1変数対訳文パターンを用いた対訳句の抽出調査

森本智喜 *1　村上仁一 *2
*1 鳥取大学 工学部 電気情報系学科
*2 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
*1s152112@eecs.tottori-u.ac.jp *2murakami@eecs.tottori-u.ac.jp

1　はじめに

通常, 対訳句の抽出には人手による作業が伴う．そのためコストがかかり, 抽出される対訳句の数に制限がある．そこで江木ら [1] は, 従来のパターン翻訳 [2] から対訳文パターンの作成を自動化することでコストを削減し, 対訳句を大量に自動作成した．しかし同時に誤った対訳句も抽出された．そこで本研究では誤った対訳句の抽出を抑制するため,1 変数対訳文パターンを用いて対訳句の抽出を試みる．

2　従来手法

江木らの従来手法では,3 つの手順で対訳句の抽出を行う．以下にその手順を示す．

手順 2-1　対訳単語の作成

GIZA++[3] を用いて対訳学習文から翻訳確率を得る．そして任意の対訳学習文 A と翻訳確率から，対訳単語を作成する．図 1 にその手順と例を示す．

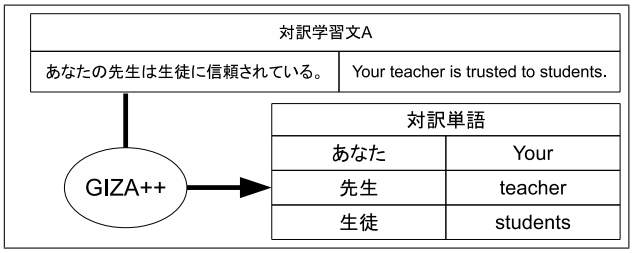

図 1　対訳単語の作成手順

手順 2-2　対訳文パターンの作成

任意の対訳学習文 A から対訳単語を全て変数化し, 対訳文パターンを作成する．図 2 にその手順と例を示す．

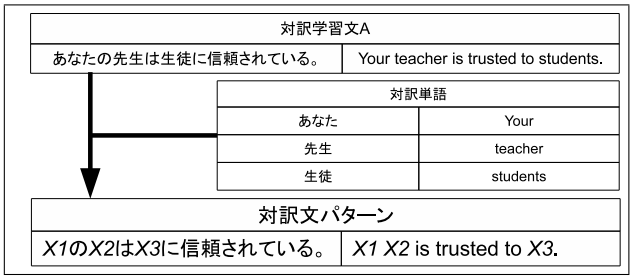

図 2　対訳文パターンの作成手順

手順 2-3　対訳句の抽出

任意の対訳学習文 B から対訳文パターンの変数部を抽出し, 対訳句とする．図 3 にその手順と例を示す．
以上を全ての対訳学習文と対訳文パターンに対して行う．

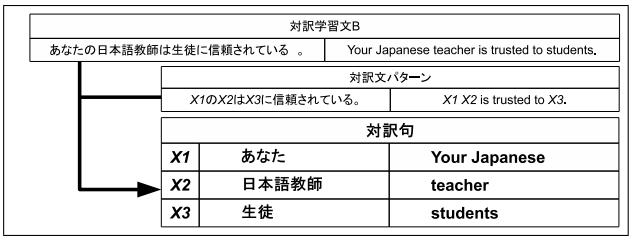

図 3　対訳句の抽出手順

3　問題点

3.1　変数の連続

従来手法の問題点の 1 つが, 対訳文パターンにおける連続した変数間の境界の曖昧性である．図 3 において, 変数 $X1$ と $X2$ が対訳文パターン中で連続している．この対訳文パターンが, 不適切な位置で区切られた対訳句が抽出される原因となる．表 1 にこのような誤った対訳句の例を示す．

表 1　誤った対訳句

$X1$: あなた	Your Japanese
$X2$: 日本語教師	teacher

3.2　語順

従来手法の別の問題点は, 対訳文パターンの語順である．対訳学習文 B と対訳文パターンの語順の違いは、誤った対訳句が抽出される原因となる．図 4 にこのような誤った対訳句の抽出例を示す．

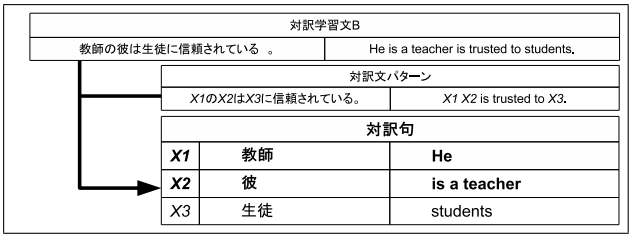

図 4　語順の異なる対訳文パターンによる対訳句の抽出手順

4　提案手法

基本的に対訳文パターン中に変数が多くなるほど, 誤った対訳句が抽出される原因となる．そこで本研究では,1 変数のみの対訳文パターンを使用し, 誤った対訳句の抽出の抑制を試みる．

提案手法は 3 つの手順で対訳句の抽出を行う．以下にその手順を示す．

手順 4-1　対訳単語の作成

手順 2-1 と同様にして，対訳単語を作成する．

手順 4-2　1 変数のみの対訳文パターンの作成

任意の対訳学習文 A から対訳単語を 1 つだけ変数化し,1 変数のみの対訳文パターンを作成する．図 5 にその作成手順と例を示す.

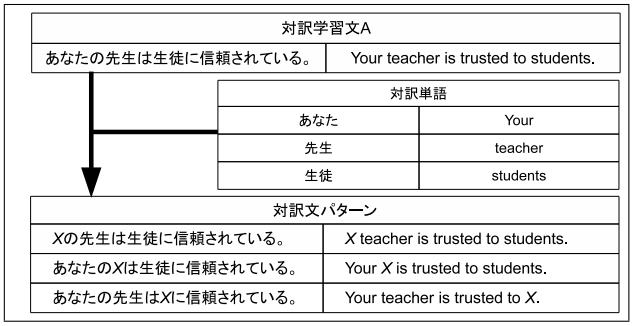

図 5　1 変数のみの対訳文パターンの作成手順

手順 4-3　対訳句の抽出

任意の対訳学習文 B から対訳文パターンの変数部を抽出し, 対訳句とする．図 6 に対訳句の抽出手順と例を示す.

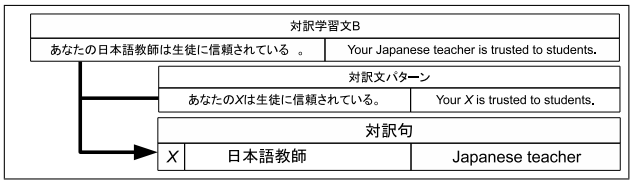

図 6　1 変数のみの対訳文パターンによる対訳句の抽出手順

以上を全ての対訳学習文と対訳文パターンに対して行う.

5　実験

5.1　実験データ

対訳学習文 159,998 文 [4] に対し, 従来手法と提案手法で対訳句の抽出を行う.

5.2　実験結果

従来手法の手順 2-3 と提案手法の手順 4-3 より, 抽出された対訳句の異なり数を表 2 に示す.

表 2　従来手法と提案手法による対訳句の抽出数

	抽出数
従来手法	5,058,468 対
提案手法	17,540 対

5.3　対訳句の抽出精度評価

5.3.1　評価基準

以下の基準に従い, 人手による評価を行う．評価基準は対訳句の日英間の対応を焦点としている.

- ○…日英の対応が適切である.
- △…日英の対応が部分的に適切である.
- ×…日英の対応が不適切である.

5.3.2　評価結果

手法ごとに, 抽出された対訳句を出現回数が 1 回か 2 回以上かで大別し, それぞれ 100 対の対訳句を評価した．表 3 に評価結果を示す．また表 4 に提案手法より得られた出現回数 1 回の対訳句の評価例を示す.

表 3　対訳句 100 対の人手評価結果

従来手法

出現回数	○	△	×
1 回	22 対	31 対	47 対
2 回以上	20 対	31 対	49 対

提案手法

出現回数	○	△	×
1 回	93 対	5 対	2 対
2 回以上	99 対	1 対	0 対

表 4　提案手法より抽出された出現回数 1 回の対訳句評価例

評価 ○

対訳句 (X)	先生のよう like a teacher
対訳学習文 B	彼は先生のように見える。 He looks like a teacher.
対訳文パターン	彼は X に見える。 He looks X.
対訳単語	神経質 nervous
対訳学習文 A	彼は神経質に見える。 He looks nervous.

評価 △

対訳句 (X)	私は辺り around
対訳学習文 B	私は辺りを見回した。 I looked around.
対訳文パターン	X を見回した。 I looked X.
対訳単語	辺り around
対訳学習文 A	辺りを見回した。 I looked around.

評価 ×

対訳句 (X)	香り good
対訳学習文 B	バラはよい香りがする。 Roses smell good.
対訳文パターン	バラはよい X がする。 Roses smell X.
対訳単語	香り sweet
対訳学習文 A	バラはよい香りがする。 Roses smell sweet.

5.4　評価 △ および評価 × となる原因の調査

表 3 において提案手法より得られた対訳句の内, 評価 △ または評価 × となった原因の調査を行う.

5.4.1　対訳学習文 A の語の省略

表 5 に評価 △ である対訳句の例を示す．この例では対訳学習文 A の日本語側において, 英語側”He”に相当する語が省略されている．一方で, 対訳学習文 B の日本語側で語の省略はない．この対訳学習文の違いが原因で, 対訳句の日本語側”彼は”に対応する英語側の語が抽出されていない．従って, 日英の対応が部分的に適切となり評価 △ となった.

表 5　語の省略により評価 △ となった対訳句の例

評価 △

対訳句 (X)	a strong voice 彼は声に張り
対訳学習文 B	彼は声に張りがある。 He has a strong voice.
対訳文パターン	X がある。 He has X.
対訳単語	そばかす fleckles
対訳学習文 A	そばかすがある。 **He** has freckles.

評価 △ である対訳句 6 対すべてで, 対訳学習文 A の日本語側に語の省略が確認された. ただし, このような語の省略がある対訳学習文 A は, 同じように語の省略がある対訳学習文 B に対して, 正しい対訳句の抽出を行うことができる.

5.4.2　誤った対訳単語

表 6 に評価 × である対訳句の例を示す. この例では, 対訳単語の日英の対応が不適切であり, 誤った対訳句が抽出される原因となっている. 評価 × である対訳句 2 対において, どちらの対訳単語にも不適切さが確認された.

表 6　不適切な対訳単語により評価 × となった対訳句の例

評価 × である対訳句

対訳句 (X)	つたって流れ Tears
対訳学習文 B	涙が彼女のほおをつたって流れた。 Tears trickled down her cheeks.
対訳文パターン	涙が彼女のほおを X た。 X trickled down her cheeks.
対訳単語	**流れ落ち** **Tears**
対訳学習文 A	涙が彼女のほおを流れ落ちた。 Tears trickled down her cheeks.

5.5　実験結果のまとめ

提案手法は従来手法に比べ対訳句の抽出数は少ないが, 抽出精度はかなり高い. これは誤った対訳句の抽出を抑制できているといえる. また, 提案手法により抽出された出現回数 1 回の対訳句は,2 回以上に比べ抽出精度が低かった.

6　考察

6.1　追加抽出

提案手法により抽出される対訳句の数は従来手法に比べて少ない. そこで対訳句の数を増加させるため, 対訳句から対訳文パターンを作成し, 対訳句の追加抽出を行う.

6.1.1　追加抽出の手順

追加抽出では, 提案手法の手順 4-3 以降に以下の手順を加える. 区別のため以降は手順 4-3 により抽出される対訳句を対訳句 A と呼び, 追加抽出により得られる対訳句を対訳句 B と呼ぶ.

手順 6.1.1-1　対訳句 A を用いた対訳文パターンの作成

任意の対訳学習文 A から対訳句 A を 1 つだけ変数化し,1 変数のみの対訳文パターンを作成する. 図 7 にその作成手順と例を示す.

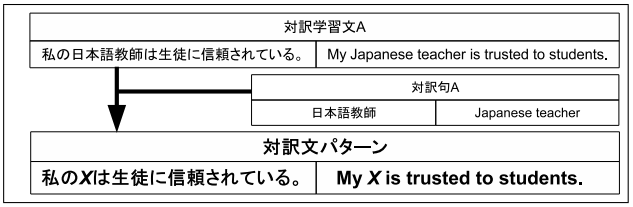

図 7　対訳句 A による 1 変数のみ対訳文パターンの作成手順

手順 6.1.1-2　対訳句 B の抽出

任意の対訳学習文 B から対訳文パターンの変数部を抽出し対訳句 B とする. 図 8 に対訳句 B の抽出手順と例を示す.

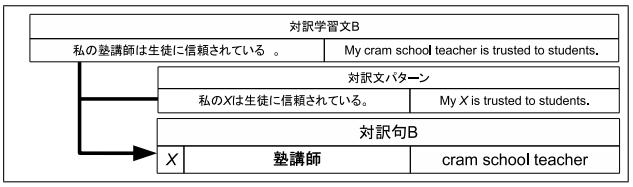

図 8　1 変数のみ対訳文パターンによる対訳句 B の抽出手順

以上を全ての対訳学習文と対訳文パターンに対して行う.

6.1.2　追加抽出の結果

追加抽出により得られた対訳句 B の異なり数を表 7 に示す.

表 7　追加抽出による対訳句 B の数

	抽出数
追加抽出	91,200 対

6.1.3　追加抽出の精度評価

追加抽出により得られた対訳句 B を 5.3.2 項と同様に評価する. 表 8 に評価結果を示す.

表 8　対訳句 B 100 対の人手評価結果

追加抽出

出現回数	○	△	×
1 回	66 対	31 対	3 対
2 回以上	68 対	32 対	0 対

6.2　評価 △ の原因調査

表 8 より, 追加抽出では評価 △ の数が増加している. この原因について表 8 中の評価 △ である対訳句 B 63 対を調査する.

6.2.1　人称代名詞

表 9 に評価 △ である対訳句 B の例を示す. この例のような対訳句 B の英語側”her”は日本語において省略することが可能であるが, 日本語側に対応する語がないため評価 △ としている. この理由で評価 △ となった対訳句 B は 63 対の内, 36 対だった.

表 9　人称代名詞により評価 △ となった対訳句 **B** の例

対訳句 B(X)	エプロンで涙を拭い dried **her** tears on **her** apron.
対訳学習文 B	彼女はエプロンで涙を拭いた。 She dried her tears on her apron.
対訳文パターン	彼女は X た。 She X.
対訳句 A	アメリカから来 came from the United States
対訳学習文 A	彼女はアメリカから来た。 She came from the United States.

6.2.2　日本語側に語の省略がある対訳句の増加

語の省略がある対訳学習文 A を用いて抽出されている対訳句が増加していた．表 10 に語の省略がある対訳学習文 A を用いて抽出された対訳句 B の例を示す．この例の対訳句 B の日本語側”彼は”に対応する語が, 英語側に存在しないため評価 △ とした．語の省略がある対訳学習文 A が原因で評価 △ となった対訳句 B は 63 対の内 32 対であった．

表 10　語の省略により評価 △ となった対訳句 **B** の例

対訳句 B(X)	彼は夜遅くまで勉強を続けた countinued working till late at night
対訳学習文 B	彼は夜遅くまで勉強を続けた。 He countinued working till late at night.
対訳文パターン	X。 He X.
対訳句 A	夜遅く帰宅した got home late at night
対訳学習文 A	夜遅く帰宅した。 **He** got home late at night.

6.3　評価基準の見直し

ここまでの評価基準では, 対訳句 B の英語側に人称代名詞が含まれており, 日本語側で対応する語が存在しない場合は評価 △ としてきた．しかしそのような場合でも, 日本語において省略することが可能ならば対訳句として適切である．よって評価 △ の評価基準を以下のように見直す．

- △$^{\bigcirc}$…日本語において省略することが可能な語以外で日英の対応が適切である．
- △$^{\times}$…日英の対応が部分的に適切である．

評価基準の見直しにより, 表 9 に示す評価 △ の対訳句 B は評価 △$^{\bigcirc}$ となる．これは対訳句 B の人称代名詞”her”に日本語で対応する語が省略されるのは一般的だからである．逆に表 10 は評価 △$^{\times}$ となる．対訳句 B の”彼は”に英語で対応する語を省略するのは不自然だからである．

表 11 に評価 △ である対訳句 B 63 対を評価し直した結果を示す．また評価 △ から評価 △$^{\times}$ となった対訳句 B 33 対の内,29 対の対訳学習文 A に主語の省略があった．

表 11　評価 △ であった対訳句 **B 63** 対の再評価結果

出現回数	△$^{\bigcirc}$	△$^{\times}$
1 回	20 対	11 対
2 回以上	10 対	22 対

6.3.1　追加抽出結果のまとめ

追加抽出では提案手法に比べ対訳句の抽出数は増加したが, 評価 △ も増加し, 抽出精度が低下した．しかし, 追加抽出と従来手法を比べると抽出精度は依然として高い．また, 表 3 と同様に出現回数 1 回の対訳句は 2 回以上に比べ抽出精度が低かった．

6.4　提案手法の改良

対訳句の抽出数に関して, 表 2 より従来手法に比べ提案手法は大きく減少している．しかし 6.1 節より追加抽出を行うことで, 対訳句を増加させられることがわかった．さらに手順 6.1.1-1 において, 追加抽出によって得られた対訳句 B を対訳句 A に代用し, 対訳文パターンを作成することが可能である．つまり抽出された対訳句から対訳文パターンを作成し, 対訳句の抽出を行うことを繰り返すことで, さらに対訳句の数を増加させることができる．

一方, 抽出精度に関しては表 3 と表 8 より従来手法に比べ提案手法はかなり高い．しかし追加抽出において評価 △ が増えた．評価 × はあまり増えなかったことから, さらに抽出精度を高めるためには評価 △ となる原因を取り除く必要がある．6.3 節より, 主語の省略がある対訳学習文 A が評価 △ となる主な原因であることがわかった．従って, そのような対訳学習文を取り除くか, 或いは同じように主語の省略がある対訳学習文 B とセットで用いる必要がある．

7　おわりに

従来手法における複数の変数を含む対訳文パターンによる対訳句の抽出には誤った対訳句が抽出される問題があった．提案手法では 1 変数の対訳文パターンのみを対訳句の抽出に用いることで, 抽出数は減少したが誤った対訳句の抽出を抑制することができた．また, 抽出された対訳句から対訳文パターンを作成し, 対訳句の抽出を繰り返すことで, 抽出精度は低下するが抽出数を増加させることができた．

参考文献

[1] 江木孝史, 村上仁一, 徳久雅人:“句に基づく対訳句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, pp.951-954, 2014.

[2] Hiroshi Maruyama:“Pattern-Based Translation: Context-Free Transducer and Its Applications to Practical NLP”, n Proc.of Natural Language Pacific Rim Symposium, pp.232-237, 1993.

[3] Franz Josef Och, Hermann Ney:“A Systematic Comparison of Various Statistical Alignment Models”, Computational Linguistics, pp.19-51, 2003.

[4] 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.