文型パターンにおける任意要素の記述方法とその効果

遠藤 久美子　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
{endo,tokuhisa,murakami,ikehara}@ike.tottori-u.ac.jp

1　はじめに

言語表現には，言語の意味が表現構造と独立に扱うことができないという非線形の問題がある．等価的類推思考の原理に基づく日英機械翻訳 [1] では，翻訳対象となる両言語の表現を「文型パターン」の対としておくことで，意味の失われない解析・生成を実現しようとしている．文型パターンは，被覆率を向上させるため，様々な改良が施されてきた．任意要素指定機能は，その１つである．

本稿では，任意要素指定機能による被覆率の向上の効果を評価することを目的とする．そこでまず，文型パターンの被覆率の向上を推定するパラメータ η（文型パターン拡大率）を提案する．そして，現在使用されているパターン辞書（基準辞書）から任意要素指定機能を除くことで，η の低下の具合を測り，その機能の効果を評価する．

2　文型パターンと任意要素

2.1　文型パターン辞書の概要

文型パターンとは，日英対訳標本文を，変数化，関数化，任意化したものである．

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要素を変数化している．文法・単語レベルパターンの例を以下に示す．

- **日本語文：** ここから目黒へ行く間にとても静かな自然教育園があります。
- **日本語文型パターン：**/ytk ここから/tcfk$N1$ へ/cf 行く間に#2[/cf$ADV3$]/f$AJV4$!$N5$ が/cf あります。
- **英文：**There is a very quiet nature study park between Meguro and here.
- **英語文型パターン：**There is #2[$ADV3$] $AJ4$ $N5$ between $N1$ and here.

日英対訳標本文 (122,316 文) 全ての文型パターンを収録したものを文型パターン辞書という．

2.2　任意要素指定機能

文型パターンで使用される任意要素は,「原文任意要素」,「文型任意要素」に分けられる．以下にそれぞれの任意要素を示す．

2.2.1　原文任意要素

文型パターン記述において，原文任意要素は，文型パターン以外の要素の挿入位置を表すものであり，離散記号（/小英字）で指定される．該当する位置に現れた入力文の要素は，別途翻訳し，英文に組み込まなければならない．原文任意要素として認める要素は，(1) 連用節，(2) 連体節，(3) 格要素，(4) 連用修飾要素，(5) 連体修飾要素の５種類とされている．原文文型任意要素として認める要素と挿入箇所を以下に示す．

- **連用節（/y）**
  節と節の間に連用節の挿入を認める．同一の節内には挿入を認めない．離散記号 /y で示される．
- **連体節（/t）**
  名詞句の直前に連体節の挿入を認める ．同一の名詞句内には挿入を認めない．離散記号 /t で示される．
- **格要素（/c）**
  文型パターン内に存在する格要素の前後に別の格要素の挿入を認める．但し，同一格要素内に別の格要素が挿入されてはならない．また，格要素のないところに格要素を挿入してはならない．離散記号 /c で示される．
- **連用修飾要素（/f）**
  挿入可能な連用修飾要素は，形容詞連用形，副詞，副詞句のいずれかとする．挿入可能な位置は，文頭，述部の直前，形容詞の直前，動詞の直前とする．離散記号 /f で示される．
- **連体修飾要素（/k）**
  挿入可能な連体修飾要素は，連体詞,「A の」・「A と」などの名詞句構成要素，形容詞連体形，動詞連体形のいずれかとし，文型パターン内の名詞句直前への挿入を認める．同一名詞句内への挿入は認めない． 離散記号 /k で示される．

2.2.2　文型任意要素

文型任意要素は，文型パターン要素のうち，省略可能な要素を示す．それが削除されても英語文型自体は変化しないが，それ自身の訳語選択の決定や訳語挿入位置の決定が困難であるなど，要素自身の翻訳に困難さが生じる要素が，文型任意要素として，任意要素記号（[]）で指定されている．文型任意要素は，英文中に訳出すべき位置情報が指定されているため，該当する部分の翻訳を組み込むことは容易である．

2.3　文型パターンの照合

文型パターンパーサは，入力文と適合する文型パターンと，適合の仕方をすべて出力するプログラムである [2].

入力文の翻訳に使用できる文型パターンは，必ずしも入力文のすべての要素が適合する文型パターンであ

る必要はなく，入力文の主要な構造が適合し，意味的に正しいパターンであればよい．適合文型は，以下の 2 種類に分類できる．

- **完全一致文型:** 入力文のすべての要素が文型パターンの要素と適合する文型パターン
- **部分一致文型：** 入力文の一部の要素が原文任意要素であり，離散記号と適合する文型パターン．

以下に入力文に適合した「完全一致文型」と「部分一致文型」の例を示す．

- **入力文：** 彼は頭がいい上に、勉強家である。
- **完全一致するパターン：** /y/tk$N1$ は/cfk$N2$ が/cf$AJ3$^rentai!上に、/tk$N4$.da。
  - $N1$ = 彼
  - $N2$ = 頭
  - $AJ3$ = いい
  - $N4$ = 勉強家
- **部分一致するパターン：** /y#1[/tk　非常に/cf$AJV2$^rentai]!$N3$ は/tcfk$N4$.da。
  （原文任意要素の対応先を<< >>の記号で表す.）
  - $N3$ = 彼
  - /tcfk = <<頭がいい上に>>
  - $N4$ = 勉強家

3　評価パラメータ

3.1　被覆率の評価パラメータ

[3] では，被覆率を評価するパラメータとして,「文型再現率 $R1$」,「平均適合パターン数 N」などを提案している．

3.1.1　文型再現率 $R1$

入力文に対して適合文型パターンが存在するかどうかを文単位で集計したもので，下式で定義される．

$$\text{文型再現率 } R1 = M/I$$

（但し，I:テスト用入力文の数 M:「自己パターン」以外の適合文型パターンが 1 つ以上存在した入力文の数）

3.1.2　平均適合パターン数 N

平均適合パターン数は，入力文に対して，完全一致あるいは部分一致する文型パターン数の平均値を表し，下式で定義される．

「完全一致文型数」の平均

$N1$ = 完全一致文型パターン数 / 入力文の数

「部分一致文型数」の平均

$N2$ = 部分一致文型パターン数 / 入力文の数

「一致文型数」の平均

$N = N1 + N2$

3.2　被覆率向上の効果の推定パラメータ

本稿では，被覆率の向上を推定するパラメータとして，文型パターン拡大率 η を提案する．η は,「評価対象の文型パターン辞書（対象辞書）が，基準文型パターン辞書（基準辞書）の文型パターン数に換算して，何倍に相当するか」を表し，以下の式で定義する．

$$\text{文型パターン拡大率} \eta = X/B \tag{1}$$

（但し，B:基準辞書の文型パターン数 ，X:対象辞書の文型パターン数の換算値）

ここで，対象辞書の文型パターン数の換算値 X の換算方法が問題となる．被覆率の評価パラメータに基づき換算値を求めることにすると,「対象辞書の被覆率に至るには，基準辞書のパターン数をどれだけにすれば良いか」という換算ができる．したがって，基準辞書の被覆率を文型パターン数 p の関数 $c(p)$ で表しておき，対象辞書の被覆率を r として，$r = c(X')$ を満たす X' を η を求める際に用いる．さらに，一般に関数 $c(p)$ は近似曲線であり，誤差を含む．そこで，基準辞書の η を 1 にするために，η の計算において，実際の値である B を用いるのではなく，基準辞書の被覆率を R として，$R = c(B')$ を満たす B' を用いることにする．

したがって，近似曲線 c で換算した文型パターン拡大率 η_c は以下の式で求める．

$$\eta_c = c^{-1}(r)/c^{-1}(R) \tag{2}$$

（但し，R:基準辞書の被覆率，r:対象辞書の被覆率，c^{-1}:近似曲線の逆関数であり，被覆率に対する文型パターン数）

本稿では，対象辞書の文型パターン数の換算値 X は次の 2 つの方法で求める．

(1) 文型再現率 $R1$ による文型パターン拡大率 η_{R1} の求め方

基準辞書（本稿では，文法・単語レベルパターン辞書（122,619 パターン収録）を使用する）を用いた実験から，文型再現率 $R1$ と文型パターン数 p の関係を，図 1 で示す．図中の縦軸は，文型再現率 $R1$ を示し，横軸は文型パターン数 p を示す．サンプル点（×）は，文型パターン数と $R1$ の実測値を示す．

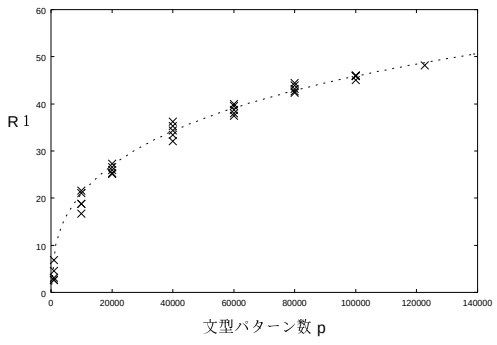

図 1: R1 と文型パターン数の関係図

非線型回帰分析より，文型再現率 $R1$ の文型パターン数 p に対する特性を，(3) 式で近似する．近似曲線を図中に点線で示す．

$$R1 = (1 - exp(-\lambda_1 p^{\lambda_2})) \times 100 \qquad (\%) \tag{3}$$

（但し，非線形回帰分析より，$\lambda_1 = 0.005038$，$\lambda_2 = 0.4171$）

実験で得られた被覆率を近似式に代入することによって得られる p を，対象辞書の文型パターン数の換算値 X とする．そして，(1) 式より，文型再現率 $R1$ による文型パターン拡大率 η_{R1} が求まる．

(2) 平均適合パターン数 N による文型パターン拡大率 η_N の求め方

同じく基準辞書を用いた実験から，平均適合パターン数 N と文型パターン数 p の関係を，図 2 に示す．図中の縦軸は，平均適合パターン数 N を示し，横軸は文型パターン数 p を示す．サンプル点（×）は，文型パターン数と N の実測値である．

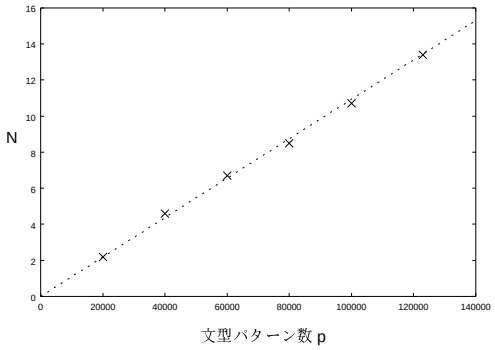

図 2: N と文型パターン数の関係図

線形回帰分析より，文型再現率 N の文型パターン数 p に対する特性を，(4) 式で近似する．近似線を図中に点線で示す．

$$N = \lambda_3 p \tag{4}$$

（但し,$\lambda_3 = 13.4085/122,619$）

同様にして，(1) 式より，平均適合パターン数 N による文型パターン拡大率 η_N が求まる．

4 任意要素の効果の調査

4.1 調査対象と目的

本調査では，2.2 節で述べた,「原文任意要素 5 種類, および，文型任意要素」の有無と被覆率の関係について調査する．

そこで，4.2 節の方法で，基準辞書から，それぞれの任意要素指定機能を除き，7 種類の対象辞書を作成する．そして，これらを調査対象とする．

4.2 対象辞書作成方法

原文任意要素，および，文型任意要素のそれぞれの効果を調査するために，基準辞書から各要素を除いた対象辞書を作成する．

例：基準パターンから，(1)「原文任意要素・連体節 (/t)」と (2)「文型任意要素」を削除した文型パターン

- **日本語文：**こんなに客が少なくては商売上がったりだ。
- **基準パターン：**/y#1[こんなに]/tk 客が/cf$AJ2$(てⅠで) は/tcfk$N3$ 上がったりだ。
- **(1) を削除：**/y#1[こんなに]/k 客が/cf$AJ2$(てⅠで) は/cfk $N3$ 上がったりだ。
- **(2) を削除：**/y こんなに/tk 客が/cf$AJ2$(てⅠで) は/tcfk$N3$ 上がったりだ。

このような方法で，任意要素別に，合計 7 種類の対象辞書を作成する．

4.3 調査方法

文型パターンパーサを用いて，テスト用入力文と 7 種類の対象辞書，および，基準辞書と照合する．その結果から，各対象辞書の被覆率を評価し，そして，被覆率向上を推定する．

テスト用入力文としては，文型パターンの作成に使用された対訳標本の日本文（123,016 文）を使用する．

4.4 調査結果

照合結果より，各評価パラメータの値を表 1 にまとめる．

表 1: 調査結果

使用するパターン辞書	$R1$ による推定		N による推定	
	$R1$(%)	η_{R1}	N	η_N
基準辞書	48.2	1.0	13.4	1.0
(i) 原文任意要素すべてなし	17.0	0.05	1.9	0.14
(ii) 連用節（/y）なし	42.1	0.65	10.9	0.81
(iii) 連体節（/t）なし	47.6	0.96	13.1	0.98
(iv) 格要素（/c）なし	39.2	0.51	6.8	0.50
(vi) 連用修飾要素（/f）なし	44.6	0.77	11.9	0.89
(v) 連体修飾要素（/k）なし	45.6	0.83	11.1	0.83
(vii) 文型任意要素すべてなし	35.6	0.39	9.7	0.72

(i) の文型再現率 $R1$ による文型パターン拡大率 η_{R1} および平均適合パターン数 N による文型パターン拡大率 η_N より，原文任意要素を文型パターンの記述に用いると，$R1$ から推定すると，基準辞書を約 20 倍（$1.0/0.05 = 20$）拡大したことに相当し，N から推定すると，約 7 倍（$1.0/0.14 \approx 7$）に拡大したことに相当する．同様に，(vii) の η_{R1}，η_N より，文型任意要素を文型パターンの記述に用いると，$R1$ から推定すると，基準辞書を約 2.6 倍拡大したことに相当し，N から推定すると，約 1.4 倍に拡大したことに相当する．

5 考察

5.1 任意要素の効果

調査結果より，任意要素指定機能は被覆率を大きく向上させることが分かった．原文任意要素をすべて使用する時，または，文型任意要素をすべて使用する時，文型再現率 $R1$ による文型パターン拡大率 η_{R1} と平均適合パターン数 N による文型パターン拡大率 η_N に差がある．これは，N は，入力文当たりの適合パターン数を測り，$R1$ は適合パターンのある入力文を測るため，適合パターンのある入力文の数が，増大しており，特定の入力文において，適合パターン数が多くなっているのではないと思われる．

5.2 原文任意要素の挿入要素別の効果

原文任意要素の挿入要素別の結果を見ると，連体節（/t）はあまり効果がなさそうである．連体節（/t）で，あまり効果が出なかった理由を考察する．

離散記号は，2.2.1 節のように挿入箇所が決められており，同じ位置に複数の離散記号が挿入されている．よって，文型パターンに，/tcfk となっているところがあるが，そこで連用節（/t）が使用されていなくても，格要素（/c），連体修飾要素（/k），連用修飾要素（/f）があれば，連体節が挿入可能になることがある．そのため，連体節（/t）の効果があまり出なかったと思われる．以下に例を示す．

- **入力文：**この本は遠藤氏が新聞に連載したコラムをまとめたものだ。
- **適合したパターン：**/y#1[GEN2]/kN3 は/tcfk$N$4 を/cf$V$5.kako^rentai!ものだ。
  - /k = この
  - $N3$ = 本
  - /tcfk = $\underset{/t}{\text{遠藤氏が新聞に連載した}}$
  - $N4$ = コラム
  - $V5$ = まとめ
- **/t を除いたパターン：**/y#1[GEN2]/kN3 は/cfk$N$4 を/cf$V$5.kako^rentai!ものだ。
  - /k = この
  - $N3$ = 本
  - /cfk = $\underset{/c}{\text{遠藤氏が}}\ \underset{/c}{\text{新聞に}}\ \underset{/k}{\text{連載した}}$
  - $N4$ = コラム
  - $V5$ = まとめ

5.3 意味的排他性の評価

任意要素を用いた文型パターンが，意味的に正しい文型パターンであるか，また，それに対応する英語パターンが訳文の生成に問題なく使用できるかを調査した．

5.3.1 原文任意要素

原文任意要素の挿入要素により，意味的に不適切なパターンと適合する割合にあまり違いはなかった．しかし，離散記号と適合した入力文の要素が，英語翻訳の際の重要な語になっている場合，もしくは，入力文のほとんどの要素が原文任意要素となる場合，得られたパターンは単純なものであったり，意味的に正しくないパターンである場合が多かった．(適合パターンのうち，約 4 割が意味的に正しくないパターンだと思われる．) 以下に例を示す．

- **日本語文：**皆いっせいに手を上げたので、先生は誰を当てたらよいか迷った。
- **英文：**They raised their hands all at once, so the teacher did not know whom he should call on.
- **適合した日本語パターン**
  /y$1^{#1[$GEN$2]}/k$N$3 は/cf$V$4.kako^katei$1 /fV5.kako。
  - /y = 皆いっせいに手を上げたので、
  - $N3$ = 先生は
  - /cf = 誰を
  - $V4$ = 当てる
  - /f = 良いか
  - $V5$ = 迷う
- **適合した日本語パターンに対応する英語パターン：**
  #1[AJ2] $N3$ $V5$.past in V4

この例文の，英語翻訳の際の重要な語となる,「ので」という部分と「よいか」という部分が原文任意要素と対応している．そのため，得られた日本語パターンに対応する英語パターンは，意味的に正しくないパターンになっている．

このような問題を解決するためには，離散記号の付与基準や，文型パターンの照合条件を見直す必要がある．

5.3.2 文型任意要素

任意要素記号を使用することにより，適合するようになったパターンは，実際の翻訳の際に使用できるかどうかを調査した．その結果，適合したパターンの中には，意味的に正しくないパターンも含まれていた．しかし，任意要素を使用したことが原因で，意味的に正しくないパターンになっているものはなかった．なお，任意要素記号は，その部分が削除されても英語文型自体は変化しないものに付与されているので，任意要素が使われているパターンと適合しても問題はない思われる．

6 おわりに

本研究では，まず，文型パターンの被覆率向上の効果を推定するパラメータとして，文型パターン拡大率 η を提案した．任意要素を用いた効果を，文型再現率 $R1$，および，平均適合パターン数 N から推定した η を用いて評価した．調査結果より，原文任意要素を文型パターンの記述に用いると，基準辞書を 7〜20 倍に拡大したことに相当し，同様に，文型任意要素を文型パターンに用いると，基準辞書を 1.4〜2.6 倍に拡大したことに相当することが分かった．よって，任意要素の指定は，被覆率を大きく向上させるといえる．

しかし，原文任意要素を使用したパターンの意味的排他性は低いので，任意要素の記述方法や挿入箇所などを見直す必要がある．

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] 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.

[3] 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, 自然言語処理, 11(4), pp.147-178, 2004.

[4] 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, 11(3), pp.69-95, 2004.