日本語文型パターンの縮退方法

片山 慶一郎　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
{kkatayam,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

1　はじめに

現在，等価的類推思考に基づく，日英機械翻訳方式が提案されている [1]．この方式は，言語表現の持つ非線形性に着目し，表現構造と意味を一体として扱う．そのために必要な日英対訳文型パターン辞書が構築されている [2]．現在，この文型パターン辞書は，24 万件の対訳パターンを持つため，規模が大きく取り扱いが困難である．また，カバー範囲に重複が見られるため，パターン数を削減することが必要である．しかし，人手による削減は困難である．そこで，本稿では日本語文型パターン間の包含関係に着目して，その数を半自動的に削減する方法を検討する．

2　パターンの削減方法

2.1　パターン間の包含関係

パターン $P1$ が受理する入力文の全てが，パターン $P0$ に受理される時，パターン $P1$ はパターン $P0$ に包含されると定義する．また，パターン $P0$ を上位のパターン，パターン $P1$ を下位のパターンと呼ぶ．両者の関係を $P0 \supseteq P1$ と表記する．

以下に包含関係にあるパターンの例を示す．

$P0$:　$N1$　は　$V2$　。
$P1$:　私　は　$V1$　。

パターン $P1$ の変数 $V1$ は動詞を表すため，パターン $P1$ は「私は 見る。」や「私は 行く。」など，「私は（動詞）。」の文を受理する．また，パターン $P0$ の変数 $N1$ は名詞を表すため，パターン $P0$ は「彼 は 見る。」や「地球 は 回る。」など，「（名詞）は（動詞）。」の文を受理する．ここで，「私」は名詞であることから，パターン $P0$ も「私は（動詞）。」の文を受理可能である．したがって，パターン $P1$ が受理する全ての日本語文は，パターン $P0$ も受理可能である．

本稿では，パターン辞書から下位パターンを削除することで，パターン数の削減を試みる．

2.2　パターン要素間の包含関係

前節の定義に従って，パターン間の包含関係を判定するためには，全ての入力文に対して受理の可否を調査する必要がある．しかし，全ての入力文に対して調査を行うことは不可能である．そこで，パターン自身が受理可能な入力文の領域を表している事に着目して，パターンが別のパターンに受理されるかどうかを調査することで，包含関係を判定する．

パターン間の包含関係を考える場合，パターンを構成する要素（変数，関数，記号，字面）の包含関係を定義する必要がある．そこで，各要素の定義 [3] に基づいて包含関係を定義する．表 1 に定義した要素間の包含関係の一部を示す．

表 1　パターン要素間の包含関係（一部）

上位の要素	下位の要素
N（名詞）	NUM（数詞），$TIME$（時詞） ND（用言性名詞）
NP（名詞句）	N，N の下位要素
VP（動詞句）	V（動詞）

3　日本語文型パターン

3.1　パターン記述言語

日本語文型パターンは，可読性があること，線形要素の対応関係が明確であること，の 2 点を主な条件として設計され，「字面」，「記号」，「変数」，「関数」の 4 種類の要素がある [3]．

変数には，名詞を表す N，動詞を表す V，名詞句を表す NP などがある．関数は，変数が受理する値の形式や，字面の指定，表現の統括を行う．記号は，パターン要素の受理方法について，任意化，選択，順序変更，記憶という制御を行う．表 2 に記号の例を示す．詳細は参考文献 [3] に示されている．

表 2　記号一覧

記号名	表記	意味
離散記号	/...	文型に無関係な要素
選択記号	(…)	いずれかの要素列を受理
任意記号	[…]	文型選択上，任意の要素
補完要素記号	<…>	ゼロ代名詞等
順序任意要素指定記号	{…,…}	順序入れ替え可能な範囲
位置変更可能要素指定記号	$n^{…}, $n	指定位置に入れ替え可能
文節境界記号	!	文節の境界を受理

3.2　パターン例

本稿で扱う日本語文型パターンは，日英対訳例文の線形要素を変数化し，関数・記号を付与したものである．今回，文型パターン辞書に収録されている文法・単語レ

ベルの 122,619 パターンを用いる．以下に文法・単語レベル文型パターンの例を示す．

日本語原文: 彼らは雪の深い冬を越した。
英語原文: They got through the winter of deep snow.
日本語パターン: */y* $1^{*/tk N1* は } */tcfk N2* の */f AJ3^rentai* ! *TIME4* を $1 */cf V5.kako*。
英語パターン: *N1 V5.past TIME4* of *AJ3 N2*.

なお，以後，日本語文型パターンをパターンと記述する．

3.3 文型パターンパーサ

文型パターンパーサ [4] は，パターンと日本語文との照合を行うプログラムである．照合方式は，ATN(Augmented Transition Network)[5] をベースとしている．本稿では，これを用いて包含関係の判定を行う．また，以後パーサと記述する．

4 包含関係による削減の実験

4.1 包含関係の判定方法の実装

パーサは日本語文とパターンとの照合を想定して作成されている．そのため，パターンをパーサの入力仕様に合わせるために次の作業を行う．

1. 要素選択記号，任意記号などを展開
   日本語文には，パターン記述に用いる，要素選択記号，任意記号等に対応する表現方法は存在しない．そこで，表 2 で示す記号の定義を元にパターンの展開を行う．以下に展開例を示す．
   パターン： */y* $1^{*/tk N1* は } */cf V2* ながら $1 */f* (*V3*|*ND3* をする)。
   展開後：
   (a) */y /tk N1* は */cf V2* ながら */f V3*。
   (b) */y /cf V2* ながら */tk N1* は */f V3*。
   (c) */y /tk N1* は */cf V2* ながら */f ND3* をする。
   (d) */y /cf V2* ながら */tk N1* は */f ND3* をする。
2. 変数等を含むパターンの形態素解析結果を作成
   パーサの入力は日本語文の形態素解析結果である．そこで，要素選択記号，任意記号などを展開したパターンを，原文とパターンの記述仕様に従って自動的に形態素解析を行い，パーサの入力仕様に適合する形式に変換する．
3. 変数・関数のオートマトン定義の修正
   入力文としての変数・関数をパターン側の変数が表 1 で示す包含関係を用いて受理する様に，パーサの ATN で用いるオートマトン定義の追加修正を行う．今回，98 個の変数・関数のオートマトン定義に対して 109 ヶ所の追加・修正を行った．

4.2 照合結果を用いた包含関係の判定

パーサの照合結果を用いて，パターン間の包含関係の判定を行う．本稿では，パターン中の記号を展開しているため，展開後のパターンの照合結果を総合して判定する必要がある．そのため，次の手順でパターン A とパターン B の間の包含関係 ($A \supseteq B$) の有無を判定する．

1. パーサの照合結果より，n 個に展開された入力パターン ($B_1 \cdots B_n$) とパターン A との照合結果 ($B_{result_1} \cdots B_{result_n}$) を抽出
2. パターン B がパターン A に含まれる割合を $Cov_{A \supseteq B}$ と定義し，次式に従い算出

$$Inc_{A \supseteq B_i} = \begin{cases} 1 & (B_{result_i} \text{ にパターン } A \text{ がある}) \\ 0 & (B_{result_i} \text{ にパターン } A \text{ が無い}) \end{cases}$$

$$Cov_{A \supseteq B} = \frac{1}{n} \sum_{i=1}^{n} Inc_{A \supseteq B_i}$$

3. 以下の 3 通りに分類
   (a) パターン A はパターン B を包含する．
   (b) パターン B の展開後パターンの一部はパターン A に含まれるが，全体では包含関係に無い．
   (c) パターン A はパターン B を包含しない．
   数式では以下のように記述する．

(a) $A \supseteq B$	($Cov_{A \supseteq B} = 1.0$)
(b) $A \not\supseteq B$ *and* $A \supseteq B_i, {}^{\exists} B_i \in B$	($0.0 < Cov_{A \supseteq B} < 1.0$)
(c) $A \not\supseteq B$	($Cov_{A \supseteq B} = 0.0$)

5 実験結果

5.1 包含関係を用いた削減結果

実験対象のパターンの包含関係を用いた削減結果を表 3 に示す．この結果，パターン数は 112,767 になった．削減率は 8.0% となった．

表 3 包含関係による削減結果

パターン数	122,619	(100.0%)
包含関係にあるパターン数	13,220	(10.8%)
上位パターン数	3,368	(2.8%)
下位パターン数（削減対象）	9,852	(8.0%)
削減後パターン数	112,767	(92.0%)

5.2 包含関係の判定結果

パーサの照合結果を用い，4.2 節に従ってパターン間の包含関係の判定した．以下に包含関係の例を示す．

- 例 1 ：$P2 \supseteq P3$
  P2: */ytk N1* は */tcfk* <u>*N2*</u> の */f* ある ! *N3.da*。
  P3: */ytk N1* は */tcfk* <u>力</u> の */f* ある ! *N2.da*。
  パターン *P3* では，名詞「力」が変数に置き換えられていない．そのため，パターン *P2* と *P3* で表記が異なっている．包含関係にあるパターンの多くは同様の理由であった．

- 例 2：$P4 \supseteq P5$

  $P4$: /y \$1^{/tk $N1$ は } /cf ($V2$|$ND2$ をし) (て | で) \$1 /cf $V3.kako$。

  $P5$: /y \$1^{/tk $N1$ は } /cf $V2$ (て | で) \$1 /cf 賛成した。

  パターン $P4$ の下線部の選択記号部分が，パターン $P5$ の下線部の変数 $V5$ を含む．包含関係は，選択記号等の機能によっても生じる．

6 考察

6.1 出現頻度と削減率の関係

包含関係により削減したパターン数と，出現頻度との関係を調査した．出現頻度は，本稿で用いたパターンが適合した原文の数で，パターンと全原文との間の照合実験により求めた．調査結果を表 4 に示す．

表 4 出現頻度データと包含関係による削減結果の関係

出現頻度	パターン数	削減パターン数
1,000 文以上	275 (0.2%)	195 (70.9%)
100 文以上 1,000 文未満	1,013 (0.8%)	502 (49.6%)
10 文以上 100 文未満	3,194 (2.6%)	983 (30.8%)
2 文以上 10 文未満	13,140 (10.7%)	3,255 (24.8%)
小計	**17,622 (14.4%)**	**4,935 (28.0%)**
1 文以下	104,997 (85.6%)	4,917 (4.7%)
合計	**122,619 (100.0%)**	**9,852 (8.0%)**

表 4 より，出現頻度が大きくなると削減したパターンが占める割合（削減率）が大きくなるが，出現頻度が小さくなると削減率も小さくなる傾向がある．

以下に，具体例を示す．

- 出現頻度が大きいパターンの例

  $P6 \supseteq P7$（出現頻度：$P6$=8,769, $P7$=8,765）

  $P6$: /y </tk $N1$ は> /$tcfk$ ($ND2$ を /cf し | $V2$) て /cf ($V3.kako$|$ND3$ をした)。

  $P7$: /y </tk $N1$ は> /$tcfk$ ($ND2$ を /cf し | $V2$) て /cf $V3.kako$。

- 出現頻度が小さいパターンの例

  $P8 \supseteq P9$（出現頻度：$P8$=2, $P9$=1）

  $P8$: /ytk $N1$ の /f $V2$^$rentai$！ことにも /$tcfk$ 一面の /k $N3$ は /cf ある。

  $P9$: /ytk きみの /f 言う！ことにも /$tcfk$ 一面の /k 真理は /cf ある。

削減率が最も小さい，出現頻度が 1 文以下のパターンの内，包含関係を持たないパターン例を以下に示す．

- /ytk $N1$ が /$tcfk$ $N2$ に /cf 対し /f 不利な！$N3$ を /cf $V4.kako$。
- /y \$1^{/tk $N1$ は } /cf 信じられないと /cf 言う (ように | 様に) \$1 /tk $N2$ を /cf $V3.kako$。

出現頻度が小さいパターンには，非線形要素（字面）が多く含まれる傾向がある．パターン化において，変数への置換によって表現構造全体の意味が変化する要素は，非線形要素として字面のまま記述する．従って，非線形要素を多く含むパターンが，パターン辞書中に占める割合が大きいことより，出現頻度が小さいパターンに特有の表現を持つものが多いと考えられる．

6.2 一部包含関係にあるパターン

パーサは，以下に示すパターン $P10$ はパターン $P2$ に一部包含されると判定した．($Cov_{P2 \supseteq P10} = 0.5$)

$P2$: /ytk $N1$ は /$tcfk$ $N2$ の /f ある！$N3.da$。

$P10$: /y </tk $N1$ は> /$tcfk$ 張合の /f ある！$N2.da$。

パターン $P10$ は，次の 2 パターンに展開される．

$P10a$: /y /$tcfk$ 張合の /f ある！$N2.da$。

$P10b$: /y /tk $N1$ は /$tcfk$ 張合の /f ある！$N2.da$。

展開後のパターン間の包含関係は，$P2 \not\supseteq P10a$, $P2 \supseteq P10b$ となる．従って，$P2 \not\supseteq P10$ と判定される．しかし，パターン $P2$ を次のパターン $P2'$ に示すように変更することで，$P2' \supseteq P10$ となる．

$P2'$: /y </tk $N1$ は> /$tcfk$ $N2$ の /f ある！$N3.da$。

このように，一部包含関係があると判定されたパターンに着目し，新たなパターンを作成することで，削減率の向上を図れる可能性がある．

6.3 下位パターン削除の影響調査

削除した下位パターンの日本語原文と，その上位対訳パターンを用いて，翻訳を行った．翻訳結果より下位パターンを上位パターンで置換可能かの検討を行う．

調査対象として，出現頻度が「1,000 文以上」と「2 文以上 10 文未満」の上位日本語パターンから，それぞれランダムに 10 件ずつ抽出した．また，抽出した上位日本語パターンそれぞれに対応する下位日本語パターンをランダムに 1 件選択し，その日本語原文を用いた．上位日本語パターンと下位日本語パターンの原文との間の照合には，文型パターンパーサを用いた．照合結果と上位英語パターンを用いて，人手による翻訳を行った．訳語選択は最も適切と考えられるものを選択した．

調査結果を表 5 に示す．

表 5 翻訳調査の結果

出現頻度	翻訳成功	翻訳失敗
1,000 文以上	5	5
2 文以上 10 文未満	10	0

また，翻訳例を以下に示す．

- 翻訳例 1

  上位パターン（出現頻度：5）

  日本語パターン:

  /ytk $N1$ が /cf $AJ2$ ば /tk $N3$ は /cf $AJ4$ 。

日本語原文: 財布が重ければ心は軽い。
英語パターン: *AJ2 N1* make *AJ4 N3.*
英語原文: A heavy purse makes a light heart.

下位パターン
日本語パターン: */ytk N1* が */cf*
(無けれ | なけれ) ば */tk* 粉は */cf* ない。
日本語原文: ひき臼がなければ粉はない。
英語パターン: No *N1*, no meal.
英語原文: No mill, no meal.

上位対訳パターンと，下位の日本語原文を用いた翻訳の結果，次の英文が得られた．

```
No mill makes no meal.
```

この英文は，翻訳に成功している．

- 翻訳例 2

上位パターン（出現頻度：5463）
日本語パターン: */y </tk N1* は*> /tcfk N2* を */cf V3* (て | で) */cf* (*V4.kako*| *ND4* をした) 。
日本語原文: 城を取り巻いて攻撃した。
英語パターン:
(*N1*|I) *V3.past N2* and (*V4*|V(*ND4*)).*past.*
英語原文:
They surrounded the castle and attacked.

下位パターン
日本語パターン: */y </tk N1* は*> /tcfk N2* を */cf* 連れて */cf V3.kako* 。
日本語原文: 女を連れて逃げた。
英語パターン: *N1 V3.past* with *N2.*
英語原文: He ran away with a girl.

この例では，次の英文が得られた．

```
I took a girl and ran away.
```

この英文は，翻訳に成功している．

- 翻訳例 3

上位パターン（出現頻度：1895）
日本語パターン:
/y </tk N1 は*> /tcfk N2* を */cf V3* (て | で) *</tk N4* は*> /tcfk N5* を */cf V6.kako*。
日本語原文: 一心を傾注して目的を遂げた。
英語パターン:
<N4|I*> V6.past <N4.pron.poss*|my*> N5* by *V3.ing <N1.poss*|my*> N2* to *N5.pron.*
英語原文: I accomplished my purpose by devoting my whole mind to it.

下位パターン
日本語パターン: */y* $1^{*/tk N1* は} */tcfk* 顔を */cf* しかめて $1 */tk* 私を */cf* 見た。
日本語原文: 彼は顔をしかめて私を見た。
英語パターン: *N1* frowned at me.
英語原文: He frowned at me.

この例では，次の英文が得られた．

```
I looked my I by frowning his face to me.
```

この英文は，重要な情報が間違って訳されている．

表 5 より，上位パターンの出現頻度が小さい場合は，下位パターンの削除を行っても，意味的に正しい翻訳が行える可能性が高いと考えられる．しかし，出現頻度が大きい場合には，下位パターンを削除することで意味的に正しい翻訳が行えなく事例が多く存在することが分かる．今後，より多くの事例について，調査を行う必要があると考えられる．

7 おわりに

本研究では，パターン間の包含関係に着目して，文型パターン数の削減方法を提案した．また，実装を行い大規模文型パターンの削減を行った．

その結果，文法・単語レベルのパターン辞書（122,619 パターン）において，8.0%（9,852 パターン）を削除することが出来た．また，出現頻度が高いパターンは削減率が大きいが，出現頻度が低いパターンは削減率が小さかった．しかし，削減を行うことで，翻訳が失敗する可能性があることも分かった．

今後，一部包含関係にあるパターンに着目し，新たなパターンを作成する手法が考えられる．また，削減が出来なかったパターンについては，縮退自体の必要性の有無も検討していく．さらに，下位パターンを削除する条件についても吟味する必要がある．

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．

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:非線形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] 池原ほか:機械翻訳のための日英文型パターン記述言語, 電子情報通信学会技術研究報告,TL2002-48, pp.1-6, 2003.
[4] 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.
[5] James Allen:Natural Language Understanding(2nd Edition), The Benjamin/Cummings Publishing Company, Inc., pp.101-106, 1994.