選択記号による文型パターンの汎化の効果

小林 和晃　村上 仁一　徳久 雅人　池原 悟
鳥取大学 工学部 知能情報工学科
{kkobayas,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

1　はじめに

近年, 機械翻訳の方式として等価的類推思考の原理に基づく機械翻訳方式が提案されている [1]. この方式の実現に向けて, 日本語の重文・複文を対象とした文型パターンを大量に蓄積した文型パターン辞書の構築が進められている [2]. 文型パターンは, 言語表現を, 字面・変数・関数・記号で記述したものであり, パターンマッチングにより入力文を解析する. 現在, 文型パターン辞書には単語レベル・句レベル・節レベルが存在する. この単語レベル文型パターンの問題点の一つに, 入力文に対し約 48% しか文型パターンが出力されておらず, 現状では適合率が低いことがあげられる. また, 現在の単語レベル文型パターン辞書には, 入力文に対する適合率を向上させる手段として, 表記のゆらぎを吸収するために, 選択記号が記述されている.

そこで, 本研究では単語レベル文型パターンにおける選択記号の効果を「文型パターン拡大率 η」, および「適合率 $R1$」を用いて, 定量的に評価し, 改良の可能性を検討する. また, 現在の単語レベル文型パターン辞書は, 選択記号になるべき箇所が記号になっていなかったり, 表現要素の表記が不足している. そこで, それらの箇所に対し, 既存の選択記号で最も表現要素数が多い選択記号による均一化, および既存の選択記号から新たに作成した選択記号による均一化を行うことで選択記号を増加したときの文型パターン拡大率と適合率も同様に評価する.

2　単語レベル文型パターンにおける選択記号

2.1　単語レベル文型パターン辞書の概要

文型パターンは, 日英対訳標本文を, 変数化および関数化, 任意化している. その中で単語レベル文型パターンは, 表現に使用される名詞, 動詞などの自立語の線形な表現要素を変数化している. また, 変数化すると対訳の訳出が困難になる部分は非線型な表現要素として字面, あるいは関数の形式で残されている. 単語レベル文型パターンの例を以下に示す.

- **日本語原文**　自分ひとりで何でもやるのが彼の主義だ。
- **日本語パターン**　$N1$(ひとりで｜一人で)(何でも｜なんでも) やるのが /#2[$N3$] の /$N4$.da
- **英語原文**　It is his principle to do anything whatever for himself.
- **英語パターン**　It is #$N2$[$N3$.$poss$]$N4$ to do anything whatever for $N1$.$reflex$.

変数には名詞や動詞の変数を表す N_n や V_n など 8 種類がある. 関数には $.da$ や $.kako$ などがあり, 字面の指定や表現を指定している. 詳細は [2] に示されている.

2.2　選択記号

選択記号とは, 表現要素のグループ化を行うため, 助詞, 助詞相当語, または, 副詞などの字面のうち, 同一の意味で異なる表記を持つものを対象に, 置き換え可能な表現として指定したものであり, (…｜…) のように表記する.

2.1 節の例において日本語パターンに (ひとりで｜一人で) や (何でも｜なんでも) という選択記号を付与することにより, 日本語原文の「自分ひとりで何でもやるのが彼の主義だ。」だけでなく,「自分一人で何でもやるのが彼の主義だ。」や「自分ひとりでなんでもやるのが彼の主義だ。」のように表現がゆらいでも文型パターンパーサで受理可能になる.

3　選択記号の効果の調査

3.1　調査方法

選択記号の汎化の効果を, [3] および [4] で示されている文型パターン拡大率 η および適合率 $R1$ を用いて定量的に評価する. 以下に, 各評価パラメータの概略を示す.

＜文型パターン拡大率 η ＞

η は「評価対象の文型パターン辞書の文型パターンが基準となる文型パターン辞書の文型パターン数に換算して, 何倍に相当するか」を表したものである. 定義を次式に示す.

$$\eta = X/B$$

X：対象文型パターン辞書の選択記号を全て展開したときの文型パターン数

B：基準文型パターン辞書の文型パターン数

＜例＞　例 1 の日本語パターンは, 例 2 のように 3 つの日本語パターンに展開できるため, このときの η は 3.00 となる.

- **＜例 1＞**　/$ytkTIME1$ も/cf あい (変わり｜かわり｜変り) ませず!お付き合いの/k ほど < /$tkN2$ は > /$tcfk$ お願い申し上げます。
- **＜例 2＞**　/$ytkTIME1$ も/cf あい変わりませず! お付き合いの/k ほど < /$tkN2$ は > /$tcfk$ お

願い申し上げます。
/$ytkTIME1$ も/cf あいかわりませず!お付き合いの/k ほど < /$tkN2$ は > /$tcfk$ お願い申し上げます。
/$ytkTIME1$ も/cf あい変りませず!お付き合いの/k ほど < /$tkN2$ は > /$tcfk$ お願い申し上げます。

＜適合率 $R1$ ＞

適合率 $R1$ は, 入力文に対して受理された文型パターンが存在する割合を文単位で集計したものである. 定義を次式に示す.

$$R1 = M/I$$

M :「自己パターン」以外に受理された文型パターンが存在する入力文の数

I : テスト用入力文の数

本研究では入力文として, 単語レベル文型パターン辞書作成に使用した日本語原文 123,451 文を使用する. 入力文と単語レベル文型パターン辞書を文型パターンパーサ jpp[5] を用いて照合を行ない, 照合結果から適合率を求める. 文型パターンパーサは入力文が受理できる文型パターンを全て出力するプログラムである.

3.2 調査対象

選択記号の効果を求めるため, 以下の単語レベル文型パターン辞書を作成し, 文型パターン拡大率と適合率を評価する.

(1) 選択記号を無くした単語レベル文型パターン辞書 (選択記号無し)

選択記号自体の効果を求めるため, 選択記号を日本語原文と同じ表現要素のみにし, 選択記号を無くした単語レベル文型パターン辞書を作成する. 作成手順を以下に示す.

手順 1 現在の選択記号で, 日本語原文と同じ表現要素だけを残し, 残りの要素を削除する.

＜例＞ /$ytkTIME1$ も /cf あい (変わり｜かわり｜変り) ませず ! お付き合いの /k ほど < /$tkN2$ は > /$tcfk$ お願い申し上げます。
→ /$ytkTIME1$ も /cf あい変わりませず ! お付き合いの /k ほど < /$tkN2$ は > /$tcfk$ お願い申し上げます。

(2) 現在の単語レベル文型パターン辞書 (オリジナル)

本研究では, [2] で作成された単語レベル文型パターン辞書 (ver.5.3.1) を使用する. なお, この単語レベル文型パターン辞書のパターン総数は 122,619 パターンである.

この単語レベル文型パターン辞書において, 選択記号の述べ数は 72,208 個, 種類数は 3,652 種類であった.

(3) 既存の選択記号で最も表現要素数が長い選択記号に均一化した単語レベル文型パターン辞書 (最長均一化)

現在の単語レベル文型パターン辞書には, 例 3 の選択記号を付与された日本語パターンがあるにも関わらず, 例 4 のような同じ表現要素を持ちながら表現要素数が少ない選択記号が付与されている日本語パターンがある. また, 選択記号となるべき表現要素が選択記号になっておらず字面で残っている日本語パターンもある. そこで該当する要素に既存の選択記号で最も表現要素数の長い選択記号を日本語パターン付与した単語レベル文型パターン辞書を作成する.

＜例 3＞ (会う｜あう｜逢う)
＜例 4＞ (会う｜あう｜逢う｜遇う｜遭う｜會う｜遘う)

作成手順を以下に示す.

手順 1 日本語パターンを形態素解析し品詞番号を付与する. なお, 選択記号内の他の単語は原文内の単語と同じ品詞番号を付与する.

手順 2 手順 1 で作成した日本語パターンから選択記号を抽出する.

手順 3 $N1$ や $V2$ などの変数はどの変数にでも受理できるように変数番号を $N*$, $V*$ のように汎化する.

手順 4 抽出した選択記号から, 単語を選択記号に, あるいは選択記号を同じ表現要素を持ちながらさらに表現要素数が多い選択記号に置き換える辞書を作成する.
ただし, $V*$ と $ND*$ に関してはさまざまに受理され置き換わってしまう可能性があるため辞書から削除した.

手順 5 辞書に従い, 選択記号を置き換える.

この単語レベル文型パターン辞書において, 選択記号の述べ数は 190,239 個であり, 種類数は変数を汎化したため 2,669 種類に減少した.

(4) 既存の選択記号から, 新たに作成した選択記号に均一化した単語レベル文型パターン辞書 (新作成)

例えば, 例 5 の選択記号と例 6 の選択記号は, 一部同じ表現要素を持っている. この 2 つの選択記号は例 7 のように一つにまとめることができる.

このように一部同じ表現要素を持っている選択記号どうしを組み合わせて新たに選択記号を作成し, 日本語文型パターンに付与することで新たな単語レベル文型パターン辞書を作成する.

＜例 5＞ (上がっ｜上っ｜あがっ｜のぼっ｜上ぼっ｜躋っ｜躋ぼっ｜隲っ｜隲ぼっ)
＜例 6＞ (挙がっ｜あがっ｜挙っ｜上がっ｜上っ｜擧がっ｜擧っ｜舉がっ｜舉っ｜驤がっ｜驤っ)
＜例 7＞ (上がっ｜上っ｜あがっ｜のぼっ｜上ぼ

っ｜躋っ｜躋ぼっ｜隲っ｜隲ぼっ｜挙がっ｜挙っ｜擧がっ｜擧っ｜擧がっ｜擧っ｜驤がっ｜驤っ)

作成手順を以下に示す.

手順 1〜3 単語レベル文型パターン辞書 (3) の作成手順 1〜3 に同じ.

手順 4 既存の選択記号のうち, 同じ表現要素を一つでも持っている選択記号があれば選択記号どうしを合わせ新たな選択記号を作成する.
ここで, 変数を含む選択記号は, 同じ表現要素を持つものが多く非常に長い選択記号となることが予想されたため, 2,669 種類の選択記号から変数を持たない選択記号 2,131 種類を使用し新たに選択記号を作成した.

手順 5 手順 4 で作成した選択記号と既存の選択記号を合わせ, 単語レベル文型パターン辞書 (3) の作成手順と同様に選択記号に置き換える辞書を作成する.

手順 6 辞書に従い, 選択記号を置き換える.

この単語レベル文型パターン辞書において, 選択記号の述べ数は 190,239 個であり, 種類数は新たに 779 種類作成し合計 3,448 種類になった.

3.3 調査結果

(1) から (4) までの単語レベル文型パターン辞書に対する, 文型パターン拡大率を表 1 に, 適合率を表 2 に示す. この結果, 現在の単語レベル文型パターンに付与されている選択記号は単語レベル文型パターン辞書の日本語パターン数を 2 倍相当にしていて, かつ適合率をおよそ 2% 向上させている. しかし, さらなる適合率の向上を狙い作成した辞書は, 付与を最も多く行った辞書 (単語レベル文型パターン辞書 (4)) で日本語パターン数が 7 倍近くに相当するにもかかわらず, 適合率がほとんど向上していない.

表 1 各辞書に対する文型パターン拡大率

辞書	展開パターン数	文型パターン拡大率
(1) 選択記号無し	122,619	1.00(122,619/122,619)
(2) オリジナル	245,850	2.00(245,850/122,619)
(3) 最長均一化	711,055	5.80(711,055/122,619)
(4) 新作成	826,758	6.74(826,758/122,619)

表 2 各辞書に対する適合率

辞書	自己以外に受理	適合率
(1) 選択記号無し	56994	46.330(56994/123451)
(2) オリジナル	60180	48.748(60180/123451)
(3) 最長均一化	60243	48.779(60243/123451)
(4) 新作成	60248	48.803(60248/123451)

4 考察

4.1 選択記号の表現要素数に関する調査

単語レベル文型パターン辞書 (4) は, 既存の選択記号から新たに選択記号を作成し, その中で最も長い選択符号に均一化している. この単語レベル文型パターン辞書 (4) の選択記号の表現要素が適合率を向上させる効果があるかを調査するため, 以下の実験を行った.

入力文 12 万文に対し文型パターンパーサで照合を行い, 各選択記号に対して照合の際に使用された頻度をとった. そして各表現要素位置における使用された頻度の平均値を求めた. 調査結果を表 3 に示す.

表 3 選択記号で使用される表現要素の位置の平均

表現要素位置	割合
第 1 要素	86.800
第 2 要素	11.655
第 3 要素	1.267
第 4 要素	0.202
第 5 要素	0.059
第 6 要素	0.013
第 7 要素	0.002
第 8 要素以降	0

表 3 より, 全体の 98% は第 2 要素までに使用している. これにより, 最も表現要素数が長い選択記号に均一化し, 選択記号の表現要素数を増加させても適合率の向上が低いことが分かった.

4.2 人手で言い換えた入力文を用いた調査

本研究で使用した入力文, すなわち単語レベル文型パターンを作成するために使用した標本文は, 辞書や語学教育用の教科書, 機械翻訳機能評価用の試験文などで構成されている. これらは, 日本語の基本的な表現で収録されているため表現のゆらぎが少ないと考えられる. そこで, 入力文に対し人手で言い換えを行い, その文における受理パターン率を調査した.

調査対象として, 入力文からランダムで 114 文を抽出し, 人手で 641 文に言い換えた. 言い換えた 641 文のうち, 形態素解析で誤った 40 文を除いた 601 文を調査対象とした. 各辞書において受理された文数を表 4 に示す.

表 4 言い換えにより受理された文数

辞書	受理された文数	受理されなかった文数
(1) 選択記号無し	401	200
(2) オリジナル	432	169
(3) 最長均一化	425	176
(4) 新たに作成	425	176

この結果を見ると, 選択記号が無い辞書 (1) に比べ現在の辞書 (2) ではわずかに受理パターン数が増加するが, 辞書 (2) と既存の選択記号を増加させた単語レベル文型パターン辞書 (3), (4) とを比較しても受理された文数は変わらない. なお, 単語レベル文型パターン辞書 (2) に比べ単語レベル文型パターン辞書 (3), (4) の受理された文数が減っている理由は, 選択記号を増加させたことによる文型パターンパーサ jpp のバグだと考えられる.

次に, 元々の日本語原文と人手で言い換えた日本文における適合率を比較した. 使用した辞書は選択記号を最も多く付与した単語レベル文型パターン辞書 (4) である. 結果を表 5 に示す.

表 5　人手で言い換えた日本文に対する適合率

入力文	入力文数	自己以外に受理	適合率
日本語原文	114	57	50.00
言い換えた日本文	601	305	50.75

この結果, 入力文の種類が変わっても適合率にそれほどの差は無いことが分かった.

次に, 単語レベル文型パターン辞書 (4) において受理されなかった 176 文と見ると, 例 8 や例 9 のように, わずかな表現のゆらぎしかないが受理不可能になる文が 63 文存在した. この 63 文の受理不可能になった箇所は選択記号に置き換えることで受理できると考えられる. また, 残りの 113 文に関しては, 例 10 や例 11 のように,「サ変名詞＋する」と動詞の変化や, 名詞と名詞の変化などがほとんどであった. この結果, 選択記号の箇所が不足していると考えられる. また, これらの箇所が全て改善されれば, 適合率はおよそ 80%$((305 + 176)/601 = 0.8003)$ まで向上が期待される.

＜例 8＞　そうするのはどうしてもいやだと言う。(受理)
そうするのはどうしても 嫌 だと言う。(受理不可)

＜例 9＞　彼はあまりなれなれしいから人に嫌われる。(受理)
彼はあまりなれなれしい ので 人に嫌われる。(受理不可)

＜例 10＞　目的地まで遠いから、時々休みながら行く。(受理)
目的地まで遠いから、時々 休憩し ながら行く。(受理不可)

＜例 11＞　頭痛は明くる日になっても直らなかった。(受理)
頭痛は 翌日 になっても直らなかった。(受理不可)

4.3　選択記号の表現要素を新たに発見する方法

選択記号内の表現要素を発見する方法として, 現在の単語レベル文型パターン中の選択記号が付与されている箇所を 2 形態素までなら文型パターンパーサで受理できるようにし, 置き換えた表現要素が選択記号の表現要素として使用可能かどうかを調査した.

＜例 12＞　日本語原文：そこで笑ってはだめだ。
日本語パターン：$/y\#1[$ そこで $]/fV2$(て｜で) は $/cf$ だめだ。
→ $/y\#1[$ そこで $]/fV2*$ は $/cf$ だめだ。

具体的な例を示す. 例 12 において, 日本語パターン中の選択記号 (て｜で) のかわりにどのような表現要素でも受理可能である $*$ の記号をつけ, 入力文 12 万文と照合を行った.

$*$ で受理された表現要素を受理された回数でソートし, 上位 100 件ほどを調べたところ,「ていて」という表現要素を発見した. この表現要素は, 例 13 の文を入力しても受理するよう (て｜で｜ていて) のように新たに選択記号の表現要素として追加しても構わないと考えられる.

＜例 13＞　そこで笑っていてはだめだ。

しかし, 選択記号 (て｜で) の箇所のかわりに受理された表現要素は本来 103,964 件あり, その中から人手で追加可能な表現要素を捜すのは困難であった.

5　おわりに

単語レベル文型パターンにおける選択記号は単語レベル文型パターン辞書の日本語パターン数を 2 倍相当にしていて, かつ適合率をおよそ 2% 向上させていることが分かった. さらなる向上を狙い, 既存の選択記号で最も表現要素数が多い選択記号による均一化, および既存の選択記号から新たに作成した選択記号による均一化を行ったが, 狙った程の効果は得られなかった.

これにより, 現在の付与されている選択記号に関しては表現のゆらぎを吸収するに十分な効果があるといえる. また, 人手で言い換えた日本文に対しても, 日本語原文と同じだけの適合率が示され, 母集団が変わっても日本語原文を入力文とした際と同じだけの効果があることが分かった. しかし, 単語レベル文型パターン中に選択記号となるべき箇所が残っていることも示された.

謝辞

本研究は, 科学技術振興事業団「JST」の戦略的基礎研究推進事業「CREST」における研究領域「高度メディア社会の生活情報技術」の研究課題「セマンティックタイポロジーによる言語の等価変換と生成技術」の支援によるものである. また, 研究に協力していただいた研究室メンバーの片山慶一郎君に感謝する.

参考文献

[1] 池原ほか:等価的類推思考の原理による機械翻訳方式, 電子情報通信学会技術研究報告, TL2002-34, pp.7-12, 2002.
[2] 池原ほか:非線型な表現構造に着目した重文と複文の日英文型パターン化, 言語処理学会論文誌, Vol.11, No.3, pp.69-95, 2004.
[3] 遠藤ほか:文型パターンにおける任意要素の記述方法とその効果, 言語処理学会第 11 回年次大会発表論文集, pp.368-371, 2005.
[4] 池原ほか:日本語重文・複文を対象とした文法レベル文型パターンの被覆率特性, 言語処理学会論文誌, Vol.11, No.4, pp.147-178, 2004.
[5] 徳久ほか:文型パターンパーサの試作, 言語処理学会第 10 回年次大会発表論文集, pp.608-611, 2004.