概要

機械翻訳の一種にパターンに基づく統計翻訳が提案されてる．この手法は，対訳句の抽出において，対訳文と単語レベル文パターンを照合して得る方法が用いられる．しかし不適切な対応をとる対訳句が多く含まれている．

そこで興梠らは，対訳句の変数全体の確率値を利用して最適な対訳句を抽出する手法を提案した．この手法により対訳句の出力数を大幅に削減し，不適切な対応をとる対訳句の数を減らした．しかし，対訳句は対訳文１文に対し複数のパターンを照合するため，対訳文に対し不適切なパターンを照合した際，不適切な対訳句が抽出される．そこで本研究では先行手法に加え，対訳句を抽出する際，単語レベル文パターンを作成する際に用いた対訳文と対訳句を作成する際に用いる対訳文との類似度を利用して，不適切な対訳句を削除する手法を提案する．この手法を用いることで対訳句の精度向上を目指す．また，出力した対訳句を用いて日英統計翻訳を行い，翻訳精度の調査を行う．

実験結果より，対訳句の精度は提案手法が先行手法よりも高く，類似度が有効であることが確認できた．しかし翻訳文 114 文より提案手法と先行手法の対比較評価を行ったが，翻訳の精度に大きな差はないという結果であった．

目 次

図 目 次

表 目 次

1　はじめに

機械翻訳の一種にパターンに基づく統計翻訳が提案されてる．この手法は，対訳句の抽出において，対訳文と単語レベル文パターンを照合して得る方法が用いられる．しかし不適切な対応をとる対訳句が多く含まれている．

そこで興梠らは，対訳句の変数全体の確率値を利用して最適な対訳句を抽出する手法を提案した．この手法により対訳句の出力数を大幅に削減し，不適切な対応をとる対訳句の数を減らした．しかし，対訳句は対訳文１文に対し複数のパターンを照合するため，対訳文に対し不適切なパターンを照合した際，不適切な対訳句が抽出される．そこで本研究では先行手法に加え，対訳句を抽出する際，単語レベル文パターンを作成する際に用いた対訳文と対訳句を作成する際に用いる対訳文との類似度を利用して，不適切な対訳句を削除する手法を提案する．この手法を用いることで対訳句の精度向上を目指す．また，出力した対訳句を用いて日英統計翻訳を行い，翻訳精度の調査を行う．

実験結果より，対訳句の精度は提案手法が先行手法よりも高く，類似度が有効であることが確認できた．しかし翻訳文 114 文より提案手法と先行手法の対比較評価を行ったが，翻訳の精度に大きな差はないという結果であった．

2　翻訳システム

2.1　概要

本章は，江木の論文 [1] を参照して既述している．翻訳システムとして“ 句に基づく統計翻訳 ”がある．句に基づく統計機械翻訳は，学習データとして対訳文を与えるだけで翻訳できるシステムであり，翻訳にかかるコストが低い．さらに，対訳文から単語辞書と単語翻訳確率を自動的に得ることが可能である．

一方，翻訳システムとして“ パターン翻訳 ”がある．パターン翻訳は大量の対訳文パターンと単語辞書を用いて，翻訳文を出力する方法である．パターン翻訳は，入力文が適切な対訳文パターンに適した場合に，翻訳精度の高い翻訳文が得られやすいという特徴がある．しかし，パターン翻訳に用いる単語辞書と対訳文パターンは人手で作成するため，開発のコストが高くなる．

そこで江木らは，単語辞書と対訳文パターンを統計的手法で自動的に作成し翻訳する方法を提案した．これを“ パターンに基づく統計機械翻訳 ”と呼ぶ．パターンに基づく統計機械翻訳は，句に基づく統計機械翻訳の特徴である対訳文から単語辞書と単語翻訳確率を自動的に取得できる点に着目し，翻訳に用いる単語辞書と対訳文パターンを統計的手法を用いて自動的に作成する手法である．

2.1.1 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．

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．

$$
\begin{aligned}
P(W_1^n) &= P(w_1) \times P(w_2|w_1) \times P(w_3|w_1^2)...P(w_n|w_1^{n-1}) && (1)\\
&\approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_1^2)...P(w_n|w_{n-(N-1)}^{n-1}) && (2)\\
&= \prod_{i=1}^{n} P(w_i|w_{i-(N-1)}^{i-1}) && (3)
\end{aligned}
$$

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．

$$
P(w_i|w_{i-(N-1)}^{i-1}) = \frac{C(w_{i-(N-1)}^{i})}{C(w_{i-(N-1)}^{i-1})} \tag{4}
$$

2.1.2 単語に基づく翻訳モデル

統計翻訳における単語対応を獲得するため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 [2] がある．IBM 翻訳モデルは，model1 から model5 までの 5 つのモデルから構成されている．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

IBM 翻訳モデルは仏英翻訳を前提としているが，本研究では日英翻訳を扱っているため，日英翻訳を前提に説明する．なお，以下の説明は藤原ら [7] の論文より引用した.

原言語の日本語文を J，目的言語の英語文を E として定義する．IBM 翻訳モデルにおいて，日本語文 J と英語文 E の翻訳モデル $P(J|E)$ を計算するため，アライメント a を用いる．以下に IBM モデルの基本的な計算式を示す.

$$P(J|E) = \sum_{a} P(J, a|E) \tag{5}$$

ここで，アライメント a は，J と E の単語の対応を意味している．IBM 翻訳モデルにおいて，各日単語に対応する英単語は 1 つであるのに対して，各英単語に対応する日単語は 0 から n 個あると仮定する．また，日単語と適切な英単語が対応しない場合，英語文の先頭に e_0 という空単語があると仮定し，日単語と対応させる.

model1

式 (3) は以下の式に置き換えられる．

$$P(j,a|E) = P(m|E)\prod_{j=1}^{m} P(a_j|a_1^{j-1}, j_1^{j-1}, m, E)P(j_j|a_1^{j}, j_1^{j-1}, m, E) \tag{6}$$

m は日本語文の文長を示す．また，a_1^{j-1} は日本語文の 1 単語目から j-1 単語目までのアライメントである．そして j_1^{j-1} は日本語文の 1 番目から j-1 番目までの単語を示す．ここで，Model1 では以下を仮定している．

- 日本語文の長さの確率 ϵ は，m と E に依存しない
  $\epsilon \equiv P(m|E)$
- アライメントの確率は英語文の長さ l にのみ依存する
  $P(a_j|a_1^{j-1}, j_1^{j-1}, m, E) \equiv (l+1)^{-1}$
- 日本語の翻訳確率 $t(j_j|e_{a_j})$ は，日単語に対応する英単語にのみ依存する
  $P(j_j|a_1^{j}, j_1^{j-1}, m, E) \equiv t(j_j|e_{a_j})$

以上の仮定を用いて，式 (4) は簡略化することができる．以下に式を示す．

$$P(J,a|E) = \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m} t(j_j|e_{a_j}) \tag{7}$$

$$P(J|E) = \frac{\epsilon}{(l+1)^m}\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\prod_{j=1}^{m} t(j_j|e_{a_j}) \tag{8}$$

$$= \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m}\sum_{i=0}^{l} t(j_j|e_i) \tag{9}$$

model1 において，翻訳確率 $t(j|e)$ の初期値が 0 でない場合， EM アルゴリズム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 $t(j|e)$ に初期値を設定する．

手順 2 日本語と英語の対訳文 $(J^{(s)}, E^{(s)})(1 \le s \le S)$ において，日単語 j と英単語 e が対応付けられる回数の期待値を求める．ここで $\delta(j, j_j)$ は日本語文 J において日単

語 j が出現する回数を表す．そして $\delta(e, e_i)$ は英語文 E において英単語 e が出現する回数を表す．

$$c(j|e; J, E) = \frac{t(j|e)}{t(j|e_0) + \cdots + t(j|e_l)} \sum_{j=1}^{m} \delta(j, j_j) \sum_{i=0}^{l} \delta(e, e_i) \tag{10}$$

手順 3 英語文 $E^{(s)}$ において，1 回以上出現する英単語 e に対して，翻訳確率 $t(j|e)$ を計算する．

- 定数 λ_e を以下の式で計算する

$$\lambda_e = \sum_j \sum_{s=1}^{S} c(j|e; J^{(s)}, E^{(s)}) \tag{11}$$

- 上式で求めた定数 λ_e を用いて $t(j|e)$ を以下の式で再計算する

$$t(j|e) = \lambda_e^{-1} \sum_{s=1}^{S} c(j|e; J^{(s)}, E^{(s)}) \tag{12}$$

$$= \frac{\sum_{s=1}^{S} c(j|e; J^{(s)}, E^{(s)})}{\sum_j \sum_{s=1}^{S} c(j|e; J^{(s)}, E^{(s)})} \tag{13}$$

手順 4 $t(j|e)$ が収束するまで，手順 2 と手順 3 を繰り返す．

model2

model1 において，アライメントの確率は英語文の長さ l にのみ依存する．そこで model2 では，英語文の長さ l に加え，j 単語目のアライメント a_j，日本語文の長さ m に依存するとし，以下の式で表す．

$$a(a_j|j,m,l) \equiv P(a_j|a_1^{j-1}, j_1^{j-1}, m, l) \tag{14}$$

よって，model1 の式 (6) は以下のように置き換えられる．

$$\begin{aligned} P(J|E) &= \epsilon \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(j_j|e_{a_j}) a(a_j|j,m,l) & (15) \\ &= \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} t(j_j|e_i) a(i|j,m,l) & (16) \end{aligned}$$

model2 において，対訳文中の英単語 e と日単語 j が対応付けされる回数の期待値である $c(j|e; J^{(s)}, E^{(s)})$ と，日単語の位置 j と英単語の位置 i が対応付けられる回数の期待値 $c(i|j,m,l; J^{(s)}, E^{(s)})$ が存在する．以下に，期待値 $c(j|e; J^{(s)}, E^{(s)})$ と $c(i|j,m,l; J^{(s)}, E^{(s)})$ を求める式を示す．

$$\begin{aligned} c(j|e; J^{(s)}, E^{(s)}) &= \sum_{j=1}^{m} \sum_{i=0}^{l} \frac{t(j|e)a(i|j,m,l)\delta(j,j_j)\delta(e,e_i)}{t(j|e_0)a(0|j,m,l) + \cdots + t(j|e_l)a(l|j,m,l)} & (17) \\ c(i|j,m,l; J^{(s)}, E^{(s)}) &= \frac{t(j_j|e_i)a(i|j,m,l)}{t(j_j|e_0)a(0|j,m,l) + \cdots + t(j_j|e_l)a(l|j,m,l)} & (18) \end{aligned}$$

model2 においても，最適解を推定するために EM アルゴリズムを用いる．しかし，計算によって複数の極大値が算出され，最適解が得られない場合が存在する．model2 の特殊な場合に，$a(i|j,m,l) = (l+1)^{-1}$ が挙げられるが，これは model1 として考えることができる．また，最適解が保証されている model1 で求められた値を初期値として用いることで，最適解を求めることができる．

model3

model1 および model2 において，日単語と英単語の対応は 1 対 1 の場合のみを考慮していた．しかし，model3 では，1 つの単語が複数の単語に対応する場合や，単語の翻訳位置の距離についても考慮する．また，モデル 3 では単語の位置を絶対位置として考えている．モデル 3 では以下のパラメータを用いる．

- $P(j|e)$
  英単語 e が日単語 j に翻訳される確率

- $n(\phi|e)$
  英単語 e が ϕ 個の日単語と対応する確率

- $d(j|i,m,l)$
  英語文の長さ l，日本語文の長さ m のとき，i 番目の英単語 e_i が j 番目の日単語 j_j に翻訳される確率

さらに，英単語に翻訳されない日本語の単語数を ϕ_0 として，そのような単語が発生する確率 p_0 を以下の式に表す．

$$P(\phi_0|\phi_1^l,e) = \left(\begin{array}{c} \phi_1 + \cdots + \phi_l \\ \phi_0 \end{array} \right) p_0^{\phi_1+\cdots+\phi_l-\phi_0} p_1^{\phi_0} \tag{19}$$

したがって，model3 は以下の式によって表される．

$$\begin{aligned} P(j|e) &= \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} P(j,a|e) && (20) \\ &= \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \left(\begin{array}{c} m-\phi_0 \\ \phi_0 \end{array} \right) p_0^{m-2\phi_0} p_1^{\phi_0} \prod_{i=1}^{l} \phi_i! n(\phi_i|e_i) \times \\ &\qquad \prod_{j=1}^{m} t(j_j|e_{a_j}) d(j|a_j,m,l) && (21) \end{aligned}$$

モデル 3 では，全ての単語対応を考慮して計算するため，計算量が膨大となる．そのため，期待値は近似によって求められる．

model4

model3 と model4 の違いは，単語の位置の考慮の仕方である．model3 において，単語の位置は絶対位置で考慮していた．それに対して，model4 では単語の位置を相対位置で考慮する．また，各単語ごとの位置も考慮している．model4 では，単語位置の歪みの確率である $d(j|i,m,l)$ を以下の 2 通りで考慮する．

- 英単語に対応する日単語が 1 以上あるときに，その中で最も文頭に近い場合

$$P(\Pi_{[i]1}=j|\pi_1^{[i]-1},\tau_0^l,\phi_0^l,E)=d_1(j-\odot_{i-1}|\mathcal{A}(e_{[i-1]}),\mathcal{B}(j_j)) \tag{22}$$

- それ以外の場合

$$P(\Pi_{[i]k}=j|\pi_{[i]1}{}^{k-1},\pi_1^{[i]-1},\tau_0^l,\phi_0^l,E)=d_{>1}(j-\pi_{[i]k-1}|\mathcal{B}(j_j)) \tag{23}$$

model5

モデル 4 では，単語の位置に関して直前の単語のみを考慮している．そのため，複数の単語が同じ位置に生じたり，単語が存在しない位置に生成されるという問題がある．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に制約が施されている．

2.2 GIZA++

GIZA++[3] 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 翻訳モデルの model1 から model5 に基づいて，単語の対応関係の確率値を計算する．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，機械翻訳の一手法である．単語に基づく統計翻訳が用いられていたが，単語よりも句に基づく統計翻訳の方が精度が高く，現在では句に基づく統計翻訳が用いられている．句に基づく統計翻訳は，学習データとして大量の対訳文を用いることで，自動的に翻訳規則を生成し翻訳を行う．

2.3.1 句に基づく統計翻訳の概要

日英統計翻訳は日本語入力文 j が与えられたとき，翻訳モデルと言語モデルの組み合わせの中から翻訳が最大となる英語翻訳文 E を検索することで翻訳を行う．基本モデルを以下に示す．

$$E = arg\max_{e} P(e|j) \tag{24}$$

$$\simeq arg\max_{e} P(j|e)P(e) \tag{25}$$

ここで，$P(j|e)$ は翻訳モデル，$P(e)$ は言語モデルを表す．翻訳モデルは対訳学習文から学習し，言語モデルは目的言語の単言語学習文から学習する．そしてデコーダを用いて，$P(j|e)P(e)$ が最大となる英語翻訳文 E を検索する．日英統計翻訳の手順を表 1 に示す．

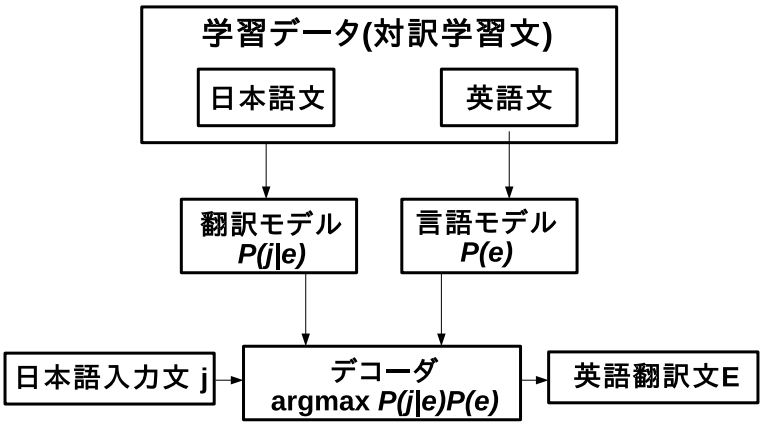

図 1: 日英統計翻訳の手順

2.3.2 フレーズテーブル作成法

GIZA++より IBM 翻訳モデルを推定することで最尤な単語確率を得る．これを日英，英日の両方向に対して行う．日本語文“ 私たちは映画を見に行く ”とその対訳英語文“ We go to watch the movie ”を例にあげ，日英方向の単語対応の例を表 2.1 に，英日方向の単語対応の例を表 2.2 に示す．また●は単語が対応した箇所を示す．

表 2.1: 日英方向の単語対応の例

	We	go	to	watch	the	movie
私たち	●					
は				●		
映画						●
を				●		
見				●		
に				●		
行く				●		

表 2.2: 英日方向の単語対応の例

	We	go	to	watch	the	movie
私たち	●					
は						
映画						●
を					●	
見		●	●	●		
に						
行く						

次に，両方向の対応付けからヒューリスティックスなルールにより，1 対多の対応を認めた単語対応の計算を行う．ここで，ヒューリスティックとは人間の日々の意思決定類似した直感的かつ発見的な思考方法である．基本のヒューリスティックとして”intersection” と”union” がある．intersection は，日英方向と英日方向の両方向に単語対応が存在する場合にその単語対応を残す．union は日英方向と英日方向のどちらか一方に単語対応が存在する場合にその単語対応を残す．intersection の例を表 2.3 に，union の例を表 2.4 に示す．

表 2.3: intersection の例

	We	go	to	watch	the	movie
私たち	●					
は						
映画						●
を						
見				●		
に						
行く						

表 2.4: union の例

	We	go	to	watch	the	movie
私たち	●					
は				●		
映画						●
を				●	●	
見		●	●	●		
に				●		
行く				●		

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 の例を表 2.5 に，grow-diag の例を表 2.6 に示す．

表 2.5: grow の例

	We	go	to	watch	the	movie
私たち	●					
は						
映画						●
を				●		
見				●		
に						
行く						

表 2.6: grow-diag の例

	We	go	to	watch	the	movie
私たち	●					
は						
映画						●
を				●	●	
見			●	●		
に				●		
行く						

grow-diag の最終処理として”final” と”final-and” がある．final は少なくとも一方の言語に対応がない場合に，union の単語対応を追加し，final-and は両言語単語に対応がない場合に，union の単語対応を追加する方法である．grow-diag-final の例を表 2.7 に，grow-diag-final-and の例を表 2.8 に示す．

表 2.7: grow-diag-final の例

	We	go	to	watch	the	movie
私たち	●					
は				●		
映画						●
を				●	●	
見		●	●	●		
に				●		
行く				●		

表 2.8: grow-diag-final-and の例

	We	go	to	watch	the	movie
私たち	●					
は						
映画						●
を				●	●	
見		●	●	●		
に				●		
行く						

得られた単語対応うち，矛盾しない全ての対訳句を得る．このとき，対訳句に対して翻訳確率を計算し，対訳句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 翻訳モデル

句に基づく翻訳モデルとは，確率的に日本語から英語の単語列へ翻訳を行うためのモデルである．統計翻訳において，句に基づく翻訳モデルとして，一般的にはフレーズテーブルが用いられている．フレーズテーブルは以下の手順で作成される．

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

詳しい作成手順については，2.3.2 節にて説明する．また，表 2.9 にフレーズテーブルの例を示す.

表 2.9: フレーズテーブルの例

突然 天気 が ||| Suddenly , the weather ||| 0.5 0.00217118 1 3.39949e-05 2.718 ||| 0-0 0-1 2-2 1-3 ||| 2 1 1
突然 天気 が 変わっ た ||| Suddenly , the weather changed ||| 0.5 9.13961e-05 0.5 4.2075e-06 2.718 ||| 0-0 0-1 2-2 1-3 3-4 4-4 ||| 2 2 1
突然 天気 が 変わっ た 。 ||| Suddenly , the weather changed . ||| 0.5 9.13961e-05 0.5 4.20734e-06 2.718 ||| 0-0 0-1 2-2 1-3 3-4 4-4 5-5 ||| 2 2 1

左から順に，日本語フレーズ，英語フレーズ，日英方向の翻訳確率 $P(j|e)$，日英方向の単語の翻訳確率の積，英日方向の翻訳確率 $P(e|j)$，英日方向の単語の翻訳確率の積，フレーズペナルティ，フレーズ内単語対応（日英方向）である．以後，フレーズペナルティは常に一定の値であるため省略する．

2.3.4 デコーダ

デコーダは，翻訳モデルと言語モデルの全ての組み合わせから確率が最大となる翻訳文を検索し出力する．代表的なデコーダとして，Moses がある．デコーダの手順を図 2 に示す.

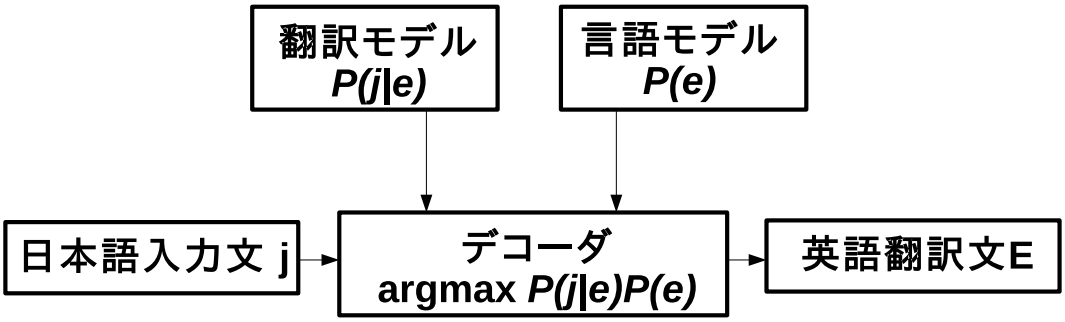

図 2: デコーダの手順

2.4 パターン翻訳

パターン翻訳は，機械翻訳の一手法であり，大量の単語辞書と対訳文パターンを用い翻訳を行う．パターン翻訳は適切に対訳文パターンが適合した場合に，文の構造を保持した翻訳精度の高い翻訳文を出力する傾向にある．しかし，単語辞書や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に翻訳ができない．

2.4.1 日英パターン翻訳の概要

手順１ 単語辞書と対訳文パターンを用意する．対訳文とは，大量の対訳文から任意の単語フレーズを変数化して得られる文パターンである．

手順２ 日本語入力文と日本語単語を単語辞書に用いて，英単語に翻訳する．

手順３ 変数部に対応する日本語単語を単語辞書を用いて，英単語に翻訳する．

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

手順５ 手順４で得た英語翻訳文を出力する．

日英パターン翻訳の手順を図 3 に示す．

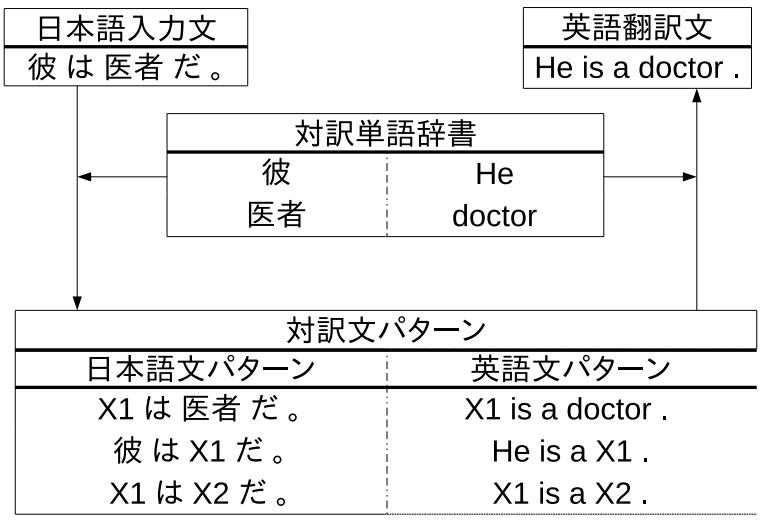

図 3: 日英パターン翻訳の流れ

2.5 パターンに基づく統計機械翻訳システム

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

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

2.5.1 パターンに基づく日英統計翻訳の概要

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

手順１ 対訳単語の作成

GIZA++を用いて，対訳単語を作成

手順２ 単語に基づく対訳文パターンの作成

対訳単語を用いて，単語に基づく対訳文パターン (以下，単語レベル文パターン) を作成．

手順３ 対訳句の作成

単語レベル文パターンを用いて，対訳句を作成．

手順４ 句に基づく対訳文パターンの作成

対訳句を用いて，句に基づく対訳文パターン (以下，句レベル文パターン) を作成．

手順５ 翻訳文作成

対訳句と句に基づく対訳文パターンを用いて，翻訳文生成を行う．

以下にそれぞれの手順の詳細を記述する．

2.5.2 対訳単語の作成

GIZA++を用いて，対訳文の単語対応を取り，対訳単語と単語翻訳確率を得る．図 4 に対訳単語作成例を示す.

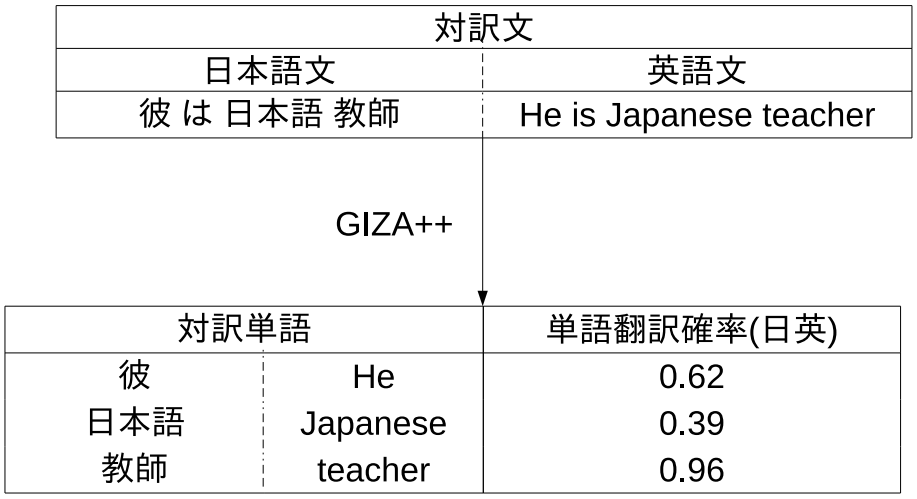

図 4: 対訳単語作成の例

2.5.3 単語レベル文パターンの作成

対訳単語と対訳文を用いて，単語レベル文パターンを作成する．まず，対訳単語と対訳文を照合する．そして，対訳文において，適合した対訳単語を変数化する．図 5 に単語レベル文パターン作成の例を示す.

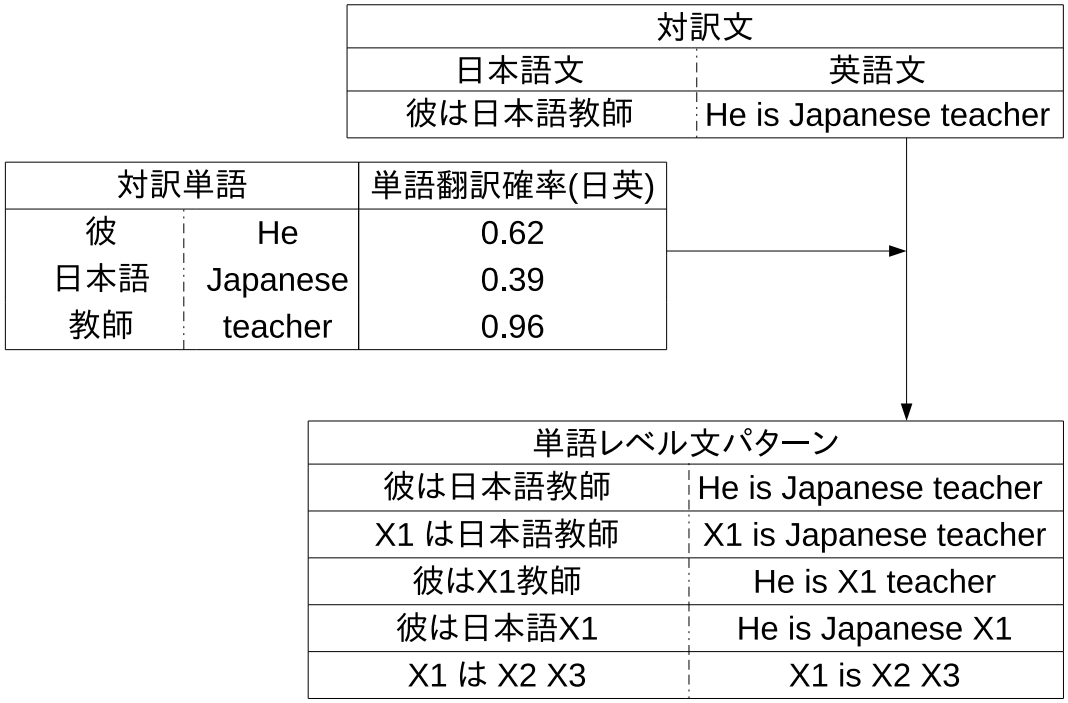

図 5: 単語レベル文パターン作成の例

2.5.4 対訳句の作成

1) 対訳句の抽出

単語レベル文パターンと対訳文を照合する．適合した場合，単語レベル文パターンの変数部に対応する単語を，対訳句として対訳文より抽出する．図 6 に対訳句抽出の流れを示す.

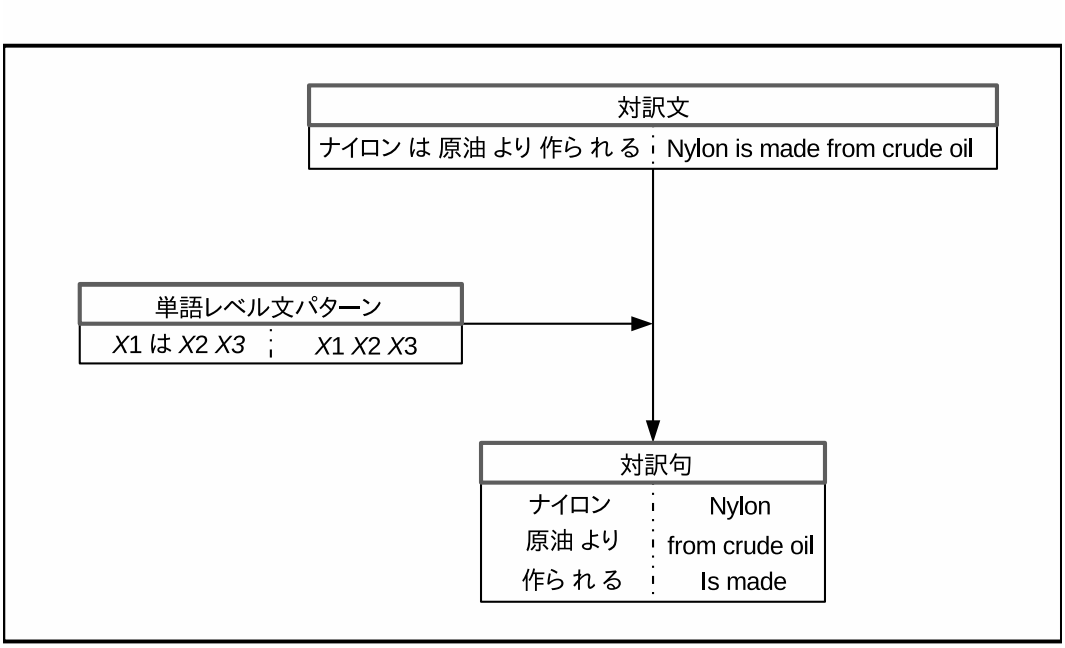

図 6: 対訳句作成の例

2) 対数フレーズ確率の付与

対訳単語と単語翻訳確率を用いて，対訳句に確率を付与する．まず，対訳句において日本語句の単語と英語句の単語の全ての組み合わせを得る．次に，日本語句の単語に対応する英語句の単語の中で，単語翻訳確率の最大値を得る．これを各日本語単語に対して行い，得られた値について対数の総和を求める．(以下，対数フレーズ確率)．同様に対訳句において，英単語に対応する日本語単語の中で単語翻訳確率の最大値を取得し，英日方向の対数フレーズ確率も求める．日英方向の対数フレーズ確率付与の例を図 7 に示す．

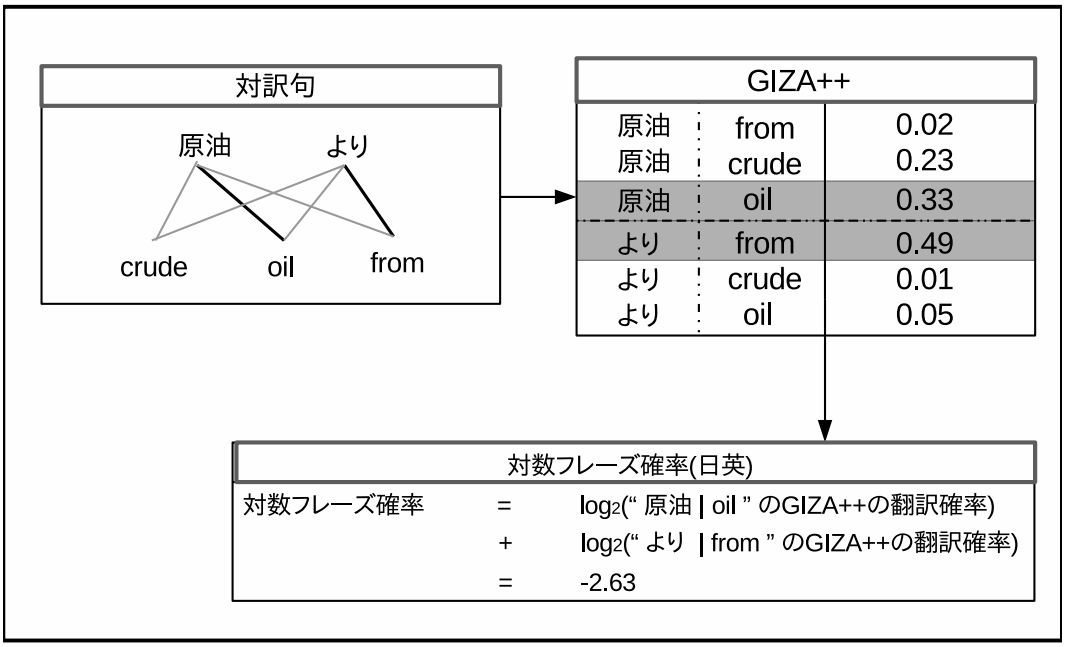

図 7: 対数フレーズ確率付与の例 (日英)

2.5.5　句レベル文パターンの作成

1) 対訳文パターンの作成

対訳句と対訳文を用いて，句レベル文パターンを作成する．作成方法は，単語レベル文パターンの作成と同様に変数の組み合わせを考慮して，句レベル文パターンを可能な限り多く作成する．句レベル文パターン作成の例を図 8 に示す.

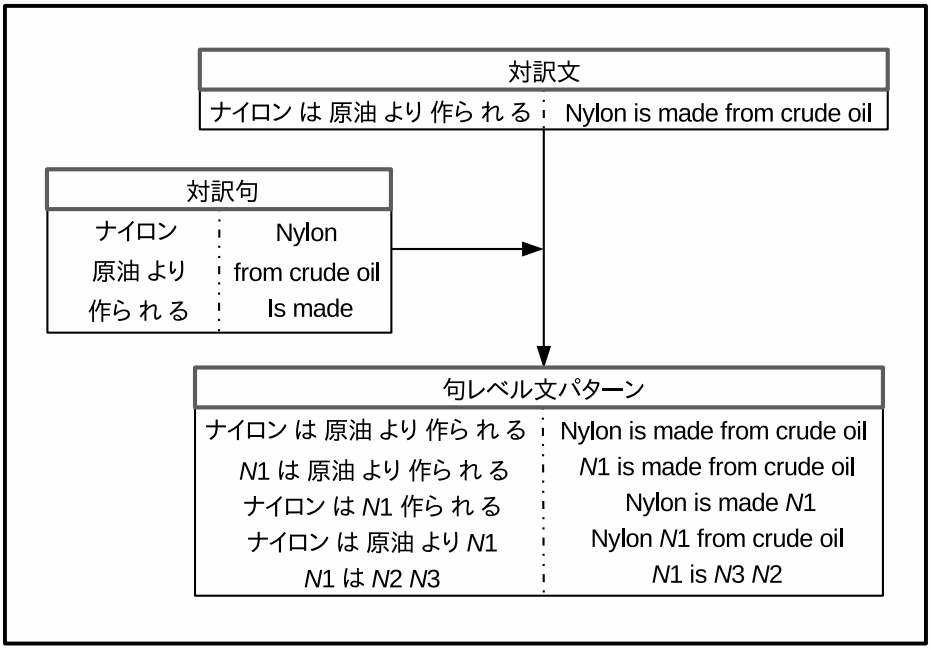

図 8: 句レベル文パターン作成の例

2) 対数文パターン確率の付与

対訳単語と単語翻訳確率を用いて，句レベル文パターンに確率を付与する．句レベル文パターンにおいて字面を用いて，対数フレーズ確率の付与と同様の計算手法で確率を求める．本研究では，この値を対数文パターン確率と呼ぶ．日英方向の対数文パターン確率付与の例を図 9 に示す.

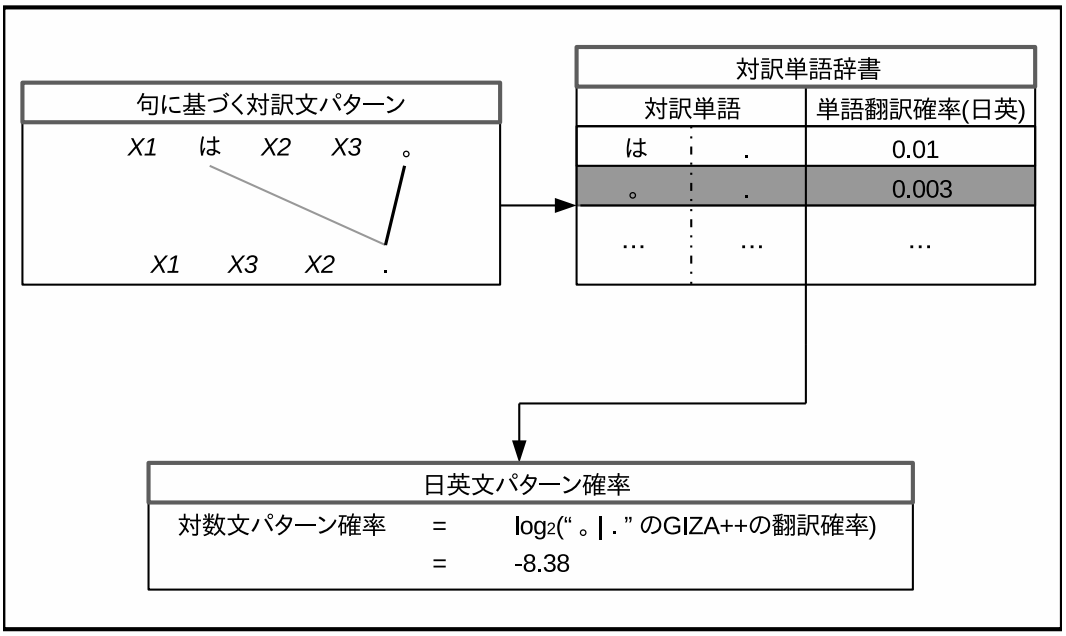

図 9: 対数文パターン確率付与の例 (日英)

2.5.6 翻訳文の作成

句レベル文パターンと対訳句を用いて，翻訳文を生成する．まず，日本語文パターンと入力文を照合し, 入力文に適合する日本語文パターンを選択する．なお，文パターンの選択には，入力文と日本語文パターンの字面を比較し，字面が多く一致する文パターンを選択する．そして，選択した文パターンにおいて，英語文パターンの変数部に対訳句を用いて英語句を置換し，翻訳候補文を生成する．この処理を各適合する文パターンに対して同様に行う．最後に，各翻訳候補文から翻訳文を選択するために，対訳文パターンの対数文パターン確率 (α) と対訳句の対数フレーズ確率 (β)，言語翻訳確率 (trigram = γ) の総和を用いる．各翻訳候補文の対訳文パターンの対数文パターン確率と対訳句の対数フレーズ確率，言語翻訳確率 (trigram) の総和を求め，翻訳候補文の中で総和が最大となる文を翻訳文として出力する．日英翻訳における翻訳文の生成例を図 10 に示す．

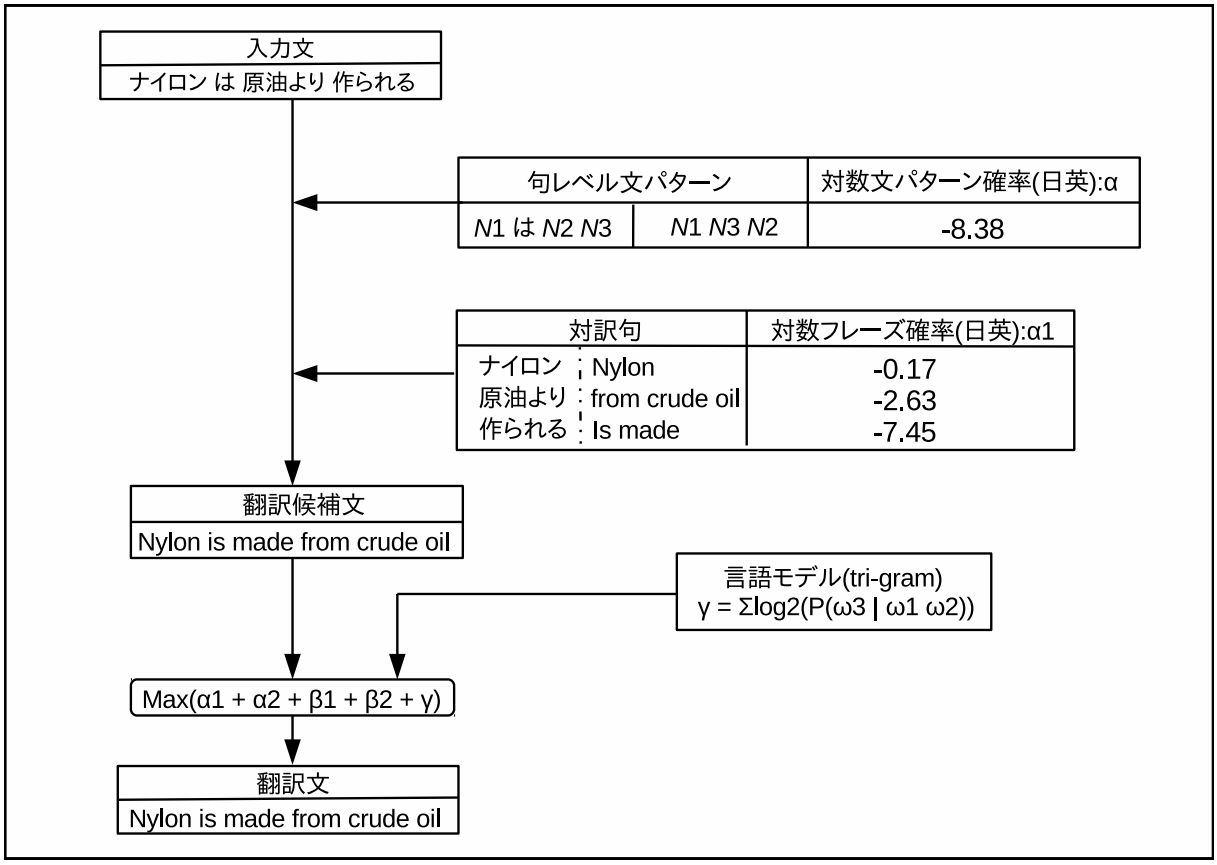

図 10: 翻訳文作成の例

3　先行手法

3.1　自動的な対訳句の作成における先行研究の概要

パターンに基づく統計翻訳で対訳句を自動作成する際，単語レベル文パターンから全ての可能な対訳句を出力するために，不適切な対応をとる対訳句が多く出力される．そこで先行研究として興梠ら [4] の研究が挙げられる．興梠らは，単語レベル文パターンごとに対数フレーズ確率の総和が最大の対訳句を出力する手法を提案した．この先行手法により不適切な対訳句を多く減らしている．

3.2　先行手法の手順

以下に先行手法の具体的な手順を示す．

手順 1 対訳文と単語レベル文パターンを照合する．

手順 2 適合した場合，単語レベル文パターンの変数部に対応する全ての組み合わせの対訳句を抽出する．

手順 3 GIZA++の単語確率を用いて，各組み合わせの中から最大となる単語確率を得る．

手順 4 得られた単語確率を用いて，対数フレーズ確率を計算する．

手順 5 各単語レベル文パターンごとに，手順 4 で計算した対数フレーズ確率の総和の最大値をとる対訳句を 1 つすつ選出する．

手順 6 手順 5 で選出した対訳句を最終的に抽出される対訳句として出力する．

具体例を図 11 に示す．対訳文と単語レベル文パターンは表 3.1 を用いる．

表 3.1: 先行手法照合例 1

対訳文 (日)	チケットショップが安売りし始めた
対訳文 (英)	The ticket shop began to sell at a bargain
パターン (日)	$X1$ が $X2$ $X3$ た
パターン (英)	The $X1$ $X3$ to $X2$

この手法により 1 つのパターンに対し，適切だと思われる 1 組の対訳句のみを出力することができる．図 11 より,「チケット ショップ」-「ticket shop」,「安売り し」-「sell at a bargain」,「始め」-「began」が抽出される．

変数	対訳句		対数フレーズ確率
X1	チケットショップ	ticket shop	-0.5
X2	安売りし	sell at a bargain	-0.2
X3	始め	began	-0.1
合計			-0.8
	:		
X1	チケットショップ	ticket	-0.7
X2	安売りし	sell at a bargain	-0.2
X3	始め	shop began	-0.7
合計			-1.6

チケットショップ	ticket shop
安売りし	sell at a bargain
始め	began

図 11: 先行手法抽出例 1

3.3 先行手法の問題点

先行手法で出力した対訳句を調査したところ，不適切な対応をとる対訳句が多く含まれていた．原因として考えられることは，対訳句抽出において対訳文 1 文につき複数パターンを適合した場合に，対訳文に対し不適切なパターンを適合してしまった場合に，不適切な対訳句が抽出されていることが挙げられる．その不適切な対訳句は先行手法では削除できない．図 12 に不適切な対訳句の抽出例を以下に示す．対訳文と単語レベル文パターンは表 3.2 を用いる．

表 3.2: 先行手法照合例 2

対訳文 (日)	チケットショップが安売りし始めた
対訳文 (英)	The ticket shop began to sell at a bargain
パターン (日)	$X1\ X2\ X3\ X4$
パターン (英)	$X1\ X2\ X4$ to $X3$

先行手法を用いて対訳句を抽出した際，変数 $X1$ の「チケットショップ」に対する対訳

変数	対訳句		対数フレーズ確率
X1	チケットショップ	The	-0.7
X2	が	ticket shop	-0.6
X3	安売りし	sell at a bargain	-0.2
X4	始めた	began	-0.05
合計			-1.55
:			
X1	チケットショップ	The ticket	-0.9
X2	が	shop	-0.6
X3	安売りし	sell at a bargain	-0.2
X4	始めた	began	-0.05
合計			-1.75

チケットショップ	The
が	ticket shop
安売りし	sell at a bargain
始めた	began

図 12: 先行手法抽出例 2

句 (英) が適切である「ticket shop」にならないパターンを適用してしまっている．その結果，$X1$ に「The」が対応してしまい，先行手法を用いた結果不適切な対訳句が出力される．図 12 より,「チケット ショップ」-「The」,「が」-「ticket shop」,「安売り し」-「sell at a bargain」,「始め」-「began」が抽出される．

4　提案手法

4.1　提案手法の概要

興梠らの先行手法は，対訳文と単語レベル文パターンを用いて対訳句を抽出する際，対訳文と照合したパターンごとに対数フレーズ確率の総和が最大の対訳句を出力し，１つのパターンにつき１組の対訳句を抽出していた．しかし，この手法では，対訳文に対し不適切な単語レベル文パターンを適合した場合に削除できないまま抽出し，不適切な対訳句を出力してしまう．

そこで本研究では，対訳句を抽出する際，単語レベル文パターンを作成する際に用いた対訳文 (以下，パターン原文) と対訳句を作成する際に用いる対訳文との類似度を利用して，先行手法で作成された不適切な対訳句を削除する手法を提案する．先行手法に加え提案手法を用いることで，不適切な対応をとる対訳句を削除することで，対訳句の精度向上を試みる．また，出力した対訳句を用いて日英統計翻訳を行い，翻訳精度の調査を試みる．

4.2　提案手法の手順

以下に具体的な手順を示す．

手順 1 対訳文と単語レベル文パターンを照合する．

手順 2 適合した場合，単語レベル文パターンの変数部に対応する全ての組み合わせの対訳句を抽出する．

手順 3 GIZA++の単語確率を用いて，各組み合わせの中から最大となる単語確率を得る．

手順 4 得られた単語確率を用いて，対数フレーズ確率を計算する．

手順 5 各単語レベル文パターンごとに，手順 4 で計算した対数フレーズ確率の総和の最大値をとる対訳句を１つずつ選出する．

手順 6 対訳句作成時の対訳文とパターン原文を抽出する．

手順 7 対訳文 (日) から見たパターン原文 (日) との類似度 A を作成する．
類似度は対訳文とパターン原文の同一の単語の出現率を表している．以下の式を用

いて計算する．

$Ps(\frac{B_1...B_N}{A_1...A_M}) = (\frac{S}{M})$

対訳文Aは単語 $A_1...A_M$ から構成されている

対訳文Bは単語 $B_1...B_N$ から構成されている

M; 対訳文A中の単語数

S; 対訳文A中の単語が対訳文Bの単語と
一致している単語数

手順 8 パターン原文 (日) から見た対訳文 (日) との類似度 B を作成する．
手順 7 と同様に計算する．

手順 9 英語文でも手順 7〜8 と同様に計算する．

手順 10 全てを掛けあわせ，類似度とする．

手順 11 類似度を用いて Nbest で対訳句を削除する．

手順 6 手順 5 で選出した対訳句を最終的に抽出される対訳句として出力する．

提案手法の具体例を表 4.1 を用いて，図 13 と 14 に示す．

表 4.1: 提案手法照合例

対訳文 (日)	チケットショップが安売りし始めた
対訳文 (英)	The ticket shop began to sell at a bargain
パターン 1 (日)	$X1$ $X2$ $X3$ $X4$
パターン 1 (英)	$X1$ $X2$ $X4$ to $X3$
パターン原文 1 (日)	雪 が 解け 始め た
パターン原文 1 (英)	The snow began to melt
パターン 2 (日)	$X1$ $X2$ $X3$ $X4$
パターン 2 (英)	$X1$ $X2$ $X4$ to $X3$
パターン原文 2 (日)	これ は 飲み やすい
パターン原文 2 (英)	This is easy to drink

対訳文(日)	チケット ショップ が 安売り し 始め た
対訳文(英)	The ticket shop began to sell at a bargain
単語レベル文パターン1(日)	X1 が X2 X3 た
単語レベル文パターン1(英)	The X1 X3 to X2
パターン原文1(日)	雪 が 解け 始め た
パターン原文1(英)	The snow began to melt

それぞれ抽出し，
右の式で計算する．

$$\text{類似度} = \frac{\text{一致している単語数}}{\text{着目している文の単語数}}$$

対訳文(日)の単語数	7	対訳文(日)から見たパターン原文1(日)との類似度	3 / 7
対訳文(英)の単語数	9	パターン原文1(日)から見た対訳文(日)との類似度	3 / 5
パターン原文1(日)の単語数	5	対訳文(英)から見たパターン原文1(英)との類似度	1 / 3
パターン原文1(英)の単語数	5	パターン原文1(英)から見た対訳文(英)との類似度	3 / 5
一致している単語(日)の数	3		
一致している単語(英)の数	3		

対訳文とパターン原文の類似度は
それぞれの類似度を掛けてあわせて計算する

対訳文とパターン原文1との類似度 ＝ 9 / 175

図 13: 提案手法抽出例 1

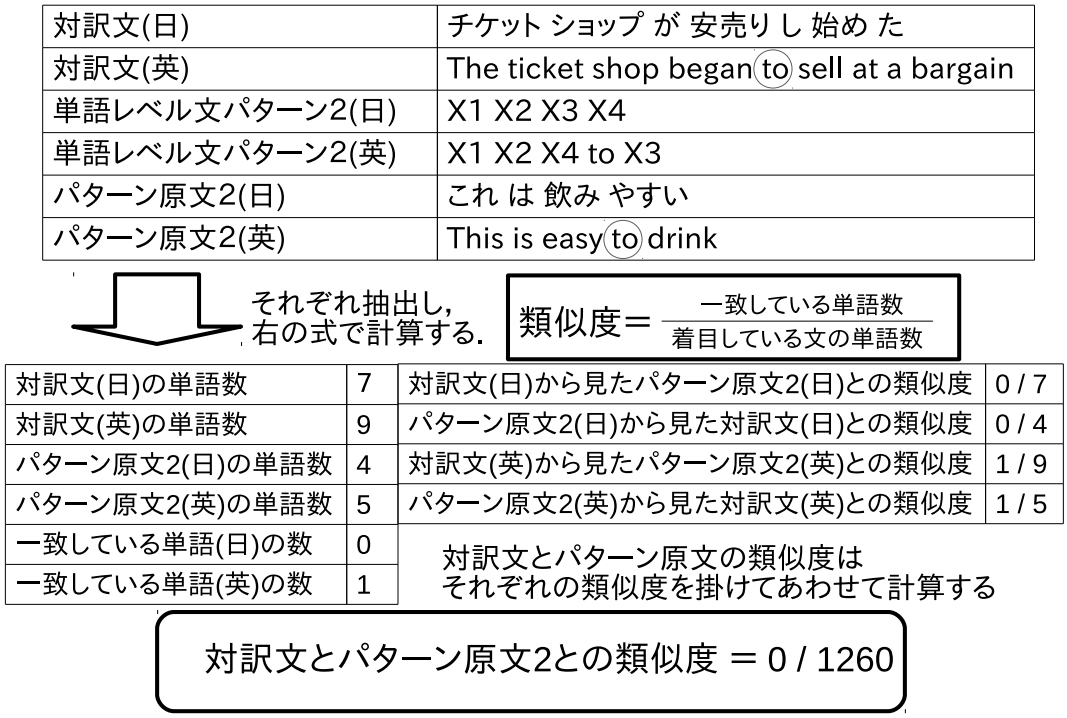

図 14: 提案手法抽出例 2

先行手法にさらに追加で提案手法を用いる．先行手法ではパターン 1 つにつき 1 組の対訳句が出力される．つまり，対訳句 (日) に対し複数の対訳句 (英) が対応する場合がある．その際に不適切な対訳句も含まれる．その不適切な対訳句を類似度を用いて削除する．図 13 では対訳文 1 とパターン原文 1 との類似度は 9/175，図 14 では対訳文 2 とパターン原文 2 との類似度は 1/1260，よって先行手法を用いて表 4.2 が抽出されていたとすると，類似度の高い対訳文 1 とパターン 1 から出力された対訳句である「チケット ショップ」-「The」が削除され,「チケットショップ」-「ticket shop」が残る．

表 4.2: 先行手法候補

チケット ショップ	ticket shop	9/175
チケット ショップ	The	0/1260

4.3 実験データ

対訳文および翻訳実験に用いるテスト文は，電子辞書から抽出した単文データを用いる [5]．なお，単文データは日本語文が単文であるのに対し，英語文は単文とは限らず，重文・複文が含まれる．対訳文および翻訳実験に用いるテスト文の例を表 4.3 に，使用するデータの内訳を表 4.4 に示す．

表 4.3: テスト文の例

日本語句	水 が 腐っ て いる 。
英語句	The water is foul .
日本語句	素行 を 改め なさい 。
英語句	You should mend your ways .
日本語句	彼 は 最後 の 断 を 下し た 。
英語句	He made a final decision .

表 4.4: 実験データ

対訳文	100,000 文
テスト文	200 文

4.4 評価方法

本研究は，対訳句の精度評価と翻訳文の精度評価を行う．具体的に対訳句の精度評価は提案手法を用いて作成した対訳句と先行手法を用いて作成した対訳句から，それぞれランダムに 100 句抽出し，人手で精度を評価する．翻訳文の精度評価はそれぞれの手法を用いて作成された対訳句を使用し統計翻訳を行い，翻訳文の精度の評価を行う．翻訳文の精度評価には，対比較評価を行う．

5　実験結果

5.1　対訳句の精度評価

各手法により抽出した対訳句からそれぞれランダムに 100 句取り出し，対訳句の精度を人手で評価した．評価基準を以下に示す．

○:適切な対応をとる

△:意味が欠落している，不必要な意味が付与されている

×:不適切な対応をとる

評価結果を表 5.1 に示す．

表 5.1: 対訳句の評価結果 (100 句)

	総数	○	△	×
先行手法	94,028	52	14	34
提案手法	77,880	61	27	12

表 5.1 より，提案手法が先行手法よりも精度が高く，提案手法が有効であることが確認できた．

また，各手法の評価基準における対訳句の例を示す．先行手法の例を表 5.2 に示す．

表 5.2: 先行手法の対訳句の例

○	△	×
母からの:from my mother	捨てた:He threw away	は この:He has
向けられ:turned on	スペースシャトル:The space	円:one million
休んだ:absent from	神経:nerves were	近づいている:A storm

また表 5.2 の対訳句において抽出に用いた単語レベル文パターンとそのパターン原文を表 5.3〜5.5 に示す．

表 5.3 は表 5.2 の評価◯である「母からの」の抽出に用いたデータである.

表 5.3: 対訳句「母からの」の詳細

対訳句 (日)	母 から の
対訳句 (英)	from my mother
対訳文 (日)	その 時 私 は 母 から の 手紙 を 読ん で い ま し た 。
対訳文 (英)	At that time , I was reading a letter from my mother .
単語レベル文パターン (日)	$X02$ は $X00$ $X01$ を $X03$ 。
単語レベル文パターン (英)	$X02$ $X03$ $X01$ $X00$.
パターン原文 (日)	彼女 は よく 英語 を 話す 。
パターン原文 (英)	She speaks English well .

表 5.4 は表 5.2 の評価△である「捨てた」の抽出に用いたデータである.

表 5.4: 対訳句「捨てた」の詳細

対訳句 (日)	捨て た
対訳句 (英)	He threw away
対訳文 (日)	彼 は 試合 を 捨て た 。
対訳文 (英)	He threw away the game .
単語レベル文パターン (日)	$X00$ を $X01$ 。
単語レベル文パターン (英)	$X01$ the $X00$.
パターン原文 (日)	距離 を 測る 。
パターン原文 (英)	Measure the distance .

表 5.5 は表 5.2 の評価×である「円」の抽出に用いたデータである.

表 5.5: 対訳句「円」の詳細

対訳句 (日)	円
対訳句 (英)	one million
対訳文 (日)	帳簿 に １００ 万 円 の 穴 が あい た 。
対訳文 (英)	There is a deficit of one million yen in the account .
単語レベル文パターン (日)	$X05$ $X02$ $X03$ $X01$ $X06$ $X00$ $X04$ た 。
単語レベル文パターン (英)	$X02$ $X05$ $X04$ $X00$ $X06$ $X01$ the $X03$.
パターン原文 (日)	車 は ホテル の 前 で 止まっ た 。
パターン原文 (英)	The car stopped in front of the hotel .

次に提案手法の例を 5.6 に示す.

表 5.6: 提案手法の対訳句の例

○	△	×
全速力:full speed	かばん の チャック:bag open	は 勇敢 に も その:expose the
よく 聞く:listen well	手 に 包帯:with a bantage	32 段:without stoping once
の 研究 に:the study of	自分 の 家 を 抵当:own house	気 が 抜け:music

また表 5.6 の対訳句において抽出に用いた単語レベル文パターンとそのパターン原文を表 5.7〜5.9 に示す.

表 5.7 は表 5.6 の評価◯である「全速力」の抽出に用いたデータである.

表 5.7: 対訳句「全速力」の詳細

対訳句 (日) 対訳句 (英)	全速力 full speed
対訳文 (日) 対訳文 (英)	彼 は 全速力 で 自動車 を 運転 し た 。 He drove his car at full speed .
単語レベル文パターン (日) 単語レベル文パターン (英)	$X03$ は $X02$ $X00$ $X01$ を $X04$ し た 。 $X03$ $X04$ $X01$ $X00$ $X02$.
パターン原文 (日) パターン原文 (英)	彼 は 大学 で 演劇 を 勉強 し た 。 He studied drama at college .

表 5.8 は表 5.6 の評価△である「かばんのチャック」の抽出に用いたデータである.

表 5.8: 対訳句「かばんのチャック」の詳細

対訳句 (日) 対訳句 (英)	かばん の チャック bag open
対訳文 (日) 対訳文 (英)	彼女 は かばん の チャック を 開け た 。 She zipped her bag open .
単語レベル文パターン (日) 単語レベル文パターン (英)	$X00$ は $X02$ を $X01$ た 。 She $X01$ $X00$ $X02$.
パターン原文 (日) パターン原文 (英)	彼女 は 目 を 伏せ た 。 She lowered her eyes .

表 5.9 は表 5.6 の評価×である「は勇敢にもその」の抽出に用いたデータである．

表 5.9: 対訳句「は勇敢にもその」の詳細

対訳句 (日)	は 勇敢 に も その
対訳句 (英)	expose the
対訳文 (日)	彼 は 勇敢 に も その スキャンダル を 暴露 し た 。
対訳文 (英)	He had the courage to expose the scandal .
単語レベル文パターン (日)	$X02$ $X00$ $X03$ $X01$ $X04$ た 。
単語レベル文パターン (英)	He $X04$ $X01$ $X03$ to $X00$ $X02$.
パターン原文 (日)	ショート に ライナー を 打っ た 。
パターン原文 (英)	He hit a liner to the shortstop .

5.2 翻訳文の精度評価

各手法により抽出した対訳句と対訳文 100,000 文を用いて統計翻訳を行い、翻訳文の精度調査を行った．人手評価には，提案手法と先行手法の翻訳文の対比較評価を行う．提案手法と先行手法との対比較評価結果を表 5.10 に示す．

表 5.10: 翻訳の人手評価

提案手法○	先行手法○	差なし	同一文
36	31	35	12

表 5.10 の結果より，提案手法と先行手法を比較して，あまり大きな差はなかった．よって本研究で提案した手法は翻訳にあまり影響を与えないことがわかった．

提案手法○の例を表 5.11〜5.12 に，先行手法○の例を表 5.13〜5.14 に，差なしの例を表 5.15〜5.16 に，同一出力の例を表 5.17 に示す．

(a) 提案手法○の出力例

表 5.11 において，先行手法は「championship」が余分に含まれているため，提案手法○とした.

表 5.11: 提案手法○の出力例 1

入力文	その 店 は 、新しい 経営 陣 の 下 で 再開 し た 。
正解文	The store has reopened under new management .
先行手法	The store resumed the under championship in the, new its management .
パターン (日)	$X03$ は $X02$ $X00$ $X01$ の $X04$ で $X05$ し た 。
パターン (英)	$X03$ $X05$ the $X04$ championship in the $X02$ $X00$ $X01$.
パターン原文 (日)	$X03$ は $X02$ $X00$ $X01$ の $X04$ で $X05$ し た 。
パターン原文 (英)	$X03$ $X05$ the $X04$ championship in the $X02$ $X00$ $X01$.
提案手法	The store has reopen under a new management .
パターン (日)	$X03$ $X01$ 新しい $X04$ $X02$ $X00$ し た 。
パターン (英)	$X01$ $X03$ has $X00$ $X02$ new $X04$.
パターン原文 (日)	先生 が 新しい 本 を お 書き し た 。
パターン原文 (英)	The teacher has written a new book .

表 5.12 において，提案手法の方が先行手法より「是非」の意味が読み取れるので，提案手法○とした.

表 5.12: 提案手法○の出力例 2

入力文	この 方法 の 是非 は 簡単 に は 決め られ ない 。
正解文	We cannot tell at once whether this method .
先行手法	This method whether not arranged that easily .
パターン (日)	$X04$ の $X02$ は $X03$ $X01$ 決め られ $X00$ 。
パターン (英)	$X04$ $X02$ $X01$ arranged $X00$ $X03$.
パターン原文 (日)	彼 の 歓迎 会 は 月曜 に 決め られ た 。
パターン原文 (英)	His reception is arranged for Monday .
提案手法	This method can't decided on the easily of right or wrong .
パターン (日)	$X03$ $X02$ $X00$ $X01$ $X05$ $X04$ られ ない 。
パターン (英)	$X03$ can't $X04$ $X01$ $X05$ $X02$ $X00$.
パターン原文 (日)	彼 は ぜいたく の 味 が 忘れ られ ない 。
パターン原文 (英)	He can't forget the taste of luxury .

(b) 先行手法○の出力例

表 5.13 において，提案手法は文の形が誤っていて，先行手法の主語が正解文と一致しているので，先行手法○とした.

表 5.13: 先行手法○の出力例 1

入力文	私 の 疑惑 は 大きく なっ た 。
正解文	My suspicion grew .
先行手法	My Suspicion became louder .
パターン (日)	私 の $X00$ は $X01$ $X02$ 。
パターン (英)	My $X00$ $X02$ $X01$.
パターン原文 (日)	私 の 時計 は 少し 進む 。
パターン原文 (英)	My watch gains a little .
提案手法	The grown my suspicion .
パターン (日)	私 の $X02$ $X00$ $X01$ た 。
パターン (英)	$X00$ $X01$ my $X02$.
パターン原文 (日)	私 の 服 の 趣味 を 当てこすっ た 。
パターン原文 (英)	He made a dig at my taste in clothes .

表 5.14 において，先行手法の主語が正解文と一致しているため，先行手法○とした.

表 5.14: 先行手法○の出力例 2

入力文	冬 は 太陽 が 早く 沈む 。
正解文	The sun sets early in winter .
先行手法	The sun sink early winter .
パターン (日)	$X03$ は $X01$ が $X02$ $X00$ 。
パターン (英)	The $X01$ $X00$ $X02$ $X03$.
パターン原文 (日)	明日 は 気温 が 高く なる でしょ う 。
パターン原文 (英)	The temperature will be high tomorrow .
提案手法	Winter sink early in the sun .
パターン (日)	$X04$ は $X02$ $X00$ $X03$ $X01$ 。
パターン (英)	$X04$ $X01$ $X03$ in $X00$ $X02$.
パターン原文 (日)	彼 は 川 に 釣り に 行っ た 。
パターン原文 (英)	He went fishing in the river .

(c) 差なしの出力例

表 5.15 において，先行手法と提案手法の双方とも “由来している” という意味が読み取れるため，差なしとした.

表 5.15: 差なしの出力例 1

入力文	その 名 は ここ に 由来 し て いる 。
正解文	Hence the name .
先行手法	The name is traceable here .
パターン (日)	$X02$ は $X00$ $X01$ し て いる 。
パターン (英)	$X02$ is $X01$ $X00$.
パターン原文 (日)	彼 は ひどく 出血 し て いる 。
パターン原文 (英)	He is bleeding badly .
提案手法	The name is origin in here .
パターン (日)	その $X02$ は $X00$ $X01$ し て いる 。
パターン (英)	The $X02$ is $X01$ in $X00$.
パターン原文 (日)	その 銅 は 炎 の 中 で 酸化 し て いる 。
パターン原文 (英)	The copper is oxidizing in the flame .

表 5.16 において，入力文が 0 型代名詞のため「私」,「彼」のどちらも明記されていないので，差なしとした.

表 5.16: 差なしの出力例 2

入力文	コーヒー を テーブル に こぼし た 。
正解文	I spilled coffee on the table .
先行手法	I spilled the coffee table .
パターン (日)	$X02$ を $X00$ に $X01$ た 。
パターン (英)	I $X01$ the $X02$ $X00$.
パターン原文 (日)	窓 ガラス を きれい に ふい た 。
パターン原文 (英)	I wiped the windowpanes clean .
提案手法	He spilled his coffee table .
パターン (日)	$X00$ を $X01$ に こぼし た 。
パターン (英)	He spilt his $X00$ $X01$.
パターン原文 (日)	インク を 机 の 上 に こぼし た 。
パターン原文 (英)	He spilt his ink on the desk .

(d) 同一出力の出力例

表 5.17 において，先行手法と提案手法の出力文が完全に同一であったため，同一出力とした.

表 5.17: 同一出力の出力例

入力文	ちょっと 待っ て 下さい 。
正解文	Wait a minute , please .
先行手法	Please wait a moment .
パターン (日)	$X00$ $X01$ て 下さい 。
パターン (英)	Please $X01$ $X00$.
パターン原文 (日)	受付 で 尋ね て 下さい 。
パターン原文 (英)	Please ask at the reception desk .
提案手法	Please wait a moment .
パターン (日)	ちょっと 待っ $X00$ 。
パターン (英)	$X00$ wait a moment .
パターン原文 (日)	ちょっと 待っ て ください 。
パターン原文 (英)	Please wait a moment .

6　考察

6.1　対訳句抽出における提案手法の有効性

実験結果より，提案手法は先行手法と比べ適切な対訳句が多く不適切な対訳句の数を大幅に減らしていることから，対訳句の精度向上に対して提案手法の有効性が確認できた．しかし，翻訳に用いて実験を行った結果，先行手法と提案手法の精度は大きく差がないことがわかった．

6.2　誤り解析

対訳句の精度が向上したが翻訳実験の結果で大きく差がないことがわかった．そこで先行手法○の解析を行った．解析の結果，いくつかの原因が確認できた．

6.2.1　0 型代名詞を含む文から作成されたパターンを利用

原因の一つとしては 0 型代名詞を含む文から作成されたパターンを照合されているため不適切な文が出力されたと考えられる．表 6.1 に詳細を示す．

表 6.1: 誤り解析文詳細 1

入力文	私 の 疑惑 は 大きく なっ た 。
正解文	My suspicion grew .
先行手法	My Suspicion became louder .
パターン (日)	私 の $X00$ は $X01$ $X02$ 。
パターン (英)	My $X00$ $X02$ $X01$.
パターン原文 (日)	私 の 時計 は 少し 進む 。
パターン原文 (英)	My watch gains a little .
提案手法	The grown my suspicion .
パターン (日)	私 の $X02$ $X00$ $X01$ た 。
パターン (英)	$X00$ $X01$ my $X02$.
パターン原文 (日)	私 の 服 の 趣味 を 当てこすっ た 。
パターン原文 (英)	He made a dig at my taste in clothes .

翻訳に使われる対訳句と句レベル文パターンには全て対数翻訳確率が付与されていて，翻訳の際，対数確率の合計が一番高いものを出力する．表 6.1 の提案手法の出力文はパターン以外の対数確率値が少し高く，パターンの対数確率が低いため出力されている．提案手法で出力されているパターンの対数確率が表 6.3 のように「-23.581」に対し，適切なパターンの対数確率は「-25.425」である．

表 6.2: 誤り解析パターンの対数確率

私の $X02$ $X00$ $X01$ た 。	$X00$ $X01$ my $X02$.	-23.581
私の $X01$ は $X00$ た 。	My $X01$ $X00$.	-25.425

上のパターンは

$$PATTERN = (\frac{私の}{my}) + (\frac{た}{my})$$

と

$$PATTERN = (\frac{my}{私の}) + (\frac{my}{た})$$

とで計算される．ここで$(\frac{私の}{my})$は「私の」が「my」に訳される対数確率を表し，PATTERNは対数文パターン確率を表す．このとき「た」が「my」に訳される対数確率は対応が不適切ではあるが付与されていて，このパターン全体の確率に影響を与える．このひらがな１文字の間違った対応の対数確率が低いとパターンそのものの対数確率は低くなり，翻訳に選ばれにくくなる．しかし，パターンにおけるひらがな１文字の対応はあまり重要ではない．よってひらがな１文字の確率の重みを適切な重みに変えることで表 6.1 の文の精度は上がるのではないかと考えられる．

6.2.2 日本語パターンには主語を含まないが英語パターンには主語を含むパターンを利用

原因の一つとして日本語パターンには主語を含まないが英語パターンには主語を含むパターンを照合されているため不適切な文が出力されたと考えられる．表 6.3 に詳細を示す．

表 6.3: 誤り解析文詳細２

入力文	患者 は ゆっくり 体 を 起こし た 。
正解文	The patient slowly raised himself in bed .
提案手法	He body slowly patient up .
パターン 1(日)	$X02$ $X00$ $X01$ 起こし た 。
パターン 1(英)	He $X01$ $X00$ $X02$ up .
パターン原文 1(日)	彼 は その 老人 を 助け 起こし た 。
パターン原文 1(英)	He (helped) (the old) (man) up .
パターン 2(日)	$X02$ は $X01$ $X03$ を $X00$ た 。
パターン 2(英)	He $X01$ $X00$ $X02$ $X03$.
パターン原文 2(日)	彼 は 突然 歩調 を 速め た 。
パターン原文 2(英)	He (suddenly) (quickened) (his) (pace) .

表 6.3 は入力文の主語「患者」に対して，出力文の主語が「He」になってしまった文である．これはパターン 1 の作成の際に「He」が残ったからだと思われる．原因としては「その老人」が「the old」に対応し,「彼」が「man」に対応してしまったからと考えられる．これによりパターン 1(日) には主語を含まないが，パターン 1(英) には主語を含むというパターン対が作られ，このパターンを適用すると不適切な文が出力されやすくなる．日本語パターンには主語を含まないが英語パターンには主語を含むパターンはパターン 2 のような例もある．パターン 2 は「彼」に対して「He」と「his」が対応しなければならないが，本システムは 1 対 2 の対応は対応できないため「彼」が「his」に対応して,「He」が残ったパターンである．パターン 1 は「その老人」が「the old man」に対応させることが出来れば「He」が残らないパターンが作られると思われる．

6.2.3 文構造の違うパターンを利用

原因の一つとして文構造の違うパターンを照合されているため不適切な文が出力されたと考えられる．表 6.4 に詳細を示す．

表 6.4: 誤り解析文詳細 3

入力文	最後 に 皆 で 校歌 を うたっ た 。
正解文	At the end we all sang our school song together .
提案手法	(In the end) sang the (school song) in a (everyone) .
パターン (日)	$X02$ $X01$ で $X00$ を うたっ た 。
パターン (英)	$X02$ sang the $X00$ in a $X01$.
パターン原文 (日)	彼女 は 低い 調子 で その 歌 を うたっ た 。
パターン原文 (英)	(She) sang the (song) in a (low tone) .

表 6.4 は入力文が主語+「で」に対し，主語でないもの+「で」のパターンを適用してしまったため，主語のない不適切な文が出力された文である．これはパターンに基づく統計翻訳では「皆」と「低い調子」の差別化が難しく，改善が困難であるといえる．以上 3 つの誤り解析は先行手法○のうちそれぞれ約 20 %ずつ占めている．対訳句の改善よりも句レベル文パターンを改善が今後の課題といえる．

7　おわりに

先行手法において，対訳文に対し不適切な単語レベル文パターンを照合した際，不適切な対訳句が出力され，それにより対訳句の精度は低かった．そこで先行手法に加え，対訳句を抽出する際に，単語レベル文パターンを作成する際に用いた対訳文と対訳句を作成する際に用いる対訳文との類似度を利用して，不適切な対訳句を削除する手法を提案した．提案手法により，不適切な対応をとる対訳句を削除することによって対訳句の精度を向上させ，翻訳の精度の向上を試みた．

対訳句の精度評価の結果，提案手法を用いて作成した対訳句は先行手法を用いて作成した対訳句より精度が良く，不適切な対訳句が削除されていることがわかった．またそれぞれの手法で翻訳を行い翻訳文の対比較評価を行った結果，精度に大きな差はなかった．誤り解析を行ったところ，翻訳に用いる句レベル文パターンに付与されている対数文パターン確率に問題があると推測された．今後は，句レベル文パターンの問題を解決していけたらと思う．また新しい V13.2 の精度は○ 69 △ 22 × 9 となった．

謝辞

最後に，一年間に渡り，本研究のご指導をいただ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，村田真樹教授に深く感謝すると共に，厚く御礼申し上げます．そして，日常の議論を通じて多くの知識や示唆を頂いた同研究室の皆様に深謝いたします．また，参考にさせていただいた論文の著者の方々に対して，深く感謝申し上げます.

参考文献

[1] 江木孝史 ”句に基づく対訳文パターンの自動作成と統計的手法を用いた英日パターン翻訳”，言語処理学会 第 20 回年次大会

[2] 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.

[3] GIZA++
http://www.fjoch.com/GIZA++

[4] 興梠 玲架 ”パターンに基づく統計翻訳において変数部の確率の総和を使った対訳句の抽出”，鳥取大学 卒業論文

[5] 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, pp.119-130. 2012.

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

[7] 藤原勇: “パターン翻訳を用いた学習データ増加手法の検討”，修士論文，pp.43-59，2013.

Franz Josef Och: “Minimum Error Rate Training in Statistical Machine Translation”, In Proceeding of the 41st Annual Meeting of the Association for Computational Linguistics, pp.160-167, 2003.

Papineni Kishore, Salim Roukos, Todd Ward, Wei-Jing Zhu: “BLEU: a method for automatic evaluation of machine translation”, 40th Annual meeting of the Association for Computational Linguistics pp. 311-318, 2002.

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, pages 177-180, June 2007.