日英機械翻訳のための文型パターン辞書の圧縮に関する検討

片山　慶一郎[†]　村　上　仁　一[†]
徳　久　雅　人[†]　池　原　　悟[†]

本稿では日英機械翻訳のための文型パターン辞書を能力を維持したまま人が利用しやすい規模に圧縮する手法を提案する．本手法では，一般的に使用される可能性の低い文型パターンやカバー範囲の重複する文型パターンおよび過剰に汎化された文型パターンなどを削除する．文型パターン辞書の圧縮の結果，122,719 件の文型パターン辞書が，17,973 件 (14.65%) となった．また，クロスバリデーションテストにおける再現率はオリジナルが 67.98%，圧縮後が 61.86%であったことから，問題なく文型パターン辞書の能力はほぼ維持出来ていることが示された．

Study of Reduction for Semantic Pattern Dictionary

Keichiro Katayama,[†] Jin'ichi Murakami,[†] Masato Tokuhisa[†]
and Satoru Ikehara[†]

This paper proposes a method of semantic pattern dictionary reduction with keeping its ability. This method removes following semantic patterns: (1) rarely used (2) fully included by other pattern (3) over trained. The number of original pattern dictionary is 122,719, and the number of reduced dictionary is 17,973(14.65%). And the matched pattern ratio of original is 67.98%, and that of reduced is 61.86%. Comparing these ratio, this method almost keeps the ability of semantic pattern dictionary.

1. はじめに

機械翻訳の方式として，古くから文型パターン翻訳方式が提案されている．この方式では，文型パターンに適合する入力文に対して品質の良い訳文が得られることから，多くの商用システムでも利用されている．しかし，使用される文型パターン数は少なく，特定の狭い分野の翻訳に用いる例が多い．

ところが，最近,「等価的類推思考の原理による機械翻訳方式」が提案され[1)]，この方式の実現のために，22 万件の重文・複文文型パターンを収録した文型パターン辞書が構築されている[2)]．この文型パターン辞書は，15 万件の日英対訳コーパスを元に作成されている．しかし，あまりにも大量のパターン数であるため，人間が文型パターンを精査することは不可能に近い．

そこで，本稿では，一般的に使用される可能性の低い文型パターンやカバー範囲が重複する文型パターンに着目し，能力を維持したまま人が利用しやすい規模に文型パターン辞書を自動的に圧縮する手法を提案する．そして，既に構築されている文型パターン辞書の圧縮を行い，圧縮後の文型パターン辞書の能力を評価する．

† 鳥取大学 工学部
Faculty of Engineering, Tottori University
{kkatayam,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

2. 文型パターン辞書と文型パターンパーサ

2.1 文型パターン辞書

本稿で扱う文型パターン辞書は，辞書などから収集された重文・複文の日英対訳コーパスから，日英の各対訳文に含まれる線形要素を変数に置き換え，各種関数・記号を付与することで作成されている．この辞書には，単語，句，節の 3 つのレベルの文型パターンが収録されている．

本稿では，単語レベル（122,719 件）を用いる．以下に例を示す．

日文：

資本主義と社会主義は相入れない概念だ。

日本語文型パターン（日パ）：

*/ytcfk N*1 と */tcfk N*2 は */cf* 相入れない！概念だ。

英文：

Capitalism is incompatible with socialism.

英語文型パターン（英パ）：

*N*1 be incompatible with *N*2.

2.2 文型パターンパーサ

文型パターンパーサ（以下パーサ）[3)] は，日本語文と日本語文型パターンとの照合を行い，日本語文に適合する日本語文型パターンを出力するプログラムである．照合方式は，ATN(Augmented Transition Network)[4)] をベースとしている．

以下に，パーサが出力した文型パターン例を示す．ここで，

バインド値とは，日本語文が文型パターンに適合するとき，文型パターン中の変数が適合した形態素である．

入力日本語文：

損害を最小限に食い止めることが大事だ。

適合日本語文型パターン：

*/ytcfk N*1 を */tcfk N*2 に */cf* (*V*3ˆ*rentai*|*ND*3 をする) ! ことが */cf AJV*4#5(*.genzai*|*.kako*) 。

バインド値：

*N*1=損害，*N*2=最小限，*V*3=食い止める，*AJV*4=大事だ

英語文型パターン：

It 'is'#5(ˆ*present*|ˆ*past*) *AJ*4 to V(*V*3|*ND*3)ˆ*past N*1 in *N*2.

2.3 評価用試験文集

本稿で用いる評価用試験文集は，日本語独特の言い回しを持つ用例[5] を集めたもので，3,851 文が収録されている．この評価用試験文集の用例は，文型パターン作成に用いた日英対訳コーパスには含まれていない．以下に例を示す．

日文： 熱を出した子どもをつききりで看病した。

英文： She paid constant attention to nursing the feverish child.

3. 適 合 頻 度

適合頻度とは，文型パターンが適合する入力文の数である．本稿では，以下の手順で求める．

(1) 文型パターン辞書作成用の用例と文型パターン辞書を用いて，クロスバリデーションテストを行う．

(2) 文型パターンそれぞれについて，適合した入力文の数を調査する．この値を適合頻度とする．

3.1 本稿におけるクロスバリデーションテスト

本稿では，クロスバリデーションテストの入力文として，文型パターン作成に用いた用例集を用いる．この入力文と文型パターン辞書をパーサを用いて照合する．そして，入力文から作成された文型パターン以外の適合文型パターンを用いる．

3.2 文型パターン辞書の適合頻度

予備調査として，本稿で扱う文型パターン辞書の適合頻度の分布を調査する．その結果を表 1 に示す．

表 1 適合頻度の度数分布表
Table 1 Pattern-Frequency Distribution

適合頻度	文型パターン数／割合	
0	97,346	79.324%
1	8,905	7.256%
2 以上 10 未満	9,124	7.435%
10 以上 100 未満	5,125	4.176%
100 以上 1,000 未満	1,684	1.372%
1,000 以上 10,000 未満	532	0.434%
10,000 以上 100,000 未満	3	0.002%
計	122,719	-

表 1 より，適合頻度 0 の文型パターンが全体の約 8 割を占めていることが分かる．

4. 文型パターン辞書の問題点

4.1 適合頻度の低い文型パターン

文型パターン辞書には，さまざまな表現に対応するため多くの文型パターンが収録されている．しかし，一般的に使用されない表現に対応する文型パターンも多く含まれている．3 章の調査結果より，そのような文型パターンは適合頻度が低い傾向があると言える．以下に例を示す．

日文： 頭をそって丸坊主になった。

日パ： <*N*1 は> (頭 | おつむ) をそって丸坊主になった。

英文： I shaved all my hair off.

英パ： <I|*N*1> shaved all <my|*N*1ˆ*pron*ˆ*poss*> hair off.

4.2 適合頻度が高い文型パターン

過剰に汎化された文型パターンは，単文などにも適合する．単文に適合する文型パターンは，適合頻度が高い傾向がある．このような文型パターンを用いても，品質の高い訳が得られる可能性は低く，辞書の能力に悪影響を与えていると考えられる．

以下に 3 章の調査において適合頻度 10,000 以上の文型パターン 3 件を示す．

- 適合頻度：24,961
  - **日文**： ニッコリ笑って承知した。
  - **日パ**： */y* <*/tk N*1 は> #2[*/cf V*3 て] */ycf* (*V*4.*kako*|*ND*4 をした) 。
  - **英文**： She smiled her consent.
  - **英パ**： <She|*N*1> *V*3ˆ*past* <her|*N*1> N(*V*4|*ND*4).
- 適合頻度：18,246
  - **日文**： 朝、旗を上げ、夕方下ろします。
  - **日パ**： */y* <*/tk N*1 は> #2[*/tcfk TIME*3 、] / #4[*N*5 を */cf N*6 、] #7[*/ytk TIME*8] */f V*9 #10[*.masu*] 。
  - **英文**： We raise the flag in the morning and lower it in the evening.
  - **英パ**： <We|*N*1> #4[V(*N*6) *N*5] #2[in *N*3] and *V*9 *N*5 #7[in *N*8].
- 適合頻度：10,077
  - **日文**： その申し出を蔑むようにはねつけた。
  - **日パ**： */y* <*/tk N*1 は> #2[*/cf GEN*3] */k N*4 を #5[*/cf* (蔑む | 貶む) (ように | 様に)] */yf V*6.*kako*。
  - **英文**： She scornfully rejected the proposal.
  - **英パ**： <She|*N*1> #5[scornfully] *V*6ˆ*past* #2[*AJ*3] *N*4.

4.3 表現のカバー範囲の重複

文型パターン辞書には，カバー範囲が重複する文型パターンがある．以下に例を示す．

- 文型パターンA
  - **日文**： 春は暖かく、秋は涼しい。
  - **日パ**： *TIME*1 は *AJ*2 、 *TIME*3 は *AJ*4 。
- 文型パターンB
  - **日文**： 夜は暗く、昼は明るい。
  - **日パ**： *TIME*1 は *AJ*2 、 昼 は 明るい 。

文型パターンBの「昼」,「明るい」が文型パターンAの時詞変数 *TIME*3，形容詞変数 *AJ*4 にそれぞれ包含されるため，文型パターンBは文型パターンAの下位の文型パターンとなる．下位の文型パターンは，上位の文型パターンのカバー範囲に含まれるので，再現率に影響を与えずに削除可能である．

5. 文型パターン辞書の圧縮方法

5.1 概　　要

本稿で提案する文型パターン辞書の圧縮方法の基本的な考え方を次に示す．

- 文型パターン辞書の能力を維持
- 適合頻度の低い文型パターンは文型パターン辞書の能力に与える影響が小さいため削除
- 適合頻度の高い文型パターンは文型パターン辞書の能力に悪影響があるため削除
- カバー範囲が重複する文型パターンを削除するために，包含関係[6] を利用
- 文型パターン辞書の作成に用いた用例集とは別の用例集を利用することで，表現の多様性を確保

以上の考え方に基づいて，文型パターン辞書の圧縮を行う．具体的な内容は 5.2 節から 5.5 節にかけて述べる．

5.2 手順１：適合頻度０の文型パターンの削除

適合頻度は，約 12 万件という大量の入力文を用いて求めているため，適合頻度 0 の文型パターンは，一般的に使用される可能性が低いと考えられる．そのような文型パターンはクロスバリデーションテストにおいて文型パターンが適合する割合に全く影響がない．そこで，文型パターン辞書 D0 から適合頻度 0 の文型パターンを削除し，文型パターン辞書 D1 を作成する．

5.3 手順２：評価用試験文集の利用した文型パターンの追加

文型パターン辞書作成時に用例として用いなかった評価用試験文集を使用して，表現の多様性を確保する．具体的には，手順 1 で削除した文型パターンのうち，評価用試験文集に適合する文型パターンを文型パターン辞書 D1 に追加し，文型パターン辞書 D2 を作成する．

5.4 手順３：日本語文型パターンの包含関係による圧縮

カバー範囲が重複する文型パターンを削除するために，日本語文型パターン間の包含関係[6] を用いる．包含関係の下位の文型パターンは，上位の文型パターンで代替可能なので，削除しても再現率に影響を与えない．ただし，本研究では簡略化のため日本文型パターンにのみ着目するため，対応する英語文型パターンは考慮しない．本手順では，文型パターン辞書 D2 において包含関係を調査し，下位の文型パターンを削除した文型パターン辞書 D3 を作成する．

5.4.1 文型パターン間の包含関係

文型パターン β に適合する入力文の全てが，文型パターン α にも適合するとき，文型パターン β は文型パターン α に包含されると定義する．また，文型パターン α を上位の文型パターン，文型パターン β を下位の文型パターンと呼ぶ．両者の関係を $\alpha \supseteq \beta$ と表記する．

5.4.2 文型パターン要素間の包含関係

前節の定義に従って，文型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適合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，文型パターン自身が適合可能な入力文の領域を表している事に着目して，文型パターンが別の文型パターンに適合するかどうかを調査することで，包含関係を判定する．

文型パターン間の包含関係を考える場合，文型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[7] に基づいて包含関係を定義する．表 2 に定義した要素間の包含関係の一部を示す．

表 2 文型パターン要素間の包含関係（一部）
Table 2 Inclusive Relations between Pattern Elements

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）

5.5 手順４：過剰に汎化された文型パターンを削除

過剰に汎化された文型パターンは，辞書の能力に悪影響を及ぼすと考えられる．このような文型パターンは適合頻度が高い傾向がみられる．そこで，文型パターン辞書 D3 から，適合頻度 10,000 以上の文型パターンを削除して，圧縮版文型パターン辞書 D4 を作成する．

6. 実 験 結 果

6.1 実 験 条 件

実験条件を以下に示す．

文型パターン辞書:
単語レベル，122,719 件 (Pat12.1.0-imi6.0.0.dat)

文型パターンパーサ: jpp2(version 5.3kc)

関数定義ファイル: version 8.4a

クロスバリデーション用入力文:
文型パターン辞書作成用の用例集（122,719 文）

評価用試験文集:
日本語文型辞典[5] から抽出した重文・複文（3,851 文）

6.2 評 価 方 法

本稿では辞書の能力を，以下に示す再現率を用いて評価する．

6.2.1 再　現　率

再現率（以下 *R*1）は，文型パターン辞書の能力を評価するためのパラメータの１つで，以下の式で求める．

$$\text{再現率 } R1 = \frac{\text{文型パターンが適合した文数}}{\text{入力文数}}$$

6.3 手順毎の結果

6.3.1 適合頻度０の文型パターンの削除結果（手順１）

オリジナルの辞書 D0（122,719 件）から適合頻度 0 の文型パターン（97,346 件）を削除して，辞書 D1（25,373 件）を作成した．

6.3.2 評価用試験文集の利用した文型パターンの追加（手順２）

手順 1 で削除した文型パターン（97,346 件）と評価用試験文集を，パーサを用いて照合した．その結果，出力された文型パターンは 73 件であった．この 73 件を必要であると判断し，辞書 D1（25,373 件）に追加して辞書 D2（25,446 件）を作成した．以下に追加した文型パターン例を示す．

日文： 緑地の少ないところには住みたくない。

日パ： */y </tk N1* は*> /tcfk N2* の */f AJ3ˆrentai /f N4* には */cf V5.tai /yf* (ない | 無い) 。

英文： I do not want to live in a place devoid of greenery.

英パ： <I|*N1*> do not *V5ˆbase.want* in *N4 AJ3* of *N2*.

6.3.3 包含関係による圧縮（手順３）

辞書 D2 の日本語文型パターンの包含関係を調査した．調査結果を文型パターンの適合頻度別に表 3 に示す．

表 3 包含関係判定結果（辞書 D2）
Table 3 Number of Inclusive Relations between Semantic Japanese Patterns(in D2)

適合頻度	パターン数	下位パターン	
0 文	73	7	9.59%
1 文	8,905	1,915	21.50%
2 文以上 10 文未満	9,124	2,598	28.47%
10 文以上 100 文未満	5,125	1,710	33.37%
100 文以上 1,000 文未満	1,684	852	50.59%
1,000 文以上 10,000 文未満	532	388	72.93%
10,000 文以上 100,000 文未満	3	0	0.00%
計	25,446	7,470	29.36%

包含関係の調査結果より，下位の日本語文型パターンが 7,470 件あることが分かった．これらの文型パターンを辞書 D2（25,446 件）から削除して，辞書 D3（17,976 件）を作成した．以下に，発見した包含関係の例を示す．

- 上位パターン
  日文： 先生は学生達を呼び寄せて注意を与えた。
  日パ： */y* $1ˆ{*/tcfk N1* は } */tcfk N2* を $1 */cf V3* (て | で) $1 */ytck N4* を */cf V5.kako* 。
  英文： The teacher called together the students to give them a warning.
  英パ： *N1 V3ˆpast N2* to *V5ˆbase N2ˆpronˆobj N4*.
- 下位パターン (1)
  日文： 子どもたちは手をつないで輪を作った。
  日パ： */y* #1{*/tcfk N1* は,*/tcfk N2* を } */cf* つないで */ytck N3* を */cf V4.kako* 。
  英文： The children joined hands to make a circle.
  英パ： *N1* joined *N2* to *V4ˆbase N3*.
- 下位パターン (2)
  日文： 彼は背を屈めて門をくぐった。
  日パ： */y* $1ˆ{*/tcfk N1* は } */tcfk* 背を */cf* 屈めて $1 */ytck N2* を */cf V3.kako* 。
  英文： He bent over to go under the gate.
  英パ： *N1* bent over to *V3ˆbase N2*.

6.3.4 過剰汎化文型パターン（手順５）

辞書 D3（17,976 件）から，３章で調査した適合頻度 10,000 以上の文型パターン３件を削除して，圧縮版文型パターン辞書（D4，17,973 件）を作成した．

6.4 辞書の圧縮結果

オリジナル辞書 D0，各手順毎の辞書（D1〜D3），および圧縮版文型パターン辞書 (D4) において，再現率を文型パターン辞書作成用の用例 ($R1_{cross}$) と，評価用試験文集 ($R1_{test}$) で求めた．結果を表 4 に示す．

表 4 パターン数と再現率の変化
Table 4 Number of Patterns and Recall

辞書	パターン数	（割合）	再現率	
			$R1_{cross}$	$R1_{test}$
D0	122,719	100.00%	67.98%	43.00%
D1	25,373	20.68%	67.98%	42.59%
D2	25,446	20.74%	67.98%	43.00%
D3	17,976	14.65%	67.64%	42.98%
D4	17,973	14.65%	61.68%	35.34%

結果より，圧縮版文型パターン辞書はオリジナル辞書の 14.65%(17,974/122,719) の規模になった．しかし，$R1_{cross}$，$R1_{test}$ はほとんど変化していないことから，提案手法が正しく適用されていることが確認できる．

しかし，適合頻度の高い，過剰に汎化された文型パターン３件を削除した辞書 D4 で $R1$ が大きく減少した．これより，適合頻度の高い文型パターンのみが適合する入力文が多く存在したことが分かる．

7. 人手による翻訳調査

各段階において文型パターン辞書の意味的な能力の変化を確認するために，前章で作成した各文型パターン辞書を用いて，人手による翻訳の評価実験を行う．

7.1 評 価 対 象

文型パターン作成に用いた例文集からランダムに抽出した 50 文を評価対象として用いる．

7.2 評 価 基 準

人手で翻訳を行った結果を以下の基準で評価する．

A 品質の高い英文
B 重要ではない要素の欠如があるが簡単に修正可能
C 部分的な訳
D 使用不可能
F 英文の生成が不可能

評価の例を以下に示す．

- 評価 A の例
  入力文：

暑いので窓を開けた。

適合日本語文型パターン：

/y #1[/tcfk TIME2 は] /cf AJ3^rentai ので *</yc N4 は>*！*N5* を */cf V6.kako* 。

英語文型パターン：

<I|*N4*> *V6^past N5 #1*[ADV(*TIME2*)] because it was *AJ3*.

翻訳結果：

I opened the window because it was hot.

- 評価 B の例

入力文：

彼の家族はみんな風変わりな連中だった。

適合日本語文型パターン：

/y $1^{*/tcfk N1* は} */cf ADV2* $1 */f AJV3^rentai*！*N4.#da.kako* 。

英語文型パターン：

N1 be^*past ADV2 AJ3 N4*.

翻訳結果：

The family were all queer lot.

「彼の」に相当する部分が欠落しているが，その点を除くと品質の高い英文なので評価 B となる．

- 評価 C の例

入力文：

病気とは全然知らずに彼を訪れた。

適合日本語文型パターン：

/y </tk N1 は*> #2[/tcfk* 容赦！なく*] #3[/ytck N4* の*]*！*N5* を */cf V6.kako* 。

英語文型パターン：

<I|*N1*> *#2*[ruthlessly] *V6^past N5 #3*[of *N4*].

翻訳結果：

I visited him.

前半部分が完全に欠落し，「彼を訪れた」部分のみ翻訳出来ているので評価 C となる．

- 評価 D の例

入力文：

彼女の足音がテラスを横切って反響した。

適合日本語文型パターン：

/y </tk N1 は*> /tcfk N2* を */cf V3* (て | で) *</ycf N4* は*>*！(*V5.kako*|*ND5* をした) 。

英語文型パターン：

<My|*N1^poss*> *N2* ADV(*V3*), <I|*N4*> V(*V5*|*ND5*) ^*past*.

翻訳結果：

My terrace across, I echoed.

入力文の意味を全く表現できていないので評価 D となる．

- 評価 F の例

入力文：

昔はいざしらず、今は会社を１０も持つ大実業家だ。

適合日本語文型パターン：

/y $1^{*/tcfk N1* は} *#2[/tcfk TIME3]* $1 *#1{/tck #4[N5* から*],/tcfk N6* を*}* */cf V7^rentai*！*N8.da* 。

英語文型パターン：

N1 V(*N8*)^*prog* to *V7^base* about *N6 #4*[from *N5*] *#2*[*N3*].

バインド値：

N1=今，*N6*=会社，*V7*=待つ，*N8*=大実業家

日本語文型パターンの名詞変数 *N8* には「大実業家」が適合しているが，英語文型パターンでは品詞変換関数 V(*N8*) が記述されている．そのため，「大実業家」を動詞として訳出しなければならないが，適切な動詞が見つからず翻訳をすることが出来ないので評価 F となる．

7.3 正 解 率

英文の評価結果をもとに，２通りの正解率 $(P1, P2)$ を以下の式で求める．正解パターンは評価が A または B の場合とする．

$$\text{正解率 } P1 = \frac{\text{正解文型パターン数}}{\text{適合した全文型パターン数}}$$

$$\text{正解率 } P2 = \frac{\text{正解文型パターンを持つ文数}}{\text{文型パターンが適合した文数}}$$

7.4 評 価 結 果

文型パターン辞書 D0〜D4 における再現率，正解率の値を表 5 に示す．

表 5 翻訳調査結果（50 文）
Table 5 Result of Experiment(Recall and Precision, 50 sentences)

辞書	パターン数	再現率	正解率	
		R1	*P1*	*P2*
D0	122,719	70%	23% (185/797)	49% (17/35)
D1	25,373	70%	23% (185/797)	49% (17/35)
D2	25,446	70%	23% (185/797)	49% (17/35)
D3	17,976	70%	18% (55/308)	46% (16/35)
D4	17,973	68%	19% (55/290)	47% (16/34)

表 5 の結果より，文型パターン数の削減幅に対して，正解率および再現率の変化は小さいことが分かる．また，辞書 D3 は辞書 D2 と比べ $P1$，$P2$ ともに低下しているが，辞書 D4 は辞書 D3 と比べ $P1$，$P2$ ともに向上している．

また，人手による各段階の評価結果を表 6 に示す．

表 6 辞書毎の翻訳評価結果（50 文）
Table 6 Translation Quality of Each Dictionary(50 sentences)

辞書＼評価	A	B	C	D	F
D0	12% (94/797)	11% (91/797)	23% (187/797)	45% (361/797)	8% (64/797)
D1	12% (94/797)	11% (91/797)	23% (187/797)	45% (361/797)	8% (64/797)
D2	12% (94/797)	11% (91/797)	23% (187/797)	45% (361/797)	8% (64/797)
D3	8% (24/308)	10% (31/308)	27% (83/308)	44% (136/308)	11% (34/308)
D4	8% (24/290)	11% (31/290)	28% (81/290)	45% (130/290)	8% (24/290)

表 6 の結果より，文型パターン数の削減幅に対して，評価 A の割合は減少したが，評価 B，C および D の割合は変化していないことが分かる．ただし，辞書 D3 において，評価

A の割合が減少している．また，辞書 D4 において，評価 C, D および F の数のみが減少していることが分かる．

8. 考　　察

8.1 包含関係による削除が与える影響

表 5 の結果より，辞書 D3 は辞書 D2 と比べて正解率 $P1$ が大きく低下した．これは，日本語文型パターンにのみ着目し，英語文型パターンを考慮せずに圧縮を行ったためと考えられる．

しかし，正解率 $P2$ は約 3%（1 文）低下するにとどまっている．文型パターン翻訳においては，正解パターンが存在することが重要なので，日本語文型パターンの包含関係による圧縮を行っても，文型パターン辞書の能力には影響が少ないと考えている．

8.2 過剰汎化文型パターンの与える影響

過剰に汎化されたと考えられる文型パターン（3 件）が与える影響を考察するために辞書 D4 と辞書 D3 を比較すると，再現率 ($R1_{cross}$，$R1_{test}$) が大きく減少したことが分かる．つまり，この 3 件の文型パターンのみが適合する入力文が多く存在することが言える．しかし，表 5 の結果から，意味的な被覆率 ($R1 \times P2$) を考えると，辞書 D3，D4 共に同じ 32%となる．さらに，正解率 $P1$，$P2$ の両方ともが向上している．したがって，過剰に汎化された文型パターンを削除することで，再現率は低下するが，意味的な能力が変化しなかったため文型パターン辞書としての能力は向上したと考えている．

そこで，適合頻度上位 10 件の文型パターンを調査した．その結果，過剰に汎化された文型パターンを，本研究で削除した 3 件に加え，新たに 4 件発見した．以下に例を示す．

- 適合頻度： 9,245
  日文： すばらしく見事な身のこなしでステージを移動した。
  日パ： */y </tk N1 は> #2[/cf すばらしく /f 見事な /f 身の /k こなしで] /tcfk N3 を /cf V4.kako* 。
  英文： He moved across the stage with marvelous deftness.
  英パ： <He|*N1*> *V4^past* across *N3* #2[with marvelous deftness].
- 適合頻度： 7,067
  日文： 聴衆は思ったように反応した。
  日パ： */y* $1^{/*tcfk N1* は} #2[/*cf* (思っ| おもっ| 惟っ| 意っ| 憶っ| 懐っ| 想っ| 念っ) た (ように | 様に)] $1 /*yf* (*V3.kako*|*ND3* をした) 。
  英文： The audience reacted predictably.
  英パ： *N1* V(*V3*|*ND3*)^*past* #2[predictably].
- 適合頻度： 7,067
  日文： 彼女は涙ぐみながらほほえんだ。
  日パ： */y* $1^{/*tcfk N1* は} #2[/*cf V3* ながら] $1 /*yf V4.kako* 。
  英文： She smiled in a mist of tears.
  英パ： *N1 V4^past* #2[in a mist of N(*V3*)].
- 適合頻度： 7,062
  日文： 街路に群がり集まった。
  日パ： */y </tk N1* は> */tcfk N2* に #3[/*cf* (群がり | 叢り | 簇り | 羣がり)] */yf V4.kako* 。
  英文： They swarmed the streets.
  英パ： <They|*N1*> *V4^past N2*.

したがって，今後さらに過剰に汎化された文型パターンを削除することで，意味的に不適切な文型パターンへの適合が減少し，辞書の能力が向上すると考えられる．

9. おわりに

本稿では，文型パターン辞書の能力を維持した上で圧縮する方法を検討し，12 万件の重文・複文文型パターン辞書の圧縮を行った．その結果，17,973 件 (14.65%) に圧縮することが出来た．また，再現率・正解率ともに文型パターン数の削減量に対してわずかな低下となっていることから，能力は維持できている．

しかし，日本語文型パターンのみの包含関係を用いると正解率が低下するため，英語文型パターンについても考慮する必要がある．今後は，再現率を維持したまま，正解率も維持する辞書の削減方法を考えていく．そのために，日英の文型パターンが共に包含関係にあるものに着目する方法がある．

謝　　辞

本研究は，独立行政法人科学技術振興機構 (JST)・戦略的創造研究推進事業 (CREST) の研究領域「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) 池原悟: 究極の翻訳方式の実現に向けて＝＝類推思考の原理に基づく翻訳方式＝＝, AAMT Journal, アジア太平洋機械翻訳協会, No.33, pp.1-7, 2002.
2) 池原悟, 阿部さつき, 徳久雅人, 村上仁一: 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, 自然言語処理, Vol.11, No.3, pp. 69–95, 2004.
3) 徳久雅人, 村上仁一, 池原悟: 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.
4) James Allen: *Natural Language Understanding(2nd Edition)*, The Benjamin/Cummings Publishing Company, Inc., pp. 101–106, 1994.
5) 砂川有里子, 駒田聡, 下田美津子, 鈴木睦, 筒井佐代, 蓮沼昭子, ベケシュ・アンドレイ, 森本順子: 日本語文型辞典, くろしお出版, 1998.
6) 片山慶一郎, 村上仁一, 徳久雅人, 池原悟: 日本語文型パターンの圧縮方法, 言語処理学会 第 12 回年次大会発表論文集, pp. 568–571, 2006.
7) 池原悟, 宮崎正弘, 佐良木昌, 池田尚志, 白井諭, 村上仁一, 徳久雅人: 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.