パターンに基づく統計翻訳における対訳句の確率の調査

松井智義 *1　村上仁一 *2
*1 鳥取大学 工学部 知能情報工学科
*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
{s142051,murakami}@ike.tottori-u.ac.jp

1 はじめに

江木らは，統計的手法を用いて自動的に対訳句と文パターンを作成して翻訳を行うパターンに基づく統計翻訳 [1][2] を提案した．この翻訳方法は，翻訳時の対訳句の選択に IBM Model1 を利用したフレーズ確率を使用している．しかし，対訳句の確率の計算方法はフレーズ確率以外も考えられる．

そこで，本研究では” フレーズ確率” の他に” フレーズ確率の総積” と”Dice 係数と類似度の積” の 3 種の確率を対訳句の選択に使用し翻訳を行う．そしてそれぞれの確率を用いて翻訳精度の差を調査する．

2 パターンに基づく統計翻訳 [1][2]

2.1 全体の流れと具体的な手順

パターンに基づく統計機械翻訳は対訳句と文パターンを用いて行う統計翻訳である．手順を以下に示す．

1. 対訳学習文と対訳単語確率を用いて，対訳単語を作成する．
2. 対訳学習文と対訳単語を用いて，単語に基づく文パターンを作成する．
3. 対訳学習文と単語に基づく文パターンを用いて，対訳句を作成する．
4. 対訳学習文と対訳句を用いて，句に基づく文パターンを作成する．
5. 対訳句と句に基づく文パターンを用いて，翻訳を行う．

対訳単語と単語に基づく文パターンと対訳句と句に基づく文パターンは確率が付与されている．一連の流れを図 1 に示す．

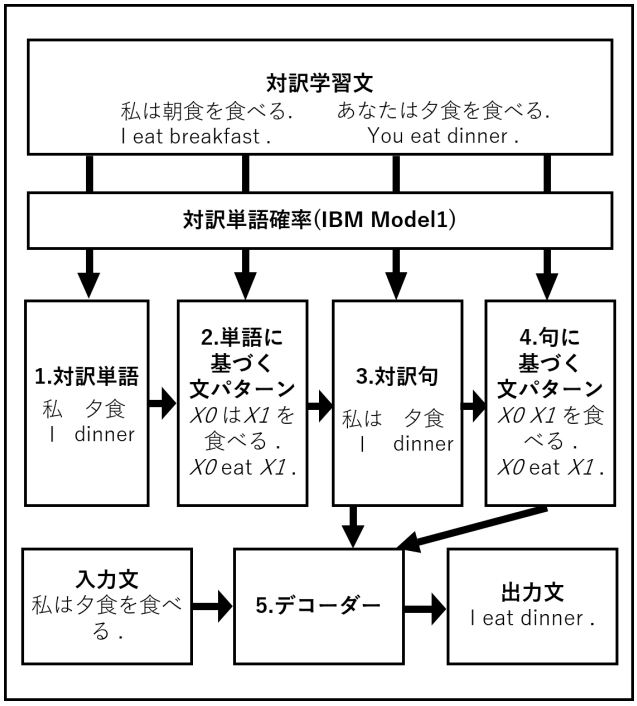

図 1　パターンに基づく統計翻訳の流れ

2.2 対訳句と対訳句確率

対訳句は日本語句と英語句の組である．対訳句確率は対訳句を選択する際に使用する確率である．従来の対訳句確率はフレーズ確率 (P_1) を使用している．フレーズ確率は IBM Model1 を利用して計算する．フレーズ確率の計算式を式 1 に示す．

$$P_1(\frac{J_0 \cdots J_{N-1}}{E_0 \cdots E_{M-1}}) = \prod_{n=0}^{M-1}\sum_{m=0}^{N-1} p(J_n|E_m)) * \prod_{m=0}^{M-1}\sum_{n=0}^{N-1}(p(E_m|J_n) \quad (1)$$

J_n; 対訳フレーズ中の日本語の単語 N; 日本語の単語数
E_m; 対訳フレーズ中の英語の単語 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 (IBM Model1 の値)

日本語単語が「この箱」，英単語が「this box」のときの日英方向のフレーズ確率の例を表 1 と式 2 に示す．

表 1　フレーズ確率の例

日本語単語	英単語	$p(E_m\|J_n)$	$p(J_n\|E_m)$
この	this	0.6	0.8
この	box	0.10	0.05
箱	this	0.05	0.01
箱	box	0.5	0.7

$$P_1(\frac{\text{この 箱}}{this\ box}) = \{P(\frac{\text{この}}{this}) + P(\frac{\text{この}}{box})\} * \{P(\frac{\text{箱}}{this}) + P(\frac{\text{箱}}{box})\} * \{P(\frac{this}{\text{この}}) + P(\frac{this}{\text{箱}})\} * \{P(\frac{box}{\text{この}}) + P(\frac{box}{\text{箱}})\} = 0.23 \quad (2)$$

3 実験目的

対訳句確率の計算方法はフレーズ確率以外にも考えられる．そこで本研究では対訳句確率を” フレーズ確率” と” フレーズ確率の総積” と”Dice 係数と類似度の積” に変更して日英翻訳を行う．そして 3 種の出力文を比較評価する．

3.1 フレーズ確率 (P_1)

フレーズ確率 P_1 は IBM Model1 を使用して求める (式 1).

3.2 フレーズ確率の総積 (P_2)

フレーズ確率の総積 P_2 は同一パターンの変数部のフレーズ確率 P_1 の総積である．計算式を式 3 に示す．

$$P_2(X_i) = \sum_{i=0}^{W} P_1(x_i) \tag{3}$$

W; 変数の数

以下に例を示す．表 2 に対訳句作成時に使用するデータの具体例を示す．

表 2　対訳学習文テストデータ

対訳学習文 (日)	この箱は木より作られる
対訳学習文 (英)	This box is made from wood
パターン (日)	$X1$ は $X2$ $X3$
パターン (英)	$X1$ $X3$ $X2$

表 2 の対訳学習文は対訳句作成に用いる文である．パターンは対訳句作成に用いるパターンである．対訳句作成の手順を以下に示す．

手順 1　対訳学習文とパターンを照合

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 3 に抽出した対訳句の具体例を示す．

表 3　手順 2 と手順 3 の具体例

変数	対訳句		フレーズ確率 (対数)
$X1$	この箱	This box	-0.4
$X2$	木より	from wood	-0.1
$X3$	作られる	is made	-0.05

手順 3　式 1 を用いて対訳句にフレーズ確率を付与

表 3 に示しているフレーズ確率は対数値である．フレーズ確率の総積はパターンにおける変数部のフレーズ確率の総積である．ここで，フレーズ確率の総積は 1 つの変数の組から作成された対訳句において同じ値となる．表 3 におけるフレーズ確率の総積の計算式を式 4 に示す．

$$\begin{aligned} & P_2(\frac{\text{この箱}}{this\ box}) \\ &= P_1(\frac{\text{この箱}}{this\ box}) * P_1(\frac{\text{木より}}{from\ wood}) * P_1(\frac{\text{作られる}}{is\ made}) \\ &= 2^{(-0.4)} * 2^{(-0.1)} * 2^{(-0.05)} = 2^{(-0.55)} = 0.683 \\ &= P_2(\frac{\text{木より}}{from\ wood}) = P_2(\frac{\text{作られる}}{is\ made}) \end{aligned} \tag{4}$$

3.3 Dice 係数と類似度の積 (P_3)

Dice 係数と類似度の積 P_3 は Dice 係数 $Dice(j,\ e)$ と類似度 P_s の積である．計算式を式 5 に示す．

$$P_3 = Dice(j,\ e) * P_S(r) \tag{5}$$

3.3.1 Dice 係数

Dice 係数を式 6 に示す．

$$Dice(j,\ e) = \frac{2 * count(j,\ e)}{count(j) + count(e)} \tag{6}$$

$count(j,\ e)$; 日本語句 j, 英語句 e が同じ対訳学習文において共起する頻度

$count(j)$; 対訳学習文の日本語文に日本語句 j が出現する頻度

$count(e)$; 対訳学習文の英語文に英語句 e が出現する頻度

日英の「この箱」と「this box」の共起頻度が 4,「この箱」の出現する頻度が 6,「this box」の出現する頻度が 8 の時の例を式 7 に示す．

$$\begin{aligned} Dice(\text{この箱},\ thisbox) &= \frac{2count(\text{この箱},\ this\ box)}{count(\text{この箱}) + count(this\ box)} \\ &= \frac{4}{6+8} = 0.28 \end{aligned} \tag{7}$$

3.3.2 類似度

類似度は対訳学習文とパターン原文の同一の単語の出現率である．類似度の計算方法を式 8 に示す．

$$P_s(r) = \frac{N_{j1}}{M_{j1}} * \frac{N_{j2}}{M_{j2}} * \frac{N_{e1}}{M_{e1}} * \frac{N_{e2}}{M_{e2}} \tag{8}$$

M_{j1}; 対訳学習文中の日本語単語数 M_{j2}; パターン原文の日本語単語数

M_{e1}; 対訳学習文中の英語単語数 M_{e2}; パターン原文の英語単語数

N_{j1}; 対訳学習文中の単語とパターン原文の単語が一致している日本語単語数

N_{j2}; パターン原文の単語と対訳学習文の単語が一致している日本語単語数

N_{e1}; 対訳学習文中の単語とパターン原文の単語が一致している英語単語数

N_{e2}; パターン原文の単語と対訳学習文の単語が一致している英語単語数

表 4 を用いた場合の類似度の例を，式 9 に示す．

表 4　類似度を求めるときのデータの例

対訳学習文 (日)	この 箱 は 木 から 作ら れる
対訳学習文 (英)	This box is made from wood
パターン (日)	$X1$ は $X2$ $X3$
パターン (英)	$X1$ $X3$ $X2$
パターン原文 (日)	この 箱 は 鉄 から 作ら れる
パターン原文 (英)	This box is made from iron

$$P_S(\frac{\text{この　箱}}{this\ box}) = (5/6) * (5/6) * (6/7) * (6/7) = 0.51 \tag{9}$$

したがって上記の例で P_3 は式 10 となる．

$$\begin{aligned} P_3 &= Dice(\text{この箱},\ thisbox) * P_S(\frac{\text{この　箱}}{this\ box}) \\ &= 0.28 * 0.51 = 0.142 \end{aligned} \tag{10}$$

4 実験

4.1 実験データ

実験には，電子辞書などの例文より抽出した単文コーパス [5] を用いる．データの内訳を表 5 に示す．

表 5　実験データ

対訳学習文	160,000 文
テスト文	1,000 文

4.2 評価方法

本研究では，出力文の翻訳精度の評価として人手評価と自動評価を行う．人手評価として従来手法の” フレーズ確率 P_1” と” 提案手法のフレーズ確率の総積 P_2” と”Dice 係数と類似度の積 P_3” を使用した出力文を比較する．なお，自動評価には BLEU, METEOR，TER，RIBES[6] を用いる．

5 実験結果

5.1 人手評価結果

3 種の確率を用いた出力文 1000 文ずつから，それぞれランダムに抽出した 100 文を用いて，人手による対比較評価を行う．評価の基準を以下に示す．

- ○$_1$：P_1 で対訳句を選択した出力文だけが意味を正しく読み取れる
- ○$_2$：P_2 で対訳句を選択した出力文だけが意味を正しく読み取れる
- ○$_3$：P_3 で対訳句を選択した出力文だけが意味を正しく読み取れる
- ×$_1$：P_1 で対訳句を選択した出力文だけが意味を正しく読み取れない
- ×$_2$：P_2 で対訳句を選択した出力文だけが意味を正しく読み取れない
- ×$_3$：P_3 で対訳句を選択した出力文だけが意味を正しく読み取れない
- □ ：全ての文で意味を読み取れない，ほぼ同一である，または同一の出力

出力文の人手評価結果を表 6 に示す．

表 6　テスト文 100 文の人手評価の結果

○$_1$	○$_2$	○$_3$	×$_1$	×$_2$	×$_3$	□
6	14	3	0	5	10	62

出力文の各例を以下の表 7 から表 12 に示す．

表 7　○$_1$ の例

入力文	日記 を つけ はじめた 。
参照文	I’ve started a journal .
出力文 P_1	I began to wite a diary .
出力文 P_2	The Army began the diary .
出力文 P_3	The Army began the diary .

5.2 自動評価結果

テスト文 1,000 文で翻訳実験を行い，出力文に対して自動評価を行う．表 13 に，それぞれの手法における自動評価の結果を示す．

表 8　○$_2$ の例

入力文	神社 の 境内 で 落ち葉 を 燃やし て いる 。
参照文	They are burning fallen leaves in the grounds of the shrine .
出力文 P_1	The shrine is Inside burned fallen leaves .
出力文 P_2	I have burned the fallen leaves in the shrine Inside the premises .
出力文 P_3	The shrine is Inside burned fallen leaves .

表 9　○$_3$ の例

入力文	わたし は 悔い の ない 生涯 を 送り たい 。
参照文	I’d like to live my life without regrets .
出力文 P_1	I want to see life is not of regret .
出力文 P_2	I I wish regretted life .
出力文 P_3	I want to lead a life no my regret .

表 10　×$_2$ の例

入力文	主人 は 客 に 椅子 を 勧め た 。
参照文	The host waved his guests to a seat .
出力文 P_1	The host offered the guest the seat .
出力文 P_2	The host offered the guest a glass of the seat .
出力文 P_3	The host offered the guest the seat .

表 11　×$_3$ の例

入力文	ハンカチ に 自分 の 名前 を 刺しゅう し た 。
参照文	I embroidered my name on the handkerchief .
出力文 P_1	She embroidered her name on a handkerchief .
出力文 P_2	She embroidered her name on her handkerchief .
出力文 P_3	She embroidered her initials on a rock with a handkerchief .

表 12　全ての文で意味を読み取れない，ほぼ同一である，または同一の出力例

入力文	契約 を 守ら ない 方 が 悪い のだ 。
参照文	The fault is with the one has not observed the contact .
出力文 P_1	not signed preserve of a bad person .
出力文 P_2	The riddance jealously lease of the contact .
出力文 P_3	is toword the bad keep their contact .

入力文	ここから 港 が よく 見える 。
参照文	We can see the whole harbor from here .
出力文 P_1	I see harbor well from here .
出力文 P_2	I see harbor well from here .
出力文 P_3	Here you can see the harbor .

入力文	彼ら は 逃げ腰 に なっ た 。
参照文	They were ready to flee .
出力文 P_1	They became 逃げ腰 .
出力文 P_2	They became 逃げ腰 .
出力文 P_3	They became 逃げ腰 .

表 13　テスト文 1000 文の自動評価の結果

パラメータ	BLEU	METEOR	TER	RIBES
フレーズ確率 (P_1)	0.1495	0.4175	0.6467	0.7309
フレーズ確率の総積 (P_2)	**0.1552**	**0.4308**	**0.6388**	**0.7354**
Dice 係数と類似度の積 (P_3)	0.1459	0.3955	0.6762	0.7139

5.3　実験結果のまとめ

表 6 と表 13 より” フレーズ確率の総積” を使用したときの結果が最も良く，”Dice 係数と類似度の積” を使用した結果が最も悪いことが分かった．しかし大きな差はないことが分かった．なお未知語が出現する文では 3 種どの確率を使用した文であっても未知語として出力されていた．

6　考察

6.1　フレーズ確率の総積の有効性

表 6 より，翻訳精度として” フレーズ確率の総積” が最も優れていることが分かった．これは非線形推定であるフレーズ確率の文の和を取ることによって値が安定したからであると考える．

6.2　差なしの原因

しかし” フレーズ確率の総積 P_2” を使用した翻訳と従来手法の” フレーズ確率 P_1” を使用した翻訳には大きな差がなかった．これは 2 つの原因が挙げられる．まず確率の変更をした場所がデコーダー部分であったこと，次に翻訳確率において言語モデルが大きな割合を占めていることである．

なお，現在対訳句から句に基づく文パターンを作成する際に対訳句確率を使用して枝刈りを行っている．よって句に基づく文パターンを作成する際の確率を変更すると本研究よりも差が出るのではないかと考える．

6.3　誤り分析

$\times_2$ の 5 文について解析すると，単語不足の文が 1 文，動詞連続文が 1 文，単語が過剰な文が 1 文，主語目的語が逆転している文が 1 文，同音異義語を選択している文が 1 文となった．単語不足の文と単語が過剰な文と同音異義語を選択している文の合計 3 文が出力されたのは誤った対訳句を選択してしまったことが原因だと考える．動詞連続文と主語目的語が逆転している文の合計 2 文が出力された原因は誤ったパターンを選択してしまったことが原因だと考える．翻訳精度の向上において対訳句選択とパターン選択の個別の問題対策が今後の課題と言える．

6.4　他の対訳句確率の計算方法

対訳句確率は他に式 11 や式 12 で計算可能である．今後比較実験を行っていきたい．

$$P(\frac{J_0 \cdots J_{N-1}}{E_0 \cdots E_{M-1}}) = \prod_{m=0}^{M-1} \arg\max_{0 \le m \le N-1}(p(E_m|J_n) * \prod_{n=0}^{N-1} \arg\max_{0 \le n \le M-1} p(J_n|E_m)) \quad (11)$$

$$P_1(\frac{J_0 \cdots J_{N-1}}{E_0 \cdots E_{M-1}}) = \prod_{m=0}^{M-1}\prod_{n=0}^{N-1}(p(E_m|J_n) * p(J_n|E_m)) \quad (12)$$

7　おわりに

本研究では，対訳句の新たな確率計算方法である” フレーズ確率の総積” と”Dice 係数と類似度の積” の方法を提案した．従来の” フレーズ確率” を加えた 3 種を使用した翻訳の結果，” フレーズ確率の総積” が最も良い翻訳精度を得ることができた．しかし従来の” フレーズ確率” と大きな差はなかった．今後は，さらなる翻訳精度の向上の手法を検討したい．

参考文献

[1] 江木孝史 “句に基づく文パターンを用いた英日翻訳 ”，2014 年修論

[2] 村上仁一 “パターンに基づく統計機械翻訳の概要と問題点について”，電子情報通信学会技術研究報告， 言語理解とコミュニケーション，NLC2017-3， pp.13-18， 2017.

[3] 江木孝史 “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， A6-2， pp.951-954， 2014.

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

[5] 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.

[6] Hideki Isozaki，“Automatic Evaluation of Translation Quality for Distant Language Pairs”， Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing，pp.944-952. 2010.