概要

パターン翻訳 [1] 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．翻訳精度の高い出力文が得られるが，対訳句辞書と対訳文パターン辞書の作成は人手で行うため，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 PatternBased SMT[3] を提案した．対訳句辞書と対訳文パターン辞書を自動的に作成により，開発コストを削減することができた．しかし，対訳文パターンに適合しても，人手評価が低い出力文があった．この問題の原因の一つは，句に基づく対訳文パターンの確率値に，パターン内の単語における GIZA++の値を利用していることにある．そこで本研究では，句に基づく対訳文パターンの確率値の計算に，パターン内の変数部の確率を利用して，翻訳精度の向上を試みる．実験として，100 文の対比較実験を行い，人手による対比較評価をした．その結果，提案手法が良かった例が 12 文，従来手法が良かった例が 11 文，差がなかった例が 23 文，同一出力が 54 文という結果になった．実験の結果，提案手法と従来手法は同等の翻訳精度であった．

目 次

第 1 章　はじめに **1**

第 2 章　従来の研究 **2**

2.1　パターン翻訳 [1] 2

2.1.1　概要 2

2.1.2　日英パターン翻訳の手順 2

2.2　統計翻訳 3

2.2.1　概要 3

2.2.2　単語に基づく統計翻訳 3

2.2.3　IBM 翻訳モデル 4

2.2.4　単語に基づく統計翻訳の問題点 9

2.2.5　GIZA++ 9

2.3　句に基づく統計翻訳 10

2.3.1　翻訳モデル 11

2.3.2　フレーズテーブル作成法 12

2.3.3　言語モデル 15

2.3.4　デコーダ 16

2.4　Pattern Based SMT 16

2.4.1　概要 16

2.4.2　Pattern Based SMT による出力文生成の手順 17

2.4.3　対訳単語辞書の作成 17

2.4.4　単語に基づく対訳文パターンの作成 18

2.4.5　対訳フレーズ辞書の作成 19

2.4.6　句に基づく対訳文パターンの作成 22

2.4.7　出力文の生成 27

第 3 章　提案手法 **29**
3.1 提案手法の概要 29
3.1.1 Pattern Based SMT の問題点 29
3.2 句に基づく文パターン辞書の作成 30

第 4 章　実験 **32**
4.1 実験データ 32
4.2 実験結果 32
4.2.1 対比較実験 33
4.2.2 提案手法○の例 34
4.2.3 従来手法○の例 37

第 5 章　考察 **43**
5.1 提案手法の有効性 43
5.2 moses と提案手法における対比較実験 43
5.2.1 moses での対比較実験 44
5.2.2 moses が提案手法より優れている例 44
5.2.3 提案手法が moses より優れている例 44
5.3 翻訳精度の問題 44
5.4 対訳文パターンの新しい計算方法 45

第 6 章　おわりに **46**

図 目 次

2.1 日英統計翻訳の枠組み 10
2.2 デコーダの動作例 16
2.3 対訳単語辞書の作成 18
2.4 単語に基づく対訳文パターンの作成 19
2.5 対訳フレーズ辞書の作成 20
2.6 日英方向の対訳フレーズ対数確率の付与 21
2.7 英日方向の対訳フレーズ対数確率の付与 22
2.8 句に基づく対訳文パターンの作成 23
2.9 日英方向の対訳文パターン対数確率の付与 24
2.10 英日方向の対訳文パターン対数確率の付与 25
2.11 句に基づく対訳文パターン辞書の作成 26
2.12 出力文生成の流れ 27

3.1 提案手法の具体例 31

表 目 次

2.1 対訳文パターンの例 2
2.2 対訳フレーズの例 3
2.3 英日方向の単語対応 9
2.4 日英方向の単語対応 9
2.5 日英方向の単語対応 12
2.6 英日方向の単語対応 12
2.7 intersection の例 13
2.8 union の例 13
2.9 grow-diag の例 14
2.10 grow-diag-final-and の例 14

3.1 明らかに妥当ではない対応を取っている出力文の例 30

4.1 実験データ 32
4.2 対訳文の例 32
4.3 入力文の例 32
4.4 人手による評価 33
4.5 提案手法○の例 1 34
4.6 提案手法○の例 2 35
4.7 提案手法○の例 3 36
4.8 従来手法○の例 1 37
4.9 従来手法○の例 2 38
4.10 従来手法○の例 3 39
4.11 差なしの例 1 40
4.12 差なしの例 2 41
4.13 差なしの例 3 42

5.1 実験データ 43
5.2 moses での対比較実験 44
5.3 moses が提案手法より優れている例の例 44
5.4 提案手法が moses より優れている例 44

第1章　はじめに

パターン翻訳 [1] は，1960 年代に提案された翻訳方法である．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．この翻訳方式は入力文が適切な対訳文パターンに適合した場合，翻訳精度の高い出力文が得られる．しかし，対訳句辞書と対訳文パターン辞書の作成は人手で行うため，開発にコストがかかる．そして，入力文が対訳文パターンに適合しない場合は，翻訳ができない．

また，1990 年代に単語に基づく統計翻訳が提案された．原言語文の単語を目的言語文の単語に翻訳する手法である．しかし，翻訳精度が低い．しかし，2000 年代始めに句に基づく統計翻訳が提案された．句に基づく統計翻訳は，単語に基づく統計翻訳よりも翻訳精度が高く，学習データとして，対訳文を与えるだけで翻訳が可能である．そのため翻訳にかかるコストが低い．

一方，江木らはパターン翻訳の問題を解決するため，GIZA++[2] を利用した Pattern Based SMT[3] を提案した．この手法は対訳フレーズ辞書と対訳文パターン辞書を対訳文から自動的に作成し，翻訳を行う．対訳文から自動的に作成するので，パターン翻訳と比較して，開発コストを低くすることができる．しかし対訳文パターンに適合しても，翻訳精度の低い出力文がある．この問題の原因の一つは，句に基づく対訳文パターンの確率値に，パターン内の単語における GIZA++の値を利用していることにある．

そこで本研究では，そこで本研究では，句に基づく対訳文パターンの確率値の計算に，パターン内の変数部の確率を利用して，翻訳精度の向上を試みる．実験として，100 文の対比較実験を行い，人手による対比較評価をした．実験の結果，提案手法と従来手法は同等の翻訳精度であった．

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??章で提案する手法について説明する．第 3 章で実験データ，実験結果と評価を示す．第 5 章で本研究の考察を述べる．

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 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が提案され，単語に基づく統計翻訳に比べて翻訳精度が高いことがわかった．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 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．

2.2.3 IBM 翻訳モデル

IBM 翻訳モデルを以下に示す．これは，力久ら [4] 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．

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．

$$P(F|E) = \sum_{a} P(F, a|E) \tag{2.1}$$

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．

・モデル 1

(2.1) 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．

$$P(F, a|E) = P(m|E) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, E) P(f_j|a_1^{j}, f_1^{j-1}, m, E) \tag{2.2}$$

(2.2) 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．

$$P(F, a|E) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.3}$$

$$P(F|E) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.4}$$

$$= \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_{a_j}) \tag{2.5}$$

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，$1 \leq s \leq S$) 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．

$$c(f|e; F, E) = \frac{t(f|e)}{t(f|e_0) + \cdots + t(f|e_l)} \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=0}^{l} \delta(e, e_i) \tag{2.6}$$

$\delta(f, f_j)$ 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．

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

2. (2.7) 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．

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

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．

・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，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．

$$a(a_j|j, m, l) \equiv P(a_j|a_1^{j-1}, f_1^{j-1}, m, l) \tag{2.9}$$

この関係からモデル 1 における (2.4) 式は，以下の式に変換できる．

$$P(F|E) = \epsilon \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) a(a_j|j, m, l) \tag{2.10}$$

$$= \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_{a_j}) a(a_j|j, m, l) \tag{2.11}$$

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．

$$c(f|e; F, E) = \frac{t(f|e)}{t(f|e_0) + \cdots + t(f|e_l)} \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=1}^{l} \delta(e, e_i) \tag{2.12}$$

$$= \sum_{j=1}^{m} \sum_{i=0}^{l} \frac{t(f|e) a(i|j, m, l) \delta(f, f_j) \delta(e, e_i)}{t(f|e_0) a(0|j, m, l) + \cdots + t(f|e_l) a(l|j, m, l)} \tag{2.13}$$

$$c(i|j, m, l; F, E) = \sum_a P(a|E, F) \delta(i, a_j) \tag{2.14}$$

$$= \frac{t(f_j|e_i) a(i|j, m, l)}{t(f_j|e_0) a(0|j, m, l) + \cdots + t(f_j|e_l) a(l|j, m, l)} \tag{2.15}$$

$c(f|e; F, E)$ 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．

・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 $P(f|e)$
  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．

$$P(\phi_0|\phi_1^l,E)=\left(\begin{array}{c}\phi_1+\cdots+\phi_l\\ \phi_0\end{array}\right)p_0^{\phi_1+\cdots+\phi_l-\phi_0}p_1^{\phi_0} \tag{2.16}$$

したがって，モデル3は以下の式で求められる．

$$\begin{aligned} P(F|E) &= \sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}P(F,a|E) && (2.17)\\ &= \sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\left(\begin{array}{c}m-\phi_0\\ \phi_0\end{array}\right)p_0^{m-2\phi_0}p_1^{\phi_0}\prod_{i=1}^{l}\phi_i!n(\phi_i|e_i)\\ &\times \prod_{j=1}^{m}t(f_j|e_{a_j})d(j|a_j,m,l) && (2.18)\end{aligned}$$

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．

・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 $d(j|i.m,l)$ 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

$$P(\Pi_{[i]1} = j|\pi_1^{[i]-1}, \tau_0^l, \phi_0^l, E) = d_1(j - \odot_{i-1}|\mathcal{A}(e_{[i-1]}), \mathcal{B}(f_j)) \tag{2.19}$$

  $\odot_{i-1}$ 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

$$P(\Pi_{[i]k} = j|{\pi_{[i]1}}^{k-1}, \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, E) = d_{>1}(j - \pi_{[i]k-1}|\mathcal{B}(f_j)) \tag{2.20}$$

  $\pi_{[i]k-1}$ 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．

・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

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

2.2.4 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

	He	went	to	kyoto	on	business
彼	●					
は						
仕事						●
で					●	
京都				●		
に						
行っ		●				
た						

表 2.4: 日英方向の単語対応

	He	went	to	kyoto	on	business
彼	●					
は			●			
仕事						●
で					●	
京都				●		
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．

2.2.5 GIZA++

GIZA++ とは，統計翻訳で用いることを前提に作られたツールである．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．

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.

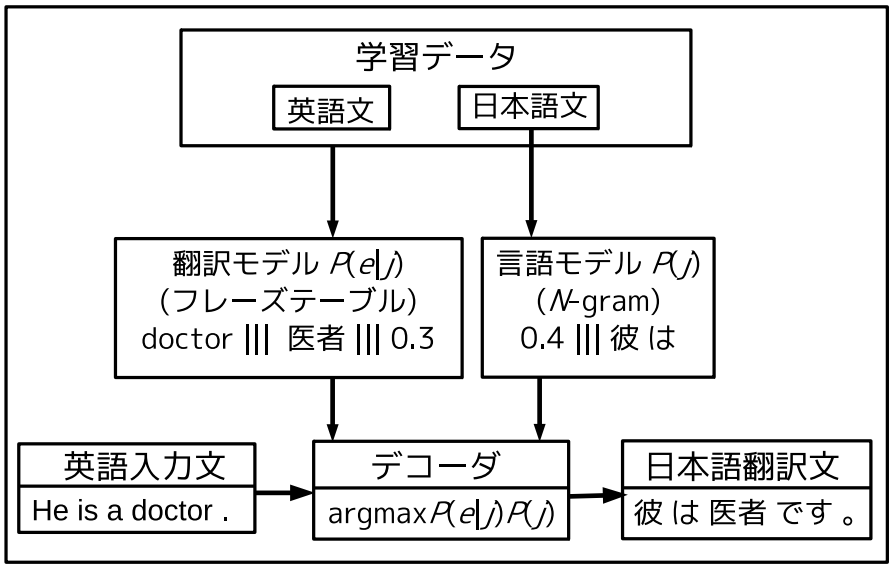

図 2.1: 日英統計翻訳の枠組み

$$E = argmax_j P(e|j) \tag{2.21}$$

$$\simeq argmax_j P(j|e)P(e) \tag{2.22}$$

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．以下にフレーズテーブルの例を示す.

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

	He	went	to	kyoto	on	business
彼	●					
は			●			
仕事						●
で					●	
京都				●		
に		●			●	
行っ		●				
た		●				

表 2.6: 英日方向の単語対応

	He	went	to	kyoto	on	business
彼	●					
は						
仕事						●
で					●	
京都				●		
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．

表 2.7: intersection の例

	He	went	to	kyoto	on	business
彼	●					
は						
仕事						●
で					●	
京都				●		
に						
行っ		●				
た						

表 2.8: union の例

	He	went	to	kyoto	on	business
彼	●					
は			●			
仕事						●
で					●	
京都				●		
に		●			●	
行っ		●				
た		●				

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” の例を表 2.9 に示す．

表 2.9: grow-diag の例

	He	went	to	kyoto	on	business
彼	●					
は						
仕事						●
で					●	
京都				●		
に		●			●	
行っ		●				
た		●				

“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 2.10 に示す.

表 2.10: grow-diag-final-and の例

	He	went	to	kyoto	on	business
彼	●					
は			●			
仕事						●
で					●	
京都				●		
に		●			●	
行っ		●				
た		●				

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.

2.3.3 言語モデル

言語モデルとは，人間が用いる言葉の自然な並びを確率としてモデル化したものであり，膨大な量の単言語データを用いて単語の列や文字の列が起こる遷移確率を付与したものである．統計翻訳では主に N-gram を用いる．言語モデルを式 2.23 に示す.

$$\mathrm{P}(\mathrm{w}_i|w_{i-1}) = P(w_1)P(w_2|w_1)P(w_3|w_2w_1)\cdots \quad (2.23)$$

2.3.4 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” [6] 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．

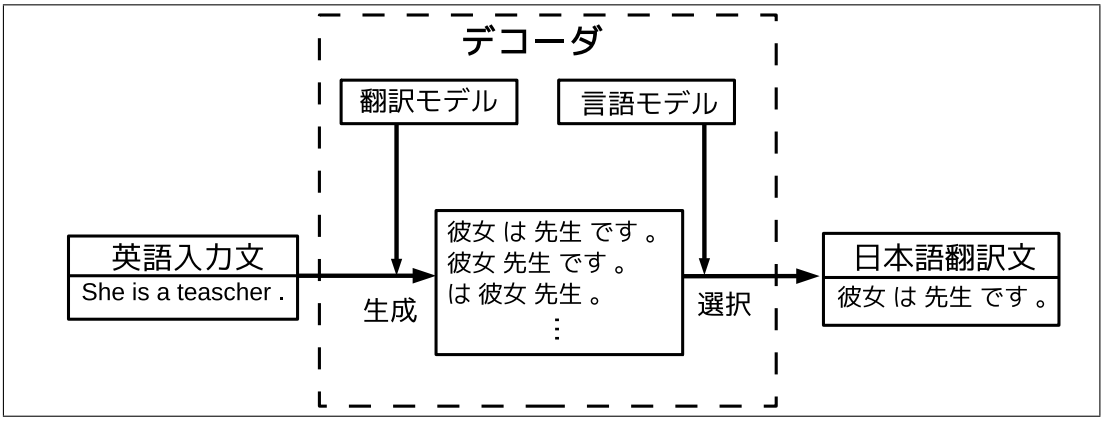

図 2.2: デコーダの動作例

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.

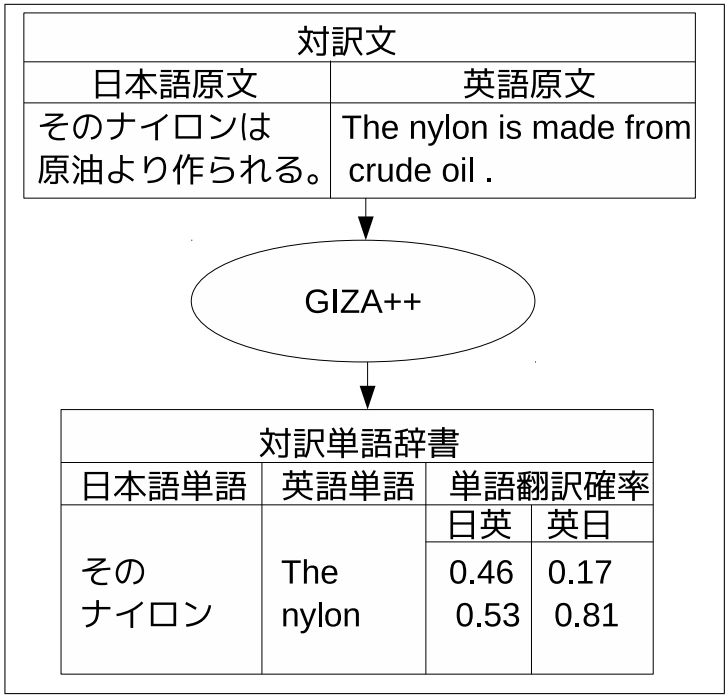

図 2.3: 対訳単語辞書の作成

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．

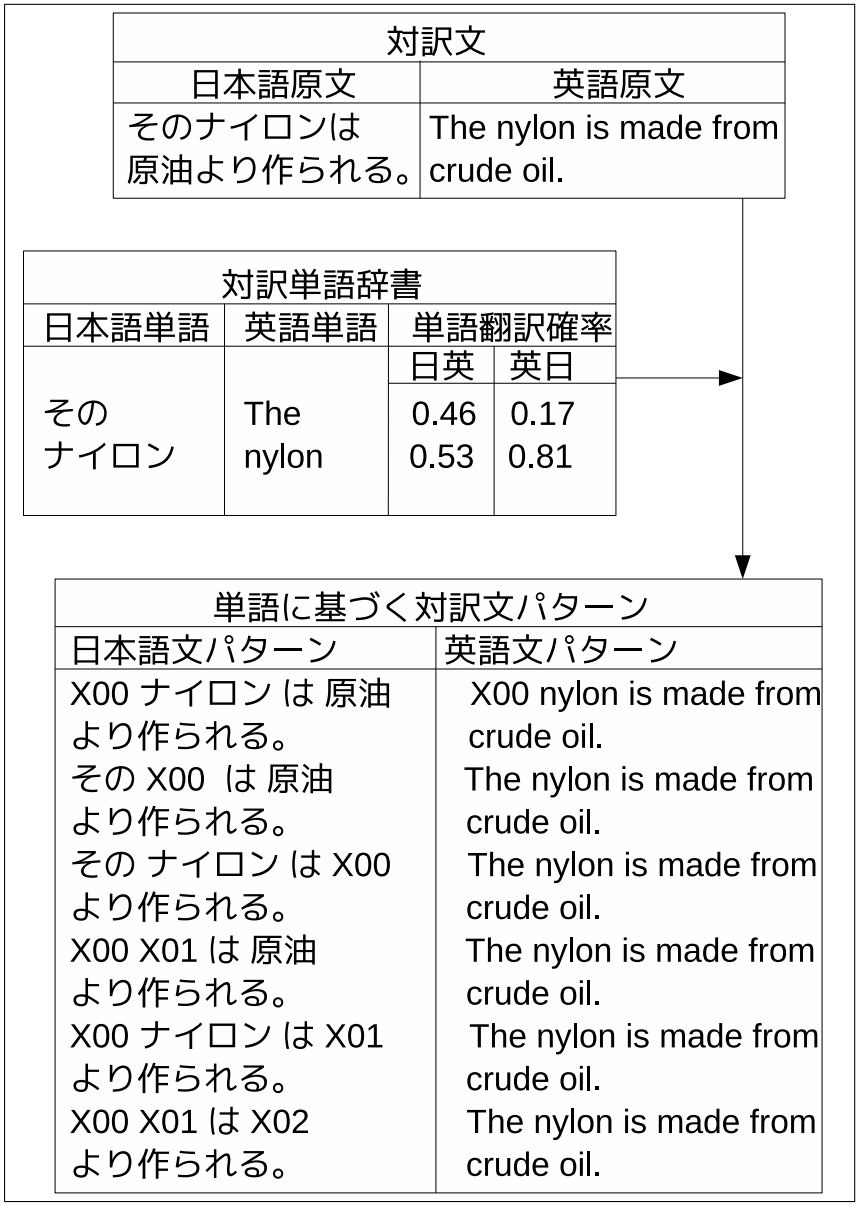

図 2.4: 単語に基づく対訳文パターンの作成

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.

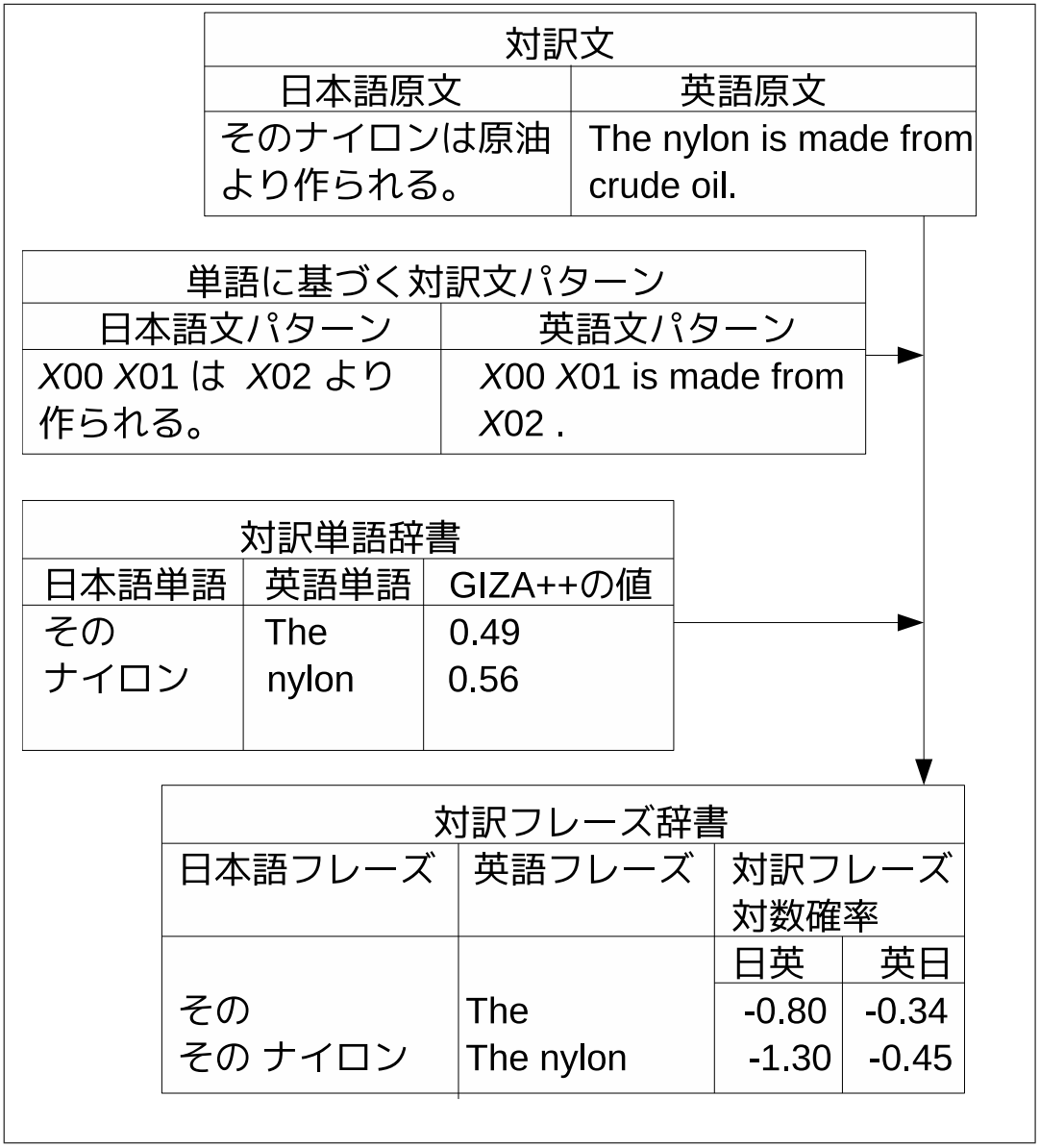

図 2.5: 対訳フレーズ辞書の作成

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.

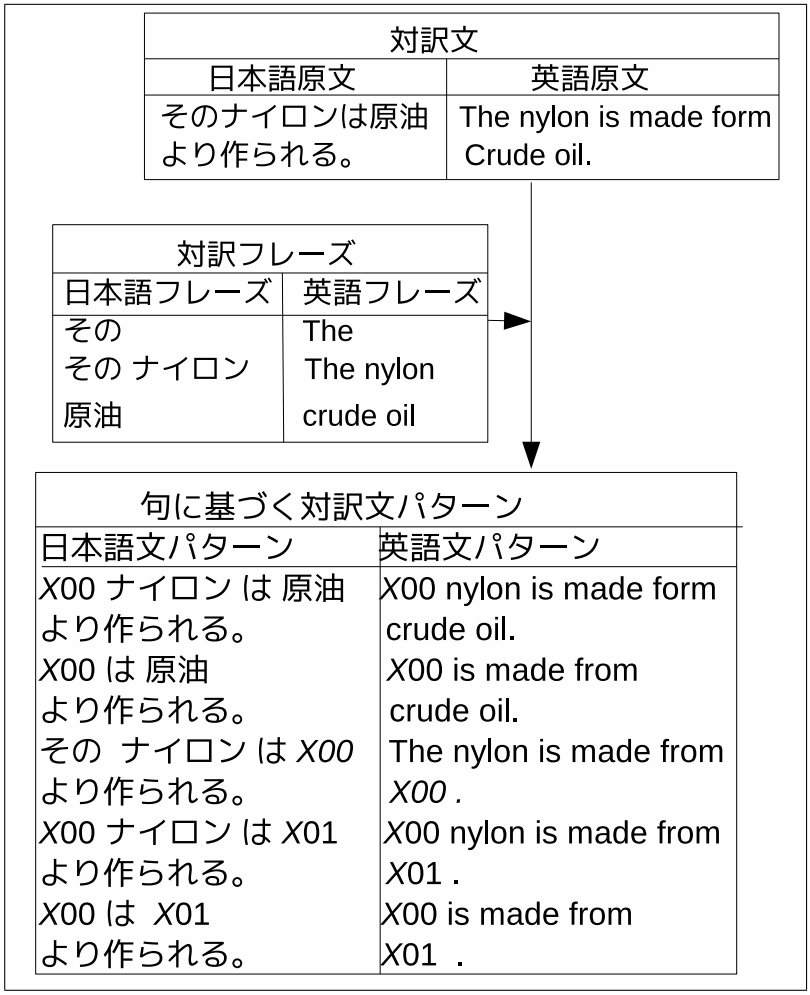

図 2.8: 句に基づく対訳文パターンの作成

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．

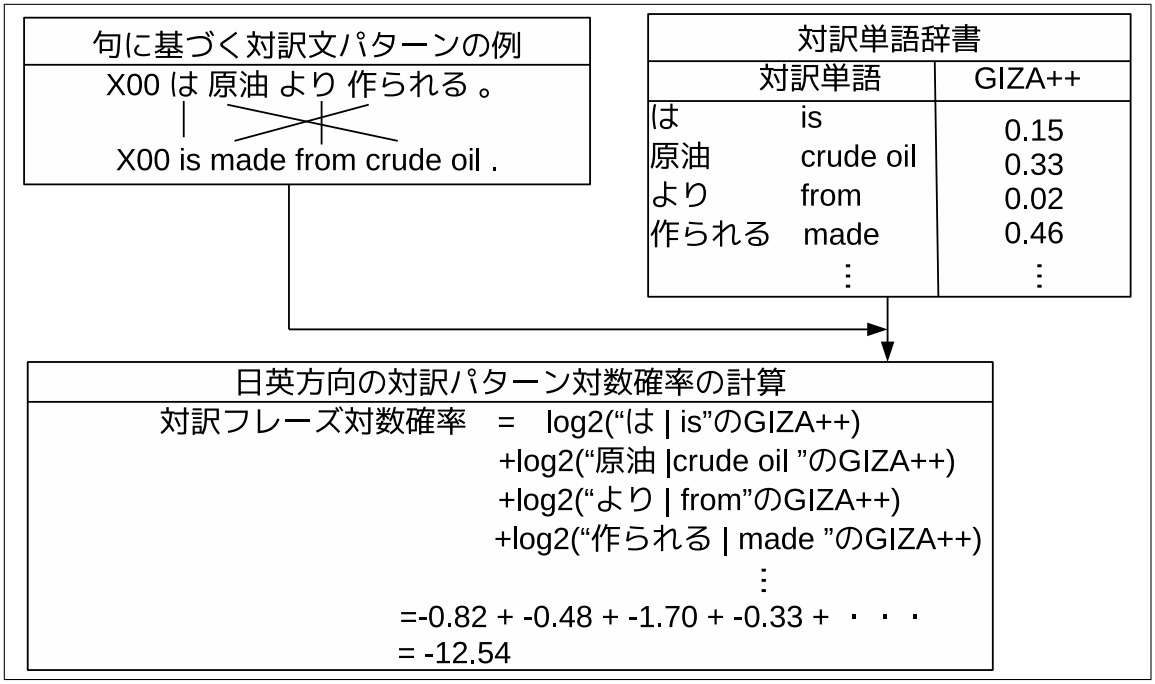

図 2.9: 日英方向の対訳文パターン対数確率の付与

$$log_2 P(\frac{J_0 \cdots J_{N-1}, JX_0 \cdots JX_{N-1}}{E_0 \cdots E_{M-1}, EX_0 \cdots EX_{M-1}}) =$$

$$\sum_{n=0}^{N-1} \arg\max_{m=0}^{M-1} (log_2(p(E_m|J_n)) + log_2(p(J_n|E_m))) \quad (2)$$

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)

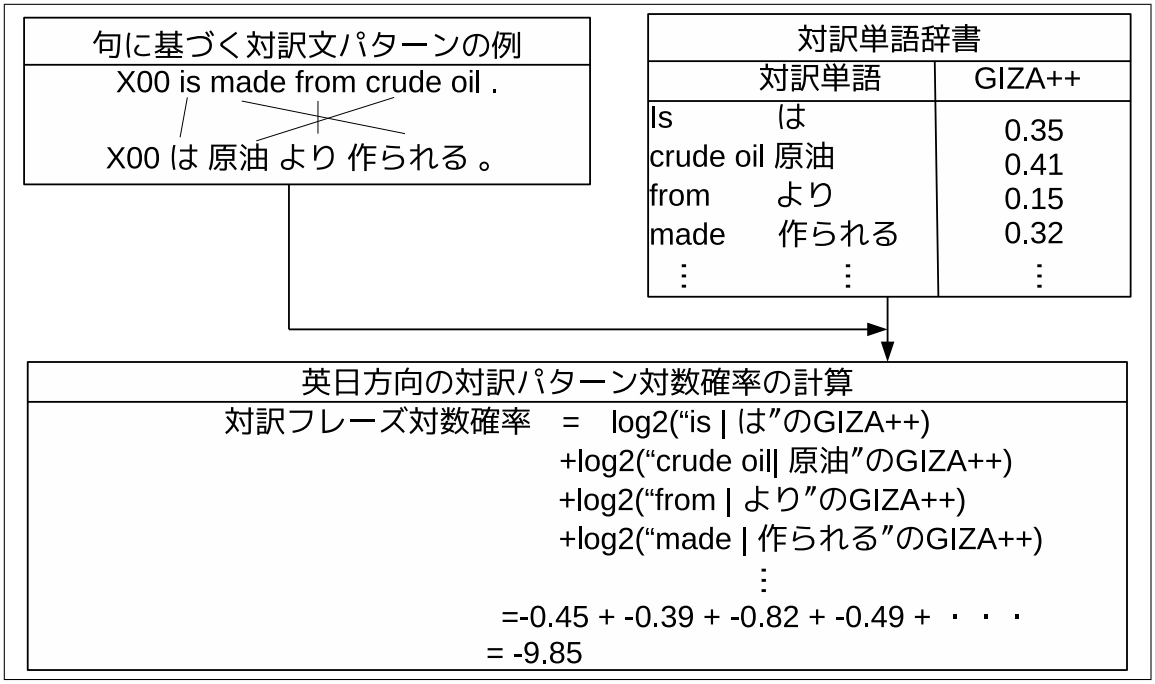

図 2.10: 英日方向の対訳文パターン対数確率の付与

$$log_2 P(\frac{J_0 \cdots J_{N-1}, JX_0 \cdots JX_{N-1}}{E_0 \cdots E_{M-1}, EX_0 \cdots EX_{M-1}}) =$$

$$\sum_{n=0}^{N-1} \arg\max_{m=0}^{M-1} (log_2(p(J_n|E_m)) + log_2(p(E_m|J_n))) \quad (3)$$

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)

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

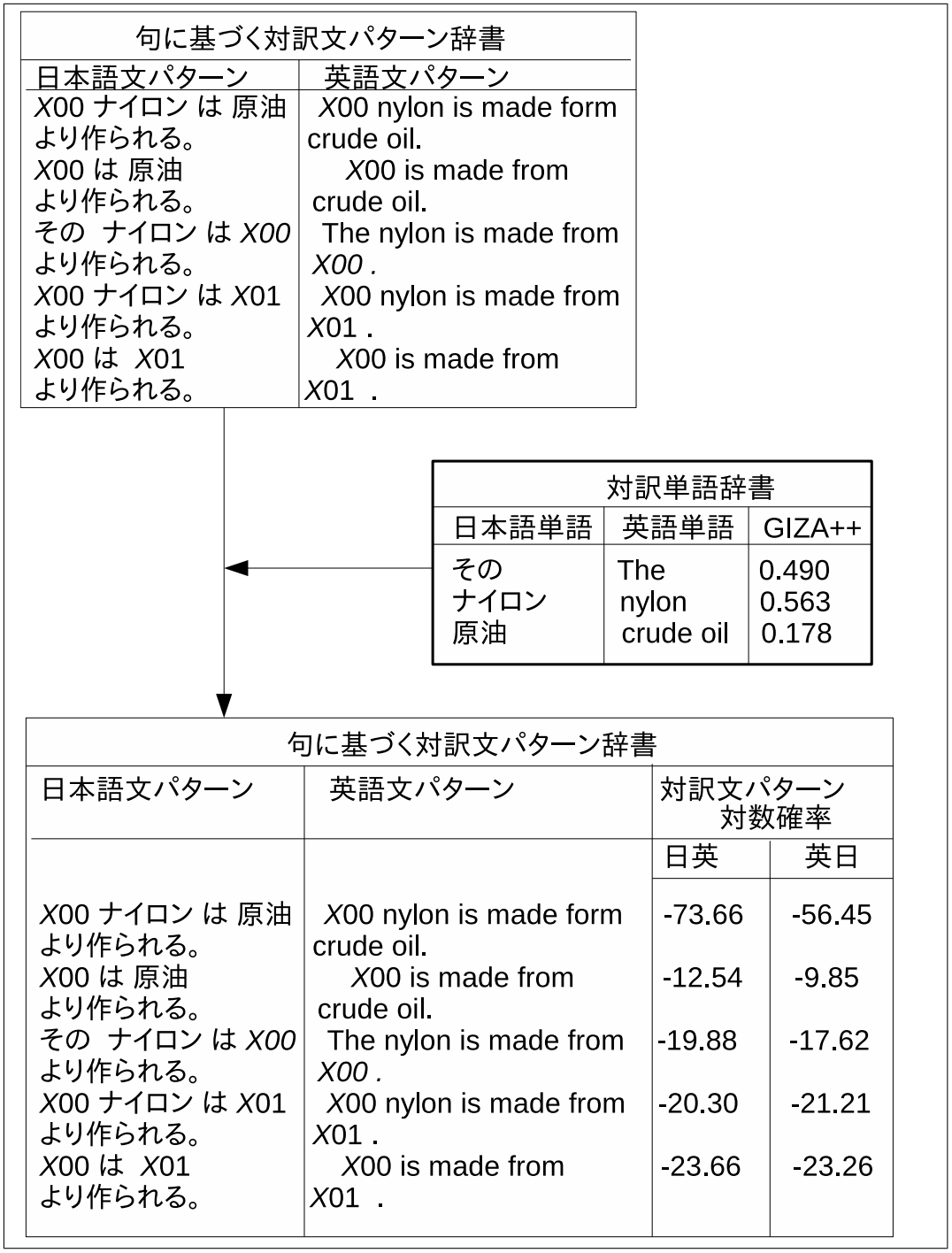

図 2.11: 句に基づく対訳文パターン辞書の作成

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．

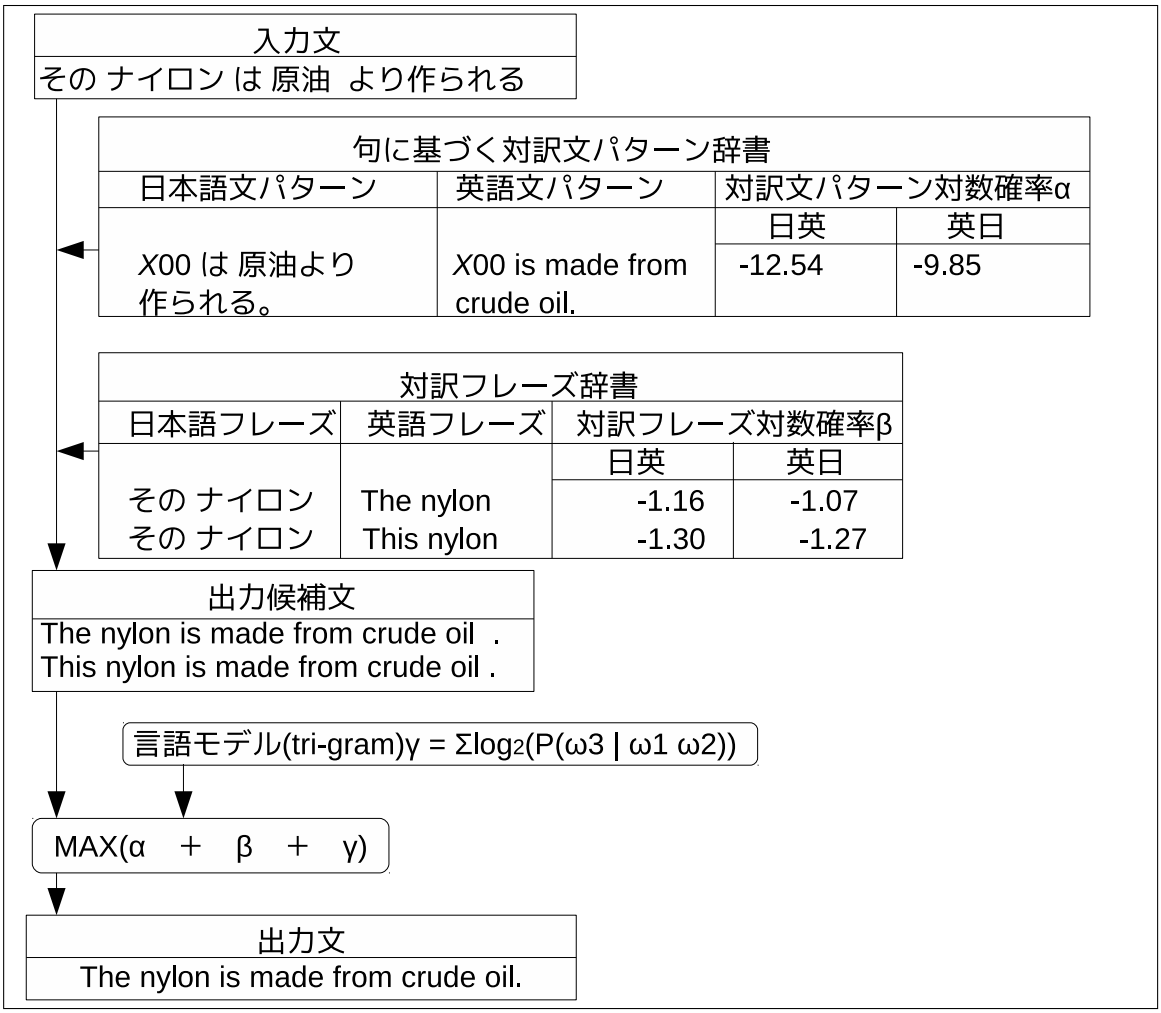

図 2.12: 出力文生成の流れ

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．

第3章　提案手法

3.1　提案手法の概要

Pattern Based SMT において人手評価が低い原因の一つは，句に基づく対訳文パターンの確率値の計算に，人が見て明らかに妥当ではない対応を利用しているためである．パターン内の単語における GIZA++の値の中には，人が見て明らかに妥当ではない対応がある．そのため，対訳文パターンの確率値に信頼性がなく，翻訳精度が低下していると考える．

そこで，句に基づく対訳文パターンの確率値の計算に，対訳フレーズ対数確率を利用する．対訳フレーズ対数確率は GIZA++の値が高い場所が選ばれる傾向にある．そして，一般に GIZA++の値が高いというのは，信頼性が高い．よって，本研究では，句に基づく対訳文パターンの確率値の計算に，対訳フレーズ対数確率を利用することで，翻訳精度の向上を目指す.

3.1.1　Pattern Based SMT の問題点

Pattern Based SMT の出力文には，人手評価の低い出力文がある．その原因として，句に基づく対訳文パターン辞書の確率値の計算方法が挙げられる．対訳文パターンの確率値の計算は，パターン内の単語における GIZA++の値を利用している．しかし，パターン内の単語における GIZA++の値の中には，明らかに妥当ではない対応を取っているものがある．そのため，対訳文パターン対数確率の値に信用性がないと考える．
明らかに妥当ではない対応を取っている出力文の例を表 3.1 に示す.

表 3.1: 明らかに妥当ではない対応を取っている出力文の例

入力文	最終 的 に 条件 面 で 合意 を み た 。
参照文	The terms were finally agreed to .
日本語文パターン	N02 に N00 N04 で N03 を N01 た 。
英語文パターン	I N 01 a N 03 to N 02 by N 00 N 04 .
日本語文パターンの原文	彼女 に 小包 郵便 で 本 を 送っ た 。
英語文パターンの原文	I sent a book to her by parcel post .
出力文	I read a agreement to Ultimately by terms sides .
「を」と「I」の対応	-8.745
「た」と「I」の対応	-3.822

Pattern Based SMT では，対訳文パターンの確率値の計算に字面を利用している．表 3.1 の例では，字面である「を」が「I」になる確率が一番高いので，-8.745 が選択されている．しかし,「を」と「I」という対応は，人が見て明らかに妥当ではない対応であることがわかる．

3.2 句に基づく文パターン辞書の作成

本研究では，句に基づく対訳文パターンの確率値の計算に，対訳フレーズ対数確率を利用することで，翻訳精度の向上を目指す．

以下に，提案手法における対訳文パターンの確率値の計算方法を式 (3) に示す．また，図 3.1 に，提案手法の具体的な計算例を示す

$$log_2 P(\frac{J_0 \cdots J_{N-1}, JX_0 \cdots JX_{N-1}}{E_0 \cdots E_{M-1}, EX_0 \cdots EX_{M-1}}) =$$

$$\sum_{n=0}^{N-1}(log_2(p(JX_n|EX_n))) \quad (3)$$

JX_n; 対訳文パターン中の日本語の対訳フレーズ

EX_n; 対訳文パターン中の英語の対訳フレーズ

$p(JX_n|EX_n)$; 英語フレーズ EX_n が日本語フレーズ JX_n に翻訳される確率 (対訳フレーズ確率)

N; 対訳フレーズの数

なお，対訳フレーズ対数確率の計算方法を式 (4) に示す．

$$log_2 P(\frac{J_0 \cdots J_{N-1}}{E_0 \cdots E_{M-1}}) =$$

$$\sum_{n=0}^{N-1}(log_2(p(J_n|E_m)) + log_2(p(E_m|J_n))) \quad (4)$$

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)

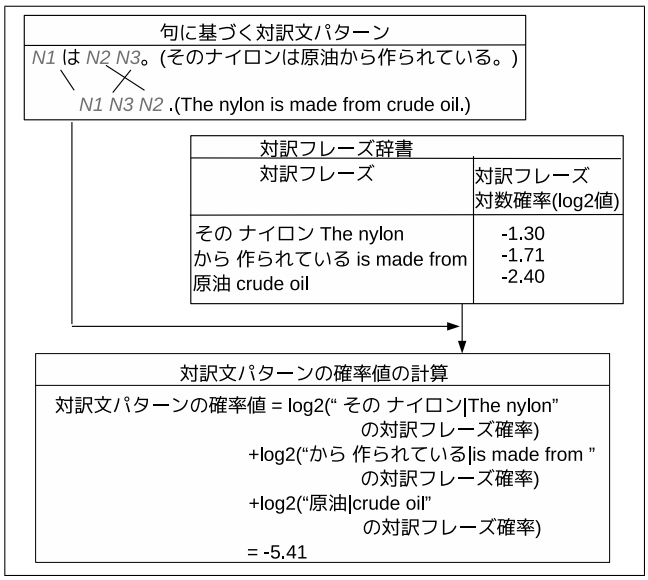

図 3.1: 提案手法の具体例

第4章　実験

4.1　実験データ

対訳文および翻訳実験に用いる入力文として電子辞書から抽出した単文データを用いる [7]．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．コーパスの内訳を表 4.1 に示す．

表 4.1: 実験データ

対訳文	100,000 文対
入力文	100 文

対訳文および，入力文の例を表 4.2 と表 4.3 に示す．

表 4.2: 対訳文の例

対訳文	
日本語原文	英語原文
英語 で は 私 は 彼 に 遠く 及ば な い 。	He is far superior in English to me .
この 町 から 悪 を 一掃 しよ う 。	Let's eradicate vice from this town .
心 は 経験 によって 育つ 。	The mind expands with experience .

表 4.3: 入力文の例

入力文	
日本語文	参照文
彼 の 姿 は 暗闇 の 中 で 見え なかっ た 。	He was hidden by the darkness .
私 は いつも 辞書 を 手近 に 置い て いる 。	I always keep a dictionary at hand .
子供 たち は 寝室 へ 立ち去っ た 。	The children disappeared to their bedrooms .

4.2　実験結果

提案手法，従来手法を用いて，翻訳実験を行う．実験の結果，入力文 100 文中，出力文 100 文を得た．

4.2.1 対比較実験

提案手法と従来手法の対比較評価を行った．従来手法には，2 章の Pattern Based SMT を用いる．評価方法は，提案手法と従来手法を伏せた状態でランダムに出力文を表示し，どちらが優れているか評価する．結果を表 4.4 に，評価例を表 4.5，表 4.6，表 4.7，表 4.8，表 4.9，表 4.10 に示す．

表 4.4: 人手による評価

提案手法○	従来手法○	差なし	同一出力
12	11	23	54

4.2.2 提案手法○の例

表 4.5: 提案手法○の例 1

入力文		彼 は その 本 を 私 の かばん の 中 に 押し込ん だ。
参照文		He shoved the book into my bag .
提案手法	日本語文パターン	*N04* は *N03 N00 N02* の 中 に *N01* だ 。
	英語文パターン	*N04 N01 N03* into *N00 N02* .
	日本語原文	彼ら は 彼 を 城 の 中 に 押し込ん だ。
	英語原文	They jostled him into the castle .
	N00	私 の my
	N01	押し込ん stuffed
	N02	かばん bag
	N03	その 本 を the book
	N04	彼 He
	パターンの確率	-17.290790
	変数の確率	-26.922323
	言語モデルの確率	-121.664923
	出力文	He stuffed the book into my bag .
従来手法	日本語文パターン	*N04* は *N03 N02* かばん の 中 に *N00 N01* 。
	英語文パターン	*N02 N00 N01 N03* in *N04* bag .
	日本語原文	彼 は 財布 を かばん の 中 に しまっ た 。
	英語原文	He put his wallet in his bag .
	N00	押し込ん stuffed
	N01	だ is
	N02	私 の I
	N03	その 本 を the book
	N04	彼 his
	パターンの確率	-18.463365
	変数の確率	-38.375487
	言語モデルの確率	-104.999003
	出力文	I stuffed is the book in his bag .

表 4.5 の例は，提案手法は，文全体が参照文と比べて，概ね合っている．また，変数の対応も人が見て妥当な対応であることがわかる．従来手法は，動詞が明らかにおかしい．これは，パターンのマッチングが問題である．従来手法のパターンの変数の対応を

見ると，*N01*が「た」と「his」，*No2*が「を」と「he」の対応を取っている．これは，人が見て明らかに妥当ではないことがわかる．よって，従来手法のパターンに問題があり，提案手法の出力文の方が優れていると評価した．

表 4.6: 提案手法○の例 2

入力文		その 飛行機 に は 何 の マーク も つい て い なかっ た 。
参照文		There were no markings on the plane .
提案手法	日本語文パターン	*N03* は 何 の *N02* も *N01 N00* た 。
	英語文パターン	*N03 N01 N00 N02* .
	日本語原文	犯人 は 何 の 手がかり も 残さ なかっ た 。
	英語原文	The criminal left no clue .
	N00	て い なかっ not
	N01	つい took
	N02	マーク Mark
	N03	その 飛行機 に The plane
	パターンの確率	-50.856352
	変数の確率	-51.369999
	言語モデルの確率	-143.609640
	出力文	The plane took not Mark .
従来手法	日本語文パターン	*N03 N01 N02* 何 の *N04* も *N00* なかっ た 。
	英語文パターン	*N03 N00* no *N04 N02 N01* .
	日本語原文	彼 は それ について 何 の 批評 も し なかっ た 。
	英語原文	He made no comment on it .
	N00	つい て い was on
	N01	飛行機 plane
	N02	に は the
	N03	その I
	N04	マーク Mark
	パターンの確率	-35.247032
	変数の確率	-72.108856
	言語モデルの確率	-143.694223
	出力文	I was on no Mark the plane .

表 4.6 の例は，提案手法は，文全体が参照文と比べて，概ね合っている．また，変数の

対応も人が見て妥当な対応であることがわかる．従来手法は，文法的に間違っていることと，意味が全く通らない．また，変数においても,「その」が「I」に対応している．これは，人が見て明らかに妥当ではない対応であることがわかる．よって，従来手法の変数に問題があり，提案手法の出力文の方が優れていると評価した．

表 4.7: 提案手法○の例 3

入力文		君 の 行動 は 常識 を はずれ て いる 。
参照文		Your actions are not in accordance with common sense .
提案手法	日本語文パターン	*N01 N02* は *N03 N00* はずれ て いる 。
	英語文パターン	*N01 N02* is out *N00 N03* .
	日本語原文	その ピアノ は 調子 が はずれ て いる 。
	英語原文	The piano is out of tune .
	N00	を of the
	N01	君 の Your
	N02	行動 act
	N03	常識 common
	パターンの確率	-17.411779
	変数の確率	-52.356724
	言語モデルの確率	-63.620014
	出力文	Your act is out of the common .
従来手法	日本語文パターン	*N00* の *N02* は *N03 N01 N04* て いる 。
	英語文パターン	*N00 N02* is *N04 N01 N03* .
	日本語原文	あなた の 考え は 私 と 似 て いる 。
	英語原文	Your idea is similar to mine .
	N00	君 Your
	N01	を in a
	N02	行動 act
	N03	常識 sense
	N04	はずれ off
	パターンの確率	-16.397056
	変数の確率	-38.531129
	言語モデルの確率	-68.963852
	出力文	Your act is off in a sense .

表 4.7 の例は，提案手法は，概ね意味がわかるものである．また，変数の対応も人が

見て妥当であると判断できる．これに対して，従来手法は全体を通して意味がわからない．しかし，変数の対応で明らかに妥当ではない対応がない，また，パターンのマッチングにも問題が見られない．この場合は出力結果が良い方を選択している．よって，提案手法の方が優れていると評価した．

4.2.3 従来手法○の例

表 4.8: 従来手法○の例 1

入力文		廊下 を じっと 見つめ た。
参照文		She peered hard down the corridor .
提案手法	日本語文パターン	*N00 N01* じっと 見つめ た 。
	英語文パターン	He stared at me hard *N01 N00* .
	日本語原文	一瞬 私 を じっと 見つめ た 。
	英語原文	He stared at me hard for a moment .
	N00	廊下 corridor
	N01	を on the
	パターンの確率	-32.988423
	変数の確率	-11.177805
	言語モデルの確率	-37.261191
	出力文	He stared at me hard on the corridor .
従来手法	日本語文パターン	*N01* を じっと 見つめ *N00* 。
	英語文パターン	*N00* gazed at the *N01* .
	日本語原文	遠く の 陸 を じっと 見つめ た 。
	英語原文	We gazed at the distant shore .
	N00	た I
	N01	廊下 hall
	パターンの確率	-22.140954
	変数の確率	-12.071409
	言語モデルの確率	-28.290798
	出力文	I gazed at the hall .

表 4.8 の例では，従来手法は，概ね意味がわかるものである．しかし，変数の対応において,「た」が「I」になっている．これは，人が見て明らかに妥当ではない対応である．出力文が良くなっている原因としては，従来手法は，パターンの字面を計算に利用して

いる．そのため，字面の対応関係がある程度よかったためだと考えられる．また，提案手法のパターンを見ると，*N00* は,「私を」が「for」に対応していることがわかる．よって，提案手法のパターンが問題となり，従来手法が良くなったと考えられる．

表 4.9: 従来手法○の例 2

入力文		彼 は 私 の 注意 を 引い た 。
参照文		He attracted my attention .
提案手法	日本語文パターン	彼 は 私 の *N01 N00 N02* た 。
	英語文パターン	He *N02 N00 N01* .
	日本語原文	彼 は 私 の 誇り を 傷つけ た 。
	英語原文	He hurt my pride .
	N00	を his
	N01	注意 attention
	N02	引い drew
	パターンの確率	-17.881104
	変数の確率	-14.232825
	言語モデルの確率	-18.586920
	出力文	He drew his attention .
従来手法	日本語文パターン	彼 は *N02* の *N00* を *N01* た 。
	英語文パターン	He *N01* his *N00* to *N02* .
	日本語原文	彼 は 辞職 の 決意 を 示し た 。
	英語原文	He indicated his resolve to resign .
	N00	注意 attention
	N01	引い drew
	N02	私 me
	パターンの確率	-22.475558
	変数の確率	-11.384461
	言語モデルの確率	-25.077716
	出力文	He drew his attention to me .

表 4.9 の例では，従来手法は，概ね意味がわかるものである．また，変数の対応関係も妥当である．そして，パターンのマッチングも妥当である．提案手法は、*N00* の「を」が「his」という対応が人が見て妥当ではないと判断できる．よって，*N00* の変数の対応に問題があり，翻訳精度が低くなってしまったと考えられる．

表 4.10: 従来手法○の例 3

入力文		仕事 の 合間 に テレビ を 見 た。
参照文		I watched television between work .
提案手法	日本語文パターン	*N03* の *N02* に *N01* を *N00* た 。
	英語文パターン	*N00 N01 N03 N02* .
	日本語原文	彼女 の 口 に キス を し た。
	英語原文	He kissed her mouth .
	N00	見 Look
	N01	テレビ TV
	N02	合間 intervals
	N03	仕事 work
	パターンの確率	-21.757533
	変数の確率	-22.270473
	言語モデルの確率	-108.024678
	出力文	Look TV work intervals .
従来手法	日本語文パターン	*N03* の *N02* に *N01* を *N00* た 。
	英語文パターン	I *N00* a *N01* for my *N03 N02* .
	日本語原文	誕生 日 の プレゼント に セーター を もらっ た 。
	英語原文	I got a sweater for my birthday present .
	N00	見 saw
	N01	テレビ TV
	N02	合間 intervals
	N03	仕事 work
	パターンの確率	-26.768407
	変数の確率	-12.919711
	言語モデルの確率	-119.287604
	出力文	I saw a TV for my work intervals .

表 4.10 の例では，従来手法は，概ね意味がわかるものである．これに対して，提案手法は文を通して意味がわからない．この原因として，まず変数を見ると，提案手法，従来手法共に変数の対応は妥当だと判断できる．また，従来手法のパターンも適切であることがわかる．一方，提案手法のパターンを見ると，パターンの *N00* の「し」が「He」に対応していることがわかる．よって，パターンのマッチングが原因だと考えられる．また，提案手法は，英語のパターンに字面が無いのに対して，従来手法は字面が残ってい

る．そして，従来手法は字面を利用して計算している．そのため，そこで差が出て従来手法の方が良くなったと考えられる．

表 4.11: 差なしの例 1

入力文		この 病気 の 治療 は 長びく 。
参照文		The disease requires lengthy treatment .
提案手法	日本語文パターン	この *N00* の *N02* は *N01* 。
	英語文パターン	This *N00 N02 N01* .
	日本語原文	この おもちゃ の 車 は ぜんまい で 動きます 。
	英語原文	This toy car moves by a spring .
	出力文	This disease cure 長びく .
従来手法	日本語文パターン	この *N00* の *N01* は *N02* 。
	英語文パターン	The *N00 N01 N02* .
	日本語原文	この 2つ の 言語 は 言語 学 上 その 起源 は 同じ で ある 。
	英語原文	The two languages have a common linguistic parent .
	出力文	The disease cure 長びく .

表 4.12: 差なしの例 2

入力文		所定 の 寸法 に 平ら な 表面 を フライス で 削る 。
参照文		Mill flat surface to size .
提案手法	日本語文パターン	*N06 N02 N05* に *N07* な *N03* を *N04 N00 N01* 。
	英語文パターン	*N01 N06 N04 N03 N00 N07 N02 N05* .
	日本語原文	弁護士 は 被告 に 有利 な 証拠 を 提出 した 。
	英語原文	The lawyer produced evidence in favor of the accused .
	出力文	Surface overran フライス surface in the flat of metric .
従来手法	日本語文パターン	*N06 N02 N05* に *N07* な *N03* を *N04 N00 N01* 。
	英語文パターン	*N06 N04 N03 N01 N07 N00 N02 N05* .
	日本語原文	弁護士 は 被告 に 有利 な 証拠 を 提出 した 。
	英語原文	The lawyer produced evidence in favor of the accused .
	出力文	Everybody フライス surface Surface flat on the metric .

表 4.13: 差なしの例 3

入力文		こういう 事態 を うまく 収拾 でき ない 。
参照文		He can't cut this kind of situation well .
提案手法	日本語文パターン	*N03 N02* を うまく *N01 N00* 。
	英語文パターン	*N03 N01 N00 N02* .
	日本語原文	彼 は 質問 を うまく はぐらかし た 。
	英語原文	He dodged the question .
	出力文	Such 収拾 not serious .
従来手法	日本語文パターン	*N03 N02 N01 N04 N00* でき ない 。
	英語文パターン	*N03* cannot *N00 N01 N02 N04* .
	日本語原文	彼 は 今 の 容体 で は 旅行 は でき ない 。
	英語原文	He cannot travel in the present condition .
	出力文	Such cannot 収拾 the situation well .

第5章　考察

5.1　提案手法の有効性

Pattern Based SMT において人手評価が低い原因の一つは，句に基づく対訳文パターンの確率値の計算に，パターン内の単語における GIZA++の値を利用しているためである．人が見て明らかに妥当ではない値がある．そのため，対訳文パターンの確率値に信頼性がなく，翻訳精度が低下していると考える．

そこで，句に基づく対訳文パターンの確率値の計算に，対訳フレーズ対数確率を利用する．対訳フレーズ対数確率は GIZA++の値が高い場所が選ばれる傾向にある．そして，一般に GIZA++の値が高いというのは，信頼性が高い．よって，本研究では，句に基づく対訳文パターンの確率値の計算に，対訳フレーズ対数確率を利用することで，翻訳精度の向上を目指す．

人手評価の結果，提案手法の有効性が確認できなかった．提案手法○の数とベースライン○の数がほぼ同等であったため，提案手法の計算方法と従来手法の計算方法では，同様の効果であると考えられる．

5.2　moses と提案手法における対比較実験

本研究では，対訳文パターン対数確率の計算に対訳フレーズ対数確率を利用することで，従来手法と同等の精度が得られることがわかった．ここで，moses[6] と提案手法における 100 文の対比較実験の結果を示す．また，moses の実験データを表 5.1 に示す．

表 5.1: 実験データ

N-gram	5
alignment	intersection
学習文	100,000 文対
dev データ	100,000 文対
テストデータ	100 文

5.2.1 mosesでの対比較実験

表 5.2 に提案手法と moses で得られた翻訳結果 100 文での対比較実験の結果を示す.

表 5.2: moses での対比較実験

提案手法○	moses ○	同一出力	差なし
27	18	2	53

5.2.2 mosesが提案手法より優れている例

表 5.2 における moses が提案手法より優れている例を表 5.3 に示す.

表 5.3: moses が提案手法より優れている例の例

日本語入力文	大 火災 が 起き た 。
正解文	A major fire broke out
英語翻訳文 (提案)	There happened great fire .
英語翻訳文 (moses)	A big fire broke out .

5.2.3 提案手法が moses より優れている例

表 5.2 における提案手法が moses より優れている例を表 5.4 に示す.

表 5.4: 提案手法が moses より優れている例

日本語入力文	お先に 失礼 し ます 。
正解文	Excuse me , I must be going now .
英語翻訳文 (提案)	I'm sorry , but I must be leaving now .
英語翻訳文 (moses)	You go first .

5.3 翻訳精度の問題

3 章より，提案手法は従来手法と同等の翻訳精度であることがわかった．moses との対比較実験を行ったところ，提案手法○を 27 文，従来手法○を 18 文得た．表 5.2 の結果より，提案手法が moses より優れているという結果になった.

5.4 対訳文パターンの新しい計算方法

本研究では，対訳文パターン対数確率の計算に対訳フレーズ対数確率を利用し，翻訳精度の調査を行った．対比較実験の結果から，提案手法の計算方法は，従来手法の計算方法と同等の精度が得られた．しかし，翻訳精度の向上は見られなかった．その理由として，以上の原因が挙げられる．日本語パターンを J_1, JX_1, J_2，英語パターンを E_1, EX_1, E_2 とした場合，従来手法は，以下の式で行う．

$P(\frac{J_1 JX_1 J_2}{E_1 EX_1 E_2})$

$= \arg\max((P\frac{J_1}{E_1}, (P\frac{J_1}{E_2})) \times \arg\max((P\frac{J_2}{E_1}), (P\frac{J_2}{E_2}))$

つまり，$P(J_1/E_1 EX_1 E_2)$ の値として $P(J_1/E_1)$, $P(J_1/E_2)$ の最大値を選択している．しかし，$P(J_1/E_1)$，$P(J_1/E_2)$ が大きな値を持つ場合がある．この場合，加算した方が妥当性があると思われる．つまり，以下の式を用いる．

$(P(J_1/E_1) + P(J_1/E_2)) \times (P(J_2/E_1) + P(J_2/E_2))$

この方法を試みたい．

第6章　おわりに

本研究では，Pattern Based SMT において，句に基づく対訳文パターンの確率値に，パターン内の単語における GIZA++の値を利用せずに，句に基づく対訳文パターンの確率値の計算に，パターン内の変数部の確率を利用して，翻訳精度の向上を試みた..実験結果より，提案手法は，従来手法と同等な精度が得られた．提案手法と moses での対比較実験の結果より，提案手法の方が優れていた，以上より，提案手法は従来手法と同等の精度の計算方法であると言える．今後は，新しい計算方法を試みたい．

謝辞

本研究を進めるにあたり，研究の説明や論文の書き方など様々なご指導を頂きました鳥取大学工学部知能情報工学科計算機工学 C 講座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，村田真樹教授に心から御礼申し上げます．また，計算機工学 C 講座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] 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.

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

[6] 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.

[7] 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.