

節レベル文型パターン翻訳方式の試作評価

伊藤 直輝*1 坂田 純*2 村上 仁一*2

*1 鳥取大学 工学部 知能情報工学科

*2 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻

{s142005,d112004,murakami}@ike.tottori-u.ac.jp

1 はじめに

過去にルールベース翻訳の一つの手法として、日英文型パターン翻訳方式が提案されている [1]. この手法は、日英重文複文文型パターン辞書 (以下、文型パターン辞書)[2] を利用し、翻訳する手法である. 文型パターン辞書は変数の範囲に応じて、単語、句、節レベルの3つのレベルで記述されている. 過去において、句レベル文型パターンを用いた翻訳手法が提案されている [3]. しかし、句レベル文型パターンを用いた翻訳手法では翻訳できる文の数が少ない. そこで本研究では、より多くの文を翻訳するために、節レベル文型パターンを用いた翻訳方式を試作する. さらに、翻訳できる文の数及び翻訳精度の観点で句レベル文型パターン翻訳方式と比較調査する.

2 日英重文複文文型パターン辞書

日英重文複文文型パターン辞書は約 12 万件の重文複文を対象として、日英機械翻訳を視野に入れ作成されている. 文型パターンは以下の3つのレベルで記述されている.

- 単語変数を用いた単語レベル文型パターン.
- 単語、句変数を用いた句レベル文型パターン.
- 単語、句、節変数を用いた節レベル文型パターン.

単語、句、節レベル文型パターンの記述の例を表 1 に示す.

表 1 文型パターン辞書の記述例

日本語原文	大阪に着くまでの間、ずっと本を読み続けた。
英語原文	I read until I reached Osaka.
単語レベル (日)	<N1 は>N2 に V3 ^{rentai} の間、<N4 は>ずっと本を V5(つづけ 続け) た。
単語レベル (英)	<N4[I]> V5 ^{past until} <N1[I]> V3 ^{past} N2.
句レベル (日)	<N1 は>VP2 ^{rentai} (まで 迄) の間、<N3 は>ずっと本を V4(つづけ 続け) た。
句レベル (英)	<N3[I]> V4 ^{past until} <N1[I]> VP2 ^{past} .
節レベル (日)	CL1 ^{rentai} (まで 迄) の間、CL2.kako。
節レベル (英)	CL2 ^{past until} CL1 ^{past} .

3 句レベル文型パターン翻訳方式

3.1 概要

過去において句レベル文型パターン翻訳方式が提案されている [3]. 翻訳は大きく分けて 6 つのステップで行う. 以下に翻訳の手順を示す. また、翻訳の例を図 1 に示す.

手順 1 入力日本語文を形態素解析

手順 2 手順 1 の結果と句レベル文型パターンを照合

(SPM[4])

手順 3 得られた複数の句レベル文型パターンについてそれぞれ文頭から文末にかけて要素を訳出

手順 3a 要素が単語変数ならば単語辞書を用いて翻訳候補を訳出

手順 3b 要素が句変数ならば、句パターン辞書と照合を行い、得られた複数の句パターンに存在する単語変数を単語辞書を用いて翻訳候補として訳出.

手順 4 英語文型パターンの変数に翻訳候補を全通り代入

手順 5 言語モデル (tri-gram) 及び、単語翻訳確率を用いて評価値を算出

手順 6 評価値の最も高い文を出力

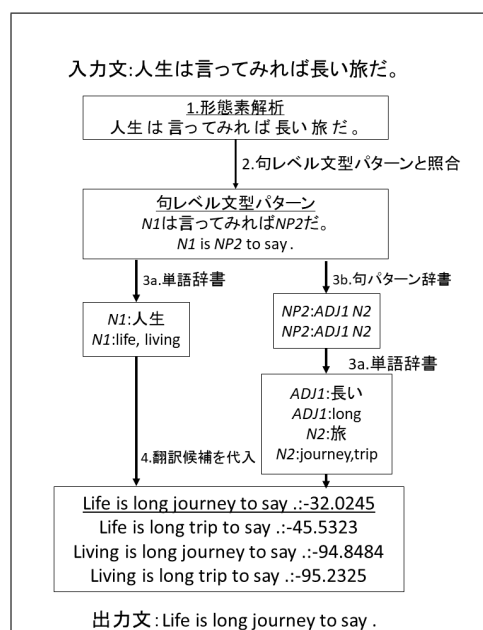


図 1 句レベル文型パターン翻訳の例

3.2 句パターン辞書

句パターン辞書とは、句レベル文型パターンに存在する句変数の構造を単語レベル文型パターンを用いて表現するための辞書である. 句パターン辞書の作成手順を以下に示す.

手順 1 文型パターン辞書の作成途中で抽出された中間ファイルより句変数に関する情報を抽出.

手順 2 手順 1 で得られた情報を用いて単語レベル文型パターンより句変数に対応する記述を抽出.

手順3 句変数の表記や句変数に対応する単語レベル文型パターンに誤りがある句パターンを削除。

手順1 で用いる中間ファイルを表2, 作成した句パターンの例を表3 に示す。

表2 中間ファイルの例

記号	日本語記述	変数	英語記述
C	私はコーヒーを飲む	CL	I drink coffee
P	私	N	I
P	コーヒーを飲む	VP	drink coffee
W	私	N	I
W	コーヒー	N	coffee
W	飲む	V	drink

中間ファイルは、単語(W)句(P)節(C)を表す記号、文型パターンに対応する記述で構成されている。句パターンを作成する際は、句を表す記号Pを手がかりにして記述を抽出する。

表3では、“コーヒー”に対応する要素をNIとして記述しており、“飲む”に対応する要素をV2として記述している。同様に、“coffee”に対応する要素をNIと記述しており、“drink”に対応する要素をV2として記述している。

表3 句パターンの例

日本語原文(一部)	コーヒーを飲む
英語原文(一部)	drink coffee
句パターン(日)	NIをV2
句パターン(英)	V2 NI

3.3 句レベル文型パターン翻訳方式の問題点

手順2の際に入力文と句レベル文型パターンを照合しているが、照合に成功する文の数が少ない。そのため、翻訳できる文の数が少ない。

4 節レベル文型パターン翻訳方式

4.1 概要

節レベル文型パターン翻訳方式では、節レベル文型パターンと入力文を照合することにより、翻訳できる文の数の向上を目指す。翻訳方法は、句レベル文型パターンと以下の2点が異なる。翻訳の例を図2に示す。

手順2(変更) 手順1の結果と節レベル文型パターンを照合(SPM[4])

手順3c(追加) 要素が節変数ならば、節パターン辞書と照合を行い、得られた複数の節パターンに存在する単語変数を単語辞書を用いて翻訳候補として訳出。

なお、節レベル文型パターンには単語変数、句変数、節変数が存在する。それぞれの変数を翻訳する際の辞書を以下に示す。

- a. 単語変数 単語辞書
- b. 句変数 句パターン辞書及び単語辞書
- c. 節変数 節パターン辞書及び単語辞書

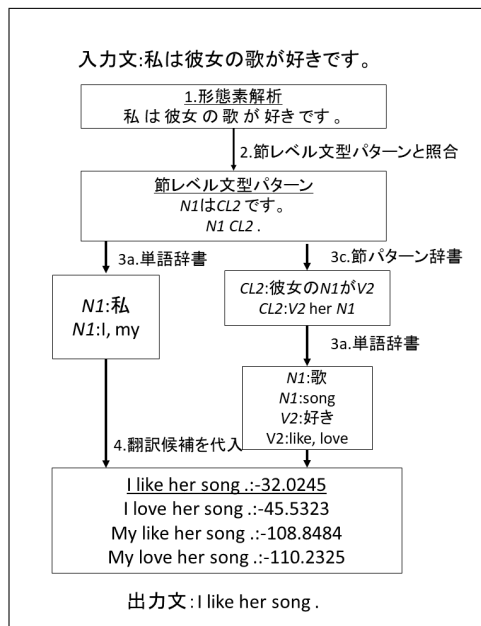


図2 節レベル文型パターン翻訳例

4.2 節パターン辞書

節パターン辞書とは、節レベル文型パターンに存在する節変数の構造を単語レベル文型パターンを用いて表現するための辞書である。節パターン辞書の作成手順を以下に示す。

手順1 文型パターン辞書の作成途中で抽出された中間ファイルより節変数に関する情報を抽出。

手順2 手順1で得られた情報を用いて単語レベル文型パターンより節変数に対応する記述を抽出。

手順3 節変数の表記や節変数に対応する単語レベル文型パターンに誤りがある節パターンを削除。

手順1 で用いる中間ファイルを表2, 作成した節パターンの例を表4に示す。

表4 節パターンの例

日本語原文(一部)	彼の行為が正しい
英語原文(一部)	his behavior is correct
節パターン(日)	#1[NIの]N2がAJ3^rentai。
節パターン(英)	#1[NI^poss]N2@beAJ3

節パターンを作成する際は、節を表す記号Cを手がかりにして記述を抽出する。

表4では、“彼”に対応する要素が“NI”となり、“行為”に対応する要素が“N2”として記述されており、“正しい”に対応する要素が“AJ3”として記述されている。また、#記号は任意要素であり、入力文に#で記述された表現が存在すれば翻訳するという意味である。また、^rentaiは該当要素が連体形であることを示す関数である。詳しくは[5]を参照。

5 評価実験

本研究では、節レベル文型パターン翻訳方式を用いることにより翻訳できる文の数(カバー率)の変化を句レベル文型パターン翻訳方式と比較調査する。また、句

レベル文型パターン翻訳方式と比較して節レベル文型パターン翻訳方式の翻訳精度の変化を自動評価 (BLEU, METEOR, RIBES, TER) 及び人手評価を用いて調査する。

5.1 実験条件

文型パターン辞書の原文対約 12 万から 10 万文対を抽出し以下に使用する。

- 翻訳モデルの学習・言語モデルの学習
- 単語辞書の作成
- 句パターン辞書の作成
- 節パターン辞書の作成

残り 2 万文対から 5,000 文対を抽出し、テスト文に用いる。節レベル文型パターン翻訳方式で照合する節レベル文型パターンは 25,475 件である。また、句レベル文型パターン翻訳方式で照合する文型パターンは 79,438 件である。また、文型パターン照合の際、得られた文型パターンの数が 10 件以上である場合は翻訳を行うのに大量の時間が必要になるため、9 件に制限して実験を行う。

5.2 カバー率

本研究で、入力文と文型パターンの照合が成功した文の数を表 5 に示す。また、翻訳できた文の数 (カバー率) を表 6 に示す。ここで、翻訳できた文の数とは、節レベル文型パターン翻訳方式では出力文に節変数に対応する記述が未知語として出力されていない文である。また、句レベル文型パターン翻訳方式では出力文に句変数に対応する記述が未知語として出力されていない文を翻訳できた文としている。表 7 に出力文に未知語が無い文の数を示す。

表 5 入力文と文型パターン照合が成功した数 (5,000 文中)

節レベル	句レベル
4,028 文 (81%)	2,666 文 (53%)

表 6 翻訳できた文の数 (5,000 文中)

節レベル	句レベル
1,659 文 (33%)	1,052 文 (21%)

表 7 未知語がない文の数 (5,000 文中)

節レベル	句レベル
1,353 文 (27%)	995 文 (20%)

表 6 より、句レベル文型パターン翻訳方式と比較すると、節レベル文型パターン翻訳方式はカバー率が向上したことが分かる。しかし、まだカバー率は低いいため更なるカバー率の向上が望まれる。

5.3 翻訳精度

提案手法の翻訳精度を自動評価 (BLEU, METEOR, RIBES, TER) 及び人手評価を用いて評価する。テスト文は表 6 を用いる。また、人手評価では、両手法で未知語がない文 (683 文) よりランダムに抽出した 100 文を用いる。

5.3.1 自動評価

自動評価に用いる文として、表 6 を用いる。自動評価の結果を表 8 に示す。

表 8 自動評価の結果

	BLEU	METEOR	RIBES	TER
節レベル	0.0798	0.2992	0.3168	1.0473
句レベル	0.1556	0.4126	0.4685	0.9205

自動評価の結果、句レベル文型パターン翻訳方式の方が節レベル文型パターン翻訳方式より翻訳精度がよいことが分かる。

5.3.2 人手評価

人手評価に用いる文として、両手法で未知語が存在しなかった文 (683 文) からランダムに抽出した 100 文を用いる*1。表 9 に結果を示す。

表 9 人手評価結果 (100 文)

節レベル○	句レベル○	差なし
30 文	58 文	12 文

節レベル○は節レベル文型パターン翻訳方式の方が句レベル文型パターン翻訳方式よりも翻訳精度が良かった文の数である。表 9 より、句レベル文型パターン翻訳方式の方が節レベル文型パターン翻訳方式よりも翻訳精度が高いことが分かる。以下に節レベル○の例及び、句レベル○の例を示す。

- 節レベル○

例を表 10 に示す。表 10 の例では、句レベルでは、“酒が回る”という表現が“a tour of the bottle”となり、入力文とは異なる意味の出力となっているため翻訳精度が低下している。

表 10 節レベル○の例

入力文	酒がまわるとしゃべり出す。
参照文	Wine loosens people's tongues .
節レベル	he fell to talking when the wine took hold .
句レベル	a tour of the bottle fell to talking .
節レベル文型パターン (日)	CL1'syushi と CL2.genzai.
節レベル文型パターン (英)	CL2 when CL1 .
節パターン CL1	(日)N3 が* 回る'。(英)N3 took hold
節パターン CL2	(日)V4:(英)he V4

- 句レベル○

例を表 11 に示す。表 11 の例では、節レベルでは、“it will rain”と入力文とは関係のない出力があるため翻訳精度が低下している。

5.4 実験結果のまとめ

カバー率は節レベル文型パターン翻訳方式の方が句レベル文型パターン翻訳方式よりもよい。翻訳精度は、自動評価、人手評価ともに句レベル文型パターン翻訳方式の方が節レベル文型パターン翻訳方式よりもよい。

*1 未知語を含む文の評価は困難であるため

表 11 句レベル○の例

入力文	10 に 5 を加えると 15 になる。
参照文	Five added to 10 makes 15 .
節レベル	when ten add five , it will rain .
句レベル	ten add five is fifteen .
節レベル文型パターン (日)	CLI ³ syushi と CL2
節レベル文型パターン (英)	When CLI , CL2 .
節パターン CL1	(日)10 に 5 を加える:(英)N1 に N2 を V3。
節パターン CL2	(日)15 になる:(英)N1 になる。

6 考察

6.1 カバー率について

表 5 より節レベル文型パターン翻訳方式の方が句レベル文型パターン翻訳方式よりも翻訳できる文が多いことが分かる。しかし、両システムとも、照合に成功した文型パターンの数(表 6)と比べると翻訳できた文の数は大幅に減少している。また、未知語無し of 文の数(表 7)も照合に成功した文型パターンの数と比べると減少している。理由として、単語辞書、句パターン辞書及び、節パターン辞書の不足が挙げられる。この問題を解決するために以下の方法が挙げられる。

- 方法 1
単語辞書、句パターン及び節パターンを人手で追加
- 方法 2
節パターン辞書の作成に句レベル文型パターンを利用

方法 1 は人手で作成するため、コストがかかる。

方法 2 は妥当性があり、より多くの節変数を表現することができ、カバー率が向上するのではないかと考える。句変数は単語変数よりも多くの表現構造を表すことができるため、単語レベル文型パターンよりも句レベル文型パターンの方がより多くの表現を表すことができる。したがって、節パターン辞書を作成する際に単語レベル文型パターンを用いるのではなく、句レベル文型パターンを用いることで多くの節変数の表現構造を表すことができるのではないかと考える。以下に句レベル文型パターンを用いた節パターンの例を示す。表 12 の日英原文の一部を単語レベル文型パターンで表現すると、“N1 が V2”となる。しかし、“美しい古城のイメージ”に対応する名詞 (N1) の単語は存在しない。同様に、“心に浮かぶ”に対応する動詞 (V2) も単語辞書に存在しない。

表 12 節パターン辞書の作成で句レベル文型パターンを用いることで良くなると思われる例

日本語原文 (一部)	美しい古城のイメージが心に浮かぶ
英語原文 (一部)	an image of a beautiful old castle come into my mind
節パターン (日)	NPI が VP2
節パターン (英)	NPI VP2

6.2 翻訳精度について

表 8, 表 13 より自動評価及び人手評価で句レベル文型パターン翻訳方式の方が節レベル文型パターン翻訳方式よりも性能が良いことが分かる。その理由の一つとして、

文型パターン選択の誤りが挙げられる。表 13 に文型パターン選択を誤った例を示す。

表 13 節レベル文型パターン翻訳方式で翻訳精度が低かった例

入力文	炉がうなりシューっと音を立てた。
出力文 (節レベル)	furnaces is groan as patter whoosh to thump .
参照文	The furnace roared and hissed .
文型パターン (日)	CL1CL2.kako.
文型パターン (英)	CL1 as CL2 .
節変数 (CL1)	(日):N1 が V2 : (英)N1 @be V2
節変数 (CL2)	(日):ADV1 N2 を V3 : (英)V3 #1[ADV1] to N2
参考情報 (句レベル)	
出力文 (句レベル)	furnaces growled and fizzled .
文型パターン (日)	NPI が VP2VP3.kako.
文型パターン (英)	NPI VP2 [^] past and VP3 [^] past . .

表 13 では、英語の節レベル文型パターン内に“as”が存在する。入力文で動作の連続を表しているが“as”の用法として動作の連続を表す表現は存在しない。そのため、表 13 の英語の節レベル文型パターンを用いると翻訳精度が低下する。このように、文型パターンの選択を誤った例が他にも多く見られた。この問題点を解決するには、適切な文型パターンを用いる必要がある。現在、適切な文型パターンを用いるためにどのような情報が必要か未調査のため、今後調査する必要がある。

7 おわりに

本研究では、節レベル文型パターン翻訳方式を試作した。句レベル文型パターン翻訳方式と比較すると、翻訳できる文の数は向上したが、翻訳精度は低下した。今後の課題として、翻訳できる文の数を増やすためにパターン辞書の数の増加が挙げられる。また、適切な文型パターンを選択するためにどのような情報が必要か調査が必要である。

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] Jun Sakata, Masato Tokuhisa, Jin' ichi Murakami. “Phrase-Level Pattern-Based Machine Translation Based on Analogical Mapping Method”. 8th International Conference on NLP(japTAL2012), Oct. 2012.
- [4] 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
- [5] 池原悟, 2009, “非線形言語モデルによる自然言語処理”, 岩波書店