ロジスティック回帰分析を用いたパターンに基づく統計翻訳

春野瑞季 村上仁一 徳久雅人
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} @ ike.tottori-u.ac.jp

1 はじめに

パターン翻訳は，対訳文パターンと対訳句を用いて翻訳を行う [1]．この翻訳方式は入力文が適切な文パターンに適合した場合，翻訳精度の高い文を出力する傾向にある．しかし，対訳文パターンと対訳句は人手で作成するため，開発にコストがかかる [2]．

そこで江木らは，対訳文パターンと対訳句を統計的手法で自動的に作成し翻訳する方法を提案した (以下，従来手法)[3]．これを “パターンに基づく統計翻訳” と呼ぶ．しかし翻訳結果を調査したところ，不適切な対応をとる対訳句が翻訳文に含まれていた．

不適切な対応をとる対訳句が翻訳文に含まれることを抑制するため，2 つの方法が考えられる．一つは対訳句の作成において，不適切な対応をとる対訳句の作成を抑制する方法である．もう一つは翻訳文の選択において，不適切な対応をとる対訳句を含む翻訳文の出力を抑制する方法である．本研究では後者の方法をとる．さらに，人手で作成した対訳句を利用して回帰分析を行う．具体的には，人手で作成した対訳句と自動作成した対訳句を比較し，ロジスティック回帰分析を用いて，複数の確率から対訳句に確率を付与する．そして翻訳文の選択において，ロジスティック回帰分析から得た確率を用いてパターンに基づく日英統計翻訳を行い，翻訳精度の調査を行う．

2 パターンに基づく統計翻訳 (従来手法)[3]

パターンに基づく統計翻訳は，大きく 5 つのステップで翻訳を行う．パターンに基づく日英統計翻訳の概要を以下に示す．

2.1 対訳単語の作成

GIZA++[4] を用いて，対訳学習文の単語アライメントを取り，対訳単語と単語翻訳確率を得る．

2.2 単語に基づく対訳文パターンの作成

対訳単語と対訳学習文を用いて，単語に基づく対訳文パターンを作成する．まず，対訳単語と対訳学習文を照合する．そして対訳学習文において，適合した対訳単語を変数化する．

なお変数の組み合わせを考慮して，単語に基づく対訳文パターンは可能な限り多く作成する．具体的には，対訳学習文 1 文に対し n 個の対訳単語が変数化できる場合，対訳単語を “変数化する・変数化しない” の 2 通りの組み合わせがあるため，全ての組み合わせを考慮し，単語に基づく対訳文パターンを 2^n 通り作成する．

2.3 対訳句の作成

a) 対訳句の抽出

単語に基づく対訳文パターンと対訳学習文を用いて，対訳句を抽出する．まず，単語に基づく対訳文パターンと対訳学習文を照合する．ただし変数に複数単語が適合することを許す．そして適合した場合，単語に基づく対訳文パターンの変数部に対応する複数単語および単語を，対訳句として対訳学習文より抽出する．対訳句抽出の例を図 1 に示す．

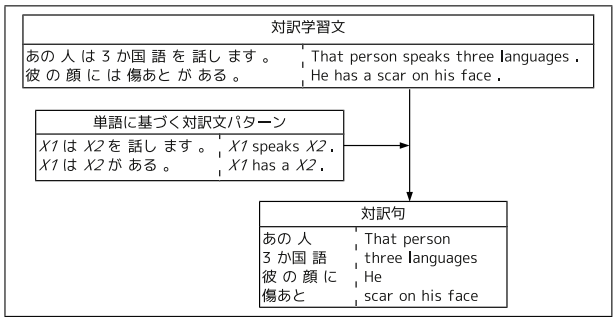

図 1 対訳句抽出の例

この方法では対訳句の抽出において，不適切な対応をとる対訳句を抽出する問題がある．図 1 の例では “彼の顔に” と “He” のような不適切な対応をとる対訳句がある．

b) 対数フレーズ確率の付与

対訳単語と単語翻訳確率を用いて，対訳句に確率を付与する．まず対訳句において，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．次に日本語単語に対応する英語単語の中で，単語翻訳確率の最大値を得る．これを各日本語単語に対して行い，得られた値について対数の総和を求める．本研究ではこの値を日英方向の対数フレーズ確率と呼ぶ．同様に対訳句において，英語単語に対応する日本語単語の中で，単語翻訳確率の最大値を取得し，英日方向の対数フレーズ確率も求める．日英方向の対数フレーズ確率付与の例を図 2 に示す．

2.4 句に基づく対訳文パターンの作成

対訳句と対訳学習文を用いて，句に基づく対訳文パターンを作成する．まず単語に基づく対訳文パターンの作成 (2.2 節) と同様に，変数の組み合わせを考慮して，句に基づく対訳文パターンを可能な限り多く作成する．

次に，対訳単語と単語翻訳確率を用いて，句に基づく対訳文パターンに確率を付与する．句に基づく対訳文

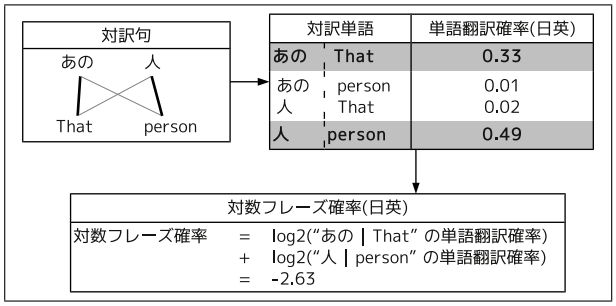

図 2 対数フレーズ確率付与の例 (日英)

パターンにおいて変数化していない部分 (以下，字面) を用いて，対数フレーズ確率の付与 (2.3 節 b) と同様の計算手法で確率を求める．本研究ではこの値を対数文パターン確率と呼ぶ．

2.5 文生成

句に基づく対訳文パターンと対訳句を用いて，文生成を行う．まず，日本語文パターンと入力文を照合し，入力文に適合する日本語文パターンを選択する．そして，日本語文パターンに対応する英語文パターンと対数文パターン確率を取得する．なお，翻訳精度向上のため，入力文と日本語文パターンの字面が多く一致する文パターンを優先して選択する．次に英語文パターンの変数部に適合する英語句と対数フレーズ確率を得る．最後に，英語文パターンの変数部を英語句に置き換え，翻訳候補文を生成する．

各翻訳候補文の対数文パターン確率と対数フレーズ確率，言語翻訳確率 (trigram) の総和を求め，翻訳候補文の中で総和が最大となる文を翻訳文として出力する．日英翻訳における文生成の例を図 3 に示す．

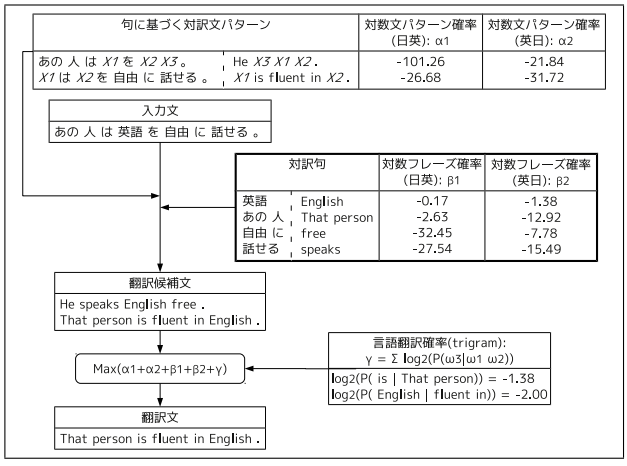

図 3 日英翻訳における文生成の例

3 提案手法

従来手法では，不適切な対応をとる対訳句が翻訳文に含まれていた．そこで，翻訳文の選択において不適切な対応をとる対訳句を含む翻訳文の出力を抑制するため，人手で作成した対訳句を利用して回帰分析を行い，自動作成した対訳句に確率を付与する．具体的には，人手で作成した対訳句と自動作成した対訳句を比較し，ロジスティック回帰分析を用いて，複数の確率から対訳句に確率を付与する．そして，従来手法における対数フレーズ確率をロジスティック回帰分析から得た確率に置き換え，パターンに基づく日英統計翻訳を行う．

3.1 ロジスティック回帰分析

ロジスティック回帰分析は，1 つあるいは複数の独立変数から，1 つの質的な従属変数 (0 または 1) を予測するための手法である．ロジスティック回帰分析を用いて，独立変数の条件下で従属変数が発生する条件付き確率を予測することができる．

独立変数 $X = \{x_1, x_2, ..., x_n\}$ の条件下で，ある事象 A が発生する条件付き確率 $P(A|X)$ は式 (1) で求められる．

$$P(A|X) = \frac{exp(a + b_1x_1 + b_2x_2 + ... + b_nx_n)}{1 + exp(a + b_1x_1 + b_2x_2 + ... + b_nx_n)} \quad (1)$$

ここで，a は定数，b_n は独立変数 x_n の回帰係数である．本研究では統計ソフト R[5] を用いて，ロジスティック回帰分析を行う．

3.2 モデルの作成

人手で作成した対訳句と自動作成した対訳句の比較により従属変数を設定する．具体的には，人手で作成した対訳句と 2.3 節 a で抽出した対訳句を照合し，一致する対訳句を 1，一致しない対訳句を 0 とし，従属変数に用いる．独立変数には 6 つの確率を用いる．

- 日英方向の対数フレーズ確率: x_1
- 英日方向の対数フレーズ確率: x_2
- 対訳学習文における日英方向の対数翻訳確率: x_3
- 対訳学習文における英日方向の対数翻訳確率: x_4
- 対訳句抽出における日英方向の対数翻訳確率: x_5
- 対訳句抽出における英日方向の対数翻訳確率: x_6

a) 対数フレーズ確率 x_1,x_2

従来手法で用いる対数フレーズ確率である．詳細は 2.3 節 b を参照のこと．

b) 対訳学習文における対数翻訳確率 x_3,x_4

対訳学習文において句 f を句 e に翻訳する確率である．対訳学習文における原言語 F の句 f から目的言語 E の句 e への翻訳確率 $P_s(e|f)$ は式 (2) で求められる．

$$P_s(e|f) = \frac{count_s(f, e)}{count_s(f)} \quad (2)$$

ここで，$count_s(f, e)$ は対訳学習文における対訳句 (f, e) の共起回数であり，$count_s(f)$ は原言語 F の学習文における句 f の出現回数である．確率 $P_s(e|f)$ の対数をとり，対訳学習文における句 f から句 e への対数翻訳確率とする．同様に，対訳学習文における句 e から句 f への対数翻訳確率も求める．

c) 対訳句抽出における対数翻訳確率　x_5,x_6

対訳句の抽出 (2.3 節 a) において，対訳句は網羅的に抽出するため，多くの場合，同一の対訳句が複数抽出される．対訳句抽出における対数翻訳確率は対訳句を重複して抽出した回数を利用した確率である．なお，文生成に用いる対訳句は重複する対訳句がないように作成する．対訳句抽出における句 f から句 e への翻訳確率 $P_v(e|f)$ は式 (3) で求められる．

$$P_v(e|f) = \frac{count_v(f,e)}{count_v(f)} \tag{3}$$

ここで，$count_v(f,e)$ は対訳句 (f,e) を抽出した回数であり，$count_v(f)$ は句 f を抽出した回数である．確率 $P_v(e|f)$ の対数をとり，対訳句抽出における句 f から句 e への対数翻訳確率とする．同様に，対訳句抽出における句 e から句 f への対数翻訳確率も求める．

3.3 確率の付与

従来手法は対訳句の確率として対数フレーズ確率を用いている．一方，本研究では対訳句の確率としてロジスティック回帰分析から得た確率を用いる．なお，ロジスティック回帰分析から得た確率についても対数をとる．ロジスティック回帰分析から得た確率の例を表 1 に示す．

表 1　ロジスティック回帰分析から得た確率の例

日本語句	英語句	ロジスティック回帰分析から得た確率 (対数)
英語	English	-0.116
あの 人	That person	-0.129
自由 に	free	-0.362
話せる	speaks	-0.136

4　実験

4.1 実験データ

対訳学習文および翻訳実験に用いるテスト文は電子辞書から抽出した単文データベースを用いる [6]．なお，単文データは日本語文が単文であるが，英語文は単文とは限らず，重文・複文が含まれる．コーパスの内訳を表 2 に示す．

表 2　コーパスの内訳

対訳学習文	100,000 文対
テスト文	100 文

人手で作成した対訳句には鳥バンク [7] の対訳句を用いる．鳥バンクは自然言語処理のための言語知識ベースを収録したデータバンクであり，日本語の重文と複文を対象とする “意味類型パターン辞書” が収録されている．本研究では鳥バンクから抽出した対訳句 329,545 句対を用いる．

4.2 分析実験

6 つの確率を独立変数に用い，ロジスティック回帰分析を行う．予備実験の結果より，句の抽出回数または対訳学習文における句の出現回数が少なくとも一方の言語で 1 回である対訳句を取り除き，モデルの作成に用いる．

鳥バンクの対訳句と自動作成した対訳句を照合した結果，一致した対訳句は 45,161 句対であった．また，鳥バンクの対訳句と一致しなかった対訳句からも一致した対訳句と同数の対訳句を無作為抽出し，合計 90,322 句対の対訳句をモデルの作成に用いる．ロジスティック回帰分析により求めたロジスティック回帰モデルを式 (4) に示す．

$$\begin{aligned} logit(P) = {} & 3.33685 + 0.01639x_1 + 0.00682x_2 \\ & + 0.12354x_3 - 0.23028x_4 - 0.01709x_5 \\ & + 0.22118x_6 \end{aligned} \tag{4}$$

4.3 翻訳実験

本研究は日英翻訳を行う．従来手法と同様に，入力文と日本語文パターンの字面を比較し，字面が多く一致する対訳文パターンを優先して選択する．

各翻訳候補文において，ロジスティック回帰分析から得た確率と対数文パターン確率，言語翻訳確率の総和を求め，翻訳候補文の中で総和が最大となる文を翻訳文として出力する．翻訳実験より，入力文 100 文から翻訳文 78 文を得た．

5　評価

人手による対比較評価を行った．ベースラインには従来手法 (2 節) を用いる．従来手法は対数文パターン確率と対数フレーズ確率，言語翻訳確率を用いて，翻訳候補文から翻訳文を選択する．なお，提案手法で出力した翻訳文 78 文中，従来手法でも翻訳文が得られた 77 文に対して評価を行った．提案手法と従来手法との対比較評価結果を表 3 に示す．

表 3　提案手法と従来手法の対比較評価結果

提案手法○	提案手法×	差なし	同一出力
7	3	59	8

提案手法○の例を表 4 に，提案手法×の例を表 5 に示す．

表 4　提案手法○の例

入力文	もっと 右 へ 寄っ て ください 。
参照文	Please move over more to the right .
従来手法	Please come to discuss .
提案手法	Move a little more to the right .

表 5　提案手法×の例

入力文	金槌 と 同様 な もの が ある 。
参照文	There is something similar to a hammer .
従来手法	There is something like a hammer .
提案手法	There’s something a view similar to a hammer .

表 3 の結果より，従来手法と比較して，提案手法が優れていることがわかる．よって，提案手法の有効性が確認された．

6　考察

6.1 翻訳実験の考察

対比較評価より，提案手法×の翻訳文を 3 文得た．この 3 文について，誤り解析を行った．以下に原因を述べる．

6.1.1 不適切な対応をとる対訳句

提案手法により対訳句の対応が改善され，翻訳精度が向上した．しかし依然として，不適切な対応をとる対訳句を含む翻訳文が残る．不適切な対応をとる対訳句を含む翻訳文の例を表 6 に示す．

表 6 不適切な対応をとる対訳句を含む翻訳文の例

入力文	彼 は たった １０００ 円 しか 持っ て い ない 。
参照文	He has only 1,000 yen on him .
翻訳文	I have no more than one thousand yen .
日本語文パターン	$X1$ は たった $X2$ 円 しか 持っ て い ない 。
英語文パターン	$X1$ have no more than one $X2$ yen .
対訳句	$X1$: 彼 I $X2$: １０００ thousand

表 6 において，翻訳文で用いた対訳句の中に不適切な対応をとる対訳句 “彼” と “I” がある．この入力文の翻訳において，翻訳候補文を調査したところ，“彼” と “He” の対訳句を含んだ翻訳候補文 “He has no more than one thousand yen .” が存在した．以上のことから，不適切な対応をとる対訳句を含む翻訳文を出力する問題は，翻訳文の選択方法の再検討により改善すると考える．

6.1.2 指示詞などの字面が残る対訳文パターン

対訳文パターンにおいて，指示詞や代名詞などの字面が残るため正しく翻訳できない問題がある．例を表 7 に示す．

表 7 指示詞などの字面が残る対訳文パターンを含む翻訳文の例

入力文	その ビル は 倒壊 の 危険 が ある 。
参照文	The building is in danger of collapsing .
翻訳文	The building is a danger of this down .
日本語文パターン	その ビル は $X1$ の $X2$ ある 。
英語文パターン	The building is $X2$ of this $X1$.
対訳句	$X1$: 倒壊 down $X2$: 危険 が a danger

表 7 の英語文パターンには指示詞 “this” が残る．対訳文パターンに指示詞などの字面が残る場合，適切な対応をとる対訳句を変数部に用いても正しく翻訳できない．以上のことから，指示詞などの字面が残る対訳文パターンの問題に対して，対訳文パターンの選択方法の再検討が必要であると考える．

6.1.3 重みの最適化

翻訳文の選択において，ロジスティック回帰分析から得た確率と対数文パターン確率，言語翻訳確率の 3 つの確率を利用した．そして，ヒューリスティックにより各確率に重みを付与している．これらの重みを最適化することでさらなる翻訳精度の向上が期待できる．なお，従来手法においても同様の効果が期待できる．

6.2 分析実験の考察

分析実験 (4.2 節) において予備実験を行った結果，不適切な対応をとる対訳句において高い確率が付与される場合があることがわかった．例を表 8 に示す．

表 8 日本語学習文における日本語句の出現回数が 1 回の例

日本語句 J	英語句 E	$C_s(J)$	$C_s(E\|J)$	$\log P_s(J\|E)$
部員 たち	members	1	1	0
部員 たち	were	1	1	0
部員 たち	The	1	1	0

ここで，$C_s(J)$ は日本語学習文における日本語句 J の出現回数であり，$C_s(E|J)$ は対訳学習文における対訳句 (J,E) の共起回数，$\log P_s(E|J)$ は対訳学習文における日英方向の対数翻訳確率である．例えば日本語句 “部員たち” の出現回数が 1 回の場合，その日本語句をもつすべての対訳句において，対訳学習文における日英方向の対数翻訳確率は 0(最大値) になる．したがって句の出現回数が 1 回の場合，不適切な対応をとる対訳句においても高い確率が付与される．以上のことから，本研究では句の抽出回数または対訳学習文における句の出現回数が少なくとも一方の言語で 1 回である対訳句を取り除き，モデルの作成を行った．

また本研究では回帰分析の従属変数として質的な値 (0 または 1) を設定したため，ロジスティック回帰分析による分析実験を行った．しかし，従属変数として量的な値 (例えば対訳句の出現回数) を設定することで重回帰分析など他の回帰分析手法による分析実験を行うことができる．今後，分析手法や分析に用いる変数の設定を再検討し，翻訳実験に最適な手法を探る．

7 まとめ

本研究では，“パターンに基づく統計翻訳” の翻訳文の選択において，不適切な対応をとる対訳句を含む翻訳文の出力を抑制するため，人手で作成した対訳句を利用してロジスティック回帰分析を行うことを提案した．そして，パターンに基づく日英統計翻訳を行い，翻訳精度を調査した．従来手法との対比較評価の結果は翻訳文 77 文において提案手法○が 7 文，提案手法×が 3 文であった．以上より，提案手法の有効性が確認できた．

しかし，機械翻訳システムには Moses[8] をはじめとする種々の翻訳システムがある．今後，これらの翻訳システムとの比較も行う予定である．

また誤り解析の結果，対訳文パターンの選択方法や翻訳文の選択方法，翻訳文選択に用いる重みについて再検討が必要であることがわかった．さらに，不適切な対応をとる対訳句が翻訳文に含まれることを抑制するための方法として，対訳句の作成において，不適切な対応をとる対訳句の作成を抑制する方法も考えられる．今後これらを検討し，さらなる翻訳精度向上を目指す．

参考文献

[1] 渡辺日出雄, 武田浩一, “パターンベース翻訳システム：PalmTree”, 情報処理学会第 55 回全国大会講演論文集, pp.80-81, 1997.

[2] 池原 悟, 宮崎 正弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良彦, “日本語語彙大系”, 岩波書店, 1997.

[3] 江木孝史, 村上仁一, 徳久雅人, “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”, 自然言語処理学会第 20 回年次大会予稿集, pp.951-954, 2014.

[4] Franz Josef Och, Hermann Ney, “A Systematic Comparison of Various Statistical Alignment Models”, Computational Linguistics, 29(1), pp.19-51, 2003.

[5] Ross Ihaka, Robert Gentleman, “R: A Language for Data Analysis and Graphics”, Journal of Computational and Graphical Statistics, 5(3), pp.299-314, 1996.

[6] 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予稿集, pp.119-130, 2012.

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

[8] Philipp Koehn, Marcello Federico, Brooke Cowan, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, Evan Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation”, Proceedings of the ACL 2007 Demo and Poster Sessions, pp.177-180, 2007.