非線形言語モデルに基づく句レベル文型パターン翻訳方式の問題分析

坂田純　村上仁一
鳥取大学大学院工学研究科
{d112004,murakami}@eecs.tottori-u.ac.jp

1　はじめに

本論文では，非線形言語モデルに基づく句レベル文型パターン翻訳方式の問題分析結果を報告する．非線形言語モデルに基づいて，重文複文文型パターン辞書が作成されている [1]．そして重文複文文型パターン辞書の句レベル文型パターンを用いる日英機械翻訳方式が，[2] の論文において実装されている．しかし論文 [2] では，動作確認が行われた段階であり，翻訳精度が非常に低い状態であった．その後いくつかの改良が行われ，翻訳精度の改善が確認されている．そこで本論文では同ドメイン内実験（オープンテスト）を行い，翻訳性能の調査と問題分析を改めて行う．

2　重文複文文型パターン辞書

非線形言語モデルに基づき，重文複文文型パターン辞書が作成されている．重文複文文型パターン辞書では，述部を 2 つか 3 つ持つ日本語文とその英訳文約 12 万文対を対象に，文型パターン化が施されている．文型パターン辞書では，変数化の範囲に合わせ，単語レベル，句レベル，節レベルの文型パターンが作成されている．句レベル文型パターン数は約 8 万である．

文型パターンは，変数，関数，字面，記号の 4 種類の記述形式を用いて記述されている [3]．句レベル文型パターンは単語変数と句変数を持つ．また日本語文型パターンの変数は，品詞と意味属性の情報を持ち，英語文型パターンの変数は品詞情報を持つ．日本語句変数の意味属性には，句変数化された要素内における，係り先の単語の意味属性のみを付与している．変数は，パターン照合において，入力文と一致しなくても適合する要素である．ただし適合において，品詞と意味属性の制約を持つ．関数は，語形の指定や，述部語尾表現の表記の揺れを吸収するための汎化等の役割を持つ．変数化または関数化されない文要素は，字面のまま文型パターンの中に記述される．また，記号は適合範囲の拡大等の役割を持つ．表 1 に句レベル文型パターンの例を示す．

表 1　句レベル文型パターンの記述例

日本語文型パターン原文	彼はパソコンを使い複雑な計算をうまく処理した。
英語文型パターン原文	He successfully made those complicated calculations by using a personal computer.
日本語文型パターン	/*N*1(他称単数男, 男) は/*VP*2(所有的移動, 他多数)^renyou/*VP*3(所有的移動, 他多数).kako。
英語文型パターン	*N*1 *VP*3^past by *VP*2^ing.

(a)　日英文型パターンの *N*1 が名詞単語変数であり，*VP*2 および *VP*3 が動詞句変数である．日本語文型パターンの変数に続く括弧内の記述が意味属性（日本語語彙大系 [4]）である．*VP*2 を例にとると，日本語原文の “パソコンを使い” が *VP*2 に変数化されており，“使い (使う)” の意味属性 “所有的移動” が *VP*2 の意味属性として記述されている．

(b)　日本語文末表現の “た” は，「過去の事例」を示す表現であり，日本語文型パターン末部の様相関数 “.kako” がこの表現に対応している．この関数は，日本語動詞句変数 *VP*3 に，過去形の語尾が続くことを示している．“.kako” に対応する英語原文の過去表現は動詞 “make” の過去形 “made” であるので，英語語形関数 “^past” で記述されている．この英語関数は，変数 *VP*3 に対応する要素の翻訳結果を，過去形に変換することを意味している．

3　句レベル文型パターン翻訳方式

本論文では，重文複文文型パターン辞書の句レベル文型パターンを使用して翻訳を行う．句レベル文型パターンは単語変数と句変数を持ち，単語辞書と句パターンを用いて翻訳文を得る [2]．また訳語選択は確率モデルを用いて行う．

システム全体のモデル図を図 1 に，手順とその概要を以下に示す[*1]．

1　入力文の形態素解析を行い，各形態素への品詞と意味属性を付与する．
2　日本語パターン検索システム [6] により，入力文に対しパターン照合を行い，入力文に適合する句レベル日英文型パターンとその変数情報を得る．
3　意味属性を用いて，日本語文型パターンの絞込みを行う．

[*1] 翻訳方法の詳細は論文 [5]，[2] を参照．

4 英文生成システムにより，日英文型パターン，単語辞書，句パターンを用いて英語文を生成する．

4.1 単語変数に対応する形態素は，単語辞書を用いて翻訳する．

4.2 句変数に対応する形態素列は，句パターンを用いて翻訳する．その際，単語レベル文型パターン翻訳方式を利用する [5].

4.3 単語翻訳確率と単語連鎖確率（tri-gram）を用いて，訳語候補から最尤の組み合わせを選択する．

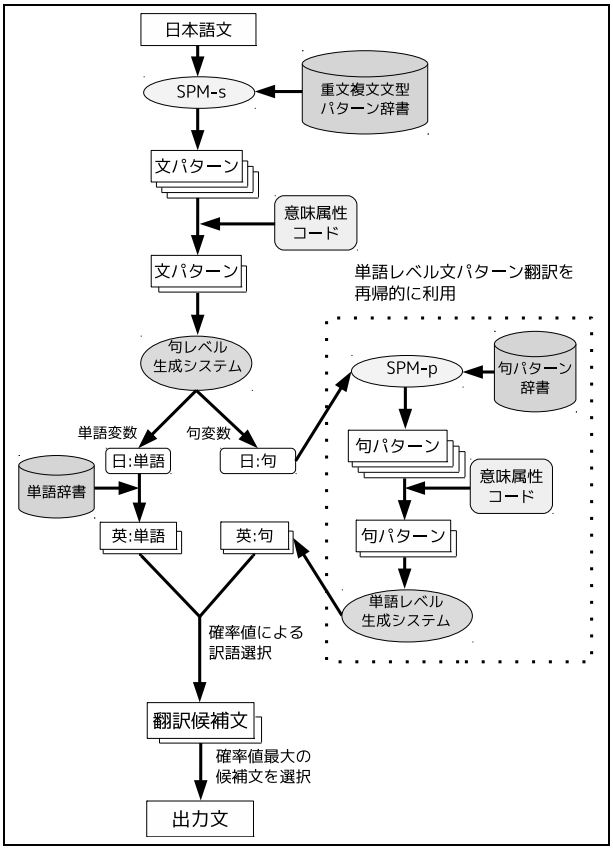

図 1　句レベル文型パターン翻訳方式

4　翻訳実験

4.1　実験条件

入力文に重文複文文型パターン辞書の日本語文型パターン原文を用いて，翻訳実験を行う．文型パターン辞書の日英文型パターン原文 10 万文対から作成された，65,171 の句レベル文型パターンを翻訳に使用する．単語辞書, 句パターン, 単語翻訳確率, 単語連鎖確率 (tri-gram) の学習は，この 10 万原文対およびこれらから作成された文型パターンを元に行う．学習に使用しない残りの日本語文型パターン原文約 2 万文から，1,000 文を抽出しテスト文に用いる．

4.2　パターン適合率

入力 1,000 文のうち，パターン照合と意味属性を用いたパターン絞込みに成功した文は，521 文であった．この 521 文のうち，全ての句変数適合部分に対し句パターンが適合した文は 277 文であった．本論文では以後パターン適合率を，入力文に対する，パターン照合とパターン絞込みの後，少なくとも 1 つの文型パターンが適合しかつ全ての句変数対応部分に対し句パターンが適合している文の割合と定める[*2]．よって本実験では，パターン適合率は約 28%(277/1,000) となる．単語レベル文型パターン翻訳方式では約 11% のパターン適合率であることが示されており [5]，この結果は，単語レベル文型パターン翻訳よりは改善されたが，まだ十分なパターン適合率には達していないことを示している．

4.3　評価結果

文型パターンが適合した 277 文中，学習文に同じ日本語文を持つのは 25 文であり，持たない文は 252 文であった．上記 252 文からランダムに 100 文抽出し人手評価を行った．人手評価は adequacy による 5 段階評価を行う．人手評価の基準を表 2 に，人手評価の結果を表 3 に示す．本翻訳方式では評価 4 と 5 を合わせても 27 文であり，翻訳精度の高い文はあまり多くない．単語レベル文型パターン翻訳方式では，約 51% の文で翻訳精度が高いことが示されており [5]，単語レベル文型パターン翻訳方式より翻訳精度が低いことがわかる．

表 2　人手評価の基準

評価 5	入力文の意味を正しく理解できる
評価 4	一部不適切な部分があるが，概ね理解できる
評価 3	入力文の意味が何となく読み取れる
評価 2	部分的な理解にとどまる，または意味が入力文と大きく異なる
評価 1	入力文の意味をほとんど理解できない

表 3　人手評価結果

	評価 5	評価 4	評価 3	評価 2	評価 1	平均値
文数	18	9	21	36	16	2.77

表 4 に本翻訳方式の人手評価の値が高い例を示す．表中の「参照文」は入力文の対訳英語文である．「日文型パターン」と「英文型パターン」は，出力文に使用した日本語文型パターンと英語文型パターンであり，「日パターン原文」と「英パターン原文」は，この文型パターンの作成元となった日本語原文と英語原文である．

5　問題分析

本節では翻訳精度が低い原因を調査する．翻訳精度が低い原因は主に，文型パターンに関する問題と，変数に

[*2] 単語変数対応部分の翻訳には単語辞書を用いるため，単語変数はパターン適合率には関係しない．

表 4　本翻訳方式において人手評価値の高い例

入力文	先生は学生にもっと勉強するようにと言いました。
参照文	The teacher told the students to study more.
出力文	The teacher told the students to study harder .
人手評価値	5
日文型パターン	/$1^{/N1 は }/*N*2 に $1!VP3.suitei と $1!*VP*4.kako。
英文型パターン	*N*1 *VP*4^past *N*2^obj to *VP*3^base .
変数情報	*N*1=先 生,*N*2=学 生,*VP*3=もっと勉強する,*VP*4=言い (言う)
日パターン原文	彼は私に中に入るようにと手で招いた。
英パターン原文	He beckoned me to come in.

対応する要素の翻訳の問題に二分できる．まず使用した文型パターンが各入力文に対し適切であったか調査する．この結果から，翻訳精度の低い文（評価値 1 から 3）のうち，「文型パターンが “不適切” な文」が文型パターンに関する問題，「文型パターンが “適切” な文」が変数に対応する要素の翻訳の問題とみなすことができる．本論文では，この二点の問題のうち，文型パターンに関する問題の分析を行う．

5.1　使用した文型パターンの適切さ

翻訳に使用した文型パターンが，その入力文に対し適切であるかどうかを調査した．英語文型パターンに対し，適切な翻訳文を出力可能だと判断できる場合を “適切”，そうでないときは “不適切” とした．人手評価を行った 100 文における，英語文型パターンの適切さと人手評価値の関係を表 5 に示す．

表 5 より，半数近い 44 文で不適切な文型パターンを使用しており，この 44 文全てで翻訳精度が低い（評価 1 から 3）．また適切な文型パターンを使用した場合でも，約半数の文（56 文中 29 文）で翻訳精度が低く，これらは変数に適合した要素の翻訳または訳語選択の問題による．

表 5　英語文型パターンの適切さと人手評価値の関係

	評価 5	評価 4	評価 3	評価 2	評価 1	合計
適切	18	9	12	14	3	56
不適切	0	0	9	22	13	44

5.2　不適切な文型パターンを使用した要因

不適切な文型パターンを使用した 44 文のうち 19 文は，形態素解析ミス等の句レベル文型パターン翻訳と直接関係しない原因による．これら 19 文も重大な問題であるが，本論文では省略する．

残り 25 文では，入力文と使用した日本語文型パターン原文の表現する事象が，大きく異なる場合がほとんどであった．このことは意味属性を用いたパターン絞込みでは，入力文に類似する文から作成された文型パターンを選択できていないことを示唆している．パターン絞込みが適切な文型パターンの選択方法として機能しないと，文型パターンが不適切となる場合が増大する．

表 6 に文型パターンが不適切な要因とその内訳を示す．表 6 の結果から，多様な要因によって，文型パターンが不適切となっていることがわかる．これらの要因のうち，表 7 に要因 (2) の例，表 8 に要因 (4) と (5) を共に持つ例を示す．

表 6　文型パターンが不適切となる要因とその内訳

要因 (大分類)	要因	数
句変数化範囲	(1) 文型パターン作成時の句変数化の範囲の問題	2
	(2) パターン照合時の句変数認識部分の範囲	11
英語文型パターンの構造	(3) 英語文型パターンで指定される語形が入力文の翻訳では不適切	7
	(4) 英語文型パターンで字面で記述されている助詞や接続詞がその入力文の翻訳には不必要	4
	(5) その他英語文型パターンの構造が不適切	5
その他	(6) 複文の入力文に重文から作られた日本語文型パターンが適合	2
	(7) 省略要素と補完記号に関連する問題	7

5.3　句変数に認識した範囲に問題がある例

句変数に認識した範囲に問題がある例を表 7 に示す．

入力文において “返事をやきもきし” が *VP*2 に適合しているが，この文の意味を考えると “やきもきしながら”，“返事を待っていた” のであり，句変数への適合の仕方に問題がある可能性がある．入力文の “返事を” は直後の “やきもきし” と述部の “待っていた” の両方に係るとみなせるが，後者との結びつきがより強い．そのため入力文に句変数をどのように適合しても，適合の仕方が不適切となる．一方，日本語文型パターン原文はその意味を考えると，“その死体をじっと見つめながら” が “立っていた” に係る文であり，英語文型パターン原文の分割の仕方と対応している．したがってこの文型パターンの句変数化の範囲は適切であり，この入力文へのこの文型パターンの適用が不適切である．この問題は係り受け解析では回避できる可能性があるが，係り受け解析の利用は，他の事例を検証してから判断すべきである．

表 7　句の認識範囲に問題のある例

入力文	返事をやきもきしながら待っていた。
出力文	She waited developing in the answer .
参照文	He was nervously waiting for a reply.
人手評価値	2
日文型パターン	/</*N*1 は >!*VP*2 ながら/*V*3.teiru.kako。
英文型パターン	<She|*N*1> *V*3^past *VP*2^ing.
変数情報	*VP*2=返事をやきもきし (やきもきする),*V*3=待っ (待つ)
日パターン原文	その死体をじっと見つめながら立っていた。
英パターン原文	She stood staring at the corpse.

5.4 英語文型パターンの構造に問題がある例

英語文型パターンの構造に問題がある例を表 8 に示す．入力文と出力文を見比べると，出力文からは入力文の意味がほとんど読み取れない．最大の原因は英語文型パターンの “@beˆpast *VP*2(群衆に向け)ˆed” の部分で，受動態への変換を指定する記述になっているため，この部分の翻訳結果が “were given to the crowd” となっている．また “into *N*3(放水する)” の部分も，この入力文では不適切である．日本語パターン原文の “兵糧攻めにあっ(会う)” という表現は受身の意味も含んでおり，英語パターン原文ではそのニュアンスが含まれるように “were starved into” の受身表現で訳されている．したがって日英文型パターンでこの表現に対応する部分は，日本語文型パターンで “*VP*2”，英語文型パターンで “@beˆpast *VP*2ˆed” と記述されている．入力文で *VP*2 に適合した “群衆に向け (向ける)” の表現は受身の意味を含まないため，上記の英語文型パターン中の該当部分の記述により受身形に変換されると，入力文とは異なる意味の表現となる．また “starve [目的語] into [名詞]” で “飢えさせて [目的語] に [名詞] させる” の意味を持つ熟語であり，“into [名詞]” は “starve” と結びついている表現形式であるため，“starve” またはこの表現形式と同じ形式を持つ英語動詞が訳語候補として存在しなければ適切な翻訳は困難である．このことからこの表現は句変数化すべきでないが，句変数化可能な要素と可能でない要素の区別をつけることが非常に難しい．

表 8　英語文型パターンの構造に問題がある例

入力文	機動隊は群衆に向けて放水した。
出力文	The riot police were given to the crowd into drainage .
参照文	The riot police turned the water cannon on the crowd.
人手評価値	1
日文型パターン	/$1ˆ{/*N*1(軍) は }!*VP*2(存在, 属性, その他多数)(て | で)$1/(*V*3.kako|*ND*3 をした)。
英文型パターン	*N*1 @beˆpast *VP*2ˆed into *N*3 .
変数情報	*N*1=機動隊,*VP*2=群衆に向け (向ける),*V*3=放水し (放水する)
日パターン原文	反乱軍は兵糧攻めにあって降服した。
英パターン原文	The rebels were starved into submission.

5.5 今後の調査課題

翻訳性能改善のために調査すべき課題を下に示す．

1. パターン適合率の改善
2. パターン絞込みの改善
3. 使用した文型パターンが不適切となる要因のさらなる調査
4. 変数対応要素の翻訳や関数処理の適切性

1 について，本翻訳方式のパターン適合率は約 28% であり，パターン適合率はまだ低いとみなせる．句レベル文型パターン翻訳方式のパターン適合率は，句パターンの適合率も含むため，句パターンの照合について分析する必要がある．また節レベル文型パターンを用いることでパターン適合率が上昇することが明らかになっていることから，句レベル文型パターンが適合しない文では，節レベル文型パターンを適用させることが現実的である．

2 の問題において，現在のパターン絞込み方法に問題がある可能性が高い．文型パターンの句変数への意味属性の付与方法と，入力文における句変数対応要素の意味属性を決定する方法を見つける必要がある．word-embedding 等の利用で改善する可能性がある．

3 の問題は 2 の問題と強く結びついているため，2 の問題とセットで調査すべきである．2 の問題の改善によっても，同様の問題が生じる可能性が高いため，3 の問題の調査も必要となる．

4 の問題においては特に，句変数に対応する要素の翻訳の精度を調査する必要がある．その精度の調査には，句パターンの抽出精度や照合の問題，句パターン内部の関数または文型パターンの句変数に付与された関数の処理方法の問題等，多数の項目の調査が必要である．

6 おわりに

本論文では，非線形言語モデルに基づく句レベル文型パターン翻訳方式の翻訳精度の調査と問題分析を行った．翻訳実験より，本翻訳方式のパターン適合率は約 28% とあまり高くないことが明らかになった．また翻訳精度の高い文の割合は 27% であり，翻訳精度もあまり高くないことがわかった．問題分析より，人手評価を行った 100 文中 44 文で不適切な文型パターンを使用しており，パターン絞込み方法に問題があると考えられる．使用した文型パターンが不適切となる要因で最も多く見られたのは，入力文の句への分割の仕方が不適切であることであった．今後はさらなる分析より問題点を明らかにし，それら問題点の改善とパターン絞込み方法の改善を行う必要がある．

参考文献

[1] 池原悟. 鳥バンク (Tori-Bank). http://unicorn/toribank.

[2] 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 回年次大会発表論文集, pp.271-274, 2012.

[3] 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.

[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] 坂田純, 村上仁一. 非線形言語モデルに基づく文型パターンを用いる日英機械翻訳方式. 電子情報通信学会, Vol.J101-D, No.1, Jan. 2018.

[6] 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自然言語処理 (NL), Vol.2006, No.124, pp.9-16, 2006.