意味的等価変換方式による句レベルパターン翻訳方式の調査

坂田 純　徳久 雅人　村上 仁一
鳥取大学大学院工学研究科情報エレクトロニクス専攻
{d112004,tokuhisa,murakami}@ike.tottori-u.ac.jp

1　はじめに

要素合成法を基本とする従来の機械翻訳方式には, 表現の構成要素への分解過程において意味が消失し, 目的言語の生成過程で意味が復元されなくなる問題がある. この問題を解決するために, 意味的等価変換方式が考えられている [1]. この方式では, 文を線型部分と非線型部分にわけて文パターンとして記述し, 線型部分に対する局所翻訳を施し, 非線型部分と組み合せて文全体の訳出を行う.

日英機械翻訳の, 意味的等価変換方式による実装には, 大量の日英文パターン対を記述する必要がある. これは日英重文複文文型パターン辞書 [2] として作成されている. なお文パターンは, 線型要素のレベルに応じて, 単語, 句, 節レベルの 3 つのレベルで記述されている. 次に, 入力文と同型の文パターンを検索する必要があり, 構造照合型パターン検索システム SPM[3] が開発されている. 最後に, 得られた文パターンを用いて英語文を生成する. そのプログラムが ITM である. 単語レベル文パターン対を用いた単語レベルパターン翻訳機能が, ITM に実装されている.

3 種類の文パターンを使用した翻訳のうち, 単語, 句, 節レベルの順に精度のよい翻訳文が得られると考えられるが, 逆に文パターン適合率は, この順に低下することが明かになっている [3]. 適合率においては, 句, 節レベルの文パターンの使用が有効であるが, その場合, 線型部分である句と節の局所翻訳が必要となる. ところで, 線型要素と非線型要素の分類が再帰的な構造を持つと考えられており [1], 線型部分に対しても再帰的に同様の処理を施すことが可能である.

現在, 句の翻訳への利用を視野に入れて, 句の表現構造を解析するための句パターン辞書が作成されている. そこで本研究では, この句パターン辞書を用いて, 再帰的な ITM による句局所翻訳を行う句レベルパターン翻訳機能を実装し, 翻訳精度を調査する.

2　パターン辞書

2.1　日英重文複文文型パターン辞書

辞書 [2] の記述例を図 1 に示す.

AC000004-00	
日原文	彼のお母さんがああ若いとは思わなかった。
英原文	I never expected his mother to be so young.
W 日パターン	/y</tk*N*1 は >/tcfk*N*2 の/k*N*3 が/tcfk ああ/f*AJ*4 とは/yf*V*5.hitei.kako。
W 英パターン	<I|*N*1> never *V*5^past *N*2^poss *N*3 to be so *AJ*4.
P 日パターン	/y</tk*N*1 は >/tcfk*NP*2 が/tcfk ああ /f*AJ*3 とは/yf*V*4.hitei.kako。
P 英パターン	<I|*N*1> never *V*4^past *NP*2 to be so *AJ*3.

図 1　重文複文文型パターン辞書の記述例

W が単語レベル, P が句レベルの文パターンを示している. それぞれのレベルにおいて, 日英文パターン対として対応している. W 日パターンを例にとる.

1. “ / ” は文節境界を示す離散記号であり, 文型パターンに記述がない要素が挿入可能なことを示している.
2. “ / ” の後ろに付与された y 等の記号は挿入可能な要素の種類を示している.
3. 動詞変数 *V*5 に付与されている “.hitei.kako” は, その動詞が否定型の過去型であることを示している.

次に W 英パターンを例にとる.

1. “ ^ ” はその直前の語を変形することを指定する語形変形関数である.
2. *V*5^past は *V*5 の動詞を過去型に変形することを意味し, 日本語パターンの “.hitei.kako” に対応している.

2.2　句パターン辞書

句パターン辞書は, 文パターン辞書の単語レベル及び句レベル文パターンを用いて, プログラムにより自動作成されている. 上の例において, P 日英パターンの *NP*2(彼のお母さん) と *NP*2(his mother) の句パターン対から, W 日英パターンの句パターンの文字列を構成する記述を抽出し, 単語レベル句パターンを作成する. 句パターンは名詞句 (*NP*), 動詞句 (*VP*), 形容詞句 (*AJP*), 形容動詞句 (*AJVP*), 副詞句 (*ADVP*) の 5 種類に分類される. 作成された句パターンのうち, 名詞句と動詞句の例を順に示す. なお動詞句パターンでは, 動詞の非線型性の強さから, 動詞は変数化せずに動詞原形で記述されている.

＜名詞句パターン例＞

日原文	彼のお母さん
英原文	his mother
日パターン	/tcfk*N*1 の/k*N*2
英パターン	*N*1^poss *N*2

＜動詞句パターン例＞

日原文	ああいう人と付き合う
英原文	associate with that kind of person
日パターン	/tcfk ああ/f いう f*N*1 と/cf’ 付き合う’
英パターン	’associate’ with that kind of *N*1

図 2　句パターン例

3　パターン検索システム SPM

SPM[3] は，文型パターン辞書を用いる日本語パターン検索システムである．動作の順序は，まず，入力日本語文を形態素解析し，適合する日本語文型パターンを検索する．そして，形態素解析の結果と，適合したパターンの線形要素に関する情報を出力する．入力文「彼のお母さんがああ若いとは思わなかった。」の出力例を, 形態素解析結果, パターン照合結果の順に示す.

```
1. /彼 (1710,{NI:23,NI:48,KR:9900k08,KR:9901s81,
     KR:0601s10,KR:9901s86,IM:11131,IM:11211})
2. + の (7410)
3. /お母さん (1100,{NI:80,NI:49,IM:11212,IM:11220})
4. + が (7410)
5. /ああ (1110)
6. /若い (3106,{NY:5,KR:8803c00,IY:A300})
7. + と (7420)
8. + は (7530)
9. /思わ (2392, 思う, 思わ,{NY:32,NY:31,KR:0601a01,
     KR:1500a00,IY:1200,IY:1262,IY:2111})
10. + なかっ (7184, ない, なかっ)
11. + た (7216)
12. +。(0110)
13. /nil
```

図 3 形態素解析結果

図 3 の 1 行目を例にとると, “彼” が形態素であり, {NI:23,NI:48, ・・・ IM:11211} が形態素の意味上の分類群である意味属性 [4] の所属コードを示している.

```
PATTERN=PJAC000004-00
        =[NP2, が, ああ,AJ3, とは,V4,.hitei,.kako,。]
        =[1,2,3,4,5,6,7,8,9,10,11,12]=12
NP2=[1,2,3]=3=3
AJ3=[6]=6=1
V4=[9]=9=1
---
PATTERN=PJAC000004-00
        =[NP2, が, ああ,AJ3, とは,V4,.hitei,.kako,。]
        =[3,4,5,6,7,8,9,10,11,12]=10
NP2=[3]=3=1
AJ3=[6]=6=1
V4=[9]=9=1
```

図 4 文パターン照合結果

図 4 の例では 2 パターンが照合されている. 1 行目の “PJAC000004-00” が日本語パターンの ID を示している. 3 行目がマッチした形態素番号を示しており, このパターンでは全ての形態素がマッチしていることがわかる. 4 行目からが形態素と照合された変数の情報である. 4 行目では, 形態素番号 1, 2, 3 の形態素 “彼”, “の”, “お母さん” が名詞句変数 *NP*2 にあたっていることを示している. 離散記号の適用により, 形態素解析による全ての形態素がパターンに照合しなくても, パターン照合可能であり, 下のパターンがこれにあたる.

句レベルパターン翻訳では, 文パターンの照合と, その照合において得られた句変数照合部分に対する句パターンの照合の, 二段階のパターン照合を行う. 文パターン, 句パターン共に, 離散記号の適用によって, 形態素がフルマッチしていないパターンも得ることができる. そのようなパターンを用いた場合, 照合されなかった形態素に対して, 別個に局所翻訳を行う必要が生じる. よって本研究では, 形態素がフルマッチしたパターンのみを翻訳に使用する.

文パターンが複数得られたときは, 翻訳に使用する文パターンを一つ, 次の手順で選択する.

手順 1 形態素解析結果の意味属性を用いた, 文パターンの絞り込み
手順 2 自己照合実験での使用頻度の最も高い 文パターンを選択
手順 3 なお複数パターン残る場合は, ランダムに一つ選択

手順 2 において, 最適なパターンを多変量解析により求める方法が検討されている [5] が, この方法ではまだ十分な精度が得られていない. よって本研究では, 上記手順 2 の方法を用いることとする.

句パターンの場合は 5 節で述べる.

4 パターン翻訳システム ITM

ITM により, 選択された文パターンを用いて訳出英文を得る. その際, 変数照合部分の日本語に対して局所翻訳を行う必要がある. 辞書引きにより得られた複数の局所翻訳結果は, 絞り込みを行わずに英語パターンに代入し, 英語の言語モデルを用いて最も尤度の高い訳語を選択し英文を生成する. ITM による翻訳の具体的な手順を以下に示す.

1. SPM による形態素解析と日本語パターン照合
2. 翻訳に使用するパターン対の選択
3. 変数に照合された日本語の局所翻訳
4. 局所翻訳結果の英語パターンへの非決定的代入
5. 英語の言語モデルによる翻訳候補の選択
6. 英語文全体の訳出

5 句レベルパターン翻訳

句レベル文パターンを用いた英語文の訳出には, 句変数照合部分に対する句局所翻訳の必要がある. 本研究では, 句変数照合部分の日本語に対して再帰的に ITM を呼び出し, 句局所翻訳では単語レベルのパターン翻訳を行う, 句レベルパターン翻訳機能を実装する.

句局所翻訳においては, 句パターン辞書のパターンを用いてパターン照合を行う. 句局所翻訳で行われる処理は次の 3 つである.

手順 1 SPM による句パターン照合
手順 2 得られた句パターン全てを使用した, ITM による句局所翻訳
手順 3 複数の句局所翻訳結果の全てを文パターンに 非決定的に代入

手順 2 において, 一つの句パターンにつき一つの句局所翻訳結果を得る. 手順 3 の後, 言語モデル (3-gram) による翻訳候補の選択を行い, 文全体の訳出文を得る. 句レベル ITM の模式図を図 5 に, 翻訳の具体例を図 6 に示す.

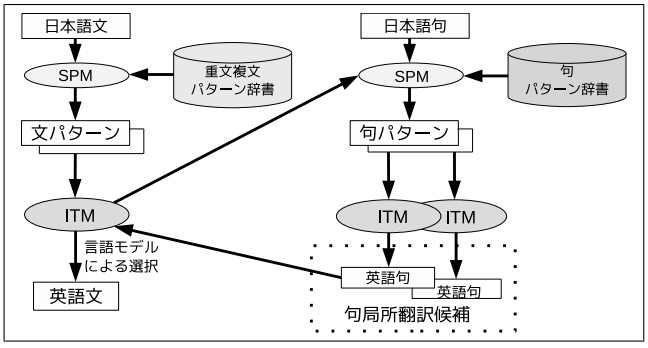

図 5 ITM による句レベルパターン翻訳

この例は, 図 3 の形態素解析結果と図 4 の照合結果を

入力文：彼のお母さんがああ若いとは思わなかった。
正解文：I never expected his mother to be so young.
英パターン：<I|*N*1> never *V*4ˆpast *NP*2 to be so *AJ*3 .
訳出文：I never expected his mother to be so young .

図 6 句レベルパターン翻訳例

使用して訳出している．図 4 の単語変数 *AJ*3 照合部分の “若い”, *V*4 照合部分の “思わ” に対して辞書引きを行い複数の局所翻訳結果を得る．英パターンの “*V*4ˆpast” は, 動詞変数 *V*4 の局所翻訳候補から, 過去形のみ選択することを指定している．具体的には, 翻訳候補「“think”, “thought”, “expect”, “expected”, ……」の中から過去形の「“thought”, “expected”, ……」を選択する．

句変数の *NP*2 照合部分 “彼のお母さん” に対して, 句パターン辞書を用いて再帰的なパターン翻訳を行っている．ここでは, 「“his mother”, “he’s mother”」等の局所翻訳結果が得られている．最後に, 変数照合部分に対して得られた局所翻訳結果を全て英語パターンに代入し, 3-gram による最尤の文を出力する．選択の例を図 7 に示す．

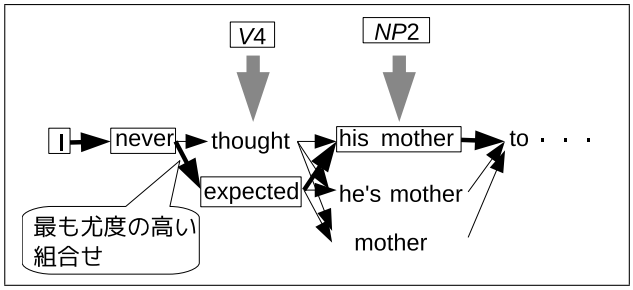

図 7 翻訳候補からの選択

6 クローズドテスト

6.1 実験方法

入力文に辞書 [2] の文を用いたクローズドテストにより, 動作確認を行う．SPM による文パターン照合を行い, 照合された自己パターンを用いて翻訳する．SPM 照合に 500 文, ITM による翻訳に 100 文を用いる．SPM 照合に用いる文は, そのパターンが *NP* を含むものを 100 文, 同様に *VP*, *AJP*, *AJVP*, *ADVP* を含む文がそれぞれ 100 文である．翻訳精度の評価は人手評価により行う．本システムとの比較には統計翻訳 (Moses) を用いる．moses の学習文には辞書 [2] の 120,000 文対を用いる．“distorshon-limit” を-1 とし, その他のパラメータは Moses のデフォルトの値を用いる．

評価基準は次の 3 通りとする．

評価 1 提案手法の方がベースライン (Moses) より優れている
評価 2 提案手法の方がベースラインより劣っている
評価 3 どちらも同程度に意味を理解できる, または理解できない

6.2 実験結果

SPM 照合結果を表 1 に示す．

訳出文の評価結果を表 2 に示す．

表 1 自己パターン照合結果

	SPM 照合失敗	自己パターンミスマッチ
NP	4/100	7/96
VP	4/100	2/96
AJP	7/100	4/93
$AJVP$	11/100	16/89
$ADVP$	20/100	27/80

表 2 人手評価の結果

	評価 1	評価 2	評価 3
文数	29	23	48

6.3 考察

6.3.1 SPM 照合結果

文パターン照合で, パターン照合の失敗と, 自己パターンに照合されない場合がみられた．主な原因は次の 3 つであった．

- SPM の照合ルールが足りない
- 文パターンの記述ミス
- 形態素解析の失敗

照合失敗数は *ADVP* が明らかに多く, 次いで *AJVP* が多かった．*ADVP* では, 照合ルールが足りない場合が多数であった．句パターンごとに句構造が異なる傾向が強く, 必要な照合ルールが多いためと思われる．*AJVP* は “名詞” + “だ” という構造と, 語尾の変化が複雑なことから, 形態素解析において形容動詞として解析されない, 形容動詞が副詞を受けるときのルールが足りないといった場合が多数みられた．

照合ルールの不足については, ルールの追加が必要になる．ただし, 不必要な適合の増加等が起らないよう, 詳細な調査が必要となる．記述ミスは人手による修正が必要である．どのような記述ミスがあるか, 調査, 分類を行った上での修正が必要と思われる．形態素解析の問題は, 現時点では不問とする．

6.3.2 評価結果

自己文パターンを用いていながら, よい翻訳精度が得られていない．最大の要因は, 局所翻訳候補からの不適切な候補の選択であった．翻訳精度の悪い例を示す．

入力文：その晩餐会は彼をたたえるために開かれた。
正解文：The dinner party was a tribute paid to him.
英パターン：*NP*1 @beˆpast *N*3 paid to *N*2ˆobj .
訳出文：That it was name paid to him .

図 8 翻訳精度の悪い例

*NP*1 に照合された形態素は「“その”, “晩餐会”」であり, 局所翻訳結果が “that it” になっている．これは句パターン “*AJ*1 *N*2ˆpron” による翻訳結果が選択されたためである．“*N*2ˆpron” は *N*2 を代名詞化することを指示している．*AJ*1 に “その”, *N*2 に “晩餐会” が照合されており, “その” の翻訳結果が “that“, “晩餐会” の翻訳結果が代名詞化されて “it” になっている．3-gram により選択されなかった候補の中には “the dinner party” のような候補もあり, 候補からの選択の失敗だとわかる．本研

究では, 複数の句パターンからの選択は行わず, それらのパターンから生成された局所翻訳結果に対して選択を行っている. 言語モデルによる選択だけでは限界があると思われ, 意味属性を用いた句パターンの選択等の処理が必要と考えられる.

7 オープンテスト

7.1 実験方法

重文複文のテスト文 300 文に対し, パターン照合と翻訳を行う. 文パターンの照合に成功し, さらに句変数照合部分がパターン照合に成功した文のみ翻訳を行う. 翻訳精度の評価は, 6.1 節の評価基準を用いた人手評価により行う.

7.2 実験結果

入力文 300 文に対してパターン照合に成功した文は 61 文であった. それら 61 文に対して句レベル ITM による訳出を行った結果, 14 文の訳出文を得た. 訳出失敗は句変数照合部分の句パターン照合失敗が原因であった. 人手評価の結果を表 3 に示す.

表 3 人手評価の結果

	評価 1	評価 2	評価 3
文数	1	2	12

クローズドテストと同様に, 句レベル ITM とベースラインの間に違いは見られない.

7.3 考察

7.3.1 SPM 照合結果

オープンテストでの文パターン照合率は約 20% であり, 文献 [3] に比較して非常に低い. 文パターン照合の正否は文末表現の違いに大きく左右されるため, 辞書 [2] と入力文の文体の違いによって照合率が大幅に低下したと思われる.

句パターン照合失敗の最大の原因は, 句変数照合部分が, 句パターンより長すぎることであった. 入力文「大阪までの間のどこかで駅弁を買って食べよう。」に対するパターン照合例を示す.

照合されたどのパターンも, 動詞句変数 *VP* として, “大阪までの間のどこかで駅弁を買っ” が照合されていた. これは, 想定していた *VP* パターンの長さに比べて長く, そのため句パターン照合に失敗する. ここで例えば, 文パターン “/y*NP*1 の/cf*VP*2 て/ycf*V*3.you” があれば, 上の入力文は, *NP*1 に “大阪までの間”, *VP*2 に “どこかで駅弁を買っ”, *V*3 に “食べ” が照合される. これならば句パターン照合にも成功し, 訳出が可能である. しかし, 十分なパターン照合精度を得るまで, 文パターン対を追加するのは非常に困難である. 節レベルのパターン翻訳機能が実装されれば, このような文でも再帰的なパターン翻訳が可能であると期待されており [1], まず節レベルパターン翻訳の実装を優先すべきと考えられる.

7.3.2 翻訳結果

クローズドテストと同様, 良い翻訳結果が得られない最大の原因は, 訳語選択における不適切な語の選択であった. また, どの文パターンを用いるかによって翻訳精度が異なるはずであるが, 本実験では明確な違いはみられなかった.

8 おわりに

本研究において ITM に句レベルパターン翻訳機能を実装した. 文パターン照合において, パターン照合ルールの不足等の理由により, 文パターン照合に失敗する場合がみられた. SPM の調整が必要と考えられる. 評価実験では, 句変数照合部分の局所翻訳において, 不適切な訳語選択により翻訳精度が低下していることが明かになった. 句パターン選択等の処理が必要と思われる. また, 句パターンの照合結果から, 句変数照合部分は, 作成された句パターン辞書の記述に対して長すぎる場合が多数あり, そのため照合結果が得られないことが多数みられた. 節レベルパターン翻訳機能の実装により, この問題が解決される可能性がある. 部分照合で得られたパターンの使用も有効と考えられるが, 照合されなかった形態素の処理方法を検討しなければならない. 今後の方針として, 本研究で明かになった問題の解決と共に, 節レベルパターン翻訳機能の実装と調査, 部分照合したパターンの使用方法の検討を行う必要がある.

参考文献

[1] 池原悟, 阿部さつき, 徳久雅人, 村上仁一：非線形な表現構造に着目した重文と復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] 池原悟：鳥バンク, 日本語表現意味辞書 -重文複文編-, 2007. (http://unicorn,ike.tottori-u.ac.jp/toribank)

[3] 徳久雅人, 村上仁一, 池原悟：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, Vol.2006, No.124, pp.9-16, 2006

[4] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦：日本語語彙大系, 岩波書店, 1997.

[5] 原真一郎, 村上仁一, 徳久雅人, 池原悟：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.