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

吉田大蔵　　村上仁一　　徳久雅人
鳥取大学大学院工学研究科 情報エレクトロニクス専攻
{s062062, murakami, tokuhisa} @ ike.tottori-u.ac.jp

1　はじめに

日英機械翻訳において，パターン翻訳方式の研究が行われている．池原らは，単文のための文型パターン辞書として，「日本語語彙大系」[1] を作成した．また，重文・複文のための文型パターン辞書として，「鳥バンク」[2] を作成した．

石上らは，鳥バンクを用いた日英パターン翻訳実験を行った [3]．単語に対応する変数部分には，辞書引きをして翻訳を行ったが，句に対応する変数部分には，ルールを作成して翻訳を行う必要があった．しかし，このルールをすべて作成するのは困難である．

一方，近年では，統計翻訳 [4] が注目され，研究が行われている．統計翻訳は，原言語と目的言語のコーパスから自動的に翻訳規則を作成して，翻訳を行う手法である．従って，統計翻訳は，システムの開発が容易である．

そこで本稿では，日英パターン翻訳における変数部分の翻訳の問題を解決するために，統計翻訳を用いることを提案する．今回の実験では，鳥バンクの重文・複文の文型パターン辞書を用いて，日本語を中間言語に変換する．作られた中間言語を用いて統計翻訳を行い，翻訳精度を調査する．

2　文型パターン辞書

「鳥バンク」の「日本語表現意味辞書（重文・複文編）」[2] には，日本語の重文・複文とその対訳英文を約 12 万文対，および，その文対から作成された「意味類型パターン（22.7 万件）」が収録されている．簡略化した例を表 1 に示す．表 1 では，日英対訳文，それに対応する文型パターン対が記述されている．また，この文型パターン対には，単語レベル，句レベル，節レベルの 3 つのレベルがある．文型パターン対は日英対訳文から生成されており，それぞれのレベルに応じて，日英対訳文において対応可能な要素が変数化されている．

変数には，要素の語形を指定する「語形関数」と，付属語類を指定する「時制様相関数」が記述されている．これらの関数の例を表 2 に示す．日本語パターン側の変数には，「一般名詞意味属性」と「用言意味属性」が記述されている．これらは，日本語語彙大系の意味属性で，変数に対応できる単語の属性を制約する．

■意味属性制約

本研究は，日本語彙大系 [1] の「一般名詞意味属性体系」と「用言意味属性体系」の意味属性を扱う．これらの体系は，それぞれ一般名詞，用言の意味的用法を上位下位・全体部分関係により体系化したものであり，ツリー構造をしている．一般名詞意味属性体系は，2,715 属性と最大 12 階層，用言意味属性体系は，36 属性と最大 4 階層で構成されている．

表 1　日本語表現意味辞書における文型パターン対の例

日英対訳文:
　勉強をしている間はラジオを切っておきなさい。
　While studying, turn o　the radio.

単語レベルパターン:
　勉強をしている間は $N1$ を $V2$ 。
　While studying, $V2$ $N1$.

句レベルパターン:
　勉強をしている間は $VP1$ 。
　While studying, $VP1$.

節レベルパターン:
　勉強をしている間は $CL1$ 。
　While studying, $CL1$.

表 2　文型パターン辞書で使用される関数の例

語形関数	$\hat{}rentai$, $\hat{}meirei$, $\hat{}poss$, ⋯
時制様相関数	$.genzai$, $.dantei$, $.teiru$, ⋯

■意味属性制約の使用例

日本語パターンの変数の意味属性制約が適合文を制限する例を表 3 に示す．表 3 において，「適合○」のパターンでは，日本語文の「彼」と「歩い」の意味属性が日本語パターンの変数の意味属性制約を満たしている．「適合×」のパターンでは，パターンの変数 $N2$ の意味属性制約 (1167 義務) を日本語文の名詞「腕」の意味属性 (592 腕) が満たしていない．

表 3　意味属性制約が適合文を制限する例

日本語文	彼ら (25 他称) は 腕 (592 腕) を 組ん (23 身体動作) で 歩い (18 物理的移動) た 。
適合○	$N1$(25 他称) は 腕 を 組ん で $V2$(18 物理的移動) 。
適合×	$N1$(25 他称) は $N2$(1167 義務) を $V3$(23 身体動作) で $V4$(18 物理的移動) 。

3　中間言語

本稿では，日本語文に対して文型パターン辞書で照合を行う．適合した文型パターン対を用いて変換した「日本語と英語の単語が混在した文」を中間言語文，および，その言語を中間言語と呼ぶ．中間言語は，文型パターン対の英語パターンの骨組みをベースとして作られるので，英語の文法構造を持っている．例を用いた中間言語への変換方法は，5.1 節で説明する．

4　統計翻訳の概要

統計翻訳では，原言語 f が与えられたとき，全ての組合せの中から確率が最大となる目的言語 $\hat{e}$ を探索して翻訳を行う．以下に基本的なモデルを示す．

$$\hat{e} = \arg\max_{e} P(e|f)$$
$$\arg\max_{e} P(f|e)P(e)$$

$P(f|e)$ は翻訳モデル，$P(e)$ は言語モデルと呼ぶ．これらのモデルは，対訳コーパスから学習する．

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

5 提案手法

本稿では，日英パターン翻訳における変数部分の翻訳の問題を解決するために，統計翻訳を用いることを提案する．その手法の概要を以下に示す．

1. 鳥バンクの重文・複文の文型パターン辞書を用いて，日本語を中間言語に変換する．
2. 1 で作成した中間言語の統計翻訳を行う．

上記の提案手法によって，日英パターン翻訳における変数部分の翻訳を統計翻訳でカバーできると考える．

今回は，日英パターン翻訳において，句レベルパターンを用いて提案手法の実験を行う．

5.1 中間言語への変換方法

日本語文を中間言語文に変換する方法を以下に説明する．

手順 1 形態素解析された日本語文に対して，文型パターン辞書を用いて照合を行う．

日本語文	彼 の お母さん が ああ 若い と は 思わ なかっ た 。

日本語パターン	$NP2$ が ああ $AJ3$ とは $V4$ *.hitei .kako* 。
英語パターン	I never $V4$ $NP2$ to be so $AJ3$.
バインド値	$NP2$=’彼 の お母さん’ $AJ3$=’若い’ $V4$=’思わ’

手順 2 手順 1 で適合した日本語パターンと対になっている英語パターンに変数の値を代入する．このとき，英語パターンの語形関数は削除する．

中間言語文	I never 思わ 彼 の お母さん to be so 若い .

以上の手順によって，日本語文を中間言語文に変換する．手順 2 では，統計翻訳におけるデータスパースネスの問題を軽減するために，英語パターンの語形関数を削除している．

5.2 提案手法における翻訳手順

本実験で行う翻訳の手順を，図 1 に従って，トレーニング部とデコーディング部に分けて説明する．

■トレーニング部

手順 T1 文型パターン辞書の日英対訳文と，その文対から作られている文型パターン対同士を照合し，日本語文を中間言語文 J' に変換する．作成した中間言語文 J' と英語文の対を統計翻訳のための学習データとする．

手順 T2 手順 T1 で作成した学習データを用いて，統計翻訳のための翻訳モデルを学習し，日英対訳コーパスの英語文を用いて言語モデルを学習する．

■デコーディング部

手順 D1 入力する日本語文を文型パターン辞書と照合して中間言語文に変換する．1 つの入力文に対して複数の文型パターンが適合した場合，すべて中間言語文に変換する．

手順 D2 手順 D1 で得たすべての中間言語文に対して統計翻訳を行う．統計翻訳のデコーダーには Moses[5] を用いる．

手順 D3 手順 D2 で得られた翻訳候補文の中から，翻訳スコアがもっとも高いものを最終的な翻訳文とする．

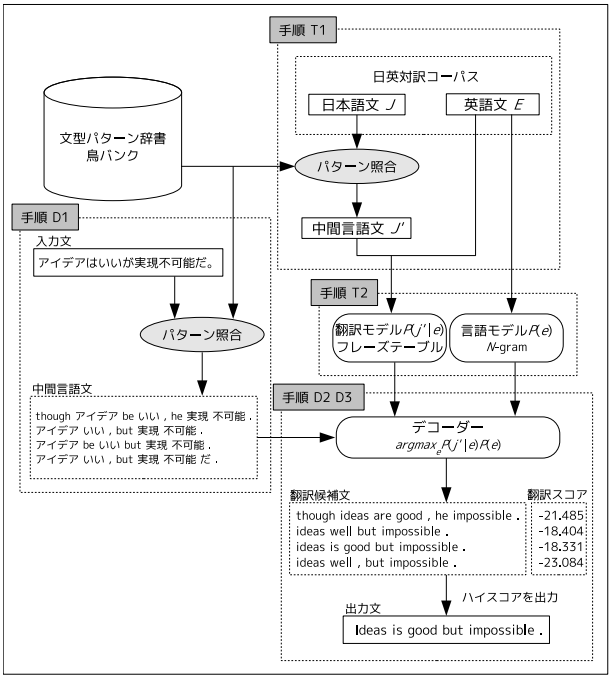

図 1 提案手法の全体の流れ

6 翻訳実験

本稿では，ベースラインと提案手法の実験を行う．テスト文として日本語文型辞典 [6] から抽出した 3,952 文を用いる．このテスト文は，鳥バンクの文型パターン辞書に対して，オープンなデータである．テスト文のうち，提案手法において中間言語に変換できたものを評価の対象とする．

6.1 ベースライン

ベースラインでは，句に基づく日英統計翻訳を行う．デコーダーには，Moses を用いる．日英対訳コーパスには，文型パターン辞書から抽出した日英対訳文 121,913 文対を用いる．

6.2 提案手法

提案手法では，日英対訳コーパスの日本語文を句レベルパターンで中間言語文に変換する．変換した中間言語と英語の対訳文 74,564 文を学習データとして用いる．言語モデルには，文型パターン辞書から抽出した英語文 121,913 文を用いる．また，入力文と文型パターン辞書を照合する際，意味属性制約を使用する．

7 実験環境

■形態素解析器 提案手法における形態素解析には，[7] で作成された形態素解析器を用いる．ベースラインにおける形態素解析には，mecab[8] を用いる．

■パターン照合器 本実験では，[9] で作成されたパターン照合器 SPM を用いる．

■翻訳モデルの学習 フレーズテーブルの作成には，Moses 付属の train-model.perl を用いる．

■言語モデルの学習 言語モデルには，N-gram モデルを用いる．N-gram モデルの学習には，SRILM [10] を用いる．本稿では，5-gram を用い，スムージングには kndiscount を用いる．

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

■デコーダのパラメータ　本実験では，パラメータチューニングを行わず．デフォルトのパラメータを用いる．ただし，翻訳時のフレーズ位置の変化に対応するために，distortion-limit を-1 とする．

8　評価方法

本稿では，ベースラインと提案手法の翻訳出力の対比較評価を行う．判断基準を以下に示す．

提案手法○　提案手法の出力がベースラインの出力より優れている場合

提案手法×　提案手法の出力がベースラインの出力より劣っている場合

差なし　提案手法の出力とベースラインの出力の表現に差がない場合

同一出力　提案手法の出力とベースラインの出力が同一の場合

9　実験結果

入力文と文型パターンを照合した結果，767 文に対して中間言語文が得られた．これらの中間言語文に対して統計翻訳を行った．出力された翻訳文の中からランダムに 100 文を選び，ベースラインと対比較評価をした結果を表 4 に示す．また，提案手法○の例を表 5，提案手法×の例を表 6，差なしの例を表 7 に示す．

表 4　対比較評価結果

提案手法○	提案手法×	差なし	同一出力
10	6	84	0

提案手法○が提案手法×より，4 件多い結果となった．

■提案手法○の例

表 5　提案手法○の例

入力文	ゆうべ飲み過ぎて頭が痛い。
正解文	I have a headache from the hongover .
ベースライン	Last night i drinking too much of a headache .
提案手法	I drank too much last night and my head aches .
中間言語文	i ゆうべ 飲み 過ぎ and my 頭 が 痛い .
適合パターン	$VP2$ て $N4$ が $AJ5$ 。
英語パターン	i $VP2$ and my $N4$ $AJ5$.

ベースラインに対して，提案手法では，「〜て頭が痛い」の部分をうまく翻訳できている．従って，提案手法○と判断した．

■提案手法×の例

表 6　提案手法×の例

入力文	春が来ると花が咲く。
正解文	hen spring comes,　owers come out.
ベースライン	When spring comes ,　owers come out .
提案手法	If the spring comes , you will of owers bloom .
中間言語文	if 春 来る , you will 花 が 咲く .
適合パターン	$N1$ が $VP2$ と $VP4$ 。
英語パターン	if $N1$ $VP2$ you $VP4$.

ベースラインに対して，提案手法では入力文の意味を捉えて翻訳できていない．従って，提案手法×と判断した．

■差なしの例

表 7　差なしの例

入力文	彼は、就職をきっかけにして、生活をかえた。
正解文	He changed his lifestyle with his employment as a turning point.
ベースライン	He was a and life away .
提案手法	He started to and lives away .
中間言語文	彼 就職 を きっかけ に し and 生活 を かえ .
適合パターン	$NP1$ は 、$VP2$ て 、 $VP3$ $.kako$ 。
英語パターン	$NP1$ $VP2$ and $VP3$.

ベースライン，提案手法，ともに入力文の意味を捉えていないと判断し，差なしとした．

10　考察

表 4 の結果から，提案手法に有効性が見られなかった．表 5 では，提案手法において，有効な文型パターン対が適合して中間言語文が作成され，良い翻訳文が出力されている．しかし，表 6，表 7 では，入力文に対し不適切な文型パターンが適合し，翻訳に悪い影響を与えている．提案手法の有効性が得られなかった原因を，10.1 節，10.2 節で考察する．

また，提案手法における，句レベルパターンのカバー率は，20 %しか得られなかった．原因として，文型パターン辞書が対応できないような表現が入力文に含まれていた，もしくは，入力文のすべての要素と対応できる文型パターンがなかったことが考えられる．この問題を解決するためには，文型パターン辞書の改良を行う，または，入力文と文型パターンとの部分的な適合を行う必要がある．

10.1　翻訳候補文の選択に用いるスコアの問題

提案手法の効果が得られない場合として，翻訳候補文の中に，適切な翻訳文があるが，統計翻訳が出力したスコアによって，最終的に不適切な翻訳文が出力される場合がある．その例を表 8 に示す．

表 8　提案手法の効果が得られなかった例

入力文	彼は立ち上がってあたりを見回した。
正解文	He rose to his feet and looked all around him.
提案手法の出力	He stood up and around.

翻訳候補文	スコア
he stood up and around .	-8.227
he stood up and looked around .	-9.162
he stood up before around .	-11.213
he looks around stood up .	-11.480
you stand up , and he around .	-13.873
he him some him and around .	-16.309
…	

表 8 では，翻訳スコアが最も高い「he stood up and around .」が出力されているが，「he stood up and looked around .」の方が翻訳出力として適切であると考える．この原因として，翻訳に適切でない文型パターン対が入力文に対して適合していることが挙げられる．また，文型パターン対が入力文に対して適合し過ぎていることが挙げられる．表 8 の例では，入力文に対して，249 パターンが適合し，中間言語文が作成されている．このような場合は，不適切な翻訳文が出力される可能性が高くなると考える．これらの問題を解決するために，文型パターン対の適合を制限する，または，別のスコアリング方法を考える必要がある．

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

10.2 学習データ量の問題

翻訳文の中に，表 9 のような，未知語を含む文が多くあった．原因として，学習データ量が不足していることが考えられる．ベースラインの学習データは，121,913 文対であるのに対し，句レベルのパターンで作成できた中間言語文と英語文の学習データは，74,564 文対である．従って，提案手法において，中間言語の日本語部分を翻訳するための情報が不足してしまったと考える．今後，提案手法における学習データを増やす方法を検討する必要がある．

表 9 未知語を含む翻訳文

翻訳文
She has been completely rose at her boy friend in プロポーズ .
Walk , you will take すくなくとも 20 mimutes .
When i spaghetti a ゆであがっ , i quickly to the sauce からめ .
I never busy ので and wait a little longer .
I went to the beach but so in attendance went ぐったり tired .
Some practice will not good くなら .

11 追加実験（単語レベルパターン）

句レベルパターンを用いた提案手法の効果があまり得られなかった．この問題に対して，単語レベルパターンでも実験を行い，調査する．学習データには，単語レベルパターンで作成された，中間言語と英語の対訳文 120,011 文対を用いる．また，入力文と文型パターン辞書を照合する際，意味属性制約を用いる．

11.1 追加実験の結果

入力文と文型パターンを照合した結果，66 文に対して中間言語文が得られた．これらの中間言語文に対して統計翻訳を行い，ベースラインと対比較評価をした結果を表 10 に示す．また，提案手法○の例を表 11，提案手法×の例を表 12 に示す．

表 10 対比較評価結果

提案手法○	提案手法×	差なし	同一出力
20	1	43	2

提案手法○が，提案手法×よりかなり多い結果となった．

■提案手法○の例

表 11 提案手法○の例

入力文	戦わずして負ける。
正解文	You could lose the war before ghting it.
ベースライン	Without ghting and beaten .
提案手法	He loses without a ght .
中間言語文	he 負ける without 戦わ .
適合パターン	$V2$ *.hitei* して $V3$ *.genzai* 。
英語パターン	he $V3$ without $V2$.

ベースラインに対して，提案手法の方が英語の文法構造を捉えていると判断し，提案手法○とした．

■提案手法×の例

表 12 提案手法×の例

入力文	事態は悪くなる一方だ。
正解文	We see the circumstances growing only worse.
ベースライン	Things are getting worse and worse .
提案手法	The situation is a bad nervous one .
中間言語文	事態 be a 悪 くなる 一方 .
適合パターン	$N1$ は $V2$ *?rentai* $N3$ *.da* 。
英語パターン	$N1$ be a $V2$ $N3$.

提案手法より，ベースラインの方が入力文の「悪くなる一方だ」の意味を捉えていると判断し，提案手法×とした．

11.2 追加実験の考察

追加実験をした結果，提案手法の効果を得ることができた．理由として，単語レベルパターンを使って入力文を中間言語文に変換した場合，句レベルパターンに比べて，より英語の文法構造に近付いたことが考えられる．しかし，66 文しか中間言語文を作成できていないことから，入力文に対する単文レベルパターンのカバー率が非常に低い結果となった．

12 おわりに

本稿では，日英パターン翻訳における変数部分の翻訳の問題を解決するために，統計翻訳を用いることを提案した．その手法として，鳥バンクの重文・複文の文型パターン辞書を用いて，日本語を中間言語に変換し，作られた中間言語を用いて統計翻訳を行った．句レベルパターンを用いた提案手法の翻訳精度を調査した結果，提案手法の有効性を得ることができなかった．原因として，1 つの入力文に適合した文型パターン対の数が多すぎたため，翻訳候補文の適切な選択ができていなかったことが挙げられる．単語レベルパターンを用いた提案手法の翻訳精度を調査した結果，ベースラインに対して，翻訳精度が向上したが，カバー率が非常に低い結果となった．今後は，これらの問題を解決するために，適切な文型パターン対を適合させながら，カバー率を上昇させる方法を検討する．また，翻訳候補文の適切な選択を行う方法を検討する．

参考文献

[1] 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.

[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] 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.

[4] Richard Zens, Franz Josef Och, Hermann Ney, “Phrase-based Statistical Machine Translation”, KI 2002, pp.35-56, 2002.

[5] Philipp Koehn, Marcello Federico, Brooke Cowan, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, Evan Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation”, ACL 2007, pp.177-180, 2007.

[6] グループ・ジャマシイ, “日本語文型辞典”, くろしお出版, 1998.

[7] 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.

[8] MeCab, “MeCab:Yet Another Part-of-Speech and Morphological Analyzer”, http://mecab.sourceforge.net/

[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, 自然言語処理, 2006-NL-176, pp.9-16, 2006.

[10] SRILM, The SRI Language Modeling Toolkit, http://www-speech.sri.com/projects/srilm/

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved