多変量解析による最適文型パターンの選択方式

岡田 敏　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
{sokada,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

1　はじめに

近年, 機械翻訳の方式として等価的類推思考の原理に基づく機械翻訳方式が提案されている [1]. この方式の実現に向けて, 日本語の重文・複文を対象とした文型パターンを大量に蓄積した文型パターン辞書の構築が進められている [2]. 現在, 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抽出する文型パターンパーサが試作されている [3].

この文型パターンパーサを使用すると, 複数の文型パターンが入力文に適合する場合が多く, 入力文の意味を考えると適合した文型パターンの中には英文生成に使用できないパターンも含まれる. よって, 入力文の訳を出力するために適合パターンの選択が必要となる.

そこで本稿では, 文型パターンパーサが出力した複数の適合パターンから, 入力文の翻訳が可能な適合パターンの選択手法を検討する. 具体的には, 入力文に適合した文型パターンを多変量解析によって分析し, 評価関数を求める. 評価関数を使用して適合パターンの得点を求め, 英文生成に使用する適合パターンの選択を行う.

2　文型パターン

2.1　文型パターンの記述形式

文型パターンは可読性と網羅性を意識して設計されており, 字面, 変数, 関数, 記号で記述されている [4]. 日英文対応の対訳コーパスの原文を, 単語レベル, 句レベル, 節レベルにパターン化した構造を持つ. 各レベルの粒度でアライメントが取れた部分は, 線形要素として変数化されている. また, 変数化すると対訳の訳出が困難になる部分は変数化されず, 非線型要素として字面, あるいは関数の形式で残されている.

以下に原文 (L) と単語レベルパターン (W) を示す. 文型パターンは, 日本語文型パターン (WJ) と, 対応する英語文型パターン (WE) で記述されており, 変数を介して両言語の要素対応付けができる.

LJ :将来は作家になりたいと思っている.
LE :I want to become a writer in the future.
WJ:$TIME1$ は/$N2$ に/$V3.tai$/と思っている.
WE:I want to $V3$ $N2$ in $TIME1$.

変数には名詞や動詞の単語を表す Nn や Vn など 8 種類がある. 関数には *.tai* や*.kako* などがあり, 字面の指定や表現の統括を行う. 記号はパターン記述要素の適合の仕方について, 任意化, 選択, 順序変更などの制御を行う (表 1).

表 1: 要素記号の一覧

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

2.2　文型パターン照合

文型パターンの照合では, 対訳文型パターン辞書から入力文に適合する文型パターンを全て検索する. 文型パターンの必須要素が指定通りに入力文と対応すれば, 適合パターンとして出力する. 文型パターンの適合の仕方が複数ある場合は複数個出力する.

2.3　文型パターンを利用した英文生成

文型パターンを利用した英文生成は, 適合パターンに対応する英語文型パターンを使用する [5]. 日本語文型パターンの変数と対応する入力文の箇所を翻訳し, 英語文型パターンの対応する箇所と置換することで英文を生成する. 置換の際, 英語パターンの記述に沿う形に単語を変形する. 単語レベルの文型パターンを利用した英文生成の例を示す.

入力文
　将来は作家になりたいと思っている.
適合パターン 1
　$WJ1$：$TIME1$ は/$N2$ に/$V3.tai$/と思っている.
　$WE1$：I want to $V3$ $N2$ in $TIME1$.
作成訳 1
　I want to become a writer in the future
適合パターン 2
　$WJ2$：$N1$ は/$N2$ に/$V3.tai$/と思っている.
　$WE2$：$N1$ be thinking of $V3.ing$ to $N2$.
作成訳 2
　The future is thinking of becoming to a writer.

この例では, 入力文に 2 種類の文型パターンが適合している. しかし, 適合パターン 2 では品質の悪い英文しか生成できない. 適合パターンに対応する英語文型パターンを使用しても, 必ずしも良質の英文を生成できるとは限らない. よって, 品質の良い翻訳が可能な適合パターンの選択が必要となる.

3　多変量解析による評価関数の作成

3.1　本稿の目的

本稿では, 複数の適合パターンの中から, 英文生成に用いる適合パターンを一意に選択する. まず, テスト入力文の適合パターンを多変量解析によって分析し, 評価関数を求める. 次に, 得られた評価関数で適合パターンの得点を求め, 英文生成に使用する適合パターンの選択を行う.

3.2　意味属性大系

適合パターン選択評価関数を求める際に, 名詞, 動詞の意味属性を使用する. 入力文と, 適合パターンの元となった原文との間で, 変数を介して対応する箇所の名詞, 動詞意味属性距離を調べ, 評価関数を求める際にパラメータとして使用する. 意味属性は, 日本語語彙大系 [6] に記載されている「一般名詞意味属性大系」および「用言意味属性大系」を使用する.

一般名詞意味属性大系は, 名詞の意味的用法に着目してシソーラスとして大系化されている. 登録単語数約 40 万語, 最大 12 段の木構造であり,2710 の意味分類に分類されている. また, 各ノードにおいて上位の意味属性の性質を下位の意味属性が継承する.

用言意味属性大系は, 動詞の意味的用法に着目してシソーラスとして大系化されている [7]. 約 6000 語, 最大 4 段の木構造であり,35 の意味分類に分類されている. また, 各ノードにおいて上位の意味属性の性質を下位の意味属性が継承する.

3.3 評価関数

まず, 適合パターンを使用してテスト入力文を人手で英文に翻訳する. 次に, 得られた英文の品質を各適合パターンの評価値とし, 入力文と適合パターンの関係からパラメータを抽出する. 最後に, テスト入力文の適合パターンのパラメータと評価値を多変量解析によって分析し, 評価関数を求める.

適合パターンの以下のパラメータを評価パラメータとする. 評価パラメータを回帰分析することで評価関数を求める. 評価関数を $\hat{y}$(式 1) とし, 評価値 y との残差 $e(e = y - \hat{y})$ の 2 乗の総和を最小にする回帰係数 $b_1, ..., b_7$ と a の値を求める.

＜評価関数＞

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_1x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 \text{ (式 1)}$$

＜評価パラメータ＞

- y:評価値
  適合パターンに対応する英語文型パターンを使用して, 入力文を人手で翻訳する. 作成した英文の品質で適合パターンを評価する. 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する.
  評価 A : 1
  品質の高い英文が生成できる
  評価 B : 0.66
  重要ではない要素の欠如はあ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

- x_1:パターン適合率
  入力文と, 適合パターンの文字単位の一致する割合をパターン適合率とする. 単語単位で計算し, パターンに適合している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," 道路" と"$N1$"," 横断する" と"$V2$^rentai"," 注意しなさい" と"$V3$^meirei" が対応しており, 入力文の全ての要素がパターンに適合している. よってパターン適合率は 1.00(16 文字/16 文字) である.
  入力文
  道路 を 横断する と き は 注意しなさい.
  適合パターン
  $N1$ を /$V2$^rentai と き は /$V3$^meirei.

- x_2:パターン字面適合率
  入力文と, 適合パターンに共通する字面の一致する割合をパターン字面適合率とする. 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 0.43(3 単語/7 単語) である.
  入力文
  道路 を 横断する とき は 注意しなさい.
  適合パターン
  $N1$ を /$V2$^rentai とき は /$V3$^meirei.

- x_3:パターン元字面適合率
  入力文と, 適合パターンを作成する際に用いた原文に共通する字面の一致する割合をパターン元字面適合率とする. 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 0.71(5 単語/7 単語) である.
  入力文
  道路 を 横断する とき は 注意し なさい.
  適合パターンの原文
  線路 を 渡る とき は 注意し なさい.

- x_4:記号の適合率
  適合パターンに含まれる要素記号が使用される割合を記号の適合率とする. 表 1 の記号を対象にする.

- 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を変数の適合率とする.

- x_6:名詞の平均意味属性距離の逆数
  入力文と, 適合パターンの元となった原文との間で, 変数を介して対応する名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. 平均意味属性距離が 0 の場合は 1 とする. 以下の例では, 名詞の平均意味属性距離の逆数は 0.25(1/4) である.
  入力文
  道路 を 横断する とき は 注意し なさい.
  適合パターン
  $N1$ を $V2$^rentai とき は $V3$^meirei.
  適合パターンの原文
  線路 を 渡る とき は 注意し なさい.

- x_7:動詞の平均意味属性距離の逆数
  入力文と, 適合パターンの元となった原文との間で, 変数を介して対応する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. 平均意味属性距離が 0 の場合は 1 とする. 以下の例では, 動詞の平均意味属性距離の逆数は 1(平均意味属性距離 0 のため) である.
  入力文
  道路 を 横断する とき は 注意し なさい.
  適合パターン
  $N1$ を $V2$^rentai とき は $V3$^meirei.
  適合パターンの原文
  線路 を 渡る とき は 注意し なさい.

3.4 評価関数作成の実験条件

入力文には, 対訳文型パターン集を作成した際に使用した原文約 12 万文からランダムに 200 文を選び, テスト入力文として評価関数作成に使用する. ただし, 入力文から作成した文型パターンが適合した場合は実験に使用しない.

文型パターンパーサは $wjpp.ver$.2.4[3] を使用する. 英文の生成は文型パターンパーサの出力を人手で修正する. 回帰分析には MicrosoftOffice の回帰分析ツールを使用する.

3.5 評価関数作成結果

テスト入力文 200 文のうち 72 文に適合パターンが存在した. 1 入力文に対して平均 26 パターンの文型パターンが適合した. 各入力文毎に最大 30 パターンまで調査し,765 パターンを評価関数作成に使用した. 得られた評価関数を (式 2) に示す.

$$\hat{y} = -0.403 + 0.122x_1 - 0.194x_2 + 0.498x_3 + 0.027x_4 + 0.208x_5 + 0.195x_6 + 0.130x_7 \quad \text{(式 2)}$$

式 2 から, 評価パラメータ x_3(パターン元字面適合率) の回帰係数が最も高い値となっていることがわかる.

3.6 各評価における作成訳の例

評価 A から D の出力英文の例を表 2 に示す. 各評価毎に, 入力文, 翻訳例, 適合パターン (適 P), 作成訳を示す.

表 2: 各評価における作成訳の例

評価 A 品質の高い英文が生成できる	
入力文	**愛情を持続させることは難しい**
翻訳例	It is hard to keep love alive.
適 P(日)	$N1$ を/($V2.sase\hat{}rentai$ | $V2\hat{}sase\hat{}rentai$)! ことは/難しい.
適 P(英)	It is quite impossible to $V2$ all $N1$.
作成訳	**It is quite impossible to continue all love.**

評価 B 重要ではない要素の欠如はあるが簡単に修正可能	
入力文	**彼女は彼にすぐ行くよう命じた**
翻訳例	She commanded him to go at once.
適 P(日)	\$1\^{}{$N1$ は }/$N2$ に\$1/$V3.suitei$\$1 /$V4.kako$.
適 P(英)	$N1$ $V4.past$ $N2.obj$ to $V3$.
作成訳	**She commanded him to go.**

評価 C 入力文を部分的に訳せている	
入力文	**これは今まで使ったなかでいちばん おもしろい辞書だ**
翻訳例	This is the most interesting dictionary I have ever used.
適 P(日)	#1[その/辺で]/いちばん/$AJ2\hat{}$rentai!$N3.da$.
適 P(英)	It is $AJ2.st$ $N3$ #1[thereabout].
作成訳	**It is most interesting dictionary.**

評価 D 入力文の訳としては使用不可能	
入力文	**母は赤ん坊をあやして笑わせた**
翻訳例	Mother played with the baby and got him to smile.
適 P(日)	\$1\^{}{$N1$ は }/$N2$ を/$V3$(て | で)\$1 /$V4.kako$.
適 P(英)	N1 be.past V4.ed at V3.ing N2.
作成訳	**Mother was made laugh at amusing baby.**

4 適合パターン選択実験

4.1 選択関数の評価

得られた評価関数を使用して, 品質の高い英文を作成できる適合パターンの選択を行う. 得られた評価関数に適合パターンの情報を代入し, 各適合パターンの評価関数の値 ($\hat{y}$) を求める. 評価 A,B の適合パターンを正解適合パターンとし, 各入力文毎に第 8 位までの累積正解率で関数を評価する.

テスト入力文 200 文を使用してクローズドテストを行う. オープンテストには, 対訳文型パターン集を作成した際に使用した原文約 12 万文から, テスト入力文以外の 200 文をランダムに抽出して使用する.

4.2 入力文の比較

クローズドテスト, オープンテストで使用する入力文の比較を表 3 に示す. 調査対象入力文は適合パターンを持つ入力文の数である. 本稿では各入力文ごとに 30 パターンまで調査に使用したため, 総適合パターン数と調査パターン数は異なる. ランダム選択における平均正解含有率は以下の式で求める.

$$\text{平均正解含有率} = \frac{\sum \text{各入力文における正解パターンの割合 (\%)}}{\text{調査対象入力文の数}} \quad \text{(式 3)}$$

表 3 の調査対象入力文, 総適合パターン数, 平均適合パターン数を比較すると, クローズドテストに使用する入力文はオープンテストに使用する入力文より, 多くの文型パターンに適合する入力文が多いとわかる.

表 3: 入力文の比較

	クローズド	オープン
入力文数	200 文	200 文
調査対象入力文	72 文	82 文
総適合パターン数	1847 パターン	1716 パターン
平均適合パターン数	26 パターン	21 パターン
調査パターン数	765 パターン	680 パターン
正解適合パターンを持つ入力文の数	29 文	29 文
正解パターンの割合	13%(96/765)	14%(95/680)
平均正解含有率	17%	23%

4.3 適合パターン選択実験結果

適合パターン選択実験を行った結果を表 4 に示す.

表 4 より, クローズドテストでは 72%, オープンテストでは 83%の入力文において, 第 1 候補に正解適合パターンが存在した. また, ほぼ全ての入力文において第 8 候補までに正解適合パターンが存在した.

表 4: 実験結果

候補	クローズドテスト	オープンテスト
第 1 候補	72%(21/29)	83%(24/29)
第 2 候補	86%(25/29)	90%(26/29)
第 4 候補	90%(26/29)	100%(29/29)
第 8 候補	97%(28/29)	100%(29/29)

5 考察

5.1 適合パターン選択失敗の原因

不正解適合パターンが第 1 候補にある入力文の大部分は, 大量の文型パターンに適合していた. 入力文に大量の文型パターンが適合するとき, 入力文と適合パターンおよび原文から得られる表面的な情報に差が少なかった. 不正解適合パターンの中にも, 名詞・動詞の平均意味属性距離が小さい適合パターンがある. よって, 正解適合パターンを第 1 候補にできなかったと考えられる. 表 2 の評価 D の入力文には文型パターンが 204 パターン適合するが, 以上の理由により正解適合パターンの選択は困難であった.

多くの文型パターンに適合する入力文がオープンテストには少なかったため, オープンテストの結果がクローズドテストの結果より良い値を示したと考えられる.

5.2 適合パターンの最大値と評価

各入力文ごとに, 評価関数による値が最大の適合パターンを調べ, 正解・不正解の関係を調査した. クローズドテストの結果を図 1 に, オープンテストの結果を図 2 に示す. 図より, 適合パターンの最大値が 5.0 以上の入力文は, 最大値を持つ適合パターンで入力文を翻訳できると推測できる. 適合パターンの最大値が 0.1 以下の入力文は, 最大値を持つ適合パターンでは入力文の翻訳は困難であると推測できる.

評価関数による値に閾値を設定し, 第 1 候補の適合パターンが正解・不正解か判別できるか調査した. 調査方法は, 第 1 候補の適合パターンが正解・不正解か正しく判別できる入力文の数を調べ, 適合パターンが存在した入力文の数で除算する. クローズドテストの結果を表 5 に, オープンテストの結果を表 6 に示す.

閾値を 0.4 に設定すると, 正解適合パターンを正しく判断できた入力文の割合は低いが, 不正解適合パターンを全て正しく判断できた. よって, 評価関数の信頼値は最も高い値を示した.

図 1: 適合パターンの最大値と評価の関係 (クローズド)

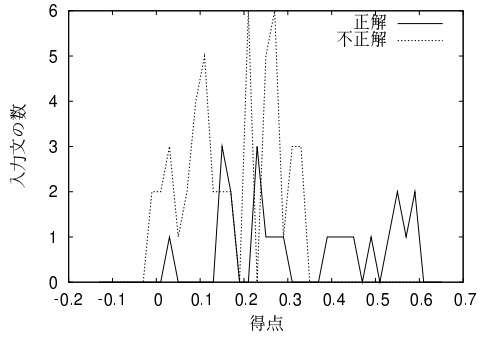

図 2: 適合パターンの最大値と評価の関係 (オープン)

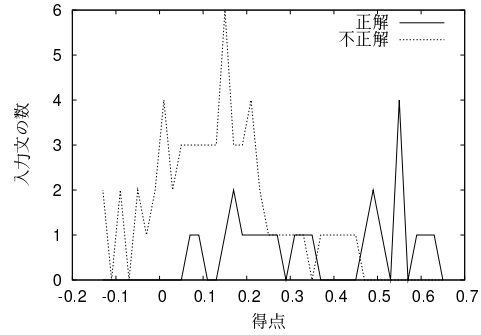

表 5: 評価関数の信頼値 (クローズドテスト)

閾値	信頼値	正しく判断できた割合	
0.6	60%	正 0%(0/29)	不 100%(43/43)
0.5	68%	正 21%(6/29)	不 100%(43/43)
0.4	73%	正 34%(10/29)	不 100%(43/43)
0.3	66%	正 38%(11/29)	不 86%(37/43)
0.2	50%	正 59%(17/29)	不 44%(19/43)

表 6: 評価関数の信頼値 (オープンテスト)

閾値	評価値	正しく判断できた割合	
0.6	67%	正 7%(2/29)	不 100%(53/53)
0.5	74%	正 28%(8/29)	不 100%(53/53)
0.4	74%	正 38%(11/29)	不 94%(50/53)
0.3	74%	正 48%(14/29)	不 89%(47/53)
0.2	67%	正 62%(18/29)	不 70%(37/53)

5.3 評価関数の説明力

本稿では評価関数を評価値 y と評価パラメータ x の関係から求めている．しかし, データのばらつきが大きければ評価関数の信頼性が低くなる問題がある．そこで, 得られた評価関数が全体のデータのばらつきをどの程度説明しているか調査した．

評価関数の信頼性は自由度修正済 R^2 値によって判断する．自由度修正済 R^2 値は, 評価値のデータの変動が評価パラメータの変動でどの程度説明できるか表している．よって, 自由度修正済 R^2 値 (R'^2) で評価関数の寄与率 (説明力) がわかる．寄与率は, 以下の式で求めた．

$$\text{寄与率 (\%)} = R'^2 \times 100 \qquad \text{(式 4)}$$

$$R'^2 = 1 - \frac{S_E/(n-p-1)}{S_T/(n-1)} \qquad \text{(式 5)}$$

$$S_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \text{(式 6)}$$

$$S_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \text{(式 7)}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad \text{(式 8)}$$

p:評価パラメータの数 (8)
n:評価関数作成に使用した適合パターン数 (765)
$\hat{y}$:評価関数の値
y:評価値

本稿で作成した評価関数の寄与率と, 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 7 に示す．本稿で作成した評価関数は適合パターンの選択に効果があったが, 寄与率は 16.5%と低い値であった．寄与率の値から, 適合パターン選択に必要な情報を最も持つ評価パラメータはパターン元字面適合率であると考えられる．

表 7: 寄与率の比較

関数	寄与率
本稿作成評価関数	16.5%
パターン適合率のみ	7.1%
パターン字面適合率のみ	5.7%
パターン元字面適合率のみ	11.4%
変数の適合率のみ	2.7%
記号の適合率のみ	1.8%
名詞意味属性のみ	3.5%
動詞意味属性のみ	4.7%

5.4 パターン元字面適合率

単回帰分析の結果, パターン元字面適合率の寄与率が最も高かった．パターン元字面適合率と評価値の関係から評価関数を求め, 適合パターン選択実験を行った．評価パラメータが単独のため, 複数の適合パターンが第 1 候補になる場合が多かった．クローズドテストでは 9 文, オープンテストでは 7 文の入力文において, 正解・不正解適合パターンが第 1 候補に混在していた．そこで, 第 1 候補に不正解適合パターンが含まれる入力文を除外し, 本稿で作成した評価関数と比較した．結果を表 8 に示す．表 8 より, パターン元字面適合率単独で求めた評価関数は, 本稿で作成した評価関数より選択精度は劣るが, 多くの入力文において正解適合パターンを一意に選択できた．

表 8: パターン元字面適合率のみによる実験結果

評価関数	クローズド	オープン
本稿作成評価関数	69%(20/29)	83%(24/29)
パターン元字面適合率	59%(17/29)	66%(19/29)

6 まとめ

本稿では, 等価的類推思考の原理に基づく機械翻訳方式の実現に向け, 文型パターンパーサが出力する複数の適合結果から, 入力文の翻訳が可能な適合パターンの選択を行った．具体的には, テスト入力文の適合パターンの情報を多変量解析によって分析し, 適合パターン選択評価関数を求めた．そして, 得られた評価関数で適合パターンの選択を行った．

実験の結果, クローズドテストでは 72%, オープンテストでは 83%の入力文において, 第 1 候補に正解適合パターンが存在した．今後は, 多くの文型パターンが適合する入力文において選択精度を上げる必要があると考えられる．

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: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.
[3] 徳久雅人，池原悟，村上仁一:文型パターンパーサの試作，言語処理学会第 10 回年次大会発表論文集，pp.608-611，2004.
[4] 池原悟：機械翻訳のための日英文型パターン記述言語，信学技報，TL2002-48/NLC2002-90，pp.1-6，2003.
[5] 前田春奈，池原悟，村上仁一：パターンを使用した重文複文の日英翻訳の精度，言語処理学会第 10 回年次大会発表論文集，pp.237-240，2004.
[6] 池原悟，宮崎正弘，白井諭:日本語語彙大系，岩波書店，1997.
[7] 中岩浩巳，池原悟:日英の構文的対応関係に着目した日本語用言意味属性の分類，情報処理学会論文誌，Vol.38，N0.2，pp.215-225，1997.
[8] 渡辺美智子，小山斉:Excel 徹底活用統計データ分析，秀和システム，2003.