機能表現言い換えデータからの言い換え規則の自動生成

土屋 雅稔　佐藤 理史　宇津呂武仁

京都大学大学院 情報学研究科
tsuchiya@pine.kuee.kyoto-u.ac.jp, {sato,utsuro}@i.kyoto-u.ac.jp

1. はじめに

日本語には,「〜に関して」「〜を通じて」などのように，付属語そのものではないが，ひとかたまりとなって付属語的に働く機能表現が多く存在する．複雑な機能表現が使われているテキストは，日本語初学者のように機能表現に慣れていない人にとっては非常に分かりにくい．そのような読者にも分かり易いテキストを記述しようとする時は，複雑な機能表現の使用を慎重に避ける必要がある．しかし，これは，機能表現を自然に使いこなすことのできる人にとっては，かなり難しい作業になる．仮に，分かりにくい機能表現を指摘し，代替表現を提案するツールがあれば，その労力を効果的に軽減することができるだろう．

そのようなツールは，(1) 分かりにくい機能表現を検出する処理，(2) 検出された機能表現の代替表現を生成する処理，という 2 つの処理からなる．これらの処理を自動的に行うためには，機能表現に関する計算機処理可能な大量の規則が必要である．日本語の機能表現に関する規則を記述するための，確立された標準的な記法と言うものは存在しない．品詞体系によって機能表現の解釈が異なるので，品詞体系に依存した方法で記述された規則は再利用性が低くなることが予想される．

品詞体系に依存しない記法としては，例文に基づいて機能表現を記述する方法が考えられる．しかし，例文に基づく記法は，検出及び言い換えに関する規則を明示しているわけではない．そのため，実際に検出・言い換えを行うためには，適切な変換を行って具体的な規則を得る必要がある．

本論文では，変換時に必要となる情報を人手で追加することにより，例文に基づいて記述された機能表現言い換えデータから，機能表現の言い換え規則を自動的に生成する方法について述べる．次に，生成された規則に基づいて機能表現の検出および言い換えを行うシステムを実装し，無作為に選んだ 50 個の機能表現を対象として検出と言い換えの実験を行い，十分な精度で検出と言い換えが可能であることを示す．

2. 機能表現

2.1 機能表現の特徴

機能表現とは,「〜について」「〜において」などのように，付属語そのものではないが，いくつかの語が複合してひとまとまりの句となって付属語的な役割を果たす表現である．どのような句を機能表現と呼ぶか，に関して明確な定義は存在しない．本研究では，森田ら [1] の複合辞についての考え方を参考にして，以下の 3 つの指標を考える．

非構成的な意味　表現全体の意味に，それぞれの語の意味の単なる組合わせではなく，それ以上の独自の意味が生じていること．

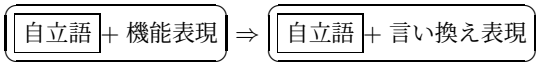
緊密な構成要素　構成要素の交替，他語の挿入，構成要素の省略などが起こりにくいこと．

形式化した自立語　表現に含まれている名詞・動詞・形容詞などの自立語が，本来の機能を失って形式化し，助詞・助動詞的な機能しか果たさなくなっていること．類語で言い換えられなかったり，活用が失われていることが目安となる．

これらの指標を定量的に測ることは非常に困難である．そのため，本研究では，この 3 つの指標を主観的に満たす表現を機能表現と呼び，それらの検出と言い換えについて検討する．

2.2 機能表現の言い換え

ある機能表現を別の表現で言い換える場合，もっとも簡単な言い換えは，その機能表現を，代替表現で単純に置き換えてしまうような言い換えである．しかし，機能表現は直前の自立語と結び付いて付属語的に働く表現であるから，機能表現を言い換える時は，直前の自立語にも適切な変更が必要となることが殆んどである．本研究では，機能表現と直前の自立語を組み合わせた句を単位として，句を置き換えることによる言い換えを対象とする．



句よりも大きな範囲の文構造の変形などが必要な言い換えについては，本研究では考慮しない．

3. 機能表現言い換えデータ

3.1 データの形式

日本語の機能表現に関する規則を記述するための，確立された標準的な記法は存在しない．ここでは，品詞体

```
<言い換えペア>
  <ラベル>〜に違いない</ラベル>
  <級>2</級>
  <言い換え元文>
    彼は<言い換え元句>知っている<機能表現>に違いない</機能表現></言い換え元句>.
  </言い換え元文>
  <言い換え先文>
    彼は<言い換え先句>知っているはずだ</言い換え先句>.
  </言い換え先文>
</言い換えペア>
```

図 1 機能表現言い換えデータ

系に依存しない記法として，例文に基づく記法について検討する．

以前の研究 [2] では，機能表現を含む文法事項を定義するために，以下の 4 項目からなるデータ形式を採用していた．

平易度	2
規格名	〜に違いない
例文	彼は知っているに違いない.
正解	知っているに違いない

しかし，このデータ形式では機械的に検出規則を生成することはできず，実際の検出規則を得るためには，機械的に生成した検出規則に人手で修正を加える必要があった．人手による修正が必要となった主な理由は，変換時に必要な情報の不足である．このデータには，例文中で機能表現の占めている部分についての明示的な情報が欠けていたから，規格名を手がかりとして，機能表現に相当する部分を推定していた．しかし，規格名は機械処理を考慮して決められているわけではないので，推定には，しばしば誤りが含まれていた．推定に誤りが含まれていると，正しい検出規則が得られず，人手による修正が必要となっていた．また，形態素解析誤りのために，正しい検出規則が得られない場合もあった．

この知見と，機能表現を含む句を単位として言い換えを行うことを考慮し，機能表現の検出と言い換えに必要な情報として以下の 7 項目を考える．

（1） 機能表現ラベル
（2） 級
（3） 言い換え元文
（4） 言い換え元句
（5） 機能表現
（6） 言い換え先文
（7） 言い換え先句

この 7 項目の情報を，図 1 の形式で表現する．

3.2 検出規則の生成

図 1 の機能表現言い換えデータは，機能表現「〜に違いない」を実例に基づいて定義しているが，この機能表現を検出するための具体的な規則を示しているわけではない．したがって，実際に機能表現を検出するためには，適切な変換を行って，計算機で処理可能な明示的な検出規則を得る必要がある．

検出された機能表現は，そのまま，代替表現を生成する言い換え規則の入力となる．そのため，機能表現と直前の自立語を含む言い換え元句の単位で検出を行っておく方が都合が良い．本研究では，以下の手順で，言い換え元文から形態素パターンを取り出し，取り出されたパターンを機能表現を含む言い換え元句の検出規則として用いる．

（1） 言い換え元文を形態素解析する．必要に応じて，形態素解析結果の誤りを人手で修正する．
（2） 言い換え元句の部分の形態素列を取り出す．

知っている	動詞	基本形
に	助詞	
違いない	形容詞	基本形

（3） 機能表現の直前の自立語部分については，品詞を検出規則とする．機能表現部分は，そのままの形態素列を検出規則とする．その結果，以下のような計算機処理可能な検出規則が得られる．

```
np( 2, "〜に違いない",
    Dm( { H => "動詞" },
        { G => "に", H => "助詞" },
        { G => "違いない", H => "形容詞",
          K => "基本形" } ) );
```

この規則は，任意の動詞に続いて，助詞「に」と形容詞「違いない」の基本形が現れた時に,「〜に違いない」という機能表現が現れたと見なすことを意味している．

機能表現の区切り位置と，形態素の区切り位置が一致しない場合がある．例えば,「〜てやまない」という機能表現では，直前の自立語の活用語尾が機能表現に含まれている．

あの人が無事で居ることを，祈ってやまない.

このような場合は，直前の自立語の活用形を検出規則に含める．

```
np( 1, "〜てやまない",
    Dm( { H => "動詞", K => "タ系連用テ形" },
        { G => "やむ", H => "動詞",
          K => "未然形" },
        { G => "ない", H => "接尾辞",
          K => "基本形" } ) );
```

3.3 言い換え規則の生成

直前の自立語と言い換え表現を組み合わせて言い換え先句を生成するには，少なくとも 2 種類の変形を考える必要がある．第 1 に，言い換え表現に応じて，自立語を

活用させる必要のある場合がある．以下の例では，動詞「困る」を意志形からタ系連用テ形に活用しなければならない．

彼がいかに困ろうが，私には関係ない．

→ 彼がいかに困っても，私には関係ない．

第 2 に，自立語の品詞が交替する場合がある．例えば，以下の言い換えでは，言い換え元句の自立語は「断り」という名詞であるのに対して，言い換え先句の自立語は「断る」という動詞である．

断りなしに入ってはいけない．

→ 断らないで入ってはいけない．

以下の言い換えでは，言い換え元句の自立語は「貧しさ」という名詞であるが，言い換え先句では「貧しい」という形容詞になっている．

母は貧しさゆえに苦労してきた．

→ 母は貧しくて苦労してきた．

これらの変形操作を考慮して，以下のような手順で言い換え規則を生成する．

(1) 言い換え先文を形態素解析する．必要に応じて，形態素解析結果の誤りを人手で修正する．

(2) 言い換え先句に相当する形態素列を取り出す．

知っている	動詞	基本形
はずだ	助動詞	基本形

(3) 変形を考慮しながら言い換え元句の形態素列と比較して，自立語と言い換え表現を分割する．この例の場合，自立語は「知っている」，言い換え表現は「はずだ」になる．

4. 実験と考察

4.1 実験条件

日本語能力試験出題基準 [3] に掲載されている機能表現から 50 種類を無作為に選択し，これらの機能表現を説明するために書かれた例文 216 文を，日本語能力試験を対象とした教科書 [4] から取り出した．これらの例文は，ある 1 つの機能表現を説明しようとして書かれた文であり，説明対象の機能表現を少なくとも 1 つ含む．取り出された例文集合を，機能表現毎の例文数がなるべく均等になるように考慮しながら，規則生成用の例文集合 (123 文) と，実験用の例文集合 (93 文) に分割した．

出題基準に掲載されている例文 (565 文) と，教科書から得られた規則生成用の例文を対象として，人手で必要な情報を追加して機能表現言い換えデータを作成した．この機能表現言い換えデータから，663 個の検出規則と 642 個の言い換え規則を生成した☆．この生成された規則群を利用したシステム (図 2) を用いて，機能表現の検出及び言い換えの実験を以下の通り実施した．

4.2 検出実験

検出規則を生成する時に参照した例文 123 文を対象として，50 種類の機能表現が正しく検出されるかを実験した (クローズドテスト)．次に，教科書の例文の残り 93 文を対象として，同じく 50 種類の機能表現が正しく検出されるかを実験した (オープンテスト)．結果を，表 1 に示す．なお，1 つの例文に複数の機能表現が出現する場合があるので，検出対象となる機能表現の総数は，文数よりも多くなっている．

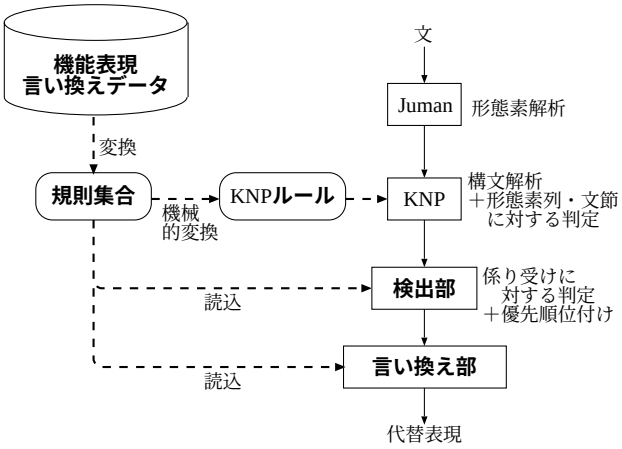


図 2 システム構成

表 1 検出結果

	機能表現数	検出された		検出されなかった
		正解	誤り	
クローズドテスト 適合率＝93%, 再現率＝87%	131	114	9	17
オープンテスト 適合率＝99%, 再現率＝73%	95	69	1	25

正解率は，クローズドテスト・オープンテストともに良好な結果を示しているのに対して，再現率では，オープンテストでの低下 (87% → 73%) が目立つ．検出できなかった機能表現を検討してみると，機能表現のバリエーションに対して，規則を生成するための例文が不足していることが分かった．例えば,「〜ならない」という機能表現には，丁寧な派生形として「〜なりません」がある．しかし，この派生形は，規則を生成するための例文には現れていなかったので，検出に失敗していた．したがって，検出規則を生成するための例文を増やせば，再現率も自然に向上すると考えられる．

4.3 言い換え実験

まず，言い換え規則を生成する時に参照した教科書の例文 123 文を対象として，これらの例文が説明している機能表現を言い換える実験を行った (クローズドテスト)．次に，教科書の例文の残り 93 文を対象として，やはり，これらの例文が説明している機能表現を言い換える実験を行った (オープンテスト)．結果を表 2,3 に示す．同時に，得られた言い換え例を図 3 に示す．

多くの機能表現は，いくつかの異なる表現に言い換えることができる．例えば，図 3 の先頭の「〜でからとい

☆ まったく同一の検出規則や言い換え規則は統合されるので，機能表現言い換えデータの例文数と規則数は一致しない

あの本を読んでからというものは、いろいろ考えることが多い. → あの本を読んでからは、いろいろ考えることが多い.
→ あの本を読んだあとは、いろいろ考えることが多い.
→ あの本を読んだおかげで、いろいろ考えることが多い.
これも勝たんがための練習だから、がんばるしかない。→ これも勝つための練習だから、がんばるしかない。
実験データに基づいた事実である。→ 実験データをもとにした事実である。
不景気とあって、デパートの人出はよくなかった。→ 不景気なので、デパートの人出はよくなかった。
→ 不景気ということで、デパートの人出はよくなかった。
この問題に関して意見が寄せられた。→ この問題について意見が寄せられた。
２０キロからある荷物を背負って７階まで階段を登った。→ ２０キロ以上もある荷物を背負って７階まで階段を登った。

図 3 機能表現の言い換え例

表 2 言い換え結果 (文単位)

	文数	言い換えた		言い換えられなかった
		成功	失敗	
クローズドテスト 適合率＝79%, 再現率＝74%	123	92	24	7
オープンテスト 適合率＝79%, 再現率＝59%	93	54	14	24

表 3 言い換え結果 (言い換え単位)

	正解		誤り	
クローズドテスト	132	(63%)	76	(37%)
オープンテスト	92	(72%)	35	(28%)

うものは」という表現は，3 種類の異なる表現に言い換えが可能である．機能表現の書き換えを支援するための言い換えという目的から考えると，複数の代替表現をそのまま利用者に示し，好みの表現を 1 つ選択してもらうことができるので，特に 1 つの表現に絞り込むという操作を行う必要はない．このような利用法では，言い換えの出力は 2 つの条件を満たせば良い．第 1 の条件は，出力に正解が含まれていることであり，第 2 の条件は，出力の数がユーザーが選択できないほど多くないことである．

本システムは，クローズドテストの場合，123 文の例文中 92 文について，正しい言い換えを含む言い換え候補を提示できる．つまり，正解を含む言い換え候補の提示の成功率は 79%である．言い換え候補は 208 文あり，1 文あたり平均して 1.8 文の言い換え候補が提示され，その内の 63%が正しい言い換えである．したがって，ユーザーが選択できる程度の量の言い換え候補を提示すると同時に，提示された言い換え候補には過半数の正しい言い換えが含まれている．このように，クローズドテストの結果は，本システムが機能表現を書き換えることを支援するために適切な言い換えを出力していることを示している．

オープンテストの結果も同様に解釈できるが，ただし，言い換え候補提示の再現率が 59%と低いことが問題である．この主な原因は，機能表現の検出に失敗していることなので，機能表現検出の再現率が向上すれば，自然に改善するものと期待される．

5. おわりに

本研究では，品詞体系に依存しない形式で記述された機能表現言い換えデータから，自動的に機能語の検出規則と言い換え規則を生成し，生成された規則に基づいて検出と言い換えを行うシステムを作成した．ここで，機能表現は，付属語と形式的な自立語がひとかたまりの句となって，付属語的に働くような表現である．機能表現の検出は，簡単な形態素列パターンに基づくパターンマッチングによって行い，言い換えは，機能表現とその直前の自立語からなる句を単位とする局所的な交換によって行う．

日本語能力試験出題基準から無作為に選択した 50 種類の機能表現について，検出・言い換えの実験を行い，提案手法の有効性を確認した．ただし，ここで実験対象とした文は，機能表現を説明するために特に注意して書かれた文であり，機能表現と良く似ているが，実は機能表現ではないような表現が殆んど現れないという特徴がある．したがって，機能表現の検出・言い換えという点から見ると，非常に特殊なコーパスだったと言える．

今後は，対象とする機能表現を拡充すると同時に，ドメインの異なるコーパスを対象として実験を行い，本手法の有効性を確認していくことを予定している．

参考文献

1) 森田良行, 松木正恵: 日本語表現文型, アルク (1989).
2) 土屋雅稔, 佐藤理史: 文法の規格とその自動判定, 言語処理学会第 9 回年次大会発表論文集, pp. 89–92 (2003).
3) 国際交流基金, 財団法人日本国際教育協会: 日本語能力試験出題基準【改訂版】, 凡人社 (2002).
4) 友松悦子, 宮本淳, 和栗雅子: どんな時どう使う 日本語表現文型 500, アルク (1996).