パターンに基づく統計翻訳における，文パターン確率の考察

西尾聡一郎 *1 村上仁一 *2 徳久雅人 *2
*1 鳥取大学 工学部 知能情報工学科
*2 鳥取大学 工学部研究科
*1,*2{s122036, murakami, tokuhisa} @ ike.tottori-u.ac.jp

1 はじめに

パターンに基づく統計翻訳 [1](以下，Pattern Based SMT) は, 統計的手法を用いて, 対訳句と文パターン辞書を自動作成して翻訳を行う手法である. この翻訳手法は，対訳フレーズ対数確率，対訳文パターン対数確率，単語の N-gram を使用する．そして，対訳文パターン対数確率は，パターン内の単語における GIZA++[2] の値を利用して得ている．しかし，パターン内の単語における GIZA++ の値の中には，明らかに異常値を取っているものがある．そのため，対訳文パターン対数確率の値に信用性がないと考える.

しかし，一方で対訳フレーズ対数確率は，一般に GIZA++ において高い値が選ばれる傾向にある．そして，高い値は信頼性が高いと考える．そこで，本研究では，対訳文パターン対数確率の計算に，対訳フレーズ対数確率を利用して，翻訳精度の向上を試みる.

2 Pattern Based SMT

江木ら [1] によって提案された Pattern Based SMT は大きく分けて 5 つのステップで翻訳を行う．翻訳の手順を以下に示す.

2.1 対訳単語辞書の作成

対訳文と GIZA++ を用いて，対訳単語辞書を作成する.

2.2 単語に基づく対訳文パターンの作成

対訳文と対訳単語辞書を用いて，単語に基づく対訳文パターンを作成する.

2.3 対訳フレーズ辞書の作成

対訳フレーズ辞書は，対訳フレーズと対訳フレーズ対数確率で構成される．以下に，この作成方法を述べる.

2.3.1 対訳フレーズ

対訳文と単語に基づく対訳文パターンを照合し，対訳フレーズを抽出する．対訳フレーズの作成の流れを図 1 に示す.

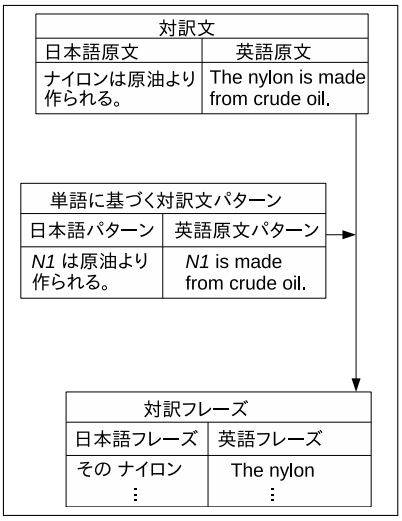

図 1 対訳フレーズの作成の流れ

2.3.2 対訳フレーズ対数確率

抽出した対訳フレーズに GIZA++ の値を用いて，対訳フレーズ対数確率を付与する．対訳フレーズ対数確率は，以下の方法で行う.

$$P(J_0 \cdots J_{N-1}|E_0 \cdots E_{M-1}) =$$

$$\sum_{n=0}^{N-1} \arg\max_{m=0}^{M-1}(log_2(p(E_m|J_n)) + log_2(p(J_n|E_m))) \quad (1)$$

J_n; 対訳フレーズ中の日本語の単語

E_m; 対訳フレーズ中の英語の単語

$p()$;GIZA++ の値

N; 日本語の単語数

M; 英語の単語数

対訳フレーズ対数確率の計算例を図 2 に示す.

2.4 句に基づく対訳文パターン辞書の作成

句に基づく対訳文パターン辞書は，対訳パターンと対訳文パターン対数確率で構成されている．以下に，この作成方法を示す.

2.4.1 句に基づく対訳文パターン

句に基づく対訳文パターンの作成の手順を以下に示す.

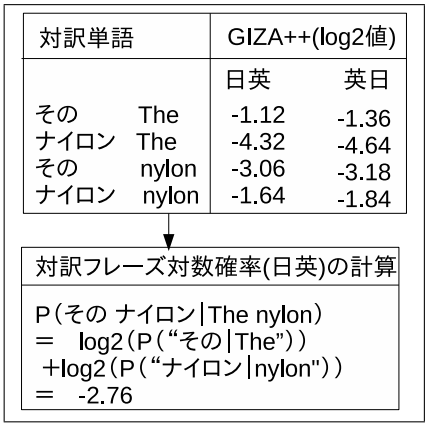

図 2 対訳フレーズ対数確率の付与の例

1. 対訳文と対訳フレーズの照合を行う．
2. 対訳フレーズが適合した対訳文のフレーズを変数化して句に基づく対訳文パターンを作成する．

句に基づく対訳文パターンの作成方法と例を図 3 に示す．

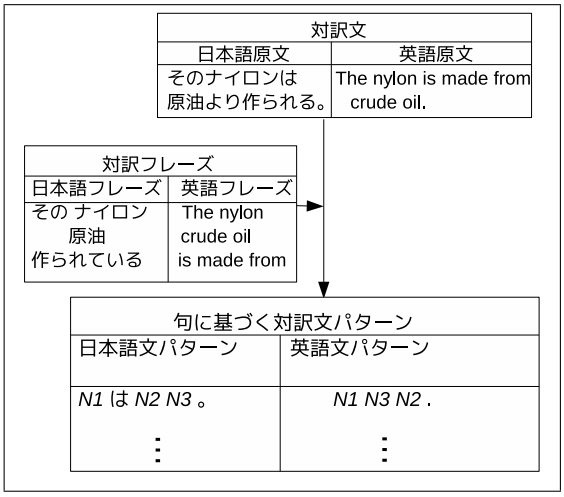

図 3 句に基づく対訳文パターンの作成の流れ

2.4.2 対訳文パターン対数確率

対訳文パターン対数確率は，GIZA++ を用いて，以下の式で行う．

$$P(J_0 \cdots J_{N-1}|E_0 \cdots E_{M-1}) = \sum_{n=0}^{N-1} \arg\max_{m=0}^{M-1}(log_2(p(E_m|J_n)) + log_2(p(J_n|E_m))) \quad (2)$$

J_n; 対訳パターン中の日本語の単語

E_m; 対訳パターン中の英語の単語

$p()$; $GIZA++$ の値

N; 日本語の単語数

M; 英語の単語数

対訳文パターン対数確率の付与を図 4 に示す．

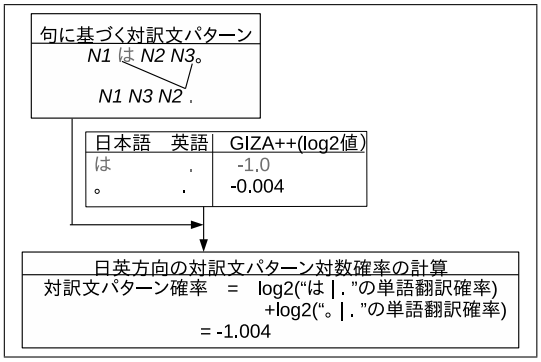

図 4 対訳文パターン対数確率の付与

2.5 出力文の生成

対訳文パターン対数確率と出力候補文の作成に用いた対訳フレーズ対数確率と単語の N-gram を用いて，出力候補文の翻訳対数確率を計算する．そして，作成した出力候補文から出力文を選択する．

3 従来手法の問題点

従来手法の対訳文パターン対数確率は，パターン内の単語における GIZA++ の値を利用している．しかし，パターン内の単語における GIZA++ の値の中には，明らかに異常値を取っているものがある．そのため，対訳文パターン対数確率の値に信用性がないと考える．

3.1 従来手法における例

図 4 において，P（は | .）の GIZA++ の値は異常値である．対訳文パターン確率の計算は，GIZA++ の値を利用しているため，GIZA++ に異常値がみられると翻訳結果に信頼が得られないと考える．

4 提案手法

GIZA++ の値が低い場合は，信頼がおけないと考えられる．一方で対訳フレーズ対数確率は GIZA++ において高い値が選ばれる傾向にある．そこで，本研究では，対訳フレーズ対数確率を利用する．具体的には，1 文パターン全体における対訳フレーズ対数確率の対数の総和を求め，その対訳パターンの翻訳確率とする．

4.1 句に基づく文パターン辞書の作成

対訳フレーズ対数確率の計算方法を示す．

$$P(JX_0 \cdots JX_{N-1}|EX_0 \cdots EX_{N-1}) = \sum_{n=0}^{N-1}(log_2(p(JX_n|EX_n))) \quad (3)$$

JX_n; 対訳パターン中の日本語の対訳フレーズ

EX_n; 対訳パターン中の英語の対訳フレーズ

$log_2p()$; 対訳フレーズ対数確率

N; 対訳フレーズの数

4.2 提案手法の具体例

図 5 は，提案手法の計算例を示す．なお，対訳文パターンは，2.4 節の表 3 と同じである．

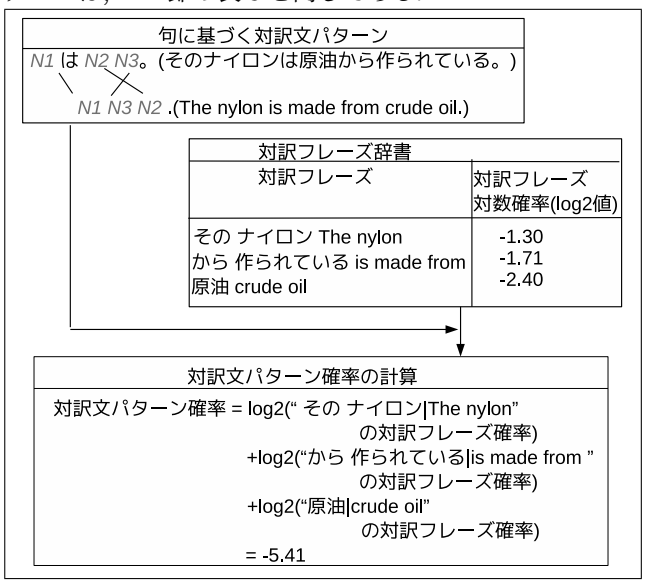

図 5 提案手法の具体例

5 実験環境

5.1 日英対訳文

本研究では，対訳文として，単文コーパス [4] を用いる．統計翻訳の前処理として，日本語文に対して，MeCab[5] を用いて形態素解析を行う．また，英語文に対して tokenizer.sed [6] を用いて分かち書きを行う．前処理を行った単文コーパスの例を表 1 に示す．

表 1 単文コーパスの例

私 は 映画 を 見 に 行く 。
I go to see a movie .

5.2 実験データ

本研究では，対訳文として単文コーパスを表 2 の内訳で用いる．

表 2 日英対訳文数

対訳文	100,000 文対
テストデータ	100 文

6 実験

本研究では，従来手法と提案手法より 100 文ずつ翻訳結果を出力し，対比較実験を行う．

7 実験結果

7.1 対比較実験

提案手法と従来手法で得られた翻訳結果 100 文に対して，対比較実験を行う．表 3 に提案手法と従来手法の対比較実験の結果を示す．表 3 中の表記方法について説明する．

- “提案手法〇”：提案手法の翻訳精度が従来手法の翻訳精度より優れている文
- “従来手法〇”：従来手法の翻訳精度が提案手法の翻訳精度より優れている文
- “差なし”：２種類の翻訳精度が同程度である文
- “同一出力”：２種類の翻訳文が完全に同一な文

表 3 100 文での対比較実験

提案手法〇	従来手法〇	同一出力	差なし
10	8	62	20

7.1.1 従来手法が提案手法より優れている例

表 3 における従来手法が提案手法より優れている例を表 4 に示す．

表 4 従来手法が提案手法より優れている例

日本語入力文	私 は 都会 に 出 た 。
正解文	I arrived in the city .
英語翻訳文 (提案)	I was in the city .
日本語文パターン (提案)	*N00* は 01 に 出 た 。
英語文パターン (提案)	*N00* was in the *N01* .
英語翻訳文 (従来)	I went to the city .
日本語文パターン (従来)	私 は *N00* に *N01* た 。
英語文パターン (従来)	I *N01* to the *N00* .

7.1.2 提案手法が従来手法より優れている例

表 3 における提案手法が従来手法より優れている例を表 5 に示す．

表 5 提案手法が従来手法より優れている例

日本語入力文	組合 は ストライキ に 参加 する 。
正解文	The union will join in the strike .
英語翻訳文 (提案)	The union is joined to strike .
日本語文パターン (提案)	*N00* は *N01* に *N02* する 。
英語文パターン (提案)	*N00* is *N02* to *N01* .
英語翻訳文 (従来)	Each part of the strike .
日本語文パターン (従来)	*N02* は ストライキ に *N01 N00* 。
英語文パターン (従来)	*N02 N01 N00* strike .

7.2 実験結果のまとめ

表 3 の結果より，提案手法は従来手法の計算と同等な精度が得られた．

8 考察

8.1 moses と提案手法における対比較実験

本研究では，対訳文パターン対数確率の計算に対訳フレーズ対数確率を利用することで，従来手法と同等の精度が得られることがわかった．ここで，moses[3] と提案手法における 100 文の対比較実験の結果を示す[*1]．

8.1.1 moses での対比較実験

表 6 に提案手法と moses で得られた翻訳結果 100 文での対比較実験の結果を示す．

表 6 moses での対比較実験

提案手法○	moses ○	同一出力	差なし
18	2	0	80

8.1.2 moses が提案手法より優れている例

表 6 における moses が提案手法より優れている例を表 7 に示す．

表 7 moses が提案手法より優れている例の例

日本語入力文	関税 は 完全 に 撤廃 さ れ た 。
正解文	Tariffs have been eliminated altogether .
英語翻訳文 (提案)	関税 It was in full .
英語翻訳文 (moses)	Tariffs are entirely to be rescinded .

8.1.3 提案手法が moses より優れている例

表 6 における提案手法が moses より優れている例を表 8 に示す．

表 8 提案手法が moses より優れている例

日本語入力文	彼 は 健康 を 損なっ た 。
正解文	He ruined his health .
英語翻訳文 (提案)	He ruined his health .
英語翻訳文 (moses)	His health 損なっ .

表 3 の結果より，提案手法が moses より優れているという結果になった．

8.2 対訳文パターンの新しい計算方法

本研究では，対訳文パターン対数確率の計算に対訳フレーズ対数確率を利用し，翻訳精度の調査を行った．対比較実験の結果から，提案手法の計算方法は，従来手法の計算方法と同等の精度が得られた．しかし，翻訳精度の向上は見られなかった．その理由として，以上の原因が挙げられる．日本語パターンを J_1，JX_1，J_2，英語パターンを E_1，EX_1，E_2 とした場合，従来手法は，以下の式で行う．

$$P(J_1JX_1J_2)/(E_1EX_1E_2)$$
$$= \arg\max(P(J_1/E_1), P(J_1/E_2)) \times \arg\max(P(J_2/E_1), P(J_2/E_2))$$

つまり，$P(J_1/E_1EX_1E_2)$ の値として $P(J_1/E_1)$，$P(J_1/E_2)$ の最大値を選択している．しかし，$P(J_1/E_1)$，$P(J_1/E_2)$ が大きな値を持つ場合がある．この場合，加算した方が妥当性があると思われる．つまり，以下の式を用いる．

$$(P(J_1/E_1) + P(J_1/E_2)) \times (P(J_2/E_1) + P(J_2/E_2))$$

この方法を試みたい．

9 おわりに

本研究では，対訳文パターン対数確率の計算に変数部を利用し，翻訳精度の調査を行った．実験結果より，提案手法は，従来手法と同等な精度が得られた．提案手法と moses での対比較実験の結果より，提案手法の方が優れていた，以上より，提案手法は従来手法と同等の精度の計算方法であると言える．今後は，新しい計算方法を試みたい．

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

[3] Moses: “Open Source Toolkit for Statistical Machine Translation”, Proceedings of the ACL 2007 Demo and Poster Sessions, pp.177-180. 2007.

[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] MeCab http://mecab.sourceforge.net/

[6] tokenizer.sed http://www.cis.upenn.edu/treebank/tokenizer.sed

[*1] ただし，実験条件が提案手法と完全に同一ではない