一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
TL2017-3,NLC2017-3 (2017-06)

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

村上　仁一†

† 鳥取大学 工学部　〒 680–8552 鳥取県鳥取市湖山町南 4-101
E-mail: †murakami@eecs.tottori-u.ac.jp

あらまし　パターン翻訳は，古くからある翻訳方法である．対訳文パターンと対訳句を利用して，元言語を目的言語に翻訳する．この方式は，元言語に適切な対訳パターンが照合できたとき，高い翻訳品質が得られる．しかし，対訳文パターンと対訳句を人手で作成するため，システムを構築するにはコストが高い．そこでこのコストを削減するため，統計翻訳で使われている方法を用いて，対訳パターンと対訳句を自動的に作成する方法を提案した [7]．本論文では，この提案したパターンに基づく統計機械翻訳の概要と，翻訳精度について述べる．また，句に基づく統計機械翻訳との比較や，提案方法の問題点について述べる．

キーワード　統計翻訳，パターン翻訳，IBM Model1 , パターンに基づく統計機械翻訳, 対訳句，パターン

Introduction of Proposed Pattern Based Statitic Machine Translation and Problem

Jin'ichi MURAKAMI†

† Faculty of Engineering, Tottori University　Koyama chou 4-101, Tottori city, Tottori, 680–8552 Japan
E-mail: †murakami@eecs.tottori-u.ac.jp

Abstract　Pattern-based machine translation needs translation pattern pairs and phrase pairs, which are usually made manually. A high-quality translation can be obtained if the input sentence matches the translation pattern pairs and these translation pattern pairs are correct. However, translation pattern pairs and phrase pairs cost a lot to make. To decrease the cost, we have developed a method to make translation pattern pairs and phrase pairs automatically. This method, called Pattern-Based Statistical Machine Translation (Pattern-Based SMT), selects translations by using a word tri-gram, translation probabilities of phrase pairs, and translation probabilities of pattern pairs. We demonstrated its effectiveness in Japanese-English machine translation experiments.

Key words　Statistical Machine Translation, Pattern Based Translation, IBM Model 1, Translation Pattern, Translation Phrase

1. ま え が き

パターン翻訳は，非常に古い歴史をもっている [3] [2]．この翻訳方式は 1950 年代に提案されている．この方式は，元言語を対訳文パターンと対訳句を利用して，目的言語に変換する．この方式の利点は，仮に，入力文が，適切な対訳文パターンに適合し，適切な対訳句に変換できれば，非常に高い翻訳性能が得られる．この翻訳方式の代表例が、1970 年代の作成された Montreal 大学の TAUM であろう。このシステムは気象情報の英仏翻訳である。また，日本で初の機械翻訳システム（ヤマト）は，この方式とも言われている．しかし，多くの問題点をもつ．この 1 つは，適切な対訳文パターンが得にくいことである．入力文に対して，多くの場合複数の対訳文パターンが照合する．この中から適切な対訳文パターンを得るために，対訳文パターンは，言語的な制約を入れる．この言語的な制約が，過剰な制約となることがあるため，入力文に対するカバー率が，低下しがちである．よって対訳文パターンの照合率と翻訳精度にはトレードオフの関係がある．そして，最大の問題は，対訳文パターンと対訳句を人手で作成するため，非常にコストがかかる．

一方，1990 年代から，対訳文だけを利用した統計翻訳 (SMT) が非常に盛んになってきた．この方式の最大の特徴は，対訳文のみを収集すれば，翻訳が可能であるため，システムの制作コストが非常に少ない．また翻訳精度も，大量の対訳文を利用することで，比較的高い翻訳精度が得られる．なお，統計的機械翻訳は，単語に基づく統計機械翻訳と，句に基づく統計機械翻訳の 2 種類に分かれる．そして，両者は，対訳単語確率を IBM model を利用して計算している．単語に基づく統計機械翻訳は，対訳単語確率を利用して，翻訳をおこなう．句に基づく統計機械翻訳は，対訳単語確率を利用して，対訳句を抽出し，対訳文から対訳句確率を計算してから翻訳を行う．通常，句に基づく統計機械翻訳は，単語に基づく統計機械翻訳と比較して，翻訳精度は高い．しかし，文を句に分割して再合成していて，文法を全く考慮しないため，文法的に非常におかしく意味の通らない文を出力することも多い．

この問題を解決するために，パターンに基づく統計機械翻訳 (Pattern Based Statistical Machine Translation) を提案してきた [7] [14] [11] [8] [9] [13]．この方法は，単語に基づく統計機械翻訳で利用される対訳単語確率を利用して，自動的に，対訳文から対訳文パターンと対訳句を抽出し，その確率値を付与する．そして，確率値が付与された対訳文パターンと対訳句を利用して，翻訳を行う．自動的に，対訳学習文から対訳文パターンと対訳句を抽出するため，パターン翻訳において問題となっていたシステム構築のためのコストを大幅に削減できる．

本論文では，パターンに基づく統計機械翻訳の概要について

述べる．そして，提案したパターンに基づく統計機械翻訳と，一般的な句に基づく統計機械翻訳の，翻訳性能を比較する．最後に，提案方法の問題点について述べる．

2. パターンに基づく統計機械翻訳

Pattern based Statistical Machine Translation

句に基づく統計翻訳は，従来のパターン翻訳と比較すると，対訳文を準備するだけでシステムを構築できるため，非常にコストが低い．しかし，この方法は，入力文を対訳句に分割し，対訳句の翻訳結果を，組み合わせて出力文を作成する．文法を考慮しないで出力文を生成するため，文法的におかしい文（非文）が出力されることが多い．一方，パターン翻訳は，入力文が，適切な対訳文パターンに適合し，適切な対訳句に変換できれば，翻訳精度が高い出力文が得られる．

そこで，パターン翻訳の特にコストの問題を解決するために，本論文では，対訳句の自動抽出に着目した．対訳句の抽出には，多くの方法が提案されているが，本論文では，IBM Model を対訳単語確率の計算に利用した．このモデルは GIZA++としてインプリメントされていて，統計翻訳において，良く利用されている．以上を考慮してパターンに基づく統計機械翻訳を提案した．パターンに基づく統計機械翻訳の概要を以下に示す．

手順 1: 対訳学習文から対訳単語確率を利用して，対訳単語辞書を作成

手順 2: 対訳学習文と対訳単語辞書を利用して，単語レベル対訳文パターン辞書を作成

手順 3: 対訳学習文と単語レベル対訳文パターン辞書を利用して，対訳句辞書を作成

手順 4: 対訳学習文と対訳句辞書を利用して，句レベル対訳文パターン辞書を作成

手順 5: 句レベル対訳文パターン辞書と対訳句辞書と言語モデルを利用して翻訳

全体の構成図を図 1 に示す．

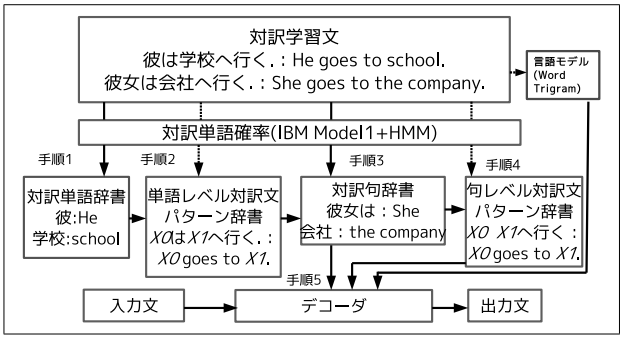

図 1　パターンに基づく統計機械翻訳

なお，パターンに基づく統計機械翻訳の根幹は，大量に正確な対訳句を持つ対訳句辞書を自動的に作成することにある．

以下にパターンに基づく統計機械翻訳の詳細を各手順ごとに述べる．

2.1 対訳単語辞書 [13]

最初に，対訳単語を対訳単語確率と対訳学習文を利用して作成する．

対訳単語確率は IBM Model1+HMM を利用して計算する．具体的には GIZA++を利用する．ただし，GIZA++を利用して直に対訳単語辞書を作成した場合，対訳単語の数が膨大になる．そこで，GIZA++から対訳単語を作成し，対訳単語から枝狩を行って対訳単語辞書を作成する．対訳単語辞書の作成方法を以下に示す [13].

手順 1: 対訳学習文から IBM Model1 + HMM を利用して，対訳単語確率を計算する．

手順 2: 全ての対訳単語と，その対訳単語確率を計算する．

手順 3: 対訳単語から，以下の閾値を用いて対訳単語の数を削減する．枝刈条件を以下に示す．

（a） 日本語単語から見た対訳英語の日英対訳単語確率の順位が A 以内

（b） 英語単語から見た日本語単語の英日対訳単語確率の順位が B 以内

（c） 対訳学習文において，日本語と英語の対訳単語の頻度が 2 以上

手順 4: 枝狩りされた対訳単語を対訳単語辞書とする．

なお，対訳単語辞書は，大量の対訳単語から成り立ち、対訳単語確率は持たない．

図 2 に，対訳単語辞書の例を示す．

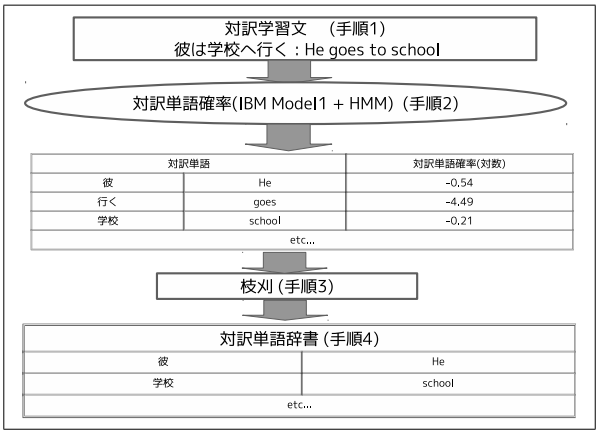

図 2　対訳単語辞書の作成例

2.2 単語レベル対訳文パターン辞書

次に，単語レベル対訳文パターンを．対訳単語辞書と対訳学習文を利用して作成する．単語レベル対訳文パターンの作成方法を以下に示す．

手順 1: 対訳学習文と，対訳単語辞書を比較する．

手順 2: 対訳学習文中の日本語と対訳単語中の日本語単語が一致し，かつ対訳学習文中の英語と対訳単語中の英単語が一致する箇所を，検索する．

手順 3: 一致した箇所を，変数化する．変数化された対訳学習文を単語レベル対訳文パターンとする．

手順 4: 対訳学習文中の日本語と対訳単語の日本語単語が一致し，かつ対訳学習文中の英語と対訳単語の英単語が一致する箇所がなくなるまで，手順 1 から 3 を繰り返す．これを単語レベル対訳文パターン辞書とする．

従って，単語レベル対訳文パターン辞書中の文パターンの多くは，複数の変数を持つ．なお，単語レベル対訳文パターン辞書は，大量の単語レベル対訳文パターンで構成されていて、確率を持たない．

図 3 に，単語レベル対訳文パターン辞書の例を示す．

2.3 対訳句辞書

対訳句辞書は，対訳学習文と単語レベル対訳文パターン辞書から作成する．ただし，対訳句の対訳の精度を向上させるため，大きく 3 つのステップで成り立っている [14], [8].

2.3.1 対訳句の抽出 [9]

初めに，可能性のある対訳句を全て抽出する．手順を以下に示す．

手順 1: 単語レベル対訳文パターンと対訳学習文を比較する．

手順 2: 照合した対訳学習文とパターンの対訳句を変数に置き換える．

手順 3: 照合する部分がなくなるまで，手順 1 と手順 2 を，繰り返す．

手順 4: 抽出した対訳句に対して，対訳単語確率を利用して，対訳フ

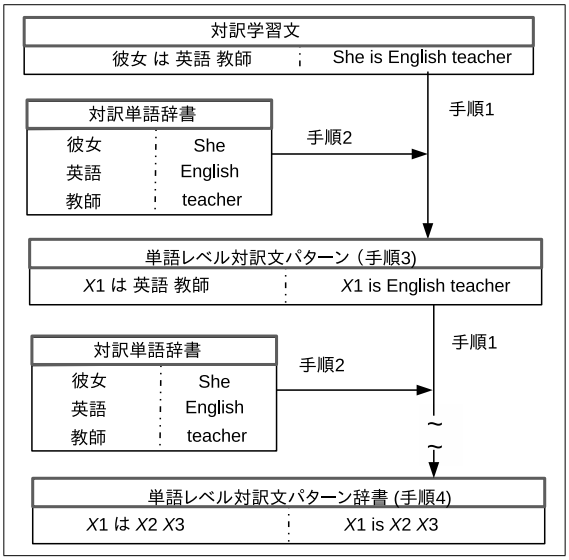

図 3　単語レベル対訳文パターン辞書の例

レーズ確率を付与する．

対訳フレーズ確率の計算式には多くの方法が考えられる。本論文では以下の式を用いた。

$$P(\text{対訳フレーズ確率}) =$$

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

$$\prod_{n=0}^{N-1} \sum_{m=0}^{M-1} p(\frac{J_n}{E_m}) \times \prod_{m=0}^{M-1} \sum_{n=0}^{N-1} p(\frac{E_m}{J_n})$$

J_n; 対訳句の日本語句の単語
E_m; 対訳句の英語句の単語
$p()$; 対訳単語確率
N; 日本語の単語数
M; 英語の単語数

なお，対訳句の抽出のアルゴリズムは CKY を変形した形になる [9]．対訳句の抽出例を図 4 に示す．この方法では，大量の対訳句が出力される．図 4 の例では，3 変数 4 組=12 対の対訳句が出力される．実際の実験では 1 億を超える対訳句が生成される．

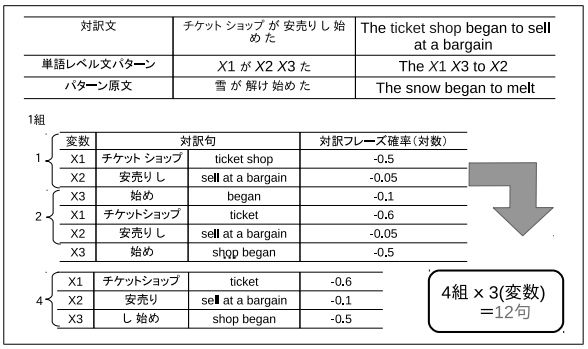

図 4　対訳句の例

2.3.2　パターン制約 [14]

2.3.1 節で抽出される対訳句は，膨大な数になる．そこで以下の制限を加える．

1 対の対訳学習文において 1 対の単語レベル対訳文パターンから出力される対訳句は，対訳句の組の対訳フレーズ確率の総和が最大の対訳句の組のみを，選択する．

本論文では，これをパターン制約と読んでいる．以下に具体的な手順を示す．

手順 1:　対訳学習文と単語レベル対訳文パターンを照合する．

手順 2:　適合した場合，単語レベル文パターンの変数部に対応する全ての組み合わせの対訳句を抽出する．

手順 3:　対訳単語確率を用いて，対訳フレーズ確率を計算する．

手順 4:　各単語レベル文パターンごとに，手順 4 で計算した対訳フレーズ確率の総和の最大値をとる対訳句を 1 つずつ選出する．

図 5 にパターン制約の例を示す．この例では，1 対の対訳学習文に対し，1 つの単語レベル対訳文パターンから 1 組 3 対の対訳句が選択されている．図 5 においては 12 対の対訳句が出力されているため，対訳句の数が削減されていることが解る．

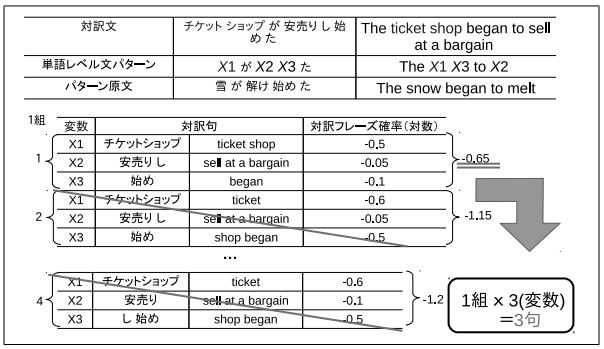

図 5　パターン制約の例

パターン制約は、主に対訳句の数を削減することに貢献している。実際の例では，対訳句の数が 4 桁削減されることがある．

2.3.3　類似制約 [8]

パターン制約（2.3.2 節）は，1 つの単語レベル対訳文パターンにつき 1 組の対訳句を抽出するため，パターン制約がない場合と比較すると、生成される対訳句の数を大幅に削減できる．しかし，単語レベル対訳文パターンは複数（通常対訳学習文の数）ある。そして、対訳学習文に対し不適切な単語レベル対訳文パターンを適合した場合に不適切な対訳句を出力する．

そこで，対訳句を抽出する際，単語レベル対訳文パターンを作成する際に用いた対訳学習文 (以下，パターン原文) と対訳句を作成する際に用いる対訳学習文との類似度を利用して，不適切な対訳句を削除する．これを類似制約と呼んでいる．

この類似制約の手順を以下に示す．

手順 1:　対訳句作成時の対訳学習文とパターン原文を抽出する．

手順 2:　対訳文中の日本文から見たパターン原文中の日本文の，文の類似度 F を計算する．

手順 3:　パターン原文中の日本文から見た対訳文中の日本文の、文の類似度 G を計算する．

手順 4:　英文でも手順 1〜3 と同様に計算し、類似度 H、I を計算する。

手順 5:　類似度 J を F, G, H, I の総積とする。$J = F \times G \times H \times I$

手順 6:　類似度 J の閾値 K を用いて対訳句を削除する．実際には K は類似度 J の N-best の値とする。

手順 7:　手順 6 で選出した対訳句を最終的に抽出される対訳句として出力する．

文の類似度は，対訳学習文とパターン原文の同一の単語の出現率（文の Cosine 類似度）で，以下の式を用いて計算する．

$$P(\text{類似度}) = P(\frac{B_1 ... B_N}{A_1 ... A_M}) = (\frac{S}{M})$$

対訳文 A は単語 $A_1 ... A_M$ で構成

対訳文 B は単語 $B_1 ... B_N$ で構成

M; 対訳文A中の単語数

S; 対訳文A中の単語が対訳文Bの単語と一致している単語数

類似制約の例を，図 6 に示す．

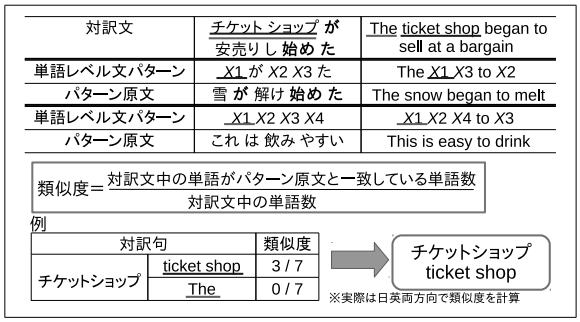

図 6　類似制約の例

類似制約は主に対訳句の精度向上に寄与している。類似度を、日本文と英文の両方向で計算することで、対訳句の誤りを大幅に削減している。日本文もしくは英文の 1 方向で計算した場合より，精度向上する．

2.3.4　対訳句辞書

対訳句辞書は枝狩した対訳句と、その対訳句確率で構成される。

（1） 枝刈条件

パターン制約および類似制約において選択された対訳句から以下の枝刈り条件を満たす対訳句を選択する。枝刈条件を以下に示す．

（a） 日本語句から見た対訳英語句の対訳フレーズ確率の順位が G 以内

（b） 英語句から見た日本語句の対訳フレーズ確率の順位が H 以内

（c） 対訳学習文において，日本語句と英語句の対訳句の頻度が 2 以上

（2） 対訳句確率

対訳句確率は，対訳学習文の出現回数から計算する（Dice 係数）．対訳単語確率 (IBM Model) や対訳フレーズ確率は用いない．理由は 5.5 節に述べる．計算式を以下に示す．

$$P(\text{対訳句確率}) = \frac{Count(\text{対訳句中の日本句が出現した対訳学習文の数})}{Count(\text{対訳句が出現した対訳学習文の数})} \times \frac{Count(\text{対訳句中の英語句が出現した対訳学習文の数})}{Count(\text{対訳文が出現した対訳学習文の数})}$$

句レベル対訳文パターン辞書の例を図 8 に示す。

2.4　句レベル対訳文パターン辞書

対訳学習文と対訳句辞書を利用して，句レベル対訳文パターン辞書を作成する．句レベル対訳文パターン辞書は句レベル対訳文パターンと句レベル対訳文パターン確率で構成される。以下にフローを示す [11].

手順 1: 対訳学習文と対訳句辞書を照合する．

手順 2: 対訳学習文中の日本語句と対訳句の日本語句が一致し，かつ対訳学習文中の英語句と対訳単語の英語句が一致する箇所を，検索する．

手順 3: 一致した対訳句を変数化する．

手順 4: 対訳学習文中の日本語句と対訳句の日本語句が一致し，かつ対訳学習文中の英語句と対訳句の英語句が一致する箇所がなくなるまで，手順 1 から手順 3 を繰り返す．

なお，句レベル対訳文パターンは，変数部の対応が取れるたびに，対応をとるパターンと元の字面のパターンを作成する．従って句レベル対訳文パターンの数は，非常に多くなる．

手順 5: 句レベル対訳文パターンに句レベル対訳文パターン確率を付与する．句レベル対訳文パターン確率は，文パターンが出現する対訳学習文の数で計算する．計算式を以下に示す．

$$P(\text{句レベル対訳文パターン確率}) = \frac{Count(\text{句レベル対訳文パターン中の日本文パターンが出現した対訳学習文の数})}{Count(\text{句レベル対訳文パターンが出現した対訳学習文の数})} \times \frac{Count(\text{句レベル対訳文パターン中の英文パターンが出現した対訳学習文の数})}{Count(\text{句レベル対訳文パターンが出現した対訳学習文の数})}$$

なお，句レベル対訳文パターンの生成のアルゴリズムは CKY の変形の形になる [9].

図 7 に句レベル対訳文パターン辞書の例を示す．

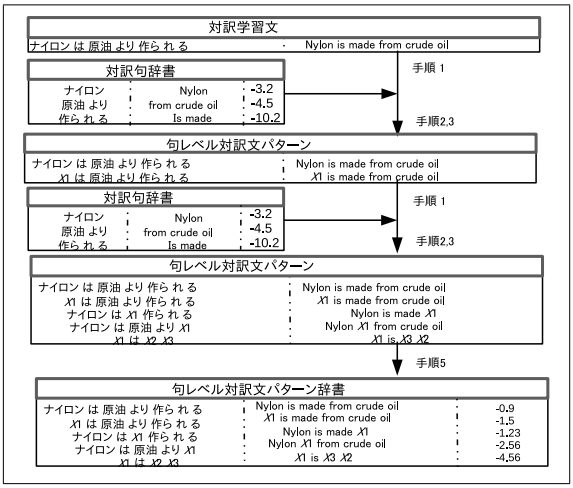

図 7　句レベル対訳文パターン辞書の例

2.5　翻訳（デコーダ）

翻訳には，句レベル対訳文パターン辞書と対訳句辞書を用いる．翻訳のステップを以下に示す．

手順 1: 入力文に照合する句レベル対訳文パターンを選択する．

手順 2: 句レベル対訳文パターン中の日本語変数部と変数に相当する日本語句を抽出する．

手順 3: 対訳句パターン（日本語）に相当する対訳句パターン（英語）を得る．

手順 4: 対訳句辞書を利用して，句レベル対訳文パターン中の日本語における変数部に相当する英語の変数を抽出する．

手順 5: 句レベル対訳文パターン中の英語における変数部に英語を代入する．

手順 6: 句レベル対訳文パターン確率と，対訳句確率と，言語モデル (英語の単語 trigram) を総計して，最尤の出力文を選択する．なお，英語の単語 trigram は，対訳学習文の英文から計算する．

翻訳の例を図 8 に示す．

2.6　実験に利用した対訳学習文

実験に利用した対訳学習文は，電子辞書などから抽出した [12]．日本文は単文である．ただし英文は単文および複文の場合もある．学習には， 対訳学習文 163,188 対を利用した．テストには

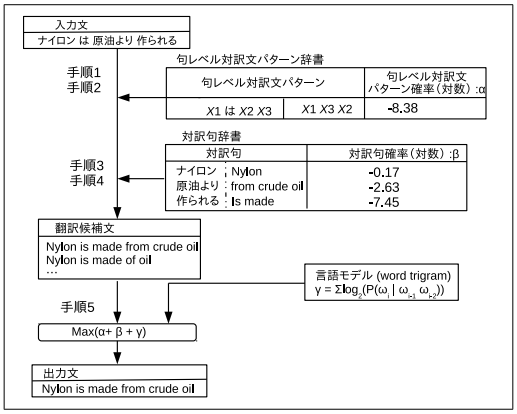

図 8　翻　　訳

1000 文利用した．また，日本語は Mecab [5] を利用して，単語分割した．英語は，Moses [1] の標準 tokenizer を利用した．表 1 に，実験に利用した対訳学習文の例を示す．

表 1　実験に利用した対訳学習文の例

	単文
J	星 が 光っ て いる .
E	The stars are twinkling .
J	信長 勢 は 再び 京 へ 上っ た .
E	Nobunaga's army went up to Kyoto again .
J	彼女 は 我 を 通し た .
E	She had her own way .
J	ぶらんこ が 揺れ て いる .
E	The swing is swinging .
J	私 は 電車 事故 で 足留め を 食っ た .
E	I was stranded as a result of the train accident .

言語モデルとしての，英語の word trigram の計算は，163,188 文の対訳学習文の英文から計算した．

3. 実　　験

3.1　対訳句と対訳句パターンの数

実験の結果以下の対訳辞書が得られた．

（1） 対訳単語辞書
対訳単語として 32,2347 対が得られた．

（2） 単語レベル対訳文パターン辞書
単語レベル対訳文パターンとして 168,136 対が得られた．

（3） 対訳句辞書
対訳句として 2,219,452 対が得られた．

（4） 句レベル対訳文パターン辞書
句レベル対訳文パターンとして 249,832,215 対が得られた．

3.2　句に基づく統計翻訳 (Moses)

実験には，提案手法に対する比較として，句に基づく統計翻訳 (Moses) を利用した．学習条件は、提案手法と同じである。ただし，Moses は言語モデルとして 5-gram を利用した．また MERT [4] を利用して開発データ (Development Data)1000 文でパラメータの最適化を行った．

4. 実 験 結 果

4.1　Translation Example

翻訳結果の例を表 2 に示す．また Moses の出力文も示す。

表 2　翻　訳　例

入力文	信号 が 青 より 赤 に 変わっ た .
出力文	The traffic lights changed from green to red .
参照文	The signal changed from green to red .
パターン (J)	X02 が X03 X00 X01 に 変わっ た .
パターン (E)	The X 02 changed X 00 X 03 to X 01 .
パターン原文 (J)	風 が 北 から 南 に 変わっ た .
パターン原文 (E)	The wind changed from north to south .
対訳句 00 (J)	より
対訳句 00 (E)	from
対訳句原文 00 (J)	５０ ページ より 続く .
対訳句原文 00 (E)	Continued from page 50 .
対訳句 01 (J)	赤
対訳句 01 (E)	red
対訳句原文 01 (J)	彼女 の ドレス は 燃え立つ よう な 赤 だ .
対訳句原文 01 (E)	Her dress is flaming red .
対訳句 02 (J)	信号
対訳句 02 (E)	traffic lights
対訳句原文 02 (J)	その 信号 は 点滅 し て い た .
対訳句原文 02 (E)	The traffic lights were flashing .
対訳句 03 (J)	青
対訳句 03 (E)	green
対訳句原文 03 (J)	信号 は 青 に なっ た .
対訳句原文 03 (E)	The signal turned to green .
TRIGRAM（対数）	-67.302279
対訳句確率の総和 （対数）	-434.807906
パターン確率（対数）	-14.684444
Moses	The signale turned red rather than blue .

このテーブルの項目の意味を以下に示す．

パターン (J)	入力文に照合した句レベル対訳文パターンの日本語パターン
パターン (E)	入力文に照合した句レベル対訳文パターンの英語パターン
パターン原文 (J)	“パターン (J)” の元となった対訳学習文の日本文
パターン原文 (E)	“パターン (E)” の元となった対訳学習文の英文
対訳句 (J)	“入力文” と “パターン (J)” が 照合した対訳句の日本語句
対訳句 (E)	“入力文” と “パターン (J)” が 照合した対訳句の英語句
対訳句原文 (J)	“対訳句 (J)” の元となった対訳学習文の日本文
対訳句原文 (E)	“対訳句 (E)” の元となった対訳学習文の英文

この例では”青”が”green” に、”より”が”from” になっていることに着目したい．通常の翻訳では，”青”は”blue”，”より”は”rather” になる．

4.2　自動評価結果

実験には自動評価方法として， BLEU, METEOR, RIBES, TER の 4 つの方法を利用した．これらの結果を表 3 に示す．なお，比較のために Moses の結果も示す．

表 3　自動評価結果

	BLEU	METEOR	RIBES	TER
Proposed	0.1537	0.4329	0.7719	0.6242
Moses	0.1375	0.4284	0.7463	0.6723

表 3 より，自動評価では，提案方法は Mose より翻訳精度が高いことが解る．

4.3　人手評価結果

翻訳精度の評価として人手評価を行った．人手評価には対比較試験を利用した．出力文からランダムに 100 文を選出し，提案手法と Mose との対比較を行った．評価結果を表 4 に示す．

表 4　人手評価　対比較試験

提案手法 ○	Moses ○	差なし	同一
30	19	14	37

表の項目の意味を以下に示す．

提案手法 ○	提案手法の翻訳品質が Mose より良い．
Moses ○	提案手法の翻訳品質が Mose より悪い．
差なし	提案手法と Moses の翻訳品質に差がない．
同一	提案手法と Moses の出力が文字単位で完全に同一．

表 4 から，パターンに基づく統計機械翻訳（提案手法）は，標準的な句に基づく統計機械翻訳 (Moses) より翻訳精度が優れていることが解る．これは，提案手法の有効性を示している．

5. 考　　察

5.1 提案手法の有効性

パターンに基づく統計機械翻訳（提案手法）は，標準的な句に基づく統計機械翻訳 (Moses) より翻訳精度以外にも以下の点で優れている．

（1） 人手による修正の容易さ

パターンに基づく統計機械翻訳は，句レベルの対訳文パターンの原文（対訳学習文）や，対訳句の原文（対訳学習文）が理解できる．したがって，出力文が間違っていても，翻訳修正が人手で明確にできる場合が多い．

（2） N-best の有効性

出力文を複数候補出力すると，第 1 候補は誤っていても，第 2 候補が合っている場合が多い．そこで，人手によって出力文を選択できる．ちなみに句に基づく統計機械翻訳 (Moses) における N-best は，第一候補に類似した出力文がでることが多いため，実質的には N-best に意味がない．

5.2 頻度 1 の対訳句 [8], [13]

対訳句辞書において，頻度 1 の対訳句は削除した．これは，対訳学習文中に頻度が 1 回しか出現しない対訳句の確率は，信頼性がないため，翻訳精度の低下をまねいたためである．しかし，特に固有名詞は，対訳学習文中に 1 回しか出現しないことが多い．また，慣用的な言い回しも同様である．そのため，対訳句のカバー率が不足がちになる．今後，対訳学習文中に頻度が 1 回しか出現しない対訳句の抽出方法を考えていきたい．

5.3 未知後処理 [10]

提案手法では，対訳句辞書において，対訳学習文中に頻度が 1 回しか出現しない対訳句は削除した．そのため，固有名詞が未知語になりやすい．なお，提案手法は，未知語は，出力文（英文）中の日本語として出力される．この未知語処理として，市販の辞書から対訳句を抽出して，対訳句辞書に登録する方法がある．また，対訳単語辞書を利用する方法もある．

5.4 自動評価と人手評価

自動評価と人手評価を比較すると，特に人手評価において提案方法の有効性が明確に解る．これは，提案手法は，パターン翻訳を基礎としているため，文法的に正しい文が出力されることが多い．一方 Moses は，対訳句をつないで文を出力するため，非文を出力することもある．この差がでている．

なお，自動評価の問題もある．自動評価では，多くの評価方法では，局所的な翻訳精度を評価して，文全体にわたる文法的な評価をしていない．そのため，ルールベースの機械翻訳と統計翻訳の翻訳精度は，人手評価と自動評価において，大きな差が生じる．これと同じ原因が，提案方法における評価結果になったと考えている．

また，基本的に BLEU は短い文において評価に信頼性が低いことが知られている [6]．本研究では単文を利用した．平均の単語数は，約 15 単語である．したがって特に BLEU において問題が出たと考えている．

5.5 対訳句確率と句に基づく対訳文パターン確率

研究では，対訳句確率と句に基づく対訳文パターン確率は，対訳学習文中に出現する回数で計算した．つまり頻度を基本に計算した．そのため線形推定法と言える．（なお，対訳句確率は dice 係数になる．）

しかし，これらの確率には別の計算方法がある．対訳句確率と句に基づく対訳文パターン確率を，対訳単語確率（IBM Model）から計算することができる [11]．例えば，対訳句確率を IBM Model を利用した対訳フレーズ確率に置き換えることができる．しかし，IBM Model は EM 推定法である．EM 推定法は非線形推定法である．非線形推定法は，特に学習データが少ないとき，値に信頼性が持てない．特に対訳学習文において頻度 1 の対訳フレーズ確率は，人手からみて異常な値を持っている．

また，MERT と BLEU を利用して，対訳句確率と句に基づく対訳文パターン確率と言語モデルの確率を最適化する方法は，BLEU 値に信頼がおけないため，問題があると考えている．

今後，他の対訳句確率と句に基づく対訳文パターン確率の計算方法を考えていきたい．

5.6 変換主導型パターンベース統計機械翻訳

この論文では、パターンに基づく統計機械翻訳を提案した。この方式において、翻訳（デコーダ）は、句に基づく文パターンを主軸に翻訳していると考えられる。しかし、これの拡張系として、句に基づく文パターンのパターン原文を主軸にして翻訳する方法が考えられる。これを、変換主導型パターンベース統計機械翻訳と呼んでいる。この方法を今後追求していきたい。

6. 結　　論

本論文では，パターンに基づく統計機械翻訳 (Pattern Based Statistical Machine Translation) を提案した．この方法は，対訳学習文から自動的に句レベル対訳文パターンと対訳句を抽出し，これを利用して翻訳を行う．実験の結果，句に基づく統計機械翻訳として代表的な Moses と比較して翻訳精度が高く，提案手法の有効性が示された．特に，文法的に整っている文が出力されやすい．そのため，この有効性は，特に人手評価において示されている．

なお，提案方法は，対訳句辞書や句レベル対訳文パターンの作成方法や確率の計算方法に，数多くの方法が考えられる．今後，これらの方法を追求していきたい．

文　　献

[1] Philipp Koehn et. al. Moses: Open source toolkit for statistical machinetranslation. *Article of the ACL 2007 Demo and Poster Sessions*, pages 177–180, 2007.

[2] W.John Hutchins and Harold L. Somers. An introduction to machine translation. *London: Academic Press*, pages 207–220, 1992.

[3] Hiroshi Maruyama. Pattern-based translation: Context-free transducer and its applications to practical nlp. *In Proc. of Natural Language Pacific Rim Symposium*, pages 232–237, DECEMBER 1993.

[4] Franz Josef Och. Minimum error rate training for statistical machine translation. *Article of the ACL*, 2003.

[5] Yuji Matsumoto Taku Kudo, Kaoru Yamamoto. Applying conditional random fields to japanese morphological analysis. *Article of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, 2004.

[6] Kishore Papineni Salim Roukos Todd Ward Wei-Jing Zhu. Bleu : a method for automatic evaluation of machine translation. *Article of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

[7] 江木 孝史. 句に基く文パターンを用いた英日翻訳. **鳥取大学修士論文**, March 2014.

[8] 栗下 尚樹. 類似度を用いた対訳句辞書の作成. **鳥取大学卒業論文**, March 2017.

[9] 春野 瑞季. 多変量解析を用いたパターンに基づく統計翻訳. **鳥取大学修士論文**, 3 2014.

[10] 村上 仁一 川原 宰. パターンに基づく統計翻訳とその未知語処理. **言語処理学会第 *23* 回年次大会**, pages 11–7, March 2017.

[11] 西尾 聡一郎. 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. **鳥取大学卒業論文**, March 2017.

[12] 藤波進 村上仁一. 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**, pages 119–130, March 2012.

[13] 中村 友哉. パターンに基づく統計翻訳における対訳単語辞書の精度調査. **鳥取大学卒業論文**, March 2017.

[14] 興梠 玲架. パターンに基づく統計翻訳において変数部の確率の総和を使った対訳句の抽出. **鳥取大学卒業論文**, March 2016.