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

原 真一朗　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
{s022040,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

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重文複文を対象とする大規模な日英対訳パターン辞書が構築された [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の選択方式

2.1　文型パターン辞書からの適合パターン検索

文型パターン辞書には翻訳対象である日英両言語の文型パターンが収録されており, その規模は単語レベル 12 万パターン, 句レベル 9.5 万パターン, 節レベル 1.2 万パターンである. すでに, 入力文と文型パターンを照合し, 入力文に適合する文型パターンを検索する文型パターンパーサが試作されている [5]. 文型パターンパーサは入力文に対して適合する全てのパターンを出力する. 句レベルでの適合文型パターン数の平均は 164.72 件である.

2.2　多変量解析による選択

多変量解析は複雑なデータを解析して有効な情報を見つけるための統計的な手法である. 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. 評価値によって最適な適合パターンを定める.

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

x_1〜x_7 は単語レベルについての評価パラメータ [3] に従った. 句レベルの実験ではパターン中に句の変数が存在するため x_8, x_9 を新たに追加した. 切片 a および回帰係数 b_i は, 評価値 y と評価パラメータ x_i の事例から, 多変量解析の重回帰分析により求める. 評価パラメータ x_i について以下で説明を行う.

- x_1:パターン適合率
  入力文の文字数のうち適合パターンに対応した文字の割合である. ただし単語単位で計算し, パターンに適合している単語と入力文の単語で適合している単語の総文字数の除算で求める. 以下の例では, “彼” と “$N1$”, “頭を打っ” と “$VP2$”, “喪失した” と “$V3.kako$” が対応しており, パターン適合率は 0.8(11 文字/14 文字) である.
  入力文
  彼 は 頭を打っ て 記憶 を 喪失した.
  適合パターン
  /$N1$ は $VP2$ て /$V3.kako$.

- x_2:パターン字面適合率
  入力文と, 適合パターンに存在する字面が入力文に占める割合である. 単語単位で計算し, 一致単語数と入力文の総単語数の除算で求める. 以下の例ではパターン字面適合率は 0.3(3 単語/10 単語) である.
  入力文
  彼 は 頭を打っ て 記憶 を 喪失した.
  適合パターン
  $N1$ は $VP2$ て /$N3$ を /$V4.kako$

- x_3:パターン元字面適合率
  適合パターンを作成する際に用いた原文 (以降単に「原文」と呼ぶ) と, 入力文とを比較し, 共通する字面の一致する割合である. 単語単位で計算し, 一致単語数と入力文の総単語数の除算で求める. 以下の例ではパターン字面適合率は 0.5(5 単語/10 単語) である.
  入力文
  彼 は 頭 を 打っ て 記憶 を 喪失し た.
  適合パターンの原文
  裁判官 は 公訴 を 却下し て 被告人 を 放免し た.

- x_4:記号の適合率
  適合パターン中の記号が使用される割合で, 表 1 の記号を用いる. 以下の例では 1.0(2 個/2 個) である.
  入力文
  彼 は 頭 を 打っ て 記憶 を 喪失した.
  適合パターン
  $1^{$N1 は }$VP2(て/で) $1/$N3$ を$1/$V4.kako$

表 1　要素記号の一覧

記号名	表記	意味
選択記号	(… | …)	いずれかの要素列と適合
任意記号	[…]	文型選択上, 任意の要素
補間記号	<…>	ゼロ代名詞等
順序任意指定記号	{… | …}	順序入れ換え可能な範囲 (例 格要素の順序)
位置変更可能指定記号	$n^{ 定義 } $n	指定位置に入れ換え可能 (例 副詞の位置)

- x_5:変数の適合率
  適合パターンに含まれる変数が, 入力文との適合に使用される割合である. 変数が複数ある場合, 異なる品詞ごとにパターンに含まれる数を調べ, 入力文で使用される割合の平均を求める. 以下の例では $N(2/3)$, $VP(1/1)$, $V(1/1)$ より平均適合率 0.9 である.
  入力文
  彼 は 頭 を 打っ て 記憶 を 喪失した.
  適合パターン
  $N1$ は / [$N3$ の] $N4$ を/$V5$ て /$V6.kako$.

- x_6:名詞の平均意味属性距離の逆数
- x_7:動詞の平均意味属性距離の逆数
  入力文と, 適合パターンの原文との間で, 変数を介して対応する名詞, 動詞箇所に関して意味属性距離を調べ, 平均値の逆数を使用する. 意味属性は, 日本語語彙大系 [6] に記載された「一般名詞意味属性体系」および「用言意味属性体系」を使用する. 平均意味属性距離が 0 の場合は 1 とする. 以下の例では “彼 [他称 (単数/男)]” と “裁判官 [裁判官]” の意味属性距離が 8 より逆数は 0.1 である.
  入力文
  彼 は 頭 を 打っ て 記憶 を 喪失し た.
  適合パターン
  $N1$ は $VP2$ て $N3$ を $V4.kako$.
  適合パターンの原文
  裁判官 は 公訴 を 却下し て 被告人 を 放免し た.

- x_8:名詞句の平均意味属性距離の逆数
- x_9:動詞句の平均意味属性距離の逆数
  入力文と, 適合パターンの原文との間で, 変数を介して対応する名詞句, 動詞句箇所に関して, 意味の中心と考えられる名詞, 動詞の意味属性距離を調べ, 平均値の逆数を使用する. 意味属性距離が 0 の場合は 1 とする. 以下の例では変数 VP2 を介して “頭を打って” と “公訴を却下して” が対応しており, “打って [身体動作]” と “却下して [思考動作]” が意味の中心と考えられる. 意味属性距離は 4 より逆数は 0.3 である.
  入力文
  彼 は 頭を打っ て 記憶 を 喪失し た.
  適合パターン
  $N1$ は $VP2$ て $N3$ を $V4$.kako.
  適合パターンの原文
  裁判官 は 公訴を却下し て 被告人 を 放免し た.

3　回帰係数の設定

3.1　回帰係数の作成条件

句レベルパターンの原文 55 文を入力文として選択し, 各文の適合パターンを収集した. ただし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使用することで 492 件を得た. 英文の生成は文型パターンパーサの出力から, 入力文に適合した日英の文型パターンを用いて, 人手により生成を行う.

3.2　評価パラメータ x_i と評価値 y の事例

各適合パターンについて付随する情報より, x_1〜x_9 の評価パラメータの値を求めた. 次に評価関数を求める. 英文の生成は, 適合したパターンに対応する英語パターンを用いて, 人手により行う. その英文の品質を評価する. 評価は以下の A〜D の 4 段階で行い, 評価関数作成の際には評価に応じた値を使用し, 適合パターンの評価値とする.

評価 A : 1
　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が簡単に修正可能
評価 C : 0.33
　入力文を部分的に訳せている
評価 D : 0
　入力文の訳としては使用不可能

作成英文の評価をを表 2 に示す.

3.3　回帰係数の作成結果

前節で求めた評価パラメータと評価値から, 重回帰分析によって回帰係数を求める. 回帰係数から求めた評価関数を式 2 に示す.

$$y = -0.566 + 0.611x_1 - 0.061x_2 + 0.195x_3 - 0.014x_4 + 0.143x_5 + 0.108x_6 + 0.309x_7 - 0.024x_8 + 0.093x_9 \quad \text{(式 2)}$$

式 2 から, 評価パラメータ x_1(パターン適合率) の回帰係数が最も高い値となっていることがわかる.

表 2　各評価における作成訳の例

評価 A	
入力文	**彼にはその任務を果たせるだけの能力がなかった**
対訳	He was not equal to the task.
適 P(日)	$N1$ には!$VP2$^$rentai$ だけの/$N3$ が /(無かっ｜なかっ) た.
適 P(英)	$N1$ had little $N3$ to $VP2$.
作成訳	**He had little ability to fulfill the duty.**

評価 B	
入力文	**この辺には気の利いた料理屋がない**
対訳	**There are no respectable restaurants in this neighbourhood.**
適 P(日)	/< $N1$ は>/$NP2$ が/$AJ3$.
適 P(英)	$N1$ $AJ3$ $NP2$.
作成訳	**No respectable restaurants in this neighbourhood.**

評価 C	
入力文	**将来は建築関係に進みたいと思っている**
対訳	I'm thinking about going into architecture in the future.
適 P(日)	$1^{$N1 は }!$VP2$.*tai* と$1/思っている.
適 P(英)	$N1$ hope to $VP2$.
作成訳	**In the future hope to going into architecture.**

評価 D	
入力文	**警官が来て騒ぎを鎮めた**
対訳	The policemen came and got the things under control.
適 P(日)	$N1$ が $VP2$(て｜で)/$N3$ を/$V4$.*kako*.
適 P(英)	$N1$ $VP2$.*past* across $N3$.
作成訳	**The policemen came across the things.**

4　評価関数を用いた適合パターン選択実験

4.1　実験目的と調査対象

本実験では適合パターン選択の精度を求め, 多変量解析の選択方式の有効性の確認を行う. 多変量解析による最適パターン選択の目標は, 評価関数による適合パターンの推定の評価値 (推定値) に基づき選択した結果, 英文として許せる評価の範囲に含まれることである. オープンテストでは, 3.1 節と同様の方法で新たに集めた 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　オープンテスト

入力文 35 文から前節の手順 1 により 341 件の適合パターンを得た. 手順 2 で各件の適合パターンの推定値を算出する. 手順 3 で各入力文における最大 10 件の適合パターンの中で, 最も推定値の高いパターンを選択する. 手順 4 では手順 3 で求めた適合パターンを用いて英文を作成する. 最後に手順 5 で英文を 4 段階に評価する.

4.4　実験結果

各入力文で推定値上位 1 位の英文が評価 A または B の場合, 適合パターンの選択に成功とし, その適合パターンを正解適合パターンとする. 評価 C または D であるものを不正解適合パターンとする. 推定値を 0.1 の階級区間に分け, 各区間において推定値の 1 位を持つ入力文が, 正解適合パターンを持つ割合を調べた. 55 文のクローズドテスト結果を図 1 に, 35 文のオープンテスト結果を図 2 に示す. 実線が正解適合パターン正解率を表し, 点線は評価の条件を変えた場合の正解率を表す. 図より, 推定値が 0.7 以上のときは正解適合パターン選択の正解率が 50% 以上で行えると推測できる. また推定値が 0.4 未満のときは正解適合パターン選択の正解率は 0% である.

図 1　推定値 1 位の正解率 (クローズドテスト)

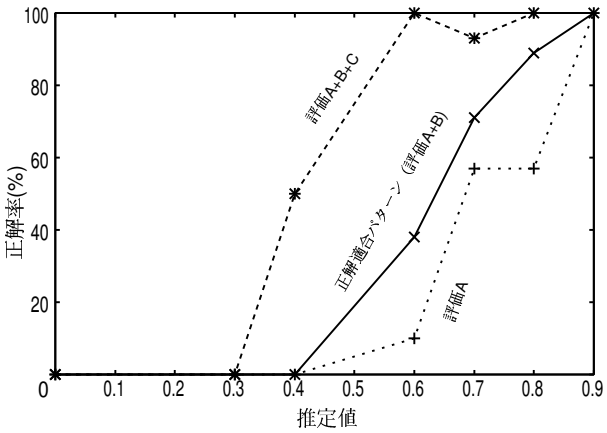
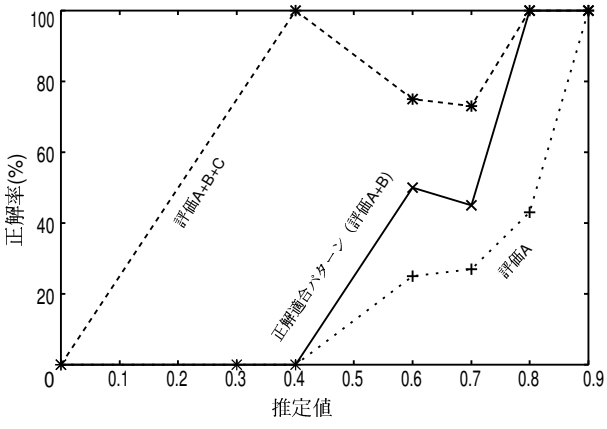


図 2　推定値 1 位の正解率 (オープンテスト)



5 考察

5.1 推定値の信頼性

入力文 35 文のうち, 評価 A を最低 1 つ持つ文は 22 文 (212 件) あった. その 22 文に対して推定値が上位 1 位の適合パターンの評価を調べてみたところ, 11 文 (50%) において評価 A の適合パターンを選択した. また 18 文 (82%) において評価 A または B を選択した (表 3).

表 3 信頼性

	評価 A	評価 A または B
1 位	50%(11/22 文)	82%(18/22 文)
1 位〜2 位	68%(15/22 文)	95%(21/22 文)
1 位〜3 位	73%(16/22 文)	100%(22/22 文)
1 位〜4 位	100%(22/22 文)	100%(22/22 文)

5.2 評価パラメータの考察

5.2.1 動詞意味属性距離について

推定値の上位 1 位が不正解適合パターンを選択したパターンについて, 各評価パラメータを調査したところ, 動詞意味属性距離の値が近いという共通の特徴があった. 動詞の意味属性距離が近いことで推定値の増加が起こり, 不正解適合パターンが推定値の上位になった.

5.2.2 寄与率

評価関数の寄与率を求めた. 寄与率によってパラメータの重要性がある程度得られる. 作成した評価関数に加え, 各評価パラメータ単独で回帰分析し, 寄与率を求めた結果を表 4 に示す.

表 4 寄与率の比較

本稿作成評価関数	15.7%
パターン適合率のみ	**8.5%**
パターン字面適合率のみ	0.2%
パターン元字面適合率のみ	1.1%
変数の適合率のみ	0.8%
記号の適合率のみ	0.7%
名詞意味属性のみ	1.2%
動詞意味属性のみ	**5.5%**
名詞句意味属性のみ	0.9%
動詞句意味属性のみ	0.5%

表 5 より, パターン適合率 (8.5%) と動詞の意味属性距離 (5.5%) について寄与率が高い. 評価関数に与える動詞意味属性距離の影響が大きいことが, 不正解適合パターンが推定値の上位となった原因と考えられる.

5.3 推定精度

評価値が離散的であるため単純な相関係数が使えないと考えて, 推定値と評価値の差が 0.33 未満 (評価の差が 1 つ未満) 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, 各文で上位 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.

表 5 実験結果

値の差	推定値 1 位	推定値上位 10 位
0.33 未満 (合)	77%(27/35 文)	79%(270/341 件)
0.66 未満	20%(7/35 文)	18%(63/341 件)
0.66 以上	3%(1/35 文)	2%(8/341 件)

6 おわりに

本研究では句レベルについて多変量解析による選択方式を用いることで, 最適な文型パターンを選択した. 実験の結果, 推定値が 0.7 以上のとき正解適合パターンを選択する正解率が 50% 以上となった. また, 推定値が 0.4 未満のときは正解適合パターン選択の正解率は 0% である. 今後は入力文と適合パターン数の収集量の増加による選択精度の調査と精度の向上が課題である. また節レベルについて多変量解析による選択方式が有効であるかを適用する必要がある.

謝辞

本研究は, 科学技術振興事業団「JST」の戦略的基礎研究推進事業「CREST」における研究領域「高度メディア社会の生活情報技術」の研究課題「セマンティックタイポロジーよる言語の等価変換と生成技術」の支援により行った.

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.

[3] 岡田敏, 村上仁一, 徳久雅人, 池原悟: 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発表論文集, pp.25-28, 2005.

[4] 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, 情報処理学会, 自然言語処理研究会,2004-NL-159,pp.139-146, 2004.

[5] 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発表論文集, pp.608-611, 2004.

[6] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林: 日本語語彙大系, 岩波書店, 1997.

[7] 上田太一郎, 苅田正雄, 本田和恵; 実践ワークショップ Excel 徹底活用 多変量解析, 秀和システム, 2003.