012022　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

計算機工学講座　池原研究室　神野 絵理

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しなければならない．そこで，本研究では，名詞句翻訳のプロトタイプシステム (Meijin) を作成し，性能の評価を行うことを目的とする．

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．本研究では，この名詞句の日英対訳コーパスから名詞句パターンを作成する．

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，以下の自立語については和英辞書を利用し，単語対応箇所を変数化した．
数詞 (NUM)，代名詞 (PRN)，一般の名詞 (N)，用言名詞 (NS)，形容詞 (AJ)，形容動詞 (AJV)，副詞 (ADV)，連体詞 (REN)，動詞 (V)

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．

3　名詞句パターン辞書の作成

日英共に同じであるパターンをまとめて，パターン辞書とする．日本語パターンは全部で 23,834 種類あった．パターンを作るために用いたコーパスの原文の分布を調べたところ，原文が一番多く使われていたパターンが $REN1N2$ であり，その原文の数は 3,719 個であった．以下原文が 1,000 件以上であった上位 4 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 1 に示す．原文が 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．

表 1: 日本語パターンに対する英語パターンの種類

日本語パターン 例の数	英語パターン		
	1 位	2 位	その他 [種類数]
$REN1N2$ 3719 個	$PRN1$ $N2$ (87.0%)	$AJ1$ $N2$ (5.6%)	[102] (7.4%)
その $N1$ 3686 個	the $N1$ (96.9%)	his $N1$ (0.4%)	[44] (2.7%)
$PRN1$ の $N2$ 1936 個	$PRN1$ $N2$ (97.2%)	$PRN1$ true $N2$ (0.2%)	[39] (2.6%)
$N1$ の $N2$ 1224 個	the $N2$ of the $N1$ (12.5%)	$N1$ $N2$ (11.8%)	[187] (75.7%)

4　翻訳実験

4.1　翻訳手順

手順を以下に示す．
(1) 入力の日本語名詞句と日本語名詞句パターンを ATN を用いたパターンパーサ照合をする [3]．
(2) 照合結果より，適応した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．

4.2　実験内容

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．選んだ入力句から作られるパターンは，照合に用いないこととする．上述の翻訳手順に従って訳出された英語を，人手で評価する．

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，意味的に違っている，または，訳出が無い場合

4.2.1　評価方法

評価値は，再現率 R，および，適合率 P を用いる．

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

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

4.3　結果

100 個中，○が 74 個，△が 1 個，×が 25 個だった．この 25 個は，全てパターンに当たらなかった名詞句であった．したがって，$R = 75\%$，$P = 98.7\%$ となった．また，1 つの入力句に対し，訳出された英語句は，平均で 7〜8 件であった．

5　考察

他の翻訳機で同様の実験を行ったところ，ALT-J/E では，$R = 100\%$，$P = 87\%$ となり，「翻訳の王様」では，$R = 100\%$，$P = 94\%$ となった．適合率 (P) で，Meijin が良かったのは，正解 (○) が，複数解出た中に 1 つでもあればいいとしていることと，意味属性の制約が効いていることにあると考えられる．

6　おわりに

本研究では，名詞句翻訳のプロトタイプシステム (Meijin) を作成し，その翻訳精度を検証した．その結果，適合した句においては高い翻訳精度が出たと言える．しかし，再現率は 75%に留まっている．

今後は，意味属性の汎化や，名詞句パターン対を増やすことによる再現率の向上を目指す．

参考文献

[1] 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, 自然言語処理, Vol.11, No.3, pp.69-95, 2004.

[2] Brill,E.: A simple rule-based part-of-speech tagger, ANLP-92, pp.152-155, 1992.

[3] 徳久ほか: 文型パターンパーサの試作, 言語処理学会第 10 回年次研究会, pp.608-611, 2004.

[4] 神野ほか: 文型パターンにおける名詞句翻訳のためのパターン辞書の構築, 言語処理学会第 11 回年次大会, 2005（発表予定）.