パターンを用いた対訳句の自動作成と翻訳精度の調査

栗下 尚樹 [*1]　村上仁一 [*2]
[*1] 鳥取大学 工学部 知能情報工学科
{s132024,murakami}@ike.tottori-u.ac.jp

1　はじめに

対訳文から自動的に対訳句を作成する方法として，Barkeley などが挙げられる．これは基本的に Och のヒューリスティック [1] を利用している．しかし従来の方法で作成された対訳句 (特に日英の対訳句) の精度は低い．

そこで本研究では，新たな手法を提案する．具体的には対訳文と文パターンを利用して対訳句を作成する．また翻訳実験を行い翻訳の精度向上を試みる．

2　関連研究

対訳句の作成や精度向上のためにいくつかの先行手法が挙げられる．Och ら [1] は，アライメントテンプレートを用いる方法の中で，２つの方向の対応関係を用いてそれぞれの方向の対応関係をヒューリスティックルールを適用して拡張している．また山田ら [2] は，同じ対訳コーパスから非対称な２つの学習によって構築された２つの翻訳モデルを利用することで，それぞれの翻訳モデルの精度を向上させた．また Chris ら [3] らは，文アライメントだけでなく単語アライメントを学習時に用いることで AER を下げる研究がなされている．

3　提案手法

提案手法としてパターンから対訳句を作成する方法を提案する．

3.1　基本手法

基本的な手法は対訳文とパターンから対訳句を作成する．例として表 1 に具体例を示す．

表 1　テストデータ

対訳文 (日)	ナイロンは原油から作られる
対訳文 (英)	Nylon is made from crude oil
パターン (日)	*X*1 は *X*2 *X*3
パターン (英)	*X*1 *X*3 *X*2
パターン原文 (日)	ワインはブドウから作られる
パターン原文 (英)	Wine is made from grape

表 1 の対訳文は対訳句作成に用いる文であり，パターンは対訳句作成に用いるパターンであり，パターン原文はパターンを作る際に用いた対訳文である．(パターン原文からパターンを作成する手法は文献 [4] を参照) 対訳句作成の手順を以下に示す．

手順１　対訳文とパターンを照合
日本語文では「ナイロンは原油から作られる」と「*X*1 は *X*2 *X*3」である．

手順２　単語レベル文パターンの変数部に対応する組み合わせの対訳句をすべて抽出
表 2 に具体例を示す．

表 2　手順２と手順３の具体例

変数	対訳句		対数フレーズ確率
*X*1	ナイロン	Nylon	-0.5
*X*2	原油より	from crude oil	-0.1
*X*3	作られる	is made	-0.05
⋮			
*X*1	ナイロン	Nylon is made from	-0.8
*X*2	原油より	oil	-0.2
*X*3	作られる	crude	-0.3

手順３　GIZA++ を用いて対数フレーズ確率を付与
対数フレーズ確率は以下の式を用いて計算する．

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

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)

手順４　全ての対訳句の対を抽出

この手法では，パターンに基づいて対訳文から対訳句を可能な限り抽出していくため，大量の対訳句を出力する. 表 2 のパターンでは 40 組 (変数の組み合わせ) × 3(変数) で 120 句が抽出される．

3.2　パターン制約

基本手法から対訳句の数を減らすために，１つの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の組み合わせを出力する．その方法をパターン制約 [5] と呼ぶ．以下に具体例を示す．

手順５　各パターンごとに，手順３で付与した各組みの対数フレーズ確率の総和が最大値をとる対訳句を１つずつ選出
表 3 に具体例を示す．

表 3　パターン制約の具体例

変数	対訳句		対数フレーズ確率
*X*1	ナイロン	Nylon	-0.5
*X*2	原油より	from crude oil	-0.1
*X*3	作られる	is made	-0.05
総和			-0.65

これにより基本手法の例では 120 句が 3 句に削減される．

3.3 類似制約

パターン制約では1つのパターンに対して対訳句が1組ずつ選ばれるが，実際には1つの対訳句 (日) に対し複数のパターンが照合するため複数の対訳句 (英) が作成される．またパターン制約を用いても誤った対訳句が存在する．そこで対訳句を作る際，対数フレーズ確率だけでなく，対訳文とパターン原文との類似度を利用する．これを類似制約と呼ぶ．この類似制約はパターン制約を適用しさらに類似度を利用して対訳句を選択する．以下に具合的な手順を示す．手順はパターン制約以降の続きである．

手順６ 対訳句作成時の対訳文 (対訳文) と適合したパターン作成時の対訳文 (パターン原文) を抽出
表 1 参照

手順７ 対訳文から見たパターン原文との類似度 A を作成
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．

$$P(r) = (\frac{N}{M})$$

M; 対訳文中の単語数
N; 対訳文中の単語がパターン原文中の単語と一致している単語数
表 4 から類似度 A の値は 3/5.

表 4　類似度 A の具体例

比較する単語	評価
ナイロン	不一致
は	一致
原油	不一致
から	一致
作られる	一致
類似度 A=一致/総語数=3/5	

手順８ パターン原文から見た対訳文との類似度 B を作成
手順 7 と同様に計算し，類似度 B の値は 3/5

手順９ 両者を掛けあわせ，類似度とする
類似度 = 類似度 A * 類似度 B = 3/5 * 3/5 = 9/25

手順１０ 類似度が最大の対訳句を選出
表 5 は「原油より」に対応する対訳句を表す．

表 5　類似制約の例

対訳句		類似度
原油より	**from crude oil**	9/25
	from oil	2/21
	than crude oil	3/14

表 5 の例に対して類似制約を適用すると表 6 が出力される．

表 6　類似制約の出力

対訳句		類似度
原油より	crude oil	9/25

なお実験においては日英両方向での類似度を計算する．

4 実験環境

4.1 実験データ

本実験では表 7 のデータを用いて行う．

表 7　実験データ

日本語学習文	100,000 文
英語学習文	100,000 文
テスト文	100 文

4.2 評価方法

対訳句の評価において，提案手法の対訳句と先行手法の対訳句から，それぞれランダムに 100 句抽出し，人手評価を行う．また表 8 に評価基準を示す．

表 8　対訳句の評価方法

評価	評価内容
○	適切な対応をとる
△	意味が欠落している，もしくは 不必要な意味が付与されている
×	不適切な対応をとる

5 実験結果

5.1 対訳句の精度結果

提案手法と先行手法において，対訳句の精度の評価を行う．ただし，先行手法は Moses の phrase-table とし，アライメントは intersection とする．また，人手対訳句 [6] も評価する．表 9 に評価結果を示す．

表 9　対訳句の評価結果 (100 句中)

手法	総数	○	△	×
先行手法	30,887,824	5	53	42
パターン制約	94,028	52	14	34
類似制約	77,880	61	27	12
人手対訳句	335,726	77	17	6

表 9 より，先行手法より 2 つの提案手法の対訳句の精度の方が高く，提案手法が有効であるといえる．また類似制約はパターン制約に比べて”不適切な対応をとる対訳句”が減少し，”意味が欠落している，もしくは不必要な意味が付与されている対訳句”が増加した．また類似制約の精度は人手対訳句に接近しているがまだ及ばない．

以下に提案手法で作成された対訳句の具体例を示す．表 10 に先行手法の例を，表 11 にパターン制約の例を，表 12 に類似制約の例を示す．

表 10　phrase-table の具体例

対訳句 (日)	対訳句 (日)	評価
途方もなく大きな恩恵	benefited tremendously	○
世界を独占している	have a world monopoly	○
煙突から煙が上って	from the chimney	△
上衣	off my coat	△
もやっと山	all downhill	×
警察署には一種	station has an atmosphere	×

先行手法は長い対訳句が多く，対訳句 (日) に完全に対応している対訳句 (英) は少ない．

表 11　パターン制約の具体例

対訳句 (日)	対訳句 (英)	評価
母からの 向けられ	from my motehr turned on	○ ○
捨てた スペースシャトル	He threw away The space	△ △
は この 円	He has one million	× ×

パターン制約は先行手法に比べ○の数が多くなっているがまだも多い．対訳句 (日) に全く対応していない対訳句 (英) が多い．

表 12　類似制約の具体例

対訳句 (日)	対訳句 (英)	評価
全速力 よく聞く	full speed listens well	○ ○
かばんのチャック 手に包帯	bag open with a bandage	△ △
は勇敢にもその 32 段	expose the without stoping once	× ×

類似制約は×がパターン制約と比べさらに減っている．また”意味が欠落している，もしくは不必要な意味が付与されている対訳句”が増加し，”不適切な対応をとる対訳句”がかなり減少した．

6　翻訳精度の調査

対訳句の精度向上による翻訳精度の影響を調べるために翻訳実験を行う．翻訳実験は提案手法を適用した対訳句を用いる．翻訳方法はパターンに基づく統計機械翻訳 (PBSMT) を用いて行う．PBSMT の概要を以下に示す [4].

手順 A　対訳文と GIZA++ から対訳単語を作成.
手順 B　対訳文と対訳単語から単語レベル文パターンを作成.
手順 C　対訳文と単語レベル文パターンから対訳句を作成.
手順 D　対訳文と対訳句から句レベル文パターンを作成.
手順 E　対訳句と句レベル文パターンを用いて翻訳を行う.

本研究で提案した 2 つの手法は手順 C で用いる.

6.1　翻訳実験データ

翻訳実験は表 13 のデータを用いる.

表 13　翻訳実験データ

日本語学習文	100,000 文
英語学習文	100,000 文
テスト文	200 文

ただし類似制約とパターン制約の実験において，枝刈りのパラメータが少し異なる.

6.2　翻訳の評価方法

翻訳文の評価方法において，パターン制約の翻訳文と類似制約の翻訳文から，ランダムに 200 文抽出し，人手による対比較評価を行う．また表 14 に評価基準を示す.

表 14　翻訳の評価方法

評価	評価内容
類似制約○	類似制約の方がパターン制約よりも良い
パターン制約○	パターン制約の方が類似制約よりも良い
差なし	どちらもあまり差がない
同一文	類似制約とパターン制約が同一

6.3　翻訳結果

翻訳を行った結果，200 文中 76 文が未知語を含み，10 文が翻訳できなかった．よって 114 文で評価を行った．結果を表 15 に示す.

表 15　翻訳の人手評価

類似制約○	パターン制約○	差なし	同一
36	31	35	12

表 15 より，どちらの提案手法を用いても翻訳精度にあまり差がないことがわかった．また表 16 に類似制約○を，表 17 にパターン制約○の出力例を示す．太字は正しいと判断した部分，下線部は間違っていると判断した部分である.

表 16　類似制約○の具体例 1

入力文 1	その 店 は 、新しい 経営 陣 の 下 で 再開 し た
参照文 1	The store has reopened under new management.
パターン制約	The store resumed the under championship in the, new its management
パターン (日)	*X*03 は *X*02 *X*00 *X*01 の *X*04 で *X*05 し た 。
パターン (英)	*X*03 *X*05 the *X*04 championship in the *X*02 *X*00 *X*01 .
類似制約○	The store has reopen under a new management.
パターン (日)	*X*03 *X*01 新しい *X*04 *X*02 *X*00 し た 。
パターン (英)	*X*01 *X*03 has *X*00 *X*02 new *X*04 .
入力文 2	この 方法 の 是非 は 簡単 に は 決め られ ない.
参照文	We cannot tell at once whether this method is right or wrong .
パターン制約	This method whether not arranged that easily.
パターン (日)	*X*04 の *X*02 は *X*03 *X*01 決め られ *X*00 。
パターン (英)	*X*04 *X*02 *X*01 arranged *X*00 *X*03 .
類似制約	This method can’t decided on the easily of **right or wrong**.
パターン (日)	*X*03 *X*02 *X*00 *X*01 *X*05 *X*04 られ ない 。
パターン (英)	*X*03 can’t *X*04 *X*01 *X*05 *X*02 *X*00 .

表 16 において，入力文 1 の結果は類似制約の方が参照文 1 に似ているため，入力文 2 の結果は「是非」が読み取れるため，類似制約○とした.

表 17 において，入力文 3 の結果は文の形として成り立っているため，入力文 4 の結果は文と主語が合っているためパターン制約○とした.

表 17　パターン制約○の具体例

入力文３	私 の 疑惑 は 大きく なっ た
参照文	My suspicion grew .
パターン制約○	My Suspicion became louder .
パターン (日)	私 の *X*00 は *X*01 *X*02 。
パターン (英)	My *X*00 *X*02 *X*01 .
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 .
類似制約	The grown my suspicion .
パターン (日)	私 の *X*02 *X*00 *X*01 た 。
パターン (英)	*X*00 *X*01 my *X*02 .
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 .
入力文４	冬 は 太陽 が 早く 沈む
参照文	The sun sets early in winter .
パターン制約○	**The sun** sink early winter .
パターン (日)	*X*03 は *X*01 が *X*02 *X*00 。
パターン (英)	The *X*01 *X*00 *X*02 *X*03 .
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 .
類似制約	Winter sink early in the sun.
パターン (日)	*X*04 は *X*02 *X*00 *X*03 *X*01 。
パターン (英)	*X*04 *X*01 *X*03 in *X*00 *X*02 .
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 .

7　考察

7.1　先行手法と類似制約の比較

提案手法の類似制約と先行手法として Moses の翻訳の評価を行う．それぞれの翻訳文からランダムに 200 文抽出する．人手による対比較評価を行う際，200 文中 63 文が未知語を含み，9 文が翻訳できなかったので 128 文で評価を行った．結果を表 18 に示す．

表 18 類似制約○は類似制約の方が Moses よりも良いことを表す．

表 18　Moses の人手評価

類似制約○	Moses ○	差なし	同一
49	14	41	4

表 18 より類似制約は Moses と比較して非常に精度が良いことがわかった．

7.2　対訳句作成の考察

表 9 より，類似制約において対訳句の精度は向上したが，不適切な対応をとる対訳句が 12 文ある．表 12 にある評価×の「は勇敢にもその」-「expose the」の対応する際に用いられたデータを調べた．表 19 にデータを示す．

類似度の値は表 12 にある評価○の「全速力」-「full speed」よりも低い．「は勇敢にもその」-「expose the」の対応はパターン制約を適用しているのでパターンにおいて最大の対数フレーズ確率値であるといえる．しかし，「スキャンダル」-「courage」や「彼」-「scandal」という誤った対応の対訳句を含んでいる．これは対数フレーズ確率を付与する際に利用する GIZA++ の推定値に問題がある．解決方法としては対数フレーズ確率の付与の方法を見直す，もしくは学習文のデータ量を増やすことであると考える．

表 19　類似制約における対訳句の考察

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 .
パターン (日)	*X*02 *X*00 *X*03 *X*01 *X*04 た 。
パターン (英)	He *X*04 *X*01 *X*03 to *X*00 *X*02 .
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 .
*X*00	「は勇敢にもその」-「expose the」
*X*01	「を」-「the」
*X*02	「彼」-「scandal」
*X*03	「スキャンダル」-「courage」
*X*04	「暴露し」-「had」

7.3　0 型代名詞を含む文からのパターン作成の問題

入力文に対して形の似ていないパターン原文から作られたパターンを用いて翻訳を行った結果，不適切な文となっている例がある．表 17 の入力文３の主語は「私の疑惑」に対してパターン原文は主語がない文である．そのため入力文に対して不適切なパターンを照合したことで不適切な文が出力されている．

8　おわりに

本研究では，対訳文と文パターンを利用して対訳句を作成する手法として，１つのパターンに対し１組ずつの対訳句を出力するパターン制約と，対訳文とパターン原文の類似度を用いて対訳句を選択する類似制約を用いた手法を提案した．その結果，対訳句の精度は大幅に向上した．しかし，これらを適用した対訳句を翻訳に用いても精度はあまり向上しなかった．

参考文献

[1] Och,F.J., Tillmann, C., and Ney, H. ”Improved Alignment Models for Statistical Machine Translation.” In Proc. of EMNLP/WVLC-99

[2] 山田節夫　永田晶明　山田賢治 ”再学習により翻訳モデルを用いた単語アライメントの向上”

[3] Chris Callison-Burch, David Talbot, and Miles Osborne. ”Statistical Machine Translation with Word- and Sentence-Aligned Parallel Corpora”

[4] 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

[5] 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

[6] 鳥バンク http://unicorn.ike.tottori-u.ac.jp/toribank/

[7] Moses：http://www.statmt.org/moses/.