概要

パターン翻訳は，対訳文パターンと対訳句を用いて翻訳を行う．この翻訳方式は入力文が適切な文パターンに適合した場合，翻訳精度の高い文を出力する傾向にある．しかし，対訳文パターンと対訳句は人手で作成するため，開発にコストがかかる．

そこで江木らは，対訳文パターンと対訳句を統計的手法で自動的に作成し翻訳する方法を提案した．これを “パターンに基づく統計翻訳” と呼ぶ．しかし翻訳結果を調査したところ，不適切な対応をとる対訳句が翻訳文に含まれていた．

本研究では不適切な対応をとる対訳句が翻訳文に含まれることを抑制するため，翻訳文の選択において，不適切な対応をとる対訳句を含む翻訳文の出力を抑制する．そこで，人手で作成した対訳句を利用して多変量解析を行う．具体的には，人手で作成した対訳句と自動抽出した対訳句を比較し，ロジスティック回帰分析を用いて，複数の確率から対訳句に確率を付与する．そして翻訳文の選択において，ロジスティック回帰分析から得た確率を用いてパターンに基づく日英統計翻訳を行う．

翻訳実験の結果，入力文 100 文から翻訳文 78 文を得た．さらに従来手法との対比較評価において，提案手法〇が 7 文，提案手法×が 3 文であり，提案手法の有効性が確認できた．

目 次

図 目 次

表 目 次

第1章　はじめに

パターン翻訳は，対訳文パターンと対訳句を用いて翻訳を行う [1]．なお，本論文における対訳句は対訳単語および対訳節などを含む，異なる言語において同じ意味を有する 1 単語以上のまとまりを指す．パターン翻訳は入力文が適切な文パターンに適合した場合，翻訳精度の高い文を出力する傾向にある．しかし，対訳文パターンと対訳句は人手で作成するため，開発にコストがかかる [2]．

そこで江木らは，対訳文パターンと対訳句を統計的手法で自動的に作成し翻訳する方法を提案した（以下，従来手法）[3]．これを “パターンに基づく統計翻訳” と呼ぶ．しかし翻訳結果を調査したところ，不適切な対応をとる対訳句が翻訳文に含まれていた．

不適切な対応をとる対訳句が翻訳文に含まれることを抑制するため，2 つの方法が考えられる．一つは対訳句の作成において，不適切な対応をとる対訳句の作成を抑制する方法である．もう一つは翻訳文の選択において，不適切な対応をとる対訳句を含む翻訳文の出力を抑制する方法である．本研究では後者の方法をとる．そこで，人手で作成した対訳句を利用して多変量解析を行う．具体的には，人手で作成した対訳句と自動抽出した対訳句を比較し，ロジスティック回帰分析を用いて，複数の確率から対訳句に確率を付与する．そして翻訳文の選択において，ロジスティック回帰分析から得た確率を用いてパターンに基づく日英統計翻訳を行い，翻訳精度の調査を行う．

翻訳実験の結果，入力文 100 文から翻訳文 78 文を得た．さらに従来手法との対比較評価において，提案手法○が 7 文，提案手法×が 3 文であり，提案手法の有効性が確認できた．

本論文の構成は以下の通りである．第 2 章で種々の翻訳システムについて説明し，第 3 章で提案手法について説明する．そして第 4 章で実験データと実験結果を示す．第 5 章で提案手法の評価を示し，第 6 章で考察を述べる．

第2章　翻訳システム

2.1　翻訳システムの概要

現在，最も主流となっている翻訳システムとして句に基づく統計翻訳がある．句に基づく統計翻訳は学習データとして対訳文を与えるだけで翻訳が行える．このため，翻訳にかかるコストが低い．さらに，対訳文から対訳単語と単語翻訳確率を自動的に取得することが可能である．

一方，翻訳システムの一手法としてパターン翻訳がある．パターン翻訳は大量の対訳文パターンと対訳句を用いて，翻訳文を得る手法である．パターン翻訳は入力文が適切な対訳文パターンに適合した場合に翻訳精度の高い翻訳文が得られやすいという特徴がある．しかし，パターン翻訳に用いる対訳文パターンと対訳句は人手で作成するため，開発コストが高くなる．そこで，江木らは，対訳文パターンと対訳句を統計的手法で自動的に作成し翻訳する方法を提案した．これをパターンに基づく統計翻訳と呼ぶ．パターン基づく統計翻訳は句に基づく統計翻訳の特徴である対訳文から対訳単語と単語翻訳確率を自動的に取得できる点に注目し，翻訳に用いる対訳文パターンおよび対訳句を統計的手法を用いて自動的に作成する．

本章ではまず，現在主流の翻訳システムである句に基づく統計翻訳について説明し，次にパターン翻訳について説明する．そして，パターンに基づく統計翻訳について説明する．

2.2　句に基づく統計翻訳システム

句に基づく統計翻訳は機械翻訳の一手法である．はじめは単語に基づく統計翻訳が提案されたが，単語に基づく統計翻訳より句に基づく統計翻訳のほうが精度が高いことから，現在は句に基づく統計翻訳が主流となっている．句に基づく統計翻訳は学習データとして大量の対訳文を用いることにより，自動的に翻訳規則を生成し翻訳を行う．

2.2.1 句に基づく日英統計翻訳の概要

日英統計翻訳は日本語入力文 j が与えられたとき，翻訳モデルと言語モデルの組み合わせの中から確率が最大となる英語翻訳文 $\hat{e}$ を検索することで翻訳を行う．基本モデルを式 (2.1) に示す.

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

ここで，$P(j|e)$ は翻訳モデル，$P(e)$ は言語モデルを表す．翻訳モデルは対訳学習文から学習し，言語モデルは目的言語の単言語学習文から学習する．そしてデコーダを用いて，$P(j|e)P(e)$ が最大となる英語翻訳文 $\hat{e}$ を検索する．日英統計翻訳の手順を図 2.1 に示す.

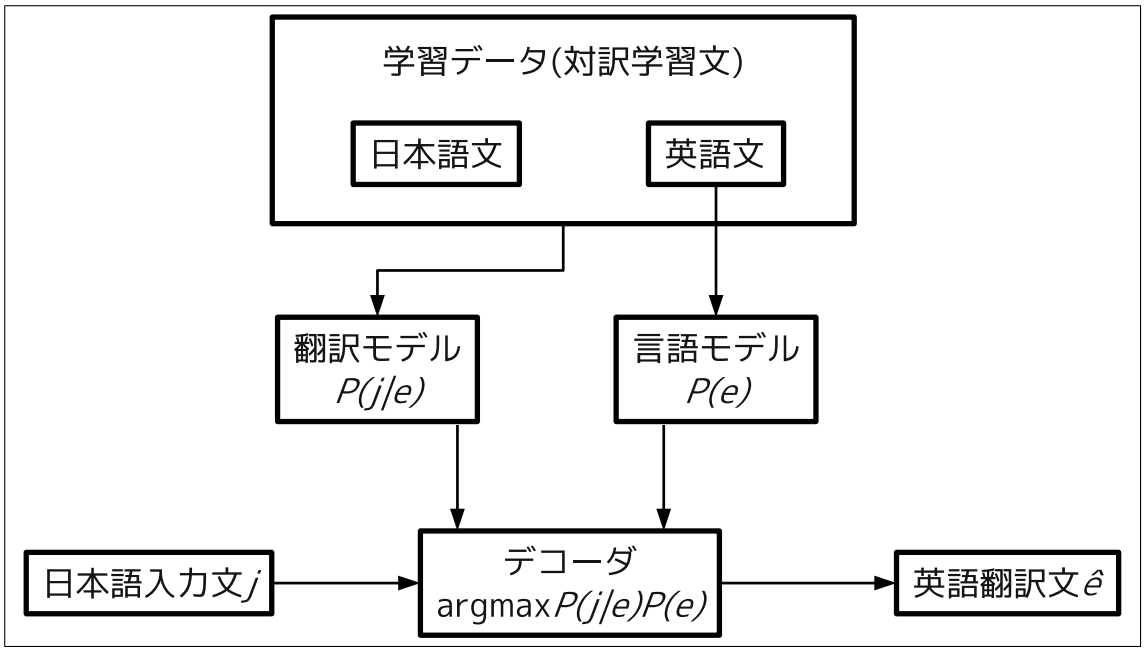

図 2.1 日英統計翻訳の手順

2.2.2 翻訳モデル

翻訳モデルは日本語単語列から英語単語列へ確率的に翻訳を行うためのモデルである．翻訳モデルには単語に基づく翻訳モデルと句に基づく翻訳モデルがある．初期の統計翻訳では，単語に基づく翻訳モデルを用いていた．しかし，翻訳精度の高さから，現在は句に基づく翻訳モデルが主流となっている．句に基づく翻訳モデルは一般的にフレーズテーブルで管理される．句に基づく翻訳モデルの作成手順を以下に示す．

手順 1 後述する IBM モデルを用いて，単語対応づけを行う．

手順 2 ヒューリスティックなルールを用いて句に基づく対応づけを行う．

手順 3 手順 2 で求めた句に基づく対応づけからフレーズテーブルを作成する．

表 2.1 にフレーズテーブルの例を示す．

表 2.1 フレーズテーブルの例

日本語句	英語句	$P(j\|e)$	$\Pi P(j\|e)$	$P(e\|j)$	$\Pi P(e\|j)$
あの 人	That person	0.716	0.182	0.157	0.012
自由 に	free	0.041	0.011	0.153	0.101
見 に 行く	go to see	0.333	0.003	0.500	0.004

ここで，フレーズテーブルは左から順に日本語句，英語句，英日方向の翻訳確率 $P(j|e)$, 英日方向の単語翻訳確率の積 $\Pi P(j|e)$，日英方向の翻訳確率 $P(e|j)$, 日英方向の単語翻訳確率の積 $\Pi P(e|j)$ である．

2.2.2.1 IBM モデル

統計翻訳における単語対応を得るための代表的なモデルとして，IBM 翻訳モデル [4] がある．IBM 翻訳モデルは仏英翻訳を前提としている．しかし，本研究は日英翻訳を扱うため，原言語文を日本語文 J，目的言語文を英語文 E と定義する．

IBM 翻訳モデルにおいて，日本語文 J と英語文 E の翻訳モデル $P(J|E)$ を計算するため，アライメント a と呼ばれる概念を導入する．アライメントはある日本語単語 j と英単語 e の対応関係を意味する．IBM モデルの基本的な計算式を式 (2.2) に示す．

$$P(J|E) = \sum P(J, a|E) \tag{2.2}$$

IBM 翻訳モデルにおいて，各日本語単語に対応する英単語は 1 つであるのに対し，各英単語に対応する日本語単語は 0 から n 個あると仮定する．また，日本語単語に対応する適切な英語単語がない場合，英語文の先頭に特殊文字 e_0 があると仮定し，日本語単語と対応させる．

モデル 1

式 (2.2) は以下の式に分解することができる．m は日本語文の長さ，a_1^{i-1} は日本語語文における，1 番目から $i-1$ 番目までのアライメント，j_1^{i-1} は日本語文における，1 番目から $i-1$ 番目まで単語を表している．

$$P(J,a|E) = P(m|E)\prod_{i=1}^{m} P(a_i|a_1^{i-1}, j_1^{i-1}, m, E)P(j_i|a_1^{i}, j_1^{i-1}, m, E) \tag{2.3}$$

式 (2.3) はとても複雑であるので計算が困難である．そこで，モデル 1 では以下の仮定により，パラメータの簡略化を行う．

- 日本語文の長さの確率 ϵ は m, E に依存しない

$$P(m|E) = \epsilon$$

- アライメントの確率は英語文の長さ l に依存する

$$P(a_i|a_1^{i-1}, j_1^{i-1}, m, E) = (l+1)^{-1}$$

- 日本語の翻訳確率 $t(j_i|e_{a_i})$ は，日本語単語 j_i に対応する英単語 e_{a_i} に依存する

$$P(j_i|a_1^{i}, j_1^{i-1}, m, e) = t(j_i|e_{a_i})$$

パラメータの簡略化を行うことで，$P(J,a|E)$ と $P(J,E)$ は以下の式で表される．

$$P(J,a|E) = \frac{\epsilon}{(l+1)^m}\prod_{i=1}^{m} t(j_i|e_{a_i}) \tag{2.4}$$

$$\begin{aligned} P(J|E) &= \frac{\epsilon}{(l+1)^m}\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\prod_{i=1}^{m} t(j_i|e_{a_i}) \\ &= \frac{\epsilon}{(l+1)^m}\prod_{i=1}^{m}\sum_{k=0}^{l} t(j_i|e_{a_i}) \end{aligned} \tag{2.5}$$

モデル 1 では翻訳確率 $t(j|e)$ の初期値が 0 以外の場合， Expectation-Maximization（EM）アルゴリズムを繰り返し行うことで得られる期待値を用いて最適解を推定する．EM アルゴリズムの手順を以下に示す．

手順 1

翻訳確率 $t(j|e)$ の初期値を設定する．

手順 2

日英対訳対 $(J^{(s)}, E^{(s)})$(但し， $1 \leq s \leq S$) において，日本語単語 j と英語単語 e が対応する回数の期待値を以下の式により計算する．

$$c(j|e; J, E) = \frac{t(j|e)}{t(j|e_0) + \cdots + t(j|e_l)} \sum_{i=1}^{m} \delta(j, j_i) \sum_{k=0}^{l} \delta(e, e_k) \tag{2.6}$$

$\delta(j, j_i)$ は日本語文 J 中で日本語単語 j が出現する回数， $\delta(e, e_j)$ は英語文 E 中で英語単語 e が出現する回数を表している．

手順 3

英語文 $E^{(s)}$ の中で 1 回以上出現する英単語 e に対して，翻訳確率 $t(j|e)$ を計算する．

1. 定数 λ_e を以下の式により計算する．

$$\lambda_e = \sum_j \sum_{s=1}^{S} c(j|e; J^{(s)}, E^{(s)}) \tag{2.7}$$

2. 式 (2.6) より求めた λ_e を用いて，翻訳確率 $t(j|e)$ を再計算する．

$$\begin{aligned} t(j|e) &= \lambda_e^{-1} \sum_{s=1}^{S} c(j|e; J^{(s)}, E^{(s)}) \\ &= \frac{\sum_{s=1}^{S} c(j|e; J^{(s)}, E^{(s)})}{\sum_j \sum_{s=1}^{S} c(j|e; J^{(s)}, E^{(s)})} \end{aligned} \tag{2.8}$$

手順 4

翻訳確率 $t(j|e)$ が収束するまで手順 2 と手順 3 を繰り返す．

モデル 2

モデル 1 では，全ての単語の対応に対して，英語文の長さ l にのみ依存し，単語対応の確率を一定としている．そこで，モデル 2 では，i 番目の日本語単語 j_i と対応する英語単語の位置 a_i は英語文の長さ l に加えて，i と，日本語文の長さ m に依存し，以下のような関係とする．

$$a(a_i|i,m,l) \equiv P(a_i|a_1^{i-1}, j_1^{i-1}, m, l) \tag{2.9}$$

この関係からモデル 1 における式 (2.5) は，以下の式に変換できる．

$$\begin{aligned} P(J|E) &= \epsilon \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{i=1}^{m} t(j_i|e_{a_i}) a(a_i|i,m,l) \\ &= \epsilon \prod_{i=1}^{m} \sum_{k=0}^{l} t(j_i|e_{a_i}) a(a_i|i,m,l) \end{aligned} \tag{2.10}$$

モデル 2 では，期待値は $c(j|e;J,E)$ と $c(k|i,m,l;J,E)$ の 2 つが存在する．以下の式から求められる．

$$\begin{aligned} c(j|e;J,E) &= \frac{t(j|e)}{t(j|e_0) + \cdots + t(j|e_l)} \sum_{i=1}^{m} \delta(j,j_i) \sum_{k=1}^{l} \delta(e,e_k) \\ &= \sum_{i=1}^{m} \sum_{k=0}^{l} \frac{t(j|e) a(k|i,m,l) \delta(j,j_i) \delta(e,e_k)}{t(j|e_0) a(0|i,m,l) + \cdots + t(j|e_l) a(l|i,m,l)} \end{aligned} \tag{2.11}$$

$$\begin{aligned} c(k|i,m,l;J,E) &= \sum_{a} P(a|E,J) \delta(k,a_i) \\ &= \frac{t(j_i|e_k) a(k|i,m,l)}{t(j_i|e_0) a(0|i,m,l) + \cdots + t(j_i|e_l) a(l|i,m,l)} \end{aligned} \tag{2.12}$$

$c(j|e;J,E)$ は対訳文中の英語単語 e と日本語単語 j が対応付けされる回数の期待値，$c(k|i,m,l;J,E)$ は英語単語の位置 k が日本語単語の位置 i に対応付けされる回数の期待値を表している．

モデル 2 では，EM アルゴリズムで計算すると複数の極大値が算出され，最適解が得られない可能性がある．モデル 1 では $a(k|i,m,l) = (l+1)^{-1}$ となるモデル 2 の特殊な場合であると考えられる．したがって，モデル 1 を用いることで最適解を得ることができる．

モデル 3

モデル 3 は，モデル 1 とモデル 2 とは異なり，1 つの単語が複数対応する単語の繁殖数や単語の翻訳位置の歪みについて考慮する．またモデル 3 では単語の位置を絶対位置として考える．モデル 3 では以下のパラメータを用いる．

- 翻訳確率 $P(j|e)$
  英語単語 e が日本語単語 j に翻訳される確率
- 繁殖確率 $n(\phi|e)$
  英語単語 e が ϕ 個の日本語単語と対応する確率
- 歪み確率 $d(i|k,m,l)$
  英語文の長さ l，日本語文の長さ m のとき，k 番目の英語単語 e_k が i 番目の日本語単語 j_i に翻訳される確率

さらに，英語単語が日本語単語に翻訳されない個数を ϕ_0 とし，その確率 p_0 を以下の式で求める．このとき，歪み確率は $\frac{1}{\phi_0!}$ で，$p_0+p_1=1$ で p_0，p_1 は 0 より大きいとする．

$$P(\phi_0|\phi_1^l,E)=\left(\begin{array}{c}\phi_1+\cdots+\phi_l\\ \phi_0\end{array}\right)p_0^{\phi_1+\cdots+\phi_l-\phi_0}p_1^{\phi_0} \tag{2.13}$$

したがって，モデル 3 は以下の式で求められる．

$$\begin{aligned}P(J|E)&=\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}P(J,a|E)\\&=\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\left(\begin{array}{c}m-\phi_0\\ \phi_0\end{array}\right)p_0^{m-2\phi_0}p_1^{\phi_0}\prod_{k=1}^{l}\phi_k!n(\phi_k|e_k)\\&\quad\times\prod_{i=1}^{m}t(j_i|e_{a_i})d(i|a_i,m,l)\end{aligned} \tag{2.14}$$

モデル 3 では，全てのアライメントを計算するため，計算量が膨大となるので期待値を近似により求める．

モデル 4

モデル 4 では，モデル 3 と異なり，単語の位置を絶対位置ではなく，相対位置で考える．またモデル 3 では考慮されていない各単語の位置，例えば形容詞と名詞の関係を考慮する．モデル 4 では歪み確率 $d(i|k,m,l)$ を 2 つの場合で考える．

- 繁殖数が 1 以上である英語単語に対応する日本語単語の中で，最も文頭に近い場合

$$P(\Pi_{[k]1} = i|\pi_1^{[k]-1}, \tau_0^l, \phi_0^l, E) = d_1(i - \odot_{k-1}|\mathcal{A}(e_{[k-1]}), \mathcal{B}(j_i)) \tag{2.15}$$

$\odot_{k-1}$ は k-1 番目の英語単語が対応する日本語単語の位置を表している．

- それ以外の場合

$$P(\Pi_{[k]x} = i|{\pi_{[k]1}}^{x-1}, \pi_1^{[k]-1}, \tau_0^l, \phi_0^l, E) = d_{>1}(i - \pi_{[k]x-1}|\mathcal{B}(j_i)) \tag{2.16}$$

$\pi_{[k]x-1}$ は同じ英語単語に対応している直前の日本語単語を表している．

モデル 5

モデル 4 では，単語の位置に関して直前の単語以外は考慮されていない．したがって，複数の単語が同じ位置に生じたり，単語の存在しない位置が生成される．モデル 5 では，この問題を避けるために，単語を空白部分に配置するよう改善が施されている．

- 繁殖数が 1 以上である英語単語に対応する日本語単語の中で，最も文頭に近い場合

$$\begin{aligned} &P(\Pi_{[k]1} = i|\pi_1^{[k]-1}, \tau_0^l, \phi_o^l, E) \\ &\quad = d_1(v_i|\mathcal{B}(j_i), v_{\odot_{k-1}}, v_m - \phi_{[k]} + 1)(1 - \delta(v_i, v_{i-1})) \end{aligned} \tag{2.17}$$

v_i は i 番目までの空白数，$\mathcal{A}$ は英語の単語クラス $\mathcal{B}$ は日本語の単語クラスを表している．

- それ以外の場合

$$\begin{aligned} &P(\Pi_{[k]x} = i|{\pi_{[k]1}}^{x-1}, \pi_1^{[k]-1}, \tau_0^l, \phi_0^l, E) \\ &\quad = d_{>1}(v_i - v_{\pi_{[k]x-1}}|\mathcal{B}(j_i), v_m - v_{\pi_{[k]x-1}} - \phi_{[k]} + x)(1 - \delta(v_i, v_{i-1})) \end{aligned} \tag{2.18}$$

2.2.2.2 GIZA++

GIZA++[5] とは統計翻訳のために作られた単語の確率の計算を行うツールである．IBM 翻訳モデルに基づいて，単語の対応関係の確率である単語翻訳確率を計算する．

2.2.2.3　フレーズテーブルの作成

GIZA++より IBM 翻訳モデルを推定することで最尤な単語対応を得る．これを日英，英日の両方向に対して行う．日本語文 “彼は犬を優しく世話した” とその対訳英語文 “He treated his dog kindly” を例に挙げ，日英方向の単語対応の例を図 2.2 に，英日方向の単語対応の例を図 2.3 に示す．ここで●は対応点を示す．

	He	treated	his	dog	kindly
彼	●				
は			●		
犬				●	
を		●			
優しく					●
世話					●
し		●			
た		●			

図 2.2　日英方向の単語対応の例

	He	treated	his	dog	kindly
彼	●				
は					
犬				●	
を			●		
優しく		●			●
世話					
し					
た					

図 2.3　日英方向の単語対応の例

次に，両方向の対応付けから，ヒューリスティックなルールにより，1 対多の対応を認めた単語対応の計算を行う．ここで，ヒューリスティックとは人間の日々の意思決定に類似した直感的かつ発見的な思考方法である．基本のヒューリスティックとして “intersection” と “union”，“grow”がある．さらに最終処理として “final” と “final-and”がある．

intersection は日英方向と英日方向の両方向に単語対応が存在する場合，その単語対応を残す．union は日英方向と英日方向のどちらか一方に単語対応が存在する場合，その単語対応を残す．intersection の例を図 2.4 に union の例を図 2.5 に示す．

	He	treated	his	dog	kindly
彼	●				
は					
犬				●	
を					
優しく					●
世話					
し					
た					

図 2.4 intersection の例

	He	treated	his	dog	kindly
彼	●				
は			●		
犬				●	
を		●	●		
優しく		●			●
世話					●
し		●			
た		●			

図 2.5 union の例

grow は intersection の対応点の縦横方向に union の対応点がある場合，その単語対応を intersection に追加する方法である．さらに，縦横方向に加え，対角方向に存在する union の対応点を intersection に追加する方法として grow-diag がある．grow-diag の例を図 2.6 に示す.

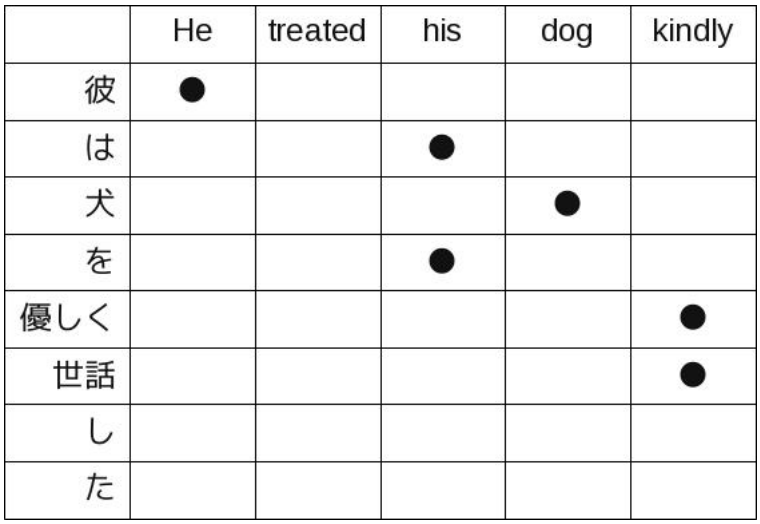

図 2.6 grow-diag の例

最終処理として final と final-and がある．final は少なくとも一方の言語の単語に単語対応がない場合，union の単語対応を追加する．final-and は両言語の単語に単語対応がない場合，union の単語対応を追加する方法である.

“grow-diag-final-and”は grow-diag において，両言語の単語に対応がない場合，union の単語対応を追加する．grow-diag-final-and の例を図 2.7 に示す.

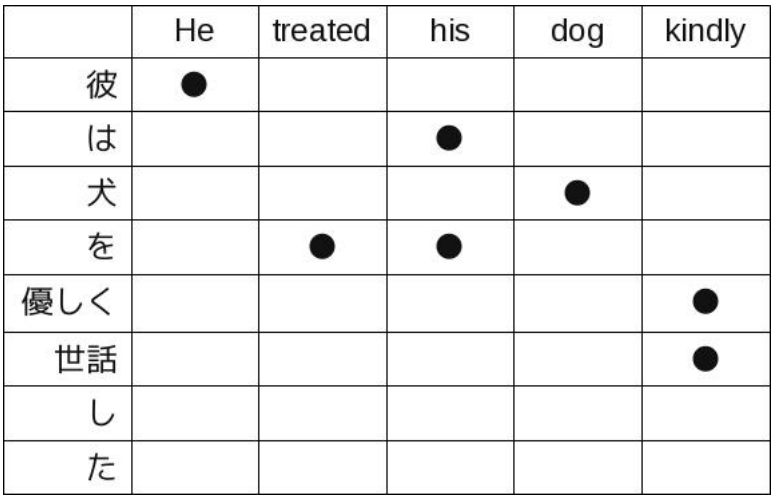

図 2.7 grow-diag-final-and の例

単語対応のうち，矛盾しない全ての対訳句を抽出する．抽出した対訳句の例を表 2.2 に示す.

表 2.2 抽出した対訳句の例

日本語句	英語句
は 犬 を	treated his dog
彼 は 犬 を	He treated his dog
優しく 世話	kindly
優しく 世話 し た	kindly

2.2.3 言語モデル

言語モデルは単語列に生成確率を付与するモデルである．言語モデルは単言語学習文から学習される．統計翻訳では一般的に，N-gram モデルを用いる．

N-gram モデルは “単語列 $\omega_1^n = \omega_1, \omega_2, \omega_3, ..., \omega_n$ の i 番目の単語 ω_i の生起確率 $P(\omega_i)$ は直前の $(N-1)$ 単語に依存する” という仮説に基づくモデルである．単語列 ω_1^n の生起確率 $P(\omega_1^n)$ の計算式を式 (2.19) に示す.

$$
\begin{aligned}
P(\omega_1^n) &= P(\omega_1) \times P(\omega_2|\omega_1) \times P(\omega_3|\omega_1^2) \times ... \times P(\omega_n|\omega_1^{n-1}) \\
&\approx P(\omega_1) \times P(\omega_2|\omega_1) \times P(\omega_3|\omega_1^2) \times ... \times P(\omega_n|\omega_n^{n-1}) \\
&= \Pi_{i=1}^{n} P(\omega_i|\omega_{i-(N-1)}^{i-1})
\end{aligned}
\tag{2.19}
$$

ここで，ω_i^j は i から j 番目までの単語列を表す．例えば，“She is a teacher” という単語列に対して 2-gram モデルを適応した場合，単語列の生起確率は式 (2.20) で計算される．

$$
P(\text{“}She\ is\ a\ teacher\text{”}) \simeq P(She) \times P(is|She) \times P(a|is) \times P(teacher|a) \tag{2.20}
$$

3-gram の場合，“She is” の単語列の次に “a” が生じる確率を考える．しかし，N-gram モデルにおいて，信頼できる値を算出するためには，大規模な対訳学習文を用いることが必要である．そこで，出現数の少ない単語列をモデルの学習から削除する手法（カットオフ）や，確率が 0 になるのを防ぐため，大きい確率を小さく，小さい確率を大きくする手法（スムージング）が提案されている．スムージングの代表的な手法にバックオフ・スムージングがある．バックオフ・スムージングは学習データに出現しない N-gram を低次の $(N-1)$-gram で推定する手法である．表 2.3 に N-gram モデルにおける 2-gram の例を示す．

表 2.3 2-gram の例

2-gram の単語列 $\omega_1\omega_2$	2-gram の確率 $log_{10}(P(\omega_2\|\omega_1))$	バックオフ・スムージングによる確率 $log_{10}(P(\omega_2\|\omega_1))$
American English	-1.885179	-0.0880824
He is	-2.023028	-0.000409741
I have	-1.509964	-0.05597086

2.2.4 デコーダ

デコーダは翻訳モデルと言語モデルの全ての組み合わせから確率が最大となる翻訳文を検索し出力する．代表的なデコーダとして，Moses[6] がある．デコーダの手順を図 2.8 に示す．

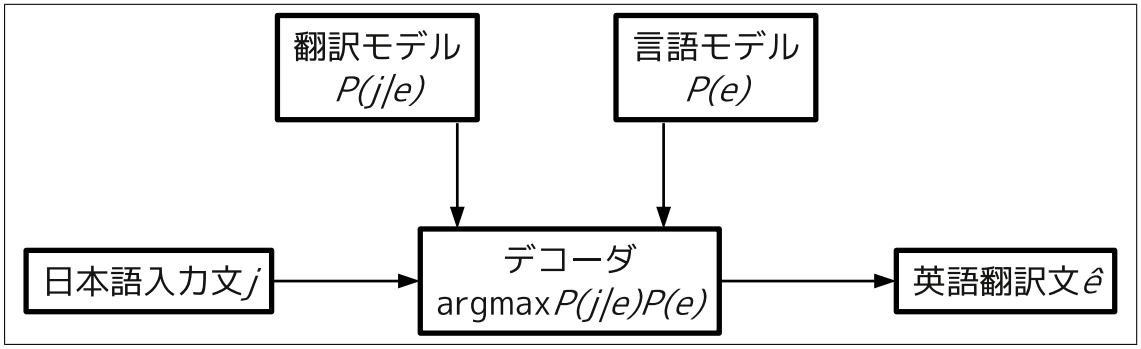

図 2.8 デコーダの手順

2.3 パターン翻訳システム

パターン翻訳は機械翻訳の一手法である．大量の対訳文パターンと対訳句を用いて翻訳を行う．パターン翻訳は適切に対訳文パターンが適合した場合，文全体の構造を保持した翻訳精度の高い翻訳文を出力する傾向にある．しかし，一般的にパターン翻訳は対訳文パターンや対訳句を人手で作成するため，開発にコストがかかる．

2.3.1 日英パターン翻訳の概要

日英パターン翻訳は日本語入力文が与えられたとき，まず日本語入力文の表現を解析する．そして解析結果を日本語文パターンと照合し，適合する日本語文パターンを抽出する．次に抽出された日本語文パターンから適切な日本語文パターンの選択を行い，対となる英語文パターンの抽出を行う．最後に抽出した英語文パターンを用いて英語翻訳を生成する．日英パターン翻訳の手順を図 2.9 に示す．

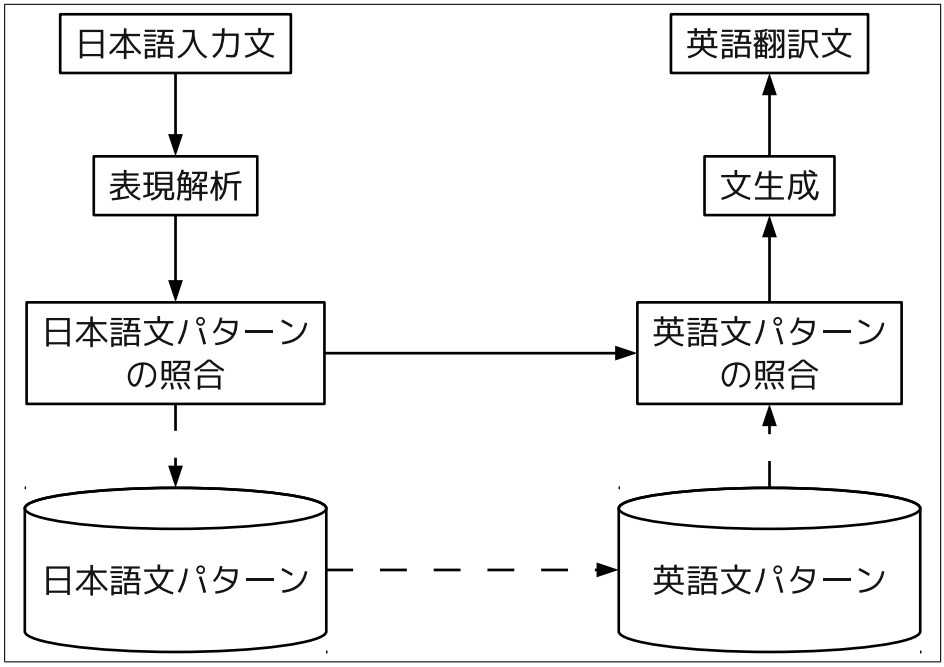

図 2.9 日英パターン翻訳の手順

2.3.2 表現解析

表現解析は日本語入力文の形態素解析を行う．形態素解析器 [7] に日本語入力文を入力すると，日本語入力文は単語に分割され，単語それぞれについて品詞や意味属性などの情報が出力される．表現解析の例を表 2.4 に示す.

表 2.4 表現解析の例

日本語入力文	ここできみを見かけるとは夢にも思わなかった。
出力	1. /ここで (4100) 2. /きみ (1710, あなた,NI:15,IM:11120,IM:11121,IM:11122,$\cdots$) 3. +を (7430) 4. /見かける (2416, 見掛ける,NY:30,KR:0400a13,IY:1110,IY:2000) 5. +と (7420) 6. +は (7530) 7. /夢にも (4100,KR:4111f29) 8. /思わ (2392, 思う,NY:32,NY:31,KR:0601a01,KR:1500a00,$\cdots$) 9. +なかっ(7184, ない) 10. +た (7216) 11. +。([P]0110) 12. /nil

2.3.3 文パターンの選択

日本語入力文の形態素解析結果と日本語文パターンの照合を行い，適合した日本語文パターンの抽出を行う．照合には検索ツール [8] を用いる．ここで，英語翻訳文の全体的な構造は英語文パターンにより決定できる．英語文パターンは日本語入力文と適合した日本語文パターンにより決まる．そこで日本語入力文に適切に適合する日本語文パターンの選択を行う必要がある．選択手法には多変量解析を用いる手法などがあるが，[1] では人手により文パターンの選択を行っている．適合した対訳文パターンの例を表 2.5 に示す.

表 2.5 適合した対訳文パターンの例

日本語入力文	雨の中に立って彼を待った。
日本語文パターン	$< N1$ は $> N2$ の中に立って $N3$ を $V4.kako$ 。
英語文パターン	$<$I $\mid N1 > V4$^$past$ for $N3$^obj standing in $N2$.

英語文の部分的構造は英語文パターンの記号と変数および関数で定義されている．記号は日本語入力文と日本語文パターンの照合結果により処理が決まる．記号は構造上の線形性を記述するために使用される．英語文パターンで使用される記号を表 2.6 に示す．記号を分類すると (1)OR 条件による線形要素の適合範囲拡大のための記号，(2) 線形要素の挿入や省略のための記号に分類できる．

表 2.6 英語文パターンで使用される記号

分類	記号名	書式	説明
(1)	要素選択記号	(パターン記述 1|パターン記述 2|$\cdots$)	複数記述された要素のいずれかを使用
	対応型要素選択記号	#数 (パターン記述 1|パターン記述 2|$\cdots$)	日英方向の要素の対応関係を保ったまま対応する順序で日英方向の要素を指定
	訳出要素選択記号	#数 $<$パターン記述 1|パターン記述 2$>$	日本語文パターンで左側のパターン記述に対応する要素が適合しなかった場合，右側のパターン記述を使用
	標準形表現記号	‘用言の標準形’	字面部分の標準形を指定
(2)	任意要素記号	#数 [$\cdots$]	日本語文パターンで適合する要素がある場合のみ訳出対象
	要素挿入記号	#数 { 挿入要素 }	英語側の要素の中に副詞等の別の要素を挿入し訳出

変数には表 2.7 の変数を用いる．

表 2.7 英語文パターンで使用される変数

分類	変数	説明	分類	変数	説明
単語	N	名詞または名詞複合語	句	NP	名詞句
	$TIME$	時詞		VP	動詞句
	NUM	数詞		AJP	形容詞句
	ND	用言性名詞		$AJVP$	形容動詞句
	V	動詞		$ADVP$	副詞句
	AJ	形容詞	節	CL	節
	AJV	形容動詞	その他	ANY	直接引用で使用され，どの要素も使用可
	ADV	副詞			
	REN	連体詞			
	GEN	限定詞			

英語文パターンの記述子は表 2.7 の変数および関数である．関数は制約の機能を持ち，先行する記述子に制約を与え，変数のみによる制約よりも厳しい絞り込みが可能である．英語文パターンで使用される関数を表 2.8 に示す．

表 2.8 英語文パターンで使用される関数

分類	関数	説明
名詞制約	^obj，^$poss$，^$pron$ ^$adposs$，^$reflex$	目的格，所有格，代名詞 独立所有格，再帰代名詞
動詞制約	^$base$，^$past$，^ing ^$present$，^ed	原形，過去形，現在分詞形 現在形，過去分詞形
形容詞制約	^er，$verb$|^st	比較級，最上級

2.3.4 文生成

対訳句を用いて文パターンの変数に適合する局所的な要素の翻訳を行う．通常日本語の要素を英語に翻訳する場合，同一の意味の表現が多数ある．また，関数による制約がない場合，活用形の表現もさまざまである．従って，日本語要素の翻訳は，変数や関数の制約による絞り込みを行い，それでもなお複数の表現や活用形がある場合，それ以上の絞り込みは行わずに対訳句を英語文パターンに代入する．これらを翻訳文の候補とする．

次に翻訳文の候補より尤もらしい候補を選択する．翻訳候補の選択には N-gram モデルを用いる．まず英語文パターンをワードグラフに変換し，ワードグラフの隣接する候補の連鎖確率を求める．日本語入力文 “彼は車を持っている。”，英語文パターン “<He|$N1$ > $V3$^$present$ (a|an) $N2$.”，変数 “$N1$=彼，$N2$=車，$V3$=持っ”である場合のワードグラフを図 2.10 に示す．図 2.10 において，太線は連鎖確率が高いつながりを表している．

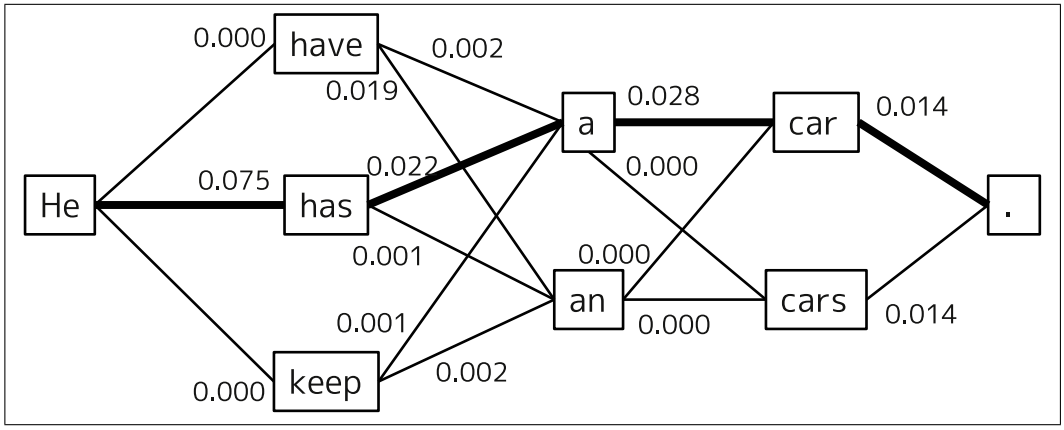

図 2.10　ワードグラフの例

2.4　パターンに基づく統計翻訳システム

パターン翻訳は対訳文パターンと対訳句を人手で作成するため，開発コストが高くなる．そこで江木らは対訳文パターンと対訳句を統計的手法で自動作成し翻訳する手法を提案した．これをパターンに基づく統計翻訳と呼ぶ．パターンに基づく統計翻訳は句に基づく統計翻訳の特徴である対訳文から対訳単語と対訳単語翻訳確率を自動的に作成できる点に注目し，翻訳に用いる対訳文パターンおよび対訳句を統計的手法を用いて自動的に作成する．

また，パターン翻訳は対訳文パターンに変数として品詞情報を付与している．一方，パターンに基づく統計翻訳システムは対訳文パターンに変数による制約がない．よって翻訳を行う際，入力文に対して形態素解析器による品詞情報の取得を行う必要がない．

2.4.1　パターンに基づく日英統計翻訳の概要

パターンに基づく統計翻訳は，大きく 5 つのステップで翻訳を行う．パターンに基づく日英統計翻訳の概要を以下に示す．

ステップ 1 対訳単語の作成

GIZA++を用いて，対訳単語を作成する．

ステップ 2 単語に基づく対訳文パターンの作成

対訳単語を用いて，単語に基づく対訳文パターンを作成する．

ステップ 3 対訳句の作成

単語に基づく対訳文パターンを用いて，対訳句を作成する．

ステップ 4 句に基づく対訳文パターンの作成

対訳句を用いて，句に基づく対訳文パターンを作成する．

ステップ 5 文生成

対訳句と句に基づく対訳文パターンを用いて，翻訳文生成を行う．

2.4.2 対訳単語の作成

GIZA++を用いて，対訳学習文の単語対応を取り，対訳単語と単語翻訳確率を得る．対訳単語作成の例を図 2.11 に示す．

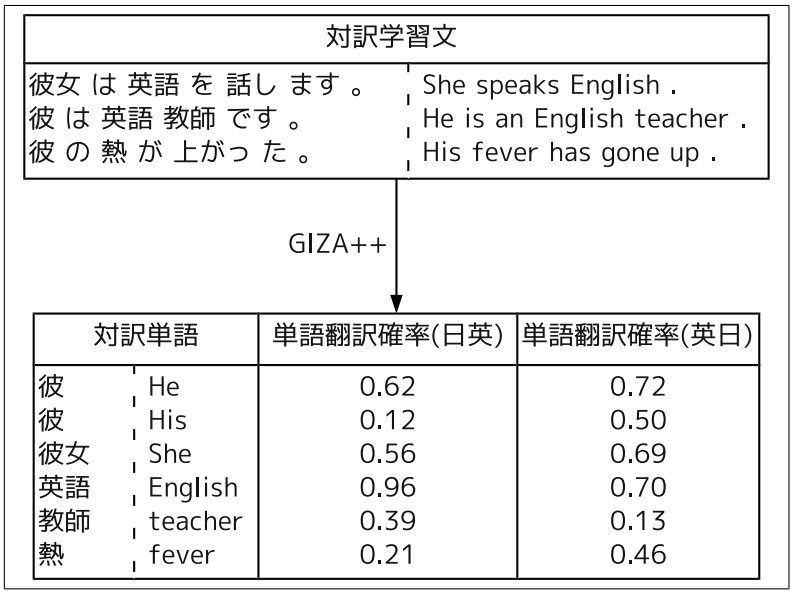

図 2.11 対訳単語作成の例

2.4.3 単語に基づく対訳文パターンの作成

対訳単語と対訳学習文を用いて，単語に基づく対訳文パターンを作成する．まず，対訳単語と対訳学習文を照合する．そして対訳学習文において，適合した対訳単語を変数化する．

なお変数の組み合わせを考慮して，単語に基づく対訳文パターンは可能な限り多く作成する．具体的には，対訳学習文 1 文に対し n 個の対訳単語が変数化できる場合，対訳単語を “変数化する・変数化しない”の 2 通りの組み合わせがあるため，全ての組み合わせを考慮し，単語に基づく対訳文パターンを 2^n 通り作成する．単語に基づく対訳文パターン作成の例を図 2.12 に示す．なお，変数は “X 数”で表される．

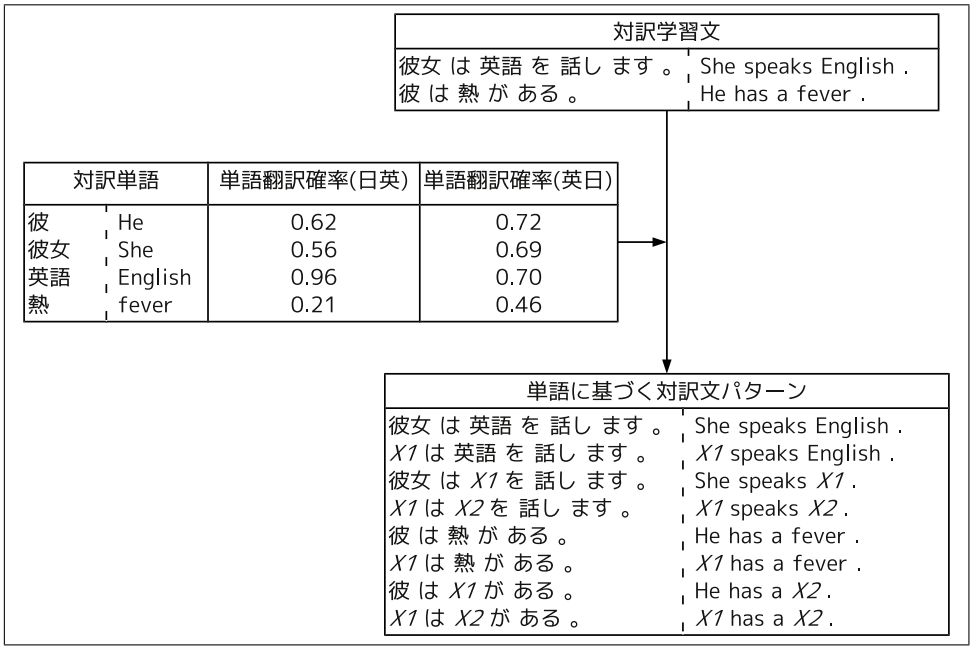

図 2.12 単語に基づく対訳文パターン作成の例

2.4.4 対訳句の作成

a) 対訳句の抽出

単語に基づく対訳文パターンと対訳学習文を用いて，対訳句を抽出する．まず単語に基づく対訳文パターンを用いて対訳学習文の構文解析を行う．構文解析にはボトムアップアルゴリズムを用いる．構文解析を行うために，単語に基づく対訳文パターンはチョムスキー標準形（式 (2.21)）の文脈自由文法に変換する．式 (2.21) において，V_X は非終端記号の集合，V_C は前終端記号の集合を表す．

$$\mathrm{A} \rightarrow \mathrm{BC} \quad \mathrm{A} \in V_X, B, C \in (V_C \cup V_X) \tag{2.21}$$

パターンに基づく統計翻訳は対訳文パターンの変数に品詞よる制約を持たないため，チョムスキー標準形の規則は変数化していない部分（以下，字面）と変数および記号で構成

される．日本語文パターン “$X1$ は $X2$ $X3$ が いる 。”のチョムスキー標準形を表 2.9 に示す．表 2.9 において，“S 数”は前終端記号を，“S”は開始記号を表す．なお，チョムスキー標準形への変換の際に，対訳文パターンに対して終端記号 “END1”と “END2”を付与する．

表 2.9 チョムスキー標準形の例

規則
S → $X1$ S1
S1 → は S2
S2 → $X2$ S3
S3 → $X3$ S4
S4 → が S5
S5 → いる S6
S6 → 。 S7
S7 → END1 END2

構文解析は三角行列 a_{ij}（$1 \leq i \leq j \leq n$，n は単語数）を用いて解析を行う．日本語入力文 “あの 人 は たくさん の 友達 が いる 。”を表 2.9 の文法規則を用いて構文解析する場合の三角行列を図 2.13 に示す．

	1	2	3	4	5	6	7	8	9	10	11
1	あの	*X1*									S
2		人									
3			は								S1
4				たくさん/*X2*	*X2*						S2
5					の	*X3*					S3
6						友達/*X3*					S3
7							が				S4
8								いる			S5
9									。		S6
10										END1	S7
11											END2

図 2.13 構文解析の例

構文解析において，変数に複数単語が適合することを許す．また，変数は品詞情報や関

数による制約を持たないため，全ての単語および複数単語が変数と適合する．文法規則を用いて要素 $a_{n-1,n}$ から $a_{1,n}$ へボトムアップアルゴリズムで文法規則の探索を行う．要素 $a_{k,n}$ において，文法規則の探索を行う場合，$a_{k,k+i}$ と $a_{k+1+i,n}$ を参照し探索を行う．ここで，i は 0 から $k+1+i=n$ になるまで 1 ずつ加算していく．例えば，図 2.13 の要素 $a_{7,11}$ の S4 は表 2.9 の文法規則より $a_{7,7}$ の “が” と $a_{8,11}$ の “S5” の文法規則より作られる．また，要素 $a_{6,11}$ の $S3$ は $a_{6,6}$ の “友達” を変数 “$X3$” とし，$a_{7,11}$ の “S4” との文法規則より作られる．なお，各記号が，どの記号，変数および字面から作られたかをポインタにより記録しておく．そして，要素 $a_{1,n}$ まで，文法規則の探索を行い，$a_{1,n}$ に開始記号 S が与えられれば，入力文の構文解析に成功したといえる．構文解析に成功した場合，開始記号 S からポインタをたどり，各変数に対応する入力文の単語列を句として抽出する．図 2.13 の構文解析により抽出される日本語句を表 2.10 に示す．

表 2.10 抽出される日本語句の例

日本語句
あの 人
たくさん
たくさん の 友達
の 友達

日本語文パターンの対となる英語文パターンを用いて日本語学習文の対となる英語学習文に対しても同様の構文解析を行い，変数に対応する英語学習文の単語列を句として抽出する．そして，得られた日本語句と英語句において，同一の変数におけるすべての組み合わせを取得し，対訳句とする．

対訳句抽出の流れを図 2.14 に示す．なお，対訳句の抽出は網羅的に行うため，不適切な対応をとる対訳句を抽出する問題がある．図 2.14 の例では “彼の顔に” と “He” のような不適切な対応をとる対訳句がある．

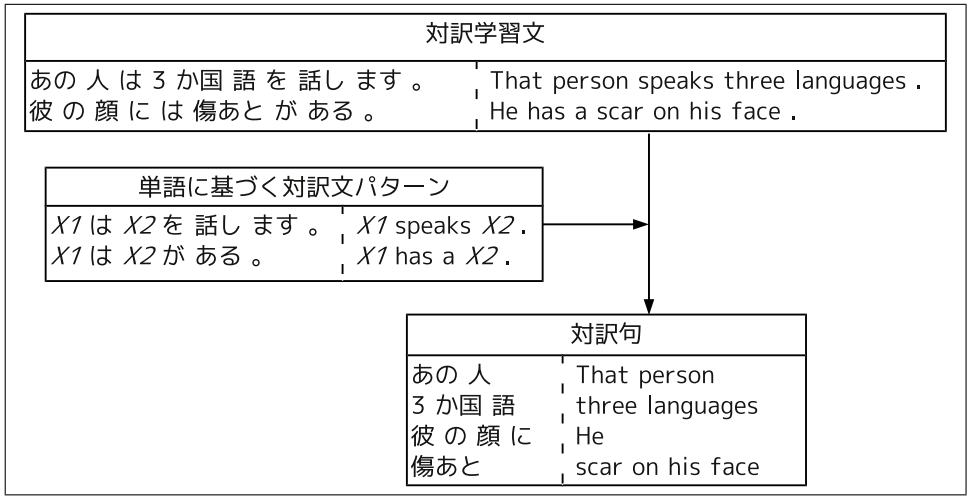

図 2.14 対訳句抽出の流れ

b) 対数フレーズ確率の付与

対訳単語と単語翻訳確率を用いて，対訳句に確率を付与する．まず対訳句において，日本語句の単語と英語句の単語の全ての組み合わせを得る．次に日本語単語に対応する英語単語の中で，単語翻訳確率の最大値を得る．これを各日本語単語に対して行い，得られた値について対数の総和を求める．本研究ではこの値を日英方向の対数フレーズ確率と呼ぶ．同様に対訳句において，英語単語に対応する日本語単語の中で，単語翻訳確率の最大値を取得し，英日方向の対数フレーズ確率も求める．日英方向の対数フレーズ確率付与の例を図 2.15 に示す．

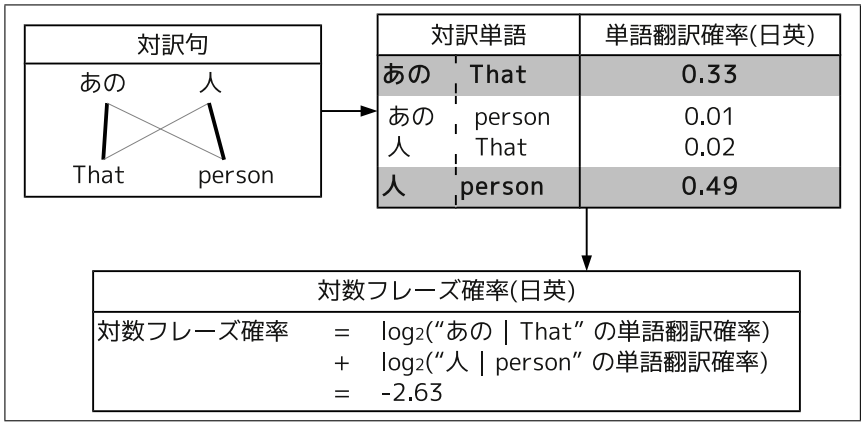

図 2.15 対数フレーズ確率付与の例（日英）

2.4.5　句に基づく対訳文パターンの作成

a) 対訳文パターンの作成

対訳句と対訳学習文を用いて，句に基づく対訳文パターンを作成する．まず単語に基づく対訳文パターンの作成（2.4.3 節）と同様に，変数の組み合わせを考慮して，句に基づく対訳文パターンを可能な限り多く作成する．句に基づく対訳文パターン作成の例を図 2.16 に示す．

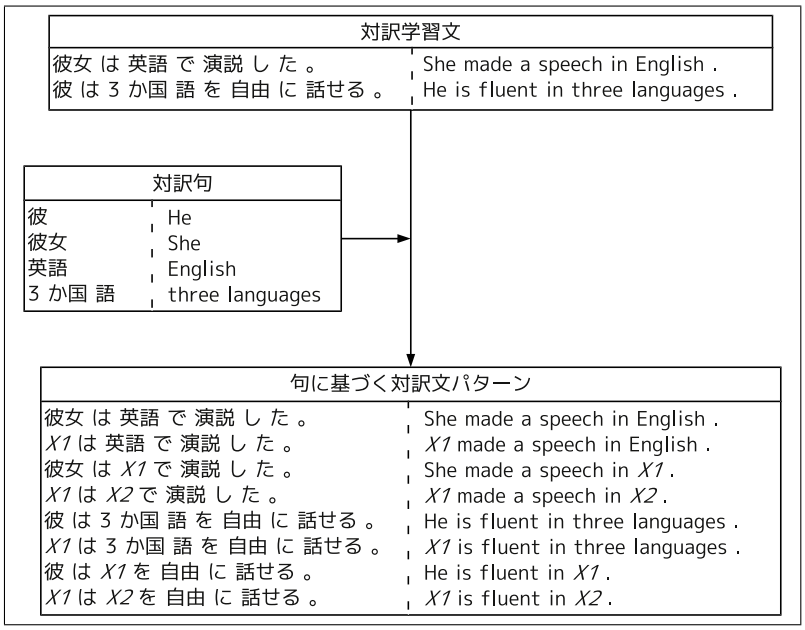

図 2.16　句に基づく対訳文パターン作成の例

b) 対数文パターン確率の付与

対訳単語と単語翻訳確率を用いて，句に基づく対訳文パターンに確率を付与する．句に基づく対訳文パターンにおいて字面を用いて，対数フレーズ確率の付与（2.4.4 節 b）と同様の計算手法で確率を求める．本研究ではこの値を対数文パターン確率と呼ぶ．日英方向の対数文パターン確率付与の例を図 2.17 に示す．

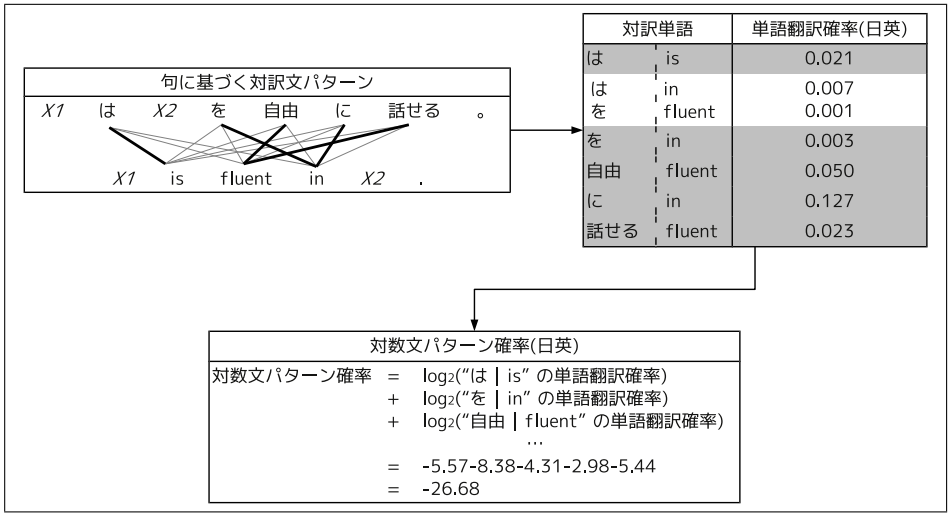

図 2.17 対数文パターン確率付与の例（日英）

2.4.6 文生成

句に基づく対訳文パターンと対訳句を用いて，文生成を行う．まず，日本語文パターンと入力文を照合し，入力文に適合する日本語文パターンを選択する．なお，日本語文パターンの選択には入力文と日本語文パターンの字面を比較し，字面が多く一致する文パターンを優先して選択する．そして，日本語文パターンに対応する英語文パターンと対数文パターン確率を取得する．次に対訳句を用いて文パターンの変数部に適合する局所的な要素の翻訳を行う．パターンに基づく統計翻訳は変数に品詞情報などの制約がないため，日本語の要素を英語に翻訳する場合，多数の翻訳候補がある．そこで，局所的な要素の翻訳は対訳句に付与された対数フレーズ確率と N-gram モデルを用いて絞り込みを行う．具体的には各要素に用いられた対訳句の対数フレーズ確率と生成された文の N-gram モデルの総和を求め，その総和が最大となる文を選択された文パターンにおける翻訳候補文とする．パターンに基づく統計翻訳において，N-gram モデルは 3-gram を用いる．各適合する文パターン対して同様の処理を行い，入力文に対する翻訳候補文を生成する．

最後に，翻訳候補文から翻訳文の選択を行う．翻訳文の選択には文生成に用いた対訳文パターンの対数文パターン確率と対訳句の対数フレーズ確率，N-gram モデル (3-gram) を用いる．各翻訳候補文の対数文パターン確率と対数フレーズ確率，N-gram モデル（3-gram）の総和を求め，翻訳候補文の中で総和が最大となる文を翻訳文として出力する．日英翻訳における文生成の例を図 2.18 に示す．

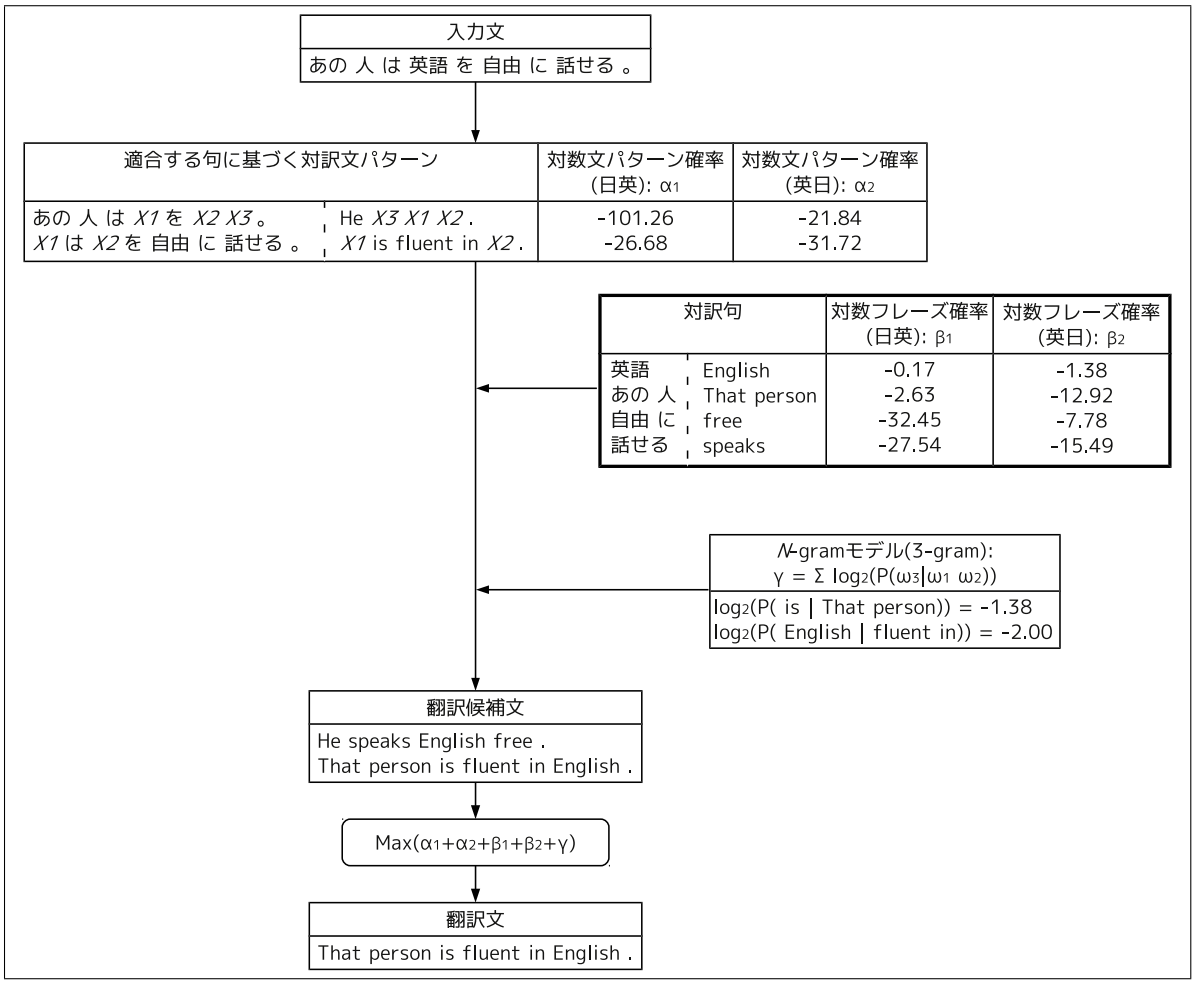

図 2.18 日英翻訳における文生成の例

第3章　提案手法

江木らの従来手法は対訳文パターンと対訳句を統計的手法を用いて自動作成することで開発コストを低くさせることに成功した．しかし，翻訳文の調査を行ったところ，不適切な対応をとる対訳句が翻訳文に含まれていた．そこで，本研究は翻訳文の選択において不適切な対応をとる対訳句を含む翻訳文の出力を抑制するため，人手で作成した対訳句を利用して多変量解析を行い，自動抽出した対訳句に確率を付与することを提案する．具体的には，人手で作成した対訳句は適切な対応をとる対訳句であると仮定し，人手で作成した対訳句との比較により，自動抽出した対訳句に従属変数を設定する．そして，複数の確率を独立変数として設定し，ロジスティック回帰分析から対訳句に確率を付与する．そして，従来手法における対数フレーズ確率をロジスティック回帰分析から得た確率に置き換え，パターンに基づく日英統計翻訳を行う．

3.1　ロジスティック回帰分析

ロジスティック回帰分析 [9] は，独立変数が観測値や確率などの量的な値で，従属変数が 2 値（0 または 1）の質的な値である場合に用いる回帰分析手法である．ロジスティック回帰分析は生物学の分野では，投薬実験による実験対象物の生存・死亡などの 2 値しかとらない結果において，その反応の割合を分析するために広く利用されている．また，ロジスティック回帰分析を用いることで独立変数の条件下である事象が発生する条件付き確率を予測することができる．このため医学の分野においても，ある病気が発生する要因を検索するために，いくつかの検査数値に対して発病する確率を推定する方法として用いられている．

ある独立変数 $X = \{x_1, x_2, ..., x_n\}$ に対して，ある事象が発生する（=1）もしくは発生しない（=0）ことを表す従属変数を Y とすると独立変数 X に対して発生する確率 p と発生しない確率 $1 - p$ はそれぞれ式 (3.1), 式 (3.2) と表すことができる．

$$p = Pr(Y = 1|X) \tag{3.1}$$

$$1 - p = Pr(Y = 0|X) \tag{3.2}$$

このとき，ロジスティック関数を用いて独立変数とある事象の発生の関係を記述するモデルとしてロジスティック回帰モデル (式 (3.3)) がある．

$$p = \frac{exp(a + b_1x_1 + b_2x_2 + ... + b_nx_n)}{1 + exp(a + b_1x_1 + b_2x_2 + ... + b_nx_n)} \tag{3.3}$$

ここで，a は定数，b_n は独立変数 x_n の回帰係数である．式 (3.3) はロジット変換により線形回帰モデル（式 (3.4)）に変換される．

$$\log \frac{p}{1 - p} = a + b_1x_1 + b_2x_2 + ... + b_nx_n \tag{3.4}$$

式 (3.4) により線形回帰モデルの枠組みでモデルの回帰係数を推定することが可能である．本研究では統計ソフト R[10] を用いて，ロジスティック回帰分析を行う．

3.2 モデルの作成

本研究は不適切な対応をとる対訳句を含む翻訳文の出力を抑制するため，ロジスティック回帰分析を用いて自動抽出した対訳句に確率を付与する．具体的には，人手で作成した対訳句は適切な対応をとる対訳句であると仮定し，人手で作成した対訳句との比較により，自動作成した対訳句に従属変数を設定する．そして，複数の確率を独立変数として設定する．

3.2.1 従属変数の設定

人手で作成した対訳句を利用して，自動抽出した対訳句に従属変数 Y を付与する．人手で作成した対訳句は適切な対応をとる対訳句であると仮定し，対訳句が適切な対応をとる確率を算出することを目的として，従属変数を設定する．まず，人手で作成した対訳句と 2.4.4 節 a で自動抽出した対訳句を照合する．そして，人手で作成した対訳句と一致した対訳句は適切な対応をとる（Y=1）と判断する．一方，人手で作成した対訳句と一致しない対訳句は不適切な対応をとる（Y=0）と判断し，自動抽出した対訳句にそれぞれ従属変数を付与する．

3.2.2 独立変数の設定

自動抽出した対訳句に独立変数を付与する．独立変数には以下の 6 つの確率を用いる．

- 日英方向の対数フレーズ確率: x_1
- 英日方向の対数フレーズ確率: x_2
- 対訳学習文における日英方向の対数翻訳確率: x_3
- 対訳学習文における英日方向の対数翻訳確率: x_4
- 対訳句抽出における日英方向の対数翻訳確率: x_5
- 対訳句抽出における英日方向の対数翻訳確率: x_6

a) 対数フレーズ確率　x_1,x_2

従来手法で用いる対数フレーズ確率である．詳細は 2.4.4 節 b を参照のこと．

b) 対訳学習文における対数翻訳確率　x_3,x_4

対訳学習文において句 j を句 e に翻訳する確率である．対訳学習文における日本語学習文 J の日本語句 j から英語学習文 E の英語句 e への翻訳確率 $P_s(e|j)$ は式 (3.5) で求められる．

$$P_s(e|j) = \frac{count_s(j,e)}{count_s(j)} \tag{3.5}$$

ここで，$count_s(j,e)$ は対訳学習文における対訳句 (j,e) の共起回数であり，$count_s(j)$ は日本語学習文 J における日本語句 j の出現回数である．確率 $P_s(e|j)$ の対数をとり，対訳学習文における日英方向の対数翻訳確率とする．同様に，対訳学習文における英日方向の対数翻訳確率も求める．対訳学習文における日英方向の対数翻訳確率の例を表 3.1 に示す．

表 3.1 対訳学習文における日英方向の対数翻訳確率の例

日本語句	英語句	$count_s(j)$	$count_s(j,e)$	$\log(P_s(e\|j))$
あなた の 健康	your health	3	2	-0.176
駅 から	from the station	21	9	-0.368
数 マイル	A few miles	7	1	-0.845

c) 対訳句抽出における対数翻訳確率　x_5,x_6

対訳句の抽出（2.4.4 節 a）において，対訳句は網羅的に抽出するため，多くの場合，同一の対訳句が複数抽出される．対訳句抽出における対数翻訳確率は対訳句を重複して抽出した回数を利用した確率である．なお，文生成に用いる対訳句は重複する対訳句がないように作成する．対訳句抽出における日本語句 j から句 e への翻訳確率 $P_v(e|j)$ は式 (3.6) で求められる．

$$P_v(e|j) = \frac{count_v(j,e)}{count_v(j)} \tag{3.6}$$

ここで，$count_v(j,e)$ は対訳句 (j,e) を抽出した回数であり，$count_v(j)$ は日本語句 j を抽出した回数である．確率 $P_v(e|j)$ の対数をとり，対訳句抽出における日英方向の対数翻訳確率とする．同様に，対訳句抽出における英日方向の対数翻訳確率も求める．対訳句抽出における日英方向の対数翻訳確率の例を表 3.2 に示す．

表 3.2 対訳句抽出における日英方向の対数翻訳確率の例

日本語句	英語句	$count_v(j)$	$count_v(j,e)$	$\log(P_v(e\|j))$
生まれ ました	was born	2233	104	-1.332
生活 の 複雑 さ	complexity of human life	2718	124	-1.34
正比例 する	are directly proportional	7761	20	-2.589

3.3 確率の付与

従来手法は対訳句の確率として対数フレーズ確率を用いている．一方，本研究では対訳句の確率としてロジスティック回帰分析から得た確率を用いる．

まず，自動抽出した対訳句の従属変数および独立変数を用いてモデルの学習を行い，線形回帰モデル（式 (3.4)）の定数および回帰係数を推定する．次に，得られた定数および

回帰係数を式 (3.3) に代入し，本研究におけるロジスティック回帰モデルを作成する．そして，自動抽出した対訳句に付与された各独立変数をロジスティック回帰モデルに代入し，確率を得る．この確率をロジスティック回帰分析から得た確率と呼ぶ．なお，対数フレーズ確率と同様に，ロジスティック回帰分析から得た確率についても対数をとる．ロジスティック回帰分析の従属変数の設定において，人手で作成した対訳句と一致する対訳句は適切な対応をとる（Y=1）と仮定しているため，ロジスティック回帰分析から得た確率は対訳句が適切な対応をとる確率を表している．最後に，従来手法における対数フレーズ確率をロジスティック回帰分析から得た確率に置き換え，パターンに基づく日英統計翻訳を行う．

第4章　実験

4.1　実験データ

対訳学習文および翻訳実験に用いるテスト文は電子辞書から抽出した単文データベースを用いる [11]．なお，単文データは日本語文が単文であるが，英語文は単文とは限らず，重文・複文が含まれる．前処理として日本語学習文に対して形態素解析エンジン MeCab[12] を用いて分かち書きを行う．また，英語学習文に対して tokenizer.perl[6] を用いて分かち書きを行う．対訳学習文および翻訳実験に用いるテスト文の例を表 4.1 に，コーパスの内訳を表 4.2 に示す．

表 4.1　対訳学習文および翻訳実験に用いるテスト文の例

日本語文	ナンシー と テニス を し た 。
英語文	I played tennis with Nancy .
日本語文	ぼく は バス の 中 で 先生 に 会っ た 。
英語文	I saw our teacher on the bus .
日本語文	彼女 は テスト で 良い 点 を とっ た 。
英語文	She got a good score on the test .

表 4.2　コーパスの内訳

対訳学習文	100,000 文対
テスト文	100 文

人手で作成した対訳句には鳥バンク [13] の対訳句を用いる．鳥バンクは自然言語処理のための言語知識ベースを収録したデータバンクであり，日本語の重文と複文を対象とする “意味類型パターン辞書”が収録されている．本研究では鳥バンクから抽出した対訳句 329,545 句対を用いる．鳥バンクから抽出した対訳句の例を表 4.3 に示す．

表 4.3 鳥バンクから抽出した対訳句の例

日本語句	ある プログラム
英語句	a program
日本語句	とても 有効 な
英語句	very useful
日本語句	家族 から
英語句	from the family
日本語句	英語 の 勉強 を し
英語句	study English

4.2 分析実験

統計ソフト R を用いてロジスティック回帰分析を行う.

4.2.1 予備実験

対訳句の特性を理解するため，予備実験を行う．予備実験の結果，不適切な対応をとる対訳句において高い確率が付与される場合があることがわかった．例を表 4.4 に示す.

表 4.4 日本語学習文における日本語句の出現回数が 1 回の例

日本語句 J	英語句 E	$count_s(J)$	$count_s(E\|J)$	$\log P_s(E\|J)$
部員 たち	members	1	1	0
部員 たち	were	1	1	0
部員 たち	The	1	1	0

ここで，$count_s(J)$ は日本語学習文における日本語句 J の出現回数であり，$count_s(E|J)$ は対訳学習文における対訳句 (J,E) の共起回数，$\log P_s(E|J)$ は対訳学習文における日英方向の対数翻訳確率である．例えば日本語句 “部員たち”の出現回数が 1 回の場合，その日本語句をもつすべての対訳句において，対訳学習文における日英方向の対数翻訳確率は 0（最大値）になる．したがって句の出現回数が 1 回の場合，不適切な対応をとる対訳句においても高い確率が付与される．以上のことから，本研究では句の抽出回数または学習文における句の出現回数が少なくとも一方の言語で 1 回である対訳句を取り除き，モデルの作成を行う.

4.2.2 モデルの作成結果

6 つの確率を独立変数に用い，ロジスティック回帰分析を行う．予備実験の結果より，句の抽出回数または学習文における句の出現回数が少なくとも一方の言語で 1 回である対訳句を取り除き，モデルの学習に用いる．

鳥バンクの対訳句と自動抽出した対訳句を照合した結果，鳥バンクと一致した自動抽出した対訳句は 45,161 句対であった．また，鳥バンクの対訳句と一致しなかった自動抽出した対訳句からも一致した対訳句と同数の対訳句を無作為抽出し，合計 90,322 句対の自動抽出した対訳句をモデルの学習に用いる．本研究は統計ソフト R を用いてモデルの学習を行う．R より得られた出力を以下に示す．

統計ソフト R の出力結果

```
Call:
glm(formula = Object_Func ~ Prob_JP + Prob_EN + Sent_JE + Sent_EJ +
    Var_JE + Var_EJ, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5548  -0.7102   0.1224   0.6379   3.8446

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.3368508  0.0311004 107.293  < 2e-16 ***
Prob_JP      0.0163943  0.0001599 102.514  < 2e-16 ***
Prob_EN      0.0068223  0.0001403  48.637  < 2e-16 ***
Sent_JE      0.1235362  0.0073223  16.871  < 2e-16 ***
Sent_EJ     -0.2302809  0.0073989 -31.124  < 2e-16 ***
Var_JE      -0.0170917  0.0062589  -2.731  0.00632 **
Var_EJ       0.2211758  0.0063798  34.668  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125213  on 90321  degrees of freedom
Residual deviance:  82244  on 90315  degrees of freedom
AIC: 82258

Number of Fisher Scoring iterations: 5
```

ここで，Prob_JP は日英方向の対数フレーズ確率を，Prob_EN は英日方向の対数フレーズ確率，Sent_JE は対訳学習文における日英方向の対数翻訳確率，Sent_EJ は対訳学習

文における英日方向の対数翻訳確率，Var_JE は対訳句抽出における日英方向の対数翻訳確率，Var_EJ は対訳句抽出における英日方向の対数翻訳確率を表す．各独立変数において，対訳句抽出における日英方向の対数翻訳確率の P 値（Pr）はやや高いものの，各 P 値より，全ての独立変数が有意であることがわかる．

モデルの学習より求めた回帰係数（Estimate）を用いた線形回帰モデルを式 (4.1) に示す．

$$\begin{aligned}\log \frac{p}{1-p} &= 3.33685 + 0.01639x_1 + 0.00682x_2 \\ &\quad + 0.12354x_3 - 0.23028x_4 - 0.01709x_5 + 0.22118x_6\end{aligned} \tag{4.1}$$

モデルの学習より求めた回帰係数を式 (3.3) に代入し，本研究におけるロジスティック回帰モデルを作成する．そして，ロジスティック回帰モデルより，各自動抽出した対訳句に確率を付与する．なお，ロジスティック回帰分析から得た確率においても対数をとる．ロジスティック回帰分析から得た確率の例を表 4.5 に示す．本研究では従来手法における対数フレーズ確率をロジスティック回帰分析から得た確率に置き換え，パターンに基づく日英統計翻訳を行う．

表 4.5 ロジスティック回帰分析から得た確率の例

日本語句	英語句	ロジスティック回帰分析から得た確率（対数）
英語	English	-0.116
あの 人	That person	-0.129
室温 を 調節	device	-3.123
この 切符 で	will admit	-5.478

4.3 翻訳実験

本研究はパターンに基づく日英統計翻訳を行う．

4.3.1 実験条件

各ステップにおける実験条件を以下に示す．

ステップ 2 単語に基づく対訳文パターンの作成

- 単語に基づく対訳文パターンの出力を抑制するため，単語に基づく対訳文パターンの作成に用いる対訳単語の単語翻訳確率の閾値を 0.1 とする．

ステップ 4 句に基づく対訳文パターンの作成

- 句に基づく対訳文パターンの出力を抑制するため，句に基づく対訳文パターンの作成に使用する対訳句に以下の条件を用いる．
  - 日本語句の単語数を基準とし，対応する英語句の単語数との差が ±7 単語以内の対訳句を用いる．
  - 日本語句が同一の対訳句は対数フレーズ確率の上位から最大 8 句までとする．
  - 対訳句の総数はロジスティック回帰分析から得た確率の上位から最大 2,000,000 句までとする．

ステップ 5 翻訳

- 翻訳に使用する対訳句に以下の条件を用いる．
  - ロジスティック回帰分析から得た確率の閾値を-10000.0 とする．
  - 日本語句の単語数を基準とし，対応する英語句の単語数との差が ±7 単語以内の対訳句を用いる．
  - 日本語句が同一の対訳句はロジスティック回帰分析から得た確率の上位から最大 64 句までとする．
- 翻訳に使用する句に基づく対訳文パターンに以下の条件を用いる．
  - 入力文と日本語文パターンの字面を比較し，字面が多く一致する対訳文パターンを優先して選択する．
  - 対数文パターン確率の閾値を-1000.0 とする．
  - 日本語文パターンの単語数を基準とし，対応する英語文パターンの単語数との差が ±7 単語以内の対訳文パターンを用いる．

– 日本語文パターンが同一の対訳文パターンは対数文パターン確率の上位から最大 64 句までとする．
– 入力文 1 文に対し，用いる対訳文パターンの数は最大 10,000 までとする．

4.3.2 翻訳実験の結果

各翻訳候補文において，ロジスティック回帰分析から得た確率と対数文パターン確率，N-gram モデルの総和を求め，翻訳候補文の中で総和が最大となる文を翻訳文として出力する．翻訳実験により得た単語数，対訳句の数，対訳文パターン数，翻訳文数を表 4.6 に示す．また翻訳文の例を表 4.7 に示す．

表 4.6 提案手法より得たデータ数

対訳単語	28,672 単語対
単語に基づく対訳文パターン	1,412,913 パターン対
対訳句	229,874,598 句対
句に基づく対訳文パターン	341,648,816 パターン対
翻訳文	78 文

表 4.7 翻訳文の例

入力文	彼女 は 彼 に 失望 し て い た 。
参照文	She was disappointed in him .
翻訳文	She was disappointed in him .
入力文	子供 たち は 寝室 へ 立ち去っ た 。
参照文	The children disappeared to their bedrooms .
翻訳文	The children were left in the bedroom .
入力文	新車 の 性能 の 試験 を し た 。
参照文	The performance test of new cars was held .
翻訳文	The performance test of the new car was in .

第5章　評価

5.1　従来手法と提案手法の対比較評価

提案手法と江木らの従来手法の人手による対比較評価を行った．従来手法は対訳句の確率として対数フレーズ確率を用いる．なお，その他の条件については提案手法と同様である．また，提案手法で出力した翻訳文 78 文中，従来手法でも翻訳文が得られた 77 文に対して評価を行った．提案手法と従来手法との対比較評価結果を表 5.1 に示す．

表 5.1　従来手法と提案手法の対比較評価結果

提案手法○	提案手法×	差なし	同一出力
7	3	59	8

提案手法○の例を表 5.2 に，提案手法×の例を表 5.3 に示す．

表 5.2　従来手法と提案手法の対比較評価：提案手法○の例

入力文	もっと 右 へ 寄っ て ください 。
参照文	Please move over more to the right .
提案手法	Move a little more to the right .
従来手法	Please come to discuss .
入力文	水 が 腐っ て いる 。
参照文	The water is foul .
提案手法	The water is rotten .
従来手法	The water is at the right .
入力文	その 計画 は 成功 の 見込み が 十分 ある 。
参照文	The plan bids fair to succeed .
提案手法	The project has a chance of success .
従来手法	He has a good chance for success .

表 5.3 従来手法と提案手法の対比較評価：提案手法×の例

入力文	金槌 と 同様 な もの が ある 。
参照文	There is something similar to a hammer .
提案手法	There's something a view similar to a hammer .
従来手法	There is something like a hammer .
入力文	彼 は たった １０００ 円 しか 持っ て い ない 。
参照文	He has only 1,000 yen on him .
提案手法	I have no more than one thousand yen .
従来手法	He has no more than one thousand yen .
入力文	その ビル は 倒壊 の 危険 が ある 。
参照文	The building is in danger of collapsing .
提案手法	The building is a danger of this down .
従来手法	The building is a danger of collapse .

表 5.1 の結果より，従来手法と比較して，提案手法が優れていることがわかる．よって，本研究で用いた入力文における提案手法の有効性が確認された．

5.2 句に基づく統計翻訳と提案手法の対比較評価

提案手法と句に基づく統計翻訳の人手による対比較評価を行った．句に基づく統計翻訳のデコーダには Moses を用いる．提案手法ではロジスティック回帰分析において人手で作成した対訳句の情報が追加されている．そこで，提案手法と句に基づく統計翻訳の実験データを平等にするため，句に基づく統計翻訳においても人手で作成した対訳句の情報を追加し翻訳を行う必要がある．ここで，句に基づく日英統計翻訳において対訳句を対訳学習文に追加し，翻訳モデルの作成を行うことで翻訳精度が向上したと報告されている [14]．よって本評価に用いる句に基づく統計翻訳においても，翻訳モデルの作成において鳥バンクから抽出した対訳句 329,545 句対を対訳学習文に追加し，翻訳モデルを作成する．なお，言語モデルの作成については 3-gram モデルを用いる．

本評価は提案手法で得られた翻訳文 78 に対して評価を行った．提案手法と句に基づく統計翻訳との対比較評価結果を表 5.4 に示す．

表 5.4 句に基づく統計翻訳と提案手法の対比較評価結果

提案手法○	提案手法×	差なし	同一出力
9	9	57	3

提案手法○の例を表 5.5 に，提案手法×の例を表 5.6 に示す.

表 5.5 句に基づく統計翻訳と提案手法の対比較評価：提案手法○の例

入力文	彼女 に は 文学 の 素養 が ある 。
参照文	She has learned a good deal of literature .
提案手法	She is versed in literature .
統計翻訳	She literary culture.
入力文	その 都市 の 大半 が 焼失 し た 。
参照文	Most of the city was burned into cinders .
提案手法	The better part of the city was burnt down .
統計翻訳	Most of the city .
入力文	彼ら は 彼ら 自身 の 力 で 家 を 建て た 。
参照文	They built the house for themselves .
提案手法	They built a house of their own .
統計翻訳	They their own strength built a house .

表 5.6 句に基づく統計翻訳と提案手法の対比較評価：提案手法×の例

入力文	４ 人 が 負傷 し た 。
参照文	Four people were wounded .
提案手法	I was injured in the four children .
統計翻訳	four were injured .
入力文	関税 は 完全 に 撤廃 さ れ た 。
参照文	Tariffs have been eliminated altogether .
提案手法	Agreement was that to the hilt .
統計翻訳	The customs were completely abolished .
入力文	この 下水 は よく 通る 。
参照文	The sewer runs well .
提案手法	This drain is heavy .
統計翻訳	The sewage carries well .

表 5.4 の結果より，本研究で用いた入力文において，句に基づく統計翻訳と提案手法は翻訳精度に差がないことがわかった.

第6章　考察

6.1　翻訳実験の考察

6.1.1　誤り解析

江木らの従来手法との対比較評価より，提案手法×の翻訳文を 3 文得た．この 3 文について，誤り解析を行った．以下に原因を述べる．

6.1.1.1　指示詞などの字面が残る対訳文パターン

対訳文パターンにおいて，指示詞や代名詞などの字面が残るため正しく翻訳できない問題がある．例を表 6.1 に示す．

表 6.1　指示詞などの字面が残る対訳文パターンを含む翻訳文の例

入力文	その ビル は 倒壊 の 危険 が ある 。
参照文	The building is in danger of collapsing .
翻訳文	The building is a danger of this down .
日本語文パターン	その ビル は $X1$ の $X2$ ある 。
英語文パターン	The building is $X2$ of this $X1$.
対訳句	$X1$: 倒壊　　down
	$X2$: 危険 が　a danger

表 6.1 の英語文パターンには指示詞 “this”が残る．対訳文パターンに指示詞などの字面が残る場合，適切な対応をとる対訳句を変数部に用いても正しく翻訳できない．以上のことから，指示詞などの字面が残る対訳文パターンの問題に対して，対訳文パターンの選択方法の再検討が必要であると考える．

6.1.1.2　不適切な対応をとる対訳句

提案手法により対訳句の対応が改善され，翻訳精度が向上した．しかし依然として，不適切な対応をとる対訳句を含む翻訳文が残る．不適切な対応をとる対訳句を含む翻訳文の例を表 6.2 に示す．

表 6.2　不適切な対応をとる対訳句を含む翻訳文の例

入力文	彼 は たった １０００ 円 しか 持っ て い ない 。
参照文	He has only 1,000 yen on him .
翻訳文	I have no more than one thousand yen .
日本語文パターン	$X1$ は たった $X2$ 円 しか 持っ て い ない 。
英語文パターン	$X1$ have no more than one $X2$ yen .
対訳句	$X1$: 彼　　I
	$X2$: １０００　thousand

表 6.2 において，翻訳文で用いた対訳句の中に不適切な対応をとる対訳句 “彼” と “I” がある．この入力文において，翻訳候補文を調査したところ，“彼” と “He” の対訳句を含んだ翻訳候補文 “He has no more than one thousand yen .” が存在した．以上のことから，不適切な対応をとる対訳句を含む翻訳文を出力する問題は，翻訳文の選択方法の再検討により改善すると考える．

6.1.2　対訳句の精度調査

翻訳精度が向上した原因を調べるため，対訳句の精度調査を行う．提案手法で用いた対訳句と江木らの従来手法で用いた対訳句を比較するため，各手法に対し，翻訳文の生成に用いた全対訳句における，適切な対応をとる対訳句の割合を調査した．調査結果を表 6.3 に示す．適切な対応をとる対訳句の例を表 6.4 に，不適切な対応をとる対訳句の例を表 6.5 に示す．

表 6.3　適切な対応をとる対訳句の割合

手法	翻訳文生成に用いた全対訳句数	適切な対応をとる対訳句数	割合
提案手法	233	104	0.45
従来手法	260	113	0.43

表 6.4 適切な対応をとる対訳句の例

日本語句	英語句	ロジスティック回帰分析から得た確率（対数）
山	the mountain	-0.245
合格 点	passing mark	-0.226
彼ら 自身 の	their own	-0.208

表 6.5 不適切な対応をとる対訳句の例

日本語句	英語句	ロジスティック回帰分析から得た確率（対数）
すぎる	is	-0.118
コロンブス	the	-0.695
手続き は	for	-0.242

表 6.3 の結果より，翻訳文の生成に用いた全対訳句における，適切な対応をとる対訳句の割合は提案手法と従来手法で差が見られなかった．また，表 6.5 より，“すぎる” と “is” や “手続き は” と “for” などの不適切な対応をとる対訳句においても，ロジスティック回帰分析から得た確率は高い値であることがわかる．よって，対訳句の精度調査からは翻訳精度の向上原因は解明できなかった．

6.1.3 翻訳精度向上の原因調査

対訳句の精度調査からは翻訳精度が向上した原因が解明できなかった．そこで，提案手法と従来手法の対比較評価において，提案手法○であった 7 文に対して提案手法の翻訳精度が向上した原因について調査を行った．翻訳精度が向上した原因を以下に述べる．

6.1.3.1 対訳句の改善

従来手法と比較して対訳句の対応が改善したため，翻訳精度が向上した翻訳文が 7 文中 3 文あった．例を表 6.6 に示す．

表 6.6 対訳句改善の例

入力文	水 が 腐っ て いる 。
参照文	The water is foul .
提案手法	
翻訳文	The water is rotten .
日本語文パターン	水 が $X1$ て いる 。
英語文パターン	The water is $X1$.
対訳句	$X1$: 腐っ　　rotten
従来手法	
翻訳文	The water is at the right .
日本語文パターン	水 が $X1$ て いる 。
英語文パターン	The water is at the $X1$.
対訳句	$X1$: 腐っ　　right

従来手法では日本語句 “腐っ” に対して英語句 “right” が対応している．一方，提案手法では日本語句 “腐っ” に対して英語句 “rotten” が対応しており，対訳句の対応が改善したといえる．

6.1.3.2 対訳文パターンの改善

従来手法と比較して選択された対訳文パターンが改善したため，翻訳精度が向上した翻訳文が 7 文中 3 文あった．例を表 6.7 に示す．

表 6.7 対訳文パターン改善の例

入力文	彼女 に は 文学 の 素養 が ある 。
参照文	She has learned a good deal of literature .
提案手法	
翻訳文	She is versed in literature .
日本語文パターン	彼女 に は $X1$ の 素養 が ある 。
英語文パターン	She is versed in $X1$.
対訳句	$X1$: 文学　　literature
従来手法	
翻訳文	She is versed in Japanese literature .
日本語文パターン	彼女 $X1$ $X2$ の 素養 が ある 。
英語文パターン	She is versed $X1$ Japanese $X2$.
対訳句	$X1$: に　　in
	$X2$: は 文学　　literature

従来手法で選択された英語文パターンには不要な字面 “Japanese”が残っているため，正しい翻訳文が得られない．一方，提案手法で選択された対訳文パターンは字面の対応が適切であるため，正しい翻訳文が得られている．よって，選択された対訳文パターンが改善したといえる．

6.1.4 重みの最適化

翻訳文の選択において，ロジスティック回帰分析から得た確率と対数文パターン確率，N-gram モデルの 3 つの確率を利用した．これらの確率は値の大きさが異なるため，ヒューリスティックにより各確率に重みを付与している．これらの重みを最適化することでさらなる翻訳精度の向上が期待できる．なお，従来手法においても同様の効果が期待できる．

6.2 分析実験の考察

6.2.1 回帰係数の調査

分析実験の結果（4.2.2 節）より，ほとんどの独立変数は有意水準 0.1%で有意であることがわかった．しかし，逸脱度（deviance）や赤池情報量基準（AIC）[15] は値が大きいことがわかる．よって，モデルの適合度が悪い可能性がある．今後，これらの値が小さくなるような独立変数の設定を行う必要があると考える．

また，本研究では回帰分析の従属変数として質的な値（0 または 1）を設定したため，ロジスティック回帰分析による分析実験を行った．しかし，従属変数として量的な値（例えば対訳句の出現回数）を設定することで重回帰分析など他の回帰分析手法による分析実験を行うことができる．

6.2.2 モデル作成に用いた対訳句の調査

モデルの作成に用いた対訳句のうち，人手で作成した対訳句と一致しなかった対訳句の精度調査を行った．人手で作成した対訳句と一致しなかった対訳句から 100 句対を無作為抽出し，適切な対応をとる対訳句であるか調査を行った．調査の結果，100 句中 4 句が適切な対応をとる対訳句であった．人手で作成した対訳句と一致しなかった対訳句のうち，適切な対応をとる対訳句を表 6.8 に示す．

表 6.8 人手で作成した対訳句と一致しなかった対訳句のうち適切な対応をとる対訳句

日本語句	英語句	ロジスティック回帰分析から得た確率（対数）
手 を 繋い	joined hands	-0.276
金	My money	-0.423
１０ 時 に	at 10	-0.244
日常	our daily lives	-0.944

ロジスティック回帰分析を用いたモデルの作成では，人手で作成した対訳句は適切な対応をとる対訳句であると仮定し，人手で作成した対訳句と一致した対訳句に従属変数を付与している．しかし，人手で作成した対訳句と一致しない対訳句においても適切な対応をとる対訳句は存在する．よって，本研究で設定した従属変数では，適切な対応をとる対訳句かどうかを明確に判定することは困難であると考える．今後，従属変数や独立変数の設定を変更し実験を行う必要があると考える．

第7章　おわりに

本研究では，“パターンに基づく統計翻訳”の翻訳文の選択において，不適切な対応をとる対訳句を含む翻訳文の出力を抑制するため，人手で作成した対訳句を利用してロジスティック回帰分析を行うことを提案した．そして，パターンに基づく日英統計翻訳を行い，翻訳精度を調査した．従来手法との対比較評価の結果は翻訳文77文において提案手法○が7文，提案手法×が3文であった．以上より，提案手法の有効性が確認できた．

しかし，句に基づく統計翻訳との対比較評価において，提案手法○と提案手法×はともに9文であった．よって，句に基づく統計翻訳と提案手法は翻訳精度に差がないことがわかった．

また従来手法との対比較評価より，提案手法×であった3文において誤り解析を行った．誤り解析の結果，対訳文パターンの選択方法や翻訳文の選択方法，翻訳文選択に用いる重みについて再検討が必要であることがわかった．さらに，翻訳精度の向上原因を探るため，翻訳文の生成に用いた対訳句の精度調査を行った．しかし，従来手法と提案手法で対訳句の精度に差は見られなかった．よって，翻訳実験を追加することにより評価文数を増やし，より正確な評価を行う必要があると考える．

本研究は回帰分析の従属変数として質的な値（0または1）を設定したため，ロジスティック回帰分析による分析実験を行った．しかし，従属変数として量的な値を設定することで重回帰分析など他の回帰分析手法による分析実験を行うことができる．また，不適切な対応をとる対訳句が翻訳文に含まれることを抑制するための方法として，対訳句の作成において，不適切な対応をとる対訳句の作成を抑制する方法も考えられる．今後これらを検討し，さらなる翻訳精度向上を目指す．

謝辞

最後に，二年間に渡り，本研究の御指導をいただきました鳥取大学工学部知能情報工学科計算機講座 C 研究室の村上仁一准教授，村田真樹教授，徳久雅人講師に深く感謝するとともに厚くお礼を申し上げます．また，ご多忙の中，助言を頂きました岩井儀雄教授に厚くお礼申し上げます．そして，計算機工学講座 C 研究室の皆様，参考にさせて頂いた論文の著者の方々に対して，深く感謝します．

参考文献

[1] 石上真理子, 水田理夫, 徳久雅人, 村上仁一, 池原悟, “関数・符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 回年次大会予稿集, pp.67-70, 1997.

[2] 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.

[3] 江木孝史, 村上仁一, 徳久雅人, “句に基づく対訳文パターンの自動作成と統計的手法を用いた英日パターン翻訳”, 言語処理学会第 20 回年次大会予稿集, pp.951-954, 2014.

[4] Peter F.Brown, Stephen A.Della Pietra, Vincent J.Della Pietra, Robert L.Mercer, “The mathematics of statistical machine translation:Parameter Estimation”, Computational Linguistics, 19(2), pp.263-311, 1993.

[5] Franz Josef Och, Hermann Ney, “A Systematic Comparison of Various Statistical Alignment Models”, Computational Linguistics, 29(1), pp.19-51, 2003.

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

[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階翻訳方式”, 情報処理学会論文誌, 28(12), pp.1269-1279, 1987.

[8] 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, 2006-NL-176(2), pp.9-16, 2006.

[9] 中村永友, “R で学ぶデータサイエンス 2　多次元データ解析法”, 共立出版, pp.79-90, 2009.

[10] Ross Ihaka, Robert Gentleman, “R: A Language for Data Analysis and Graphics”, Journal of Computational and Graphical Statistics, 5(3), pp.299-314, 1996.

[11] 村上仁一, 藤波進, “日本語と英語の対訳文対の収集と著作権の考察”, 第一回コーパス日本語学ワークショップ予稿集, pp.119-130, 2012.

[12] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis”, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , pp.230-237, 2004.

[13] 鳥バンク
http://unicorn.ike.tottori-u.ac.jp/toribank/

[14] 日野聡子, 村上仁一, 徳久雅人, 村田真樹, “日英統計翻訳における対訳句コーパスの効果”, 言語処理学会第 19 回年次大会予稿集, pp.572-575, 2013.

[15] Hirotogu Akaike, “Information theory and an extension of the maximum likelihood principle”, Proceedings of the 2nd International Symposium on Information Theory, pp267-281, 1973.