概要

対訳句は翻訳において重要な要素である．対訳句を手動で抽出する場合，コストが高く，作成数に制限がかかる．

江木は統計的手法を用いて，大量の対訳文から，対訳句を自動で抽出する手法を提案した．自動であるため，対訳句を手動で抽出する場合に比べコストが低く，大量の対訳句を抽出した．しかし，対訳句の抽出精度はまだ低い．その原因の一つは，変数が多い対訳文パターンにあると考えられる．なぜなら，対訳文パターン中に占める変数の割合が多くなるほど，対訳文と対訳文パターンが一致しやすくなるが，文構造を一致させることは困難になる．また，対訳言語が一対一で対応していない場合，対訳文パターン中で連続した変数部において，適切な位置で変数部を区切って対訳句を抽出することが困難になる．

本研究では，対訳句の抽出精度の向上を目指し，変数が 1 つの対訳文パターンを用いて，対訳句を抽出する．また，変数が 1 つの対訳文パターンを用いて抽出した対訳句に基づき，さらに変数が 1 つの対訳文パターンを作成した．そして，増加させた対訳文パターンを用いて，対訳句の抽出を行った．最後に，抽出した対訳句を人手評価し，対訳句の抽出精度を調査した．

評価の結果，変数が 1 つの対訳文パターンを用いた対訳句の抽出は，抽出精度が非常に高いことがわかった．加えて，抽出した対訳句から，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，ある程度の抽出精度を維持しながら，抽出数を増加させることができた．しかし，変数が複数の対訳文パターンを用いた対訳句の抽出に比べ，抽出数が非常に少なくなった．

今後は，対訳句の抽出をさらに繰り返すことや，変数が 2 つの対訳文パターンを用いて，対訳句の抽出を行うなどして，抽出数を増加させることを考える必要がある．

目 次

図 目 次

表 目 次

第1章　はじめに

対訳句は翻訳において重要な要素である．対訳句を手動で抽出する場合，コストが高く，抽出数に制限がかかる．

江木は統計的手法を用いて，大量の対訳文から対訳句を自動で抽出する手法 [1] を提案した．自動であるため，対訳句を手動で抽出する場合に比べコストが低く，大量の対訳句を抽出した．しかし，対訳句の翻訳精度はまだ低い．その原因の一つは，変数が複数の対訳文パターンにあると考えられる．なぜなら，対訳文パターン中に占める変数の割合が多くなるほど，対訳文と対訳文パターンが一致しやすくなるが，文構造を一致させることは困難になる．また，対訳言語が一対一で対応していない場合，対訳文パターン中で連続した変数部において，適切な位置で変数部を区切って対訳句を抽出することが困難になる．

本研究では，対訳句の抽出精度の向上を目指し，変数が 1 つの対訳文パターンを用いて，対訳句の抽出を行う．また，変数が 1 つの対訳文パターンを用いて抽出した対訳句に基づき，さらに変数が 1 つの対訳文パターンを作成した．そして，増加させた対訳文パターンを用いて，対訳句の抽出を行う．最後に，抽出した対訳句からランダムに 100 対を人手評価し，対訳句の抽出精度を調査する．

本論文の構成を以下に示す．第 2 章で，対訳句の自動抽出手法について説明する．第 3 章で，変数が 1 つの対訳文パターンを用いた対訳句の自動抽出手法について説明する．第 4 章で，実験条件を述べる．第 5 章で，実験結果を示す．第??章で，追加実験条件を述べる．第??章で，追加実験結果を示す．第 11 章で，本研究の考察を述べる．

第2章　対訳句の抽出

対訳句の抽出について，日本語と英語の場合を例にして説明する．なお，本章は江木孝史，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”第 4 章を参考にしている．

2.1　概要

対訳句とは，異なる二言語間において，それぞれ対訳関係にある語の対である．この対訳句を抽出する際，対訳文と対訳文パターン辞書が必要である．対訳文とは，異なる二言語間において，それぞれ対訳関係にある文の対である．対訳文パターンは，対訳文を任意の対訳単語または対訳句単位で変数化することで得られる．

対訳句を手動で抽出する場合，抽出精度は最高であるがコストが高く，抽出数に制限がかかる．これに対して，対訳句を自動で抽出する場合，コストが低く，抽出数は多くなる．しかし，対訳句の抽出精度は低い．

2.2　自動抽出

古典的な翻訳手法に用いる対訳単語辞書は，従来的に手動で作成していた．そこで江木らは，手動で作成していた対訳単語辞書を，統計的手法を用いて，自動で作成した．具体的には，IBM 翻訳モデル [2] を用いて，対訳文から対訳単語辞書を自動作成する．その後，対訳単語辞書を用いて，対訳文から対訳文パターン辞書を作成する．最後に，対訳文パターンを用いて，対訳文から対訳句を抽出する．

2.3　翻訳モデルの概要

翻訳モデルは，ある言語の単語列から別の言語の単語列へと確率的に翻訳を行うためのモデルである．

2.3.1 IBM 翻訳モデル

統計翻訳で代表的な翻訳モデルとして「，Brown らが提案したフランス語英語翻訳モデル（通称，IBM 翻訳モデル）がある．IBM 翻訳モデルは，Model1 から Model5 までの５つのモデルからなる．原言語をフランス語 f, 目的言語を英語 e と想定して説明を行う．

IBM 翻訳モデルでは，フランス語文 f と英語文 e の翻訳モデル $P(f|e)$ を計算するために，アライメント α を用いる．(2.1) に IBM 翻訳モデルの基本的な計算式を示す．

$$P(f|e) = \sum_{\alpha} P(f, \alpha \ |e) \tag{2.1}$$

2.4 GIZA++

GIZA++[3] は，対訳文から対訳単語と単語翻訳確率を自動的に得ることができる．単語翻訳確率とは，原言語と目的語における単語の対応関係 (Word Alignment) の確率である．単語翻訳確率を IBM Model1〜5 を用いて計算する．GIZA++を用いることで，以下のファイルを得る．

1. T TABLE(Translation Table)

T TABLE は，IBM Model1〜3 により作成された翻訳確率 $P(f|e)$ のデータである．f は原言語で，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)，原言語の単語 ID(f_id)，原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 (p^0)，繁殖数が 1 である確率 (p^1)，…，繁殖数が n である確率 (p^n) で構成される．

2.5 抽出手順

対訳句を自動で抽出するには，以下に示す 3 つの手順を踏む．

手順 1 対訳単語辞書の作成

GIZA++を用いて，対訳文から対訳単語辞書を作成する．

手順 2 対訳文パターン辞書の作成

対訳単語辞書を用いて，対訳文から対訳文パターン辞書を作成する．

手順 3 対訳句の抽出

対訳文パターン辞書を用いて，対訳文から対訳句を抽出する．

2.6 抽出手順の詳細

2.6.1 対訳単語辞書

対訳文から対訳単語辞書を作成する．まず，GIZA++を用い，対訳文の日英方向と英日方向で，対訳単語と単語翻訳確率をそれぞれ得る．そして，両者の単語翻訳確率を掛け合わせ，単語確率と呼ぶ確率値を得る．最後に対訳単語辞書を，対訳単語と単語確率で構成する．図 2.1 に日英方向の対訳単語辞書の作成例を示す．

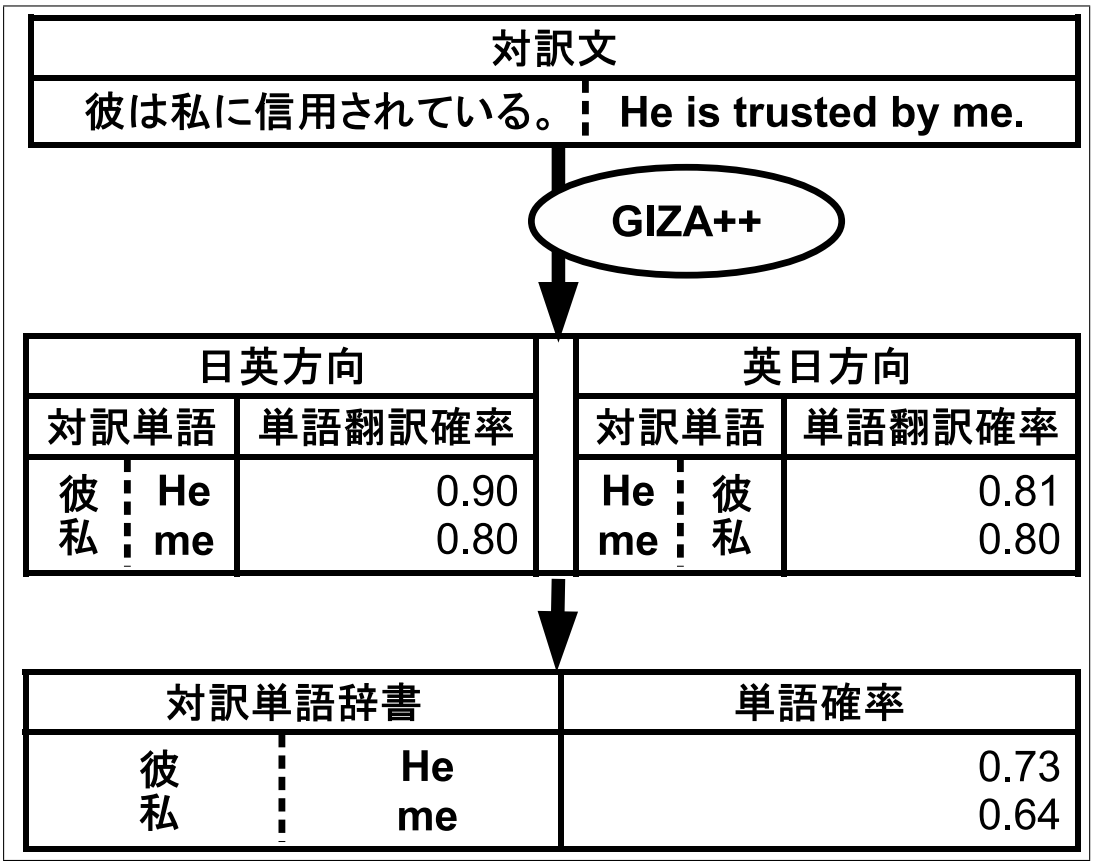

図 2.1: 日英方向の対訳単語辞書の作成

2.6.2　対訳文パターン辞書

対訳文から対訳文パターン辞書を作成する．まず，対訳文の各日本語単語と，対訳単語辞書の日本語単語を照合する．次に，対訳単語辞書の日本語単語に対応する英語単語と，対訳文の各英単語を照合する．最後に，日英両方の単語が照合に成功した場合，該当箇所を変数化する．図 2.2 に対訳文パターン辞書の作成例を示す．

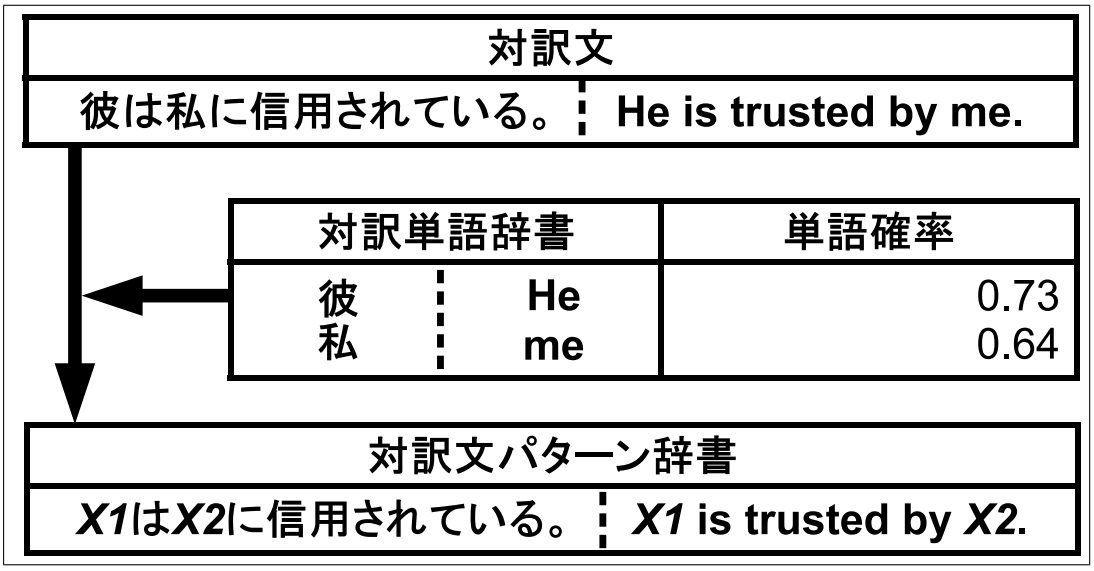

図 2.2: 対訳文パターン辞書の作成

2.6.3 対訳句の抽出

対訳文から対訳句を抽出する．まず，対訳文と対訳文パターン辞書を照合する．次に，対訳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訳文の語を，対訳句として抽出する．図 2.3 に対訳句の抽出例を示す．

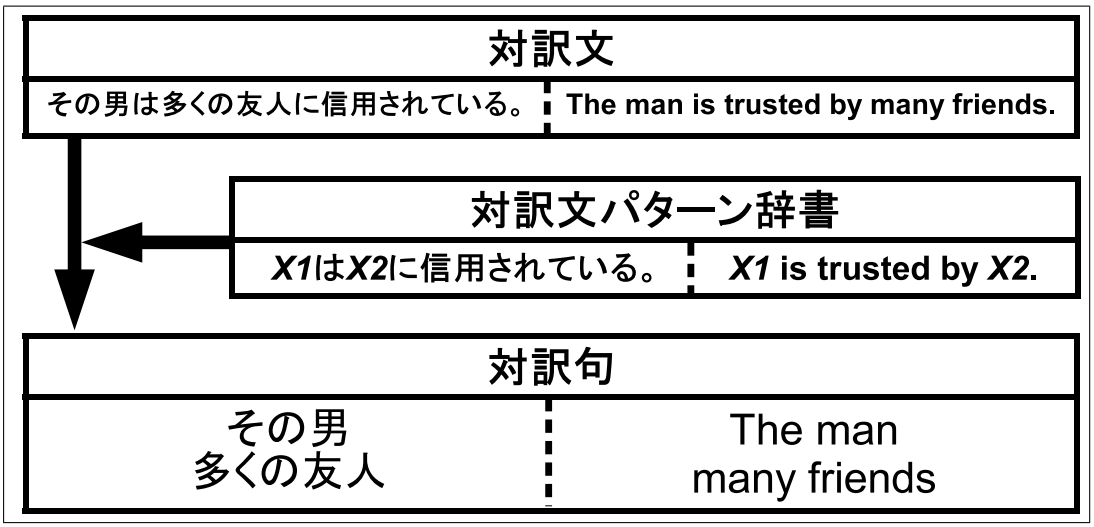

図 2.3: 対訳句の抽出

2.7 問題点

対訳句の自動抽出は，抽出精度が低い．抽出精度が低くなる原因の一つは，変数が複数の対訳文パターンである．対訳文パターン中の変数の数が多くなると，対訳文と対訳文パターンが適合しやすくなるが，対訳文と対訳文パターンを適切に組み合わせることが困難になる．図 2.4 に誤った対訳句が抽出される例を示す．

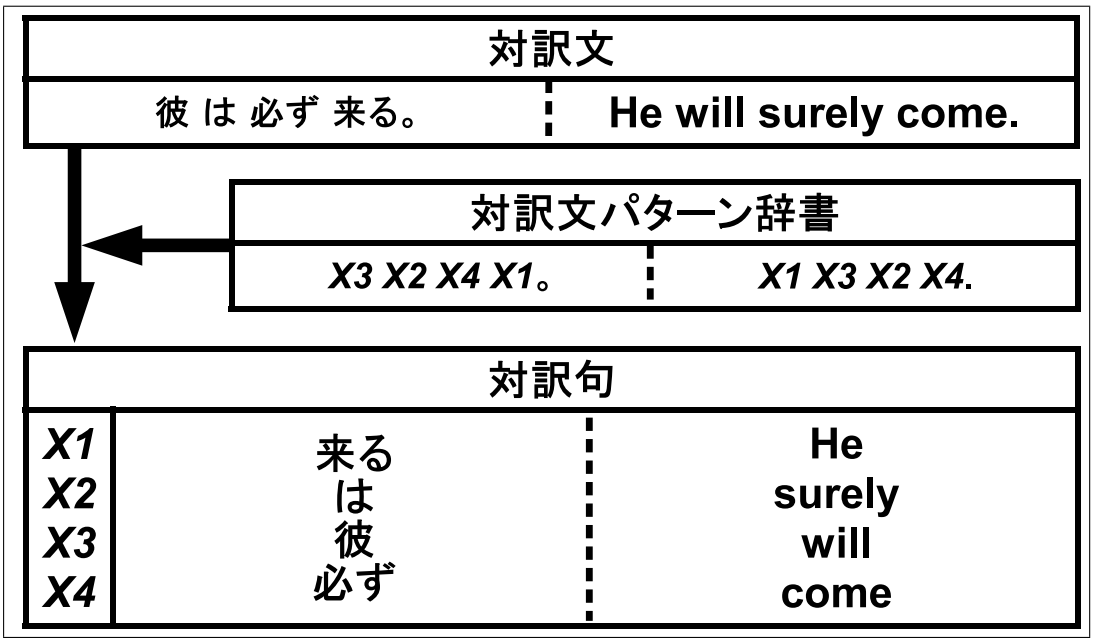

図 2.4: 誤った対訳句の抽出

第3章　提案手法

3.1　概要

変数が多い対訳文パターンほど，対訳句の抽出精度が低い．そこで本研究では，変数が 1 つの対訳文パターンを用いて，対訳句の抽出を行う．具体的には，第 2 の 2.6.2 項で対訳文パターン辞書を作成する際に，変数化する箇所を 1 つだけにし，変数が 1 つの対訳文パターン辞書を作成する．その他は第 2 と同様の手順である．図 3.1 に変数が 1 つの対訳文パターン辞書の作成例を示す．

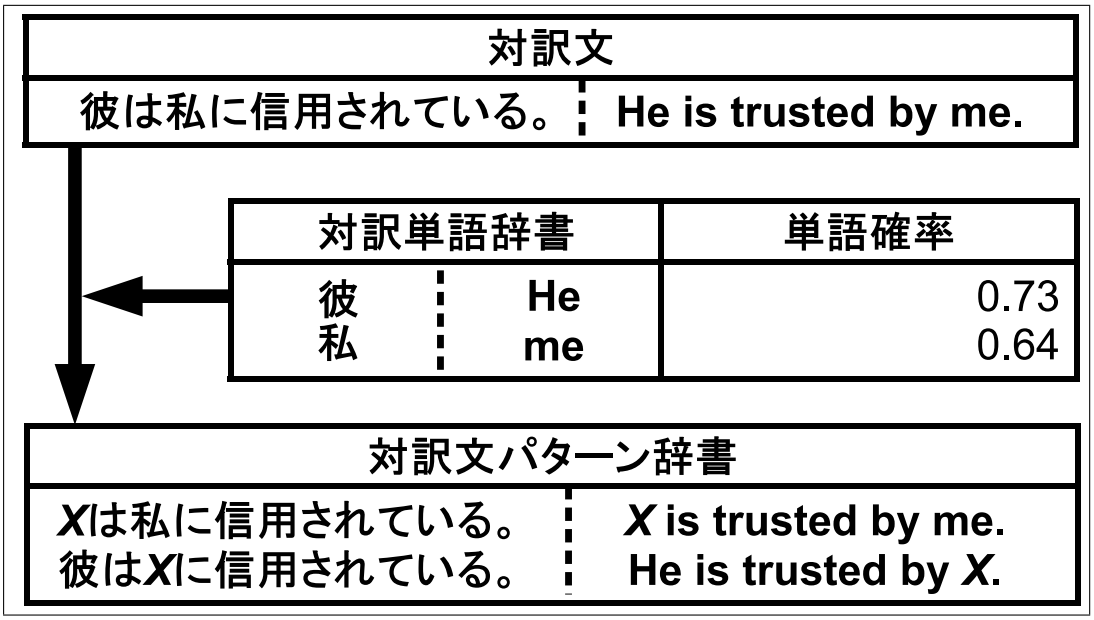

図 3.1: 変数が 1 つの対訳文パターン辞書の作成

第4章　実験概要

対訳句の抽出の際に対訳文パターン辞書が，変数が 1 つの対訳文パターンのみの場合 (以下，提案手法) と，変数が複数の対訳文パターンを含む場合 (以下，従来手法) を考慮して実験を行う．

4.1　実験目的

変数が多い対訳文パターンは，対訳句の抽出精度が低い．そこで本研究では，変数が 1 つの対訳文パターンを用いて，対訳句の抽出を行う．これにより，対訳句の抽出精度を高めることを目的とする．

4.2　実験データ

実験には対訳文 159,998 文 [4] を使用する．また，表 4.1 に，対訳文パターンの作成に用いる対訳単語数と，対訳句の抽出に用いる対訳文パターン数を示す[1]．従来手法で用いる対訳文パターンは，ある対訳文中から全ての対訳単語を変数化した，最も変数が多い対訳文パターンである．

表 4.1: 対訳単語，対訳文パターン数

	対訳単語	対訳文パターン
提案手法	404,144	1,176,922
従来手法	404,144	165,051

[1]
対訳単語:~/P1-PHRASE/p1-word.giza/p8-word-format/output.txt
提案手法:~/P2-PATTERN/PATTERN-N/p7-pattern-only-N0/output.txt
従来手法:~/P2-PATTERN/PATTERN-C/p6-pattern-check/output.txt

4.3 評価方法

抽出する対訳句に対し，ランダムな 100 対を以下の評価基準に従い，人手で評価する．評価基準は，対訳句の日本語を基に，英語の対応を焦点としている．

- ◯…日英の対応が適切
- △…日英の対応が部分的に適切
- ×…日英の対応が不適切

第5章　実験結果

5.1　抽出結果

表 5.1 に対訳句の抽出結果を示す[1].

表 5.1: 対訳句の抽出結果

	◯	△	×	抽出数	異なり数
提案手法	90	7	3	26,243	17,540
従来手法	41	21	38	27,301,971	5,058,468

5.1.1　評価 ◯ の抽出例

表 5.2 に評価 ◯ の対訳句の抽出例を示す.

表 5.2: 評価 ◯ の対訳句の例

提案手法

対訳句 (X)	成功の見込み	a chance of success
対訳文	彼は成功の見込みがある。	He has a chance of success.
対訳文パターン	彼は X がある。	He has X.
対訳単語	根性	guts
対訳文	彼は根性がある。	He has guts.

従来手法

対訳句 ($X3$)	身が締まっている	has firm flesh
対訳文	この魚は身が締まっている。	This fish has firm flesh
対訳文パターン	$X2$ $X1$ $X3$。	$X1$ $X2$ $X3$.
対訳単語	飛ぶ	flies
対訳文	鳥が飛ぶ。	A bird flies.

[1]
提案手法:~/P3-TABLE/table-N-DP.normalize/p14-phrase-format/output.txt
従来手法:~/P3-TABLE/table-C-DP.normalize/p14-phrase-format/output.txt

5.1.2 評価 △ の抽出例

表 5.3 に評価 △ の対訳句の抽出例を示す.

表 5.3: 評価 △ の対訳句の例

提案手法

対訳句 (X)	彼女はお産	childbirth
対訳文	彼女はお産で死んだ。	She died in childbirth.
対訳文パターン	X で死んだ。	She died in X.
対訳単語	お産	childbirth
対訳文	お産で死んだ。	She died in childbirth.

従来手法

対訳句 ($X3$)	ケンが	is
対訳文	ケンが来ています。	Ken is here.
対訳文パターン	$X3$ $X1$ $X2$。	$X1$ $X3$ $X2$.
対訳単語	山	mountains
対訳文	山がそびえる。	The mountains tower.

5.1.3 評価 × の抽出例

表 5.4 に評価 × の対訳句の抽出例を示す.

表 5.4: 評価 × の対訳句の例

提案手法

対訳句 (X)	た	My
対訳文	時計が止まった。	My watch has stopped.
対訳文パターン	時計が止まっX。	X watch has stopped.
対訳単語	た	The
対訳文	時計が止まった。	The watch has stopped.

従来手法

対訳句 ($X1$)	彼は後ろ	back
対訳文	彼は後ろを向いた。	He looked back.
対訳文パターン	$X1$ $X3$ $X2$。	$X2$ $X3$ $X1$.
対訳単語	すばやく	swiftly
対訳文	素早く受け入れた。	He accepted swiftly.

5.2 実験結果のまとめ

表 5.1 より，提案手法による対訳句の抽出精度は高いことがわかる．しかし，抽出数は従来手法に比べ非常に少なくなった．

基本的に，対訳文パターンを増加させれば抽出できる対訳句も増加する．だが，表 4.1 より，提案手法は従来手法よりも多くの対訳文パターンを用いて対訳句を抽出しているにもかかわらず，対訳句の抽出数が非常に少なくなっている．これは，対訳文パターン中の変数の数が少ないほど，対訳文と対訳文パターンが適合が困難になるためである．

第6章　追加実験　概要

提案手法は，従来手法に比べ対訳句の抽出数が非常に少ない．そこで，抽出した対訳対訳句を対訳単語辞書の代わりに用いて，対訳文パターンを作成し，対訳句の抽出を行う．
抽出した対訳句を用いて，対訳文パターンを作成するため，対訳句の抽出を繰り返すことができる．本実験では 2 回まで対訳句の抽出を繰り返す．

6.1　実験目的

変数が 1 つの対訳文パターンは，対訳句の抽出精度が高い．つまり，対訳単語辞書の代わりに抽出された対訳句を用いて，対訳文から適切な対訳文パターン辞書を作成することができる．この対訳文パターン辞書を用いて，提案手法の抽出精度を維持しながら，抽出数を増加させることを目指す．

6.2　追加実験の手順

追加実験では，第 5 章で提案手法により抽出した対訳句を用いて，変数が 1 つの対訳文パターンを作成する．図 6.1 に，追加実験における対訳文パターンの作成例を示す．
対訳文パターン辞書の作成を除いた手順は，提案手法と同様である．

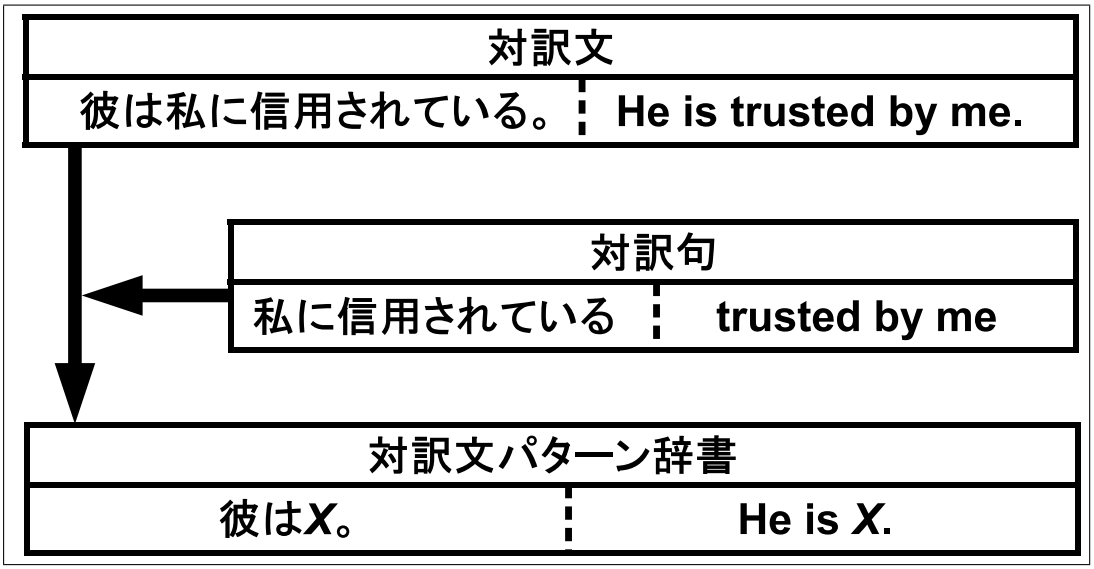

図 6.1: 対訳句を用いた対訳文パターンの作成

第7章　追加実験　1回目

7.1　概要

本章では，第5章の提案手法により抽出した対訳句を用いて，対訳文パターン辞書を作成する．

7.2　実験データ

実験には対訳文 159,998 文 [4] を使用する．また，表 7.1 に，対訳文パターンの作成に用いる対訳句数と，対訳句の抽出に用いる対訳文パターン数を示す[1]．

表 7.1: 対訳句，対訳文パターン数

	対訳句（対訳文パターン作成）	対訳文パターン
追加実験　1回目	17,540	735,717

[1] 対訳句:~/P3-TABLE/table-N-DP.normalize/p14-phrase-format/output.txt
対訳文パターン:~/P4-LOOP/P01-PATTERN-N/p7-pattern-only-N/output.txt

第8章　追加実験　1回目　実験結果

8.1　抽出結果

表 8.1 に対訳句の抽出結果を示す[1].

表 8.1: 対訳句の追加抽出　1回目結果

	○	△	×	抽出数	異なり数
追加実験　1回目	99	1	0	11,571,370	91,200

8.1.1　評価 ○ の抽出例

表 8.2 に評価 ○ の対訳句の抽出例を示す.

表 8.2: 評価 ○ の対訳句の例

対訳句 (X)	しばしば遅刻する	often late
対訳文	彼はしばしば遅刻する。	He is often late.
対訳文パターン	彼は X。	He is X.
対訳単語	自由にそこへ行ける	free to go there
対訳文	彼は自由にそこへ行ける。	He is free to go there.

[1] 追加実験 1回目:~/P4-LOOP/P02-TABLE-N0/p14-phrase-format/output.txt

8.1.2　評価 △ の抽出例

表 8.3 に評価 △ の対訳句の抽出例を示す.

表 8.3: 評価 △ の対訳句の例

対訳句 (X)	私は壁に寄り掛かった	leaned against the wall
対訳文	私は壁に寄り掛かった。	I leaned against the wall.
対訳文パターン	X。	I X.
対訳単語	彼女との婚約を解消した	broke off my engagement to her
対訳文	彼女との婚約を解消した。	I broke off my engagement to her.

8.2　実験結果のまとめ

表 8.1 より，追加実験 1 回目により抽出した対訳句は，精度が非常に高いことがわかる．加えて，抽出数が増加した．しかし，抽出した対訳句が長文化した.

第9章　追加実験　２回目

9.1　概要

本章では，第 8 章の追加実験　1 回目により抽出した対訳句を用いて，対訳文パターン辞書を作成する．

9.2　実験データ

実験には対訳文 159,998 文 [4] を使用する．また，表 9.1 に，対訳文パターンの作成に用いる対訳句数と，対訳句の抽出に用いる対訳文パターン数を示す[1]．

表 9.1: 対訳句，対訳文パターン数

	対訳句（対訳文パターン作成）	対訳文パターン
追加実験　1 回目	91,200	812,199

[1]
対訳句:~/P4-LOOP/P02-TABLE-N0/p14-phrase-format/output.txt
対訳文パターン:~/P4-LOOP/P11-PATTERN-N/p7-pattern-only-N/output.txt

第10章　追加実験　２回目　実験結果

10.1　抽出結果

表 10.1 に対訳句の抽出結果を示す[1].

表 10.1: 対訳句の追加抽出　２回目結果

	○	△	×	抽出数	異なり数
追加実験　１回目	67	33	0	456,955,607	163,525

10.1.1　評価 ○ の抽出例

表 10.2 に評価 ○ の対訳句の抽出例を示す.

表 10.2: 評価 ○ の対訳句の例

対訳句 (X)	昨日偶然彼に会った	chanced on him yesterday
対訳文	昨日偶然彼に会った	I chanced on him yesterday.
対訳文パターン	X。	I X.
対訳単語	休み中に北海道に行った	went to Hokkaido on vacation
対訳文	休み中に北海道に行った。	I went to Hokkaido on vacation.

[1] 追加実験 ２回目:~/P4-LOOP/P12-TABLE-N0/p14-phrase-format/output.txt

10.1.2　評価 △ の抽出例

表 10.3 に評価 △ の対訳句の抽出例を示す.

表 10.3: 評価 △ の対訳句の例

対訳句 (X)	彼は大口を叩いた	talked big
対訳文	彼は大口を叩いた。	He talked big.
対訳文パターン	X。	He X.
対訳単語	彼はひどいけいれんを起こす	has a bad twitch
対訳文	彼はひどいけいれんを起こす。	He has a bad twitch.

10.2　実験結果のまとめ

表 10.1 より，追加実験 2 回目により抽出した対訳句は，精度が高いことがわかる．加えて，抽出数がさらに増加した．しかし，評価△が増加し，抽出した対訳句がさらに長文化した.

第11章　考察

対訳句の抽出を行うためには，2つの対訳文が必要である．以下，区別のために，対訳文パターンを作成するために用いる対訳文を対訳文 α , 対訳句を抽出する対訳文を対訳文 β とする．

11.1　誤り解析

第4章，第7章，第9章の実験結果より，変数が1つの対訳文パターンを用いて抽出した対訳句から，評価 △ または評価 × である対訳句を対象に，誤り解析を行う．

11.1.1　語の省略

表11.1に評価 △ の対訳句の例を示す．この例では，対訳文 α の日本語側で，英語側“He”に対応する語が省略されている．一方，対訳文 β には語の省略はない．この語の省略の有無により，抽出される対訳句の日英の語の対応が部分的に不適切になっている．

表 11.1: 評価 △ の対訳句の例

対訳句 (X)	彼は声に張り	a strong voice
対訳文 β	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	freckles
対訳文 α	そばかすがある。	**He** has freckles.

11.1.2　不適切な対訳単語

表 11.2 に評価 × の対訳句の例を示す．この例では，対訳単語が不適切である．日本語の助詞などの付属語や，英語の冠詞などの指示詞は，適切な対応をとるのが困難であり，不適切な対訳文パターンを作成する原因となる．

表 11.2: 評価 × の対訳句の例

対訳句 (X)	た	My
対訳文 β	時計が止まった。	The watch has stopped.
対訳文パターン	時計が止まっX。	X watch has stopped.
対訳単語	た	The
対訳文 α	時計が止まった。	The watch has stopped.

11.1.3　対訳文パターンの汎化

追加実験において，追加実験　2 回目では，追加実験　1 回目に比べ評価 △ が増加した．これは，対訳文パターンの過度な汎化が原因である．表 11.3 に例を示す．

語の省略がある対訳文から，長文化した対訳句を用いて，対訳文パターンを作成する場合，日本語文の全てを変数化した対訳文パターンが作成可能になる．このような，対訳文パターンは，語の省略がない対訳文パターンに比べ，広範な対訳文に合致し，誤った対訳句を抽出する．

表 11.3: 過度に汎化した対訳文パターンを用いた対訳句の抽出例

対訳句 (X)	彼は借金を清算した	cleared up his debt
対訳文 β	彼は借金を清算した。	He cleared up his debt.
対訳文パターン	X。	He X.
対訳句	夜遅く帰宅した	got home late at night
対訳文 α	夜遅く帰宅した。	He got home late at night.

11.2　語の出現頻度による精度調査

大量の対訳文から対訳句を抽出する際，対訳文中で出現頻度が低い語（低頻度語）は，高い語に比べ，正確な対訳句を得ることが難しい．そこで，対訳文中での出現頻度を基に，各実験による対訳句の抽出精度を再考する．具体的には，出現頻度が 1 回と 2 回以上の対訳句を分け，それぞれ評価する．表 11.4 に各実験の再考結果を示す[1]．．

表 11.4: 対訳句の抽出結果

出現頻度 1 回

	○	△	×	抽出数	異なり数
従来手法	37	26	37	13,464,695	3,676,582
提案手法	96	4	0	21,290	13,923
追加実験 1 回目	100	0	0	11,234,992	85,209
追加実験 2 回目	67	32	0	451,199,887	157,039

出現頻度 2 回以上

	○	△	×	抽出数	異なり数
従来手法	43	15	42	13,837,276	1,381,886
提案手法	91	8	1	4,953	3,617
追加実験 1 回目	89	7	4	336,378	5,991
追加実験 2 回目	91	6	3	5,755,720	6,486

[1]
従来手法:~/P3-TABLE/table-C-DP.normalize/p14-phrase-format/output.txt
提案手法:~/P3-TABLE/table-N-DP.normalize/p14-phrase-format/output.txt
追加実験 1 回目:~/P4-LOOP/P02-TABLE-N0/p14-phrase-format/output.txt
追加実験 2 回目:~/P4-LOOP/P12-TABLE-N0/p14-phrase-format/output.txt

11.3 抽出した対訳句による翻訳実験

各実験で作成した対訳文パターン及び抽出した対訳句を用いて，安場らの変換主導型統計機械翻訳 [5] による翻訳実験を行う．表 11.5 に，日本語 100 文を翻訳した実験結果を示す[2].

表 11.5: 変換主導型統計機械翻訳による翻訳可能文数

従来手法	29 文
提案手法	4 文
提案手法 +追加実験 1 回目 +追加実験 2 回目	4 文
従来手法 +提案手法 +追加実験 1 回目 +追加実験 2 回目	30 文

[2] ~/P8-DECORDER_V2.C-DP.type5/p8-decode_V4.unknown.V09/TMP/target.txt

第12章　おわりに

従来手法では，大量の対訳文から対訳句を自動抽出した．自動抽出により，対訳句を手動で作成する場合に比べコストが低く，大量の対訳句を抽出した．しかし，対訳句の翻訳精度はまだ低い．その原因の一つは，変数が複数の対訳文パターンにあると考えられる．

そこで，対訳句の抽出精度の向上を目指し，変数が１つの対訳文パターンを対訳句の抽出に用いた．実験結果より，提案手法は対訳句の抽出精度が高いことがわかった．しかし，抽出数は従来手法に比べ非常に少なくなった．これを受け，変数が１つの対訳文パターンを用いて抽出した対訳句に基づき，対訳文パターンを作成することで対訳句の抽出数の増加を試みた．最終的に，ある程度の抽出精度を維持しながら，抽出数を増加させることができた．

提案手法による対訳句の抽出数では，未だ実用的ではない．今後は，対訳句の抽出をさらに繰り返すことや，変数が２つの対訳文パターンを用いて，対訳句の抽出を行うなどして，抽出数を増加させることを考える必要がある．

謝辞

最後に，本研究を遂行するにあたり，ご指導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，村田真樹教授をはじめ, 自然言語処理研究室の方々に厚く御礼申し上げます.

また，参考にさせていただいた論文の著者の方々に，深く感謝致します.

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] 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] 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.

[5] 安場裕人，村上仁一，: “変換主導型統計機械翻訳の提案”, 言語処理学会第 24 回年次大会，pp7-9，2018

付録
/mnt/auto/hatter/usr18/backup/2018/s152112/~

●ランダム 100 対の人手評価結果
【従来手法】
全頻度:~/P3-table/table-C-DP.normalize/p14-phrase-format/score.sort
頻度 1 :~/P3-table/table-C-DP.normalize/p14-phrase-format/f1.sort
頻度 2 以上:~/P3-table/table-C-DP.normalize/p14-phrase-format/f2.sort

【提案手法】
全頻度:~/P3-table/table-N-DP.normalize/p14-phrase-format/score.sort
頻度 1 :~/P3-table/table-N-DP.normalize/p14-phrase-format/f1.sort
頻度 2 以上:~/P3-table/table-N-DP.normalize/p14-phrase-format/f2.sort

【追加実験 1 回目】
全頻度:~/P4-LOOP/P02-table-N0/p14-phrase-format/score.sort
頻度 1 :~/P4-LOOP/P02-table-N0/p14-phrase-format/f1.sort
頻度 2 以上:~/P4-LOOP/P02-table-N0/p14-phrase-format/f2.sort

【追加実験 2 回目】
全頻度:~/P4-LOOP/P12-table-N0/p14-phrase-format/score.sort
頻度 1 :~/P4-LOOP/P12-table-N0/p14-phrase-format/f1.sort
頻度 2 以上:~/P4-LOOP/P12-table-N0/p14-phrase-format/f2.sort

●変換主導型統計機械翻訳による実験結果
従来手法:~/P8-DECODER_V2.C-DP.type5/p8-decoder_V4.unknown.V09/target.C
提案手法:~/P8-DECODER_V2.C-DP.type5/p8-decoder_V4.unknown.V09/target.N
提案手法+追加実験 1 ・ 2 回目:
~/P8-DECODER_V2.C-DP.type5/p8-decoder_V4.unknown.V09/target.N.P02.P12
従来手法+提案手法+追加実験 1 ・ 2 回目:
~/P8-DECODER_V2.C-DP.type5/p8-decoder_V4.unknown.V09/target.C.N.P02.P12