032005 多変量解析による最適文型パターンの選択方式

計算機工学講座　池原研究室　岡田 敏

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. 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. 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, 品質の良い英文が生成できる文型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　評価関数

評価関数の作成には, 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．
- 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した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元字面適合率は 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.
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 [6], 平均値の逆数を使用する．
- 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 [6], 平均値の逆数を使用する．

3.3 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．

3.4 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．

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

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, 品質の高い翻訳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 適合パターン選択実験結果

テスト入力文 200 文およびオープンテスト入力文 200 文には, 正解適合パターンが 29 文存在した．適合パターン選択実験を行った結果を表 2 に示す．

表より, 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 2: 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

不正解適合パターンが第 1 候補にな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．

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少なかったため, オープンテストの結果はクローズドテストの結果より良い値を示していた．

さらに品質を向上させるには, より多くの評価パラメータで評価関数を作成する必要があると考えられる．

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, より多くのパラメータを用いて評価関数を作成する必要があると考えられる．

参考文献

[1] 池原ほか：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.

[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 回年次大会発表論文集,pp.608-611,2004.

[4] 池原ほか：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] 前田ほか：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] 池原ほか:日本語語彙大系, 岩波書店,1997.