文型パターン辞書により中間言語に変換した統計翻訳の検討

坂田純　村上仁一　徳久雅人　村田真樹
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,murakami,tokuhisa,murata}@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, 文全体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. この問題の解決を目的として, 意味的等価変換方式が考案されている [1]. この変換方式においては, 文の構造が抽象的な意味を包含するとみなし, それらの意味が保存されるように文パターンを記述する. 具体的には, 他の要素に置き換えても抽象的な意味の変わらない線型部分と, 置き換えにより抽象的な意味が変容してしまう非線型部分との区別を付けて, 文パターン化を施す. 機械翻訳への利用にあたっては, 線型部分に対する局所翻訳を行い, 非線型部分と組み合せて文全体の訳出を行うこととなる.

意味的等価変換方式の日英機械翻訳への利用を目的として, 現在, 日英重文複文文型パターン辞書が作成されている [2]. 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. これら 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, 逆に文パターン照合率は, この順で低いことが明かになっている. 照合率においては, 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. 文献 [3] において, 句パターンを用いて句局所パターン翻訳を行う, 句レベルパターン翻訳方式が実装されている. しかしながら, 句パターン照合率が低い, 句局所翻訳精度が低いという問題が明らかになっている. これらの問題に対して, 変数照合部分を統計翻訳によって局所翻訳を行う方式が可能であり, これは用例翻訳等で試みられている [4].

一方, 文献 [5] において, 変数照合された日本語文字列を英語パターンへ挿入することによる, 入力文を中間言語へと変換し統計翻訳を行う方式が試みられている. [5] は主に, 中間言語による方式の実装を目的としたものであり, 十分な問題解析はなされていない. そこで本研究において, 中間言語方式の問題解析を行い, パターン翻訳の一方式としての可能性を検討する.

2　日英重文複文文型パターン辞書

重文, 複文の日英対訳文約 12 万文対を対象に, 3 レベルの文型パターン対が作成されている. 本研究では, 句レベル文パターン約 8 万対を用いる. 文パターンは, 字面, 変数, 関数, 記号の 4 種類の要素を用いて記述されている. これらの要素の内, 主に字面と関数が非線型部分にあたり, 変数が線型部分にあたる. また, 記号は適合範囲の拡大等の役割を持つ. 表 1 に句レベルパターンの例を示す.

日本語の助詞や接続詞は, 文の構造と密接な関係を有しており, 文末表現 “なかっ た” が否定形かつ過去形といった抽象的意味を持つように, 人がその文を理解するのに決定的な働きをもつ場合が多い. 英語文との対応においては, 語型の変化や語順に対応することも多く, その場合, 単純な語と語の対応をとることはできなくなる. このような部分を非線型部分と呼び, 字面, 関数, さらに語順として, 文パターン対の中に記述することによって, 抽象的な意味を保存することができる.

具体的には, 日本語パターンの “.hitei” は動詞変数 $V4$ が否定形であることを意味し, 英語パターンの字面 “never” と対応している. また, “.kako” は過去形であることを意味し, 英語関数 “^past” と対応している. この例では, $V4$ 照合部分の局所翻訳結果を, 過去形に変換することを指示している. 日本語パターンの “$<N1$ は $>$” と英語パターンの “$<$I$|N1>$” は, 主語省略文に対応するための記号である. 主語 $N1$ に対応する語が入力文にあるときは, 英語パターンにおいて $N1$ が選択され, そうでないときは, “I” が選択されることを意味している.

表 1　句レベルパターン例

日本語文	彼のお母さんがああ若いとは思わなかった。
英語文	I never expected his mother to be so young.
日パターン	$<N1$ は $>NP2$ がああ $AJ3$ とは $V4$.hitei.kako。
英パターン	$<$I$\|N1>$ never $V4$^past $NP2$ to be so $AJ3$.

3　句レベル文パターンを用いたパターン翻訳

現在, 単語レベルまたは句レベル文パターンを用いたパターン翻訳方式が実装されている. 単語レベルパターン翻訳は, 翻訳精度は高いがパターン照合率が低いことが明かになっている [6]. そこでパターン照合率増加のために, 句パターンを用いて句局所翻訳を行う, 句レベルパターン翻訳方式が実装された [3]. パターン照合率は 5% から 20% に改善されたが, 句パターン照合の失敗による翻訳失敗が多数, 句パターン照合に成功しても句局所翻訳精度が低い, という結果となった.

一方, 句レベル文パターンを用いて, 中間言語に変換してから統計翻訳を行う, 中間言語方式も実装された [5]. しかし, この方式の問題解析は十分には行われていない. そこで本研究において, 問題解析を行い, 本手法が句レベルパターン翻訳の一手法として有効であるかどうか検討する.

4　中間言語を使用する統計翻訳

本手法は, パターン照合に次いで統計翻訳による訳出を行う, 直列のハイブリッド翻訳とみなすことができる. また, 中間言語に変換し文全体に統計翻訳を行うことにより, 日本語から英語への構造変換をして統計翻訳を行って

いるとみなすこともできる．

ところで, パターン翻訳には, 照合された複数パターンから適切な使用パターンを選択することが難しい, という問題がある．そこで統計翻訳により, 局所翻訳のみならずパターンの選択も行うこととする．

4.1 中間言語作成手順

ここで言う中間言語とは, 英語パターンに変数部分の日本語文字列を挿入した日英混じり文のことである．日本語文を “J”, 英語文を “E”, パターンから作成された中間言語文を “J′” で表す．

中間言語を用いて統計翻訳を行うには, 翻訳モデルの学習を J′–E 間で行う必要がある．学習に用いる中間言語文抽出には, 文パターン作成元になった日本語文と, それに対応する日英文パターンを用いる．英語パターンの変数部分に, 対応する日本語文字列を代入し作成する．学習に用いる中間言語文抽出の例を表 2 に示す．なお, 翻訳モデル作成に使用するため, パターン中の各種記号は削除してある．また, 英語関数に関しては, 時制等の重要な情報を持つ非線型要素であるため, 削除しないこととする．

入力文の中間言語文への変換は, パターン照合を行った後, 学習用中間言語文と同様の手順により行うことができる．

表 2 中間言語文の抽出例

日本語文	彼のお母さんがああ若いとは思わなかった。
日パターン	*NP*2 がああ *AJ*3 とは *V*4.hitei.kako.。
英パターン	I never *V*4^past *NP*2 to be so *AJ*3 .
変数部分	*NP*2 = 彼 の お母さん *AJ*3 = 若い *V*4 = 思わ
中間言語文	I never 思わ^past 彼 の お母さん to be so 若い .

4.2 翻訳手順

次に学習と翻訳の手順を示す．

学習 1 日英パターンから J′ 文を抽出する．

学習 2 J′–E 間で翻訳モデルを, E から言語モデルを作成する．

翻訳 1 入力文にパターン照合を行う．

翻訳 2 照合パターン全てに対して, 一パターンにつき一つの中間言語文を作成する．

翻訳 3 これら中間言語文に対して統計翻訳を行う．

翻訳 4 一入力文につき複数の候補文が得られた場合は, 統計翻訳における翻訳スコアが最大の候補文を出力文とする．

文献 [5] では, 翻訳 2 の前に, 意味属性制約 [7] による使用パターンの絞り込みを行っている．しかしながら, パターンの絞り込みを行わない方が翻訳精度が高いという結果となっている．そこで本研究では, 翻訳スコアにより候補文から適切な出力文を選択できると仮定し, パターンの絞り込みを行わないこととする．

本手法全体の流れ図を図 1 に, 翻訳の具体例を表 3 に示す．

5 翻訳実験

オープンテストにより, 翻訳精度の調査と解析を行う．

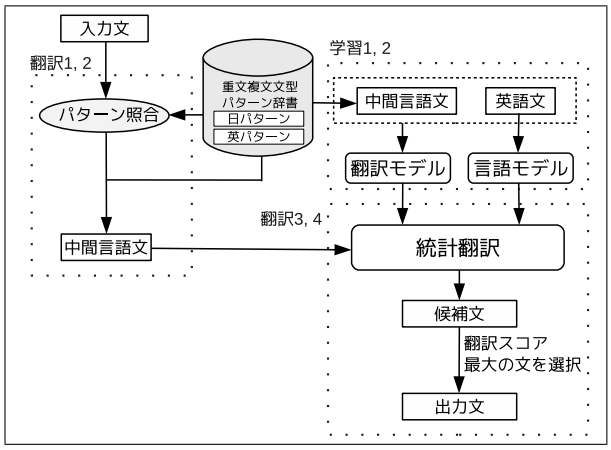

図 1 翻訳の流れ図

表 3 翻訳例

入力	アイデアはいいが実現不可能だ。	翻訳スコア
J′ 文	though アイデア be いい , he 実現不可能 . アイデア いい , but 実現 不可能 . アイデア be いい but 実現 不可能 .	
候補	though ideas are good , he impossible . ideas well but impossible . ideas is good but impossible .	-21.49 -18.40 -18.33
出力	ideas is good but impossible .	-18.33

5.1 実験方法

統計翻訳にはフレーズベースの moses を用いる．moses 付属の mert-moses.pl を用いてパラメータチューニングを行う．なお, 中間言語への変換により構造変換が行われたと仮定し, distortion-limit の値は 6 のままとする．

翻訳モデル作成に使用する, 中間言語文と英語文の対は, 77,432 文対である．また言語モデル作成には, 文型パターン辞書の英語原文 121,913 文を使用する．形態素解析は文献 [8] のものを使用し, パターン照合には文献 [9] のパターン照合器 SPM を使用する．

テスト文は重文複文 3951 文を用いる．パターン照合に成功した文のみ翻訳を行う．文献 [5] との実験方法の違いを表 4 にまとめる．

表 4 実験方法の違い

	本研究	文献 [5]
意味属性による絞り込み	行わない	行う
英語語型関数	付与	削除
distortion-limit	6	-1

5.2 評価方法

評価は人手評価により行う．ランダム抽出した訳出文 100 文を評価の対象とする．評価基準を下に示す．

評価 1 文全体の意味が, 問題無く理解できる．

評価 2 文法的に何かしらの問題があるが, 元の文の大意は理解できる．

評価 3 文の意味が理解できない, もしくは意味が異なる．

評価の例を表 5 に示す．

表 5 評価例

< 評価 1>

入力	事件が起きてから十年たった。
参照	Ten years have passed since the accident occurred.
出力	ten years have passed since the incident occurred .

< 評価 2>

入力	彼女が気づいたとき、彼はもう彼女の写真をとっていた。
参照	When she noticed it, he had already taken her picture.
出力	when she realized , he had already her photo .

< 評価 3>

入力	うちへ帰るとすぐテレビのスイッチを入れた。
参照	When I came back home, I turned on the TV right away.
出力	on my home , he the tv .

5.3 実験結果

3951 文の入力文のうち, 943 文がパターン照合に成功した. ランダム抽出した 100 文に対する人手評価の結果を, 表 6 に示す.

評価 3 がおよそ 8 割を占めており, 翻訳精度が非常に低いことがわかる.

表 6 人手評価結果

	評価 1	評価 2	評価 3
文数	5	12	83

6 考察

6.1 翻訳精度の低い原因

翻訳精度が低い原因として推測されるものを, 次の 3 つにまとめることができる.

1. 複数の候補文から, 適切な候補文が存在するにも関らず不適切な候補文を選択
2. 候補文の元となった中間言語文の構造が不適切
3. 中間言語文の構造は適切であるが局所翻訳の精度が低

候補文からの出力文の選択, 中間言語文の構造, 局所翻訳精度の順に分析を行う.

6.1.1 候補文からの出力文の選択

表 6 における, 評価 2 の文 12 文と評価 3 からランダム抽出した 20 文に対して, 候補文中に適切な候補が存在するかどうか調査を行った. 表 7 に結果を示す.

表 7 の結果より, 適切な候補文があるにも関らず不適切な候補文を選択している事例は存在しなかった. つまり 32 例全てが, 不適切な候補文しか存在しないため, 適切な候補文を選択できないことを示している.

調査を行った文に限れば, 適切な候補文を持つ 5 例において全て, 適切な候補文を出力文として選択できたことになる. したがって, わずか 5 例ではあるが, 翻訳スコアによる出力文選択の効果が確認された.

表 7 適切な候補文の有無

	評価 2	評価 3
適切な候補文あり	0	0
適切な候補文なし	12	20

6.1.2 使用された中間言語文の構造

不適切な候補文が生成される原因を調査するため, 出力文に使用された中間言語文が適切であるかどうか調査を行った. 調査には, 評価 1 の 5 文, 6.1.1 節で用いた評価 2, 3 の 32 文を用いた. 具体例を表 8 に, 調査結果を表 9 に示す.

表 8 文構造が妥当な文と妥当でない文の例

< 文構造が妥当 >

入力文	彼は立ち上がってあたりを見回した。
参照文	He rose to his feet and looked all around him.
J′ 文	彼 立ち上がっ^past and あたり を 見回し^past .
出力文	he stood up and looked around .

< 文構造が妥当でない >

入力文	あの映画は何回見ても面白い。
参照文	That movie is interesting, no matter how many times I watch it.
J′ 文	it will be 面白い to あの 映画 は 何 回 見 .
出力文	it will be interesting to see the movie in several times .

表 9 出力文に使用された中間言語文の構造の調査結果

	J′ 文の構造が妥当	構造が妥当でない
評価 1	5/5	0/5
評価 2	9/12	3/12
評価 3	7/20	13/20

評価 1 の 5 文は全て妥当な構造を持ち, 評価 2 は 75% が妥当な構造を持つ. しかし評価 3 は 35% しか妥当な構造を持たない. この結果から, 文構造が妥当であるほど翻訳精度が高くなっていることがわかる.

よって, 不適切な構造の中間言語文を用いていることが, 不適切な候補文が生成される原因の一つと考えることができる.

6.1.3 統計翻訳による局所翻訳精度

表 9 の評価 2, 3 において, 文の構造が妥当である文は合計 16 文である. この数は, “文構造は妥当だが局所翻訳精度が低い文” の数であり, 21 文中 16 文は局所翻訳精度が低いことを示している. よって統計翻訳による局所翻訳精度が非常に低いことも, 不適切な候補文生成の主な原因の一つである.

表 10 に局所翻訳精度が低い例を示す. 動詞句照合部分の局所翻訳精度が低い場合が多く, 特に日本語動詞に対応する英語表現が消失する傾向がみられる. 動詞は文全体の意味に大きく作用することから, 文の翻訳精度向上には, 動詞句の局所翻訳精度の向上が必要となる.

表 10 文構造は妥当だが局所翻訳精度が低い文の例

入力文	まずボタンを押して、次にレバーを引いてください。
参照文	First, push the button, and then pull the lever.
J′ 文	まず ボタン を 押し and 次 に レバー を 引い .
出力文	first button and then the lever .

6.2 未知語

局所翻訳精度が非常に低い理由の一つに, 未知語が多数出現していることが挙げられる. 6.1.1 節の分析に用いた 32 文において, 未知語は 10 語含まれていた. そのうち 7 語が “動詞^語形変換関数”(例 : 急落し^ing) の形であった. 関数を付与した状態で統計翻訳を行った結果, データ数減少により未知語となった可能性が高い. しかし, 関数の付与により正しい語形が選択される効果も, 多数確認されている. 改善の最大の手段は学習文の追加であるが, パ

ターンの大量の追加は難しいことから, 翻訳モデルの学習用中間言語文を追加することも困難である. よって, 多数の未知語の原因となる英語語型関数に対して異なる処理を用いなければ, 未知語を減少させることは困難である.

6.3 意味属性制約によるパターンの絞り込み

本研究においては, 意味属性制約によるパターンの絞り込みを行っていない. 中間言語文は照合されたパターンから作成しているので, パターンの絞り込みにより不適切な中間言語文の使用を減少させることができる. 不適切な中間言語文のみを持つ場合は翻訳精度の低い出力文となることから, そのような場合はパターン翻訳を行わない方がよいと思われる. 意味属性制約によるパターンの絞り込みにより, 不適切な中間言語文のみを持つ文のパターン翻訳を減少させることができる.

ただし, 絞り込みにより不適切なパターンのみならず適切なパターンをも刈り取ってしまう可能性がある. なお, 意味属性の階層性を利用して, 制約の強さを調整することが可能である. 意味属性をパターンの絞り込みに利用するために, 制約の適度な強さを調査する必要がある.

6.4 文型パターン辞書を用いた中間言語方式の問題

[2] の文型パターンは, 線型部分 (変数照合部分) を独立に翻訳することが可能だと仮定して作成されている. そして各種関数および記号は, 変数ごとに個別で処理を行うことを想定して記述されている. また, 文献 [9] のパターン照合器は, 変数照合部分の構文解析を行ってからパターン照合を行っている. したがって各種変数ごとに, 照合部分の日本語文字列が固有の構造を持つ.

中間言語方式では, 照合パターンから中間言語への変換後, 文全体を統計翻訳により翻訳を行っている. そのため関数や変数ごとに個別の処理を行うことが難しい. 変数照合部分ごとに局所翻訳を行えば, 関数の機能や各種変数固有の構造を利用することができる. 文献 [3] は, 句パターンによって句局所翻訳を行う方式である. この方式でも句局所翻訳の精度は低いという結果であったが, その原因は明らかになっており, 翻訳精度の改善が可能である.

よって句レベル文パターンを用いたパターン翻訳は, 句変数照合部分に対して, 句パターンが照合されれば句パターンを用いて, そうでなければ統計翻訳を用いて個別に局所翻訳を行う方式が最善と考えられる.

6.5 考察のまとめ

6.1.1 節の結果から, 翻訳スコアによる出力文選択の効果が確認された.

よって次の 2 点が, 文の翻訳精度が低い理由となる.

1. 不適切な照合パターンを用いた翻訳.
2. 変数照合部分の局所翻訳精度が非常に低.

特に 2 の結果から, 中間言語方式を継続する必要はないと考えられる.

以上を踏まえた, 句レベルパターン翻訳の今後の方針を下に示す.

1. 意味属性制約によるパターン絞り込みの効果を調査する.
2. 句局所翻訳を, 句パターンを用いたパターン翻訳, 統計翻訳の順に句変数ごとに行う, 並列型のハイブリッド翻訳方式を実装する.

7 おわりに

本研究において, 文型パターンを用いて中間言語に変換した統計翻訳を行い, 問題解析を行った.

実験結果から翻訳精度が非常に低いことが明かになった. その理由は, 不適切なパターンを用いて翻訳を行ったことと, 適切なパターンであっても統計翻訳による局所翻訳精度が非常に低いことであった. 中間言語方式の有効性は確認されず, 改善も困難であると考えられる. 一方, 翻訳スコアによる訳出文の選択は, 効果が確認された.

今後の方針として, 句パターンまたは統計翻訳を用いて, 句変数照合部分に対して並列型のハイブリッド翻訳を行う方式を実装する.

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] 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] 坂田純, 徳久雅人, 村上仁一, “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”, 言語処理学会第 18 回年次大会, pp.271-274, 2012.

[4] Philipp Koehn and Jean Senellart, “Convergence of translation memory and statistical machine translation”, In Proceedings of AMTA Workshop on MT Research and the Translation Industry, pp.21-31, 2010.

[5] 吉田大蔵, 村上仁一, 村田真樹, 徳久雅人, “文型パターン辞書により原言語を中間言語に変換した日英統計翻訳”, 言語処理学会第 18 回年次大会, pp.487-490, 2012.

[6] 石上真理子, 水田理夫, 徳久雅人, 村上仁一, 池原悟, “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, pp.67-70, 2007.

[7] 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.

[8] 池原悟, 宮崎正弘, 白井諭, 林良彦, “言語における話者の認識と多段翻訳方式”, 情報処理学会論文誌, 28(12), pp.1269-1279, 1987.

[9] 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.