重文・複文文型パターン辞書による意訳の可能性

The feasibility of a non-literal translation
by the pattern dictionary for complex and compound sentences

水田理夫 Michio Mizuta　徳久雅人 Masato Tokuhisa　村上仁一 Jin'ichi Murakami　池原悟 Satoru Ikehara

鳥取大学 工学部 知能情報工学科
Department of Information and Knowledge Engineering, Tottori University

1　はじめに

従来の要素合成法を用いた日英翻訳方式では，統語構造の単位で一律に直訳が行われるため，その過程において意味が失われる問題がある．文型パターン翻訳方式では，表現構造と意味の関係に着目し，要素合成法により翻訳できる所を区別している．そのため，簡潔な表現としての意訳が期待される．本稿では [1] の文型パターン辞書に基づく翻訳プロトタイプシステム ITM[2] において，重文・複文の意訳性能を評価する．

2　日英パターン翻訳プロトタイプシステム ITM

ITM は，文型パターン辞書の照合結果より半自動で日英翻訳を行うツールである．文型パターン辞書は，単語，句，節の 3 つのレベルで分類された約 22.7 万件の日英文型パターン対で構成されている [1]．日本語文と文型パターン辞書の照合は，パターン検索プログラムにより実行される [3]．パターン検索の結果として，適合する全てのパターンが得られるので，人手で，翻訳に使用する日英パターン対，および，適合状態を指定すると，ITM がそれらを用いて英文を生成する [2]．なお，本稿では，文型パターン辞書から，単語レベル，句レベル，節レベルの順にパターンを参照し，最初に適合したレベルの結果を使用する．

以下に入力文，日英文型パターン対，および，訳出文の例を示す．

入力文　人は年をとると大人になる。
(単) 日パ　/ytcfk$N1$ は/tcfk 年を/cf(取る | とる) と /tck(大人 | おとな) に/cf(成る | なる)。
(単) 英パ　As $N1$ grow older, $N1$ will grow in moderation.
(句) 日パ　/ytcfk$N1$ は $VP2$ と $VP3$。
(句) 英パ　As $N1$ $VP2$, $N1$ will $VP3$.
(節) 日パ　/ytcfk$CL1$ と/tck$CL2$。
(節) 英パ　As $CL1$, $CL2$.
訳出文　As you grow older, it will grow in moderation.

N, VP, CL は，それぞれ名詞，動詞句，節の変数である．これらは，線形要素と呼ばれる部分であるので，変数に対応する日本語表現を局所的に翻訳した結果を，英語パターンの対応箇所に挿入することができる．

/ytcfk や /cf などは，離散記号である．各離散記号と適合する日本語要素の詳細を表 1 に示す．現在の ITM には未実装だが，離散記号処理機能が稼働すれば離散記号に適合する部分の訳出が可能となる．例えば，格要素に適合する記号 /c が「〜で」という日本語表現に適合した場合に,「by 〜」,「in 〜」等の知識を用いて訳出する．これらは [1] の辞書における対訳の関係から抽出できる．

表 1　各離散記号と適合する日本語要素

/y	連用節と適合
/t	連体節と適合
/c	格要素と適合
/f	連用修飾要素と適合
/k	連体修飾要素と適合

3　意訳とは

本研究における意訳文の定義は「日本語文の節数よりも，翻訳後の英文の節数が少なく，かつ，それが正しい英文であること」とする．意訳文は直訳文よりも簡潔で明瞭な場合が多く，品質の高い訳文と考えている．

4　評価実験

4.1　意訳文の評価方法

[1] の辞書からランダムに意訳文を 135 文抽出し，実験に用いる．これらを ITM で翻訳した結果を評価することで，文型パターン辞書による意訳の可能性を調査する．

ITM の翻訳結果を，手動で評価する．評価基準は，日英文の意味的な対応関係の正否と英文法上の正誤を判断するものである．具体的には，以下の 4 段階とする．

〈評価基準〉

評価 4	動作主，動作，対象の意味的関係が正しく，統語的誤り * がない
評価 3	上記の意味的関係が正しいが，統語的誤り * がある
評価 2	上記の意味的関係が部分的に正しい
評価 1	上記の意味的関係が正しくない

* 冠詞と複数形については誤りとしない

〈評価例〉

評価 4	入力文	夕方になってやっと暑さが和らいだ。
(意訳)	正解文	The heat finally let up in the evening.
	訳出文	Heat finally let up in evening.
評価 3	入力文	うまくいったので意気揚々としている。
(意訳)	正解文	He is flushed with success.
	訳出文	He be flushed with success.
評価 2	入力文	人が自分のことを話していたから気を利かして場をはずした。
	正解文	They were talking about me, and so I discreetly kept out of the way.
	訳出文	He used his mind to leave room.
評価 1	入力文	彼は頭もよく勤勉でもある。
	正解文	He is as diligent as bright.
	訳出文	We and is <n5>.

4.2 ITM の翻訳実験結果

意味的な保存を重視するため，評価値 3，4 の訳出文を「翻訳成功」とし，さらに「意訳成功」と「直訳成功」に分類する．評価値 1，2 の訳出文を「翻訳失敗」とし，パターンに適合しなかった文を「適合無し」とする．

基本的な性能を評価するために，単語レベルパターンのみを使用し，かつ，離散記号の処理を想定しない，という場合について評価実験を行う．評価結果を表 2 に示す．

クローズドテストとは，入力文から作られたパターンを用いて英文を生成する場合である．クロスバリデーションテストとは，入力文から作られたパターンは使用せずに，別のパターンを用いて英文を生成する場合である．

表 2　翻訳性能：単語レベル・離散記号未処理の場合

実験条件	意訳成功	直訳成功	翻訳失敗	適合無し
ITM (closed test)	96% (129/135)	1% (2/135)	2% (3/135)	1% (1/135)
ITM (cross validation)	13% (17/135)	10% (14/135)	41% (55/135)	36% (49/135)

表 2 の翻訳失敗の文の多くは離散記号が原因であった．例を以下に示す．

入力文： とても優遇されて英文学の講座を担当した。

日本語パターン： /**y**</tk$N1$ は>#1[/tcfk 容赦!なく] #2[/ytck$N4$ の]!$N5$ を/cf$V6$.kako。

バインド値： #2=$N4$ の，$N4$=英文学，$N5$=講座，$V6$=担当し /y= とても優遇されて

英語パターン： <I|$N1$> #1[ruthlessly] $V6$^past $N5$ #2[of $N4$].

訳出文： I covered chair of english literature.

評価： 2

正解訳： He filled the chair of English literature with great distinction.

入力文の英語パターン： <He|$N1$> filled the chair of $N2$ with great distinction.

上記の例では,「とても優遇されて」の部分が /y に適合しているため，訳出文では対応する意味の文が出力されていない．自己英語パターンを参考に人手による離散記号処理を示す．対応する部分を挿入した例を以下の下線部に示す．

離散記号処理をした訳出文： I covered chair of english literature with great distinction.

評価： 4，意訳

「翻訳失敗」となった 55 文（表 2 を参照）について，離散記号処理を想定して再実験を行った結果を表 3 の (a) 行に示す.「意訳成功」が+15 件,「直訳成功」が+8 件と増加した．さらに「適合無し」の 54 文（表 3(a) を参照）について，句，節レベル文型パターンによる照合と翻訳を行い，かつ離散記号処理を行った結果を表 3 の (b) 行に示す．表 2 と比較して「意訳成功」が+25 件,「直訳成功」が+9 件と増加した．

表 3　翻訳性能：句，節レベル・離散記号処理の場合

実験条件	意訳成功	直訳成功	翻訳失敗	適合無し
(a) 単語レベル 離散記号処理	24% (32/135)	16% (22/135)	20% (27/135)	40% (54/135)
(b) 句・節レベル 離散記号処理	31% (42/135)	17% (23/135)	47% (64/135)	4% (6/135)

4.3 一般翻訳システムとの比較

ITM と一般の翻訳システム（システム 1，2）の性能比較を行う．意訳率を「意訳成功文数／入力文数」とし，意訳率の比較結果を表 4 に示す．

表 4　ITM と一般翻訳システムにおける意訳率の比較

ITM	システム 1	システム 2
31% (42/135)	13% (17/135)	26% (35/135)

ITM の意訳率がシステム 1，2 よりも高いため，文型パターン辞書は意訳文の訳出において優位と言える．

5　今後の課題

第 4.3 節で文型パターン辞書による意訳の可能性が確認できたが，全自動で行うには以下の問題を解決しなければならない．

- ITM に使用するパターンの選択
  現在，ITM で使用する文型パターンは入力日本語文と文型パターンの照合の後に，手動で選択している．この自動化に向けて，多変量解析の手法により最適なパターンの選択が検討されている [4, 5].
- 離散記号の訳出処理
  離散記号の訳出処理の自動化を行うためには，挿入する英訳の形成と，英訳の挿入位置の決定を行う必要がある．英訳の形については，日英対訳文の対応関係から抽出の見込みがある．挿入位置は英文の構文情報を用いる必要があり，検討をすすめている．

6　おわりに

文型パターン辞書を用いたパターン翻訳方式において，重文・複文の意訳の出力可能性が確認できた．

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, 11(3), pp.149-164, 2004.

[2] 石上, 水田, 徳久, 村上, 池原: 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, 言語処理学会第 13 回年次大会発表論文集, pp.67-70, 2007.

[3] 徳久, 村上, 池原: 重文・複文文型パターン辞書からの構造照合型パターン検索, 情報処理学会研究報告, 自然言語処理, 2006-NL-176, pp.9-16, 2006

[4] 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 回年次大会発表論文集, pp.25-28, 2005.

[5] 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 回年次大会発表論文集, pp.268-271, 2006.