概要

パターン翻訳は，人手で作成した大量の対訳文パターンと対訳句 (単語や節を含む) を用いて翻訳を行う方法である．パターン翻訳は入力文が文パターンに適合した場合は翻訳精度の高い文が得られる．しかし，対訳文パターンと対訳句を人手で大量に作成するには時間がかかる．

また近年，機械翻訳において，統計的機械翻訳が注目されている．統計的機械翻訳は対訳文から自動的に翻訳規則を生成し，翻訳を行う方法である．統計的機械翻訳における対訳句の抽出方法として，Och らの方法がある．しかし，この方法は人間が見ると不自然な対訳句を抽出してしまう問題がある．

本研究では，対訳文パターンを用いた対訳句の抽出方法を提案する．対訳文パターンを人手で大量に作成するにはコストがかかる．そこで本研究では対訳文パターンを自動作成する．そして，対訳文パターンを用いて，対訳テスト文から対訳句を抽出する．実験の結果，6,264 句を抽出した．

目 次

図 目 次

表 目 次

第1章　はじめに

パターン翻訳は，人手で作成した大量の対訳文パターンと対訳句 (単語や節を含む) を用いて翻訳を行う方法である．パターン翻訳は入力文が文パターンに適合した場合は翻訳精度の高い文が得られる．しかし，対訳文パターンと対訳句を人手で大量に作成するには時間がかかる．

その問題に対して，道祖尾らは，日本語英語間において，N-gram を利用して，日英対訳パターンの候補を自動抽出した [1]．道祖尾らの日英対訳パターンとは，熟語や連語のような意味的まとまりを持つ表現である．実験の結果，人手評価より，約 8 割の候補において，日英対訳パターンの作成が可能であると報告した．北村らは，日本語英語間において，Dice 係数と単語の出現回数による閾値を用いて，日英対訳の表現を自動抽出した [2]．その結果，閾値が低下した場合においても 80〜90%の適合率で対訳表現の抽出を報告した．

また近年，機械翻訳において，統計的機械翻訳 (以下，SMT と表記) が注目されている．SMT は対訳文から自動的に翻訳規則を生成し，翻訳を行う方法である．SMT における対訳句の抽出方法として，Och らの方法 [3, 4] や，BerkeleyAligner[5] における抽出方法がある．Och らの方法はまず，IBM モデル [6] を用いて単語対応を求める．そして，単語対応よりヒューリスティックを用いて，網羅的に対訳句を抽出する．しかし，この方法は人間が見ると不自然な対訳句を抽出してしまう問題がある．

本研究では，対訳文パターンを用いた対訳句の抽出方法を提案する．対訳文パターンを人手で大量に作成するにはコストがかかる．そこで本研究では対訳文パターンを自動作成する．具体的には，対訳文パターンの自動作成方法として，西村らの方法 [7] を用いる．そして，対訳文パターンを用いて，対訳テスト文から対訳句を抽出する．実験の結果，6,264 句を抽出した．さらに，人手評価において，Och らの方法よりも優れていることを示した．

本論文の構成は以下の通りである．第 2 章で日英パターン翻訳システムについて説明し，第 3 章で Och らの方法による対訳句の抽出方法を説明する．第 4 章で提案する対訳句の抽出方法について説明し，第 5 章で本研究で使用するデータベースや閾値について

説明する．第6章で実験結果を示し，第7章で考察を述べる．

第2章　日英パターン翻訳システム

2.1　パターン翻訳の概要

パターン翻訳は，大量の対訳文パターンと対訳句 (単語や節を含む) を用いて，対訳文パターンの照合を行い翻訳文を出力する方法である．パターン翻訳は対訳文パターンが適合した場合，文全体の構造を保持した翻訳精度の高い翻訳文を得ることができる．しかし，一般的なパターン翻訳は対訳文パターンや対訳句を人手で作成するため，開発に時間がかかる．

2.2　日英パターン翻訳の手順

一般的な日英パターン翻訳の手順を以下に示す．

手順 1 対訳文パターンと対訳句を用意する．

手順 2 対訳文パターンと対訳句を用いて，入力文と日本語文パターンを照合する．

手順 3 照合に成功した場合，日本語文パターンに対応する英語文パターンを得る．

手順 4 英語文パターンの変数部を，対訳句を参照し英語句に置き換える．

手順 5 手順 4 で生成した翻訳文を出力する．

日英パターン翻訳の手順を図 2.1 に示す．

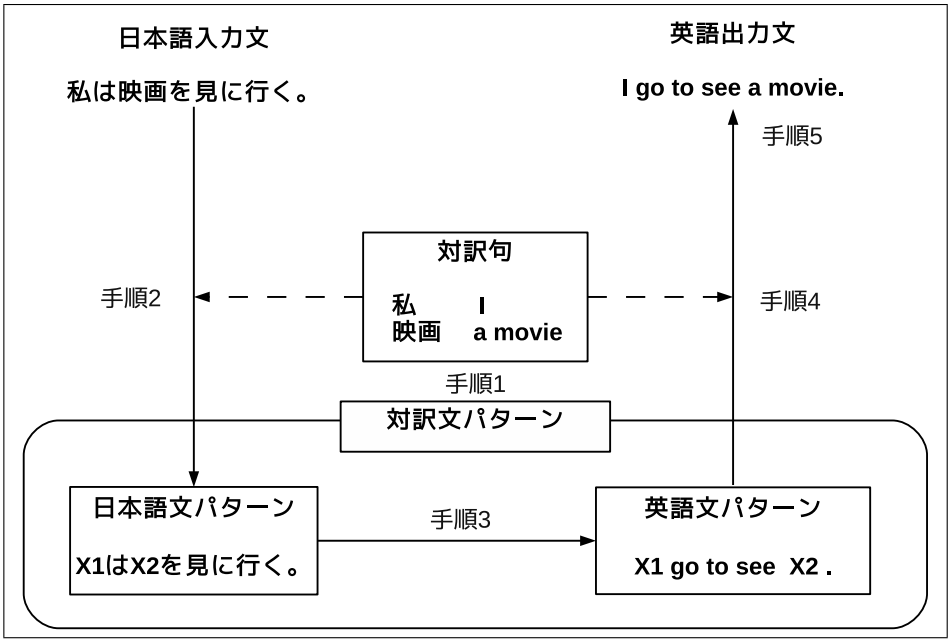

図 2.1 日英パターン翻訳の手順

2.3 対訳文パターン

対訳文パターンとは，大量の対訳文から単語を変数化により置き換えることで得られる文パターンである．表 2.1 に例を示す．なお，本研究における対訳文パターンの作成手順は 4.1.2 節で示す．

表 2.1 対訳文パターンの例

対訳文	(日本語)	彼女 は 熱 が ある 。
	(英語)	She has a fever .
対訳文パターン	(日本語)	$X1$ は $X2$ が ある 。
	(英語)	$X1$ has a $X2$.

2.4 対訳句

対訳句とは，異なる言語において，同じ意味を有する単語のまとまりの対である．日英対訳句の例を表 2.2 に示す．

表 2.2 日英対訳句の例

日本語句	英語句
あの 人	That person
とても よく 似合う	very becoming
月 の 光	The moonlight

なお，本研究では，少なくとも一方の言語が 2 単語以上で構成される対を対訳句として用いる．

第3章　Ochらの方法による対訳句の抽出

Och らの方法による対訳句の抽出は，まず IBM モデルを学習し，対訳単語の対応 (以下，単語対応と表記) を得る．次に，単語対応からヒューリスティックを用いて網羅的に対訳句を抽出する．Och らの方法による対訳句の抽出手順を以下に示す．

3.1　単語対応の獲得

IBM モデルを用いて，単語対応を得る．具体的には GIZA++[8] を用いて IBM モデルを学習し，単語対応を得る．

3.1.1　IBM モデル

SMT における単語対応を得るための代表的なモデルとして，IBM モデルがある．IBM モデルは model1 から model5 までの 5 つのモデルからなる．IBM モデルでは原言語の日本語文 J，目的言語の英語文 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．

$$P(J|E) = \sum_{a} P(J, a|E) \tag{3.1}$$

ここで，アライメント a は，日本語単語 j と英単語 e の対応関係を意味している．IBM モデルにおいて，各日本語単語に対応する英単語は 1 つであるのに対して，各英単語に対応する日本語単語は 0 から n 個あると仮定する．また，日本語単語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，日本語単語を対応させる．

3.1.2 GIZA++

GIZA++とは，日英方向と英日方向の対訳文において最尤な単語対応を得るツールである．対訳文を用いて IBM モデルを学習し，日英方向と英日方向の単語の翻訳確率を得る．

日英方向の単語対応の例を表 3.1 に示す．表 3.1 は左から順に日本語単語，英語単語，翻訳確率を示している．

表 3.1 GIZA++を用いた日英方向の単語対応の例

貿易	trade	0.5119
工場	factory	0.9057

3.1.3 単語対応の例

3.1.2 節で説明した GIZA++を用いて，対訳文から日英方向と英日方向の最尤な単語対応を得る．日英方向の単語対応の例を図 3.1 に，英日方向の単語対応の例を図 3.2 に示す．また，●は単語が対応した箇所を示す．

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

図 3.1 日英方向の単語対応の例

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

図 3.2 英日方向の単語対応の例

3.2 対訳句の抽出

ヒューリスティックを用いて単語対応から網羅的に対訳句を抽出する．具体的には日英方向と英日方向の単語対応を用いて，ヒューリスティックな方法により “対称な単語対応”を求める．そして，“対称な単語対応”のうち，矛盾しない全ての対訳句を抽出する．

3.2.1 ヒューリスティック

ヒューリスティックな方法は主に，“intersection”,“union”,“grow”がある．さらに，最終処理として，“final”と “final-and”がある．

3.2.1.1 intersection

intersection(積集合) は日英方向と英日方向の両方に単語対応が存在する場合，その単語対応を “対称な単語対応”とする方法である．図 3.3 に例を示す．

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

図 3.3 intersection の例

3.2.1.2 union

union(和集合) は日英方向と英日方向のどちらか一方に単語対応が存在する場合，その単語対応を “対称な単語対応” とする方法である．図 3.4 に例を示す．

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

図 3.4 union の例

3.2.1.3 grow

grow は intersection の “対称な単語対応” の縦横方向に union の “対称な単語対応” が存在する場合，その単語対応を intersection の “対称な単語対応” に追加する方法である．さらに，縦横方向に加え，対角方向に存在する union の “対称な単語対応” を intersection の “対称な単語対応” に追加する方法として “grow-diag” がある．図 3.5 に “grow-diag” の例を示す．

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

図 3.5 grow-diag の例

3.2.1.4 grow-diag-final-and

“grow-diag-final-and” は “grow-diag” において，日英方向と英日方向の単語対応がない場合，union に “対称な単語対応” があれば，union の “対称な単語対応” を追加する方法である．図 3.6 に “grow-diag-final-and” の例を示す．

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

図 3.6 grow-diag-final-and の例

3.2.2 対訳句の例

“対称な単語対応”のうち，矛盾しない全ての対訳句を抽出する．抽出した対訳句の例を表 3.2 に示す．

表 3.2 Och らの方法を用いて抽出した対訳句の例

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

第4章　提案手法

本研究では対訳文パターンを用いて対訳句を抽出する．対訳文パターンを用いることで文法構造を考慮した対訳句の抽出ができると考える．また，本研究の抽出方法では人間が見て不自然な対訳句を抽出する場合がある．そこで，対訳句の選別を行う．提案手法の手順を以下に示す．

4.1　対訳文パターンを用いた対訳句の抽出方法

4.1.1　対訳単語の作成

3.1.2 節で説明した GIZA++を用いて，対訳単語を作成する．手順を以下に示す．

手順 1 GIZA++を用いて対訳文から日英方向と英日方向の単語対応を得る．

手順 2 単語対応より対訳単語を得る．

手順 3 日英方向と英日方向の単語の翻訳確率を掛け合わせ，対訳単語の翻訳確率 (以下，対訳単語翻訳確率と表記) を得る．

手順 4 対訳単語翻訳確率が一定の閾値 (α) 以上である対訳単語を抽出する．

対訳単語の作成の例を図 4.1 に示す．

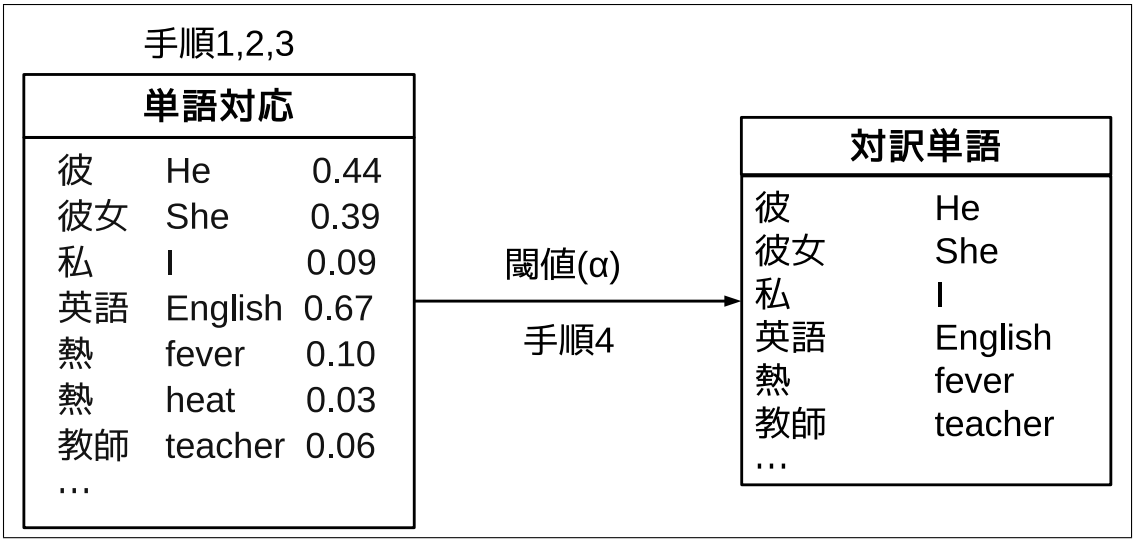

図 4.1 対訳単語作成の例

4.1.2 対訳文パターンの作成

対訳単語と対訳文を用いて対訳文パターンを作成する．手順を以下に示す．

手順 1 4.1.1 節で抽出した対訳単語が対訳文中で適合した場合，変数化を行い，対訳文パターンを得る．

手順 2 対訳文パターンの英文パターンにおいて，変数の直前に冠詞がある場合，冠詞を除去する．

なお，変数が連続しない対訳文パターンのみを本研究で用いる対訳文パターンとする．対訳文パターンの作成の例を図 4.2 に示す．

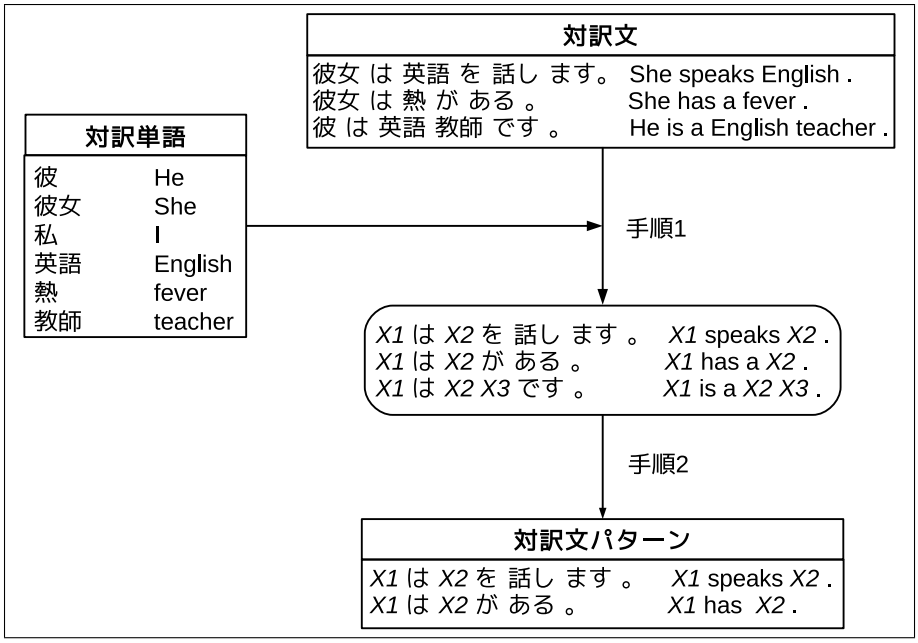

図 4.2 対訳文パターン作成の例

4.1.3 対訳句の抽出

対訳文パターンと対訳テスト文を用いて対訳句を抽出する．手順を以下に示す．

手順 1 対訳テスト文と対訳文パターンを照合する．

手順 2 対訳テスト文が対訳文パターンに適合した場合，対訳文パターンの変数部に対応する対を対訳句として抽出する．

なお，日本語句または英語句の少なくとも一方が複数単語で構成される対訳句のみを本研究で用いる対訳句とする．対訳文パターンの作成の例を図 4.3 に示す．

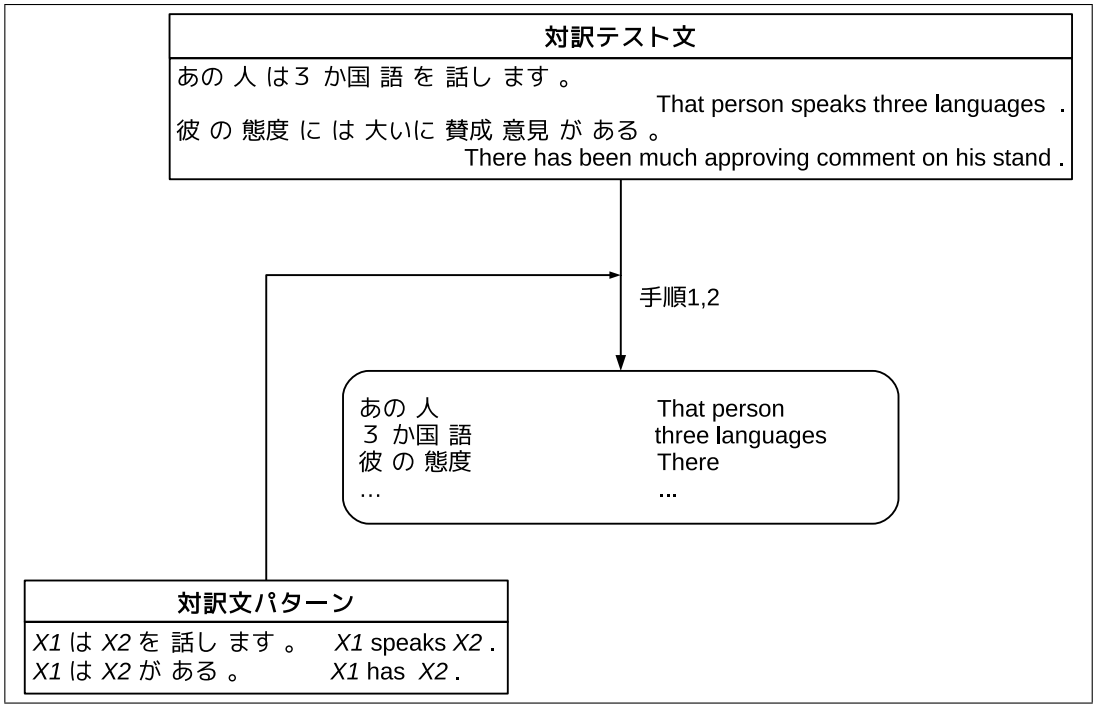

図 4.3 対訳句抽出の例

4.2 対訳句の選別方法

本研究では，人間が見て自然な対訳句を抽出することを目的とする．しかし，4.1 節の抽出方法では図 4.3 の例で示す “彼 の 態度” と “There” のような不自然な対訳句を抽出する．そこで，対訳句の選別を行う．

4.2.1 句の翻訳確率の付与

対訳句の選別には，対訳句の翻訳確率 (以下，句の翻訳確率と表記) を利用する．句の翻訳確率の付与の手順を以下に示す．

手順 1 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．

手順 2 GIZA++を用いて，各組み合わせの対訳単語翻訳確率を得る．

手順 3 各組み合わせの対訳単語翻訳確率の対数をとり，総和を求める．そして，総和の値を句の翻訳確率とする．なお，対訳単語翻訳確率が存在しない場合，ペナルティーとして-1000 を付与する．

対訳句の例を表 4.1 に，対訳単語翻訳確率の例を表 4.2 に示す．

表 4.1 対訳句の例

日本語句	彼 の 耳
英語句	his ear

表 4.2 対訳単語翻訳確率の例

日本語単語	英語単語	対訳単語翻訳確率
彼	his	0.018
彼	ear	-
の	his	0.003
の	ear	0.001
耳	his	0.001
耳	ear	0.073

表 4.1 の句の翻訳確率を以下の式で求める．

$$
\begin{aligned}
\text{句の翻訳確率} \quad = \quad & \log_2(\text{“彼”と“his”の対訳単語翻訳確率}) \\
+ \quad & \log_2(\text{“彼”と“ear”の対訳単語翻訳確率}) \\
+ \quad & \log_2(\text{“の”と“his”の対訳単語翻訳確率}) \\
+ \quad & \log_2(\text{“の”と“ear”の対訳単語翻訳確率}) \\
+ \quad & \log_2(\text{“耳”と“his”の対訳単語翻訳確率}) \\
+ \quad & \log_2(\text{“耳”と“ear”の対訳単語翻訳確率})
\end{aligned}
$$

上式と表 4.2 より，句の翻訳確率を求める．なお，表 4.2 より，“彼”と “ear”の対訳単語翻訳確率は存在しない．よって，ペナルティーとして-1000 を付与する．表 4.1 の対訳句において，句の翻訳確率は-1037.884 となる．

4.2.2 対訳句の選別

対訳句の選別は閾値 (β) を用いる．4.2.1 節で付与した句の翻訳確率が閾値 (β) 以上の対訳句を選別する．

第5章　実験環境

5.1　実験データ

実験データは，辞書の例文から抽出した日英対訳の単文データ [9] から，対訳文および対訳テスト文として，100,000 文を用いる．なお，対訳文と対訳テスト文は同一の単文データである．英語文には moses[10] に付属する tokenizer.perl を用いてわかち書きを行う．また，日本語文には MeCab[11] を用いて形態素解析を行う．なお，日英対訳の単文データは日本語文が単文であるため，英語文には重文・複文が含まれる場合がある．

5.2　閾値

4.1.1 節の対訳単語の作成に用いる閾値 (α) は，α=0.05 とする．また，4.2.2 節において，信頼度が高い対訳句を選別するために，閾値 (β) は，$\beta = -2000$ とする．

第6章　実験結果

6.1　対訳句の抽出結果

実験結果を以下に示す.

- GIZA++を用いて得た単語対応から，対訳単語を 17,182 語得た.
- 対訳文 100,000 文から，対訳単語を用いて，対訳文パターンを 54,417 文得た.
- 対訳テスト文 100,000 文から，対訳文パターンを用いて，対訳句を 19,504 句得た.
- 対訳句 19,504 句から，閾値 (β) を用いて，6,264 句選別した.
- 選別した対訳句 6,264 句中，4,312 句において，対訳テスト文と，適合した対訳文パターンの原文は同一であった.

抽出した対訳句の例を以下に示す．対訳句の例において，対訳句 1 および 2 は，対訳テスト文が対訳文パターンに適合して抽出した対訳句である．対訳テスト文は対訳句の抽出に用いた入力文である．対訳文パターンは対訳句の抽出に用いた文パターンである．対訳文パターンは対訳文パターンの原文から作成された.

表 6.1 対訳文パターンを用いた対訳句の例 1

対訳句 1	(日本語)	金 だ らい
	(英語)	a basin
対訳句 2	(日本語)	水 を つい だ
	(英語)	He poured water
対訳テスト文	(日本語)	金 だ らい に 水 を つい だ 。
	(英語)	He poured water into a basin .
対訳文パターン	(日本語)	$X1$ に $X2$ 。
	(英語)	$X2$ into $X1$.
対訳文パターンの原文	(日本語)	壁 に ぶつかる 。
	(英語)	Crash into a wall .

表 6.2 対訳文パターンを用いた対訳句の例 2

対訳句 1	日本語	風
	英語	The wind
対訳句 2	日本語	雨戸
	英語	the shutters
対訳テスト文	日本語	風 で 雨戸 が がたがた いう 。
	英語	The wind is shaking the shutters and rattling them .
対訳文パターン	日本語	$X1$ で $X2$ が がたがた いう 。
	英語	$X1$ is shaking $X2$ and rattling them .
対訳文パターンの原文	日本語	風 で 雨戸 が がたがた いう 。
	英語	The wind is shaking the shutters and rattling them .

表 6.3 対訳文パターンを用いた対訳句の例 3

対訳句 1	(日本語)	大きな 差
	(英語)	There
対訳句 2	(日本語)	ある
	(英語)	a big difference
対訳テスト文	(日本語)	大きな 差 が ある 。
	(英語)	There is a big difference .
対訳文パターン	(日本語)	X1 が X2 。
	(英語)	X1 is X2 .
対訳文パターンの原文	(日本語)	板 が 歪む 。
	(英語)	The board is distorted .

6.2 人手評価結果

選別した対訳句を用いて評価を行う．対訳句からランダムに 50 句抽出し，人間が見て対訳句が自然であるかを評価した．評価○は，対訳句が人間が見て自然であることを示す．評価×は，対訳句が人間が見て不自然であることを示す．評価結果を表 6.4 に示す．

表 6.4 対訳句の人手評価結果

評価○	評価×
42	8

表 6.4 より，人間が見て自然である対訳句は 50 句中 42 句 (84%) であった．

6.3 対訳句の例

6.3.1 評価○の例

人手評価における評価○の対訳句の例を以下に示す．

表 6.5 評価○の例 1

対訳句	(日本語)	友人
	(英語)	a friend of mine
対訳テスト文	(日本語)	通り で 友人 に 会っ た 。
	(英語)	I met a friend of mine on the street .
対訳文パターン	(日本語)	$X1$ で $X2$ に 会っ た 。
	(英語)	I met $X2$ on $X1$.
対訳文パターンの原文	(日本語)	通り で 彼 に 会っ た 。
	(英語)	I met him on the street .

表 6.6 評価○の例 2

対訳句	(日本語)	試験
	(英語)	The examination
対訳テスト文	(日本語)	間もなく 試験 が 始まり ます 。
	(英語)	The examination will begin shortly .
対訳文パターン	(日本語)	間もなく $X1$ が 始まり ます 。
	(英語)	$X1$ will begin shortly .
対訳文パターンの原文	(日本語)	間もなく 試験 が 始まり ます 。
	(英語)	The examination will begin shortly .

表 6.7 評価○の例 3

対訳句	(日本語)	虫 に 食わ れ た
	(英語)	moth-eaten
対訳テスト文	(日本語)	セーター が 虫 に 食わ れ た 。
	(英語)	My sweater is moth-eaten .
対訳文パターン	(日本語)	$X1$ が $X2$ 。
	(英語)	$X1$ is $X2$.
対訳文パターンの原文	(日本語)	板 が 歪む 。
	(英語)	The board is distorted .

表6.5，表6.6および表6.7より，人間が見て自然な対訳句を抽出していることがわかる．

6.3.2 評価× の例

人手評価における評価×の対訳句の例を以下に示す．

表 6.8 評価×の例 1

対訳句	(日本語)	縦横
	(英語)	The sewer
対訳テスト文	(日本語)	下水 が 市内 を 縦横 に 貫通 し て いる 。
	(英語)	The sewer system runs in all directions through the city .
対訳文パターン	(日本語)	下水 が 市内 を $X1$ に 貫通 し て いる 。
	(英語)	$X1$ system runs in all directions through the city .
対訳文パターンの原文	(日本語)	下水 が 市内 を 縦横 に 貫通 し て いる 。
	(英語)	The sewer system runs in all directions through the city .

表 6.9 評価×の例 2

対訳句	(日本語)	母
	(英語)	My mother's hair
対訳テスト文	(日本語)	母 は 髪 が 白く なっ て き た 。
	(英語)	My mother's hair is getting gray .
対訳文パターン	(日本語)	$X1$ は $X2$ て き た 。
	(英語)	$X1$ is getting $X2$.
対訳文パターンの原文	(日本語)	彼女 は 太っ て き た 。
	(英語)	She is getting fat .

表 6.10 評価×の例 3

対訳句	(日本語)	まだ 時間
	(英語)	There
対訳テスト文	(日本語)	まだ 時間 が ある 。
	(英語)	There is yet time .
対訳文パターン	(日本語)	$X1$ が $X2$ 。
	(英語)	$X1$ is $X2$.
対訳文パターンの原文	(日本語)	板 が 歪む 。
	(英語)	The board is distorted .

表 6.8，表 6.9 および表 6.10 より，不自然な対訳句を抽出していることがわかる．

6.4 句の翻訳確率を用いた対訳句の選別結果

句の翻訳確率の閾値 (β) を $\beta = -1000, -2000, -3000, -4000, -5000$ として，対訳句の数を調査した．対訳句の数を表 6.11 に示す．

表 6.11 閾値 (β) で選別した対訳句の数

閾値 (β)	対訳句の数
-1000	1,860
-2000	6,264
-3000	7,637
-4000	8,706
-5000	10,019

また，閾値 $\beta = -5000$ の対訳句において，ランダムに 50 句抽出し，人手評価を行った．結果を表 6.12 に示す．

表 6.12 閾値 $\beta = -5000$ を用いた対訳句の人手評価結果

評価○	評価×
36	14

表 6.4 と表 6.12 を比較すると，閾値 $\beta = -2000$ の対訳句の方が優れていることがわかる．

第7章　考察

7.1　対訳句の精度の考察

表 6.4 の評価×である 8 句において，誤り解析を行った．解析の結果，誤りの原因を表 7.1 に示す 5 種類に分類した．

表 7.1　誤り解析の結果

誤り原因	対訳句の数
不適切な対訳単語	4
対訳文パターンの不適切な適合	1
主語の省略	1
主語の違い	1
対訳文パターンの不足	1

評価×の対訳句を以下に示す．

7.1.1　不適切な対訳単語

表 7.2 において，対訳文パターンと，対訳文パターンの原文を比較すると，日本語単語 “オリエンテーション”は英単語 “freshmen”と誤って対応したことがわかる．よって，誤った対訳文パターンを作成した．その結果，人間が見て不自然な対訳句を抽出した．なお，表 7.3 から表 7.5 においても，同様の理由で人間が見て不自然な対訳句を抽出した．

不適切な対訳単語の問題は，対訳単語の作成に用いる閾値 (α) の調整により改善できると考えている．

表 7.2 不適切な対訳単語の例 1

対訳句	(日本語)	オリエンテーション
	(英語)	the freshmen
対訳テスト文	(日本語)	４月 １ ０ 日 に 新入生 の オリエンテーション が 行わ れ ます 。
	(英語)	They will give guidance to the freshmen on April 10 .
対訳文パターン	(日本語)	$X1$ １ ０ 日 に 新入生 の $X2$ が 行わ れ ます 。
	(英語)	They will give guidance to $X2$ on $X1$ 10 .
対訳文パターンの原文	(日本語)	４月 １ ０ 日 に 新入生 の オリエンテーション が 行わ れ ます 。
	(英語)	They will give guidance to the freshmen on April 10 .

表 7.3 不適切な対訳単語の例 2

対訳句	(日本語)	岩崎
	(英語)	the stock's
対訳テスト文	(日本語)	岩崎 前 社長 は 、 株価 急騰 へ の 関与 について は 否定 し た 。
	(英語)	Former Nikko chairman Iwasaki denied any involvement in the stock's share price rise .
対訳文パターン	(日本語)	$X1$ 前 社長 は 、 株価 急騰 へ の 関与 について は 否定 し た 。
	(英語)	Former Nikko chairman Iwasaki denied any involvement in $X1$ share price rise .
対訳文パターンの原文	(日本語)	岩崎 前 社長 は 、 株価 急騰 へ の 関与 について は 否定 し た 。
	(英語)	Former Nikko chairman Iwasaki denied any involvement in the stock's share price rise .

表 7.4 不適切な対訳単語の例 3

対訳句	(日本語)	跋扈
	(英語)	The gangster's
対訳テスト文	(日本語)	暴力団 の 跋扈 が 目だつ 。
	(英語)	The gangster's rampage is remarkable .
対訳文パターン	(日本語)	暴力団 の $X1$ が 目だつ 。
	(英語)	$X1$ rampage is remarkable .
対訳文パターンの原文	(日本語)	暴力団 の 跋扈 が 目だつ 。
	(英語)	The gangster's rampage is remarkable .

表 7.5 不適切な対訳単語の例 4

対訳句	(日本語)	縦横
	(英語)	The sewer
対訳テスト文	(日本語)	下水 が 市内 を 縦横 に 貫通 し て いる 。
	(英語)	The sewer system runs in all directions through the city .
対訳文パターン	(日本語)	下水 が 市内 を $X1$ に 貫通 し て いる 。
	(英語)	$X1$ system runs in all directions through the city .
対訳文パターンの原文	(日本語)	下水 が 市内 を 縦横 に 貫通 し て いる 。
	(英語)	The sewer system runs in all directions through the city .

7.1.2 対訳文パターンの不適切な適合

表 7.6 不適切な対訳文パターンの例

対訳句	(日本語)	立つ
	(英語)	skilled at Kendo
対訳テスト文	(日本語)	彼 は 剣道 の 腕 が 立つ 。
	(英語)	He is skilled at Kendo .
対訳文パターン	(日本語)	$X1$ が $X2$ 。
	(英語)	$X1$ is $X2$.
対訳文パターンの原文	(日本語)	板 が 歪む 。
	(英語)	The board is distorted .

表 7.6 において，対訳テスト文の日本語単語 “は” は，英単語 “is” と対応する．しかし，適合した対訳文パターンは，対訳テスト文と不適切に適合し，対訳句を抽出した．

7.1.3 主語の省略

表 7.7 主語の省略の例

対訳句	(日本語)	彼 の アリバイ
	(英語)	I
対訳テスト文	(日本語)	彼 の アリバイ に は 不審 な 点 が ある 。
	(英語)	I have some doubts about his alibi .
対訳文パターン	(日本語)	$X1$ に は $X2$ が ある 。
	(英語)	$X1$ have $X2$.
対訳文パターンの原文	(日本語)	私 に は 夢 が ある 。
	(英語)	I have a dream .

表 7.7 において，対訳テスト文の日本語文は主語を省略した文である．日本語文の主語を省略すると，英語文の主語と対応する日本語がなくなる．よって，対訳テスト文は，対訳文パターンと適切に適合することが困難である．

7.1.4 主語の違い

表 7.8 主語が異なる例

対訳句	(日本語)	母
	(英語)	My mother's hair
対訳テスト文	(日本語)	母 は 髪 が 白く なっ て き た 。
	(英語)	My mother's hair is getting gray .
対訳文パターン	(日本語)	$X1$ は $X2$ て き た 。
	(英語)	$X1$ is getting $X2$.
対訳文パターンの原文	(日本語)	彼女 は 太っ て き た 。
	(英語)	She is getting fat .

表 7.8 において，対訳テスト文の日本語文の主語が “母”であるのに対し，英語文の主語は “My mother's hair”である．そのため，対訳テスト文は，対訳文パターンに適合しても，人間が見ると不自然な対訳句を抽出する．

7.1.5 対訳文パターンの不足

表 7.9 対訳文パターンの不足の例

対訳句	(日本語)	まだ 時間
	(英語)	There
対訳テスト文	(日本語)	まだ 時間 が ある 。
	(英語)	There is yet time .
対訳文パターン	(日本語)	$X1$ が $X2$ 。
	(英語)	$X1$ is $X2$.
対訳文パターンの原文	(日本語)	板 が 歪む 。
	(英語)	The board is distorted .

表 7.9 において，対訳テスト文は不適切な対訳文パターンと適合している．また，本研究で用いた対訳文パターンには対訳テスト文と適切に適合する対訳文パターンは存在しなかった．よって，対訳文パターンの不足の問題がある．

7.2 人手評価の考察

7.2.1 Och らの方法で抽出した対訳句との比較

Och らの方法で抽出した対訳句において，ランダムに 50 句抽出し，人手評価を行った．結果を表 7.10 に示す．

表 7.10 Och らの方法による対訳句の人手評価結果

評価○	評価×
21	29

表 7.10 より，人間が見て自然であると評価した対訳句は，50 句中 21 句 (42%) であった．表 6.4 と表 7.10 を比較すると，提案手法は，Och らの対訳句の抽出方法よりも優れていることがわかる．なお，Och らの方法で抽出した対訳句においても，句の翻訳確率が付与されている．しかし，本節において，Och らの方法で抽出した対訳句は句の翻訳確率による選別を行っていない．今後，Och らの方法で抽出した対訳句においても，句の翻訳確率による選別を行い，提案手法と比較を行う必要があると考える．

7.2.2 Och らの方法で抽出した対訳句の例

Och らの方法で抽出した対訳句の評価○の例を表 7.11 に，評価×の例を表 7.12 に示す．

表 7.11 Och らの方法による対訳句の人手評価○の例

日本語句	英語句
危険 性 が 増す	The risk is on the increase
彼 は 英気 を 養っ た	He stored up his energy
帽子 を かぶっ て い た	were wearing a hat

表 7.12 Och らの方法による対訳句の人手評価×の例

日本語句	英語句
婦人 服 が	are popular
て ほしい 。	other .
に 立っ て いる 。	in the league tables .

7.3 抽出した対訳句の考察

実験の結果より，選別した対訳句 6,264 句中，4,312 句において，対訳テスト文と，適合した対訳文パターンの原文は同一であった．対訳テスト文と，適合した対訳文パターンの原文が同一である例を表 7.13 に示す．

表 7.13 対訳テスト文と，適合した対訳文パターンの原文が同一である例

対訳句 1	(日本語)	石
	(英語)	The stone
対訳句 2	(日本語)	水
	(英語)	the water
対訳テスト文	(日本語)	石 が 水 の 中 に ザブン と 落ち た 。
	(英語)	The stone splashed into the water .
対訳文パターン	(日本語)	X1 が X2 の 中 に ザブン と 落ち た 。
	(英語)	X1 splashed into X2 .
対訳文パターンの原文	(日本語)	石 が 水 の 中 に ザブン と 落ち た 。
	(英語)	The stone splashed into the water .

この場合英語句は，表 7.13 の英語句 “The stone”や “the water”のような，単語に冠詞がついた対訳句である．よって，日本語句および英語句において，冠詞以外の複数単語で構成される対訳句は，少ない．今後，冠詞以外の複数単語で構成される大量の対訳句の抽出を試みる．

7.4 今後の課題

今後の課題を以下に示す．

- 7.1.5 節において，対訳文パターンの不足の問題を示した．対訳文パターンの不足の問題を解決するために，新たな対訳文パターンを作成する必要がある．そこで，抽出した対訳句を用いて，対訳文から新たな対訳文パターンを作成することを検討する．そして，新たな対訳文パターンを用いて，対訳テスト文からさらに対訳句を抽出することを試みる．
- 対訳句の抽出方法は Och らの方法の他に，BerkeleyAligner など，種々の方法が存在する．今後，これらの方法とも比較を行う予定である．また，Och らの方法で抽

出した対訳句においても，句の翻訳確率を用いた選別を行い，比較を行う必要があると考える．

- 提案手法で抽出した対訳句のパターン翻訳への応用を検討する．

第8章　おわりに

パターン翻訳は対訳文パターンと対訳句を人手で大量に作成するためコストがかかる問題がある．また，統計的機械翻訳における対訳句の抽出方法として，Och らの方法がある．Och らの方法は自動で対訳句を抽出するためコストが低くなるが，人間がみると不自然な対訳句を抽出してしまう問題がある．

本研究では，対訳文パターンを用いて対訳句の抽出を行った．対訳文パターンを用いることで人間が見て自然な対訳句の抽出ができると考える．しかし，対訳文パターンを人手で大量に作成するにはコストがかかる．そこで本研究では西村らの方法を用いて対訳文パターンの自動作成を行った．実験の結果，対訳テスト文 100,000 文から閾値 $\beta = -2000$ を用いて，6,264 句の対訳句を抽出した．また，人手評価において，人間が見て自然であると評価した対訳句は 50 句中 42 句 (84%) であり，Och らの方法よりも優れていることを示した．今後，対訳句をパターン翻訳で用いることを目指す．

謝辞

最後に，一年間に渡り，本研究の御指導をいただきました鳥取大学工学部知能情報工学科計算機講座 C 研究室の村田真樹教授，村上仁一准教授，徳久雅人講師に深く感謝するとともに厚くお礼を申し上げます．また，計算機工学講座 C 研究室の皆様に厚くお礼を申し上げます．また，参考にさせて頂いた論文の著者の方々に対して，深く感謝します．

参考文献

[1] 道祖尾太祐, 村上仁一, 徳久雅人, 池原悟. n-gram を利用した日英対訳パターンの自動抽出. 言語処理学会第 10 回年次大会発表論文集, pp. 241–244, 2004.

[2] 北村美穂子, 松本祐治. 対訳コーパスを利用した対訳表現の自動抽出. 情報処理学会論文誌, Vol. 38, No. 4, pp. 727–736, 1997.

[3] Franz Josef Och and Germann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, 2003.

[4] Phillip Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *International Workshop on Spoken Language Translation(IWSLT)*, 2005.

[5] BerkeleyAligner: A word alignment software package. http://code.google.com/p/berkeleyaligner/.

[6] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[7] 西村拓哉, 村上仁一, 徳久雅人, 池原悟. 文単位のパターンを用いた統計翻訳. 言語処理学会第 16 回年次大会発表論文集, pp. 676–679, 2010.

[8] GIZA++: Training of statistical translation models. http://www.fjoch.com/GIZA++.

[9] 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.

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

[11] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.