

日本語機能表現から多様な類義表現への言い換え*

梶田 達也[†]松吉 俊^{††,‡}佐藤 理史[‡][†] 名古屋大学工学部電気電子・情報工学科^{††} 京都大学大学院情報学研究科[‡] 名古屋大学大学院工学研究科

{masuda,matuyosi}@sslslab.nuee.nagoya-u.ac.jp, ssato@nuee.nagoya-u.ac.jp

1 はじめに

言い換えとは、ある表現をその意味内容を変えずに別の表現に置き換えることである。これを機械的に実現する言い換え技術には、例えば、高齢者や子供、外国人、障害者など、利用者の言語能力にあわせて文章を言い換えるなどの応用がある。

日本語の表現は「犬」、「動く」、「楽しい」のような内容語と、「が」、「けれど」、「に対して」、「なければならぬ」のような機能表現に分類できる。これまでの言い換え研究は、内容語を別の内容語に言い換えることに焦点を当てたものが多かったが、近年、機能表現の言い換えの研究 [2, 1, 4] も見られるようになってきた。

言い換えでは、通常、内容語は内容語に、機能表現は機能表現に言い換えることが多い。しかし、一部の機能表現には、例えば、「走る よりほかにしかたがない」を「走る 以外に方法はない」に言い換えるように、内容語を含んだ表現を用いて言い換えることができるものがある。このような言い換えを実現した研究は、敬語表現を言い換える研究 [3] を除いてほとんど例がなく、松吉らのシステム [2] でも、このような言い換えは実現されていない。

本研究では、松吉らのシステムを拡張し、機能表現をより多様な類義表現へ言い換えるシステムを実現する。具体的には、内容語を含んでいるが、機能表現とほぼ等価な意味を表し、機能表現の言い換えとして利用できるように表現（準機能表現）を収集し、このような表現への言い換えを実現した。本稿では、作成した準機能表現辞書と、システムの拡張点、および評価について述べる。

2 ベース・システム

本研究では、松吉らによって提案された機能表現言い換えシステム [2] をベース・システムとして利用する。

このシステムが利用する機能表現辞書は、16,771 表現からなる機能表現の表記のリストを含んでおり、それぞれの表現に対して、左接続情報、属する意味的等価クラス、難易度、文体等の情報が付与されている。ベース・システムは、言い換えるべき文節と、出力すべき代替表現の難易度と文体指定を入力として、次のような手順で代替表現のリストを生成する。

Step 1. 文節解析

- 形態素解析器 MeCab¹を用いて入力文節を形態素解析する。
- 機能表現辞書を参照し、入力文節を内容語部と機能語部に分割する。機能語部に含まれるそれぞれの機能表現に対して、属する意味的等価クラスを得る。

Step 2. 類義表現の獲得

- 得られた意味的等価クラスと、指定された難易度、文体の情報を利用し、機能語部に含まれる機能表現の類義表現を列挙する。

Step 3. 代替表現の生成

- 列挙された類義表現の左接続情報をもとに、内容語部と機能表現を接続する。
- 内容語部を活用させる必要があれば、活用させて接続する。
- 語の挿入が必要な場合、その語を挿入し、全体を接続する。
- 入力文節内に機能表現が複数存在する場合、前の機能表現の末尾要素と、後ろの機能表現の左接続から、それらを接続する。

Step 4. 順位付け

- 毎日新聞 1991-2005 年版（計 15 年分）における機能表現の出現回数に基づき、順位付けする。

3 準機能表現辞書

本研究では、内容語を含んでいるが、機能表現とほぼ等価な意味を表し、機能表現の言い換えとして利用できるように表現を、**準機能表現**とよぶ。我々は、準機能表現に以下で述べる情報を付与し、準機能表現辞書を作成した。

¹ <http://mecab.sourceforge.jp/>

* Automatic Paraphrasing of Japanese Functional Expressions to Various Similar Expressions.

Tatsuya Masuda[†], Suguru Matsuyoshi^{††,‡}, Satoshi Sato[‡][†] Department of Electrical and Electronic Engineering and Informatics, School of Engineering, Nagoya University^{††} Graduate School of Informatics, Kyoto University[‡] Graduate School of Engineering, Nagoya University

3.1 準機能表現の収集

非母語話者向けの日本語の教科書では、機能表現の意味を説明する際に、内容語を用いた等価な表現を示す場合がある。このような表現は、準機能表現として利用することができる。我々は、「どんな時どう使う日本語表現文型 500 中・上級」[5]を対象に、準機能表現の収集を行なった。

この本では、機能表現が意味的に 30 種類に分類され、それぞれの意味に分類される機能表現に対して、その意味的特徴、文法的性格、例文等がまとめられており、ほとんどの機能表現に対しては、その意味を説明する代替表現が記述されている。例えば、「～からすると」に対しては、「～の立場から考えると」が代替表現として記述されている。ここでは、まず、機能表現と代替表現の組をすべて収集した。総数は、494 組であった。

こうして得られた機能表現と代替表現の組には、次の 3 種類のものが混在している。

1. 機能表現から機能表現への言い換え
2. 機能表現から準機能表現への言い換え
3. 呼応表現の言い換え

本研究の焦点は準機能表現への言い換えであるので、収集したものの中から 2 に該当するものを選んだ。さらに、ベース・システムとの統合を前提としているので、言い換え元をベース・システムの機能表現辞書に含まれる機能表現に限定した。この 2 つの条件を満たすものは 128 組であり、そこに含まれる準機能表現の異なりは、97 種類であった。

3.2 辞書情報

こうして得られた 97 個の準機能表現に対して、以下の 3 つの情報を記述した。

3.2.1 左接続情報

機能表現辞書と同じように、左接続情報を記述した。左接続情報は、「名詞」、「動詞連用形」、「助動詞“た”」など、左に隣接する語の形態素で記述する。例えば、「～の立場から考えると」という表現の左接続は、「名詞」となる。これは「～」の部分に名詞が入るということの意味する。

3.2.2 意味的等価クラス

ベース・システムの機能表現辞書に登録されている機能表現は、199 の意味的等価クラスに分類されている。準機能表現に対しても同じ分類を採用し、準機能表現のそれぞれに対して、言い換え元の機能表現を手がかりに意味的等価クラスを付与した。付与した意味的等価クラスは、66 種類であった。

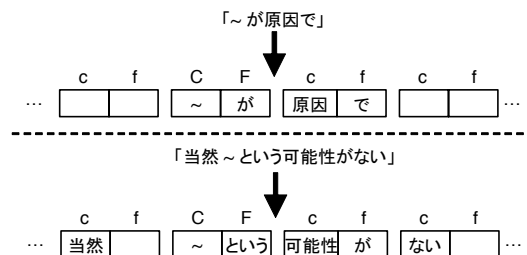


図 1: 文節情報

3.2.3 文節情報

機能表現を別の機能表現に言い換える場合は、1 文節内の機能語部のみを変更すればよい。しかしながら、準機能表現は内容語を含むので、次の例に示すように、変更の対象は複数の文節に及ぶ。 (“/” は文節区切りを表す。)

- 文節数の増加
例) あなた からすると
→ あなた の/立場から/考えると
- 内容語部前に副詞等の挿入
例) 走ら ざるを得ない
→ どうしても/走る/必要が/ある

このような複数の文節に及ぶ変更操作を実現するために、準機能表現に、(1) 文節境界、(2) 内容語部・機能語部の境界、の情報を付与した。その例を図 1 に示す。この図において、例えば、「～が原因で」という表現には、「～が」と「原因で」の間に文節の境界があり、さらに、言い換え元の内容語部を表す「～」と「が」、「原因」と「で」の間にそれぞれ内容語部と機能語部の境界があることを意味する。

4 提案システム

本研究で作成したシステムの全体像を図 2 に示す。このシステムは、機能語部が 1 つあるいは複数の機能表現からなる文節を入力とし、代替表現の順位付きリストを出力する。代替表現は、1 つの文節、または、文節列のいずれかである。

このシステムの特徴は次の 2 つである。

- 可能な限り多くの代替表現を出力する
- 一般的に使用される表現を上位に出力する

本システムの代替表現生成手順は、2 節で述べたベース・システムとほぼ同様で、4 つのステップから構成される。本システムは、このうち、最初のステップを除く 3 つのステップにおいて、変更・拡張されている。主な変更点は、以下の 3 つである。

- 準機能表現辞書を統合する
- 言い換えにより生じる文節数増加に対応する

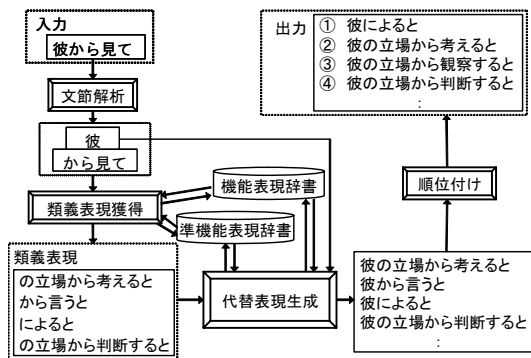


図 2: システムの全体像

- 順位付け手法を改良する

4.1 類義表現の獲得

第2のステップである類義表現の獲得では、文節解析で得られた機能表現の意味的等価クラスから、同じクラスを持つ機能表現を列挙する。本システムでは、機能表現辞書に加えて準機能表現辞書も検索し、同等の意味を持つ準機能表現も列挙する。

4.2 代替表現の生成

本システムの第3ステップは、代替表現生成である。ステップ2で列挙された類義表現が準機能表現である場合、このステップは、(1) 内容語部の右に隣接する語の接続、(2) 前後の文節の付与、から構成される。また、入力文節に機能表現が複数含まれる場合は、この処理を再帰的に適用する。

4.2.1 入力の内容語部とその右に隣接する語の接続

類義表現の左接続情報をもとに、入力の内容語部と、その右に隣接する語を接続する。この処理は、ベース・システムの内容語部と機能語部の接続の処理に相当する。この処理には、ベース・システムのアルゴリズムをそのまま使い、必要があれば内容語部を活用したり語を挿入したりする。例えば、「動く」と「～が原因で」を接続する場合、「動く」と「が」の間に、「こと」という語を挿入し、「動くことが」という文節を生成する。

4.2.2 前後の文節の付与

上記の処理で、準機能表現の中心となる文節が生成されるので、次に、この文節の前後に、準機能表現の残りの文節を接続する。例えば、上記の処理で「動くことが」が生成されるので、その後ろに「原因で」を接続し、「動くことが原因で」という代替表現を生成する。また、「動く」と「当然～という可能性がない」の接続のような、前に副詞等が付く場合は、前節で生成された「動くという」に、まず、「可能性が」と「ない」を接続して「動くという可能性がない」を生成し、最後に「当然」を一番前に接続することで「当然動く」と

いう可能性がない」を生成する。

4.2.3 複数の機能表現が含まれる場合

入力文節内に複数の機能表現が含まれる場合は、上記の処理を再帰的に適用する。例えば、「寂しくてしかたがないにきまっている」という表現を言い換える過程は、以下になる。

「寂しくてしかたがないにきまっている」

↓ (文節解析)

「寂しく」「てしかたがない」「にきまっている」

↓ (類義表現獲得)

「寂しく」「我慢できないほど～」 「～に違いがない」

↓ (接続 1)

「我慢できないほど寂しい」「～に違いがない」

↓ (接続 2)

「我慢できないほど寂しいに違いがない」

4.3 順位付け

ベース・システムと同様に、毎日新聞 1991-2005 年版 (計 15 年分) における出現回数をもとに順位付けする。

準機能表現を含む代替表現の多くは、表現の文字数が多いため、表現全体が新聞に一度も出現しない。これを考慮して、出力が2文節以上の場合と、1文節の場合に分け、以下の方法でスコア付けをする。

出力が2文節以上の場合は、次の方法でスコアを計算する。表現 E の先頭から i 番目 (先頭を 0 とする) の文節の内容語部を c_i 、機能語部を f_i とする。ここで、 n_i を次のように定義する。

$$n_i = \text{“} c_i f_i c_{i+1} \text{ の出現回数”} \quad (1)$$

この n_i を用いて、代替表現 E のスコアを次のように定義する。

$$S(E) = \frac{\sum_{k=0}^{N-2} n_k}{N} \quad (2)$$

ここで、 N は表現 E の文節数を表す。

出力が1文節の場合、代替表現 E のスコアを次のように定義する。

$$S(E) = \text{“} c_0 f_0 \text{ の出現回数”} \quad (3)$$

このようにして得られたスコアにより、生成された代替表現を順位付けする。

4.4 システムの実行例

以下に、システムの実行例を示す。

- 動く わけがない
 - (1位) 当然 動く という可能性がない
 - (2位) 動く まい
 - (3位) 動く はずがない
 - (4位) 決して 動か ない
- 走る とともに
 - (1位) 走る につれて
 - (2位) 走る と、それと一緒に
 - (3位) 走る と、だんだん
 - (4位) 走る にしたがって

5 システムの評価

以下の点に着目し、言い換えシステムのクローズドテストを行った。

- 接続の誤りはないか
- 出力される表現は、入力表現の代替表現として適切であるか

この評価では、「どんな時どう使う日本語表現文型 500 中・上級」の例文から、機能表現辞書に記述されている機能表現を含み、かつ、準機能表現を用いた代替表現が出力されると期待できる文節、341 文節を抽出し、システムの入力とした。評価基準は、以下のよう

に定めた。

評価A 代替表現として適切である

評価B 接続は誤っていないが、代替表現として適切ではない

評価C 接続が誤っている

なお、評価は、1 人の作業者が主観に基づいて行った。また、出力されたすべての代替表現を評価の対象としたかったが、例えば、「寂しくてしかたがないにきまっている」という表現に対して、4,664 個の代替表現が出力され、評価に多大な時間がかかるため、出力の上位 4 位までを評価の対象とした。クローズドテストの結果を表 1 と図 3 に示す。図 3 は、入力全体の 341 文節に対して、1 位の代替表現が評価 A の割合、1 位と 2 位が評価 A の割合、1 位から 3 位までがすべて評価 A の割合、1 位から 4 位までがすべて評価 A の割合をグラフで示したものである。

評価 C の出力はクローズドテストを繰り返すことにより取り除くことができた。図 3 に示されるように、1 位が評価 A となるものは全体の 84% とよい精度であったが、1 位から 4 位がすべて評価 A となるものは全体の 44% であり、改善の余地がある。評価が B となる主な理由は、限られた表現にしか適用できない言い換えの問題である。一般に、準機能表現は内容語を含むため、機能表現に比べてより限定された意味を持つ。例えば、「あなたにしても」は、「あなたの気持ちでは」と言い換えることができるが、「東京にしても」は「東京

表 1: クローズドテスト結果

	1 位	2 位	3 位	4 位	合計
評価 A	285(84%)	263(82%)	246(83%)	198(75%)	992(81%)
評価 B	56(16%)	56(18%)	50(17%)	65(25%)	227(19%)
総数	341(100%)	319(100%)	296(100%)	263(100%)	1,219(100%)

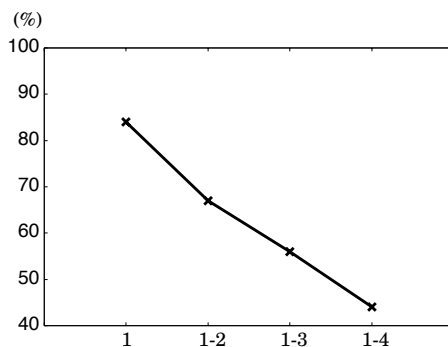


図 3: クローズドテスト結果

の気持ちでは」と言い換えることはできない。「～の気持ちでは」という準機能表現は、内容語部が人物を表す語の場合のみ、言い換え先の表現として使用可能である。

6 おわりに

本研究では、内容語を含んでいるが、機能表現とほぼ等価な意味を表す表現を、準機能表現と位置づけて整理し、準機能表現辞書を作成した。さらに、松吉らが提案する日本語機能表現言い換えシステムを拡張し、機能表現を多様な類義表現へ言い換えるシステムを実装した。

今後は、準機能表現辞書の規模を拡大するとともに、5 節で示した課題を解決し、言い換えることができる表現の増加と、言い換えシステムの精度向上を目指す。

謝辞 本研究は、科学研究費補助金 基盤研究 (A) 「円滑な情報伝達を支援する言語規格と言語変換技術」(課題番号 16200009) の支援を受けた。

参考文献

- [1] 伊佐治和哉, 山田佳裕, 石原吉晃, 高松大地, 松本忠博, 池田尚志. 文節構造解析システム ibukic. 言語処理学会 第 11 回年次大会発表論文集, pp. 719–722, 2005.
- [2] S. Matsuyoshi and S. Sato. Automatic Paraphrasing of Japanese Functional Expressions Using a Hierarchically Organized Dictionary. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp. 691–696, 2008.
- [3] K. Ohtake and K. Yamamoto. Paraphrasing Honorifics. In *Proceedings of NLP2001 Workshop on Automatic Paraphrasing: Theories and Applications*, pp. 13–20, 2001.
- [4] K. Shudo, T. Tanabe, M. Takahashi, and K. Yoshimura. MWEs as Non-propositional Content Indicators. In *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing (MWE-2004)*, pp. 32–39, 2004.
- [5] 友松悦子, 宮本淳, 和栗雅子 (編). どんな時どう使う日本語表現文型 500 中・上級. 株式会社アルク, 1996.