単文文型パターン辞書の構築

西山 七絵　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
{nisiyama,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

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.

2　単文抽出

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」であると認識されているが, 定義は曖昧であることが多い. そこで, 本研究では, 単文の条件を日本語側からみて以下のように定義する.

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。

3　日英対訳パターンの作成

3.1　作成手順

2 章で定義した単文の条件に従い,CREST 対訳例文 100 万件 [3] より抽出した単文 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. 形態素解析には,ALT ー JAWS[4] を使用する.
   - 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.
2. **変数の決定**
   ALT の日英対訳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 つの日英対訳辞書を用いる. 各々の辞書は品詞別に単語が登録されており, 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. 対訳辞書によって対応関係が決定できた単語を日英同一の変数に置き換える. 変数に置き換えられた順番に変数に番号を付ける. 表 1 に例を示す.

表 1: 変数の例

品詞	日本語単語	英語単語	変数
名詞	妹	sister	$N1$
代名詞	私	I	$PRO2$

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.
   - 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

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) とする. 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. 形態素調整の記号は, 英語パターンを用いて英文生成する場合に利用できる. 表 2 に, 変数の定義を示す.

表 2: 変数の定義

品詞	日本語側	英語側
名詞	$N1$	$N1$
名詞の複数形	$N1$	$N1\hat{}pl$
名詞の所有格	$N1$	$N1\acute{}s$
固有名詞	$PNOUN1$	$PNOUN1$
副詞	$ADV1$	$ADV1$
連体詞	$REN1$	$REN1$
代名詞	$PRO1$	$PRO1$
形容詞	$ADJ1$	$ADJ1$
形容詞の比較級	$ADJ1$	$ADJ1\hat{}er$
形容詞の最上級	$ADJ1$	$ADJ1\hat{}est$
動詞	$VERB1$	$VERB1$
動詞の三単現	$VERB1$	$VERB1\hat{}s$
動詞の過去形	$VERB1$	$VERB1\hat{}kako$

4 得られた日英対訳パターン

4.1 変数化された単語

単文 215,242 件を変数化し, 単文の文型パターンを作成する. 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る. また, 日英対訳辞書によって変数化された単語数を数える. 変数化できた単語の割合を品詞別に表 3 に示す.

表 3: 変数化された単語の割合

品詞	単語数	変数化単語数	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
すべての品詞	900,277	385,087	42.8

すべての品詞を変数化した結果, 日英対訳辞書によって 42.8%の単語が自動的に変数化されたことが示された.

4.2 文型数の調査

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. 表 4 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 4 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%になった.

表 4: 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

実験の結果, 日本語パターンにおいて重複するパターンを削減しても, 削減率が 7%に満たないことが示された.

4.3 翻訳精度の調査

文型パターンを用いた翻訳精度を検証するため, 単文 215,242 件 (2 節参照) よりランダムに 100 件の単文を抽出した. 各々の日英対訳文から作成された日本語パターンに対する英語パターンを調査した. その結果, ひとつの日本語パターンに対して自己以外の英語パターンを持つ単文は 9 件あった. この英語パターンから頻度の高いパターンを選択して英文生成した所,9 件とも精度の高い英文が得られた. 調査結果の一部を表 5 に示す.

表 5: 調査結果の一部

日本語文	英語文	
雨がやんだ。	The rain has left off.	
日本語パターン	英語パターン	頻度
*N*1 がやんだ。	The *N*1 has passed.	2
作成された英文	The rain has passed.	

日本語文	英語文	
この肉は堅い。	This meat is tough.	
日本語パターン	英語パターン	頻度
*REN*1*N*2 は *ADJ*3。	*RNE*1 *N*2 is *ADJ*3.	72
	*REN*1 *N*2 tastes *ADJ*3.	4
	*REN*1 is a *ADJ*3 *N*2.	2
作成された英文	This meat tastes tough.	

この例では, 日本語文「この肉は堅い。」に対して自己以外の英語パターン「*RNE*1 *N*2 is *ADJ*3.」の変数 *REN*1 に英語単語 'This' を *N*2 に 'meat' を *ADJ*3 に 'tough' を人手で代入して英文を得た.

5 考察

5.1 変数化の問題点

5.1.1 変数化失敗の原因

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.

1. **分類Ａ：対訳単語**
   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. 例文に関して, 辞書には, 株価⇔ stock price と載っている.
   - (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.
2. **分類Ｂ：複数単語**
   対訳の英語単語が複数単語である場合. 例文に関して辞書には, 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.
   - (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.
3. **分類Ｃ：表現の異なり**
   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. 例文に関して「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.
   - (Cの例)
     - 日本文…上半身裸であった。
     - 英語文…She was nakes to the waist.
4. **分類Ｄ：不足単語**
   日本語単語が全く辞書に載っていない場合. 例文の名詞「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.
   - (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。
5. **分類Ｅ：形態素解析ミス**
   形態素解析ミスの場合. 例文の「非常」が形態素解析によって名詞と判断されていた.
   - (Eの例)
     - 日本文…その戦闘では非常に多くの死傷者が出た。

5.1.2 変数化の改良点

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 6 に変数化できなかった単語 111 個を (5.1.1 節) のＡ～Ｅに分類した結果を示す.

表 6: 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	32
B	9	8
C	53	48
D	7	6
E	7	6

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.

5.2 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.

- 英語文で人称代名詞を 2 つ含む文
  - 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.

5.3 汎化によるパターンの縮退

5.3.1 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. 以下に結果を示す.

- 日本語パターン [A] … *N*1 が *ADV*2*ADJ*3 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 *N*1*pl* have become *ADV*2 *ADJ*3^*er*.
  - 英語の原文
    - The skies have become considerably brighter.
- 日本語パターン [B] … *N*1 が *ADV*2 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 *N*1 has *ADV*2 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. また,10 文を動的計画法で調査した所, 同様の結果が得られた. 従って, 日本語パターンを自動的に汎化することは困難であると考えている. 今後は, 人手を含めた半自動化の方法も考えている.

5.3.2 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 *VERB*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.
  - *PRO*1 *VERB*3^*kako* the *N*2.
- 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.
  - *PRO*1 *VERB*3^*kako* our *N*2.
- 単数形と複数形が異なる場合
  - *PRO*1 *VERB*3^*kako* his *N*2.
  - *PRO*1 *VERB*3^*kako* his *N*2^*pl*.

6 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.

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).