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 Pattern Based 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，不適切な対訳文パターンの選択であった．そこで本研究では，日英 Pattern Based SMT において，対訳文パターンの日本語原文と入力文とのレーベンシュタイン距離 [4](以下 LsD) を求める．この距離を利用して，対訳文パターンの日本語原文と入力文が類似した対訳文パターンの選択を行い，翻訳精度の向上を目指した．実験の結果，入力文 100 文中出力文 81 文を取得．人手による対比較評価をした結果，提案手法の翻訳精度の向上はあまり見られなかった．

目 次

表 目 次

図 目 次

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，不適切な対訳文パターンの選択である．

そこで本研究では，日英 Pattern Based SMT において，対訳文パターンの日本語原文と入力文とのレーベンシュタイン距離 [4](以下 LsD) を求める．この距離を利用して，入力文と対訳文パターンの日本語原文との類似度を求め，対訳文パターンを選択する際に，対訳文パターン対数確率の代わりに使用する．そして，入力文と類似した日本語原文から作成された対訳文パターンを選択することにより，翻訳精度の向上を目指した．しかし，翻訳精度の向上はあまり見られなかった．

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 3 章で今回使用する Pattern Based SMT について説明する．第 4 章で提案する手法について説明する．第 5 章で実験データを示す．第 6 章で実験結果と評価を示す．第 7 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

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, 隣接する複数のフレーズを用いて翻訳を行う方法である. 日英統計翻訳は，日本語入力文 J が与えられた場合に，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 E を探索することで翻訳を行う．以下にその基本モデルを示す.

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

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

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.

2.4 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.

2.5　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.

2.6 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．例として「京都　に　行く 。」という文に対する 2-gram モデルを式 2.23 に示す.

$$P(\text{京都　に　行く　。}) = P(\text{京都})P(\text{に} \mid \text{京都})P(\text{行く} \mid \text{に})P(\text{。} \mid \text{行く}) \tag{2.23}$$

2.7 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” [5] がある．

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．

第3章　Pattern Based SMT

3.1　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.

3.2　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辞書と対訳単語辞書を用いて“ 対訳フレーズ辞書 ”を作成する.

手順 4 対訳文と対訳フレーズと対訳単語辞書を用いて“ 句に基づく対訳文パターン辞書 ”を作成する.

手順 5 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 6 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.

3.2.1　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 3.1 に示す.

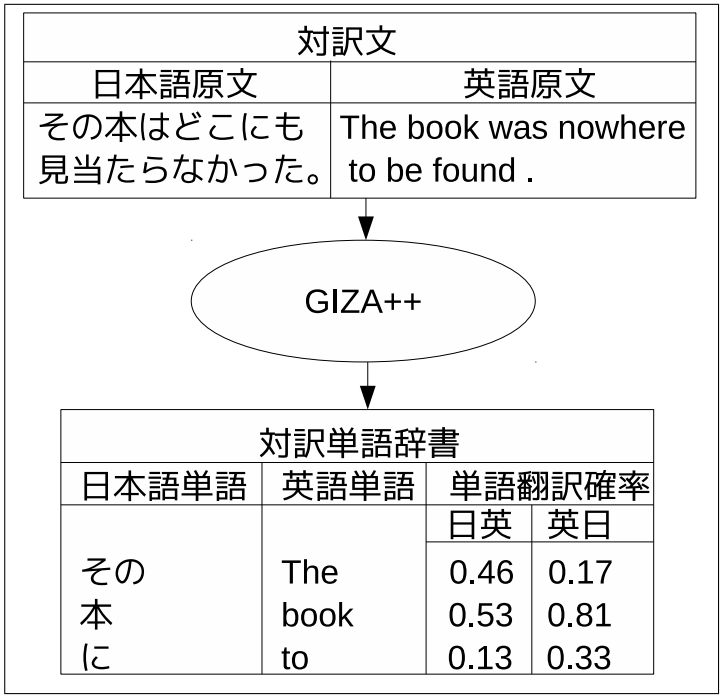

図 3.1: 対訳単語辞書の作成

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．

3.2.2 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 3.2 に示す．

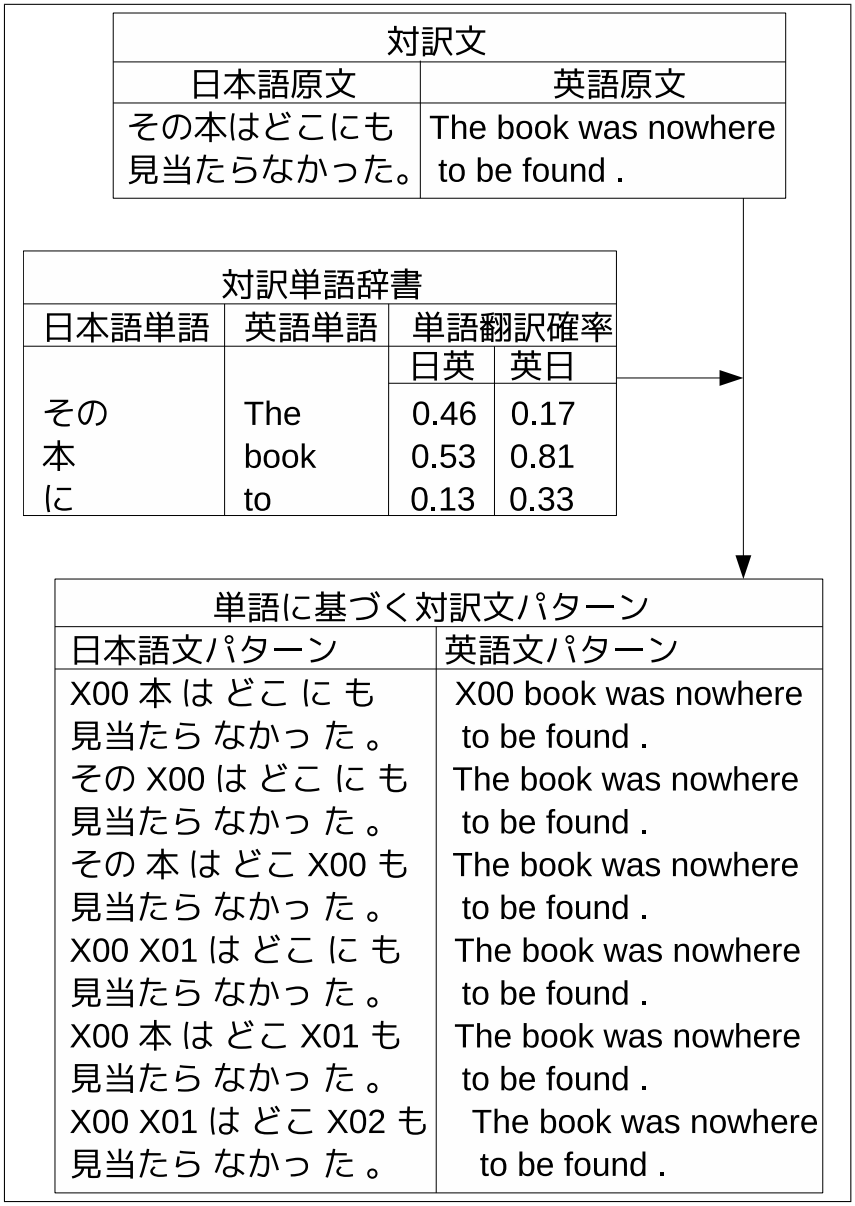

図 3.2: 単語に基づく対訳文パターンの作成

3.2.3 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 3.3 にに示す.

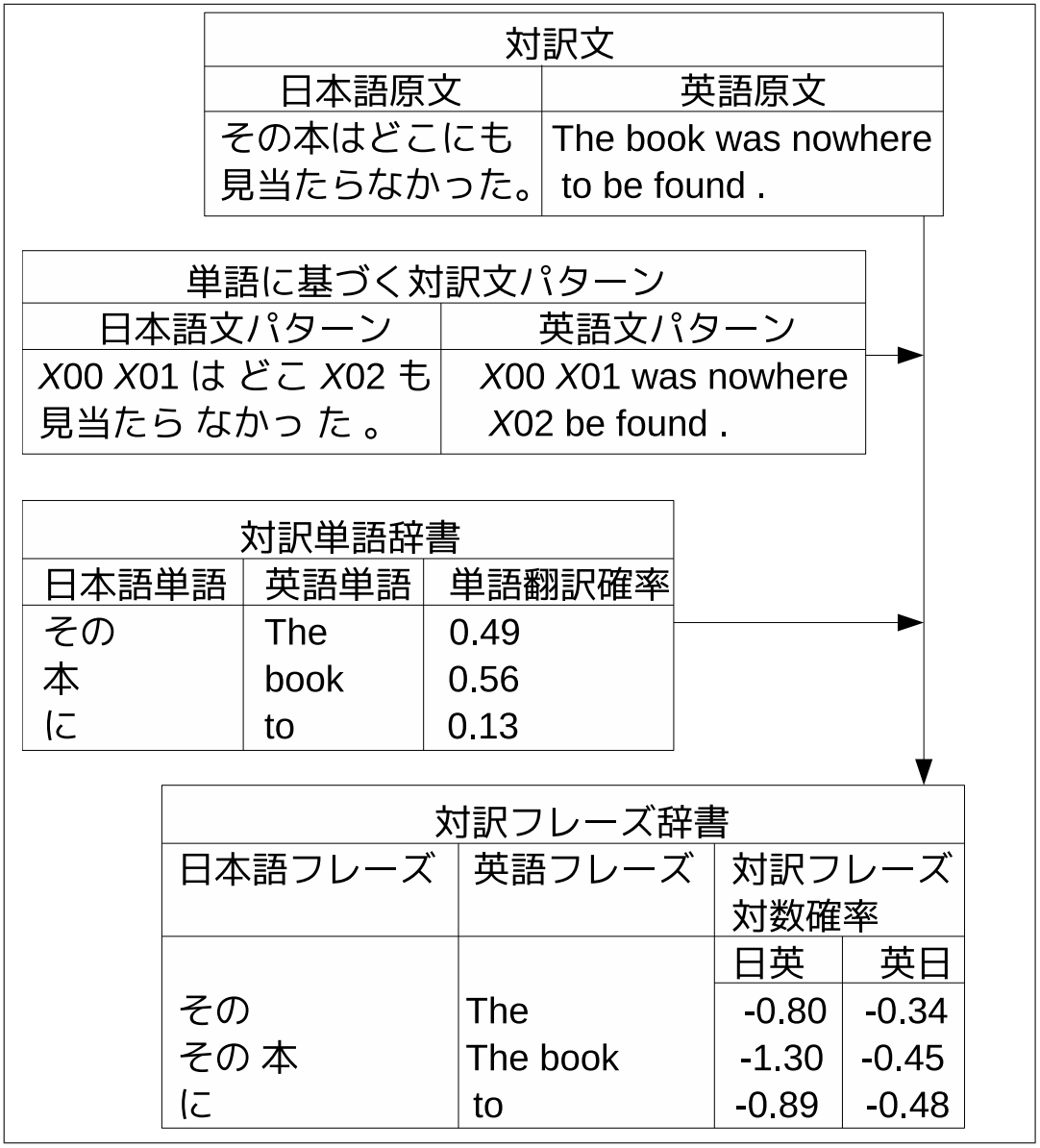

図 3.3: 対訳フレーズ辞書の作成

対訳フレーズ対数確率にも，3.2.1 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 3.4 に，英日対訳フレーズ対数確率の付与を図 3.5 に示す．

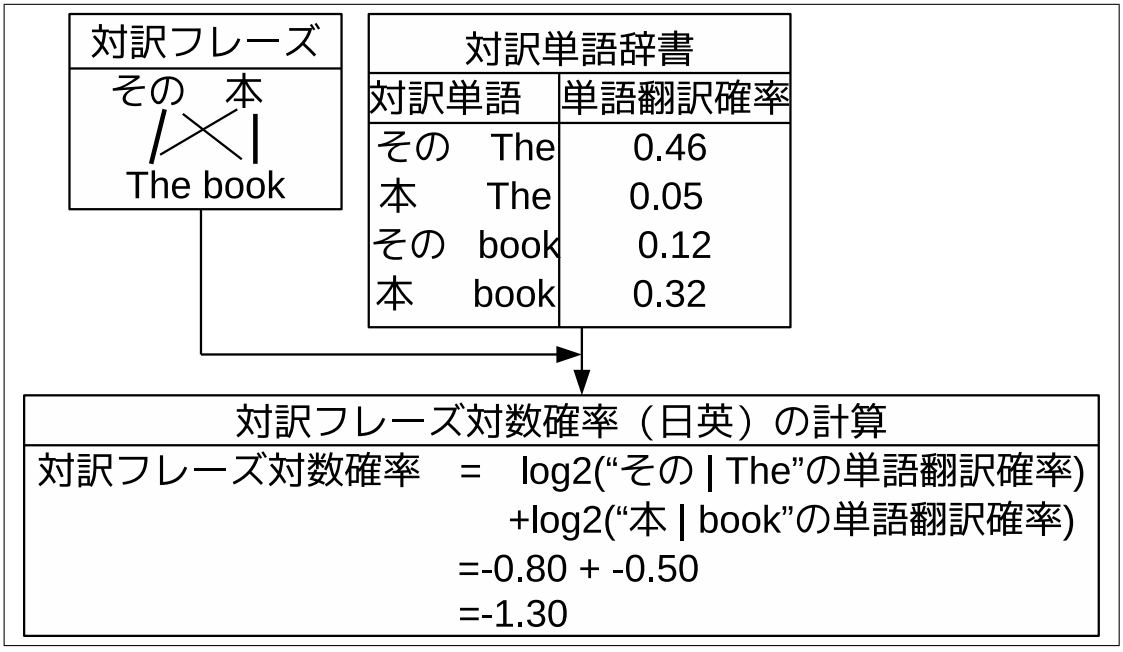

図 3.4: 日英方向の対訳フレーズ対数確率の付与

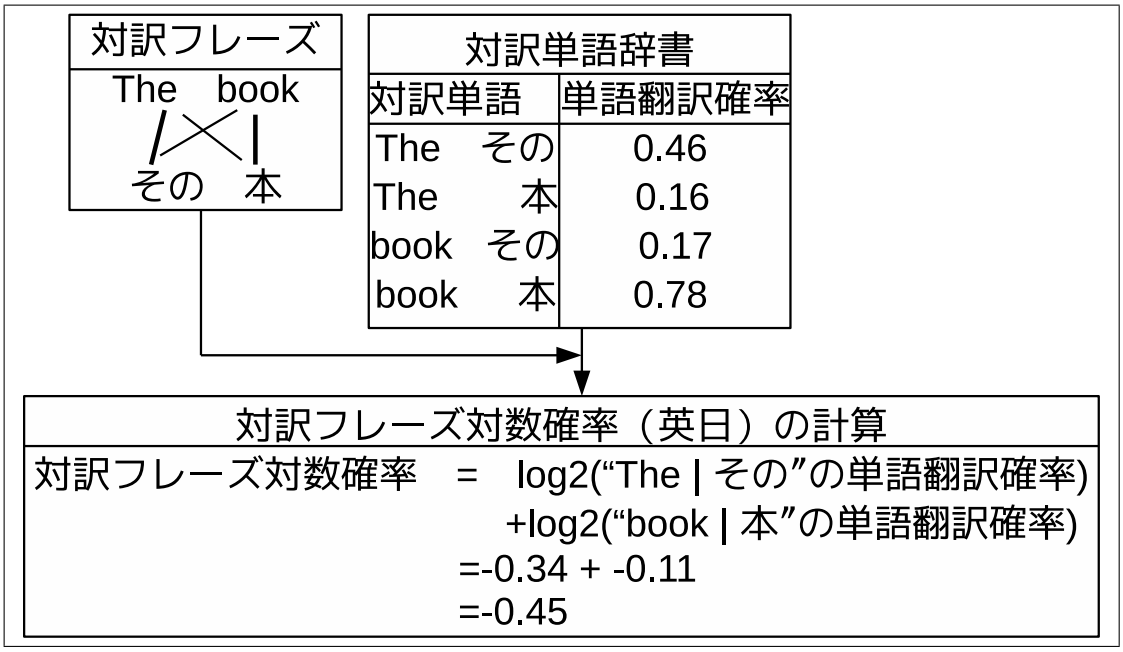

図 3.5: 英日方向の対訳フレーズ対数確率の付与

3.2.4 句に基づく対訳文パターン辞書の作成

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．その後，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．句に基づく対訳文パターン辞書の作成を図 3.6 に示す.

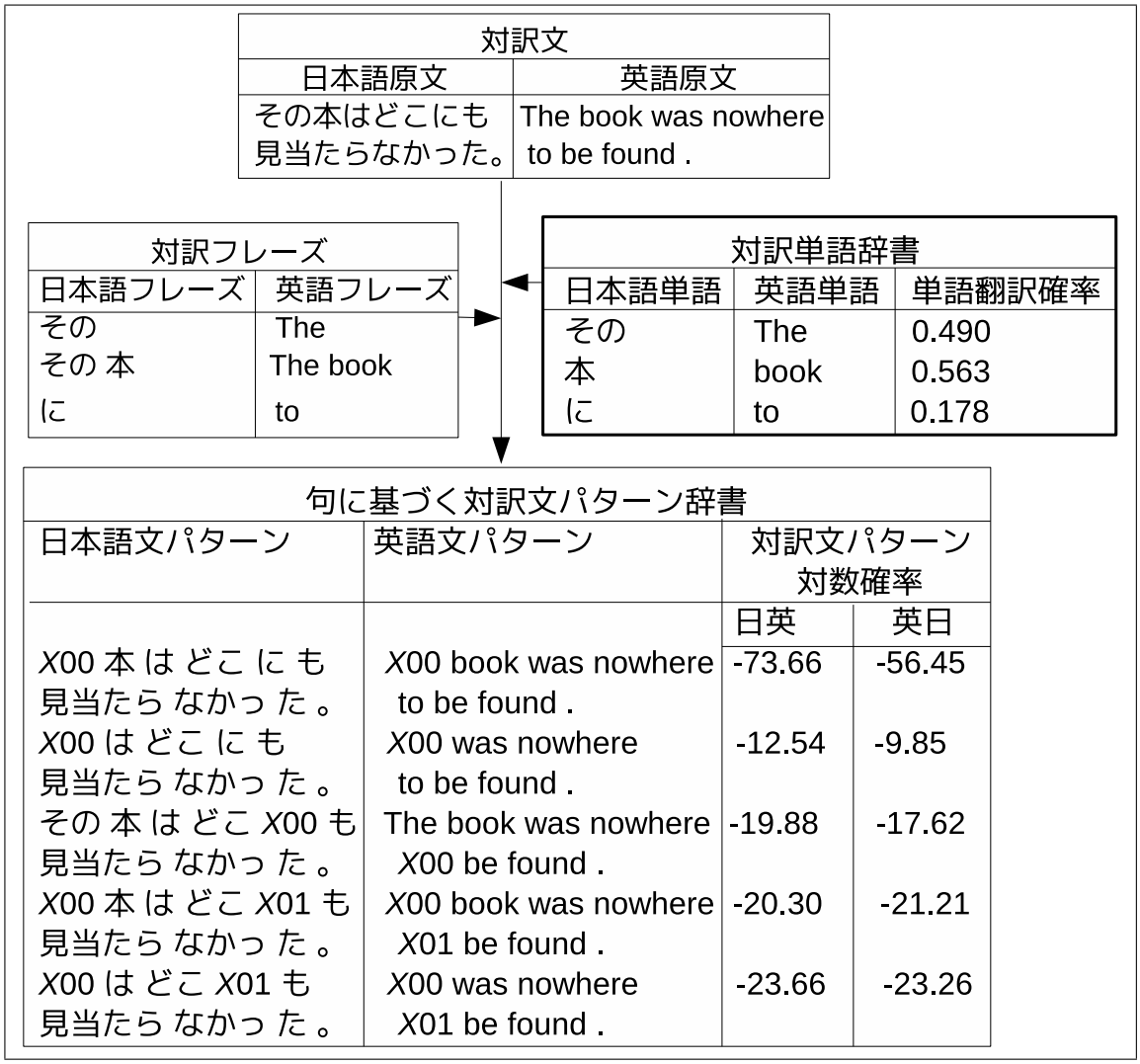

図 3.6: 句に基づく対訳文パターン辞書の作成

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 3.7 に，英日対訳文パターン対数確率の付与を図 3.8 に示す．

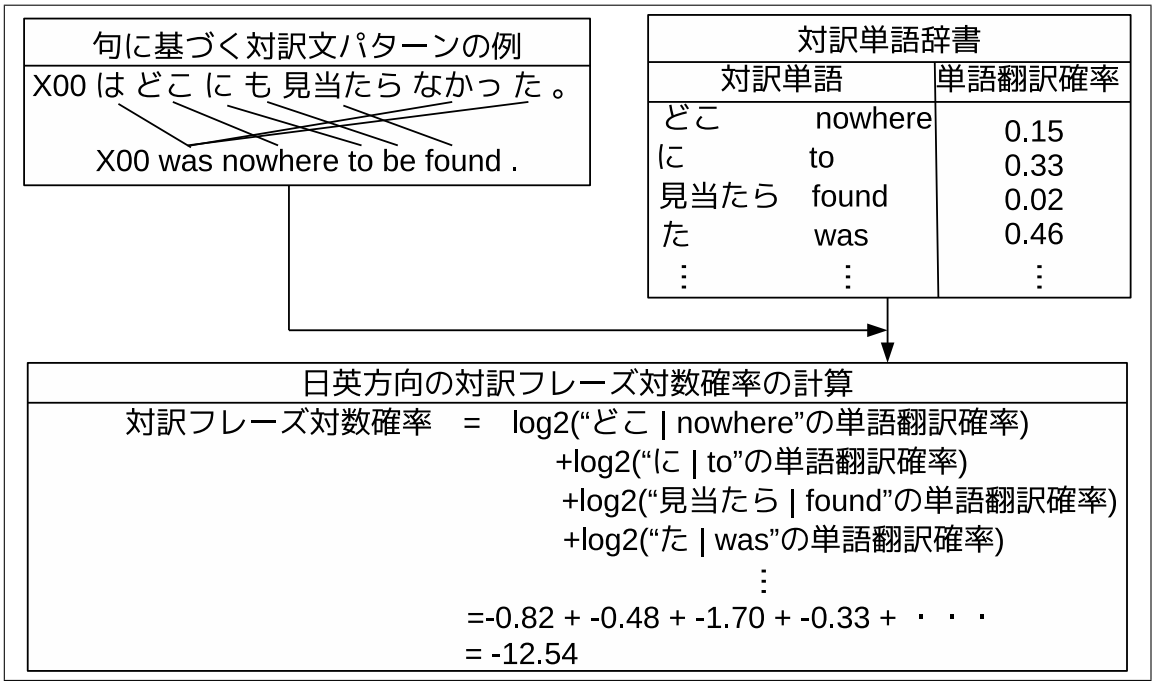

図 3.7: 日英方向の対訳文パターン対数確率の付与

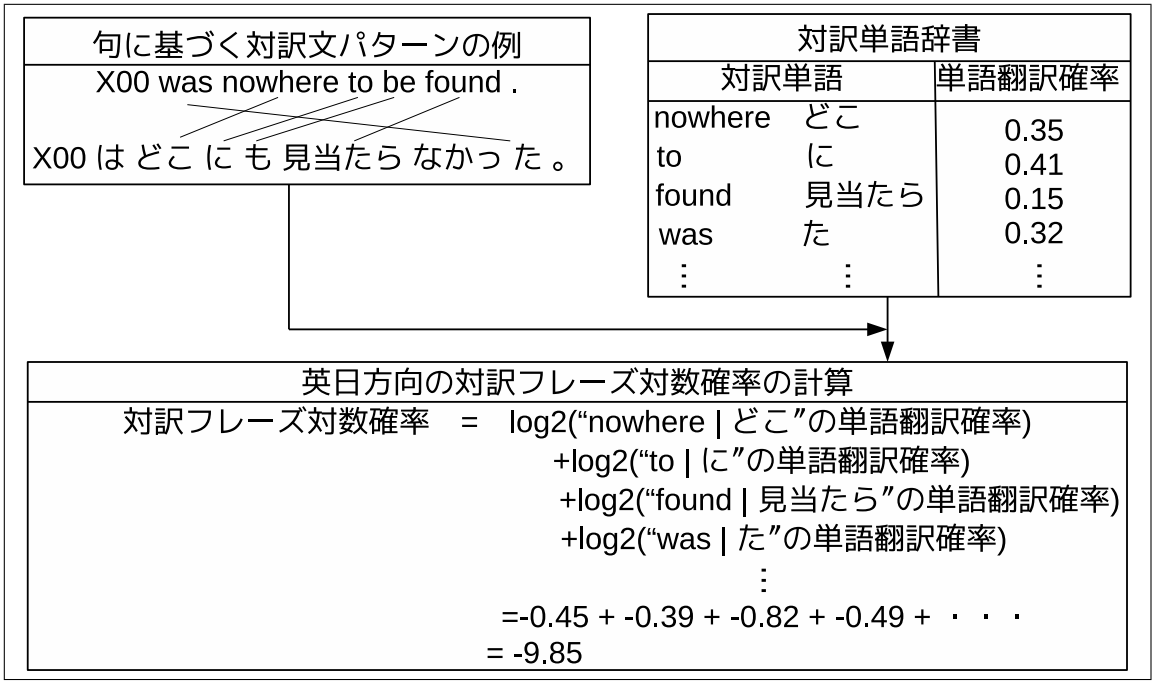

図 3.8: 英日方向の対訳文パターン対数確率の付与

3.2.5 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 3.9 に示す．

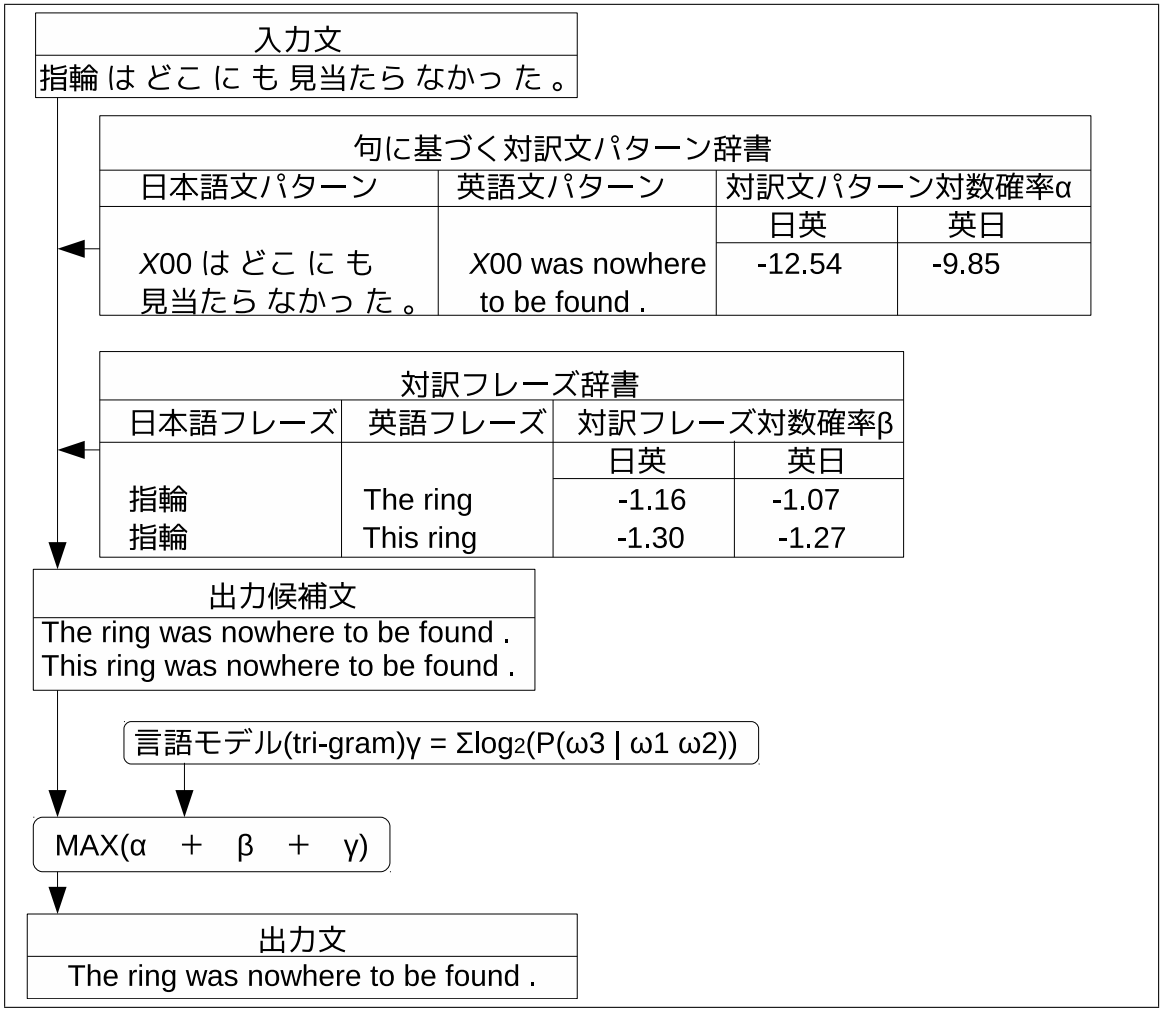

図 3.9: 出力文生成の流れ

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．

3.3 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．誤り解析を行った結果，不適切な対訳文パターンの選択が一つの原因であった．人手評価の低い入力文に対して，適切な対訳文パターンを使用した結果，人手評価の高い出力文が得られた．不適切な対訳文パターンを使用した翻訳の例を表 3.1 に，適切な対訳文パターンを使用した翻訳の例を表 3.2 に示す．

表 3.1: 不適切な文パターンを使用した翻訳の例

入力文	指輪 は どこ に も 見当たら なかっ た 。
参照文	The ring was nowhere to be found .
日本語文パターン	$X00$ は $X01$ に $X02$ なかっ た 。
英語文パターン	It took $X00$ forever to $X02$ $X01$.
日本語文パターンの原文	彼 は なかなか 仕事 に 取りかから なかっ た 。
英語文パターンの原文	It took him forever to get down to work .
出力文	It took the ring forever to be found anywhere .

表 3.2: 適切な文パターンを使用した翻訳の例

入力文	指輪 は どこ に も 見当たら なかっ た 。
参照文	The ring was nowhere to be found .
日本語文パターン	$X00$ は どこ に も 見当たら なかっ た 。
英語文パターン	$X00$ was nowhere to be found .
日本語文パターンの原文	その 本 は どこ に も 見当たら なかっ た 。
英語文パターンの原文	The book was nowhere to be found .
出力文	This ring was nowhere to be found .

表 3.1 の例では，日本語文パターンの原文は入力文とは類似しておらず，生成した出力文に対する人手評価は低い．

表 3.2 の例では，表 3.1 と異なり，日本語文パターンの原文は入力文と類似している．よって表 3.2 は，生成した出力文に対する人手評価は表 3.1 の例と比較して高い．

第4章　提案手法

4.1　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，入力文に対して不適切な対訳文パターンを選択することである．人手評価が低い入力文に対して対訳文パターンの日本語原文と入力文が類似した対訳文パターンを与えた結果，人手評価が高い出力候補文が選択された．

よって，対訳文パターンの日本語原文と入力文の類似した対訳文パターンを選択するために，入力文と日本語文パターンの原文との類似度を利用する．類似度は LsD を用いて求め，3.2.5 節での出力候補文を選択する際に，対訳文パターン対数確率 (α) の代わりに使用する．

4.2　レーベンシュタイン距離 (Levenshtein Distance)

文字列同士の類似度にはいくつかあるが，今回は LsD を用いた類似度を使用する．LsD とは, 一つの文字列をもう一つの文字列にするための編集回数である. 編集には, 挿入編集 (Insertion), 削除編集 (Deletion), 置換編集 (Substitution) の三つがある．一つの文字列を入力文，もう一つの文字列を日本語原文として編集内容を説明する．挿入編集は入力文に存在せず，日本語原文に存在する必要な単語を挿入する編集である．削除編集は入力文に存在し，日本語原文に存在しない不必要な単語に対して行われる編集である．置換編集は，入力文の不必要な単語を，日本語原文の必要な単語に置き換える，挿入編集と削除編集を同時に行う編集である．この編集回数の総和がレーベンシュタイン距離となる．　レーベンシュタイン距離を求める手順を表 4.1 の入力文と日本語原文を用いて示す.

表 4.1: レーベンシュタイン例:データ

入力文	彼 は 車 で 海 に 行 く 。
日本語原文	今日 彼 は 山 に 行 く 。

手順 1 入力文の単語と日本語原文の単語の対応を取得する．対応は動的計画法より取得する．対応を取った結果を表 4.2 に示す．

表 4.2: レーベンシュタイン例:対応

入力文	彼 は 車 で 海 に 行 く 。
日本語原文	今日 彼 は 山 に 行 く 。

手順 2 入力文を日本語原文にするための編集を行う．編集内容を図 4.1 に示す．

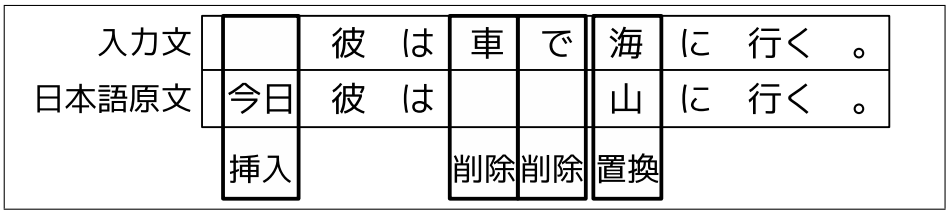

図 4.1: レーベンシュタイン距離:編集

手順 3 編集回数の総和を求め，それをレーベンシュタイン距離とする．今回の例では挿入編集一回，削除編集二回，置換編集一回となり，入力文と日本語原文のレーベンシュタイン距離は 4 となる．

4.3 類似度

本研究では， 削除編集 D，置換編集 S と入力文の単語数 N を用いて入力文と日本語文パターンの原文との類似度を式 (4.1) の計算式により求める．

$$\text{類似度} = \frac{N - D - S}{N} \tag{4.1}$$

4.2 節の例で使用した入力文と日本語原文を使用して類似度の計算を行う．入力文の単語数は 8，削除編集回数は 2，置換編集回数 1 であり，類似度は 0.63 となる．計算式を式 4.2 に示す．

$$\frac{8 - 2 - 1}{8} = 0.63 \tag{4.2}$$

4.4 実験手順

本研究では，日英翻訳を行う．提案手法は出力文の生成において，入力文と日本語文パターンの原文との類似度を求め，対訳文パターン対数確率の代わりに使用する．

4.5 レーベンシュタイン距離を用いた類似度の付与

入力文と対訳文パターンの日本語原文との LsD を求める．次に，LsD より“ 類似度 ”を求める．最後に，出力候補文の生成に用いた対訳フレーズ対数確率と言語モデル (tri-gram) と“ 類似度 ”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．提案手法における出力文生成の流れを図 4.2 に示す．

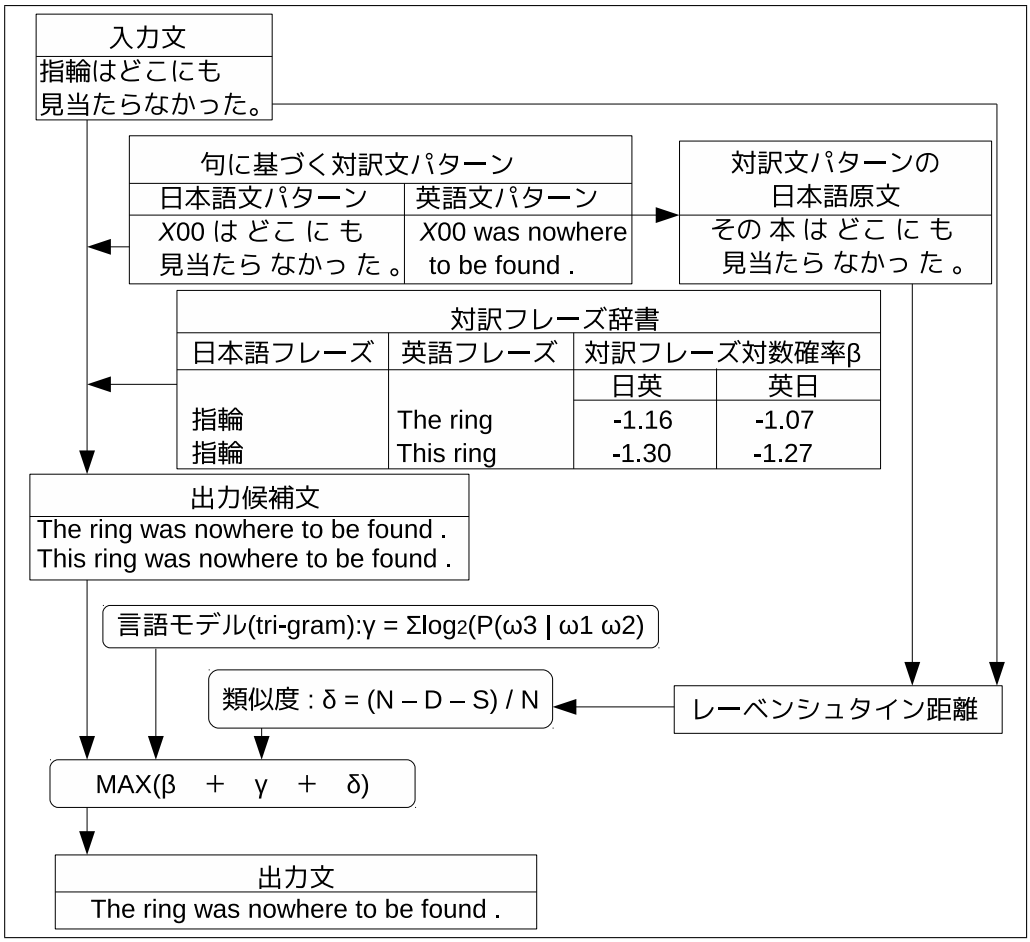

図 4.2: 提案手法における出力文生成の流れ

第5章　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 [6]. 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 5.1 に示す.

表 5.1: 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 5.2 と表 5.3 に示す.

表 5.2: 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 5.3: 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 .

第6章　実験結果

提案手法を用いて，翻訳実験を行う．実験の結果，入力文 100 文中，出力文 81 文を得た．

6.1　対比較評価

提案手法とベースラインの対比較評価を行った．ベースラインには，3 章の対訳文パターン対数確率を使用した Pattern Based SMT を用いる．評価方法は，提案手法とベースライン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 6.1 に，評価例を表 6.2，6.3，6.4，6.5，6.6，6.7 に示す．

表 6.1: 人手による評価

提案手法○	ベースライン○	差なし	同一出力
8	7	18	48

6.1.1 提案手法○の例

表 6.2: 提案手法○の例 1

入力文		彼 は 仕事 で 京都 へ 行っ た 。
参照文		He went to Kyoto on business .
提案手法	日本語文パターン	彼 は 仕事 で $X00$ 行っ た 。
	英語文パターン	He went $X00$ on business .
	日本語原文	彼 は 仕事 で ロンドン へ 行っ た 。
	英語原文	He went to London on business .
	類似度	0.89
	出力文	He went to Kyoto on business
ベースライン	日本語文パターン	彼 は $X00$ で $X01$ 行っ た 。
	英語文パターン	He went $X01$ by $X00$.
	日本語原文	彼 は シベリア 経由 で ヨーロッパ へ 行った 。
	英語原文	He went to Europe by via Siberia .
	対訳文パターン対数確率	-116.36
	出力文	He went to Kyoto by work

表 6.2 の例は，提案手法とベースラインの出力文は似ているが,「by」では意味合いが異なると考え,「on」である提案手法の出力文が優れていると評価した.

表 6.3: 提案手法○の例 2

入力文		彼 の 胸 は 興奮 で 躍動 し て い た 。
参照文		His heart was throbbing with excitement .
提案手法	日本語文パターン	彼 の $X00$ は $X01$ $X02$ し て い た 。
	英語文パターン	His $X00$ $X02$ $X01$.
	日本語原文	彼 の 発言 は 彼 の 愚か さ を 暴露 して い た 。
	英語原文	His remarks revealed his stupidity .
	類似度	0.67
	出力文	His heart danced with excitement .
ベースライン	日本語文パターン	彼 の $X00$ は 興奮 で $X01$ い た 。
	英語文パターン	His $X00$ was $X01$ excitement .
	日本語原文	彼 の 声 は 興奮 で 震え て い た 。
	英語原文	His voice was shaking with excitement .
	対訳文パターン対数確率	-167.64
	出力文	His heart was in excitement .

表 6.3 の例は，提案手法は概ね意味がわかるものに対して，ベースラインは動詞が抜けていたので，提案手法の出力文が優れていると評価した.

表 6.4: 提案手法○の例 3

入力文		時計 が かちかち と 時 を 刻ん で いる 。
参照文		The clock is ticking away the time .
提案手法	日本語文パターン	時計 が かちかち と $X00$ $X01$ いる 。
	英語文パターン	The clock $X01$ making $X00$.
	日本語原文	時計 が かちかち と 、音 を 立て て いる。
	英語原文	The clock is making a ticking sound .
	類似度	0.70
	出力文	The clock ticked away the making of a time .
ベースライン	日本語文パターン	時計 が $X00$ $X01$ $X02$ で いる 。
	英語文パターン	The clock is $X00$ $X01$ $X02$.
	日本語原文	時計 が 5 分 進ん で いる 。
	英語原文	The clock is five minutes fast .
	対訳文パターン対数確率	-58.11
	出力文	The clock is frozen in time with his .

表 6.4 の例は，提案手法は「making of a time」が怪しいですが概ね意味が合っており，ベースラインは時計以外意味が通じなかったので，提案手法の出力文が優れていると評価した．

6.1.2 ベースライン○の例

表 6.5: ベースライン○の例 1

入力文		彼女 の 心臓 は どきどき し て い た 。
参照文		Her heart thudded .
提案手法	日本語文パターン	彼女 の $X00$ $X01$ し て い た 。
	英語文パターン	$X00$ $X01$
	日本語原文	彼女 の 目 は 疲労 で ピクピク し て い た 。
	英語原文	Her eyes twitched with fatigue .
	類似度	0.80
	出力文	The heart beats
ベースライン	日本語文パターン	彼女 の $X00$ は $X01$ し て い た 。
	英語文パターン	Her $X00$ was $X01$.
	日本語原文	彼女 の 声 は はつらつ と し て い た 。
	英語原文	Her voice was fresh as springtime .
	対訳文パターン対数確率	-188.69
	出力文	Her heart was pounding .

表??の例では，提案手法は「Her」が抜けており，過去形でもない．対して，ベースラインは意味が入力文と同じと考え，ベースラインの出力文が優れていると評価した．

表 6.6: ベースライン○の例 2

入力文		雨 で 試合 が 流れ た 。
参照文		The game was rained out .
提案手法	日本語文パターン	雨 で $X00$ が 流れ た .
	英語文パターン	The $X00$ the poster ran in the rain .
	日本語原文	雨 で ポスター の 色 が 流れ た 。
	英語原文	The colors on the poster ran in the rain .
	類似度	0.86
	出力文	The match the poster ran in the rain .
ベースライン	日本語文パターン	雨で $X00$ が $X01$.
	英語文パターン	The $X00$ $X01$ in the rain .
	日本語原文	雨 で 木々 が 潤っ た 。
	英語原文	The trees became wet in the rain .
	対訳文パターン対数確率	-65.31
	出力文	The game was rained out in the rain .

表 6.6 の例では，提案手法は,「poster」という不要な字面が残っている．対してベースラインは入力文と意味が同じと考え，ベースラインの出力文が優れていると評価した．

表 6.7: ベースライン◯の例 3

入力文		彼女 は 直感 的 に 物事 を とらえる 。
参照文		She perceives things intuitively .
提案手法	日本語文パターン	$X00$ は $X01$ 的 に $X02$ を $X03$.
	英語文パターン	$X00X03$ $X01$ $X02$.
	日本語原文	捕鯨 反対 運動 は 国際 的 に 勢い を 増して いる 。
	英語原文	The campaign against whaling is gaining increasing international momentum .
	類似度	0.56
	出力文	Her views of things .
ベースライン	日本語文パターン	彼女 は $X00$ に $X01$ を $X02$ 。
	英語文パターン	She $X02$ the $X01$ $X00$.
	日本語原文	彼女 は 壁 に 皿 を 投げつけ た 。
	英語原文	She smashed the plate against the wall .
	対訳文パターン対数確率	-48.08
	出力文	She obtained the things of it intuitively .

表 6.7 の例では，提案手法は動詞もなく，全く意味がわからない．対して，ベースラインは過去形ではあるが，概ね意味が入力文と同じと考え，ベースラインの出力文が優れていると評価した．

表 6.8: 差なしの例 1

入力文		幸福 で 彼女 の 目 は 輝い た 。
参照文		Happiness kindled her eyes .
提案手法	日本語文パターン	$X00$ で 彼女 の $X01$ は $X02$ た 。
	英語文パターン	The $X00$ $X02$ her $X01$.
	日本語原文	強い 風 で 彼女 の 髪 は くしゃくしゃ に なっ た 。
	英語原文	The strong wind tousled her hair .
	出力文	The happiness with her eyes .
ベースライン	日本語文パターン	$X00$ で 彼女 の $X01$ は $X02$ た 。
	英語文パターン	$X00$ made her $X01$ $X02$.
	日本語原文	恐怖 で 彼女 の ひざ は がくがく した 。
	英語原文	Fear made her knees knock together .
	出力文	She made her eyes spakled .

表 6.9: 差なしの例 2

入力文		人 は 誰 しも 間違い を する もの だ 。
参照文		One is certain to make slips .
提案手法	日本語文パターン	$X00$ は $X01$ $X02$ する もの だ 。
	英語文パターン	$X00$ $X02$ $X01$.
	日本語原文	子ども たち は おもしろ 半分 に いたずら を する もの だ 。
	英語原文	Children play tricks for fun .
	出力文	Two mistakes in his foible .
ベースライン	日本語文パターン	$X00$ は $X01$ $X02$ する もの だ 。
	英語文パターン	$X00$ is $X02$ $X01$.
	日本語原文	それ は 教会 を 冒涜 する もの だ 。
	英語原文	It is blasphemy against the church .
	出力文	Two is of mistakes in his foible .

表 6.10: 差なしの例 3

入力文		噴水 は 四方八方 に 水 を まき散らし て い た 。
参照文		The fountain sprayed water in every direction .
提案手法	日本語文パターン	$X00$ は $X01$ に $X02$ を $X03$ $X04$ い た 。
	英語文パターン	$X00$ $X03$ $X04$ $X02$ $X01$.
	日本語原文	彼 は 左 の 目 に 眼帯 を し て い た 。
	英語原文	He had a patch over his left eye .
	出力文	The tissue is water in all directions .
ベースライン	日本語文パターン	$X00$ は $X01$ $X02$ $X03$ を $X04$ て い た 。
	英語文パターン	She had a $X04$ $X02$ $X00$ $X01$ $X03$.
	日本語原文	彼女 は 母 の 帰り を 待ちわび て いた 。
	英語原文	She had a tedious wait for her mother's return .
	出力文	She had a tissue in the fountain in water .

第7章　考察

7.1　提案手法の有効性

人手評価の結果，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，入力文と対訳文パターンの日本語原文との類似度の効果と，対訳文パターン対数確率の効果が同様であると考えられる．

よって，提案手法は，対訳文パターンの選択において，対訳単語辞書を用いない新たな対訳文パターンの選択手法であると言える．

7.2　誤り解析

対訳文パターンの選択の際に，類似度を用いることにより，対訳文パターンの日本語原文と入力分が類似した対訳文パターンの選択をすることができた．しかし，評価を行った結果，ベースラインと比較して，精度は向上しなかった．そこで，ベースライン○と評価された 7 文に対して，誤り解析を行った．誤り解析の結果，選択した対訳文パターンにおいて，対訳文パターンの日本語原文と入力文は類似していたが，日本語文パターンの字面と英語文パターンの字面の対応が取れていない対訳文パターンがあることがわかった．字面の対応の取れていない対訳文パターンの例を表 7.1 に示す．

表 7.1: 字面の対応が取られていない日英文パターンの例

入力文		彼女 は 思いがけない 質問 に まごつい た よう だっ た 。
参照文		She seemed to be embarrassed at the unexpected question .
提案手法	日本語文パターン	$X00$ $X01$ $X02$ に $X03$ よう だっ た 。
	英語文パターン	I came across a very helpful person when I $X03$ $X01$ $X02$ $X00$.
	対訳文パターンの日本語原文	地獄 で 仏 に あっ た よう だっ た 。
	対訳文パターンの英語原文	I came across a very helpful person when I was in dire trouble .
	出力文	I came across a very helpful person when I was upset at the unexpected question to her .
ベースライン	日本語文パターン	$X00$ $X01$ $X02$ に $X03$ た $X04$ だっ た 。
	英語文パターン	$X00$ was $X03$ $X04$ $X01$ $X02$.
	対訳文パターンの日本語原文	それ は 浮き彫り に し た キューピッド の 像だっ た 。
	対訳文パターンの英語原文	It was a figure of Cupid in relief .
	出力文	She was perplexed by such a stroke of questions .

7.3 対訳文パターンにおける字面の対応

表 7.1 において，日本語文パターンの字面と英語文パターンの字面の数が著しく異なる対訳文パターンを，対訳文パターンの選択の際に除外することにより，人手による対比較評価が改善すると考える．また，本研究では，対訳フレーズ対数確率と対訳文パターン対数確率，類似度が等しくなるように重みを設定した．この重みを最適化することで人手評価が向上すると考える．

7.4 出力文の生成における字面の一致と類似度

対訳文パターンの選択の際に，まず入力文の字面と対訳文パターンの字面の照合を行い，多く字面適合する対訳文パターンを優先的に選択する．この字面の照合による優先的な選択が類似度の効果と同じではないかと考えられる．

第8章　追加実験

Pattern Based SMT における類似度の有効性を更に確認するために，追加実験を行う．追加実験は Pattern Based SMT の条件を変え類似度を用いた翻訳実験を行い，翻訳精度の調査を行う．追加実験は全部で二つ行い，一つ目を追加実験 A，二つ目を追加実験 B として扱う．以下に実験内容を示す.

8.1　追加実験 A

8.1.1　実験内容

追加実験 B は，Pattern Based SMT における対訳フレーズの自動作成方法を変更し，類似度の有効性を確認する．新たな自動作成の流れを図 8.1 に示す.

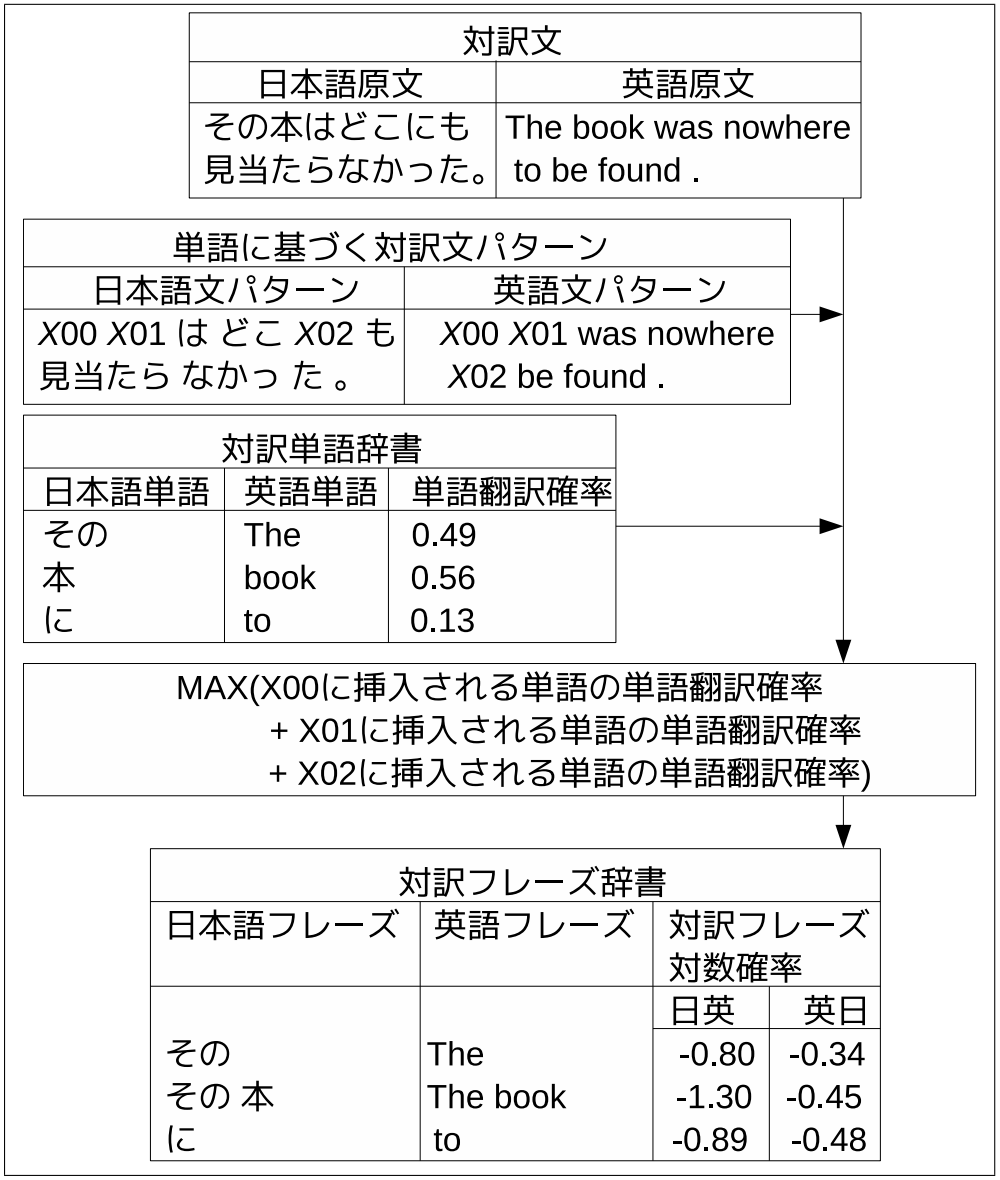

図 8.1: 新たな対訳フレーズ辞書の作成方法

新たな対訳フレーズの自動作成方法は, 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の抽出までは, 従来の作成方法と同じである. ここで, 一つの単語に基づく対訳文パターンに対して抽出された一組の対訳フレーズの対訳フレーズ対数確率の総和を取る. この総和が最大となった一組の対訳フレーズのみを対訳フレーズ辞書に出力する.

8.1.2 実験結果

新たな対訳フレーズ辞書を使用し, 類似度を使用したものと類似度を使用しないもので実験を行う. 学習を行う対訳文及び, 翻訳を行う入力文は 5 節と同じものを使用する. 入力文 100 文に対して, 追加実験 A は出力文 71 文を取得した. 出力文に対して, 対比較評価を行う.

8.1.3 対比較評価

追加実験 A の類似度を使用したものと類似度を使用しないもので対比較評価を行う. 評価結果を表 8.1 に示す.

表 8.1: 追加実験 A の対評価

類似度あり○	類似度なし○	差なし	同一出力
1	3	14	53

この結果から, 追加実験 A での類似度の有効性は確認できなかった.

8.2 追加実験 B

8.2.1 実験内容

追加実験 B は，追加実験 A の対訳フレーズの自動作成方法の変更に加え, 句に基づく対訳文パターンの自動作成方法を変更し, 類似度の有効性を確認する. 新たな句に基づく文パターン辞書の自動作成方法の流れを図 8.2 に示す.

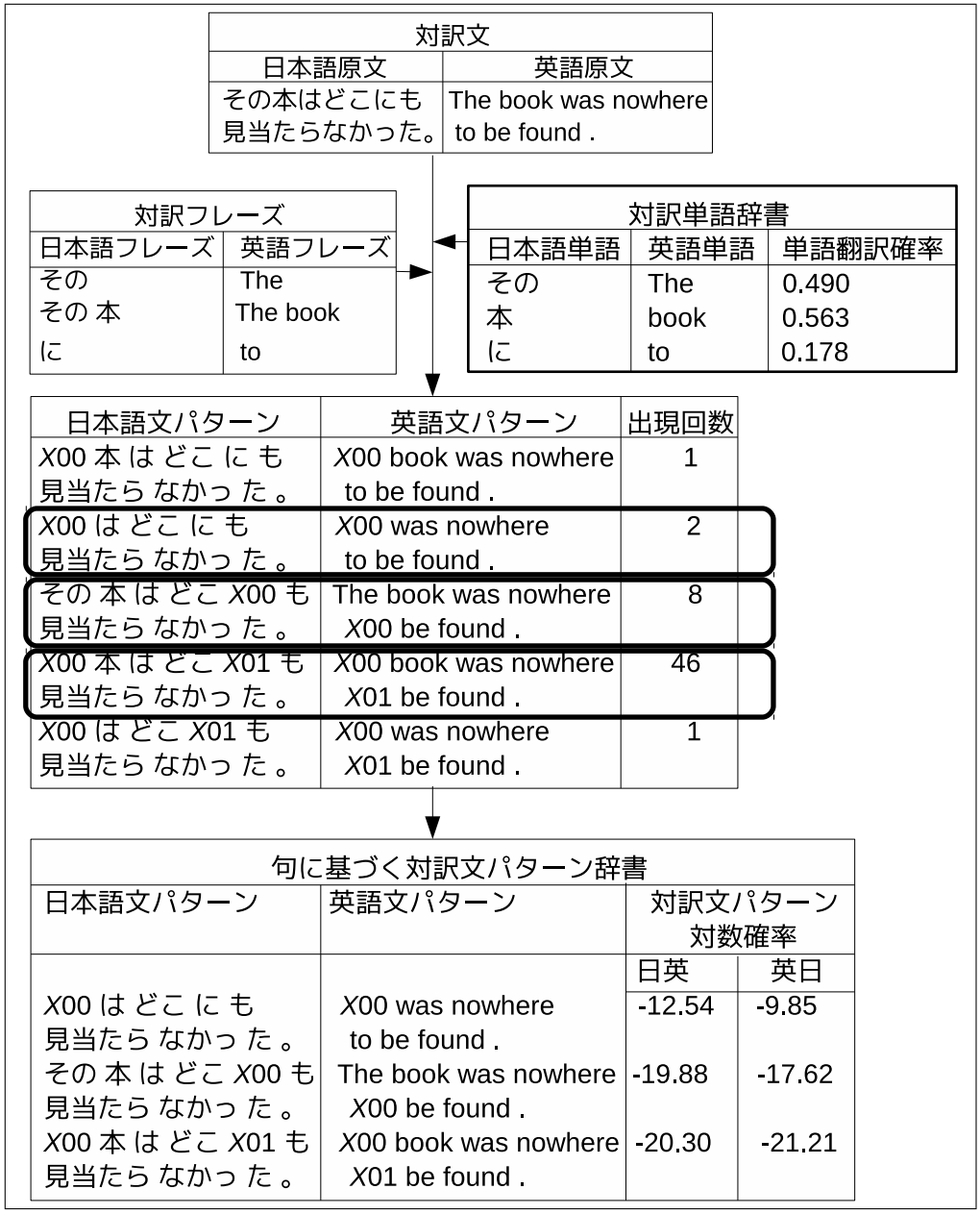

図 8.2: 新たな句に基づく対訳文パターン辞書の作成

新たな対訳文パターン辞書の自動作成の変更点は, 対訳文パターンの出現回数を数え, 一度しか出現していないものを削除する.

8.2.2 実験結果

新たな句に基づく対訳文パターン辞書を使用し, 類似度を使用したものと類似度を使用しないもので実験を行う. 学習を行う対訳文及び, 翻訳を行う入力文は 5 節と同じものを使用する. 入力文 100 文に対して, 追加実験 A は出力文 71 文を取得した. 出力文に対して, 対比較評価を行う.

8.2.3 対比較評価

追加実験 B の類似度を使用したものと類似度を使用しないもので対比較評価を行う. 評価結果を表 8.1 に示す.

表 8.2: 追加実験 B の対評価

類似度あり○	類似度なし○	差なし	同一出力
1	3	13	51

この結果から, 追加実験 B での類似度の有効性は確認できなかった.

第9章　おわりに

本研究では，Pattern Based SMT において，不適切な対訳文パターンの選択を抑制するため，対訳文パターンを選択する際に，対訳文パターン対数確率の代わりとして，入力文と対訳文パターンの日本語原文との LsD を使用することを提案した．人手による評価をした結果，出力文 81 文中，提案手法○ 8 文，ベースライン○ 7 文であり，提案手法の有効性は確認できなかった．また，誤り解析を行った結果，日本語文パターンの字面と英語文パターンの字面において，対応が取れていない対訳文パターンがあることがわかった．よって，日本語文パターンの字面と英語文パターンの字面の数が著しく異なる対訳文パターンを，除外することにより，翻訳精度が向上すると考える．また，今後対訳フレーズ対数確率と対訳文パターン対数確率と類似度の重みを最適化し，実験を行う．

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，徳久雅人講師，春野瑞季先輩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] Vladimir Iosifovich Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals”, Soviet Physics Doklady, 10(8), pp.707-710, 1966.

[5] 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.

[6] 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.