

# 文型パターンによる日英翻訳のための名詞句パターン辞書の構築

神野 絵理 徳久 雅人 村上 仁一 池原 悟

鳥取大学工学部知能情報工学科

{jinno,tokuhisa,murakami,ikehara}@ike.tottori-u.ac.jp

## 1 はじめに

重文・複文の日英機械翻訳に文型パターンを用いる手法が提案されている [1]. その手法では, パターンの記述要素に名詞句変数が使われており, その変数に代入された日本語表現を英訳する必要がある.

そこで, 本稿では, 名詞句翻訳においてもパターン翻訳による実現を目的とする. 具体的には, 大規模名詞句コーパスより, 名詞句パターンを自動生成し, 名詞句翻訳プロトタイプシステム (Meijin) を用いて, 名詞句パターン辞書の性能評価を行う.

## 2 名詞句パターン化の方法

### 2.1 名詞句の日英対訳コーパス

[1] では, 15 万文対の日英対訳コーパスから文型パターンを作成した. その作成過程では, 対応関係の見い出された名詞句が約 4.5 万対存在する. 本稿では, この名詞句対から名詞句パターン対を作成する.

表 1: 名詞句対訳コーパスの一部

日本語	英語
彼のお母さん	his mother
あの建物	that bilding
あの男	that man
古い民謡の一つ	one of the old folk songs
息子の話	son's story

### 2.2 パターン化の手順

パターン化には, 単語アライメントによる対応要素の変数化, 変数への意味属性制約の付与, 形態素調整用タグの付与の大きく 3 つの手順がある.

単語アライメントでは, ALT-JAWS, および, Brill パーサ [2] を用いて日英形態素解析を行い, 和英辞書を利用することにより単語が対応した箇所を表 2 に従って変数化する. その中で, 一般の名詞 ( $N$ ) と用言性名詞 ( $NS$ ) に対し意味属性 [3] を付ける. また, 形態素調

整用タグとして, 英単語が所有格である場合 “ $\sim$ poss” を, 複数である場合, “ $\sim$ pl” を付与する.

次にパターン化の流れを示す.

元の句

日本語名詞句: 彼のお母さん

英語名詞句: his mother

単語アライメント

彼  $\longleftrightarrow$  his お母さん  $\longleftrightarrow$  mother

変数化

彼  $\rightarrow PRN$  お母さん  $\rightarrow N$

his  $\rightarrow PRN$  mother  $\rightarrow N$

意味属性の付与と形態素調整

変数化する際, “お母さん” は一般名詞 ( $N$ ) なので, 意味属性を付与する. また, “his” は, “he” の所有格となっているので, “his” の変数 “ $PRN$ ” に “ $\sim$ poss” を付与する.

パターン

日本語パターン:  $PRN1$  の  $N2$ (男女, 親)

英語側パターン:  $PRN1^{\sim$ poss  $N2$

() の中には意味属性を意味を表す.

表 2: 自立語の変数化

単語 の品詞	品詞変数	
	日本語	英語
用言性名詞	$NS$	無し
数詞	$NUM$	$NUM$
代名詞	$PRN$	$PRN$
一般の名詞	$N$	$N$
動詞	無し	$V$
形容詞	$AJ$	$AJ$
形容動詞	$AJV$	無し
副詞	$ADV$	$ADV$
連体詞	$REN$	無し
数助詞	$UNIT^*$	無し

(\* $UNIT$  は, 辞典 [4] に収録されている語を対象とする)

## 2.3 名詞句パターン化の結果

名詞句コーパスから変数化できた句は、36,729 対、字面の句は、8,947 対であった。後者は字面パターンとみなす。

## 3 名詞句パターン辞書の作成

日英名詞句パターン対において、同じ記述のパターン対を 1 つにまとめて、パターン辞書とする。日本語パターンは、字面パターンを含め、全部で 23,834 種類あった。日本語名詞句の圧縮率は、52%であった。

パターンを作るために用いたコーパスの名詞句の分布を調べたところ、パターン化の元の名詞句が一番多く使われていたパターンが  $REN1N2$  であり、その名詞句の数は 3,719 個であった。以下、上位 10 位までの日本語パターンと、それに対する英語パターンを表 3 に示す。

コーパスの名詞句が 1,000 個以上であったパターンが 4 件、999 ~ 100 個であったパターンが 14 件、99 ~ 20 個であったパターンが 56 件、29 ~ 1 個であったパターンが 23,735 件であった。

## 4 Meijin の翻訳手順

手順を以下に示す。

1. 入力日本語名詞句と日本語パターンをパターンパ - サを用いて照合する [5]。
2. 照合結果より、適合した日本語パターンを抽出する。
3. 抽出した日本語パターンに対応する英語パターンを名詞句パターン辞書から検索する。
4. 抽出された英語パターンの変数部に対応する英単語を代入し、出力する。

以下に具体例を示す。

- 入力句：この新聞
- 模範訳：this newspaper
- 日本語パターン： $REN1N2$ (本)  
この  $\rightarrow REN1$   
新聞  $\rightarrow N2$ (本)
- 英語パターン： $PRN1 N2$   
ここで、変数部に訳語が代入される。  
 $PRN1 \rightarrow this$   
 $N2 \rightarrow newspaper$
- 出力：this newspaper

## 5 翻訳実験

### 5.1 実験の目的

作成した名詞句パターン辞書の性能を評価することを目的とする。具体的には、次の 2 つの実験を行う。

(実験 1) 既存の翻訳機 2 種類(以下、システム 1、システム 2 と称す)の翻訳精度と Meijin の翻訳精度を比較する。

(実験 2) Meijin で訳出の無い名詞句を、システム 1、または、システム 2 で訳出し直すという「2 段階翻訳」を行い、それぞれの総合の性能を評価する。

### 5.2 実験対象

実験の入力データは、3 章で述べた日本語名詞句をランダムに選んだ 100 件を対象とする。

Meijin においては、入力された名詞句から作られるパターンは、照合に用いないこととする。4 章で述べた翻訳手順に従って訳出する。ただし、Meijin は、複数の訳出があるが、名詞句パターン辞書の性能を調べることがねらいなので、その選択は人手で行うこととする。

システム 1、システム 2 については、Meijin と同様の名詞句を入力し訳出する。

### 5.3 評価方法

評価基準は、以下の通りとする。

... 訳出された英語が、文法的に正しく、意味も理解できる場合(英語の訳語、冠詞、句の外の情報は考慮しない)

... 訳出された英語が、文法的に間違っているが、意味は理解できる場合

×... 訳出された英語が、意味的に間違っている、または、訳出が無い場合

以上の評価を、再現率  $R$ 、および、適合率  $P$  を用いて集計する。

$$\text{再現率 } R = \frac{\text{出力パターンが一つ以上ある回答の数}}{\text{出題数}}$$

$$\text{適合率 } P = \frac{\text{評価 ( ) のある回答数}}{\text{出力パターンが一つ以上ある回答数}}$$

以下に評価の例を示す。

評価 の例

(入力句)別の機会

表 3: 日本語パターンに対する英語パターンの種類 (日本語の上位 10 位まで)

日本語パターン 句の数	英語パターン			
	1 位	2 位	3 位	その他 [種類数]
<i>REN1N2</i> 3,719 個	<i>PRN1 N2</i> (87.0%)	<i>AJ1 N2</i> (5.6%)	the <i>AJ1 N2</i> (1.7%)	その他 [101] (5.7%)
その <i>N1</i> 3,686 個	the <i>N1</i> (97.2%)	his <i>N1</i> (0.4%)	this <i>N1</i> (0.2%)	その他 [41] (2.2%)
<i>PRN1 の N2</i> 1,936 個	<i>PRN1 N2</i> (97.2%)	<i>PRN1 true N2</i> (0.2%)	<i>N2 of PRN1</i> (0.1%)	その他 [38] (2.5%)
<i>N1 の N2</i> 1,224 個	the <i>N2 of the N1</i> (12.5%)	<i>N1 N2</i> (11.8%)	the <i>N1 N2</i> (10.1%)	その他 [186] (65.6%)
この <i>N1</i> 719 個	the <i>N1</i> (95.3%)	<i>N1</i> (0.7%)	those <i>N1</i> (0.4%)	その他 [17] (3.6%)
<i>PRN1 の NS2</i> 661 個	<i>PRN1 N2</i> (99.2%)	<i>PRN1 own N2</i> (0.2%)	<i>N2 of PRN1</i> (0.2%)	その他 [6] (0.4%)
<i>AJ1N2</i> 524 個	<i>AJ1N2</i> (46.1%)	a <i>AJ1 N2</i> (34.9%)	the <i>AJ1 N2</i> (8.0%)	その他 [35] (11.0%)
その <i>NS1</i> 496 個	the <i>N1</i> (97.3%)	their <i>N1</i> (0.4%)	my <i>N1</i> (0.4%)	その他 [8] (1.9%)
<i>REN1NS2</i> 461 個	<i>PRN1 N2</i> (76.4%)	<i>AJ1 N2</i> (10.6%)	the <i>AJ1 N2</i> (2.0%)	その他 [21] (11.0%)
<i>AJV1N2</i> 381 個	<i>AJ1N2</i> (45.1%)	a <i>AJ1 N2</i> (25.7%)	the <i>AJ1 N2</i> (12.1%)	その他 [34] (10.8%)

(解答例) another opportunity

(出力句) a different opportunity

(理由) 訳出された句 “a different opportunity” は、  
文法的にも意味的にも正しいので評価 となる。

#### 評価 の例

(入力句) 新幹線の旅

(解答例) The trip by Shinkansen

(出力句) the trip of a sinkansen

(理由) 訳出された句 “the trip of the Sinkansen” は、  
“of” が誤りであるために、評価は となる。

#### 評価 × の例

(入力句) あの俳優

(解答例) that actor

(出力句) that sumo actor

(理由) 訳出された句 “that sumo actor” は、意味が  
明らかに異なるので、評価 × となる。

## 5.4 実験結果

### 5.4.1 実験 1

評価結果を表 4 に示す。Meijin では、25 個が訳出で  
きなかった。また、入力句 1 個に対し、出力句は、平  
均で 7~8 件であった。

再現率と適合率を表 5 に示す。Meijin の再現率は、  
低いが、適合率は他より高かった。Meijin の再現率が  
低かった理由は、今回作成したパターンの辞書の作成に  
用いた標本が、[1] の重文・複文から抽出した名詞句の  
みであったためと考えられる。

表 4: 評価結果

	評価	評価	評価 ×
Meijin	74%(74 個)	1%(1 個)	25%(25 個)
システム 1	87%(87 個)	12%(12 個)	1%(1 個)
システム 2	94%(94 個)	5%(5 個)	1%(1 個)

表 5: 各翻訳機の性能

	再現率 <i>R</i>	適合率 <i>P</i>
Meijin	75%	98.7%
システム 1	100%	87%
システム 2	100%	94%

### 5.4.2 実験 2

実験 1 で訳出の無かった 25 個の名詞句について、シ  
ステム 1、システム 2 で 2 段階翻訳を行った結果を表  
6、および、表 7 に示す。この結果から、総合性能の向  
上がみられた。

表 6: Meujin で訳出の無かった名詞句の結果

	評価	評価	評価 ×
システム 1	68%(17 個)	28%(7 個)	4%(1 個)
システム 2	92%(23 個)	8%(2 個)	0%(0 個)

表 7: 2 段階翻訳の性能

	再現率 $R$	適合率 $P$
Meijin とシステム 1	100%	91%
Meijin とシステム 2	100%	97%

## 6 考察

実験 1 で Meijin の訳出が無かった名詞句についての考察を 6.1 節に、実験 1 および実験 2 において評価が、または、×となった名詞句について 6.2 節で考察する。

### 6.1 実験 1 の考察

実験 1 で Meijin の訳出が無かった名詞句の一部を以下に示す。

- このテ - ブルの位置
- 前科のある男
- 見え透いたうそ
- 世界中の少年たちの伝統的な夢
- 高い地位及び名声への道

この原因を以下に示す。

1. 名詞句パタ - ン辞書に日本語パタ - ン自体が存在しない場合

(入力句 1) 高い地位および名声への道

この場合、名詞句パターン辞書の標本を増やすことにより解決できると考えられる。なお、入力句 1 に、類似する日本語パターンはなかった。

2. 日本語パタ - ン、および、英語パタ - ンが存在するが、名詞意味属性制約で一致しない場合

(入力句 2) 湖の表面

(正解例) the surface of the lake

この入力句 2 に一番近いパターンは次の例である。

(日本語) N1(その他, 池) の N2(面, 表)

(英語) the N2 of the N1

この場合、入力句の“湖”とパターンの意味属性の“池”，および、入力句の“表面”とパターンの意味属性の“面, 表”は、単語の意味属性の距離が近い。そこで、名詞の汎化を考慮することによって、パターンを適合できると考えられる。

### 6.2 パタ - ン化の問題

以下にパタ - ン化の誤り例を示す。

<元の句>

(日本語) あの力士

(英語) That sumo wrestler

<パタ - ン>

(日本語) REN1N2(競技者, 男)

(英語) PRN1 sumo N2

本来、“力士 = sumo wrestler”となる箇所が、今回単語アライメントを行ったことで“力士 = wrestler”となっていた。この問題を解決するためには、日本語名詞に対し、英語の単語をどこまで含むのかを検討しなければならない。また、英語パタ - ンにおいて、字面で残っている他のパタ - ンに対しても同様である可能性があるため、見直す必要がある。

## 7 おわりに

本稿は、大規模名詞句コーパスより、名詞句パタ - ンを自動生成し、名詞句翻訳プロトタイプシステム (Meijin) を用いて、名詞句パタ - ン辞書の性能評価を行った。この結果は、再現率は 75%、適合率は 98.7%であった。2 段階翻訳を行うと、再現率は 100%、適合率が 91~97%となり、総合的に精度を高めることができた。このことから、作成した名詞句パタ - ン辞書の有効性が確認できた。

今後の課題は、意味属性の汎化や名詞句パターンのアライメントの精密化、および、標本を増やすことによる新たなル - ルの作成である。

## 参考文献

- [1] 池原悟, 阿部さつき, 徳久雅人, 村上仁一: 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, 自然言語処理, Vol.11, No.3, pp.69-95, 2004.
- [2] 飯田朝子, 町田健: 数え方の辞典, 小学館, 2004.
- [3] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦: 日本語語彙大系, 岩波書店, 1997.
- [4] Brill, E.: A simple rule-based part-of-speech tagger, ANLP-92, pp.152-155, 1992.
- [5] 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 回年次研究会, pp.608-611, 2004.