関数・記号付き文型パターンを用いた機械翻訳の試作と評価

石上 真理子　水田 理夫　徳久 雅人　村上 仁一　池原 悟
鳥取大学工学部知能情報工学科
{isigami,s032054,tokuhisa,murakami,ikehara}@ike.tottori-u.ac.jp

1　はじめに

重文・複文の日英機械翻訳のために，文型パターン辞書が構築されている [1]．これまでの開発で重視された点は，文型パターンの被覆率の確保であり，文型パターンを用いた翻訳は未着手であった．そこで，本稿では，[1] の文型パターン辞書を用いた日英パターン翻訳方式を実装するとともに，実装したシステムの評価実験を行う．

2　提案する日英パターン翻訳方式

2.1　文型パターンと照合

文型パターン辞書は，約 12 万文の対訳コーパスから単語，句，節の 3 つのレベルで約 23 万件の対訳パターンがすでに作成されている．このうちの単語レベル文型パターンの例を以下に示す．

日本語パターン：#1{$N1$ が,$N2$ に } なってから $N3$ を ($V4$^$meirei$|$V4$.$meireigo$)。

英語パターン：$V4$ $N3$ after $N1$ turn $N2$.

文型パターンの記述子は，字面，変数，関数，記号である．この例で，N および V は変数，^$meirei$ および .$meireigo$ は関数，#数 {···，···} および (···|···) は記号である．

以下の日本語文と，上記の日本語パターンを照合すると，変数や記号にはバインド値が代入される．

日本語文：信号が青になってから道路を渡りなさい。

バインド値：#1=$N1$ が $N2$ に，$N1$=信号，$N2$=青，$N3$=道路，$V4$=渡り

2.2　提案する日英パターン翻訳方式の骨組み

文型パターン辞書を用いた日英翻訳は，英語パターンにより英語文の構造が決まるため，英文全体の意味を粗く保つ利点がある．しかし，英語パターンの記述子の局所的な翻訳に対する制約は弱く，何らかの方法で訳出の選択を行わなければならない．

そこで本稿で提案する方式では，まず記述子の局所的な翻訳を局所翻訳と定義し，複数の局所翻訳結果を絞り込まずに英語パターンに代入する．次に，英語の言語モデルを用いて尤度の高い訳語の組みを選択して，英文を生成する．その骨組みを以下に示す．

1. 英語文の構造決定
2. 局所翻訳の実行および英語パターンへの非決定的な代入
3. 英語パターンの記述子制約による絞り込み
4. 英語の言語モデルによる翻訳候補の選択

(1) 英語文の構造決定

英語文の構造決定は，英語文の全体的な構造の決定および部分的な構造の決定がある．

(1-1) 英語文の全体的な構造の決定

英語文の全体的な構造は，英語パターンにより決まる．

英語パターンは，入力日本語文と適合した日本語パターンにより決まる．入力日本語文が適合する日本語パターンは複数あり，パターンの選択をしなければならない．パターンの選択問題については，[2]，[3] に解決方法が示されている．

(1-2) 英語文の部分的な構造の決定

英語文の部分的な構造は，英語パターン中の記号で定義されており (表 1)，基本的には入力日本語文と日本語パターンの照合結果により記号の処理が決まる．例えば，訳出要素選択記号 “<$N1$|I>” は，日本語文で $N1$ に適合する日本語要素がある場合は “$N1$”，そうでない場合は “I” を使用する．

表 1: 英語パターンで使用される記号一覧

分類	記号例	機能説明
要素選択記号	(a|an)	複数記述された要素のいずれかを使用
対応型要素選択記号	(日)#1(.$genzai$|.$kako$) (英)#1(^$present$|^$past$)	日英での要素の対応関係を保ったまま，対応する順序で日英での要素を指定
訳出要素選択記号	<$N1$|I>	日本語文型パターンで左側のパターン記述に対応する要素が適合しなかった場合右側のパターン記述を使用
任意要素記号	#1[$ADV1$]	日本語文型パターンで適合する要素があった場合のみ訳出の対象となる
標準形表記記号	‘is’	字面部分の標準形を指定
要素挿入記号	#1{never}	英語側の要素の中に副詞等の別の要素を挿入し訳出

(2) 局所翻訳の実行および英語パターンへの非決定的な代入

英語変数の翻訳を行うと，訳語およびその活用形などの候補が複数得られるので，これらを絞り込まずに英語パターンに代入する．

文型パターンで使用される変数を表 2 に示す．日本語のみの変数として，$TIME$，NUM，ND，AJV，REN，GEN，$AJVP$ がある．これらは英語パターンでは，N，AJ に対応する．

表 2: 文型パターンで使用される変数

分類	変数	機能説明
単語	N	名詞または名詞複合語
	$TIME$	時詞
	NUM	数詞
	ND	用言性名詞
	V	動詞
	AJ	形容詞
	AJV	形容動詞
	ADV	副詞
	REN	連体詞
	GEN	限定詞
句	NP	名詞句
	VP	動詞句
	AJP	形容詞句
	$AJVP$	形容動詞句
	$ADVP$	副詞句
節	CL	節を表す
その他	ANY	直接引用で使用され，どのような要素が来ても良い事を示す

(3) 英語パターンの記述子制約による絞り込み

制約の機能を持つ英語パターンの記述子は，変数および表 3 で示した関数であり，(2) の候補の絞り込みができる．例えば，名詞変数 N であれば品詞が名詞の候補が選ばれ，関数 "$\hat{}past$" であれば，過去形の候補が選ばれる．

表 3: 英語パターンで使用されるプリミティブな関数

分類	関数	機能説明
名詞制約	$\hat{}obj$，$\hat{}poss$，$\hat{}pron$ $\hat{}adposs$，$\hat{}reflex$	目的格，所有格，代名詞 独立所有格，再帰代名詞
動詞制約	$\hat{}base$，$\hat{}past$，$\hat{}ing$ $\hat{}present$，$\hat{}ed$	原形，過去形，現在分詞形 現在形，過去分詞形
形容詞制約	$\hat{}er$，$\hat{}st$	比較級，最上級

(4) 英語の言語モデルによる翻訳候補の選択

(2) と (3) の結果を用いて英語パターンをワードグラフに変換し，英語の言語モデルを用いて翻訳候補を選択する．ワードグラフの例を図 1 に示す．

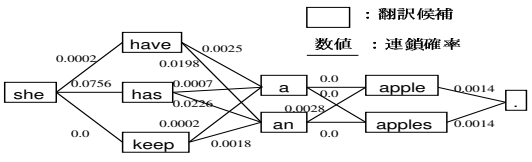

図 1: ワードグラフの例

2.3 複雑な関数

英語パターンで使用される複雑な関数を表 4 に示す．

表 4: 英語パターンで使用される複雑な関数

分類	関数	機能説明
様相関数	$.will$，$.would$，$.can$ $.not$，$.must$，$.may$ $.should$，$.better$	意志+試行動作，可能 否定，義務，許可 選択的判断
語形関数	$\hat{}future$，$\hat{}prp$，$\hat{}psp$ $\hat{}prog$，$\hat{}passive$，$\hat{}grn$	未来形，現在完了形 過去完了形，進行形 受身形，動名詞
品詞変換関数	$N()$，$V()$，$AJ()$ $ADV()$，$NP()$，$VP()$ $AJP()$，$ADVP()$	(名詞，動詞，形容詞 副詞，名詞句，動詞句 形容詞句，副詞句) 変換

様相関数は，助動詞や否定表現に使用する．例えば関数 "$.can$" は，"$V.can$" のように動詞変数に付属する．助動詞の直後の動詞は原形でなければならないので，プリミティブな関数に置き換えると "can $V\hat{}base$" となる．

語形関数は，時制や相の表現に使用する．例えば関数 "$\hat{}prp$" は，"$V\hat{}prp$" のように動詞変数に付属する．この場合も同様にプリミティブな関数に置き換えると "have $V\hat{}ed$" となる．

品詞変換関数は，"変換後変数 (変換前変数)" と記述され，"変換前変数" が日本語変数を，"変換後変数" が英語変数を意味する．日本語変数は照合結果より参照できる．そのため，"変換後変数 (変換前変数)" を "変換後変数" に書き換える．

3 翻訳実験システムの実装

提案方式を実装する上の要点を説明する．なお，本稿で実装するシステムを「ITM」と称す．

3.1 文型パターンの書き換え

複雑な関数を処理するために文型パターンの書き換えを行う．英語パターンファイルにおいて，5 回以上出現する表現を対象に，書き換え規則を作成して，機械的に書き換える．本稿では 472 個の書き換え規則を作成した．書き換え例を表 5 に示す．

表 5: 文型パターンの書き換え例

要素	書き換え前	書き換え後
様相関数	$V.will$	will $V\hat{}base$
語形関数	$V\hat{}passive.should$	should be $V\hat{}ed$
品詞変換関数	$ADV(AJ)$	ADV
要素挿入記号	$V\hat{}passive\#\{*\}$ $V\hat{}psp\#\{*\}$	@be$\hat{}present$ * $V\hat{}ed$ had * been $V\hat{}ed$

3.2 局所翻訳

変数翻訳は既存の辞書引きプログラムを使用する．辞書引きは,「日本語要素」,「適合する日本語変数」および「対応する英語変数」の情報を元にする．日本語変数と英語変数の種類が違う場合，日本語要素の末尾を変化させて辞書引きを行う．具体例を表 6 に示す．

表 6: 日本語変数，英語変数の種類が違う辞書引き例

日本語変数	英語変数	日本語要素	末尾変化後	辞書引き結果
ADV	*N*	うまく	うまさ	felicity，skill，$\cdots$
AJ	*V*	欲しい	欲しがる	desire，want，$\cdots$
AJV	*ADV*	確かだ	確かに	admittedly，$\cdots$

また，訳語の複数の活用形を候補として得るために，活用形辞書を用いる．例えば，“go” の活用形は，“go”，“goes”，“went”，$\cdots$ が得られる．

3.3 翻訳候補の選択

[1] の文型パターン作成に用いた英文集約 12 万文を英語の言語モデルとして，局所翻訳で生成した翻訳候補の選択を行う．本稿では，2 単語の連鎖確率を用いる．

4 評価実験

実装したシステムの評価項目は，(1) 記述子処理の正確性，(2) 英語原文の復元，(3) 出力英文の意味的な正しさとする．

実装したシステムの評価するために，基本的にクローズドテストをする．ただし，局所翻訳は既存の辞書引きプログラムを用いるので，局所翻訳処理はオープンテストになる．

4.1 記述子処理の正確性

4.1.1 評価方法

文型パターン作成の際に使用した日本語原文を用いて，記述子処理の正確性を評価する．単語レベル文型パターンに記述される変数，関数，記号を網羅する入力文をランダムに抽出し計 184 文で行う．評価基準について，変数は訳出が英語パターンの品詞を満たす場合は○，それ以外は×とし，また，関数および記号は，絞り込み後の訳出が指定通りの場合は○，それ以外は×とする．

4.1.2 結果

結果を表 7 に示す．変数および記号は，全て正確に処理できた．語形関数は，ほぼ正確に処理できた．品詞変換関数は，約 30%が×となった．なお，品詞変換関数において品詞変換できない場合は，他の英語パターンの使用が適切である．しかし，本稿では処理の可能性を評価するため他の英語パターンは使用しない．

表 7: 記述子処理の正確性

調査対象	○	×
変数	100.0%(40/40)	0.0%(0/40)
品詞変換関数	66.7%(60/90)	33.3%(30/90)
語形関数	97.6%(41/42)	2.4%(1/42)
記号	100.0%(12/12)	0%(0/12)

4.1.3 誤り分析

表 7 において×の割合が最も多かった品詞変換関数について分析する．

1. 品詞変換不可能：24 件
   他の英語パターンを使用する必要がある．例えば，バインド値が “たばこ”，日本語変数が “*N*” そして英語変数が “*V*” の場合，一般的には “smoke” だが，“たばこ” では “smoke” が辞書引きできない．
2. 辞書引き方法の追加により可能：6 件
   例えば，バインド値が “はっきり”，日本語変数が “*ADV*” そして英語変数が “*AJ*” の場合，“はっきりした” で辞書引きすれば可能．

4.2 英語原文の復元

4.2.1 評価方法

文型パターンの作成に使用した日本語原文からランダムに抽出した計 250 文に対して行う．評価基準は，出力英文が英語原文と等しい場合は○，それ以外を×とする．ただし，“the”，“a”，“an” の冠詞は，本稿では実装していないため，評価の対象外とする．

4.2.2 結果

結果を表 8 に示す．復元できたのは約 10%であった．ただし，ITM の出力英文に英語原文以外の単語を含むと，×とするためかなり厳しい成功判定である．

表 8: 英語原文の復元成功率 (全 250 件)

○	×
11.6%(29/250)	88.4%(221/250)

4.2.3 誤り分析

表 8 において×であった 221 件について，提案方式のどの段階で×となったのかを分析した (表 9)．表中の (1)，(2)，(3)，(4) は，2.2 節の翻訳方式のステップを示す．

表 9: 復元失敗の原因の分布

×となった段階	(1) の段階	(2) の段階	(3) の段階	(4) の段階
割合 (全 221)	25.3% (56/221)	56.1% (124/221)	0.0% (0/221)	18.6% (41/221)

(1) は，英語パターン表記の問題である．字面表記の動詞が原形で書かれる英語パターンがあるため活用形の変形が不可能であった．

(2) は，辞書引きで失敗した．局所翻訳処理はオープンテストであるので，辞書引き結果に復元できる単語が少なかった．

(4) は，原文の単語が候補に含まれていたが，尤度による選択で失敗した．

なお，(1) については，関数を付与する対策が現在なされている．また，(4) の原因として，文型パターンに付与される意味属性を用いていない事が考えられる．

4.3 出力英文の意味的な正しさ

4.3.1 評価方法

出力英文の意味的な正しさを評価するために，クローズドテストを行い，人による評価をする．翻訳対象は，4.2 節と同じ計 250 文を対象とする．評価基準を以下に示す．

評価値 4： SVOC の関係が正しく構成されており，単語の誤訳がない

評価値 3： SVOC の関係が正しく構成されているが，単語の誤訳がある

評価値 2： SVOC の関係が部分的に構成されている

評価値 1： SVOC の関係が部分的に構成されていない

4.3.2 結果

人による評価の平均値と内訳を表 10 に示す．前述の実験によると局所翻訳に問題があったが，代わりの英単語によりある程度適切に英訳できていると言える．

評価 4 となった事例を以下に示す．

入力文 1：欠点はあるがそれでも彼が好きだ。

理想解：I like him in spite of his faults .

日パターン：$N1$ はあるが (それ | 其) でも<$N2$ は> $N3$ が $AJV4$。

英パターン：<I|$N2$> $V(AJV4)$ $N3\hat{}obj$ in spite of $N3\hat{}poss$ $N1$.

出力：I like him in spite of his faults .

入力文 2：足音が遠くて聞こえなくなった。

理想解：His footsteps died away in the distance .

日パターン：$N1$ が $AJ2$(て | で) 聞こえなくなった。

英パターン：$N1$ died away in $N(AJ2)$.

出力：Footsteps died away in a long way .

表 10: 人による評価の平均値と内訳

平均値	内訳			
	評価値 4	評価値 3	評価値 2	評価値 1
3.5	54.4% (136/250)	42.8% (107/250)	2.8% (7/250)	0.0% (0/250)

参考として，3 種類の一般の翻訳システムを用い，同じ入力文，評価基準にて人による評価を行った．平均値を表 11 に示す．

表 11: 一般の翻訳システムの評価

システム	システム 1	システム 2	システム 3
平均値	2.7	3.2	3.4

4.3.3 誤り分析

表 10 の内訳において，評価 3 が多かった．原因として，尤度による選択で正しくない単語が選択された場合と，英語パターン表記が不十分な場合がある．

尤度による選択で正しくない単語を選択した例

入力文：事件が漏れて世間の評判になっている。

理想解：The affair has got wind, and is the talk of the town.

出力：Events have fallen , and is rumor of people .

出力は，単数，複数の整合が取れていないため誤訳と判断され，評価値 3 となった．しかし，候補には “Events” だけでなく，“Event，incident，affair···” があり，“have” だけでなく，“has” もあった．単数形が選択されると，評価値 4 になると考えられる．

英語パターン表記が不十分な場合の例

入力文：彼女は田舎に子どもを残して来た。

英語パターン：$N1$ have left $N1$ $N3$ behind $N1$ in $N2$.

出力：She have left her children behind her in country.

動詞 “have” は字面表記であるため (下線部)，活用形の変形ができない．

これら二つの原因が解決できると，評価値 4 の割合が 90%以上になると考えられる．

5 おわりに

本稿では，[1] の文型パターンを用いて英文生成を実現するために，[1] の文型パターンを用いた日英パターン翻訳方式の提案と実装を行い，実装したシステムの評価実験を行った．

評価項目は，(1) 記述子処理の正確性，(2) 英語原文の復元，(3) 出力英文の意味的な正しさとした．結果として，(1) は，変数および記号は正確に，語形関数はほぼ正確に処理できた．品詞変換関数は，変換が困難なものが多いが，その場合は他の英語パターンを使用する事で対応が取れる．(2) は，約 10%しか復元できなかったが，絞り込みの処理で×になった例はなく，辞書引きの失敗，尤度による選択や英語パターンの表記が問題であった．なお，英語パターンの表記については現在対策がなされている．(3) は，評価値 3 が多かったが，尤度による選択，英語パターン表記が問題であった．この二つの問題が解決すると，評価値 4 の割合が 90%以上になると考えられる．

今後は，句変数の翻訳システムの組み込み [4][5]，また，オープンテストでも良い英文を生成できるように，改良を行う．

謝辞

本研究は，独立行政法人科学技術振興機構 (JST)・戦略的創造研究推進事業 (CREST) の研究領域「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による言語の等価変換と生成技術」の支援によるものである．

参考文献

[1] 池原, 阿部, 徳久, 村上：非線形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] 岡田, 村上, 徳久, 池原：多変量解析による最適文型パターンの選択方式, 言語処理学会第 11 回年次大会, pp.25-28, 2005.

[3] 原, 村上, 徳久, 池原：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, 言語処理学会第 12 回年次大会, pp.268-271, 2006.

[4] 吉岡, 徳久, 村上, 池原：日英対訳パターンを用いた名詞句翻訳, 言語処理学会第 12 回年次大会, pp.580-583, 2006.

[5] 石上, 徳久, 村上, 池原：結合価パターンを用いた動詞句の翻訳可能性の調査, 言語処理学会第 11 回年次大会, pp.364-367, 2005.