概要

現在，パターン翻訳 [1] や単語に基づく統計翻訳・句に基づく統計翻訳が提案されている．パターン翻訳は人手により作成した対訳句辞書と対訳文パターン辞書を用いて翻訳を行う方法である．翻訳精度は高い方法であるが対訳句辞書と対訳文パターン辞書の作成を人手で行うため，開発にコストがかかる．この問題を解決するために江木らは，GIZA++[2] を利用したパターンに基づく統計翻訳 (Pattern　Based SMT)[3] を提案した．パターンに基づく統計翻訳では対訳句辞書と対訳文パターン辞書を自動的に作成することで開発コストの解消を試みた．しかし，まだまだ翻訳精度は低い方法である．精度の低い原因の一つが翻訳に用いる対訳句辞書と句レベル文パターン辞書の作成の起点である対訳単語辞書の精度が低いことが考えられる．そこで本研究では，対訳単語辞書の精度調査をおこなった．対訳学習文と GIZA++を利用して対訳単語を作成し，全対訳単語から枝刈りを行い，対訳単語辞書を作成している．対訳単語において精度調査を行い，調査結果から対訳単語辞書を作成する時の枝刈り条件を変更し，対訳単語辞書の変更を行う．その後変更前後の対訳単語辞書の精度を調査する．実験の結果，対訳単語ではひらがな 1 文字・数字・記号において不適切な対訳単語が多く作成されていた．よってひらがな 1 文字・数字・記号を削除する条件を枝刈り条件に加えて対訳単語辞書を変更した．変更前後の対訳単語辞書の評価結果では対訳単語数と精度に差はなかった．本研究論文第 2 章は西尾 [4] の研究論文を抜粋・参照し，一部説明を加えて作成した．

目 次

図 目 次

表 目 次

第1章　はじめに

現在翻訳には様々な手法が利用されている．パターン翻訳 [1] や単語に基づく統計翻訳・句に基づく統計翻訳が提案されている．パターン翻訳は人手で作成した対訳句辞書と対訳文パターン辞書を用いた翻訳であるためコストがかかる問題があった．単語に基づく統計翻訳は翻訳精度が低い問題があった．句に基づく統計翻訳では単語に基づく統計翻訳と比べて翻訳精度は高く翻訳コストも低い方法である．

一方，江木らはパターン翻訳のコストに関する問題を解消するためパターンに基づく統計翻訳 (Pattern Based SMT)[3] を提案した．パターンに基づく統計翻訳ではコスト解消のために自動的に対訳句辞書と対訳文パターン辞書を作成し，翻訳に用いる方法である．しかしパターンベース統計翻訳の翻訳精度はまだまだ低い．そしてパターンベース統計翻訳は対訳単語を起点として対訳文から対訳句辞書と対訳文パターン辞書を作成して翻訳を行う．翻訳精度が低い原因としてこの対訳単語の精度が低いことが問題と考える．対訳単語辞書の精度が低いため精度の低い対訳句辞書と対訳文パターン辞書が翻訳に用いられている可能性がある．そこで本研究では対訳単語辞書の精度を調査した．対訳学習文と GIZA++を用いて対訳単語を作成し，全対訳単語から枝刈りを行い，対訳単語辞書を作成する．本研究では対訳単語辞書作成に用いる対訳単語の精度の調査を行う．

本論文の構成は以下の通りである．第 2 章で従来の研究について説明し，第 3 章で提案する手法について説明する．第 4 章で実験データ，実験結果と評価を示す．

第2章　従来の研究

2.1　パターン翻訳 [1]

2.1.1　概要

パターン翻訳 [1] とは，機械翻訳手法の一種である．パターン翻訳は，原言語文と目的言語文の対訳文に対して，任意の単語やフレーズを変数化した“ 対訳文パターン ”と“ 対訳フレーズ ”が必要である．原言語入力文と原言語文パターンを照合し，適合する原言語文パターンに対応する目的言語文パターンを得る．そして，文パターンの変数部に対応する単語やフレーズを，対訳フレーズを挿入し文生成を行い，目的言語翻訳文を出力する．

パターン翻訳は適切な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出力文を得ることができる．しかし，一般的なパターン翻訳は対訳文パターンを人手で作成するため開発にコストがかかる．また，対訳文パターンに適合しない場合は翻訳ができないため，問題点として，入力文に対するカバー率が低い．

2.1.2 日英パターン翻訳の手順

手順 1 対訳文パターンと対訳フレーズを用意する．対訳文パターンとは，大量の対訳文から任意の単語やフレーズを変数化して得られる．対訳フレーズとは，対訳言語において，同じ意味を有する単語のまとまりの対である．日英対訳文パターンの例を表 2.1 に，日英対訳フレーズの例を表 2.2 に示す．

表 2.1: 対訳文パターンの例

日本語原文	私 は 海 に 行く 。
英語原文	I go to the sea .
日本語文パターン	私 は $X00$ に 行く 。
英語文パターン	I go to $X00$.

表 2.2: 対訳フレーズの例

日本語フレーズ	英語フレーズ
田園 生活	country life
子供 たち	The children's
下水 管	sewage pipe

手順 2 日本語入力文と日本語文パターンを照合する．

手順 3 変数部に対応する日本語単語を対訳フレーズを用いて英語単語に翻訳する．

手順 4 日本語文パターンに対応する英語文パターンの変数部を，翻訳した英語単語に置き換える．

手順 5 手順 4 で生成した英語文を出力する．

2.2 日英統計翻訳

2.2.1 概要

統計翻訳とは，機械翻訳手法の一種である．原言語と目的言語の対訳文を大量に収集した対訳文より，自動的に翻訳規則を獲得し翻訳を行う．

統計翻訳には単語に基づく統計翻訳と句に基づく統計翻訳があり，初期の統計翻訳では単語に基づく統計翻訳が用いられていたが，翻訳精度は高くなかった．しかし近年，句に基づく統計翻訳 [9] が提案され，語順の並び替えや文脈における訳語の選択や翻訳精度において，単語に基づく統計翻訳に比べて優れている．このため現在は句に基づく統計翻訳が主流となっている．

2.2.2 言語モデル

言語モデルは，単語列の生成確率を付与するモデルである．日英翻訳では，翻訳モデルを用いて生成された翻訳候補から，英語として自然な文を選出するために用いる．統計翻訳では一般的に，N-gram モデルを用いる．

2.2.3 N-gram モデル

N-gram モデルとは “単語列 $P(W_1^n) = w_1^n = w_1, w_2, w_3, ...w_n$ の i 番目の単語 w_i の生起確率 $P(w_i)$ は直前の $(N-1)$ の単語列 $w_{i-(N-1)}, w_{i-(N-2)}, w_{i-(N-3)}, ...w_{i-1}$ に依存する” という仮説に基づくモデルである．計算式を以下に示す．

$$
\begin{aligned}
P(W_1^n) &= P(w_1) \times P(w_2|w_1) \times P(w_3|w_1^2)...P(w_n|w_1^{n-1}) && (2.1)\\
&\approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_1^2)...P(w_n|w_{n-(N-1)}^{n-1}) && (2.2)\\
&= \prod_{i=1}^{n} P(w_i|w_{i-(N-1)}^{i-1}) && (2.3)
\end{aligned}
$$

また，$P(w_i|w_{n-(N-1)}^{i-1})$ は以下の式で計算される．ここで $C(w_1^i)$ は単語列 w_1^i が出現する頻度を表す．

$$
P(w_i|w_{i-(N-1)}^{i-1}) = \frac{C(w_{i-(N-1)}^{i})}{C(w_{i-(N-1)}^{i-1})} \tag{2.4}
$$

2.2.4 単語に基づく統計翻訳

単語に基づく統計翻訳は単語対応の翻訳モデルを用いている．例として，ある日本語文を英語文に翻訳する場合を考える．日本語単語を英語に翻訳し，日本語単語の語順と同じ並びで英単語を並べて翻訳する．単語に基づく統計翻訳は単語対応の確率を得る IBM 翻訳モデルが用いられている．

2.2.5 単語に基づく統計翻訳の問題点

以下に，IBM 翻訳モデルを用いて得た英日方向における単語対応の例と，日英方向における単語対応の例を示す．また，●は単語が対応した箇所を示す．

表 2.3: 英日方向の単語対応

	She	went	to	Tokyo	on	travel
彼女	●					
は						
旅行						●
で					●	
東京				●		
に						
行っ		●				
た						

表 2.4: 日英方向の単語対応

	She	went	to	Tokyo	on	travel
彼女	●					
は			●			
旅行						●
で					●	
東京				●		
に		●			●	
行っ		●				
た		●				

表 2.3 は日本語単語 “は” と “に” と “た” に対応する英単語が存在しない．一方で，表 2.4 は全ての単語に対して対応がとれている．単語に基づく統計翻訳は対応する単語が存在しない場合，何も無い状態から単語の発生確率を計算する．このため単語翻訳確率の

信頼性が問題となっている．よって現在句に基づく統計翻訳が行われている．

2.2.6 IBM 翻訳モデル

IBM 翻訳モデルを以下に示す．これは，力久ら [5] の抜粋である．統計翻訳の代表的なモデルとして，IBM の Brown らによる仏英翻訳モデル [10] がある．IMB 翻訳モデルは，単語に基づく統計翻訳を想定して作成された，単語対応の確率モデルである．この翻訳モデルは順に複雑な計算を行うモデル 1 から 5 の 5 つのモデルで構成される．各モデルの概要を以下に示す.

model1 目的言語のある単語が原言語の単語に訳される確率を用いる

model2 model1 に加えて，目的言語のある単語に対応する原言語の単語の原言語文中での位置の確率（以下，permutation 確率と呼ぶ）を用いる（絶対位置）

model3 model2 に加えて，目的言語のある単語が原言語の何単語に対応するかの確率を用いる

model4 model3 の permutation 確率を改良（相対位置）

model5 model4 の permutation 確率を更に改良

本章では．原言語であるフランス語文を F，目的言語である英語文を E として定義する．

IBM モデルでは，フランス語文 E，英語文 F の翻訳モデル $P(F|E)$ を計算するために，アライメント a を用いる．以下に IBM モデルの基本式を示す.

$$P(F|E) = \sum_{a} P(F, a|E) \tag{2.5}$$

アライメントとは仏単語と英単語の対応を意味している．IBM モデルのアライメントでは，各仏単語 f に対応する英単語 e は 1 つあり，各英単語 e に対応する仏単語は 0 から n 個ある．また仏単語 f において適切な英単語と対応しない場合，英語文の先頭に空単語 e_0 があると仮定し，その仏単語 f と空単語 e_0 を対応づける．

model1

(2.5) 式は以下の式に分解することができる．m はフランス語文の長さ，a_1^{j-1} はフランス語文における，1 番目から $j-1$ 番目までのアライメント，f_1^{j-1} はフランス語文におけ

る，1 番目から $j-1$ 番目まで単語を表している．

$$P(F,a|E) = P(m|E)\prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, E)P(f_j|a_1^{j}, f_1^{j-1}, m, E) \tag{2.6}$$

(2.6) 式で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．

- フランス語文の長さの確率 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

$$P(a_j|a_1^{j-1}, f_1^{j-1}, m, E) = (l+1)^{-1}$$

- フランス語の翻訳確率 $t(f_j|e_{a_j})$ は，仏単語 f_j に対応する英単語 e_{a_j} に依存する

$$P(f_j|a_1^{j}, f_1^{j-1}, m, e) = t(f_j|e_{a_j})$$

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

$$P(F,a|E) = \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.7}$$

$$P(F|E) = \frac{\epsilon}{(l+1)^m}\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.8}$$

$$= \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m}\sum_{i=0}^{l} t(f_j|e_{a_j}) \tag{2.9}$$

モデル 1 では翻訳確率 $t(f|e)$ 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 $t(f|e)$ の初期値を設定する．

手順 2 仏英対訳対 $(F^{(s)}, E^{(s)})$(但し，$1 \le s \le S$) において，仏単語 f と英単語 e が対応する回数の期待値を以下の式により計算する．

$$c(f|e;F,E) = \frac{t(f|e)}{t(f|e_0)+\cdots+t(f|e_l)}\sum_{j=1}^{m}\delta(f,f_j)\sum_{i=0}^{l}\delta(e,e_i) \tag{2.10}$$

$\delta(f,f_j)$ はフランス語文 F 中で仏単語 f が出現する回数，$\delta(e,e_i)$ は英語文 E 中で英単語 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 $t(f|e)$ を計算する．

1. 定数 λ_e を以下の式により計算する．

$$\lambda_e = \sum_f \sum_{s=1}^{S} c(f|e; F^{(s)}, E^{(s)}) \tag{2.11}$$

2. (2.11) 式より求めた λ_e を用いて，翻訳確率 $t(f|e)$ を再計算する．

$$\begin{aligned} t(f|e) &= \lambda_e^{-1} \sum_{s=1}^{S} c(f|e; F^{(s)}, E^{(s)}) \\ &= \frac{\sum_{s=1}^{S} c(f|e; F^{(s)}, E^{(s)})}{\sum_f \sum_{s=1}^{S} c(f|e; F^{(s)}, E^{(s)})} \end{aligned} \tag{2.12}$$

手順 4 翻訳確率 $t(f|e)$ が収束するまで手順 2 と手順 3 を繰り返す．

model2

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，j 番目の仏単語 f_j と対応する英単語の位置 a_j は英語文の長さ l に加えて，j と，フランス語文の長さ m に依存し，以下のような関係とする．

$$a(a_j|j, m, l) \equiv P(a_j|a_1^{j-1}, f_1^{j-1}, m, l) \tag{2.13}$$

この関係からモデル 1 における (2.8) 式は，以下の式に変換できる．

$$P(F|E) = \epsilon \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) a(a_j|j, m, l) \tag{2.14}$$

$$= \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_{a_j}) a(a_j|j, m, l) \tag{2.15}$$

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

$$c(f|e; F, E) = \frac{t(f|e)}{t(f|e_0) + \cdots + t(f|e_l)} \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=1}^{l} \delta(e, e_i) \tag{2.16}$$

$$= \sum_{j=1}^{m} \sum_{i=0}^{l} \frac{t(f|e) a(i|j, m, l) \delta(f, f_j) \delta(e, e_i)}{t(f|e_0) a(0|j, m, l) + \cdots + t(f|e_l) a(l|j, m, l)} \tag{2.17}$$

$$c(i|j,m,l;F,E) = \sum_a P(a|E,F)\delta(i,a_j) \tag{2.18}$$

$$= \frac{t(f_j|e_i)a(i|j,m,l)}{t(f_j|e_0)a(0|j,m,l)+\cdots+t(f_j|e_l)a(l|j,m,l)} \tag{2.19}$$

$c(f|e;F,E)$ は対訳文中の英単語 e と仏単語 f が対応付けされる回数の期待値，$c(i|j,m,l;F,E)$ は英単語の位置 i が仏単語の位置 j 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 $a(i|j,m,l) = (l+1)^{-1}$ 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．

model3

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 $P(f|e)$
  英単語 e が仏単語 f に翻訳される確率
- 繁殖確率 $n(\phi|e)$
  英単語 e が ϕ 個の仏単語と対応する確率
- 歪み確率 $d(j|i,m,l)$
  英語文の長さ l，フランス語文の長さ m のとき，i 番目の英単語 e_i が j 番目の仏単語 f_j に翻訳される確率

さらに，英単語が仏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．

$$P(\phi_0|\phi_1^l,E) = \left(\begin{array}{c}\phi_1+\cdots+\phi_l\\ \phi_0\end{array}\right) p_0^{\phi_1+\cdots+\phi_l-\phi_0} p_1^{\phi_0} \tag{2.20}$$

したがって，モデル 3 は以下の式で求められる．

$$P(F|E) = \sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l} P(F,a|E) \tag{2.21}$$

$$
\begin{aligned}
&= \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \left(\begin{array}{c} m-\phi_0 \\ \phi_0 \end{array} \right) p_0^{m-2\phi_0} p_1^{\phi_0} \prod_{i=1}^{l} \phi_i! n(\phi_i|e_i) \\
&\times \prod_{j=1}^{m} t(f_j|e_{a_j}) d(j|a_j, m, l) \qquad (2.22)
\end{aligned}
$$

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．

model4

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 $d(j|i.m, l)$ を 2 つの場合で考える．

- 繁殖数が 1 以上である英単語に対応する仏単語の中で，最も文頭に近い場合

$$
P(\Pi_{[i]1} = j|\pi_1^{[i]-1}, \tau_0^l, \phi_0^l, E) = d_1(j - \odot_{i-1}|\mathcal{A}(e_{[i-1]}), \mathcal{B}(f_j)) \qquad (2.23)
$$

  $\odot_{i-1}$ は i-1 番目の英単語が対応する仏単語の位置を表している．

- それ以外の場合

$$
P(\Pi_{[i]k} = j|{\pi_{[i]1}}^{k-1}, \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, E) = d_{>1}(j - \pi_{[i]k-1}|\mathcal{B}(f_j)) \qquad (2.24)
$$

  $\pi_{[i]k-1}$ は同じ英単語に対応している直前の仏単語を表している．

model5

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 以上である英単語に対応する仏単語の中で，最も文頭に近い場合

$$
\begin{aligned}
P(\Pi_{[i]1} &= j|\pi_1^{[i]-1}, \tau_0^l, \phi_o^l, E) \\
&= d_1(v_j|\mathcal{B}(f_j), v_{\odot_{i-1}}, v_m - \phi_{[i]} + 1)(1 - \delta(v_j, v_{j-1}))
\end{aligned}
$$

  v_j は j 番目までの空白数，$\mathcal{A}$ は英語の単語クラス $\mathcal{B}$ はフランス語の単語クラスを表している．

- それ以外の場合

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

2.2.7 GIZA++

GIZA++ とは，日英方向と英日方向の対訳文から最尤な単語対応を得るための計算を行うツールである．IBM 翻訳モデルを用いて，対訳文 (原言語文と目的言語文の対) から対訳単語と単語翻訳確率を自動的に得る．

GIZA++を用いた場合，以下の 2 つのファイルが出力される．

1. **T TABLE (Translation Table)** T TABLE は，Model1 から Model3 により作成された翻訳確率 $P(f|e)$ のデータである．f は翻訳する言語で，e は目的言語である．T TABLE は各行が，目的言語の単語 ID(e_id)，翻訳する言語の単語 ID(f_id)，翻訳する言語の単語から目的言語の単語へ翻訳する確率 ($P(f_id|e_id)$) で構成される．

2. **N TABLE (Fertility Table)** N TABLE は，目的言語の単語における繁殖数を表したデータである．N TABLE は各行が，目的言語の単語 ID(e_id)，繁殖数が 0 である確率 ($p0$)，繁殖数が 1 である確率 ($p1$)，…，繁殖数が n である確率 (pn) で構成される．

2.3 句に基づく統計翻訳

句に基づく統計翻訳は句対応の翻訳モデルを用いる．原言語文を目的言語文に翻訳する場合に, 隣接する複数の単語 (フレーズ) を用いて翻訳を行う方法である．本研究では日英方向の翻訳を行うため，日英統計翻訳を説明する．日英統計翻訳システムの枠組みを図 2.1 に示す.

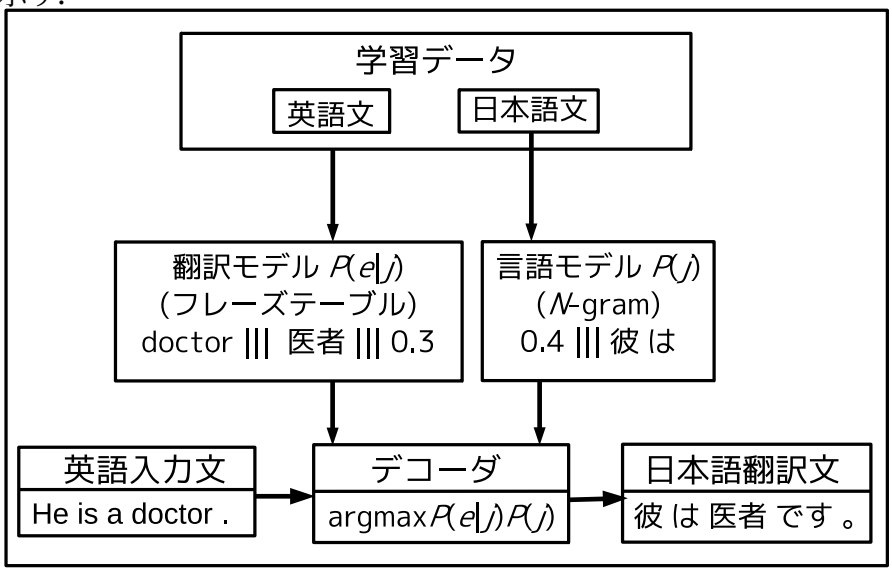

図 2.1: 日英統計翻訳の枠組み

$$E = argmax_j P(e|j) \tag{2.25}$$

$$\simeq argmax_j P(j|e)P(e) \tag{2.26}$$

ここで $P(j|e)$ は翻訳モデル，$P(e)$ は言語モデルを示す．$P(e)$ が単語であれば “単語に基づく統計翻訳” のモデル，$P(e)$ が句であれば，“句に基づく統計翻訳” のモデルとなる．

また，学習データとは対訳文 (英語文と日本語文の対) を大量に用意したものである．学習データに含まれる各々のデータから，翻訳モデルと言語モデルを学習する．

2.3.1 翻訳モデル

翻訳モデルとは，膨大な量の対訳データを用いて英語のフレーズが日本語のフレーズへ確率的に翻訳を行うためのモデルである．この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以下の手順で作成される．また，フレーズテーブルの例も以下に示す.

手順 1 IBM モデルを用いて，単語の対応を得る

手順 2 ヒューリスティックスなルールを用いて句に基づく対応を得る

手順 3 手順 2 で求めた句対応から，フレーズテーブルを作成する

フレーズテーブルの例

The flower ||| その 花 ||| 0.428571 0.0889909 0.428571 0.0907911 2.718
Tonight's concert is ||| 今晩 の コンサート は ||| 0.5 0.000223681 0.5 0.0124601 2.718

左から英語フレーズ，日本語フレーズ，フレーズの英日方向の翻訳確率 $P(j|e)$，英日方向の単語の翻訳確率の積，フレーズの日英方向の翻訳確率 $P(e|j)$，日英方向の単語の翻訳確率の積，フレーズペナルティ(値は常に自然対数の底 e=2.718) である．

2.3.2 フレーズテーブル作成法

まず，GIZA++を用いて学習文から英日，日英方向の双方向で最尤な単語アライメントを得る．英日方向の単語対応の例を表 2.5，日英方向の単語対応の例を表 2.6 に示す．また，●は単語が対応した箇所を示す．

表 2.5: 日英方向の単語対応

	She	went	to	Tokyo	on	travel
彼女	●					
は			●			
旅行						●
で					●	
東京				●		
に		●			●	
行っ		●				
た		●				

表 2.6: 英日方向の単語対応

	She	went	to	Tokyo	on	travel
彼女	●					
は						
旅行						●
で					●	
東京				●		
に						
行っ		●				
た						

次に，得られた双方向の単語アライメントを用いて，複数単語のアライメントを得る．このアライメントは双方向の単語対応の和集合と積集合から求める．ヒューリスティックスとして双方向ともに対応する単語対応を用いる “intersection”，双方向のどちらか一方でも対応する単語対応を全て用いる “union” がある．表 2.5 と表 2.6 を用いた “intersection” の例を表 2.7，に “union” の例を表 2.8 に示す．

また “intersection” と “union” の中間のヒューリスティックスとして “grow” と “grow-diag” がある．これら 2 つのヒューリスティックスでは “intersection” の単語対応と “union” の単語対応を用いる．“grow” は縦横方向，“grow-diag” は縦横対角方向に，“intersection” の単語対応から “union” の単語対応が存在する場合にその単語対応も用いる．“grow-diag”

表 2.7: intersection の例

	She	went	to	Tokyo	on	travel
彼女	●					
は						
旅行						●
で					●	
東京				●		
に						
行っ		●				
た						

表 2.8: union の例

	She	went	to	Tokyo	on	travel
彼女	●					
は			●			
旅行						●
で					●	
東京				●		
に		●			●	
行っ		●				
た		●				

の例を表 4.2 に示す.

“grow-diag” の最後に行う処理として “final” と “final-and” がある．“final” は少なくとも片方の言語の単語対応がない場合に，“union” の単語対応を追加する．また，“final-and” は，両側言語の単語対応がない場合に，“union” の候補対応点を追加する．“grow-diag-final-and” の例を表 4.2.2 に示す.

得られた単語アライメントから，全ての矛盾しないフレーズ対を得る．このとき，そのフレーズ対に対して翻訳確率を計算し，フレーズ対に確率値を付与することでフレーズテーブルを作成する.

表 2.9: grow-diag の例

	She	went	to	Tokyo	on	travel
彼女	●					
は						
旅行						●
で					●	
東京				●		
に		●			●	
行っ		●				
た		●				

表 2.10: grow-diag-final-and の例

	She	went	to	Tokyo	on	travel
彼女	●					
は			●			
旅行						●
で					●	
東京				●		
に		●			●	
行っ		●				
た		●				

2.3.3 デコーダ

デコーダは，翻訳モデルと言語モデルを用いて，確率が最大となる翻訳候補を探索し，出力を行う変換器のことである．代表的なデコーダとして，“Moses” [7] がある．

入力文として “She is a teacher .” が与えられたときの翻訳例を図 2.2 に示す．

日英統計翻訳において，$argmax_e P(e|j)P(j)$ の確率が最大となる英語文を出力するために，適切な順序で日本語と英語の単語対応を得る必要がある．しかし，適切な日本語文を決定するためには，計算量が膨大となり，かつ莫大な時間が必要となる．そこで計算量を削減するために，ビームサーチ法を用いる．

ビームサーチ法とは，翻訳候補の探索において，翻訳確率の低い翻訳候補を枝刈りし，探索範囲を減退する方法である．探索領域の中で一定の確率以上の翻訳候補のみを残し，それ以外の翻訳候補は除外する．

ただし，ビームサーチ法は，切り捨てられた翻訳候補が文章全体で見たときに，最大

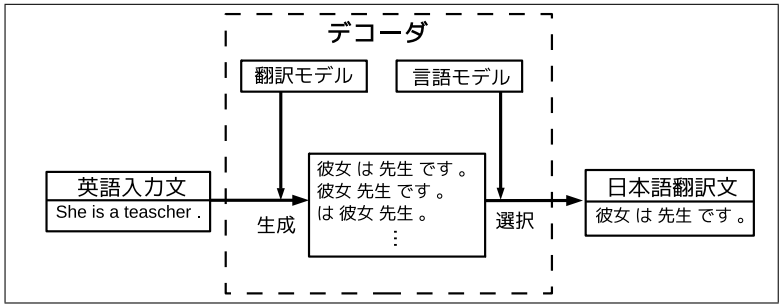

図 2.2: デコーダの動作例

の確率を持つ翻訳候補であったという可能性がある．そのため選択した翻訳文が最適解であるとは限らないという問題がある．

2.4 Pattern Based SMT

2.4.1 概要

Pattern Based SMT は，原言語と目的言語の対訳フレーズから成る“ 対訳フレーズ辞書 ”と，対訳文に対して，任意の句を変数化した“ 句に基づく対訳文パターン辞書 ”を統計的手法を用いて自動作成し，翻訳を行う．辞書の自動作成により，開発コストが削減できる．以下に Pattern Based SMT の手順を示す.

2.4.2 Pattern Based SMT による出力文生成の手順

手順 1 対訳文と GIZA++を用いて“ 対訳単語辞書 ”を作成する．

手順 2 対訳文と対訳単語辞書を用いて,“ 単語に基づく対訳文パターン辞書 ”を作成する．

手順 3 対訳文と単語に基づく対訳文パターンを照合し，変数部に対応する対訳フレーズを抽出し,“ 対訳フレーズ ”を作成する．

手順 4 抽出した対訳フレーズに対訳単語辞書を用いて，対訳フレーズ対数確率を付与した,“ 対訳フレーズ辞書 ”を作成する．

手順 5 対訳文と対訳フレーズの照合を行い，対訳フレーズが適合した対訳文のフレーズを変数化して句に基づく対訳文パターンを作成する．

手順 6 対訳単語辞書を用いて，対訳文パターン対数確率を付与した,“ 句に基づく対訳文パターン辞書 ”を作成する．

手順 7 入力文と対訳フレーズ辞書と句に基づく対訳文パターン辞書を用いて，出力候補文を生成する．

手順 8 選択された句に基づく対訳文パターンの対訳文パターン対数確率と挿入された対訳フレーズの対訳フレーズ対数確率と言語モデルの総和を取り最も高い出力候補文を，出力文とする．

2.4.3 対訳単語辞書の作成

対訳文と GIZA++を用いて，対訳単語に単語翻訳確率を付与した,“ 対訳単語辞書 ”を作成する．対訳単語辞書の作成を図 2.3 に示す.

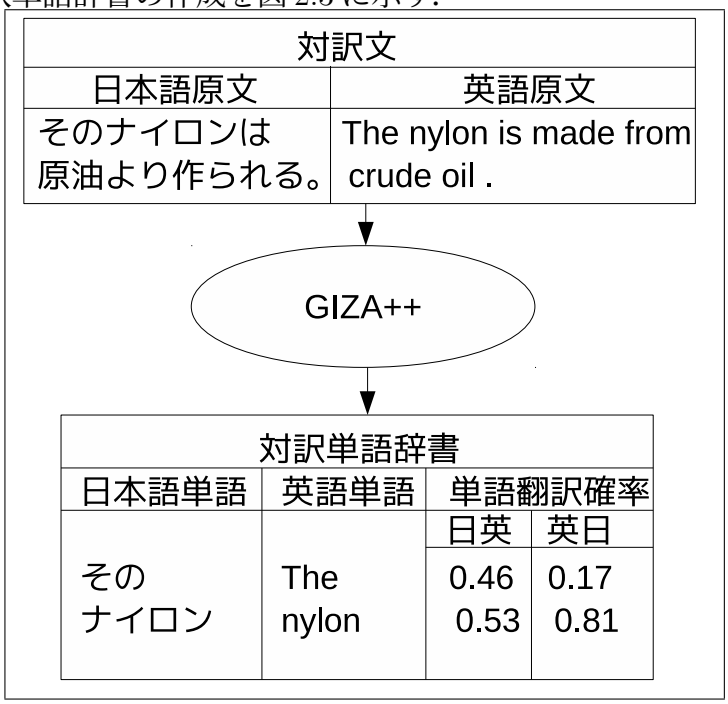

図 2.3: 対訳単語辞書の作成

単語翻訳確率には，日英方向の単語翻訳確率と，英日方向の単語翻訳確率があり，付与するにはまず，対訳文と GIZA++から日英方向の単語対応と英日方向の単語対応を取得する．そして，取得した単語対応から単語翻訳確率を得る.

2.4.4 単語に基づく対訳文パターンの作成

対訳文と対訳単語の照合を行う．対訳単語と適合した対訳文の単語を変数化して単語に基づく対訳文パターンを作成する．単語に基づく対訳文パターンの作成を図 2.4 に示す．

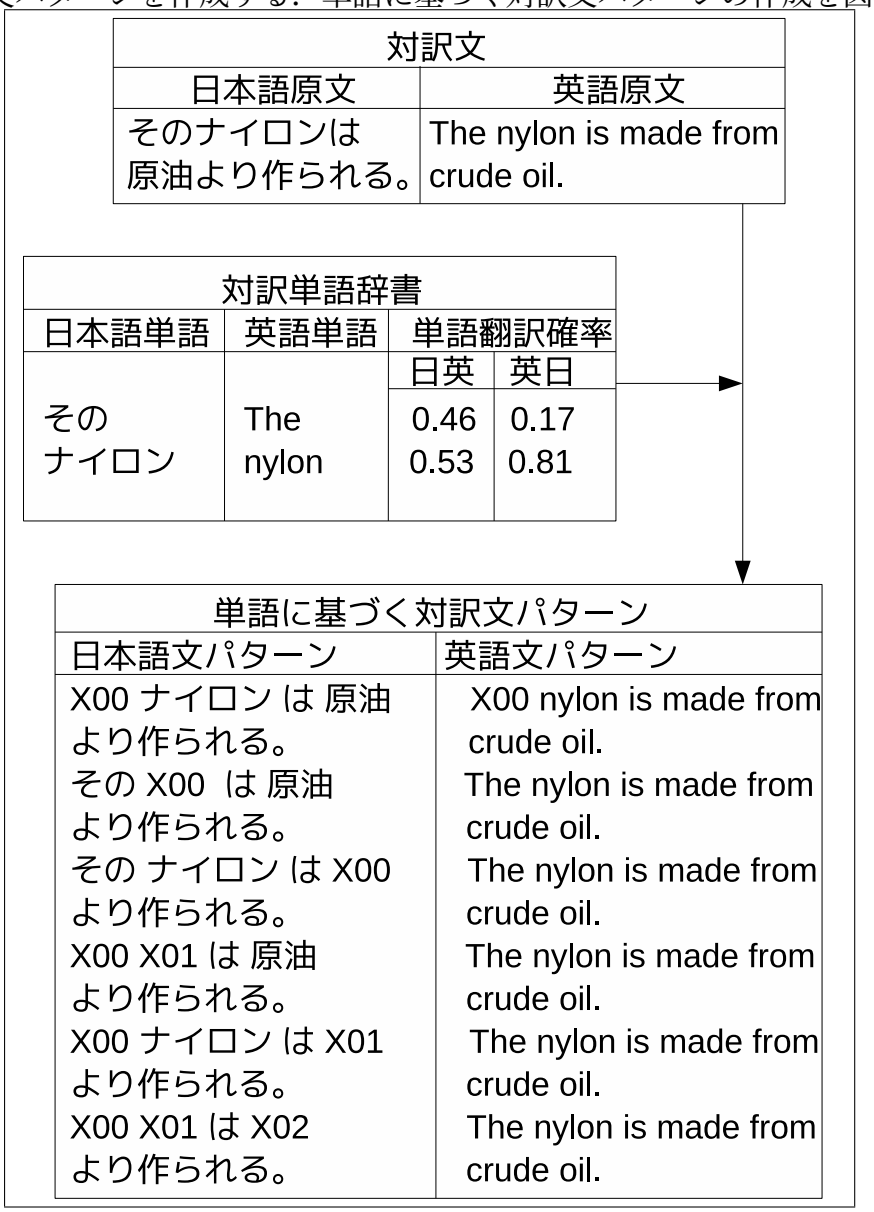

図 2.4: 単語に基づく対訳文パターンの作成

2.4.5　対訳フレーズ辞書の作成

対訳文と単語に基づく対訳文パターンを照合し，変数部に対応する対訳フレーズを抽出する．抽出した対訳フレーズに対訳単語辞書を用いて，対訳フレーズ対数確率を付与した，“ 対訳フレーズ辞書 ”を作成する．対訳フレーズ辞書の作成を図 2.5 に示す．

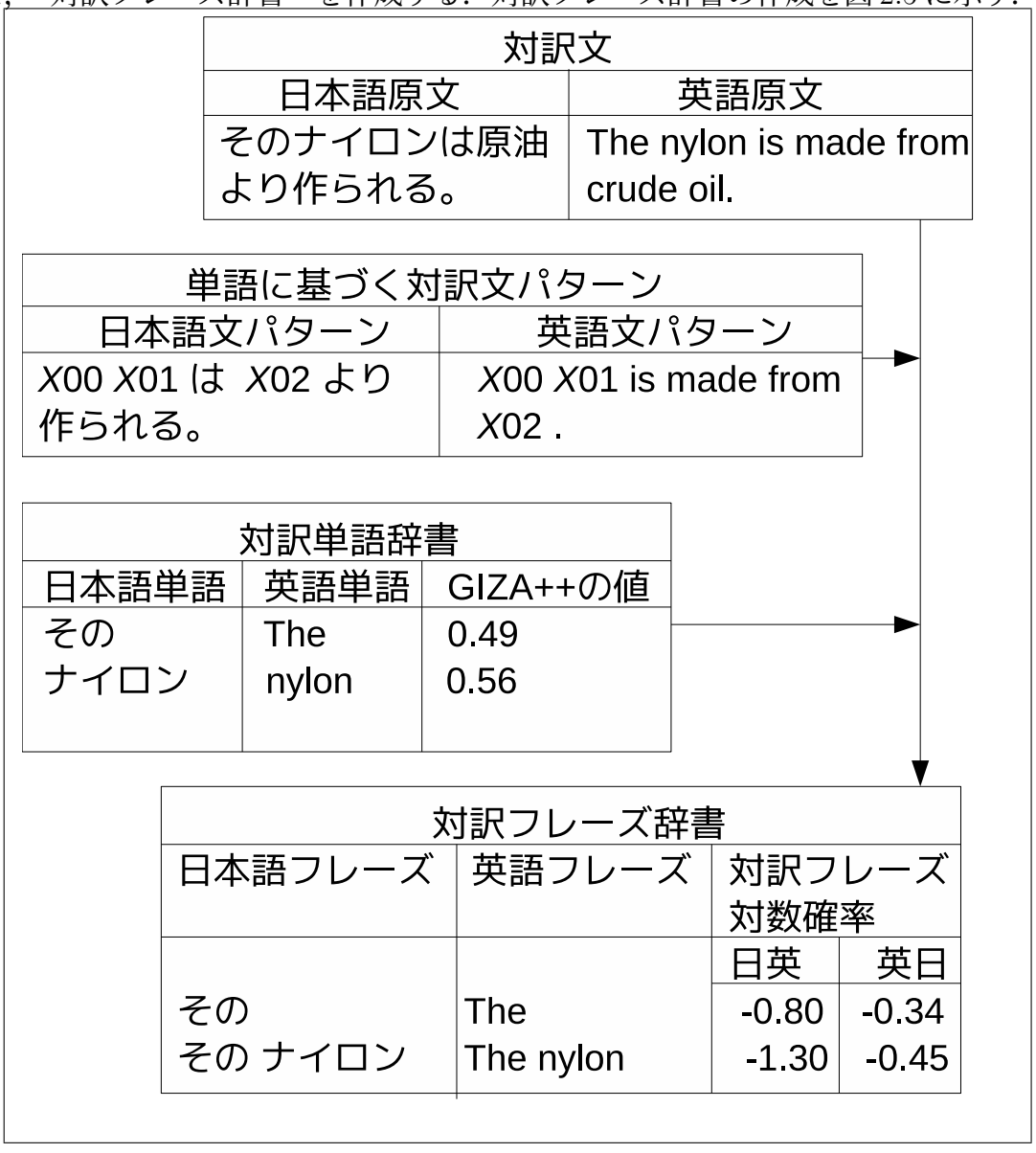

図 2.5: 対訳フレーズ辞書の作成

対訳フレーズ対数確率

抽出した対訳フレーズにGIZA++の値を用いて，対訳フレーズ対数確率を付与する．対訳フレーズ対数確率は，以下の式 (1) に示す．

$$log_2 P(\frac{J_0 \cdots J_{N-1}}{E_0 \cdots E_{M-1}}) = \sum_{n=0}^{N-1} \arg\max_{m=0}^{M-1}(log_2(p(J_n|E_m)) + log_2(p(E_m|J_n))) \quad (1)$$

J_n; 日本語の単語 N; 日本語の単語数

E_m; 英語の単語 M; 英語の単語数

$p(J_n|E_m)$; 英単語 E_m が日本単語 J_n に翻訳される確率 (GIZA++の値)

対訳フレーズ対数確率にも，2.4.3 節の単語翻訳確率と同じように日英方向と英日方向がある．日英対訳フレーズ対数確率を付与する方法は，抽出した対訳フレーズの日本語単語と英語単語の日英方向の全ての組み合わせを得る．単語辞書の単語翻訳確率を用いて，各組み合わせから最大となる単語翻訳確率を得る．そして，単語翻訳確率の対数をとり総和を求める．この総和が日英対訳フレーズ対数確率となる．同様の処理を，英日方向に対しても行い，英日対訳フレーズ対数確率を得る．日英対訳フレーズ対数確率の付与を図 2.6 に，英日対訳フレーズ対数確率の付与を図 2.7 に示す．

図 2.6: 日英方向の対訳フレーズ対数確率の付与

図 2.7: 英日方向の対訳フレーズ対数確率の付与

2.4.6 句に基づく対訳文パターンの作成

対訳文と対訳フレーズの照合を行う．対訳フレーズが適合した対訳文のフレーズを変数化して句に基づく対訳文パターンを作成する．以下に，句に基づく対訳文パターンの作成を図 2.8 に示す.

対訳文パターン対数確率

対訳文パターン対数確率にも日英方向と英日方向がある．日英対訳文パターン対数確率を付与する方法は，作成した句に基づく対訳文パターンの日本語文パターンと英語文パターンの全ての組み合わせを得る．単語辞書の単語翻訳確率を用いて，各組み合わせから最大となる単語翻訳確率を得る．そして，単語翻訳確率の対数をとり総和を求める．この総和が日英対訳文パターン対数確率となる．同様の処理を，英日方向に対しても行い，英日対訳フレーズ対数確率を得る.

日英対訳文パターン対数確率の付与を図 2.9 に，英日対訳文パターン対数確率の付与を図 2.10 に示す．また，日英における句に基づく対訳文パターンの確率値の計算方法を式 (2) に，英日における句に基づく対訳文パターンの確率値の計算方法を式 (3) に示す.

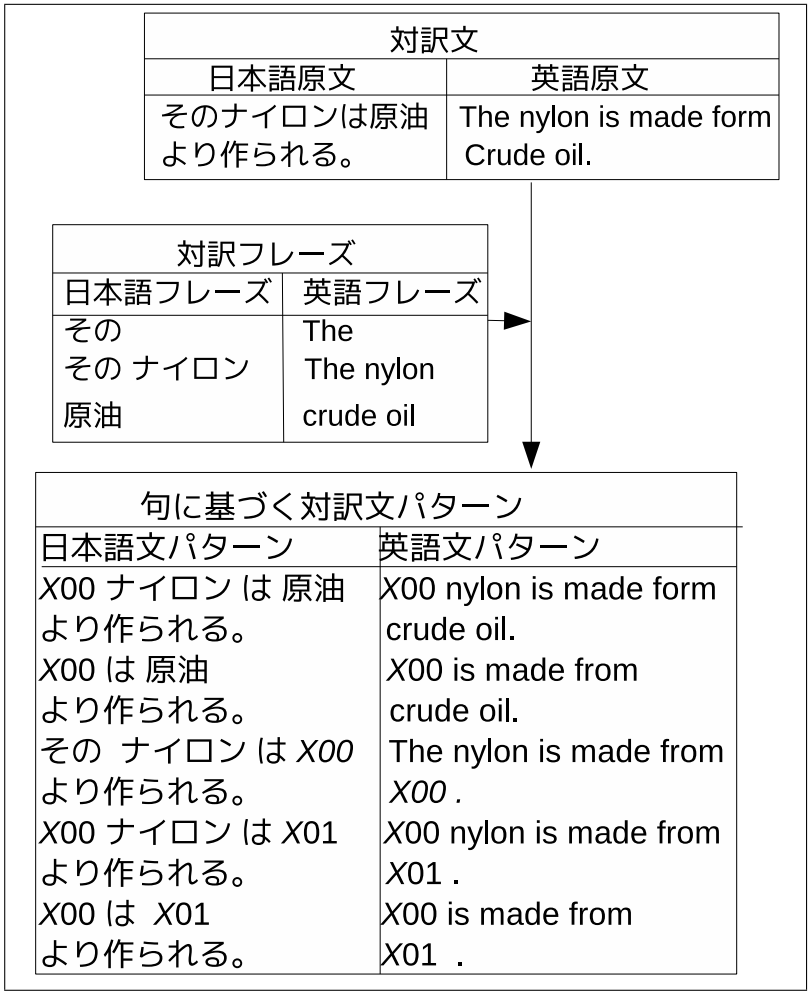

図 2.8: 句に基づく対訳文パターンの作成

$log_2 P(\frac{J_0 \cdots J_{N-1}, JX_0 \cdots JX_{N-1}}{E_0 \cdots E_{M-1}, EX_0 \cdots EX_{M-1}}) =$
$\sum_{n=0}^{N-1} \arg\max_{m=0}^{M-1}(log_2(p(E_m|J_n)) + log_2(p(J_n|E_m)))$ (2)
J_n; 対訳フレーズ中の日本語の単語 N; 日本語の単語数
E_m; 対訳フレーズ中の英語の単語 M; 英語の単語数
$p(J_n|E_m)$; 英単語 E_m が日本単語 J_n に翻訳される確率 (GIZA++の値) $log_2 P(\frac{J_0 \cdots J_{N-1}, JX_0 \cdots JX_{N-1}}{E_0 \cdots E_{M-1}, EX_0 \cdots EX_{M-1}}) =$
$\sum_{n=0}^{N-1} \arg\max_{m=0}^{M-1}(log_2(p(J_n|E_m)) + log_2(p(E_m|J_n)))$ (3)
J_n; 対訳フレーズ中の日本語の単語 N; 日本語の単語数
E_m; 対訳フレーズ中の英語の単語 M; 英語の単語数
$p(J_n|E_m)$; 英単語 E_m が日本単語 J_n に翻訳される確率 (GIZA++の値)

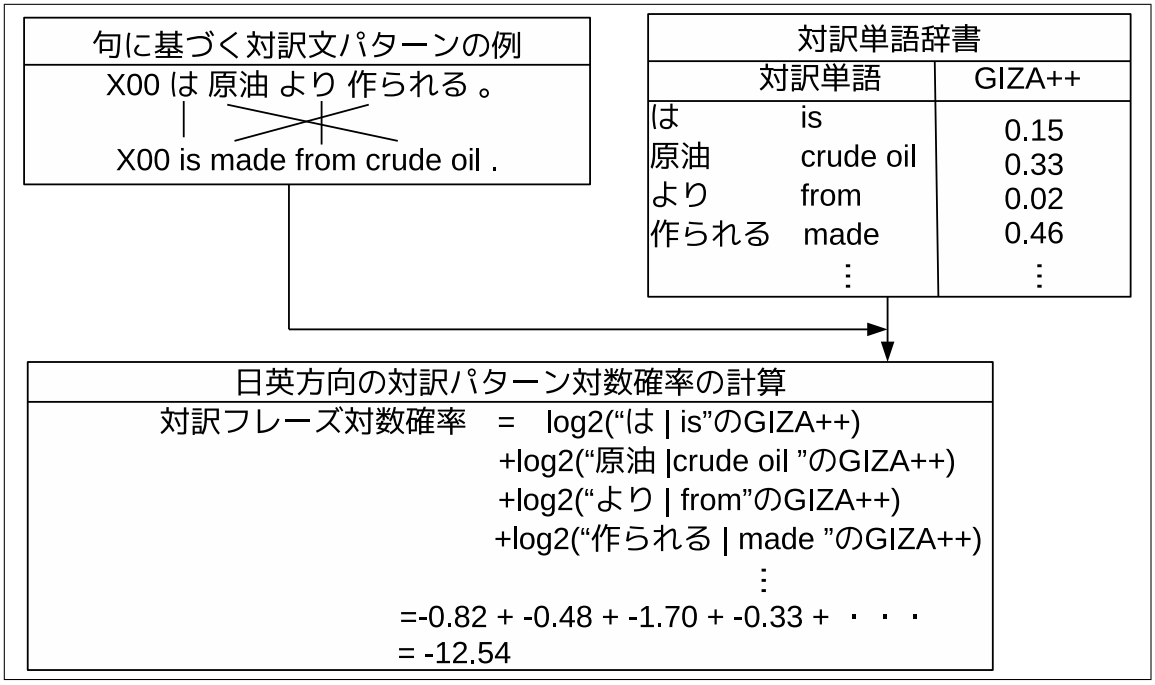

図 2.9: 日英方向の対訳文パターン対数確率の付与

句に基づく対訳文パターン辞書

句に基づく対訳文パターンの変数化していない部分 (以下字面) と，対訳単語辞書を用いて，対訳文パターン対数確率を付与した,“ 句に基づく対訳文パターン辞書 ”を作成する．以下に，句に基づく対訳文パターン辞書の作成を図 2.11 に示す.

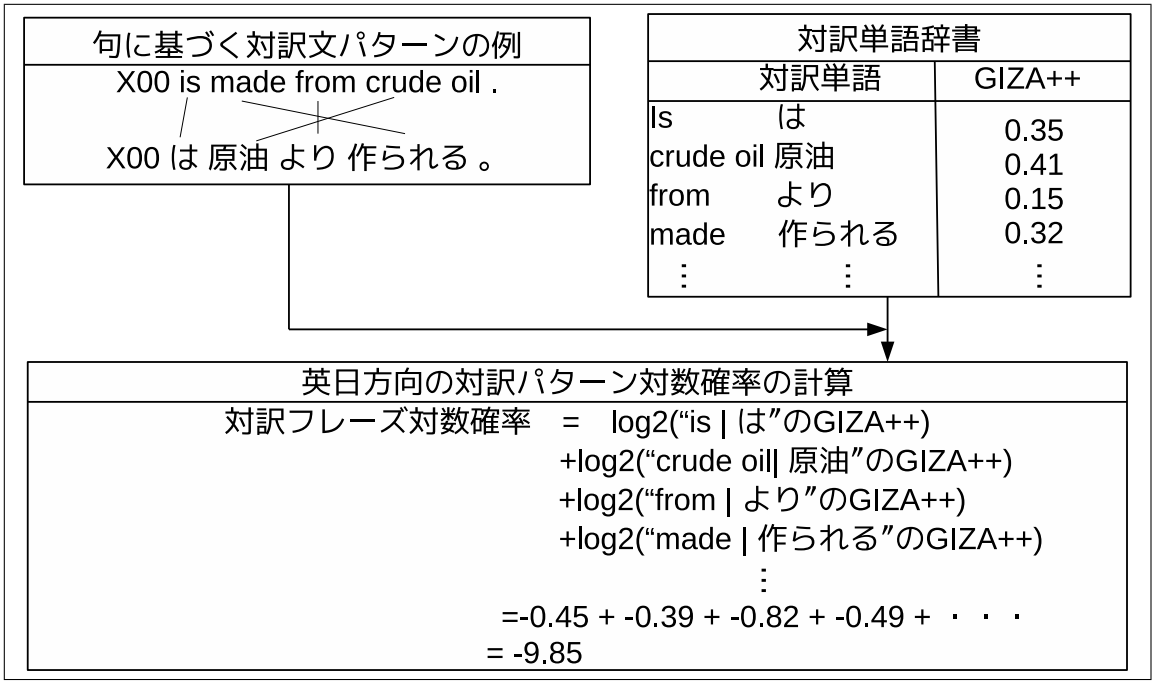

図 2.10: 英日方向の対訳文パターン対数確率の付与

2.4.7 出力文の生成

句に基づく対訳文パターン辞書と対訳フレーズ辞書を利用して出力候補文を生成する．次に，作成した出力候補文から出力文を選択する．出力文の生成方法を以下に，出力文の生成の流れを図 2.12 に示す.

a) 句に基づく日本語文パターンの選択

入力文と，句に基づく日本語文パターンの字面を照合する．字面が多く一致した日本語文パターンを持つ対訳文パターンを優先して選択する.

b) 出力候補文の作成

選択した対訳文パターンにおいて，英語文パターンの変数部に対訳フレーズを用いて英語フレーズを挿入し, 出力候補文を生成する.

c) 出力文の選択

対訳文パターン対数確率 (α) と出力候補文の作成に用いた対訳フレーズ対数確率 (β) と言語モデル (tri-gram)(γ) を用いて，出力候補文の翻訳対数確率を計算する．出力候補文の翻訳対数確率が最も高い出力候補文を“ 出力文 ”として出力する．

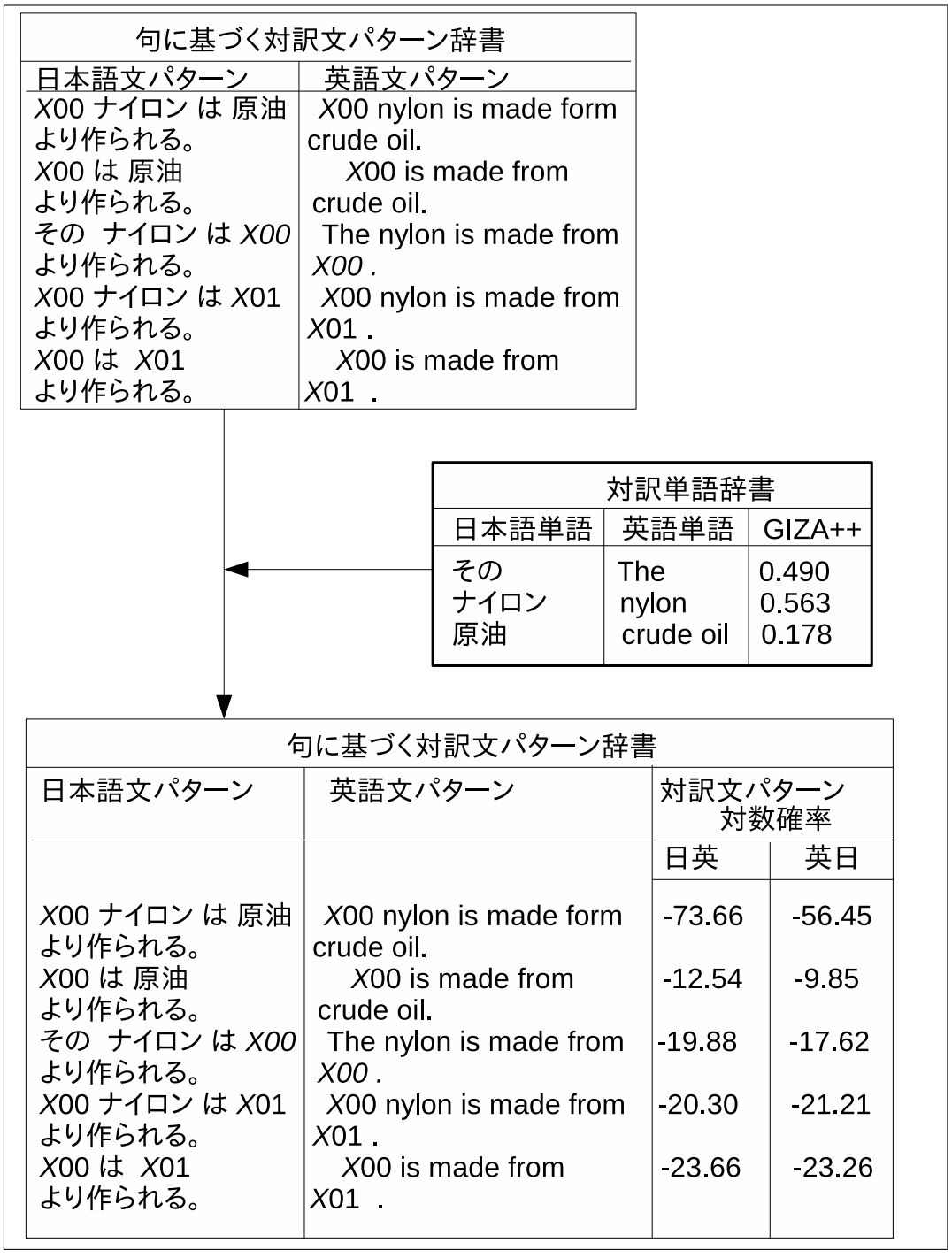

図 2.11: 句に基づく対訳文パターン辞書の作成

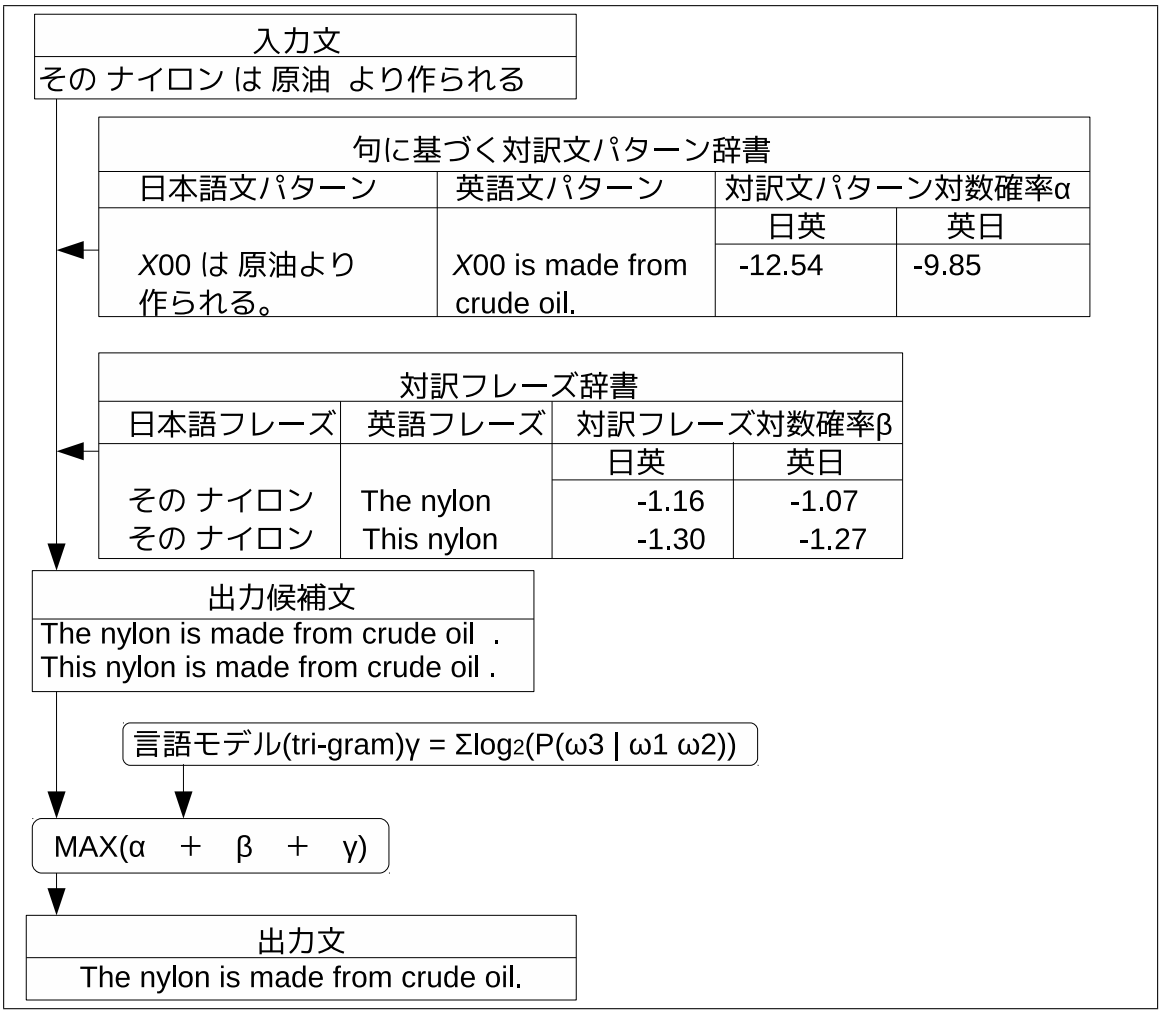

図 2.12: 出力文生成の流れ

第3章　対訳単語辞書の精度調査

3.1　本研究の実験の概要

パターンに基づく統計翻訳は対訳単語辞書を起点として対訳句辞書と句レベル文パターン辞書を作成する．そして作成した対訳句辞書と句レベル文パターン辞書を用いて翻訳を行う．しかしパターンベース統計翻訳の翻訳精度はまだまだ低い．翻訳精度が低い原因としてこの対訳単語の精度が低いことが問題と考える．対訳単語辞書の精度が低いため精度の低い対訳句辞書と対訳文パターン辞書が翻訳に用いられている可能性がある．そこで本研究では対訳単語辞書の精度を調査する．対訳学習文と GIZA++を用いて対訳単語を作成し，全対訳単語から枝刈りを行い，対訳単語辞書を作成する．不適切な対訳単語を調査し，調査結果をふまえて枝刈り条件を変更し，現在翻訳に利用されている対訳単語辞書の精度向上を試みる．

3.1.1 パターンに基づく統計翻訳の問題点

パターンに基づく統計翻訳では対訳単語辞書を起点として翻訳を行う．翻訳精度が低い理由として対訳単語辞書の精度の低いことが問題であると考える．翻訳精度低下の原因を以下に示す.

1. 対訳単語に誤りが含まれることにより単語レベル文パターンに誤りが含まれる.
2. 単語レベル文パターンに誤りが含まれることで対訳句と句レベル文パターンに誤りが含まれる.
3. 対訳句と句レベル文パターンに誤りが含まれることで翻訳において翻訳精度が低下する.

3.1.2 研究の目的

対訳単語辞書は対訳単語確率を基にして作成する．対訳単語確率は対訳文と IBM 翻訳モデルにより計算される．本研究では対訳単語に関して日本語単語と英語単語の適切な対応がとられているかの調査を行う.

第4章　実験：精度調査

4.1　対訳単語辞書の調査条件

学習文 100,000 文と IBM 翻訳モデルで作成した対訳単語 327,604 単語において，日英の対応の精度を調査する．対訳単語の計算に用いる GIZA++のパラメータを以下に示す．

- m1=4，m2=0，mh=4，m3=0，m4=0，t1=4

4.2　対訳単語の調査

評価は全てランダム 100 単語を取り出して行う．

評価基準を 4.1 に示す．

表 4.1: 評価基準

○	適切な対訳単語
×	不適切な対訳単語

また，表中の $P(E/J)$ は日本語単語が英語単語に訳される GIZA++の対訳単語確率，$P(J/E)$ は英語単語が日本語単語に訳される GIZA++の対訳単語確率である．

4.2.1 全対訳単語

全対訳単語を調査した結果を以下に示す.

表 4.2: 全対訳単語の評価

評価対象数	○	×
327,604	17	83

表 4.3: 全対訳単語の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
ご馳走	feast	-3.47	-4.18	○
ら	Association	-21.43	-17.70	×
’	’hour’	-3.16	-1.70	×
1	in	-5.33	-16.21	×

表 4.3 よりひらがな 1 文字・記号・数字において不適切な対訳単語が多く作成されていると考え，調査を行なった．また，アルファベットや頻度においても適切な対訳単語が作成されているか調査を行った.

4.2.2 記号

全対訳単語中の記号を調査した結果を以下に示す.

表 4.4: 記号の評価

評価対象数	○	×
317	0	100

表 4.5: 記号の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
°	triangle	-1.76	-19.60	×
「	brand	-23.12	-4.58	×

表より記号において適切な対訳単語が作成されていないことがわかった.

4.2.3 ひらがな 1 文字

全対訳単語中のひらがな 1 文字を調査した結果を以下に示す.

表 4.6: ひらがな 1 文字の評価

評価対象数	○	×
630	0	100

表 4.7: ひらがな 1 文字の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
き	blow	-20.44	-17.98	×
は	This	-14.61	-3.84	×

表 4.6 よりひらがな 1 文字の対訳単語においても記号と同様に全て不適切な対訳単語が作成されていることがわかる.

4.2.4 日本語単語の頻度 1

全対訳単語中の日本語単語の頻度 1 の対訳単語を調査した結果を以下に示す.

表 4.8: 日本語単語の頻度 1 の単語の評価

評価対象数	○	×
30,468	16	84

表 4.9: 日本語単語の頻度 1 の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
ぜん息	asthma	-2.32	-10.81	○
希少	stamps	-8.18	-18.64	×

日本語単語の頻度 1 の対訳単語の精度において全対訳単語 (表 4.2) とあまり差がない結果となった.

4.2.5 英語単語の頻度 1

全対訳単語中の英語単語の頻度 1 の対訳単語を調査した結果を以下に示す.

表 4.10: 英語単語の頻度 1 の単語の評価

評価対象数	○	×
43,001	14	86

表 4.11: 英語単語の頻度 1 の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
むっつり	sullenly	-1.56	-1.87	○
自称	professedly	-2.16	-1.59	×

英語単語の頻度 1 の対訳単語の評価 (表 4.10) は全対訳単語の評価 (表 4.2) と比べて差があまりない結果となった.

4.2.6 日本語単語と英語単語が両方同時に含まれる対訳文の頻度 1

全対訳単語中の日本語単語と英語単語が両方同時に含まれる対訳文の頻度 1 の対訳単語を調査した結果を以下に示す.

表 4.12: 頻度 1 の対訳単語の評価

評価対象数	○	×
246,217	14	86

表 4.13: 頻度 1 の対訳単語の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
こい	carp	-14.47	-16.19	○
どんどん	better	-13.29	-8.45	×

この種の対訳単語の評価は全対訳単語の評価 (表 4.2) と差が小さい.

4.2.7 日本語単語の頻度 1 かつ英語単語の頻度 1

全対訳単語中の日本語単語の頻度 1 かつ英語単語の頻度 1 の対訳単語を調査した結果を以下に示す.

表 4.14: 日本語単語の頻度 1 かつ英語単語の頻度 1 の対訳単語の評価

評価対象数	○	×
7,083	37	63

表 4.15: 日本語単語の頻度 1 かつ英語単語の頻度 1 の対訳単語の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
マストドン	Mastodons	-1.07	-1.59	○
ダンツァス	Hamburg	-14.61	-3.87	×

評価より適切な対訳単語が 4 割あり，精度は高いと考えた．しかし対訳単語全体を見たところ固有名詞が大半であった.

4.2.8 日本語単語と英語単語が両方同時に含まれる頻度≧2

全対訳単語中の日本語単語と英語単語が両方同時に含まれる文の数が頻度2以上の対訳単語を調査した結果を以下に示す.

表 4.16: 頻度2以上の対訳単語の評価

評価対象数	○	×
83,017	40	60

表 4.17: 頻度2以上の対訳単語の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
船底	bottom	-1.04	-5.86	○
川	crossed	-14.70	-16.14	×

日本語単語の頻度1かつ英語単語の頻度1の対訳単語 (表 4.14) と頻度2以上の対訳単語 (表 4.16) は精度が高い.

4.2.9 数字

全対訳単語中の対訳単語において数字を調査した結果を以下に示す.

表 4.18: 数字の評価

評価対象数	○	×
569	8	92

表 4.19: 数字の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
６	6	-3.29	-0.64	○
１	to	-9.33	-22.33	×

数字においては”６”と”6”のように対応がとれている対訳単語以外不適切であった.

4.2.10 アルファベット大文字 (日本語単語)

全対訳単語中の対訳単語において日本語単語がアルファベット大文字の対訳単語を調査した結果を以下に示す.

表 4.20: アルファベット大文字 (日本語単語)

評価対象数	○	×
238	12	88

表 4.21: アルファベット大文字 (日本語単語) の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
A	A	-0.30	-7.73	○
A	Company	-6.55	-3.02	×

4.2.11 アルファベット小文字 (日本語単語)

全対訳単語忠の日本語単語がアルファベット小文字の対訳単語を調査した結果を以下に示す.

表 4.22: アルファベット小文字 (日本語単語)

評価対象数	○	×
92	11	81

表 4.23: アルファベット小文字 (日本語単語) の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
x	x	-1.77	-1.33	○
z	Hz	-3.68	-1.28	×

日本語単語がアルファベット 1 文字の時は数字と同様に対応がとれている対訳単語以外は不適切な対訳単語であることがわかった.

4.2.12 アルファベット大文字 (英語単語)

全対訳単語中の英語単語がアルファベット大文字の対訳単語を調査した結果を以下に示す.

表 4.24: アルファベット大文字 (英語単語)

評価対象数	○	×
271	11	89

表 4.25: アルファベット大文字 (英語単語) の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
私	I	-0.91	-0.75	○
から	I	-14.03	-18.84	×

4.2.13 アルファベット小文字 (英語単語)

全対訳単語中の英語単語がアルファベット小文字の対訳単語を調査した結果を以下に示す.

表 4.26: アルファベット小文字 (英語単語)

評価対象数	○	×
99	5	94

表 4.27: アルファベット小文字 (英語単語) の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
x	x	-1.77	-1.33	○
です	a	-7.31	-20.85	×

英語単語がアルファベット小文字 1 文字の時ほとんどが不適切な対訳単語であることがわかった.

4.3 対訳単語辞書の評価

対訳単語辞書は全対訳単語から枝刈りを行い，作成している．本節では現在翻訳に用いられている対訳単語辞書と枝刈り条件変更後の対訳単語辞書において精度の比較を行う．

4.3.1 翻訳に用いる対訳単語辞書の評価

翻訳に用いる対訳単語は以下の条件で枝刈りを行って作成する．

- 付与した対訳単語確率を用いて作成した対訳単語の順位が日本語・英語ともに 8 位以上の単語
- 対訳単語確率が $log2(0.20)$ より高い単語
- 日本語単語と英語単語が両方同時に含まれる文の数 (頻度) が 2 以下の単語

枝刈りした後の対訳単語辞書を調査した結果を以下に示す．

表 4.28: 翻訳に用いる対訳単語辞書の評価結果

評価対象数	○	×
4340	90	10

表 4.29: 翻訳に用いる対訳単語辞書の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
興味	interested	-2.24	-1.72	○
十郎	Danjuro	-1.22	-1.46	×

現在の枝刈り条件である程度の不適切な対訳単語は枝刈りされて対訳単語辞書を作成されていることがわかった．

4.3.2 対訳単語辞書の変更と評価

現在翻訳に用いられている対訳単語辞書を作成する枝刈り条件に以下を追加し，対訳単語辞書の変更を行った.

- ひらがな 1 文字・記号・数字の削除

変更後の対訳単語辞書の調査を行なった結果を以下に示す.

表 4.30: 変更後の対訳単語辞書の評価結果

評価対象数	○	×
4322	91	9

表 4.31: 変更後の対訳単語辞書の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
質	quality	-1.06	-2.20	○
ヤクルト	Swallows	-1.08	-1.19	×

従来の対訳単語辞書と比べて対訳単語数は 18 単語減少した.

4.3.3 変更前後の対訳単語辞書の比較

変更前の従来の対訳単語と変更後の対訳単語辞書を比較した結果を以下に示す.

表 4.32: 変更前後の対訳単語辞書の比較結果

	対訳単語数	○	×
従来の対訳単語辞書	4340	90	10
変更後の対訳単語辞書	4322	91	9

変更前後の対訳単語辞書の比較を行なったところ対訳単語数も対訳単語の精度にもあまり差がなかった．差がない結果に関してはひらがな 1 文字・数字・記号の対訳単語中の見出し語数が少ないため対訳単語数にあまり変化がなかったことが原因であると考える.

4.3.4 考察

対訳単語の精度調査の結果をふまえ，対訳単語辞書の変更を行なった結果，対訳単語数にも対訳単語の精度にも差がなかった．差がない結果になったことに関してひらがな 1 文字や数字の見出し語数が少なかったために対訳単語数にあまり変化がなかったことが原因であると考えられる．しかし，ひらがな 1 文字や数字は対訳学習文中においては出現頻度が高い対訳単語である．よって変更後の対訳単語辞書を翻訳に用いることで翻訳結果に変化が起きる可能性がある．

4.3.5 追加実験

変更前 (従来) と変更後 (提案) の対訳単語辞書を用いた翻訳実験を 100 文の対比較実験でおこなった. 表 4.33 において評価は以下の通りである.

- 従来○…従来の対訳単語辞書での翻訳精度 高
- 提案○…提案手法の対訳単語辞書での翻訳精度 高
- 差なし…従来・提案の比較で精度に差なし
- 同一…従来・提案どちらとも同一出力
- 未知語…未知語を含む文
- 翻訳×…翻訳ができなかった文

表 4.33: 翻訳実験評価結果

従来○	提案○	差なし	同一	未知語	翻訳×
12	17	12	32	18	9

以上より提案手法の対訳単語辞書を用いた翻訳実験では少しではあるが精度が向上することがわかった.

第5章　おわりに

パターンに基づく統計翻訳において翻訳精度の低い原因の一つが翻訳に用いる対訳句辞書と句レベル文パターン辞書の作成の起点である対訳単語辞書の精度が低いことが考えられる．そこで本研究では，対訳単語辞書の精度調査を行なった．対訳学習文と GIZA++ を利用して対訳単語を作成し，全対訳単語から枝刈りを行い，対訳単語辞書を作成している．対訳単語において精度調査を行い，調査結果から対訳単語辞書を作成する時の枝刈り条件を変更し，対訳単語辞書の変更を行う．その後変更前後の対訳単語辞書の精度を調査する．実験の結果，対訳単語ではひらがな 1 文字・数字・記号において不適切な対訳単語が多く作成されていた．よってひらがな 1 文字・数字・記号を削除する条件を枝刈り条件に加えて対訳単語辞書を変更した．変更前後の対訳単語辞書の評価結果では対訳単語数と精度に差はなかった．しかしひらがな 1 文字・数字において見出し語数は少ないが対訳学習文中での出現頻度は高い．そのため変更後の対訳単語辞書を使用した翻訳で変化がある可能性があると考えた．

追加実験で実際に翻訳実験を行ったところ少しではあるが翻訳精度は向上した．

謝辞

本研究の作成にあたり，研究の進め方や論文の書き方など丁寧なご指導を頂きました鳥取大学工学部知能情報工学科自然言語処理研究室の村上仁一准教授に心から御礼申し上げます．また，本研究を進めるにあたり，御指導，御助言を頂きました，同じく自然言語処理研究室の村田真樹教授に心から御礼申し上げます．また，研究に協力して頂いた自然言語処理研究室の皆様へ心から感謝の気持ちと御礼を申し上げたく，謝辞にかえさせて頂きます．

参考文献

[1] 渡辺日出雄, 武田浩一, “パターンベース翻訳システム PalmTree”, 情報処理学会第 55 回全国大会講演論文集, pp.80-81, 1997.

[2] Franz Josef Och, Hermann Ney, “A Systematic Comparison of Various Statistical Alignment Models”, Computational Linguistics, 29(1), pp.299-314, 1996.

[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”, 自然言語処理学会第 20 回年次大会予稿集, pp.951-954, 2014.

[4] 西尾聡一郎，村上仁一，徳久雅人，”パターンに基づく統計翻訳における，文パターン確率の考察”，言語処理学会第 20 回年次大会，p8-8，pp.231-234，2016.

[5] 力久 剛士，“レーベンシュタイン距離を用いた翻訳精度の向上”，平成 26 年度 卒業論文，pp.3-15，February 2015.

[6] Vladimir Iosifovich Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals”, Soviet Physics Doklady, 10(8), pp.707-710, 1966.

[7] Philipp Koehn, Marcello Federico, Brooke Cowan, Richard Zens, Chris Dyer, Ondřej 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, June 2007.

[8] 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.

[9] Franz Josef Och, Hermann Ney: ”A Systematic Comparison of Various Statistical Alignment Models”, Computational Linguistics, volume 29, number 1, pp.19-51, March 2003.

[10] Peter F.Brown, Stephen A.Della Pietra, Vincent J.Della Pietra, Robert L.Mercer: "The mathematics of statistical machine translation: Parameter Estimation", Computational Linguistics, 1993.