文パターンを用いた句の抽出方法の検討

春野瑞季 村上仁一 徳久雅人 村田真樹
鳥取大学 工学部 知能情報工学科
鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
{s092051, murakami, tokuhisa, murata} @ ike.tottori-u.ac.jp

1 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．

本実験では，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高い．そこで本実験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．

2 GIZA++

GIZA++[8] とは，日英方向と英日方向の対訳文において最尤な単語の対応 (以下，単語対応と表記) 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．本実験では対訳単語の作成 (4.1 節) と句の翻訳確率の付与 (4.2 節) に GIZA++を用いる．

日英方向の単語対応の例を表 1 に示す．表 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 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 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手順を以下に示す．

手順 1 IBM モデルを用いて，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 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．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．(詳細は Koehn ら [4] 参照) “対称な単語対応” の例を図 1 に示す．図 1 は “grow-diag-final-and” を用いて “対称な単語対応” を求めた例であり，●部分は “対称な単語対応” を表している．

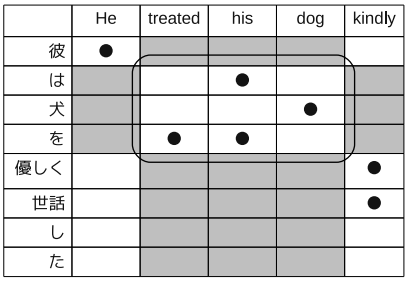

図 1 対称な単語対応の例 (grow-diag-final-and)

手順 3 “対称な単語対応” を用いて対訳句を抽出する．具体的には，対訳句の単語において，対訳句以外の単語との “対称な単語対応” がない対訳句を抽出する．図 1 において，枠で囲まれた部分が対訳句の例である．灰色部分には対訳句以外の単語との “対称な単語対応” が存在しない．よって，枠で囲まれた部分を抽出する．

抽出した対訳句の例を表 2 に示す．

表 2 Och らの方法を用いた対訳句の例

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

表 2 より，人間が見ると不自然な対訳句が抽出していることが確認できる．

4 提案手法

4.1 対訳文パターンを用いた対訳句の抽出方法

本実験では，対訳文パターンを用いて対訳句を抽出する．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．

手順 5 手順 4 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 6 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．

手順 7 変数が連続しない対訳文パターンのみを本実験で用いる対訳文パターンとする．

手順 8 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．

対訳文パターンを用いた対訳句の抽出方法の手順を図 2 に示す．

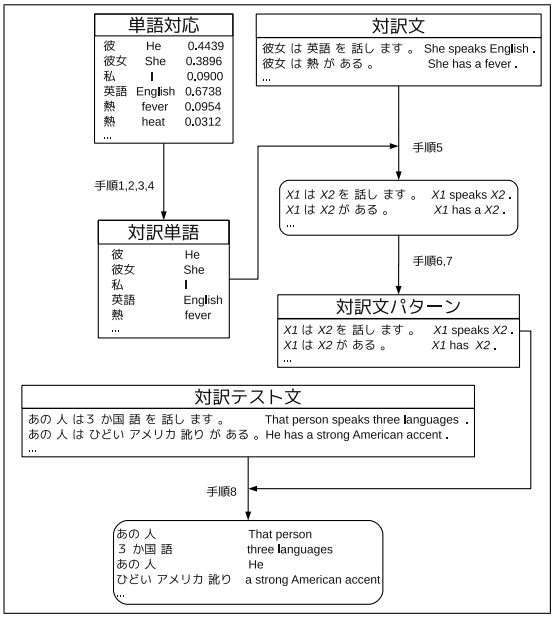

図 2 対訳文パターンを用いた対訳句の抽出方法の手順

4.2 対訳句の選別

4.1 節の抽出方法では，人間が見て不自然な対訳句を抽出する．そこで，対訳句の選別を行う．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．対訳句の選別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．
対訳句の例と対訳単語翻訳確率の例を表 3，4 に示す．

表 3 対訳句の例

日本語句	彼 の 耳
英語句	his ear

表 4 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 3 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 より，句の翻訳確率を求める．なお，表 4 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 3 の対訳句において，句の翻訳確率は-1037.884 となる．

手順 4 句の翻訳確率が一定の閾値 (β) 以上である対訳句を選別する．

5 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 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 節の手順 3 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．

6 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表 5 に示す．表 5 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 5 対訳文パターンを用いた対訳句の例

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 .

6.2 対訳句の人手評価結果

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 に示す．

表 6 対訳句の人手評価結果

評価○	評価×
42	8

表 6 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

人手評価における評価○と評価×の対訳句の例を以下に示す．

表 7 評価○の例

対訳句	(日本語)	まったく 無一物
	(英語)	utterly penniless
対訳テスト文	(日本語)	私 は まったく 無一物 に なっ た 。
	(英語)	I became utterly penniless .
対訳文パターン	(日本語)	$X1$ は $X2$ に なっ た 。
	(英語)	$X1$ became $X2$.
対訳文パターンの原文	(日本語)	彼 は 医者 に なっ た 。
	(英語)	He became a doctor .

表 8 評価×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 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．一方，表 8 は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 9 に示す．

表 9 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 10 に示す．

表 10 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 と表 10 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．

7 考察

7.1 対訳句の精度の考察

表 6 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 11 に示す 5 種類に分類した．

表 11 誤り解析の結果

a) 不適切な対訳単語
b) 対訳文パターンの不適切な適合
c) 主語の省略
d) 主語の違い
e) 対訳文パターンの不足

評価×の対訳句の例を以下に示す．

a) 不適切な対訳単語

表 12 不適切な対訳単語の例

対訳句	(日本語)	き ま し た
	(英語)	has started
対訳テスト文	(日本語)	雨 が 降っ て き ま し た 。
	(英語)	It has started raining .
対訳文パターン	(日本語)	雨 が 降っ て $X1$ 。
	(英語)	It $X1$ raining .
対訳文パターンの原文	(日本語)	雨 が 降っ て いる 。
	(英語)	It is raining .

表 12 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 “いる”は英単語 “is” 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．

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．

b) 対訳文パターンの不適切な適合

表 13 不適切な対訳文パターンの例

対訳句	(日本語)	立つ
	(英語)	a good writer
対訳テスト文	(日本語)	彼 は 筆 が 立つ 。
	(英語)	He is a good writer .
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 13 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．

c) 主語の省略

表 14 主語の省略の例

対訳句	(日本語)	彼 の 誠意 について
	(英語)	I
対訳テスト文	(日本語)	彼 の 誠意 について は 疑問 が ある 。
	(英語)	I have some doubts about his sincerity .
対訳文パターン	(日本語)	$X1$ は $X2$ が ある 。
	(英語)	$X1$ have $X2$.
対訳文パターンの原文	(日本語)	私 は 熱 が ある 。
	(英語)	I have a fever .

表 14 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．

表 14 の場合，対訳テスト文は，日本語文に主語 “私は”を補うと，対訳文パターンと適切に適合することができる．

d) 主語の違い

表 15 主語が異なる例

対訳句	(日本語)	彼 は 風邪 で 声
	(英語)	His voice
対訳テスト文	(日本語)	彼 は 風邪 で 声 が かすれ て いる 。
	(英語)	His voice is harsh because of the cold .
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 15 において，対訳テスト文の日本語文の主語が “彼”であるのに対し，英語文の主語は “His voice”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．

e) 対訳文パターンの不足

表 16 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 16 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，本実験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

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 17 に示す．

表 17 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 17 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 と表 17 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．

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．この場合，英語句は，例えば “The stone”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複数単語で構成される対訳句は，少ない．今後，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

7.1 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．

また，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．

さらに，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している．

8 おわりに

本実験では，自動で作成した対訳文パターンを用いて，対訳テスト文 100,000 文から，対訳句の抽出を試みた．実験の結果，閾値 $\beta = -2000$ を用いた場合，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．

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.