概要

近年, 機械翻訳において言語表現の構造を意味のまとまる単位にパターン化した文型パターン翻訳方式の研究が行われてきた. しかし, これらの文型パターン翻訳で使用されている文型パターン数は少なく (200〜300 パターン程度), 狭い分野の文章に適用されることが多かった. パターン翻訳を行う際, 大量の文型パターンを用意する必要がある. 現在, 日本語の重複文 12 万文に対してパターンが作成されている [2].

しかし, 日本語の翻訳精度向上のためには, 重複文の基本構造ともいえる単文の文型辞書が必要である.

そこで, 本研究では, 単文の文型パターンを自動的に作成した. まず, 単文の条件を定義し, 日英対訳文 100 万件より単文 215,242 件を抽出した. 次に, 日英対訳辞書を用いて変数化を行い, 日英対訳パターンを作成した.

具体的には, 日本語文を形態素解析にかけ, 日英対訳辞書によって対応関係が決定できる単語を同じ変数に置き換えた. このようにして得られた文型パターンの日本語パターンにおいて重複する単文を削減した結果, 日本語パターンは 201,754 件となり, 日本語パターンの削減率は,6.36%であった.

また, 得られた文型パターンでの翻訳精度を検証するため, 日英対訳文より, ランダムに 100 件の単文を抽出した. 各々の文型パターンを用いて翻訳精度を調査した. その結果, 一つの日本語パターンに対して複数の英語パターンを持つ単文は,100 件中 9 件であった. 単文 9 件の英語パターンの翻訳精度を検証した所, 自己パターン以外の英語パターンを用いても精度の高い英文が得られた.

目 次

表 目 次

1　はじめに

日英機械翻訳において, 要素合成方式が用いられてきた. この方式は,「原文の構文構造を目的言語の構造に変換する過程」と「原文の各要素を翻訳する過程」を持ち, 訳文は両者の結果を合成することによって得られる点に特徴がある. これは, 構文構造と表現の意味を別々に変換するものであり, 表現の構造と意味が線形であることを前提としている [1]. しかし, 言語表現には意味的に非線形なものが多く, 表現を分解して行く過程で全体の意味が失われることが問題であった.

この問題を解決するには,「文構造とその意味を一体的に扱う仕組み」が重要である. この仕組みとして, 古くから,「テンプレート翻訳」と呼ばれる「文型パターン翻訳」の方法が試みられてきた. そして, 大量の日英対訳例文から重複文を対象に文型パターンを作成する方法が提案された [2]. 現在, 日本語の重複文 12 万文に対してパターンが人手によって作成されている. しかし, 単文は対象外であったため, 未だに単文の文型辞書は得られていない. また, 人手によるパターン作成にはコストがかかる.

そこで, 本研究では, 単文の文型パターンを自動的に作成し, 翻訳精度を検証する. まず,CREST 対訳例文 100 万件 [3] から単文を抽出する. 次に日英対訳辞書を用いて日英の単語の対応関係を発見し, 変数化する. 最後に得られた文型パターンを用いて英文生成し, 翻訳精度を検証する.

以下, 第 2 章では単文の条件と抽出方法を示す. 第 3 章では, 日英対訳パターンの作成手順について述べる. 第 4 章では, 得られた日英パターンの削減率のついて, また, 第 5 章では, 文型パターンの翻訳精度を評価する. 最後に第 6 章では, 変数化の問題点と汎化による文型パターンの縮退について述べる.

2　単文抽出とデータベース作成

2.1　単文の条件

本研究では, 単文の条件を定義して CREST 対訳例文 100 万件 [3] より単文を自動的に抽出する. 一般的には、単文とは「述語が一つだけから成る文」であると述べられているが, 本研究では、単文の条件を日本語側からみて定義する.

1. 文中に動詞がひとつだけある文.
    - (例１) 彼は毎日自転車に乗る。
2. 文中に動詞がなく, 複合動詞がひとつだけある文.
    - (例２) ドイツは新しい歴史への一歩を踏みだした。
3. 文中に動詞, 複合動詞, 形容詞がひとつもなく, 形容詞がひとつだけある文.
    - (例３) この林檎はややすっぱい。
4. 文中に動詞, 複合動詞, 形容詞, 形容動詞がひとつもなく, 形容動詞がひとつだけある文.
    - (例４) 企業の経営戦略は大切だ。
5. 文中に動詞, 複合動詞, 形容詞, 形容動詞がひとつもなく, 文末が' 名詞＋付属語' で終わっている文.
    - (例５) あの人こそ真の英雄だ。
6. 疑問文, 命令文, 会話文は対象外とする.
    - (例６-１) 疑問文：この本は何について書いてあるか。
    - (例６-２) 命令文：そこに私のテントを張れ。
    - (例６-３) 会話文：きのうどこかへ行ったかい。

2.2 単文抽出手順

1. CREST 対訳例文 100 万件 [3] の日本語文を形態素解析する. 形態素解析には,ALT ー JAWS[4] を使用する. 本語例文を形態素解析にかけた結果の一部を表 1 に示す.

   - 日本語例文：彼女はセーターを編み上げた。

表 1: 形態素解析結果の一部

単語	品詞コード	品詞	標準表記
彼女	1710	人称代名詞	彼女
は	7530	付属語副助詞	は
セーター	1100	一般名詞	セーター
を	7430	付属語格助詞	を
編み上げ	2413	動詞	編み上げ
た	7216	付属語助動詞	た
。	0110	文末記号	。

2. 形態素解析結果の品詞コードを用いて日本語文が, 単文の条件 (2 . 1 節参照) にあてはまるか判定する. 表 1 の品詞コード'2413' は単語「編み上げ」が動詞であることを表しており, 単文の条件 1 にあてはまるので日本語例文は単文である. 他の条件についても同様にして, 形態素解析の結果を利用して単文を抽出する.

3. 抽出した単文が述部を一つ持つか CL の定義を用いて確認する.CL とは, 節を表す変数で,CL の定義は全部で 230 種類用意する. 以下に CL の定義の一部を示す. 詳細は付録の 1 に掲載する.

   - CL1.darou。
   - CL1.kako.darou。
   - CL1.suitei.kako。
   - CL1.kaishi.teinei。
   - CL1.teiku.darou。
   - CL1.reru.teinei.kako。
   - CL1.sugiru.teiru。

2.3 データベース

定義した単文の条件 (2.1 節参照) に従って CREST 対訳例文 100 万件 [3] より抽出された単文は, 全部で 215,242 件であった. 以下に抽出された単文の例を示す.

- 東京の旅は楽しかった。
- 私は大満足だった。
- 赤い旗は止まれの印です。
- 私は英語の先生が好きだ。
- あの人はいつもきげんがよい。
- 蜘蛛の巣は不思議な現象だ。
- 環境汚染は、全地球的な問題だ。
- 住宅地の開発は公的性格の強い事業だ。

3 日英対訳パターンの作成

3.1 作成手順

2 章で得られた単文 215,242 件に対して日英対訳パターンを作成する. 日英対訳パターンの作成手順を以下に示す.

1. **日英対訳文を用意**
   日英対訳文を用意し, 日本語文を形態素解析にかける [4]. 日英対訳例文と日本語文の形態素解析結果の一部を表 2 に示す.
   - 日本語例文 = 妹は私と同じくらい一所懸命勉強する。
   - 英語例文 = My sister studies as hard as I.

表 2: 形態素解析結果の一部

単語	品詞コード	品詞	標準表記
妹	1100	一般名詞	妹
は	7530	付属語副助詞	は
私	1710	人称代名詞	私
と	7420	付属語格助詞	と
同じ	3217	形容動詞	同じだ
くらい	7520	付属語副助詞	くらい
一所懸命	1240	用言性名詞	一所懸命
勉強	1230	用言性名詞	勉強
する	2636	動詞	する
。	0110	文末記号	。

2. **変数の決定**

日英対照辞書を用いて日本語単語に対応する英語単語を見つける. 本研究では,7 つの日英対照辞書を用いる. 表 3 に日英対照辞書の例を示す. 各々の辞書には日本語単語と対応する英語単語が掲載されており, 形態素解析結果の品詞コードで使用する辞書を判断する. 例えば, 名詞「妹」を日英対照辞書の名詞辞書において発見し,「妹」の訳語が「sister」であり, 英語文中の「sister」と対応関係がとれるので「妹」「sister」を同じ変数「N」と決定する. 代名詞「私」と「I」についても同様にして変数「PRO」と決定する. 本研究では, 日英対訳辞書によって対応関係がとれる単語のみを変数化する. 表 4 に変数の例を示す.

表 3: 変数の例

辞書名	日本語単語	英語単語
名詞辞書 (noun_dic)	災害	disaster
固有名詞辞書 (pnoun_dic)	ギリシア共和国	Hellenic Republic
副詞辞書 (adv_dic)	どちらにしても	anyhow,anyway,in either case
連体詞 (adj_dic)	あくせく	restlessly,in a fidget,busily
代名詞 (pro_dic)	あなた	you
形容詞 (adj_dic)	だるい	be tired,be dull,dull,feel weary,tired
動詞 (verb_dic)	はしゃぐ	be in high spirits,frolic,romp about

表 4: 変数の例

品詞	日本語単語	英語単語	変数
名詞	妹	sister	N
代名詞	私	I	PRO

3. **単語を変数に置換**
   日英対訳文中の単語を決定された変数に置き換える. そのとき, 対応付けのため, 変数化された順番に変数に番号を付ける.

   - 日本語パターン = $N1$ は $PRO2$ と同じくらい一所懸命勉強する。
   - 英語パターン = My $N1$ studies as hard as $PRO2$.

3.2 変数定義

本研究では, 日本語側からみて形態素解析によって判断された品詞の変数化を行う. 変数化する品詞は体言 5 品詞 (名詞, 固有名詞, 副詞, 連体詞, 代名詞) と用言 2 品詞 (形容詞, 動詞) である. 名詞の複数形と所有格, 形容詞の比較級と最上級, 動詞の三単現と過去形については, 英語側の変数に形態素調整の記号を付ける. 例えば, 「一冊の本」⇔”a book “と「多くの本」⇔”many books” のように日本語側では単数形も複数形も同じ表現「本」であるが英語側では違った表現”book” と” books” になる. 英語パターンを用いて英文生成する場合, 形態素調整の記号は有効である. 表 5 に, 変数の定義を示す.

表 5: 変数の定義

品詞	品詞番号の先頭	日本語側	英語側
名詞	11〜15,17,18	N	N
名詞の複数形	11〜15,17,18	N	$N\hat{}pl$
名詞の所有格	11〜15,17,18	N	$N's$
固有名詞	19,1A	$PNOUN$	$PNOUN$
副詞	41	ADV	ADV
連体詞	42	REN	REN
代名詞	17	PRO	PRO
形容詞	31	ADJ	ADJ
形容詞の比較級	31	ADJ	$ADJ1\hat{}er$
形容詞の最上級	31	ADJ	$ADJ\hat{}est$
動詞	21〜28	$VERB$	$VERB$
動詞の三単現	21〜28	$VERB$	$VERB\hat{}s$
動詞の過去形	21〜28	$VERB$	$VERB\hat{}kako$

表中の「名詞」には, 未整理名詞, 一般名詞, 用言性名詞, 転生名詞, 時詞, 形式名詞を含む. また名詞の複数形を変数化する場合や形容詞の比較級, 最上級を変数化する場合は, 名詞辞書の他に英語辞書を使用する. 英語辞書の一部を表 6 に示す.

表 6: 英語辞書の一部

英語単語	所有格	目的格	複数形	動詞派生
book	book's	book	books	bookish
leaf	leaf's	leaf	leaves	
man	man's	man	men	
schoolyard	schoolyard's	schoolyard	schoolyards	
underclothes	underclothes's	underclothes	underclotheses	
英語単語	比較級	最上級	名詞派生	副詞派生
beautiful	more beautiful	most beautiful	beauty	beautifully
superior	more superior	most superior	superiority	superiorly
happy	happier	happiest	happiness	happily
unlucky	more unlucky	most unlucky	unluckiness	unluckily
small	smaller	smallest	smallness	small

例えば,「本」の複数形「books」を変数化する場合は, 日英対照辞書の名詞辞書で「本」を発見し, その訳語「book」が英語文中に存在し対応関係があるか判断する. しかし, 英語文中では「books」となっているため,「book」を見出し語として英語辞書において発見し,「book」の複数形「books」と対応関係がとれるので, 最終的に「本」と「books」を変数に置き換える. このとき,「本」は変数「N」に,「books」は変数「N$\hat{\text{p}}$l」に置き換える. 形容詞の比較級や最上級についても同様にして変数化する.

4　得られた日英対訳パターン

4.1　変数化された単語

4.1.1　変数化された単語数の割合

単文 215,242 件を変数化し, 単文の文型パターンを作成した. 変数化された単語の割合を調査するため, 日英対訳文中の日本語文の形態素解析によって判断された品詞の単語数を数えた. また, 日英対訳辞書によって変数化された単語数を数えた. 変数化できた単語の割合を品詞別に表 10 に示す.

表 7: 変数化された単語の割合

品詞	単語数	変数化された単語数	変数化の割合 [%]
名詞	495,720	229,781	46.35
固有名詞	25,802	10,963	42.48
副詞	31,477	10,011	31.80
連体詞	50,588	21,159	41.82
代名詞	81,218	48,932	60.24
形容詞	38,343	12,497	32.59
動詞	177,129	51,744	29.21

4.1.2　変数化された単語の例

- 名詞の例
  - 変数化された単語：コーヒー⇔ coffee
    - 私は何よりも N1 が好きだ。
    - I love N1 more than anything else.
- 名詞の複数形の例
  - 変数化された単語：リンゴ⇔ apples
    - ふじは N1 のなかでもとりわけおいしい。
    - Fuji is especially good tasting among all brands of N1p̂l.

- 固有名詞の例
  - 変数化された単語：日本⇔ Japan
    - PNOUN1 の学者は皆貧乏だ。
    - Scholars in PNOUN1 are all poor.
- 副詞の例
  - 変数化された単語：だいぶ⇔ considerably
    - ２学期は成績が ADV1 下がった。
    - My grades dropped ADV1 in the second term.
- 連体詞の例
  - 変数化された単語：こうした⇔ Such
    - REN1 人物には「ピックウィックペーパーズ」以外ではめったにお目にかからない。
    - REN1 a character is rarely to be met with outside the pages of Pickwick Papers.
- 代名詞の例
  - 変数化された単語：彼らは⇔ They
    - PRO1 は新しい世界でのよりよい生活に賭けた。
    - PRO1 took a chance on a better life in the new world.

- 形容詞の例
  - 変数化された単語：若い⇔ young
    - その ADJ1 医師は近くの診療所で６か月の学外研修をした。
    - The ADJ1 doctor served 6 months of externship at a nearby clinic.
- 形容詞の比較級, 最上級の例
  - 変数化された単語：短⇔ short
    - 日がだんだん ADJ1 くなってきた。
    - The days are getting ADJ1êr.
- 動詞の例
  - 変数化された単語：出席し⇔ attended
    - 当選一回の所属国会議員が VERB1 た。
    - Newly elected Diet members VERB1k̂ako.

4.2 文型数の調査

4.2.1 文型の削減率の調査

変数化によって得られた文型パターンにおいて日本語パターンの異なり数を調査し, 重複する日本語パターンを削減した. ここで, 総文数に対して削減された文数の割合を削減率とする. 表 8 に, 日英対訳文数に対する品詞ごとの変数化後のパターンの削減率を示す. 表 8 より, 名詞の変数化によるパターンの削減率が高いことがわかる. また, 体言 (名詞, 固有名詞, 副詞, 連体詞, 代名詞) を変数化した場合は, 削減率が 4.73%であったが, 用言 (形容詞, 動詞) を変数化することで 6.36%に上がった.

表 8: 重複するパターンの削減率

品詞	総文数 [件]	削減後の文数 [件]	削減率 [%]
名詞のみ	215,242	206,246	4.27
固有名詞のみ	215,242	209,980	2.54
副詞のみ	215,242	210,037	2.51
連体詞のみ	215,242	210,042	2.51
代名詞のみ	215,242	209,511	2.66
体言	215,242	205,257	4.73
形容詞のみ	215,242	209,840	2.51
動詞のみ	215,242	209,969	2.54
すべての品詞	215,242	201,754	6.36

4.2.2 作成された文型パターンの例

作成した文型パターンにおいて出現頻度の高い日本語パターンの上位 3 番目までのパターンとそれに対応する出現頻度の高い英語パターンを示す. 括弧内の数字は出現回数を示す。

- 日本語パターン：*PRO*1 は *N*2 を *VERB*3 た。(278)
  - 対応する英語パターン
    - *PRO*1 *VERB*3ˆ*kako* the *N*2.(53)
    - *PRO*1 *VERB*3ˆ*kako* his *N*2.(41)
    - *PRO*1 *VERB*3ˆ*kako* a *N*2. (34)
- 日本語パターン：*N*1 が *VERB*2 た。(198)
  - 対応する英語パターン
    - The *N*1 *VERB*2ˆ*kako*. (68)
    - The *N*1 has *VERB*2ˆ*kako*.(22)
    - A *N*1 *VERB*ˆ*kako*. (11)
- 日本語パターン：*PRO*1 は *N2* が *ADJ*3。(141)
  - 対応する英語パターン
    - *PRO*1 has a *ADJ*3 *N*2. (52)
    - *PRO*1 has *ADJ*3 *N*2ˆ*pl*.(27)
    - *PRO*1 has *ADJ*3 *N*2. (12)

5　文型パターンの調査

5.1　調査対象

作成したパターンを用いての翻訳精度を調査するため、単文 215,242 件よりランダムに 100 件を抽出し、調査対象とした。単文 100 件は表 10 のように分類された。また, 分類された単文の例を以下に示す.

表 9: ランダム 100 件の分類

分類	対象となる文 重複文の基本構造ともいえる単文の文型辞書が必要である.	件数 [件]
A	変数化できずに原文のままである文	9
B	日本語パターンに対して英語パターンが一つある文	82
C	日本語パターンに対して英語パターンが複数ある文	9

- 分類 A の例
  - フランス語は正課の一部となっている。
  - 夏休みをまるまる無駄に過ごしてしまった。
  - 不思議な縁で彼と知り合った。
- 分類 B の例
  - 日本文…彼はその戦闘で傷を受けた。
  - 英語文…He was wounded in that battle.
    - 日本語パターン…*PRO*1 は *REN*2*N*3 で傷を受けた。
    - 英語パターン…*PRO*1 was wounded in *REN*2 *N*3.
  - 日本文…予備校生は夜も昼も勉強する。
  - 英語文…Yobiko students study day and night.
    - 日本語パターン…予備校生は *N*1 も昼も *VERB*2。
    - 英語パターン…Yobiko students *VERB*2 day and *N*1.

- 日本文…人間の寿命は年々延びている。
- 英語文…Man's life span has been growing longer from year to year.
  - 日本語パターン…人間の *N*1 は年々延びている。
  - 英語パターン…Man's *N*1 span has been growing longer from year to year.

- 分類Cの例
  - 日本文…朝食の用意ができました。
  - 英語文…Breakfast is ready.
    - 日本語パターン…*N*1 の用意ができました。
    - 英語パターン (1)…*N*1 is served.
      - 英語パターン (1) の日本語原文…夕食の用意ができました。
      - 英語パターン (1) の英語原文…Dinner is served.
    - 英語パターン (2)[自己パターン]…*N*1 is ready.
  - 日本文…男はその場で逮捕された。
  - 英語文…The man was arrested on the spot.
    - 日本語パターン…*N*1 はその *N*2 で *VERB*3 れた。
    - 英語パターン (1)…The *N*1 was *VERB*3ˆ*kako* on the *N*2.
      - 英語パターン (1) の日本語原文…加害者はその場で逮捕された。
      - 英語パターン (1) の英語原文…The assailant was arrested on the spot.
    - 英語パターン (2)[自己パターン]…The *N*1 was *VERB*3ˆ*kako* on the *N*2.

5.2 翻訳精度の評価

5.2.1 評価基準

一つの日本語パターンに対して複数の英語パターンを持つ単文 9 件 (4.1 節参照) について英語パターンの翻訳精度を調査した. 英語パターンにおいて自己パターン以外のパターンを用いて英文生成できるか評価した. 評価基準は以下の 4 段階とする [5].

- A … 意味的に正しくそのまま英文生成に使用できる
- B … 重要でない情報が欠落しているか, 文法的に正しくない所があるが英文生成に使用できる
- C … 重要な情報が欠落しているが英文生成に使用できる部分もある
- D … 重要な情報が間違っており, 英文生成に全く使用できない

5.2.2 評価結果

一つの日本語パターンに対して複数の英語パターンを持つ単文 9 件の英語パターンは, 全文で 32 件であった. 表 10 に評価結果を示す. 表 11, 表 12, 表 13 に例文に対する複数の英語パターンの評価結果を示す. 表中の波線は自己パターンを表す. 評価結果より, 自己パターン以外の英語パターンを用いて英文を生成した所, 精度の高い英文が得られた.

表 10: 評価結果

調査パターン数	A [%]	B [%]	C [%]	D [%]
32	40.6%	6.2%	15.6%	37.5%

- 日本語例文 1…この肉は堅い。
- 英語例文 1…This meat is tough.

表 11: 例文 1 の英語パターンの評価結果

日本語パターン	出現回数	英語パターン	出現回数	評価
*REN*1*N*2 は *ADJ*3。	89	*RNE*1 *N*2 is *ADJ*3.	73	A
		*REN*1 *N*2 tastes *ADJ*3.	4	A
		*REN*1 is a *ADJ*3 *N*2.	3	D
		*REN*1 is a *ADJ*3 *N*2.	2	A
		There are *ADJ*3 *REN*1 *N*2*pl*.	1	B
		*REN*1 kind of *N*2*pl* are *ADJ*3 to answer.	1	C
		*REN*1 *N*2*pl* are *ADJ*3.	1	A
		*REN*1 *N*2 smells *ADJ*3.	1	C
		*REN*1 *N*2 is *ADJ*3 to understand.	1	C
		*REN*1 *N*2 is *ADJ*3 to solve.	1	D
		*REN*1 *N*2 is *ADJ*3 in his movements.	1	D

- 日本語例文 2…雨がやんだ。
- 英語例文 2…The rain has left off.

表 12: 例文 2 の英語パターンの評価結果

日本語パターン	出現回数	英語パターン	出現回数	評価
*N*1 がやんだ。	10	The *N*1 has passed.	2	A
		The *N*1 has dropped.	2	D
		The *N*1 died away.	2	D
		The *N*1 is over.	1	D
		The *N*1 has left off.	1	
		The *N*1 has fallen.	1	D
		The *N*1 has died down.	1	D

- 日本語例文 3…タイヤが外れた。
- 英語例文 3…The tire has come off.

表 13: 例文 3 の英語パターンの評価結果

日本語パターン	出現回数	英語パターン	出現回数	評価
*N*1 が外れた。	11	The *N*1 has come off.	2	A
		The *N*1 got unhinged.	2	D
		The *N*1 came off.	2	A
		The sliding *N*1 slipped off its rail.	1	D
		The *N*1 has got unhinged.	1	C
		My *N*1 ticket did not win.	1	D
		My *N*1 came unlink.	1	C
		It fellshort of my *N*1.	1	D

6　考察

6.1　変数化の問題点

6.1.1　変数化の失敗例

本研究では, 単語の変数化を自動的に行った. 変数化に失敗した原因を検証するため, ランダムに抽出した 100 件の単文において、変数化できなかった単語を調査した. 以下に名詞について変数化に失敗した原因の分類を示す.

- A…日本語単語は辞書にあるが, 対応する英語単語が英語文の単語と異なっている場合.
  - (Aの例)
    - 日本文…株価が２割５分下がった。
    - 英語文…Stocks went down 25 percent.

    辞書には, 株価⇔ stock price と載っている.
- B…対訳の英語単語が複数単語である場合.
  - (Bの例)
    - 日本文…吉田選手が一塁を守っている。
    - 英語文…Yoshida plays first base.

    辞書には, 一塁⇔ first base と載っているが, 本研究では複数単語の変数化に対応できていない.

- C…日本語側では名詞扱いだが, 英語文では動詞など異なる品詞で表現されている場合.
  - (Cの例)
    - 日本文…上半身裸であった。
    - 英語文…She was naked to the waist.

    「裸」が日本語側では名詞と判断されているが英語側では”be naked” と表現されている.
- D…日本語単語が全く辞書に載っていない場合.
  - (Dの例)
    - 日本文…西洋流の教育によって村人たちのしきたりは衰えている。

    日本文の名詞「しきたり」は, 本研究で用いた日英対訳辞書に載っていない.
- E…形態素解析ミス
  - (Eの例)
    - 日本文…その戦闘では非常に多くの死傷者が出た。

    日本文の「非常」が形態素解析によって名詞と判断されていた.

単文 100 件において形態素解析によって名詞と判断された単語は, 222 個であった. 変数化できた単語は,111 個であった. 以下の表 14 に変数化できなかった単語 111 個を上記のＡ～Ｅに分類した結果を示す.

表 14: 変数化できなかった名詞単語の分類

分類	単語の個数 [個]	割合 [%]
A	35	31.5
B	9	8.1
C	53	47.7
D	7	6.3
E	7	6.3

上記の分類のＡとＤは辞書を強化することで、Ｂは変数化プログラムを改良することで変数化できる. しかし,C の問題を自動的に解決するのは難しく, 人手による判断が必要である.

また, その他の品詞について検証した所, 固有名詞, 副詞, 形容詞は 100 件中, 単語の出現回数が 20 回未満と少なかったため, 変数化できない原因の傾向がつかめなかった. 連体詞については, 辞書を強化すること, 代名詞については, 代名詞の所有格も変数化することでほぼすべての単語が変数化できる. 動詞に関しては, 複合動詞の変数化と辞書の強化が必要である.

従って, 辞書の強化と変数化プログラム改良で変数化できる単語の割合が全体的に約 50%増加すると予想される.

6.1.2 変数番号の問題点

本研究では, 変数化の際に変数化された順番に単語に変数番号を付けた. しかし, 変数化された単語の順番を自動的に決定するのは困難であることがわかった. 特に形態素解析によって品詞番号”1710” と判断された人称代名詞を変数化する場合には注意が必要である.

まず, 日本語文には存在しない代名詞が英語文では存在している. 特に, 名詞の直前の代名詞の所有格である場合が多く, 日本語と英語の文法的な違いが生じるためだと考えられる.

さらに, 代名詞の主格と代名詞の所有格が同時に英語文中に存在するが, 日本語文中では代名詞が一つの場合がある. この場合, どちらの代名詞が日本語単語に対応する訳語であるかを自動的に判断するのは, 困難である. 以下に例を示す.

- 日本語文…彼はバッグを肩に、旅に出た。
- 英語文…He went on a trip in with a bag on his shoulder.

日本語文において代名詞は「彼」のみである. しかし, 英語文中には,”He” と”his” のように代名詞が二つ存在する. 代名詞の所有格も変数化できるが, 変数番号に不具合が生じるため, 本研究では, 代名詞の主格のみを変数化した. 今後は,”his” のような名詞の直前の代名詞の所有格は, 名詞に含めて変数化する方法も考えられる.

6.2 汎化によるパターンの縮退

6.2.1 日本語パターンの汎化

単文は, 文構造が簡単であるため, 単文の文型パターンは, かなりの割合で同一化できると予想していた. しかし, 本研究で得られた文型パターンにおいて重複する日本語パターンを削減した所, 原文 215,242 件に対して削減できたパターン数は,13,488 件と低かった. また, 同一化できそうな日本語パターンについて検証するため, 動的計画法で類似の日本語パターンを検索した. 以下に例を示す.

- 日本語パターン [A] … $N1$ が $ADV2$ $ADJ3$ くなってきた。
  - 日本語の原文
    - 空がだいぶ明るくなってきた。
  - 対応する英語パターン [a]
    - The $N1pl$ have become $ADV2$ $ADJ3\hat{}er$.
  - 英語の原文
    - The skies have become considerably brighter.
- 日本語パターン [B] … $N1$ が $ADV2$ 弱くなってきた。
  - 日本語の原文
    - 脈がだんだん弱くなってきた。
  - 対応する英語パターン [b]
    - His $N1$ has $ADV2$ weakened.
  - 英語の原文
    - His pulse has gradually weakened.

日本語パターン [A] と [B] は, 類似したパターンであるが, 英語パターン [a] と [b] は意味的に全く異なっている. そのため, 英語パターン [b] を用いて日本語パターン [A] の英語の原文'The skies have become considerably brighter. を訳出することは難しい. 従って, 日本語パターンを汎化することは困難であると考えられる.

6.2.2 英語パターンの汎化

一つの日本語パターンに対する複数の英語パターンにおいて検証した所, 名詞の直前の単語を名詞に含めて考えることで, パターンの縮退ができることがわかった. 例えば名詞の直前の”a” や”the” などの冠詞や”my”,”his”,”her” などの人称代名詞の所有格である. しかし, 英文生成する場合は, 人称代名詞の所有格は英語文において省略できない要素である.

また, 本研究では名詞の単数形と複数形は別の変数に置き換えたが, 日本語では名詞は単数形で表現されることが多い. 従って, 英語パターンにおいて名詞は単複同じ変数で表すことができる. 以下に日本語パターン「*PRO*1 は *N*2 を V*ERB*3 た。」に対して同一化できそうな英語パターンの例を示す.

- 冠詞のみが異なる場合
  - *PRO*1 *VERB*3^*kako* a *N*2.
    - 英語原文：I used a toothbrush.
    - 日本語原文：私は歯ブラシを使った。
  - *PRO*1 *VERB*3^*kako* an *N*2.
    - 英語原文：I peeled an apple.
    - 日本語原文：私は林檎の皮を剥いた。
  - *PRO*1 *VERB*3^*kako* the *N*2.
    - 英語原文：I peeled the banana.
    - 日本語原文：私はバナナの皮を剥いた。
- 人称代名詞の所有格のみが異なる場合
  - *PRO*1 *VERB*3^*kako* his *N*2.
    - 英語原文：He bowed his head.
    - 日本語原文：彼は頭を下げた。
  - *PRO*1 *VERB*3^*kako* her *N*2.
    - 英語原文：She opened her hand.
    - 日本語原文：彼女は手を開いた。

- *PRO*1 *VERB*3ˆ*kako* our *N*2.
    - ∗ 英語原文：We continued our quest.
    - ∗ 日本語原文：われわれは捜索を続けた。

- 単数形と複数形が異なる場合
    - *PRO*1 *VERB*3ˆ*kako* his *N*2.
        - ∗ 英語原文：He teased his brother.
        - ∗ 日本語原文：彼は弟をいじめた。
    - *PRO*1 *VERB*3ˆ*kako* his *N*2*pl*.
        - ∗ 英語原文：He bent his knees.
        - ∗ 日本語原文：彼は膝を曲げた。

7　おわりに

本研究では, 自動的に単文の文型パターンを作成した. 日英対訳文 215,242 件より, 日英文型パターンを作成した. 単文は文構造が簡単であるため, 単文の文型パターンは, かなりの割合で同一化できると予想していた. しかし, 得られたパターンの日本語パターンにおいて重複するパターンを削減した結果, 異なる日本語パターンは,201,754 件であった. なお, 自動的に変数化できた単語は, 約 50%であった.

また, 得られた文型パターンを用いて英文生成した所, 良い翻訳精度が得られた. 今後は, すべて自動的に英文生成を行い, 翻訳精度を調査したい.

謝辞

本論文作成に際して, 多大なる検討と助言をしてくださった池原悟教授ならびに村上仁一助教授, 徳久雅人助手そして計算機工学C研究室の方々に深く感謝します. また, 参考にさせて頂いた文献の著者の方々に対して感謝します.

参考文献

[1] 長尾真, 黒橋貞夫, 佐藤理史, 池原悟, 中尾洋, 言語情報処理, 岩波講座「言語の科学」, 9 巻, 岩波書店.

[2] 池原 : 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] 村上ほか:日本語英語の文対応の対訳データベース, 「言語・認識・表現」, 第 7 回年次研究会, 2002-12

[4] 白井論, 池原悟, 横尾昭男, 奥山信輔, 宮崎正弘 : 多段解析による日本語形態素解析の精度, 情報処理学会, 第 50 回全国大会 (1995.3.15-18).

[5] Sumita Eiichiro:Example-based machine translation using DP-matching between word sequences,DDMT workshop of 39th ACL,2001

付録1：品詞コード表

付録2：CLの定義(230種類)

付録3：作成したパターン

- 表1に日本語パターンの出現頻度の上位15位のパターンを掲載する.
- 表2に英語パターンの出現頻度が10位以上のパターンを掲載する.

付録 4：翻訳実験結果