022040
日英機械翻訳における多変量解析を用いた最適パターンの選択

計算機工学講座 C　　原 真一朗

1　はじめに

パターンに基づく日英機械翻訳方式の実現に向けて, 単語レベル, 句レベル, 節レベルの 3 レベルで構成された大規模な日英対訳パターン辞書が構築された (各レベルの規模は 12 万, 9.5 万, 1.2 万パターンである)[1]. 大規模なパターン辞書を用いると, 入力文に適合するパターン数が多いため, 翻訳に適したパターンを選択する必要がある [2]. 最適なパターンを選択する方法として, 単語レベルのパターンを対象とした場合, 多変量解析を用いた方法が有効である [3]. しかし句レベルのパターンを対象とした場合, 適合するパターン数が単語レベルに比べて非常に多いため, 同様の方法は効果が不明である.

そこで, 本研究では句レベルの場合について多変量解析による選択方式を適用し, 有効性を調査する.

2　多変量解析による最適パターン選択方式

2.1　適合パターン

パターン翻訳方式では, パターンパーサが入力日本語文に適合するパターン全てをパターン辞書から検索する [4]. 適合パターンには英語パターンや表 1 に示すデータが付随している.

2.2　多変量解析による選択

次の評価関数を使用して, 適合パターンに評価値 y を与える. この評価値で最適な適合パターンを定める.

$$y = a + \sum_{n=1}^{9} b_i x_i$$

評価パラメータ x_i について表 1 に示す. x_1〜x_7 は [3] に従った. x_8, x_9 は句の変数が存在するため新たに追加した. a および b_i の回帰係数は, 評価値 y と評価パラメータ x_i の事例から, 重回帰分析により求める.

表 1　評価パラメータ

x_1	パターン (P) 適合率	x_6	名詞平均意味属性距離
x_2	P 字面適合率	x_7	動詞平均意味属性距離
x_3	P 元字面適合率	x_8	名詞句平均意味属性距離
x_4	記号の適合率	x_9	動詞句平均意味属性距離
x_5	変数の適合率		

3　回帰係数の設定

3.1　評価パラメータ x_i

句レベルパターンを作成した際に使用した原文から 55 文を選択し, 各文から適合パターンを収集した (入力文から作られたパターンは除く). 1 入力文につき最大 10 件の適合パターンを収集し, 492 件を得た. 各件につき評価パラメータの値を求めた.

3.2　評価値 y

492 件の適合パターンから英語文を作成し, 次の評価基準で適合パターンに評価値 y を与えた.

評価 A	1.00	情報や文法に問題がない
評価 B	0.66	重要でない情報が欠如している
評価 C	0.33	入力文を部分的に訳せている
評価 D	0.00	入力文の訳としては使用不可能

3.3　回帰係数の作成結果

3.1 節, 3.2 節の結果より回帰係数を得た.

$$y = -0.566 + 0.611x_1 - 0.061x_2 + 0.195x_3 - 0.014x_4 + 0.143x_5 + 0.108x_6 + 0.309x_7 - 0.024x_8 + 0.093x_9$$

4　適合パターン選択の実験

4.1　目的と調査対象

本実験は適合パターン選択の精度を求める. その精度は, 評価関数の精度に対応する. 評価関数による適合パターンの推定の評価値と, 適合パターンの英訳に対する評価値との差が小さいほど, 評価関数の精度が良い. 3.1 節の方法集めた新たな 35 文を対象とする.

4.2　実験手順

各対象文に対して以下を行う.

1. 対象文の適合パターンをパターンパーサで検索
2. 評価関数を適用 (推定値を得る)
3. 推定値の最も高い適合パターンを選択
4. 適合パターンから英文を作成
5. 3.2 節と同一の基準で英文を評価

4.3　実験結果

推定値と評価値の差が 0.3 未満ならば「合」とした. 推定精度は 77%(合の数/推定数 = 27/35) であった. また, 手順 3 において, 上位 10 位までの推定値について推定精度を求めたところ, 79%(270/341) であった. 推定値をもとに適合パターンを選択し, そこから英文を生成すると, その推定値の品質で英文の得られる可能性が 79% であることを意味する.

5　考察

本実験では各入力文に対して適合するパターンが 10 件より多い場合や, ランダムで適合パターンを選んだ場合の調査は行っていない. 今後は調査文を増やし, 精度の評価をより精密にする必要がある.

6　おわりに

本研究では多変量解析による選択方式を句レベルに適用し, 最適な文型パターンの選択を行った. 推定値の上位 10 位について 79% という選択精度を得た.

参考文献

[1] 池原悟：等価的類推思考の原理による機械翻訳方式, 信学技報, TL2002-34, pp.7-12, 2002.
[2] 池原悟, 徳久雅人, 竹内 (村本) 奈央, 村上仁一: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, Vol.11, No.4, pp.147-178, 2004.10
[3] 岡田敏:多変量解析による最適文型パターンの選択方式, 言語処理学会年次大会, pp.25-28, 2005
[4] 徳久雅人, 池原悟, 村上仁一:文型パターンパーサの試作, 言語処理学会年次大会, pp.608-611, 2004
[5] 池原悟:非線形な言語表現と文型パターンによる意味の記述, 情処研報,2004-NL-159,pp.139-146, 2004-1