002038 単文文型パターン辞書の構築

計算機工学講座 池原研究室　西山 七絵

1 はじめに

日英機械翻訳において, 従来, 要素合成法を基本とした機械翻訳方式が用いられてきた. この方式を超える方法として, 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を用いて翻訳する方法が提案された [1]. しかし, あらゆる分野の文章を翻訳するために充分な文型パターン数は, 未だに得られていない. 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, 単文の日英対訳パターンを自動的に作成し, 翻訳精度を検証する.

2 単文抽出

本研究で利用する単文を CREST 対訳例文 100 万件 [2] より抽出する. 本研究では, 動詞, 複合動詞, 形容詞が文中に一つだけ存在する文と文末が「名詞＋付属語」で終わっている文を単文とする. また, 疑問文, 命令文, 会話文は対象外とする. なお, 原文の日本語側を形態素解析にかけ, 定義した単文の条件から単文を抽出する. 抽出した単文 215,242 件を研究対象とする.

3 日英対訳パターンの作成

3.1 作成手順

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する.
   - 日本語文 = 妹は私と同じくらい一所懸命勉強する.
   - 英語文 = My sister studies as hard as I.
2. 変数の決定
   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. 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

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. 表 2 に, 変数の定義の一部を示す.

表 2: 変数の定義の一部

品詞	日本語側	英語側
名詞	N	N
動詞	$VERB$	$VERB$
動詞の三単現	$VERB$	$VERB\hat{}s$
動詞の過去形	$VERB$	$VERB\hat{}kako$

4 得られた日英対訳パターン

4.1 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 3 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の結果を示す. 全ての品詞を変数化してもパターンの削減率は 6.36%と低かった. 表中の体言とは, 名詞, 固有名詞, 副詞, 連体詞, 代名詞を表す. また, 変数化できた単語の割合を表 3 の備考欄に示す.

表 3: 重複するパターンの削減率 (総文数 215,242 件)

品詞	削減後文数 [件]	削減率 [%]	備考*[%]
名詞のみ	206,246	4.27	46.4
固有名詞のみ	209,980	2.54	42.5
副詞のみ	210,037	2.51	31.8
連体詞のみ	210,042	2.51	41.8
代名詞のみ	209,511	2.66	60.2
体言	205,257	4.73	46.9
形容詞のみ	209,840	2.51	32.6
動詞のみ	209,969	2.54	29.2
すべての品詞	201,754	6.36	42.8

*日英対訳辞書によって変数化された単語の割合

4.2 文型パターンの調査

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. 各々の文型パターンを調査した所, 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, 精度の高い英文が得られた. 調査結果の一部を表 4 に示す. 表中の波線は自己パターンを表す.

表 4: 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**
$N1$ がやんだ。	The $N1$ has passed.	2
	The $N1$ has left off.	1
作成された英文	The rain has passed.	

5 考察

5.1 変数化の問題点

本研究では, 単語の変数化を自動的に行った. 変数化できなかった原因を探るため, ランダム 100 件の名詞 222 個を調査した. その結果, 変数化できていない名詞 111 個について検証した所, 対訳辞書に単語が載っていないので変数化に失敗している名詞が 42 個 (38%) であった.

また, 他の品詞についても調査した結果, 対訳辞書を強化することで変数化できる単語の割合が約 50%増加すると予想できる.

5.2 汎化によるパターンの同一化

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を検証したところ, 原文 215,242 件に対して同一化できたパターン数は,13,488 件と低かった. また, 日本語側のパターンで同一化できそうなパターンがあった. 例えば「$PRO1$ は $N2$ の $N3$ だ。」と「$PRO1$ が $N2$ の $N3$ である。」というパターンにおいて「は」と「が」,「だ」と「である」を汎化してパターンを同一化できる. しかし, 日本語パターンを同一化するためには, 英語側のパターンが同一化可能であるか検討する必要がある.

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. 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, 汎化による文型の同一化を行う予定である.

参考文献

[1] 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.

[2] 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.

[3] 西山ほか:単文文型パターン辞書の構築, 言語処理学会第 11 回年次大会, 発表予定, 2005-3.