を行うためのパターンが記述できなかった。

「それも」、「そこを」、
「ところだ　“ moment ”」、「うえに」、
「の上では」、「うえで」

[1] 以下、スペースの制約で本稿内に示せなかった XML パターンは https://github.com/MyroshnykRoman/bunkei/blob/master/文型別の評価.pdf を参照いただきたい。

表 1: 日本語能力試験 N1 に出題するテキストに対する性能評価

| | TP | TN | FP | FN | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|---|---|
| テキスト 1 | 154 | 0 | 1 | 5 | 0,96 | 0,99 | 0,97 | 0,981 |
| テキスト 2 | 101 | 0 | 0 | 1 | 0,99 | 1,00 | 0,99 | 0,995 |
| テキスト 3 | 245 | 0 | 1 | 5 | 0,98 | 1,00 | 0,98 | 0,988 |
| 合計 | 500 | 0 | 2 | 11 | 0,97 | 1,00 | 0,98 | 0,987 |

表 2: 文型別の評価

| | TP | TN | FP | FN | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|---|---|
| ことから | 18 | 2 | 0 | 0 | 1,00 | 1,00 | 1,00 | 1,000 |
| 果たして | 9 | 11 | 0 | 0 | 1,00 | 1,00 | 1,00 | 1,000 |
| ところから | 30 | 0 | 0 | 5 | 1,00 | 1,00 | 1,00 | 1,000 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| につけ | 12 | 36 | 4 | 1 | 0,91 | 0,75 | 0,92 | 0,828 |
| くわえて | 2 | 37 | 1 | 0 | 0,98 | 0,67 | 1,00 | 0,800 |
| から言って | 24 | 2 | 0 | 14 | 0,65 | 1,00 | 0,63 | 0,774 |
| 合計 | 961 | 538 | 68 | 47 | 0,93 | 0,93 | 0,95 | 0,944 |

表 3: 助詞「を」に対する曖昧性

| | |
|---|---|
| [direct object marker] | 車を買う |
| [space marker] | 道を歩く |
| [a point of detachment] | 家を出る |
| [emotive marker] | 入学を喜ぶ |

文型と文型の間の曖昧性が解消不可能な文型では,「で」「も」「か」、「なんて」、「に」、「を」など、助詞の意味役割に関するものを中心に 32 種類が扱えない。

これらの文型の曖昧性を解消するためには補助的な知識が必要である。例えば、助詞「を」に対して 4 つのタイプが存在し、その間の曖昧性を解消する必要がある。それぞれのタイプの助詞「を」はそれぞれ異なる動詞を要求する（表 3)。したがって、助詞「を」の文型と文型の間の曖昧性を解消するには後続する動詞に関する援助知識が必要である。可能な方法として辞書に動詞をそれぞれのタイプに分けて記述し、必要な時その情報を引く処理が考えられる。もう一つの援助知識としては、名詞に関して物体であるか概念であるかなどの意味的な知識で、それによって「に当たって」、「に従って」、「に応じて」、「に向けて」などの文型の曖昧性解消の正確さを向上できると思われる。例えば「会議に当たって…」と「壁に当たって…」という文型と内容表現との曖昧性を解消するために対象となる表現の前の名詞が概念であるか物体であるかが必要である。この例では「会議」は概念であって表現が文型とされ、「壁」は物体であって表現が内容表現とされる。

今後はそのような援助知識の導入によって残っている曖昧性を解消することを目指していく。

# 6 おわりに

文型同定システムにおける曖昧性解消について報告した。文型同定システムは形態素解析器の解析結果を利用して 3 段階の XML パターンによって文型要素を同定するが、曖昧性解消はその最終段階で、文型の前後の二つ形態素と文型内の前後一つ形態素を参照するようなパターンを用いて行なわれる。

文型同定システムの総合的な評価によって 97 %の正解率と 98.7 %の F 値を得られた。曖昧性があり得る文型別の評価でも 93%の正解率と 93.3 の F 値を得られた。いくつかの文型については、同定のために援助知識が必要なことが明らかになった。

# 参考文献


[1] 国立国語研究所, 2012. 現代日本語書き言葉均衡コーパス.

[2] 牧野 成一, 筒井 通雄. 1989, 1995, 2008. 日本語基本文法辞典, 日本語文法辞典 [中級編], 日本語文法辞典 [上級編] The Japan Times.

[3] 松吉 俊, 佐藤 理史, 宇津呂 武仁. 2007. 日本語機能表現辞書の編纂. 自然言語処理, 14(5), pp.123-146.

[4] 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. 2007. 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, 14(1), pp.111-138.

[5] ミロシニク ロマン, 加藤 恒昭. 2017. 日本語文型同定システムの構築. 言語処理学会第 23 回年次大会発表論文集, pp.851-854.