概要

鳥バンクの文型パターン辞書を用いた翻訳において，句変数部分の翻訳ルールをすべて作成するのは困難であるという問題がある．一方，現在，主流な翻訳方式である統計翻訳は，原言語と目的言語の対訳コーパスから自動的に翻訳のモデルを作成することが可能である．したがって，システムの開発が容易である．本研究ではこの点に着目し，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う．本研究では，この提案手法の翻訳精度の調査を目的とする．

本研究では，鳥バンクの文型パターン辞書の句レベルパターンを用いて，中間言語と英語の対訳文 74,564 文対を作成した．そして，その対訳文を学習文として，統計翻訳を行う．

提案手法において，入力文 3,952 文に対して，中間言語に変換できたものは，767 文であった．この 767 文において統計翻訳を行った．その結果，ベースラインとしての日英統計翻訳と比べ，提案手法の方が翻訳精度がよかった．しかし，英語として正しく翻訳できたものは少なかった．

目 次

図 目 次

表 目 次

第1章　はじめに

従来の代表的な機械翻訳として，トランスファー方式による翻訳が行われている．トランスファー方式による翻訳は，構文構造を変換する構文処理と，要素を翻訳する意味処理を行う．そして，それらの結果を合成することによって翻訳を行う．このような従来の自然言語処理方法を要素合成法と呼ぶ．要素合成法は，言語表現の意味は線形であると仮定し，表現全体の意味を部分的な意味の合成で説明する．しかし，この方法では，元の意味が再現されないという問題があった．この問題に対して，表現全体の意味をすくい取るための仕組が必要であるとし，構造と意味を一体化したパターン翻訳の研究が行われた．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．したがって，統計翻訳は，システムの開発が容易である．

そこで本研究では，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．

第2章　先行研究

2.1　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．日英対訳文と文型パターン対の例を表 2.1 に示す．表 2.1 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．

文型パターンは様々な記述要素で構成されている．その詳細は，2.1.1.3 節で説明する．また，日本語パターン側の変数には，意味属性制約が記述されている．この制約については，2.1.2 節で説明する．

2.1.1　文型パターン記述要素の構成

文型パターンは,「字面」,「変数」,「関数」,「記号」の 4 種類の記述要素で構成されている．この節では，これらの記述要素について説明する．また，例を示す際,「変数」以外の説明に不要な記述要素は省略する．

表 2.1: 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
日本語文	勉強をしている間はラジオを切っておきなさい。
英語文	While studying, turn off the radio.
	単語レベルパターン
日本語パターン	/ytcfk 勉強を/cf している/f 間は/tcfk$N1$(OR:ラジオ, NI:970,$\cdots$) を/cf(V2(OR:切っ,NY:0506,NY:0701,NY:2001,$\cdots$).$teoku\hat{}meirei|V2$(NY:0506,NY:0701,NY:2001,$\cdots$).$teoku.meireigo$)。
英語パターン	While studying, $V2$(OR:turn off) $\hat{}baseN1$(OR:the radio).
	句レベルパターン
日本語パターン	/ytcfk 勉強を/cf している/f 間は/tcfk($VP1$(OR:ラジオを切っ, NY:0506,NY:0701,NY:2001,NY:2002,$\cdots$).$teoku\hat{}meirei|VP1$(NY:0506,NY:0701,NY:2001,$\cdots$).$teoku.meireigo$)。
英語パターン	While studying, $VP1$(OR:turn off the radio) $\hat{}base$.
	節レベルパターン
日本語パターン	/ytcfk 勉強を/cf している/f 間は/tcfk($CL1$(OR:ラジオを切っ,NY:0506,NY:0701,NY:2001,$\cdots$).$teoku\ hatmeirei|CL1$(NY:0506,NY:0701,NY:2001,$\cdots$).$teoku.meireigo$)。
英語パターン	While studying, $CL1$(OR:turn off the radio).

2.1.1.1 字面による記述

文型パターンにおける，字面の記述は，字面によって適合する単語を指定する．字面のみよって適合を指定する「出現形字面」と，字面と関数によって任意の様相時制が対応できる「終止形字面」と「原形字面」がある.「終止形字面」は日本語パターン,「原形字面」は英語パターンで用いられる．表 2.2 に字面の記述がある文型パターンの例を示す.

表 2.2: 字面による記述の例

日本語パターン	$N1$ は $N2$ と $N3$ に $V4$ $N5$ にしたいと’思う’♯7[.$masu$]。

表 2.2 において，単語の字面で表記されている「は，と，に，にしたいと」の部分は「出現形字面」である．単語の字面に関数と記号が付与されている「’思う’♯7[.$masu$]」の部分の「思う」は,「終止形字面」で，後に続く関数と記号によって,「思う」，または,「思います」に適合できる.

2.1.1.2 変数による記述

文型パターンにおける，変数の記述は，線形な自立語的要素（単語，句，節）を指定する．表 2.3 に，変数の種類を示す．

表 2.3: 変数の種類

分類	変数種別
単語変数	Nn:名詞一般＋複合名詞，$TIME$n:時間の名詞，NUMn:数詞，NDn:用言性名詞，RENn:連体詞，GENn:限定詞，Vn:動詞，AJn:形容詞，AJVn:形容動詞，ADVn:副詞，ANY:任意
句変数	NPn:名詞句，VPn:動詞句，AJPn:形容詞句，$AJVP$n:形容動詞句，$ADVP$n:副詞句
節変数	CLn:節
＜注＞すべての変数には，パターン内の通し番号 n が付与されている．	

2.1.1.3 関数による記述

文型パターンにおける，関数の記述には，要素の語形指定をする「語形関数」（ハット関数）と，付属語類の指定をする「時制様相関数」（ドット関数）がある．表 2.4 にそれらの関数の例を示す．また，表 2.5 にそれらの関数の記述がある文型パターンを示す．

表 2.4: 文型パターンにおける関数の例

語形関数	付与された用言の語形を指定する
例	$\hat{}rentai$:連体形, $\hat{}meirei$:命令形, $\hat{}poss$:所有格変形, …
時制様相関数	指定された付属語の存在を意味する
例	$.genzai$:現在, $.dantei$:断定, $.teiru$:ている, …

表 2.5: 関数の記述がある文型パターンの例

日本語パターン	$N1$ は $AJ2\hat{}rentai$ $N3.da$ というのに $N4$ はあいかわらず酒を飲んでいる。

表 2.5 において,「$AJ2\hat{}rentai$」には,「重い」のような連体形の形容詞が適合する．また,「$N3.da$」には,「病気だ」のような,「名詞＋だ（断定）」の表現が対応する．

2.1.1.4　記号による記述

文型パターンにおいて用いられる記号の意味を説明する．

要素選択のための記号

文型パターンにおける，要素選択型の記号は，表記と表現の揺らぎを指定するものである．日英パターンにおいて，互いに影響しない「要素選択記号」と，互いに対応関係のある言い換え表現を指定する「対応型要素選択記号」がある．表 2.6 にその記号の書式を示す．また，表 2.7 に「対応型要素選択記号」がある文型パターン対の例を示す．

表 2.6: 要素選択型の表現

要素選択記号	(パターン記述 1| パターン記述 2|…)
対応型要素選択記号	♯数 (パターン記述 1| パターン記述 2|…)

表 2.7: 対応型要素選択記号がある文型パターンの例

日本語パターン	$N1$ では $N2$ は $VP3$ $\hat{}rentai$ ことを $V4\sharp5(.genzai\|.kako)$。
英語パターン	In $N1$ $N2$ $V4\sharp5(\hat{}present\|\hat{}past)$ to $VP3$.

表 2.7 において，日本語パターンの「$V4\sharp5(.genzai|.kako)$」の部分は,「$V4$」の時制が,「現在」，または,「過去」のどちらかに対応できる．また，それらの時制に応じて，英語パターン側で「$\hat{}present$」，または,「$\hat{}past$」が使用される．

要素補完のための記号

文型パターンにおける，要素補完のための記号は，省略可能な主語・目的語などを指定するものである．英語側にあるが日本語側にはなくてもよい要素を指定する「補完要素記号」と，日本語側の要素の有無に応じて表現を指定する「訳出要素選択記号」がある．表 2.8 にそれらの記号の書式を示す．また，表 2.9 に要素補完のための記号がある文型パターン対の例を示す．

表 2.9 において，日本語パターンの「$< N1$ は $>$」に適合する日本語要素があった場合は「$N1$」を，なければ「I」を英語側で用いる．

表 2.8: 要素補完のための表現

補完要素記号	$<$ パターン記述 1 $>$
訳出要素選択記号	$<$ パターン記述 1$|$ パターン記述 2 $>$

表 2.9: 要素補完のための記号がある文型パターン対

日本語パターン	$<N1$ は $>NP2$ がああ $AJ3$ とは $V4$。
英語パターン	$<I\|N1>$ never $V4$ $NP2$ to be so $AJ3$.

任意要素を指定する記号

文型パターンにおいて，任意要素を指定するものとして,「離散記号」と「任意要素記号」がある.「離散記号」は，連体修飾句などの任意要素とその位置を指定している.「任意要素記号」は，日英パターンで対応関係を持つ要素で，双方から削除できる要素を指定している．表 2.10 に「離散記号」の種類と,「任意要素記号」の書式を示す．また，表 2.11 にそれらの記号がある文型パターン対の例を示す.

表 2.10: 任意要素を指定する記号

離散記号	/y:連用節，/t:連体節，/c:格要素，/f:副詞，/k:形容（動）詞連体形・連体詞
任意要素記号	♯数 [$\cdots$]

表 2.11: 任意要素を指定する記号がある文型パターン対

日本語パターン	$N1$ から/tcfk$N2$ へ/cf 行く/f 間に ♯3[/cf$ADV4$]/f$AJV5$/f$N6$/cf ありま す。
英語パターン	There is ♯3[$ADV4$] $AJ5$ $N6$ between $N2$ and $N1$.

表 2.11 において，日本語パターンの「♯3[/cf$ADV4$]」の部分に日本語要素が適合した場合は，英語パターン側で「$ADV4$」を用いる．また，日本語パターンにおいて，離散記号がある位置に，それらの記号が指定する要素が適合してもよい.

語順と位置変更可能要素を指定する記号

文型パターンにおいて，語順と位置の変更が可能な要素を指定する記号がある.「順序任意要素指定記号」は，格要素などの要素をグループ化し，順序を任意化している.「位置変更可能記号」は，副詞などの変更可能な位置を指定している．表 2.12 にそれらの記号の書式を示す．また，表 2.13 に「任意変更可能記号」がある文型パターンを示す.

表 2.12: 語順と位置変更可能要素を指定する記号

順序任意要素指定記号	♯1{ パターン記述 1, パターン記述 2, ⋯}
任意要素記号	$1 ^{ パターン記述 }.....$1

表 2.13: 位置変更可能要素がある日本語パターン

日本語パターン	1N1$ と$1 ^$N2$ は会っても $VP3$ 間柄ではありません。

表 2.13 において,「$1」で示されている「$N2$ は」は，ほかの「$1」の場所に移動してもよい.

2.1.2 意味属性制約

鳥バンクの文型パターン辞書では，日本語パターンの名詞変数と用言変数部分に，意味属性制約が付与されている．鳥バンクの文型パターン辞書で用いられる意味属性体系の種類とその種類を表す記号を表 2.14 に示す.

表 2.14: 意味属性の種類

辞書タイプ記号	意味属性体系の種類
NI	日本語語彙大系一般名詞意味属性体系
NK	日本語語彙大系固有名詞意味属性体系
NY	日本語語彙大系用言意味属性体系
IY	用言意味分類
IM	名詞意味分類体系

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.

■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 2.15 に示す．表 2.15 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 2.15: 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) 。

2.2 パターン照合

本研究では，日本語文と日本語パターンの照合を行う．日本語文に対して，形態素解析を行い，その後，パターン照合を行う．本研究で用いる形態素解析器，パターン照合器は，それぞれ，2.2.1 節，2.2.2 節で説明する．

2.2.1 形態素解析器

本研究では，日本語パターンと照合する日本語文の形態素解析に，[5] で作成された形態素解析器を用いる．その形態素解析器の入出力を表 2.16 に示す．

表 2.16: 形態素解析の例

入力	勉強をしている間はラジオを切っておきなさい。
出力	1. /勉強 (1230,NI:1388,NI:1379,NI:1888,KR:1806a01,…) 2. +を (7430) 3. /し (2433, する, し,NY:16,NY:21,NY:20,NY:32,NY:5,NY:31,…) 4. +ている (2817) 5. /間 (1500,NI:2690,NI:2691,…)(1800)(1100,NI:2660,NI:2444,…) 6. +は (7530) 7. /ラジオ (1100,NI:970,KR:0410k35,KR:2105k37,IM:13680) 8. +を (7430) 9. /切っ(2384, 切る, 切っ,NY:23,NY:7,NY:36,NY:20,NY:32,NY:22,…) 10. +ておき (2713, ておく, ておき) 11. +なさい (2789) 12. +。(0110) 13. /nil

表 2.16 の出力では，入力文は形態素に分割されている．形態素ごとに，要素番号，要素，品詞番号，意味属性が出力されている．

2.2.2 パターン照合器

本研究では，[6] で作成されたパターン照合器パターン照合器は，日本語文の形態素解析結果と日本語パターンの照合を行い，適合したパターンの抽出および各変数に対応するバインド値などを出力する．パターン照合器の出力の例を表 2.17 に示す.

表 2.17: 表 2.16 に対する SPM の出力における適合パターンの一例

適合 パターン 1	PATTERN=PJAC000090-00=[勉強を, している, 間は,$VP1$, $.teoku$, $.meireigo$,。]=[1,2,3,4,5,6,7,8,9,10,11,12]=12 $VP1$=[7,8,9]=9=3
適合 パターン 2	PATTERN=PJAC000090-00=[勉強を, している, 間は,$VP1$, $.teoku$, $.meireigo$。]=[1,2,3,4,5,6,9,10,11,12]=10 $VP1$=[9]=9=1

表 3.2 では，日本語文に対して，2 つの句レベルパターンが適合している.「適合パターン 1」において,「PATTERN=」に続く部分は，順番に，パターン ID, パターンの適合要素，適合した入力文の要素番号，適合した入力文の要素数の合計である.. また,「$VP1$=」に続く部分は，変数に対応した入力文の要素番号，変数に対応した主体となる入力文の要素番号，適合した入力文の要素数の合計である．この場合では，$VP1$ に’ ラジオ を 切っ’ がバインドされている.

2.3 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．

$$\begin{aligned} \hat{e} &= \arg\max_{e} P(e|f) \\ &\approx \arg\max_{e} P(f|e)P(e) \end{aligned} \tag{2.1}$$

$P(f|e)$ 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．また，日英統計翻訳の流れを図 2.1 に示す．

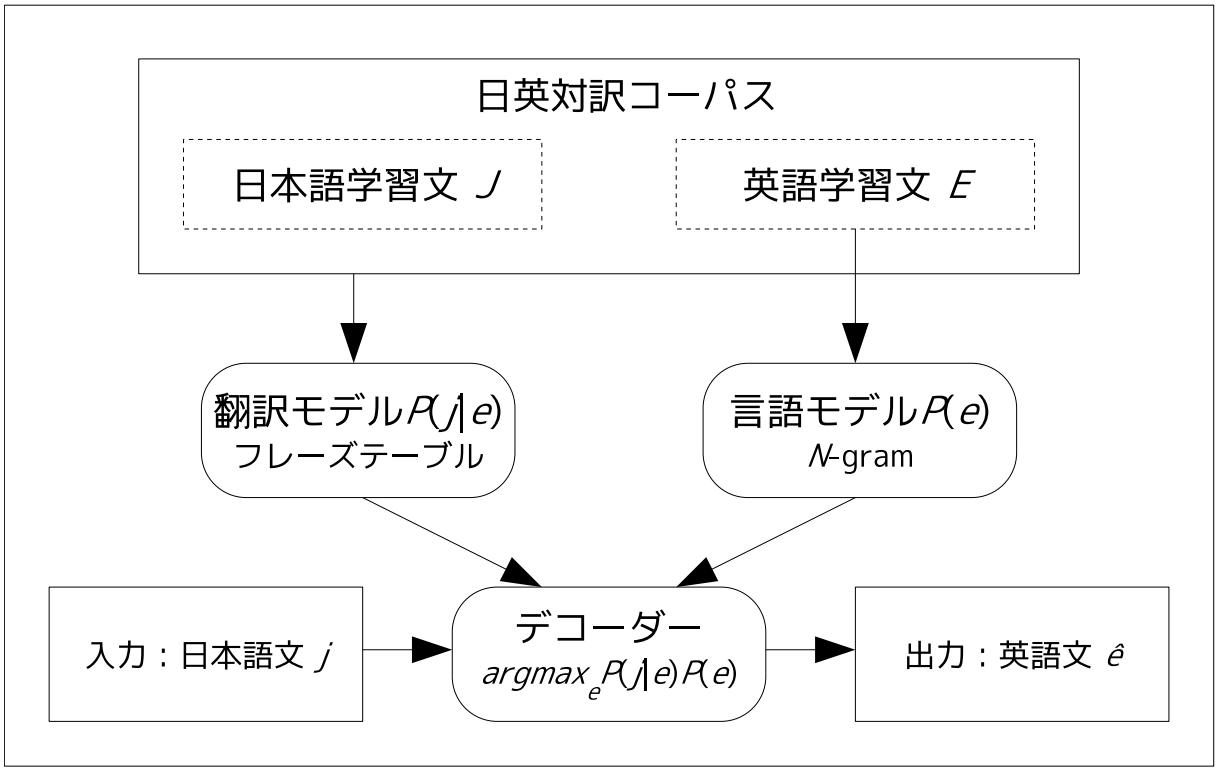

図 2.1: 日英統計翻訳の流れ

2.3.1 翻訳モデル

翻訳モデルは日本語の単語列から英語の単語列へ確率的に翻訳を行うためのモデルである．翻訳モデルはフレーズテーブルで管理されている．フレーズテーブルは，日英対訳コーパスの日英対訳文において，GIZA++[7] によって取られた単語アライメントから自動的に作成する．図 2.2 にフレーズテーブルの例を示す．

あなた に 会う ||| seeing you ||| 0.0833 0.0004 0.5 0.0187
あなた の 意見 ||| your opinion ||| 0.053 0.002 0.5 0.031

図 2.2: フレーズテーブルの例

左から日本語フレーズ，英語フレーズ，フレーズの英日翻訳確率，英日方向の単語の翻訳確率の積，フレーズの日英翻訳確率，日英方向の単語の翻訳確率の積である．

2.3.2 言語モデル

2.3.2.1 言語モデルの概要

言語モデルは単語列の生じる確率を与えるモデルである．日英機械翻訳では翻訳候補から英語として尤もらしい文を選出する．統計翻訳では，一般に，N-gram モデルを用いる．図 2.3 に言語モデルの例を示す．

-2.321799	about ten	-0.0283015
-1.625605	about that	-0.2182165
-0.8188105	about the	-0.1321755
-2.026232	about their	-0.01974426
-2.730913	about them	-0.4373209
-1.601604	about this	-0.1513322
-2.378062	about three	-0.03816521
-2.627227	about time	-0.1131216
-1.923688	about to	-0.0686989

図 2.3: 2-gram の例

図 2.3 において，1 列目が単語の連鎖確率を常用対数で表したもの，2 列目が連鎖する単語，3 列目がバックオフスムージングで推定した単語の連鎖確率である．バックオフスムージングについては，2.3.2.2 節で述べる．

2.3.2.2 N-gram モデル

代表的な言語モデルとして，N-gram モデルがある，N-gram モデルは，1 次元の単語列 $w_1, w_2, ..., w_n = {w_1}^n$ における i 番目の単語 w_i の生起確率が，直前の単語列 $w_{i-(N-1)}, w_{i-(N-2)}, ..., w_{i-1}$ に依存するという仮説に基づくモデルである．これは，以下の式にで表せる．

$$P({W_1}^n) \quad = \quad P(w_1)P(w_2|w_1)...P(w_n|w_1^{n-1})\cdots \tag{2.2}$$

$$\approx \quad P(w_1)P(w_2|w_1)...P(w_n|w_{n-(N-1)}^{n-1}) \tag{2.3}$$

$$= \quad \prod_{i=1}^{n} P(w_i|w_{i-(N-1)}^{i-1}) \tag{2.4}$$

また，$P(w_i|w_{i-(N-1)}^{i-1})$ は以下の式で計算される．$C()$ は単語列の出現数である．

$$P(w_n|w_{i-(N-1)}^{i-1}) \quad = \quad \frac{C(w_{i-(N-1)}^{i})}{C(w_{i-(N-1)}^{i-1})} \tag{2.5}$$

例えば,「He is a teacher .」という単語列に対して 2-gram の言語モデルを適用した場合，単語列が生成される確率は以下の式で計算される．

$$P(\text{“}He\ is\ a\ teacher\ .\text{”}) \quad = \quad P(He)P(is|He)P(a|is)P(teacher|a)P(.|teacher) \tag{2.6}$$

しかし，2.4 から信頼できる値を算出するためには，各単語列の出現率が高い必要がある．しかし，実際には，多くの単語列の出現率が 0 となることが多いため，信頼できる値を算出できない場合が多い．したがって，確率値を平滑化する手法であるスムージングを行う．代表的な手法にバックスオフスムージングがある，バックオフスムージングでは学習データに出現しない N-gram を (N-1)-gram の値から推定する．例として，3-gram の場合の確率は以下の式で推定される．

$$P(w_i|w_{i-2}^{i-1}) = \begin{cases} \alpha \times p(w_i|w_{i-2}^{i-1}) & 3-gram \text{が存在する} \\ \beta \times p(w_n|w_{n-1}) & 3-gram \text{がなく} 2-gram \text{が存在する} \\ p(w_n|w_{n-1}) & \text{それ以外} \end{cases} \tag{2.7}$$

ここで，α をディスカウント係数，β をバックオフ係数と呼ぶ．

第3章　提案手法

本研究では，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.

3.1　中間言語

本研究では，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，英語の文法構造に近い構造を持っている．今回，提案手法で用いる統計翻訳は，言語間の文法構造が近い場合，翻訳精度が高いという特徴がある．したがって，提案手法における中間言語は，英語への統計翻訳に有効であると考える．例を用いた中間言語への変換方法は，3.2 節で説明する.

3.2 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．ここでは，文型パターン対における表記を簡略化して，日本語から中間言語への変換を説明する．

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 。

適合文型パターン対	
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バインドした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．

3.3 提案手法における翻訳手順

本研究で行う翻訳の手順を，図 3.1 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．このとき，名詞変数と用言変数の意味属性制約を使用する．また，入力文のすべての要素が文型パターンに適合した場合のみ，中間言語文へ変換する．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．以下に入力文を中間言語文に変換した例を示す．

入力文	アイデアはいいが実現不可能だ。
中間言語文	though アイデア be いい , he 現実 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 . アイデア いい , but 実現 不可能 だ .

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[8] を用いる．デコーダーは，中間言語文を翻訳すると同時に，出力の選択に用いた尤度のスコアを出力する．ここで得られた中間言語文の翻訳結果を翻訳候補文，尤度のスコアを翻訳スコアとする．以下に手順 D1 の中間言語文を翻訳して得られた，翻訳候補文と翻訳スコアを示す．

翻訳候補文	翻訳スコア
though ideas are good , he impossible .	-21.485
ideas well but impossible .	-18.404
ideas is good but impossible .	-18.331
ideas well , but impossible .	-23.084

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．以下に手順 D2 における翻訳出力の例を示す．

出力	Ideas is good but impossible .

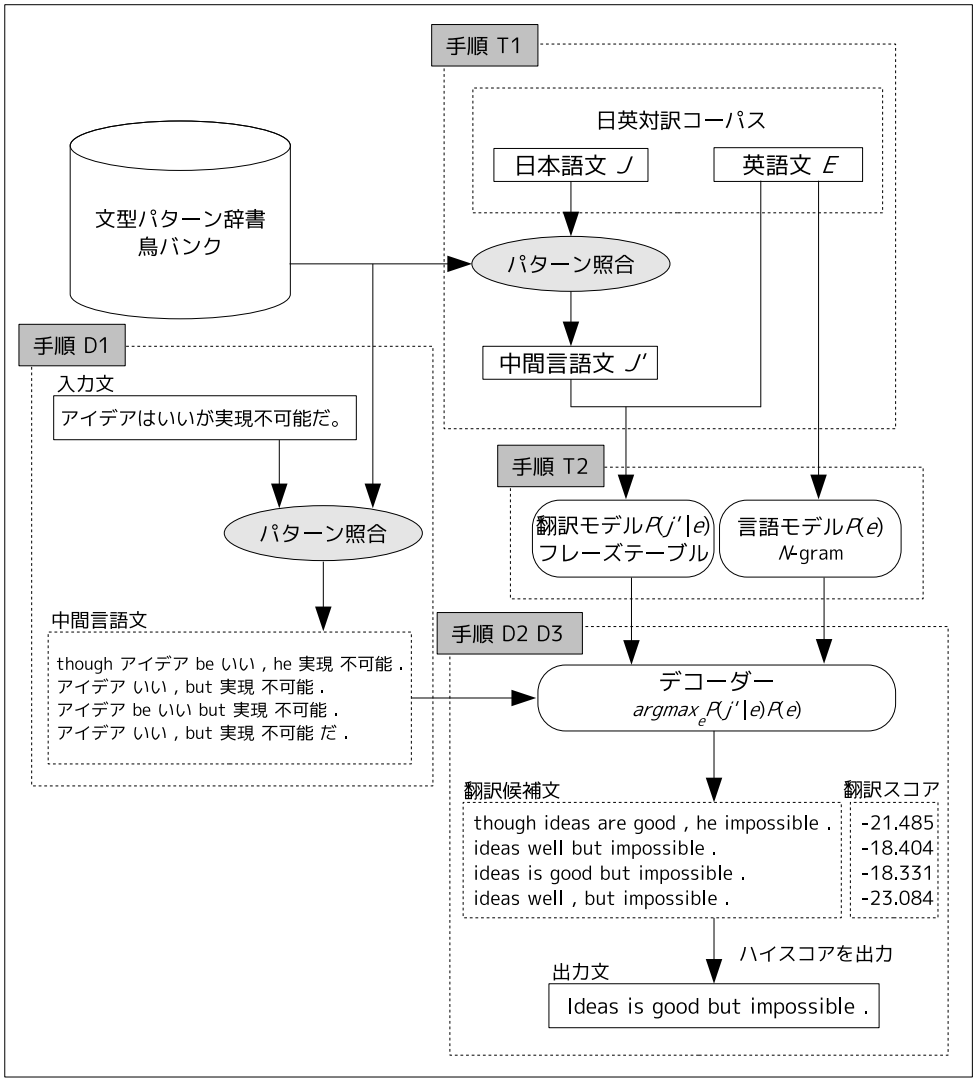

図 3.1: 提案手法の全体の流れ

第4章　翻訳実験

本研究では，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 [9] 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．

4.1　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．

4.2　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 文対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．

4.3 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．

■評価基準

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

4.4 実験環境

形態素解析器 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[10] 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 [11] 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．

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．

4.5 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 4.1 に示す．また，提案手法○の例を表 4.2，提案手法×の例を表 4.3，差なしの例を表 4.4 に示す．

表 4.1: 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 4.2: 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．したがって，提案手法○と判断した．

■提案手法×の例

表 4.3: 提案手法×の例

入力文	春が来ると花が咲く。
正解文	hen spring comes, flowers come out.
ベースライン	When spring comes , flowers come out .
提案手法	If the spring comes , you will of flowers 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．したがって，提案手法×と判断した.

■差なしの例

表 4.4: 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.

第5章　考察

表 4.1 の結果から，ベースラインと比べ，提案手法の翻訳精度の方が良かった．表 4.2 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 4.3，表 4.4 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．また，表 4.4 と同様に，そのほかの「差なし」における提案手法の出力も，誤った翻訳がされていた．したがって，全体的な提案手法の有効性が得られなかった．有効性が得られなかった原因を 5.1 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 % (767/3,952) 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．また，今回の実験では，日本語文と文型パターン対の適合に，意味属性制約を使用していた．したがって，必然的にカバー率が下がってしまったと考える．この問題を改善するために，意味属性制約を使用しない場合の提案手法の調査を行った．この調査については，6 章で述べる．

5.1　提案手法の誤り解析

提案手法の有効性を検証するために，対比較評価した 100 文に対して，以下の基準で解析を行った．解析を行った結果を表 5.1 に示す．

■誤り解析の基準

翻訳○　文法的に正しく翻訳できている

スコア×　スコアによる不適切な出力文の選択

統計×　統計翻訳による翻訳の失敗

中間×　適切な中間言語文の候補なし

表 5.1 の結果から，提案手法で文法的に正しく翻訳できた場合は少なかった.「スコア×」は 5.1.1 節,「統計×」は 5.1.2 節,「中間×」は 5.1.3 節でさらに詳しく考察する．

表 5.1: 誤り解析結果

翻訳○	スコア×	統計×	中間×
6	4	29	61

5.1.1 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 5.2 に示す．

表 5.2: 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 5.2 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 5.2 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．

5.1.2 統計翻訳による翻訳失敗の問題

日本語文に対して，翻訳に有効な文型パターン対が適合し，正しく中間言語文が作成されたが，統計翻訳がその中間言語文を正しく翻訳できなかった場合がある．表 5.3 に例を示す.

表 5.3: 統計翻訳で失敗した例

入力文	祖父は老いてもなお精力的に仕事を続けている。
中間言語文	even though 祖父 老い, he なお 精力的に 仕事 を 続け .
翻訳結果	even though my grandfather old , he is still vigorously continue to work .

表 5.3 では，翻訳に有効な中間言語文への変換ができていると判断した．しかし，統計翻訳を行った結果,「祖父 老い」の部分が「my grandfather old」と翻訳され，動詞がない表現となってしまった．このように，統計翻訳が中間言語文の日本語部分を正しく翻訳できない原因として，日本語と英語の対訳の学習量に問題があると考える．学習に用いた中間言語と英語のコーパスでは，日本語と英語が対応する部分が少ない．したがって，統計翻訳に必要な学習量が得られなかったことが考えられる．この問題に対しては，提案手法の学習文に日英対訳文を追加することによって，学習量を増やさなければならない.

5.1.3 適切な中間言語文の候補がない問題

表 5.1 の結果より，提案手法において一番多い問題は，入力文に対して，適切な中間言語文が 1 つも作成されていないことであった．この問題は，以下の要因があると考える.

1. 入力文に対し翻訳に有効な文型パターン対が辞書に存在しない

2. 入力文に対し翻訳に有効な文型パターン対が適合していない

3. 日本語から中間言語への適切な変換ができていない

1 の問題に対しては，文型パターン辞書の文型パターン対を増加させる必要がある．2 の問題に対しては，文型パターン対に用いているルールを変更する必要がある．3 の問題においては，今回の提案手法に用いた以外の手順を加えて，正しい変換を行わなければならない.

5.1.4 学習データ量の問題

翻訳文の中に，表5.4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．したがって，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．5.1.2節の場合と同様に，提案手法における学習データを増やす必要がある．

表 5.4: 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 .

第6章　意味属性制約を使用しない提案手法の実験

入力文と文型パターン辞書を照合した際に，意味属性制約を使用した場合，入力文3,952文のうち文型パターン対と適合したものは，767文であった．今回，カバー率を上げるために，意味属性制約を使用しない実験を行った．その他の実験条件，および，評価基準は，4章の翻訳実験と同じである．

6.1　実験結果

入力文と文型パターン辞書を照合した際に，意味属性制約をしない場合，入力文3,952文のうち文型パターン対と適合したものは，995文であった．この995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行った．意味属性制約を使用する場合としない場合を明瞭に比較するために，今回の対比較評価では，4章の対比較評価で用いた入力文で比較を行った．表6.1に結果を示す．

表 6.1: 対比較評価結果

提案手法○	提案手法×	差なし	同一出力
15	5	80	0

提案手法○が提案手法×より10件多い結果となった．

6.2　考察

入力文と文型パターン辞書を照合する際に，意味属性制約を使用する場合としない場合の提案手法の翻訳実験を行った結果，入力文に対するカバー率が25 % (955/3,952) に上昇した．表6.1より，意味属性制約を使用しない方が良い結果となった．この要因を調査するために，意味属性制約を使用する場合としない場合で翻訳結果が変わった8件の

場合を考察する．翻訳結果が変わった入力文の翻訳において，5.1 節で用いた誤り解析の基準で考察を行った結果を表 6.2 に示す．

表 6.2: 意味属性制約の有無による違いの解析結果

	翻訳○	スコア×	統計×	中間×
意味属性制約あり	1	1	0	6
意味属性制約なし	4	1	3	0

表 6.2 の結果から，意味属性制約を使用する場合では，翻訳に有効な中間言語文が作成されていないが，意味属性制約をしようする場合では作成されていることがわかる．このことは，入力文の要素の意味属性と文型パターンの意味属性制約が異なっても，翻訳に有効な中間言語文への変換が行われればよい可能性がある．また別の要因として，意味属性の制約が強すぎたことによって，適合されてほしい文型パターン対が適合していなかった場合が考えられる．

第7章　単語レベルパターンの実験

句レベルパターンを用いた提案手法では，統計翻訳による変数の翻訳の有効性があまり確認できなかった．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．

7.1　実験結果

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7.1に示す．また，提案手法○の例を表7.2，提案手法×の例を表7.3に示す．

表 7.1: 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

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

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．

表 7.2: 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 fighting it.
ベースライン	Without fighting and beaten .
提案手法	He loses without a fight .
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 $V2$.

表 7.3: 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 $V2$ $N3$.

7.2 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．

第8章　おわりに

本研究では，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，ベースラインと比べ，翻訳精度の翻訳精度の方が良かった．しかし，英語として正しく翻訳できたものは少なかった．主な原因として，入力文に対して適切な文型パターン対の適合が 1 つもない場合があったことと，統計翻訳による誤った訳出が挙げられる．また，入力文に対するカバー率は 20 %であった．カバー率の問題に対して，入力文と文型パターンの適合に意味属性制約を使用しないで翻訳実験を行った結果，カバー率が 25 %に上昇し，翻訳精度も上昇した．このことで，パターンの適合に用いる制約が，有効な翻訳を妨げる可能性があることがわかたった．また，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，ベースラインと比べ，提案手法の翻訳精度の方が良かった．しかし，入力文に対する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る方法と，統計翻訳の精度向上のために，学習量を増やす方法を検討する．

謝辞

本研究を進めるに当たり，種々の御助言を頂きました村田真樹教授，および，徳久雅人講師に御礼申し上げます．本論文をまとめるにあたって，ご指導頂きました松村幸輝教授にお礼申し上げます．また，村上仁一准教授には，終始に渡り研究の進め方や本論文の書き方など，細部に渡る御指導を頂きました．ここに深く感謝いたします．

その他様々な場面で御助力をいただいた計算機工学講座 C 研究室の皆様に感謝の意を表します．

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] 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.

[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, 2006-NL-176, pp.9-16, 2006.

[7] GIZA++, http://www.fjoch.com/GIZA++

[8] 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.

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

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

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