概要

日英統計翻訳において，日本語文は複数のフレーズ対を用いてフレーズ単位で変換される．そして，そのフレーズの順序を並び替え，英語文に翻訳される. しかし，重文複文といった複雑な日本語文を翻訳する場合，多くのフレーズ対が必要となる．そのため，フレーズの並び替えの候補数が膨大になり，翻訳精度が低くなる傾向がある．

そこで本研究では，長いフレーズを持つフレーズ対を増やすことで，出力文が利用するフレーズ対の数を減らし，並び替えの候補を減らす手法を提案する．具体的には，3種類の学習データから得られたフレーズテーブルをそれぞれ従来の単語区切りフレーズテーブルと併用し，翻訳精度の向上を目指す.

1つ目は,「日本語文を文節区切り，英語文を単語区切りとした学習データ」から生成されたフレーズテーブル．2つ目は,「日本語文を単語区切り，英語文をフレーズ単位に統合した」学習データから生成されたフレーズテーブル．そして，3つ目は,「日本語文を文節区切り，英語文をフレーズ単位に統合した学習データ」から生成されたフレーズテーブルである．

実験の結果，従来手法の翻訳精度と比較して，どのフレーズテーブルを併用した場合でも翻訳精度は向上した．さらに，提案手法により得られた3つのフレーズテーブルを全て併用することで，従来手法と比較して，BLEU スコアが単文の翻訳で 0.71%，重文複文の翻訳で 0.51%向上した.

目 次

図 目 次

表 目 次

1　はじめに

国際化が進む現代社会において，言語の違いはコミュニケーションの大きな障害となっている．そのため，他言語間のコミュニケーションを容易にする，機械翻訳の技術の必要性が高まっている．従来のルールベース法を用いた機械翻訳方式では，1 つの言語間の翻訳システムを作るために，長い時間をかけて翻訳規則を構築する必要がある．また，言語によって，文法規則が異なるため，多言語への拡張が難しい．そこで，現在，対訳データから自動的に翻訳規則を獲得し，翻訳を行う統計翻訳 [1] が注目されている．統計翻訳は，対訳データがあれば翻訳規則を構築できるため，短い時間で翻訳システムを構築でき，多言語への拡張が容易である．

日英統計翻訳において，日本語文は複数のフレーズ対を用いてフレーズ単位で変換される．そして，そのフレーズの順序を並び替え，英語文に翻訳される. しかし，重文複文といった複雑な日本語文を翻訳する場合，多くのフレーズ対が必要となる．そのため，フレーズの並び替えの候補数が膨大になり，翻訳精度が低くなる傾向がある [2].

そこで本研究では，長いフレーズを持つフレーズ対を増やすことで，出力文が利用するフレーズ対の数を減らし，並び替えの候補を減らす手法を提案する．具体的には，3 種類の学習データから生成されたフレーズテーブルをそれぞれ従来の単語区切りフレーズテーブルと併用し，翻訳精度の向上を目指す．1 つ目は,「日本語文を文節区切り，英語文を単語区切りとした学習データ」から生成されたフレーズテーブル．2 つ目は,「日本語文を単語区切り，英語文をフレーズ単位に統合した学習データ」から生成されたフレーズテーブル．そして，3 つ目は,「日本語文を文節区切り，英語文をフレーズ単位に統合した学習データ」から生成されたフレーズテーブルである．また，3 つのフレーズテーブルを全て併用した場合の翻訳実験も行う．

実験の結果，従来手法と比較して，BLEU スコアが単文で 0.43%，重文複文で 0.38%向上した．また，同等の手法を学習データの英語文に対して適用した場合の実験も行った．英語文に適用した場合も，従来手法と比較して，翻訳精度が向上した．さらに，従来手法のフレーズテーブルと提案手法により得られた 3 つのフレーズテーブルを併用した実験も行った．4 つのフレーズテーブルを用いた場合，従来手法と比較して，BLEU スコアが単文の翻訳で 0.71%，重文複文の翻訳で 0.51%向上した．

本論文の構成は以下の通りである．まず，2 章で日英統計翻訳についての概要を示し，各モデルの学習について述べる．3 章では本研究の提案手法について述べる．4 章では実験に用いるデータやツールといった実験環境について述べる．5 章では，提案手法の効

果を示す．6 章では，提案手法を学習データの英語文に適用した場合の効果を示す．7 章では，本研究で生成したフレーズテーブルを全て併用した場合の効果を示す．そして，8 章で，考察を行い，最後に 9 章で結論を述べ，まとめる．

1.1 関連研究

関連研究として，鏡味ら [3] は長いフレーズを持つフレーズ対として，人手で作成された日本語フレーズと英語フレーズの対応に確率を付与した．しかし，人手でフレーズ対応を作成するには，多くの時間がかかり，低コストで翻訳システムを構築できるという統計翻訳の長所が損なわれる．また，フレーズ対応に適当な確率を付与することは難しい．

村上ら [4] は，長いフレーズを持つフレーズ対を生成するために，学習に用いられるオプション “max-phrase-length” の値を変更した．“max-phrase-length” は，フレーズテーブル内の日本語と英語のフレーズ中の単語数の上限値であり，7 がデフォルトの値である．デフォルトの値はフランス語と英語の翻訳のために設定された値である．村上らは対応する単語の位置が大きく変化する日本語と英語の翻訳では，フレーズが長いほどフレーズ対の精度がよくなるとし，“max-phrase-length” の値を 20 とした．本研究でも同様の考えから，“max-phrase-length” の値は学習データの単語数の上限である 100 とする．

2　日英統計翻訳システム

2.1　基本的な考え方

日英統計翻訳は，日本語文 J が与えられたとき，全ての組み合わせの中から確率 $P(E \mid J)$ が最大になる英語文 $\hat{E}$ を探索することによって翻訳を行う．以下に基本モデルを示す．

$$\hat{E} = \underset{E}{argmax} P(E \mid J) \quad (1)$$

$$\simeq \underset{E}{argmax} P(J \mid E)P(E) \quad (2)$$

$P(J \mid E)$ は翻訳モデル，$P(E)$ は言語モデルと呼ぶ．また，$\hat{E}$ を探索する翻訳システムをデコーダと呼ぶ．式 1 は，ベイズの法則により，式 2 に展開される．これは，2 言語間のモデル $P(E \mid J)$ を正確に推定することが困難なためである．そのため，一般に，高い信頼性を持つことが知られている言語モデルを翻訳モデルと併用し，翻訳の精度を高めている．

2.2　翻訳モデル

翻訳モデルは日本語の単語列から英語の単語列へ確率的に翻訳を行うためのモデルである．翻訳モデルには，大きくわけて語に基づく翻訳モデルと句に基づく翻訳モデル [5] がある．初期の統計翻訳は，語に基づく翻訳モデルを用いていた．しかし，翻訳精度の高さから，現在は句に基づく翻訳モデルが主流となっている．句に基づく翻訳モデルは表 1 に示すフレーズテーブルと呼ばれる表で管理される．

表 1: フレーズテーブルの例

庭 ||| The garden ||| 0.0434 0.3869 0.0036 0.0261 2.718
庭 ||| a garden ||| 0.5 0.7737 0.0073 0.1612 2.718
庭 から ||| from the garden ||| 0.5 0.1728 1 0.0327 2.718
庭 から 入る ドア || door from the garden ||| 1 0.0001 1 0.0147 2.718
庭 が ある ||| There is a garden ||| 1 0.0421 1 0.0036 2.718
庭 が 荒れる ||| The garden is wasted ||| 1 0.0014 1 0.0001 2.718
庭 で ||| garden at ||| 0.3333 0.16471 0.0833 0.0480 2.718

左から，日本語フレーズ，英語フレーズ，フレーズの英日翻訳確率 $P(j \mid e)$，英日方向の単語の翻訳確率 (IBM モデル) の積，フレーズの日英翻訳確率 $P(e \mid j)$，日英方向の

単語の翻訳確率 (IBM モデル) の積，フレーズペナルティである．以後，フレーズペナルティは常に一定の値であるため省略する．本稿では，日本語フレーズ，英語フレーズ，各種確率の 3 つをまとめて，フレーズ対と呼ぶ．フレーズ対は，日本語と英語の単語の対応付けを行い，その対応から日本語フレーズと英語フレーズの対応を抽出し生成される．

2.2.1 両方向の単語対応の作成

日本語と英語の単語の対応付けは，IBM モデル [1] を用いて行う．IBM モデルにより得られる単語の対応関係は，日英方向や英日方向といった，片方向の対応関係である．統計翻訳では，より精度の高い単語の対応関係として，片方向の対応関係から両方向の対応関係を作成する．両方向の単語の対応関係の作成例を図 1 に示す．

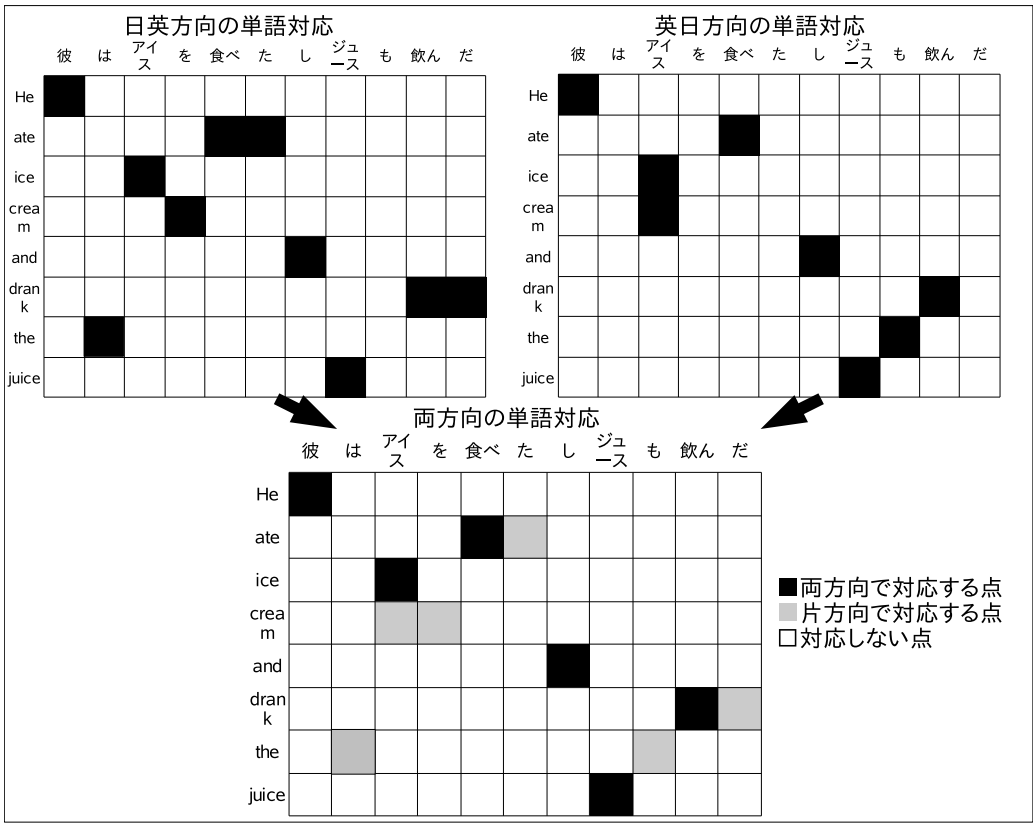

図 1: 両方向の単語の対応関係

英日方向と日英方向の単語の対応関係から，両方向の単語の対応関係を求めるためのオプションに “intersection” がある．これは，英日方向と日英方向の両方に対応がある点

を，両方向の対応点にするオプションである．図 1 の両方向の単語の対応において，黒いマスが “intersection” の対応点である．

また，“intersection” を拡張したオプションに “grow-diag” がある．これは，“intersection” の対応点に加えて，片方向で対応点があり，かつ，“intersection” の対応点と隣り合う点を，両方向の対応点にするオプションである．図 1 の両方向の単語の対応において，“は” と “the” の対応を除く灰色のマスが “grow-diag” で拡張した対応点である．

さらに，“grow-diag” を拡張としたオプションに，“grow-diag-final” と “grow-diag-final-and” がある．“grow-diag-final” は，片方向で対応点があり，かつ，“grow-diag” において少なくとも片方の言語の単語の対応がない点を，“grow-diag” の対応点に追加するオプションである．図 1 の両方向の単語の対応において，“は” と “the” の対応が “grow-diag-final” で拡張した対応点である．“grow-diag-final-and” は，片方向で対応点があり，かつ，“grow-diag” において両方の言語の単語の対応がない点を，“grow-diag” の対応点に追加するオプションである．図 1 の両方向の単語の対応において，“は” と “the” の対応が “grow-diag-final” では拡張されるが，“grow-diag-final-and” は “the” に対応する点があるため，拡張されない．

2.2.2 フレーズ対応の抽出

両方向の単語対応から，フレーズ対応を抽出する．例を図 2 に示す．

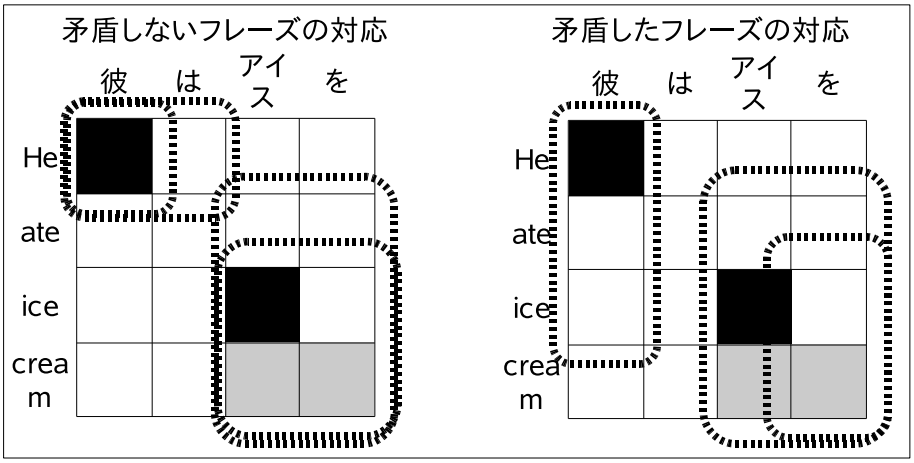

図 2: フレーズ対応の抽出例

抽出されるフレーズ対応の条件は，フレーズ対応が “矛盾しない” ことである．“矛盾しない” とは，フレーズ対応に含まれている単語は互いに対応しており，かつ，他の単

語に対応しないということを表す．例えば，フレーズ対応 “アイス を ||| ice cream” において，日本語の単語 “アイス” と “を” と，英語の単語 “ice” と “cream” は，それぞれフレーズ対応に含まれる単語にしか対応していない．そのため，“アイス を ||| ice cream” は “矛盾しない” と判断し，抽出できる．

フレーズ対応 “彼 ||| He ate ice” は，英語の単語 “ice” が，フレーズ対応に含まれない日本語の単語 “アイス” と対応している．そのため，“彼 ||| He ate ice” は “矛盾した” と判断し，抽出しない．

図 1 の両方向の単語の対応から抽出した “矛盾しない” フレーズ対応を，表 2 に示す．

表 2: 両方向の単語の対応関係から抽出したフレーズ対応

彼 ||| He
彼 は アイス を 食べた し ジュース も 飲ん だ ||| He ate ice cream and drank the juice
アイス を ||| ice cream
アイス を 食べた ||| ate ice cream
アイス を 食べた し ||| ate ice cream and
食べた ||| ate
し ||| and
ジュース ||| juice
飲ん だ ||| drank

2.2.3 抽出したフレーズ対応の確率付け

抽出したフレーズ対応に対して，確率付けを行う．日本語フレーズ J_{phrase} と英語フレーズ E_{phrase} からなるフレーズ対応の確率は以下の式で計算される．

$$P(J_{phrase} \mid E_{phrase}) = \frac{\text{学習データ中で } J_{phrase} \text{と } E_{phrase} \text{が同時に出現した数}}{\text{学習データ中で } E_{phrase} \text{が出現した数}} \quad (3)$$

$$P(E_{phrase} \mid J_{phrase}) = \frac{\text{学習データ中で } J_{phrase} \text{と } E_{phrase} \text{が同時に出現した数}}{\text{学習データ中で } J_{phrase} \text{が出現した数}} \quad (4)$$

2.3 言語モデル

言語モデルは単語列に対して，それらが生成される確率を与えるモデルである．日英翻訳では，言語モデルを用いて，訳文候補の中から英語として自然な文を選出する．言語モデルとして代表的なものに N-gram モデルがある．N-gram モデルは，“単語列 $P(W_1^n) = w_1, w_2, ..., w_n$ の i 番目の単語 w_i の生起確率 $P(w_i)$ は直前の単語列 $w_{i-(N-1)}, w_{i-(N-2)}, ..., w_{i-1}$ に依存する”，という仮説に基づくモデルである． これは，以下の式で表せる．

$$
\begin{aligned}
P(W_1^n) &= P(w_1)P(w_2 \mid w_1)...P(w_n \mid w_1^{n-1}) && (5)\\
&\approx P(w_1)P(w_2 \mid w_1)...P(w_n \mid w_{n-(N-1)}^{n-1}) && (6)\\
&= \prod_{i=1}^{n} P(w_i \mid w_{i-(N-1)}^{i-1}) && (7)
\end{aligned}
$$

また，$P(w_i \mid w_{i-(N-1)}^{i-1})$ は以下の式で計算される．ここで，$C()$ は単語列の出現数である．

$$
P(w_i \mid w_{i-(N-1)}^{i-1}) = \frac{C(w_{i-(N-1)}^{i})}{C(w_{i-(N-1)}^{i-1})} \tag{8}
$$

例えば,「He is japanese .」という単語列に対して 2-gram の言語モデルを適用した場合，単語列が生成される確率は以下の式で計算される．

$$
P(\text{“}He\, is\, japanese\,.\text{”}) = P(He) \times P(is \mid He) \times P(japanese \mid is) \times P(. \mid japanese) \tag{9}
$$

しかし，式 8 から信頼できる値を算出するためには, 大規模なコーパスを用いて，各単語列の出現数を高める必要がある．そこで，出現数の少ない単語列をモデルの学習から削除 (カットオフ) する手法や，確率が 0 となるのを防ぐために，大きい確率を小さく，小さい確率を大きくするスムージング手法が提案されている．スムージングの代表的な手法にバックオフ・スムージングがある．バックオフ・スムージングは学習データに出現しない N-gram の値をより低い次数の N-gram の値から推定する．trigram の例を以下に示す.

$$
P(w_i \mid w_{i-2}^{i-1}) = \begin{cases} \alpha \times p(w_i \mid w_{i-2}^{i-1}) & trigram \text{ が存在する} \\ \beta \times p(w_n|w_{n-1}) & trigram \text{ がなく } bigram \text{ が存在する} \\ p(w_n|w_{n-1}) & \text{それ以外} \end{cases} \tag{10}
$$

ここで，α をディスカウント係数，β をバックオフ係数と呼ぶ.

2.4 デコーダ

デコーダは翻訳候補から，翻訳モデルと言語モデルの確率が最大となる英語文を探索し，出力する．入力文として「彼はアイスを食べた」を与えた場合の探索の例を図 3 に示す.

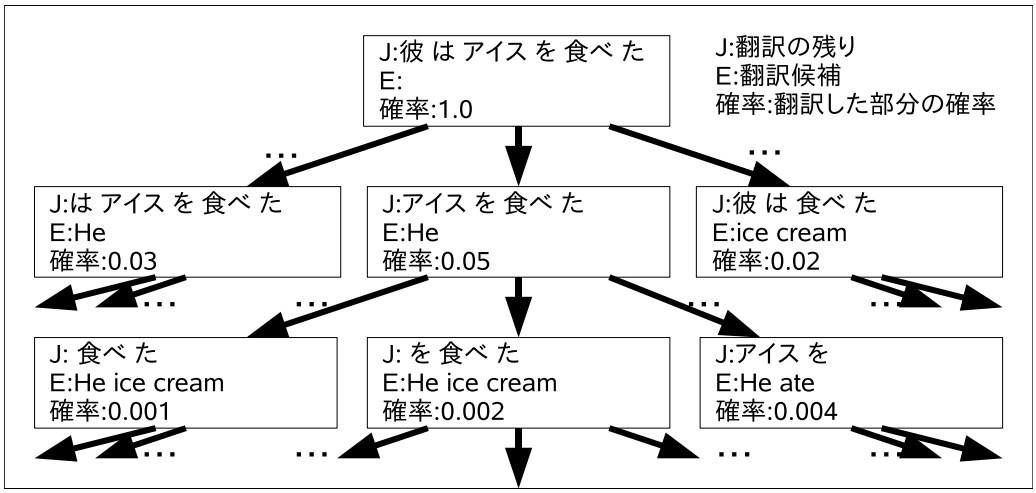

図 3: 入力文「彼はアイスを食べた」の探索例

デコーダはまったく翻訳されていない仮説から探索を始める．そして，入力文 J の中から翻訳されていないフレーズを選択し，翻訳モデルのフレーズ対を用いて翻訳候補を生成する．このとき，翻訳候補の確率を計算し，その翻訳候補のスコアとする．これを，繰り返して翻訳を行う.

しかし，入力文が長くなると，用いるフレーズ対の組み合わせは膨大な数になる．そのため，全ての翻訳候補を探索し，最適な出力文を決定することは困難である．そこで，現在，計算量を減らす手法として，ビームサーチ法とマルチスタック法を組み合わせた手法が一般に用いられている.

2.4.1 ビームサーチ法

ビームサーチ法は，探索の計算量を減らすために用いられる．ビームサーチ法は，翻訳候補の探索木において，翻訳確率の低い翻訳候補を枝刈りし，探索の範囲を限定する．枝刈りは “histogram pruning” と “threshold pruning” によって行う．“histogram pruning” は確率の高い翻訳候補のみを一定数残す枝刈り法である．“threshold pruning” は一定の

確率以上の翻訳候補のみを残す枝刈り法である．この 2 つの枝刈り法を用いて，探索範囲を限定する．しかし，3 からもわかるように，翻訳が進むほど，翻訳候補の確率は小さくなる．そのため，翻訳が進んだ翻訳候補と翻訳が進んでいない翻訳候補を比較したとき，翻訳が進んだ翻訳候補ほど枝刈りの対象となる可能性が高い．

2.4.2 マルチスタック法

ビームサーチ法の問題を解決するために，ビームサーチ法とマルチスタック法を組み合わせる．マルチスタック法は翻訳候補を翻訳した単語の数毎に分ける．そして，分けた翻訳候補の中で，ビームサーチ法を適用する．図 3 の例にマルチスタック・ビームサーチ法を適用した例を図 4 に示す．

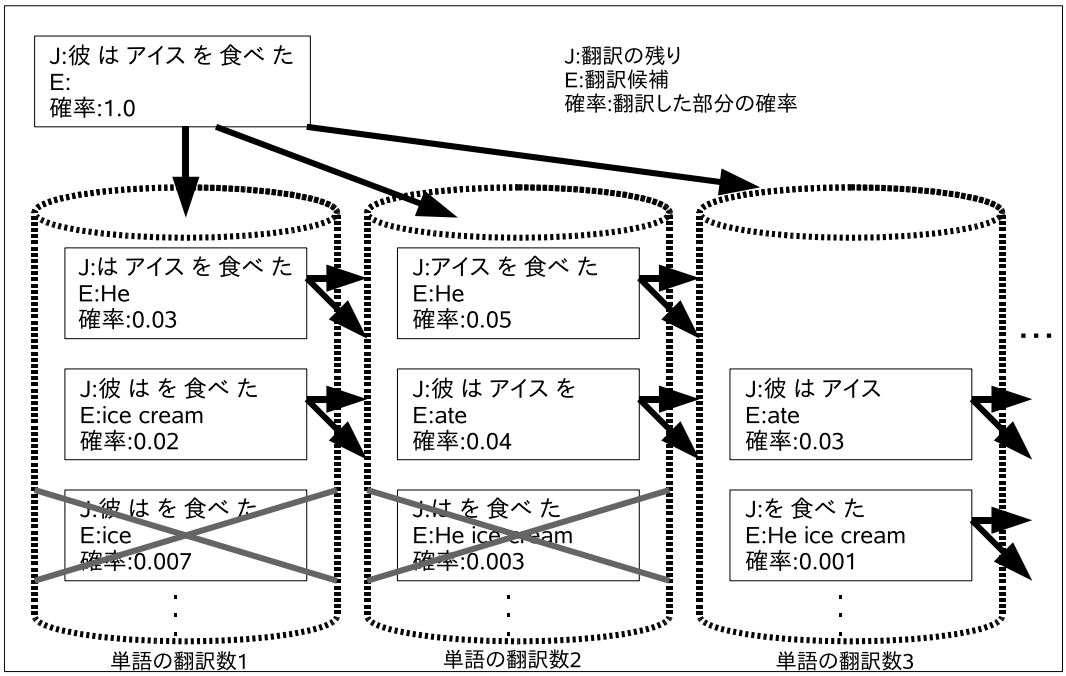

図 4: マルチスタック・ビームサーチ法の適用例 (入力文「彼はアイスを食べた」)

図 4 において，“histogram pruning” の残す翻訳候補を 2 とした場合，×のついた翻訳候補からは，探索が行われない．

2.5 デコーダのパラメータの最適化

デコーダは言語モデルや翻訳モデルに対して重みを与えることができる．例えば，言語モデルに対して高い重みを与え，翻訳モデルに低い重みを与えた場合，デコーダは言語モデルの確率 $P(e)$ を重視した出力を行う．しかし，各モデルに与える重みを，決定することは難しい．そこで，Minimum Error Rate Training(MERT)[14] という手法を用いて重みの最適化を行う．MERT は後述する自動評価法 BLEU が最大となる翻訳結果が選ばれる重み $\hat{\lambda}$ を計算する．n 個の重みの最適化は以下の式で表せる．

$$\hat{\lambda_1^n} = \underset{\lambda_1^n}{argmax} BLEU(smt(\lambda_1^n), e_{ref}) \tag{11}$$

ここで，$smt(\lambda)$ はパラメータ λ が与えられたときの，デコーダの出力文である．また，$BLEU()$ は BLEU のスコアであり，デコーダの出力文と，入力文に対してあらかじめ用意された正解文 e_{ref} から計算される．重みの最適化は，具体的には，以下の手順で行われる．

1. λ に初期値を与える
2. λ を用いてデコーディングを行い，確率の高い上位 N 文を出力する
3. 上位 N 文の中で BLEU スコアが高い文が上位にくるよう λ を最適化する
4. 重みが収束するまで 2，3 を繰り返す

3　提案手法

本章では，提案手法である文節区切りの学習データを用いたフレーズテーブルの学習について説明する．本手法の枠組みを図 5 に示す．

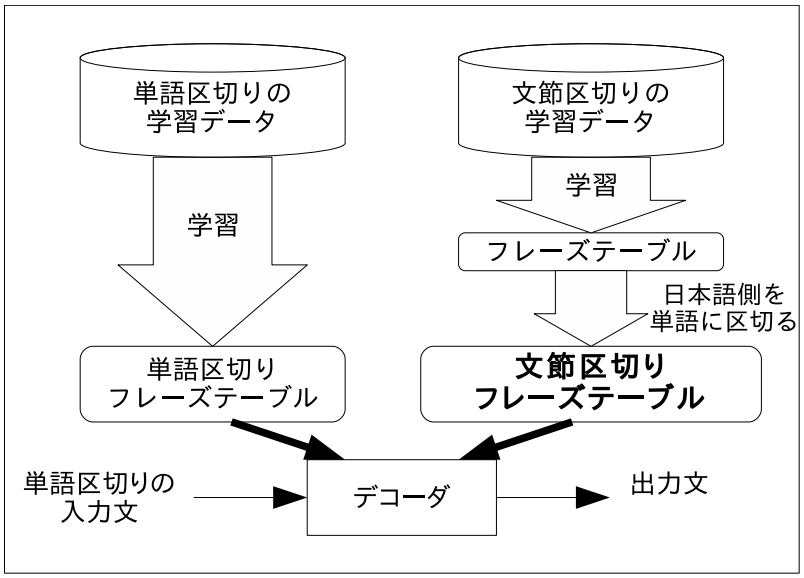

図 5: 提案手法の枠組み

日英統計翻訳において，一般に学習データの日本語文は形態素解析を用いて，単語に区切られる．そして，単語区切りの学習データを用いて，フレーズテーブルを学習する．本稿では，単語区切りの学習データから学習されるフレーズテーブルを単語区切りフレーズテーブルと呼ぶ．しかし，単語区切りフレーズテーブルは単語対応のフレーズ対や短いフレーズを持つフレーズ対が多いため，出力文は多くのフレーズ対を必要とする．そのため，並び替えの候補が膨大になり，翻訳精度が低下する．

この問題を解決するために，長い日本語フレーズを持つフレーズ対を増やすことで，出力文が利用するフレーズ対の数を減らす手法を提案する．具体的には，学習データの日本語文を文節に区切り，長い日本語フレーズを持つフレーズテーブルを学習する．例えば，図 1 で用いた学習データの日本語文を文節に区切った場合，単語の対応関係は図 6 になる．

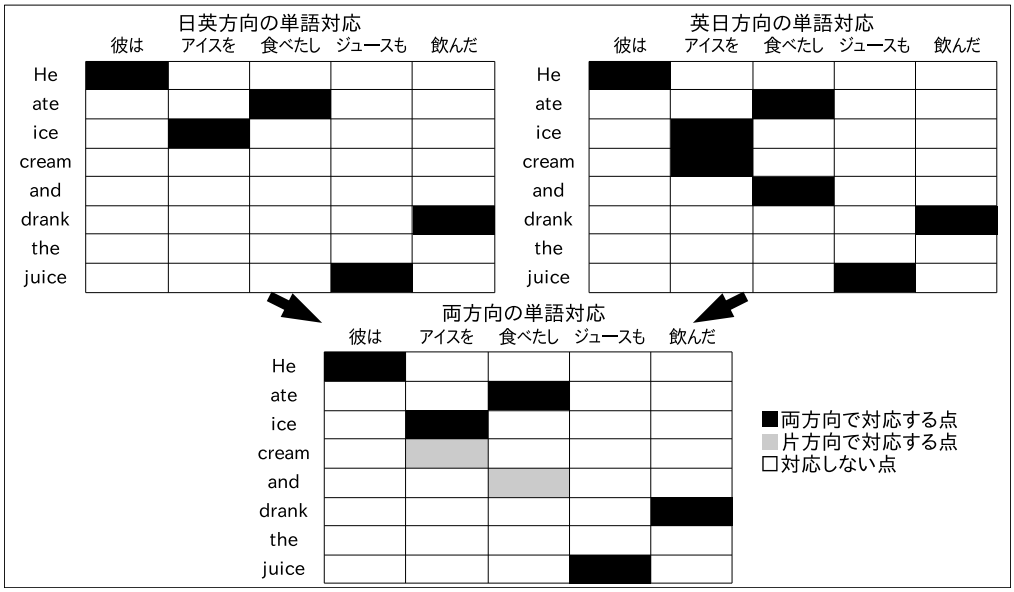

図 6: 提案手法により得られた両方向の単語の対応関係

そして，図 6 において，抽出されるフレーズ対応は表 3 になる．

表 3: 提案手法の両方向の対応関係から抽出したフレーズ対応

彼は ||| He
彼は アイスを 食べたし ||| He ate ice cream and
彼は アイスを 食べたし ジュースも 飲んだ ||| He ate ice cream and drank the juice
アイスを ||| ice cream
アイスを 食べたし ||| ate ice cream and
ジュースも ||| juice
ジュースも ||| the juice
ジュースも 飲んだ ||| drank the juice
飲んだ ||| drank
飲んだ ||| drank the

表 2 と比較して，表 3 では，“し ||| and” や “ジュース ||| juice” といった短いフレーズ対応がない．本稿では，文節に区切った学習データから学習されるフレーズテーブルを文節区切りフレーズテーブルと呼ぶ．そして，この文節区切りフレーズテーブルを従来の単語区切りフレーズテーブルと併用し，翻訳を行う．

3.1 文節区切りフレーズテーブルの生成手順

文節区切りフレーズテーブルの生成手順を以下に示す.

1. **日本語文の文節区切り**
   学習データの日本語文を文節に区切り，文節区切りの学習データを生成する.

文節区切り日本語文の例

彼-の お母さん-が ああ 若い-と-は 思わ-なかっ-た 。
ここ-で きみ-に 会お-う-と-は 夢にも 思わ-なかっ-た 。
あした 返-すから ３-，-０-０-０-円 貸し-て-ください 。
彼女-は 怠け者-で 自分-の 部屋-の 掃除-も し-ない 。
これ-は 人々-に 愛唱-さ-れ-て-いる 古い 民謡-の 一つ-です 。

2. **フレーズテーブルの生成**
   文節区切りの学習データから，フレーズテーブルを生成する.

1 から生成されたフレーズテーブルの例

道路-の ||| of the road ||| 1 0.018 0.167 0.002
読ん-だ ||| have read ||| 1 0.013 1 0.030
２-０-人-の ||| 20 people ||| 1 0.002 0.5 0.003
贅沢-に 暮らし-て-いる ||| lives in luxury ||| 1 0.0001 1 0.0097
霧-に かくれ-て 、 ||| become hidden in the mists and ||| 1 0.0002 0.5 0.0002

3. **日本語フレーズの処理**
   従来手法のフレーズテーブルと区切りを統一するために，生成されたフレーズテーブルの日本語フレーズを単語に区切る.

文節区切りフレーズテーブルの例

道路 の ||| of the road ||| 1 0.018 0.167 0.002
読ん だ ||| have read ||| 1 0.013 1 0.030
２ ０ 人 の ||| 20 people ||| 1 0.002 0.5 0.003
贅沢 に 暮らし て いる ||| lives in luxury ||| 1 0.0001 1 0.0097
霧 に かくれ て 、 ||| become hidden in the mists and ||| 1 0.0002 0.5 0.0002

3.2　2つのフレーズテーブルの併用法

2つのフレーズテーブルを併用するために，Moses のパラメータファイル “moses.ini” を編集する．編集の例を以下に示す．例の太字部分が新たに追加した部分である．

moses.ini

```
[ttable-file]
0 0 5 phrase-table1
0 0 5 phrase-table2

[weight-t]
0.5
0
0.5
0
0
0.5
0
0.5
0
0

[mapping]
0 T 0
1 T 1
```

“[ttable-file]” はフレーズテーブルのパスを指定するオプションである．例では，従来手法の “phrase-table1” と提案手法で生成した “phrase-table2” の2つのフレーズテーブルを指定している．また，“0 0 5” はフレーズテーブルが5つの確率を持つことを示している．“weight-t” は，フレーズテーブルの確率に与えられる重みを指定するオプションである．2つのフレーズテーブルはそれぞれ5つずつ確率を持つと “[ttable-file]” で指定しているため，それぞれに5つずつ，計10の重みを指定する．ここで，上の5つの値が “phrase-table1” の確率に与える重みであり，下の5つの値が “phrase-table2” の確率に与える重みである．例では，2つのフレーズテーブルに同じ値が与えられているが，パラメータチューニングにより別の値を用いるよう変更される．“[mapping]” はフレーズテーブルのマッピングを指定するオプションである．例では，2つのフレーズテーブルをそれぞれマッピングするように指定している．

4　実験環境

4.1　実験データ

実験には，辞書の例文から抽出した，単文コーパス 181,988 文 [6] と重文複文コーパス 121,719 文 [7] を用いる．単文コーパスから，Open テストデータ 9,000 文と development データ 1,000 文をランダムに抽出し，残りの 171,988 文を学習データに用いる．また，重文複文コーパスからも同様に，Open テストデータ 9,000 文と development データ 1,000 文をランダムに抽出し，残りの 111,719 文を学習データに用いる．単文コーパスと重文複文コーパス中の対訳文の例を表 4 に示す.

表 4: 単文コーパスと重文複文コーパスの例

単文コーパス	
日本語文	彼は有能な商人です。
英語文	He is an able merchant.
日本語文	ぶどう酒は葡萄より作られる。
英語文	Wine is made from grapes.
日本語文	花子は、悲しそうに俯いていた。
英語文	Hanako appeared sad and downcast.
日本語文	生徒は半径５ｃｍの円を描いた。
英語文	A student drew a circle with a radius of 5 cm.
重文複文コーパス	
日本語文	彼は偏見がありそのため信頼できなかった。
英語文	He was biased, and so unreliable.
日本語文	パチンコはわたしの好きな遊びの一つです。
英語文	Pachinko is one of my favorite pastimes.
日本語文	その鳥は山を越えて飛んでいった。
英語文	The bird winged its flight over the hills.
日本語文	急いでいて彼女に大事なことを言い忘れた。
英語文	I was in such a hurry I forgot to tell her the most important thing.

一般に，日英統計翻訳では，前処理として各コーパスの日本語文を形態素解析を用いて単語に区切る．本研究では，形態素解析器として “MeCab[8]” を用いる．また，文節区切りフレーズテーブルの学習のために，構文解析器 “CaboCha[9]” を用いて，文節区切りの学習データも生成する．また，英語文に対しては句読点の前後にスペースを入れる．一般に，英語文に対しては，大文字の小文字化を行うが，本研究では行わない.

4.2 フレーズテーブルの学習

フレーズテーブルの学習には，多くの方法がある．本研究では，“train-phrase-model.perl[10]” を用いる．このプログラムはIBM モデル 1〜5 に基づく，“GIZA++[11]” を利用している．また，フレーズテーブルを生成する際，フレーズテーブル内の日本語と英語のフレーズ中の単語数の上限値として，max-phrase-length が定義されている．例えば，max-phrase-length の値が 7 の場合，日本語か英語のいずれかのフレーズ中の単語数が，8 以上のフレーズ対は生成されない．本研究では，max-phrase-length の値として，100 を用いる．

4.3 $N-$gram モデルの学習

言語モデルは，*N*-gram モデルを用いる．*N*-gram モデルの学習には “SRILM[12]” の ngram-count を用いる．*N*-gram モデルの次数は，先行研究により 5-gram が有効であることがわかっている．そのため，本研究でも，5-gram の言語モデルを用いる．なお，本研究ではスムージングに “-ndiscount” を用いる．

4.4 デコーダのパラメータ

デコーダは “Moses[13]” を使用する．本研究では，“Moses” が用いる言語モデルの重みやフレーズテーブルの重みを最適化する．最適化は “Moses” 付属の “mert-moses.pl” を用いて行う．“mert-moses.pl” は MERT[14] を用いて，最適な翻訳結果が出力文として選ばれるように，重みを調整する．なお，提案手法は，2 つのフレーズテーブルを併用して用いる．最適化により，2 つのフレーズテーブルには異なる重みが与えられる．また，単文の翻訳には単文の development データを，重文複文の翻訳には重文複文の development データを用いる．

4.5 評価方法

出力文の評価には自動評価法である BLEU[15] と METEOR[16] を使用する．BLEU は予め用意された正解文と比較して，語順が正しい場合に高いスコアを出す．BLEU は以下の式で計算される．

$$BLEU_{score} = BP \times \sqrt[N]{\prod_{i=1}^{N} P_n} \tag{12}$$

$$P_n = \frac{\sum_i \text{出力文} i \text{と正解文} i \text{で一致した} N-gram \text{の数}}{\sum_i \text{出力文} i \text{の} N-gram \text{の数}} \tag{13}$$

ここで，P_n は出力文と正解文の N-gram の一致率を表している．BLEU はこの一致率を 1-gram から 4-gram まで計算し，その幾何平均をとる．また，出力文が正解文より短い場合，“$\sum_i$ 出力文 i の $N-gram$ の数” が小さくなり，不当にスコアが高くなる可能性がある．そこで，正解文より短い文に対するペナルティとして，BP を用いる．BP は出力文が正解文より長い場合は 1 をとなり，出力文が正解文より短い場合は 1 未満の値をとなる

METEOR は予め用意された正解文と比較して，単語属性が正しい場合に高いスコアを出す．METEOR は以下の式で計算される．

$$METEOR_{score} = F_{mean} \times (1 - Pen) \tag{14}$$

$$F_{mean} = \frac{P \times R}{\alpha \times P + (1-\alpha) \times R} \tag{15}$$

$$Pen = \gamma \times (\frac{c}{m})^{\beta} \tag{16}$$

METEOR はまず再現率 R と適合率 P に基づく F 値を求め，次に，単語の非連続性に対するペナルティとして関数 Pen を与える．ペナルティ関数 Pen において，m は出力文と正解文の単語の一致率を表す．そして，c は一致した単語を対象に，正解文と語順が同じものを 1 つのまとまりとして統合した場合の，まとまりの数を表す．そのため，出力文と正解文が同じ文であるとき c=1 となる．また，一致率の計算において，WordNet による類義語を用いて，似た意味を持つ単語は同一であると判断される．α，β，γ の値はパラメータである．本研究では，α=0.9，β=3.0，γ=0.5 の値を用いる．

両評価法とも 0 から 1 の間で評価され，出力文と正解文が同じ文であるとき 1 となり，最も良い評価である．本研究では，入力文 1 文に対して正解文 1 文を用いて評価を行う．

また，人手による評価として，対比較評価も行う．

5　実験

本章では，各フレーズテーブルを用いたときの，単文と重文複文の翻訳実験の結果について述べる．翻訳実験は従来手法である単語区切りフレーズテーブルのみを用いた実験，文節区切りフレーズテーブルのみを用いた実験，そして提案手法である単語区切りフレーズテーブルと文節区切りフレーズテーブルを併用した実験の3つを行う．また，人手による評価として，対比較実験も行う．

5.1　フレーズテーブルのフレーズ対の数

フレーズテーブルの学習には，4.1節で示した，単文171,988文と重文複文111,719文，計283,707文を用いる．学習データの単語数と，単語区切りフレーズテーブルと文節区切りフレーズテーブルのフレーズ対の数を表5に示す．

表 5: 各フレーズテーブルのフレーズ対の数 (学習データ:283,707 文)

	日本語文の単語数	英語文の単語数	フレーズ対の数
単語区切り (従来手法)	3,377,811	2,828,062	1,742,020
文節区切り	1,695,658	2,828,062	1,041,805

表5から，単語区切りフレーズテーブルと比較して，文節区切りフレーズテーブルのフレーズ対の数が，約6割であることがわかる．これは，文節区切りの学習データの文節数が，単語区切りの学習データの単語数と比較して，半分程度であることが原因である．

また，単語区切りフレーズテーブルと文節区切りフレーズテーブルには，確率は異なるが，日本語フレーズと英語フレーズの対応が同じフレーズ対が存在する．例を以下に示す．

単語区切りフレーズテーブル

１人ずつ ||| one by one ||| 0.2 0.0022 1 0.0416
２ ０ 歳 に なる ||| will be twenty years old ||| 0.3333 8.2163e-7 0.3333 1.7498e-6
お茶 の ||| tea ||| 0.0097 0.1110 0.2 0.5492
その 問題 について ||| about the problem ||| 0.4 0.0059 0.0435 0.0310
世界 的 に ||| a worldwide ||| 0.3333 0.0019 0.5 0.0006

文節区切りフレーズテーブル

１人ずつ ||| one by one ||| 0.2727 0.001 0.75 0.0416
２ ０ 歳 に なる ||| will be twenty years old ||| 1 0.0002 1 0.0002
お茶 の ||| tea ||| 0.0577 0.0280 0.75 0.45
その 問題 について ||| about the problem ||| 1 0.0009 0.0263 0.0008
世界 的 に ||| a worldwide ||| 0.5 0.0179 1 0.0181

このような，単語区切りフレーズテーブルと文節区切りフレーズテーブルにおいて，日本語フレーズと英語フレーズの対応が同じフレーズ対は，696,644 件存在する．そのため，本手法で生成したユニークなフレーズ対の数は 355,161 件である．

5.2 翻訳精度の評価

テストデータに単文と重文複文を用いて翻訳実験を行う．単文と重文複文の翻訳実験の評価結果を表 6 に示す．

表 6: 従来手法と提案手法の翻訳精度の比較 (テストデータ:各 9,000 文)

単文		
	BLEU	METEOR
従来手法	0.2015	0.4437
文節区切り	0.1522	0.3118
提案手法	**0.2058**	**0.4455**
重文複文		
	BLEU	METEOR
従来手法	0.1746	0.4034
文節区切り	0.1425	0.3066
提案手法	**0.1784**	**0.4148**

表 6 から，提案手法の翻訳精度が従来手法の翻訳精度と比較して向上していることがわかる．また，文節区切りフレーズテーブルのみを用いた結果が従来手法と比較して大きく低下している．これは，文節区切りフレーズテーブルのみでは，カバー率が低く，未知語が多く発生することが原因である．

5.3 対比較実験

表 6 の単文と重文複文の翻訳結果に対して人手による対比較実験を行う.

5.3.1 評価基準

対比較実験は従来手法の結果と提案手法の結果から，それぞれ 100 文を抽出し，どちらの文が入力文の翻訳結果として適切であるかを判断する．評価基準を以下に示す.

提案手法○ 提案手法の結果が従来手法の結果より入力文の翻訳として優れている

提案手法○の例 1(単文)
入力文　これ は 妥当 な 結論 で あろ う 。
正解文　This is an appropriate and reasonable conclusion .
従来手法　It would be appropriate decision .
提案手法　This would be appropriate conclusion .

提案手法○の例 2(単文)
入力文　私 は 2 万 円 の 謝礼 を もらっ た 。
正解文　I was given a reward of 20,000 yen for my services .
従来手法　I got a reward for two thousand yen .
提案手法　I got a reward of twenty thousand yen .

提案手法○の例 3(重文複文)
入力文　人並み の 品位 を 保と う と 試み た 。
正解文　They attempted to maintain ordinary decency .
従来手法　保と attempted a decent dignity .
提案手法　I tried to keep ordinary dignity .

提案手法○の例 4(重文複文)
入力文　余震 が 続く 中 で 我々 は 不安 な 一夜 を 過ごし た 。
正解文　We passed an uneasy night as the aftershocks continued .
従来手法　We remain in 余震 spent an uneasy night .
提案手法　Continued aftershocks in We spent an uneasy night .

提案手法× 提案手法の結果が従来手法の結果より入力文の翻訳として劣っている

提案手法×の例 1(単文)
入力文　科学 技術 は 私 たち の 社会 を 一変 さ せ た 。
正解文　Technology has transformed our society .
従来手法　Scientific technology has transformed our society .
提案手法　Our society has transformed the technology .

提案手法×の例 2(単文)
入力文　今日 は 沢山 の 客 が あっ た 。
正解文　There were many customers today .
従来手法　There was a lot of guests today .
提案手法　We had a lot of today .

提案手法×の例 3(重文複文)
入力文　決して その 行為 を 繰り返す まい と 決心 し た 。
正解文　He made a resolution never to repeat the act .
従来手法　I decided that the deed will not be repeated .
提案手法　The conduct would never have made up my mind to repeat .

提案手法×の例 4(重文複文)
入力文　僕 も それ と 同じ よう な 覚え が ある 。
正解文　I have had a similar experience myself .
従来手法　I remember that I have the same type .
提案手法　I like I remember the same as that .

差がない どちらの結果も同程度に意味が理解できる
または，同程度に意味が理解できない

差がないの例 1(単文)
入力文　あいつ の 鼻柱 を 挫 い て やれ 。
正解文　I'll humble his pride .
従来手法　He was 鼻柱 of the 挫 him .
提案手法　Let the 挫 I was 鼻柱 .

差がないの例 2(単文)
入力文　私 は いつも 路上 駐車 を する 。
正解文　I always park on the street .
従来手法　I always gives a parked bumper .
提案手法　I always road to the park .

差がないの例 3(重文複文)
入力文　環境 汚染 は ますます 深刻 な 問題 に なっ て き て いる 。
正解文　Environmental pollution is becoming an even more serious problem .
従来手法　Environmental pollution is getting increasingly a serious problem .
提案手法　The environmental pollution problem is getting more and more serious .

差がないの例 4(重文複文)
入力文　　不意 に 彼 を たずね て は 会え まい 。
正解文　　You can not possibly see him without previous appointment .
従来手法　He is not able to see me unexpectedly .
提案手法　He is not able to see me suddenly .

同一 翻訳結果が同一の文である

同一の例 1(単文)
入力文　　警察 は その 銀行 に 踏み込ん だ 。
正解文　　The police made a raid on the bank .
従来手法　The police raided the bank .
提案手法　The police raided the bank .

同一の例 2(単文)
入力文　　ひどい ショック を 受け まし た 。
正解文　　I've had a nasty shock .
従来手法　I've had a nasty shock .
提案手法　I've had a nasty shock .

同一の例 3(重文複文)
入力文　　安く 買っ て 高く 売れ 。
正解文　　Buy cheap and sell dear .
従来手法　buy cheap and sell .
提案手法　buy cheap and sell .

同一の例 4(重文複文)
入力文　　不朽 の 名声 を 得る もの は きわめて 少ない 。
正解文　　Immortality evades all but a few .
従来手法　Immortality evades all but a few .
提案手法　Immortality evades all but a few .

5.3.2 評価結果

単文の対比較実験と重文複文の対比較実験の評価結果を表 7 に示す.

表 7: 対比較実験の結果

	単文	重文複文
提案手法○	18/100	24/100
提案手法×	8/100	10/100
差がない	37/100	44/100
同一	37/100	22/100

表 7 から，単文においても，重文複文においても，提案手法○の数が提案手法×の数より多い．このことから，提案手法の翻訳結果が従来手法の翻訳結果よりも優れていることがわかる．また，単文は重文複文に比べ，同一の数が多い．これは，単文には短文が多く，長いフレーズ対の効果が少ないためである．

6　学習データの英語文に対する提案手法の適用

5 章では，学習データの日本語文を文節区切りとし，フレーズテーブルの生成を行った．本章では，学習データの英語文に対して，同等の処理を行う．そして，日本語文の文節区切りと同様に，効果の有無を調べる．

6.1　英語文に対する提案手法の適用法

日本語文の文節区切りに相当する処理として，英語構文解析器 “Apple Pie Parser[17]” を用いて学習データの英語文の単語をフレーズ単位に結合する．フレーズテーブルの生成手順を以下に示す．

1. **英語文をフレーズ単位に結合**
   学習データの英語文をフレーズ単位に統合する．

   フレーズ単位の結合の例

   He is an-able-merchant .
   The-bird winged its-flight over the-hills .
   His-room is approached by a-flight of steps .
   I could not recognize the-girl from the-vague-description of her you gave me .
   Father loves using his-new-word-processor .

2. **フレーズテーブルの生成**
   フレーズ単位に統合された学習データから，フレーズテーブルを生成する．

   1 から生成されたフレーズテーブル

   有能 な 商人 ||| an-able-merchant ||| 1 0.028 1 0.003
   丘 を ||| the-hill ||| 0.33 0.25 1 0.15
   丘 を 越え て ||| over the-hill ||| 0.5 0.004 0.33 0.005
   贅肉 が つき ||| put on superfluous-flesh ||| 1 0.0009 1 0.0013
   鋏 ||| a-pair of scissors ||| 1 0.1801 0.5 0.003

3. **英語フレーズの処理**
   従来手法のフレーズテーブルと区切りを統一するために，生成されたフレーズテーブルの英語フレーズを単語に区切る．

文節区切りフレーズテーブル

有能 な 商人 ||| an able merchant ||| 1 0.028 1 0.003
丘 を ||| the hill ||| 0.33 0.25 1 0.15
丘 を 越え て ||| over the hill ||| 0.5 0.004 0.33 0.005
贅肉 が つき ||| put on superfluous flesh ||| 1 0.0009 1 0.0013
鋏 ||| a pair of scissors ||| 1 0.1801 0.5 0.003

生成されたフレーズテーブルは，5 章と同様に，従来手法のフレーズテーブルと併用し翻訳を行う．また，実験には以下の 2 種類の学習データから生成されたフレーズテーブルを用いる．

1. **日本語文:単語区切り，英語文:フレーズ単位**

   単語区切りの日本語文とフレーズ単位の英語文からフレーズテーブルを生成する．これは，一方の言語にのみ処理を行った場合の効果を調べるために用いる．

2. **日本語文:文節区切り，英語文:フレーズ単位**

   文節区切りの日本語文とフレーズ単位の英語文からフレーズテーブルを生成する．これは，両言語に処理を行った場合の効果を調べるために用いる．このとき，日本語フレーズは文節区切りであるため，手順 3 で単語に区切る．

6.2 フレーズテーブルのフレーズ対の数

フレーズテーブルの学習には，5 章と同様に，単文 171,988 文と重文複文 111,719 文を用いる．学習データの単語数と，生成されたフレーズテーブルのフレーズ対の数を表 8 に示す．

表 8: 英語文をフレーズ単位に統合した場合の各フレーズテーブルのフレーズ対の数 (学習データ:283,707 文)

	日本語文の単語数	英語文の単語数	フレーズ対の数
日:単語，英:単語 (従来手法)	3,377,811	2,828,062	1,742,020
日:単語，英:フレーズ	3,377,811	2,215,378	1,147,845
日:文節，英:フレーズ	1,695,658	2,215,378	798,124

表 8 から，従来手法のフレーズテーブルと比較して，単語区切りの日本語文とフレーズ単位に統合された英語文から生成されたフレーズテーブルのフレーズ対の数は約 7 割，文節区切りの日本語文とフレーズ単位に統合された英語文から生成されたフレーズテーブルのフレーズ対の数は約 5 割であることがわかる．

また，文節区切りフレーズテーブルと同様に，本章で生成したフレーズテーブルにも，従来手法のフレーズテーブルのフレーズ対と，確率は異なるが，日本語フレーズと英語フレーズの対応が同じフレーズ対が存在する．これは，単語区切りの日本語文とフレーズ単位に統合された英語文から生成されたフレーズテーブルでは 718,893 件あり，文節区切りの日本語文とフレーズ単位に統合された英語文から生成されたフレーズテーブルでは 546,054 件ある．

6.3 翻訳精度の評価

テストデータに単文と重文複文を用いて翻訳実験を行う．単文と重文複文の翻訳結果の評価結果を表 9 に示す．

表 9: 英語文をフレーズ単位に統合した場合の翻訳精度の比較 (テストデータ:各 9,000 文)

単文		
	BLEU	METEOR
日:単語，英:単語 (従来手法)	0.2015	0.4437
従来手法+日:単語，英:フレーズ	**0.2063**	**0.4493**
従来手法+日:文節，英:フレーズ	0.2045	0.4444
重文複文		
	BLEU	METEOR
日:単語，英:単語 (従来手法)	0.1746	0.4034
従来手法+日:単語，英:フレーズ	0.1760	**0.4072**
従来手法+日:文節，英:フレーズ	**0.1783**	0.4068

表 9 から，提案手法が英語文に対しても，有効であることがわかる．また，両言語に対して提案手法を適用した場合，重文複文の翻訳精度は高いが，単文の翻訳精度はやや低い．

6.4 単文の翻訳結果の例

表 9 の単文の翻訳結果の例を以下に示す.

- **「日:単語，英:フレーズ」の追加で翻訳精度が向上**

単文の翻訳結果の例 1:「日:単語，英:フレーズ」の追加で翻訳精度が向上

入力文	彼 の 説教 は 果てしなく 長く 感じ られ た 。
正解文	His sermon seemed interminably long .
従来手法	He felt that of 果てしなく a long time .
日:単語，英:フレーズ	His sermon was felt endlessly long .

単文の翻訳結果の例 2:「日:単語，英:フレーズ」の追加で翻訳精度が向上

入力文	わたし は 神田 さん を 個人 的 に 知っ て い ます 。
正解文	I personally know Kanda .
従来手法	Mr . I am personally known to Kanda .
日:単語，英:フレーズ	I know I personally , Kanda .

- **「日:単語，英:フレーズ」の追加で翻訳精度が低下**

単文の翻訳結果の例 3:「日:単語，英:フレーズ」の追加で翻訳精度が低下

入力文	それ は 午後 の 3 時 と 4 時の間 に 起こっ た 。
正解文	It happened between three and four o'clock in the afternoon .
従来手法	It happened between three and four o'clock in the afternoon .
日:単語，英:フレーズ	It happened between three and four the afternoon .

単文の翻訳結果の例 4:「日:単語，英:フレーズ」の追加で翻訳精度が低下

入力文	苦しい 立場 に ある 。
正解文	He is in a painful position .
従来手法	He is in a difficult position .
日:単語，英:フレーズ	is in a difficult position .

- **「日:単語，英:フレーズ」の追加で翻訳精度に差がない**

単文の翻訳結果の例 5:「日:単語，英:フレーズ」の追加で翻訳精度に差がない

入力文	彼 は 7 0 の 坂 を 越え た 。
正解文	He is over seventy .
従来手法	He crossed the right side of 70 .
日:単語，英:フレーズ	He crossed the slope of 70 .

- **「日:文節，英:フレーズ」の追加で翻訳精度が向上**

単文の翻訳結果の例 6:「日:文節，英:フレーズ」の追加で翻訳精度が向上

入力文	日本 は 多く の 貿易 上 の 利点 を 享受 し て いる 。
正解文	Japan enjoys many trade advantages .
従来手法	Many Japanese enjoy the benefits of trade .
日:文節，英:フレーズ	Japan enjoy the benefits of trade .

単文の翻訳結果の例 7:「日:文節，英:フレーズ」の追加で翻訳精度が向上

入力文	彼 は 普通 の 作家 と 違う 。
正解文	He is different from ordinary writers .
従来手法	He usually does not agree with the author .
日:文節，英:フレーズ	He is different from ordinary writer .

- **「日:文節，英:フレーズ」の追加で翻訳精度が低下**

単文の翻訳結果の例 8:「日:文節，英:フレーズ」の追加で翻訳精度が低下

入力文	スーパーコンピュータ は 新しい 可能 性 の 世界 を 開く 。
正解文	Supercomputers open up new worlds of possibility .
従来手法	The supercomputer opens a new world of possibility .
日:文節，英:フレーズ	The possibility of supercomputers open a new world .

単文の翻訳結果の例 9:「日:文節，英:フレーズ」の追加で翻訳精度が低下

入力文	この 虫歯 は 取る べき だ 。
正解文	This decayed tooth should come out .
従来手法	This tooth , you should take .
日:文節，英:フレーズ	This should take a tooth .

- **「日:文節，英:フレーズ」の追加で翻訳精度に差がない**

単文の翻訳結果の例 10:「日:文節，英:フレーズ」の追加で翻訳精度に差がない

入力文	彼女 は 男性 と の 交際 を 望ん だ 。
正解文	She craved male companionship .
従来手法	She is a man with the entertainment for it .
日:文節，英:フレーズ	She is for my acquaintance with men .

6.5 重文複文の翻訳結果の例

表 9 の重文複文の翻訳結果の例を以下に示す.

- **「日:単語，英:フレーズ」の追加で翻訳精度が向上**

重文複文の翻訳結果の例 1:「日:単語，英:フレーズ」の追加で翻訳精度が向上

入力文	あの 人 は 愛情 の 深い 人 だ 。
正解文	He is a man of strong affection .
従来手法	He is a man of affection .
日:単語，英:フレーズ	That person is a man of deep affection .

重文複文の翻訳結果の例 2:「日:単語，英:フレーズ」の追加で翻訳精度が向上

入力文	そう し て くれれ ば 大 助かり です 。
正解文	That will be of great help to me .
従来手法	I will keep a great help to do so .
日:単語，英:フレーズ	It is a great help you would do it .

- **「日:単語，英:フレーズ」の追加で翻訳精度が低下**

重文複文の翻訳結果の例 3:「日:単語，英:フレーズ」の追加で翻訳精度が低下

入力文	１ ０ ０ 年 後 の 未来 を 予測 する こと は 困難 だ 。
正解文	It is difficult to predict what the future will be like one hundred years from now .
従来手法	It is difficult to predict the future of 100 years later .
日:単語，英:フレーズ	It is difficult to predict the future of 100 years .

重文複文の翻訳結果の例 4:「日:単語，英:フレーズ」の追加で翻訳精度が低下

入力文	彼 は 奥さん を 捨て て 、 他 の 女 の 人 と 一緒 に 逃げ て 行っ た 。
正解文	He left his wife and ran away with another woman .
従来手法	He abandoned his wife and ran away with another woman .
日:単語，英:フレーズ	He ran away with his wife away and other woman .

- **「日:単語，英:フレーズ」の追加で翻訳精度に差がない**

重文複文の翻訳結果の例 5:「日:単語，英:フレーズ」の追加で翻訳精度に差がない

入力文	生 から 死 へ の 移行 の 苦痛 を やわらげる ため に 私 たち に できる こと を し ます 。
正解文	We do what we can to ease the transition from life to death .
従来手法	We can shift to make a fresh from dulls pain of death in order .
日:単語，英:フレーズ	I will give you a dulls pain of death , we can do to shift to life .

- **「日:文節，英:フレーズ」の追加で翻訳精度が向上**

重文複文の翻訳結果の例 6:「日:文節，英:フレーズ」の追加で翻訳精度が向上

入力文	侮辱 さ れ て 黙っ て いる よう な 意気地無し だ 。
正解文	He has no spirit to resent an insult .
従来手法	It is silence 意気地無し such insult .
日:文節，英:フレーズ	has no spirit to resent an insult .

重文複文の翻訳結果の例 7:「日:文節，英:フレーズ」の追加で翻訳精度が向上

入力文	私 は どんな 結果 に なっ て も 責任 を 負い ませ ん 。
正解文	I am unaccountable for any result .
従来手法	I don't have any claim a result , the responsibility .
日:文節，英:フレーズ	I claim is not responsible for any results .

- **「日:文節，英:フレーズ」の追加で翻訳精度が低下**

重文複文の翻訳結果の例 8:「日:文節，英:フレーズ」の追加で翻訳精度が低下

入力文	いろいろ 考え た が よく 理解 でき なかっ た 。
正解文	I thought it over in a variety of ways but could not understand it well .
従来手法	I could not understand well of thought .
日:文節，英:フレーズ	There are a lot of thought I could .

重文複文の翻訳結果の例 9:「日:文節，英:フレーズ」の追加で翻訳精度が低下

入力文	大人 に なっ たら 音楽家 に なり たい 。
正解文	I want to be a musician when I grow up .
従来手法	I want to become musicians , man .
日:文節，英:フレーズ	If you want to become a great man to a musician .

- **「日:文節，英:フレーズ」の追加で翻訳精度に差がない**

重文複文の翻訳結果の例 10:「日:文節，英:フレーズ」の追加で翻訳精度に差がない

入力文	彼 は 酒 を 飲み ながら その 話 を し た 。
正解文	He told the story over his cup .
従来手法	He made a drinking , the story .
日:文節，英:フレーズ	He is drinking , the story .

7　全てのフレーズテーブルを用いた実験

5 章と 6 章では，従来手法である単語区切りフレーズテーブルと，提案手法により生成された 3 つのフレーズテーブルをそれぞれ併用し翻訳実験を行った．提案手法により生成されたフレーズテーブルは，それぞれ別のフレーズ対を含んでいるため，3 つ全てを用いることでさらに翻訳精度が向上する可能性がある. そこで，本章では，4 つのフレーズテーブルを併用した実験を行う．

7.1　フレーズテーブルのフレーズ対の数

従来手法のフレーズテーブルと 5 章と 6 章で生成された 3 つのフレーズテーブルを併用する．フレーズテーブルのフレーズ対の数を 10 に示す．

表 10: 英語文をフレーズ単位に統合した場合の各フレーズテーブルのフレーズ対の数 (学習データ:283,707 文)

	日本語文の単語数	英語文の単語数	フレーズ対の数
日:単語，英:単語 (従来手法)	3,377,811	2,828,062	1,742,020
日:文節，英:単語 (文節区切り)	1,695,658	2,828,062	1,041,805
日:単語，英:フレーズ	3,377,811	2,215,378	1,147,845
日:文節，英:フレーズ	1,695,658	2,215,378	798,124
4 つを併用			4,739,794

4 つのフレーズテーブルにおいて，確率は異なるが，日本語フレーズと英語フレーズの対応が同じフレーズ対は 2,220,356 件存在する．そのため，4 つのフレーズテーブルを併用した場合のユニークなフレーズ対の数は 2,519,438 件である．

7.2 翻訳実験

テストデータは単文と重文複文を用いる．単文と重文複文の翻訳実験の評価結果を表11に示す．

表 11: 全てのフレーズテーブルを用いた場合の翻訳精度の比較 (テストデータ:各 9,000 文)

単文		
	BLEU	METEOR
従来手法	0.2015	0.4437
4 つを併用	0.2086	0.4473
重文複文		
	BLEU	METEOR
従来手法	0.1746	0.4034
4 つを併用	0.1797	0.4099

表11から，単文と重文複文共に，翻訳精度が向上していることがわかる．また，BLEUスコアは5章と6章の結果と比較して，最も高い値となった．

本研究では，語のまとまりとして，日本語文には文節を，英語文にはフレーズを用いて長いフレーズを生成した．語のまとまりには，文節やフレーズ以外にも，句や節といった区切りがある．それらを本手法と同様に利用することで，さらに翻訳精度が向上すると考えている．

7.3 単文の翻訳結果の例

表 11 の単文の翻訳結果の例を以下に示す.

- **従来手法と比較して翻訳精度が向上**

  単文の翻訳結果の例 1:翻訳精度向上

  入力文　　彼 は 普通 の 作家 と 違う 。
  正解文　　He is different from ordinary writers .
  従来手法　He usually does not agree with the author .
  4 つを併用　He is different from ordinary writer .

  単文の翻訳結果の例 2:翻訳精度向上

  入力文　　新た に 発表 さ れる 小説 の 数 は きわめて 多い 。
  正解文　　The output of new novels is vast .
  従来手法　new novels is vast .
  4 つを併用　The output of new novels is vast .

  単文の翻訳結果の例 3:翻訳精度向上

  入力文　　われわれ は まだ 正式 に は 結婚 し て い ませ ん 。
  正解文　　We are not legally married yet .
  従来手法　We have not yet been officially married .
  4 つを併用　We still have not officially married .

- **従来手法と比較して翻訳精度が低下**

  単文の翻訳結果の例 4:翻訳精度低下

  入力文　　若い 娘 が 電話 に 出 た 。
  正解文　　A young girl answered the call .
  従来手法　The young girl answered .
  4 つを併用　A telephone call came to a young girl .

  単文の翻訳結果の例 5:翻訳精度低下

  入力文　　彼 は 彼女 の 要求 について は ひと言 も 触れ なかっ た 。
  正解文　　He made no mention of her request .
  従来手法　He did not say a word about her request .
  4 つを併用　He demands on her He never mentioned .

- **従来手法と比較して翻訳精度に差がない**

  単文の翻訳結果の例 6:翻訳精度に差がない

  入力文　　この 本 は 造本 が しっかり し て いる 。
  正解文　　This book is solidly bound .
  従来手法　This book is 造本 is sound .
  4 つを併用　This book is sound 造本 .

7.4 重文複文の翻訳結果の例

表 11 の重文複文の翻訳結果の例を以下に示す.

- **従来手法と比較して翻訳精度が向上**

重文複文の翻訳結果の例 1:翻訳精度向上

入力文　彼 は 手 を 伸ばし て 受話器 を 取っ た 。
正解文　He reached out and picked up the telephone receiver .
従来手法　He reached up and picked it up .
4 つを併用　He reached up and picked up the receiver .

重文複文の翻訳結果の例 2:翻訳精度向上

入力文　普段 は 忙しい から 、 土曜日 の 午後 か 日曜日 に 来 て 下さい 。
正解文　I am usually busy , so please come on Saturday afternoon or on Sunday .
従来手法　I usually or on Sunday , so please come on Saturday afternoon .
4 つを併用　I am so busy usually come , on Saturday afternoon or on Sunday .

重文複文の翻訳結果の例 3:翻訳精度向上

入力文　日本 に 来 た うえ は 、 一 日 も 早く 日本 の 習慣 に 慣れる つもり だ 。
正解文　Having come to Japan , I intend to become familiar with Japanese customs as early as possible .
従来手法　I will soon get used to Japanese customs came to Japan , and day .
4 つを併用　I came to Japan , and I will soon get used to Japanese customs .

- **従来手法と比較して翻訳精度が低下**

重文複文の翻訳結果の例 4:翻訳精度低下

入力文　私 は 自分 が ひどい 仕打ち を し た こと は わかっ て いる 。
正解文　I know I was a brute .
従来手法　I know that I gave her a raw deal .
4 つを併用　I know he has been a raw deal .

重文複文の翻訳結果の例 5:翻訳精度低下

入力文　彼 は 自分 の し た こと に 後悔 の 念 を 感じ なかっ た 。
正解文　He felt no regret for what he had done .
従来手法　He did not feel the regrets for what he had done .
4 つを併用　He did not feel regret to a sense of what he had done .

- **従来手法と比較して翻訳精度に差がない**

重文複文の翻訳結果の例 6:翻訳精度に差がない

入力文	大変 申し訳 ござい ませ ん が 、 当行 は 外国 為替 業務 を 取り 扱っ て おり ませ ん 。
正解文	We are very sorry but we do not handle foreign exchange transactions .
従来手法	I'm sorry , but we are not very handle foreign exchange business .
4 つを併用	I am sorry , I am very We handle foreign exchange business .

8　考察

8.1　提案手法において翻訳精度が向上した文の解析

表 7 において，提案手法で翻訳精度が向上した 42 文 (単文:18 文，重文複文:24 文) に対して，翻訳精度が向上した理由を解析した．解析結果の結果を表 12 に示す．

表 12: 翻訳精度が向上した理由

翻訳精度向上の理由	単文 18 文	重文複文 24 文	合計
翻訳に用いるフレーズ対の減少	7	9	16
適切なフレーズ対の増加	6	11	17
未知語の減少	2	3	5
単語区切りフレーズテーブルのみ	3	1	4

表 12 において,「単語区切りフレーズテーブルのみ」は,「文節区切りフレーズテーブルのフレーズ対を用いずに翻訳を行い，精度が向上した文」である．本実験において，従来手法が用いる各モデルの重みと，提案手法が用いる各モデルの重みは異なる．そのため，文節区切りフレーズテーブルのフレーズ対が利用されない場合も翻訳結果が異なる．

「翻訳に用いるフレーズ対の減少」は 8.1.1 項で,「適切なフレーズ対の増加」は 8.1.2 項で，「未知語の減少 」は 8.1.3 項で，それぞれ述べる．

8.1.1 翻訳に用いるフレーズ対の減少

「翻訳に用いるフレーズ対の減少」は,「長いフレーズ対により出力文が利用するフレーズ対を減り，翻訳精度が向上した文」である．提案手法において翻訳精度が向上した 42 文中に，単文の翻訳結果で 7 文，重文複文の翻訳結果で 9 文あった．例を以下に示す.

入力文	山 の 懐 に 小さな 村 が あっ た 。
従来手法	There was a small mountain village of his inside pocket .
提案手法	There was a small village bosom of the mountain .

ここで，従来手法の出力文が用いたフレーズ対と提案手法の出力文が用いたフレーズ対を表 13 に示す.

表 13: 出力文が用いたフレーズ対 (翻訳に用いるフレーズ対の減少)

従来手法が用いたフレーズ対	提案手法が用いたフレーズ対
山 ||| mountain	**山 の ||| of the mountain**
の ||| of	**懐 に ||| bosom**
懐 ||| his inside pocket	小さな 村 ||| a small village
に 小さな ||| a small	が あっ た ||| There was
村 ||| village	。 ||| .
が あっ た ||| There was	
。 ||| .	

従来手法において，“山” と “の” は異なるフレーズ対で翻訳されている．そのため，出力文は適切な並び替えを行えず，翻訳精度が低下した.

一方で，提案手法の出力文では，1 個のフレーズ対 “山 の ||| of the mountain” を用いて翻訳されている．また，従来手法で不適切な訳がされていた “懐” が，フレーズ対 “懐 に ||| bosom” を用いて適切に翻訳されている．さらに，従来手法は 7 個のフレーズ対を用いて翻訳しているが，提案手法は 5 個のフレーズ対を用いて翻訳を行っており，並び替えの候補数が少ない．そのため，従来手法と比較して翻訳精度が向上した.

本手法の目的は，長いフレーズ対を増やすことで，フレーズの並び替えの候補数を減らし，翻訳精度を向上させることである.「翻訳に用いるフレーズ対の減少」が翻訳精度が向上した大きな理由の 1 つであったことから，本研究の目的が達成できたと考えている.

8.1.2 適切なフレーズ対の増加

「適切なフレーズ対の増加」は,「従来手法において，不適切なフレーズ対で翻訳されていたフレーズが，提案手法によって改善され，翻訳精度が向上した文」である．提案手法において翻訳精度が向上した 42 文中に，単文の翻訳結果で 6 文，重文複文の翻訳結果で 11 文あった．例を以下に示す.

入力文	あなた 方 は 収入 に 応じ て 暮らす の が よい 。
従来手法	Living for you to meet a good income .
提案手法	You ought to live in accordance with the income .

ここで，従来手法の出力文が用いたフレーズ対と提案手法の出力文が用いたフレーズ対を表 14 に示す.

表 14: 出力文が用いたフレーズ対 (適切なフレーズ対の増加)

従来手法が用いたフレーズ対	提案手法が用いたフレーズ対
あなた ||| you	**あなた 方 は ||| You**
方 ||| to	収入 ||| income
は 収入 ||| income	に 応じ て ||| in accordance with
に 応じ て ||| meet	暮らす ||| to live
暮らす の ||| Living for	の ||| the
が よい ||| a good	が よい ||| ought
。 ||| .	。 ||| .

従来手法の出力文では，“あなた” と接尾辞である “方” が異なるフレーズ対で翻訳されている．しかし，日本語側の “あなた” の接尾辞 “方” に対応する単語は英語には存在しない．統計翻訳では，言語間の単語の対応の差をフレーズ対とその並び替えで補うが，本出力文では補いきれず，適切な翻訳ができなかった.

一方で，提案手法の出力文では，文節区切りフレーズテーブルのフレーズ対 “あなた 方 は ||| You” により，“あなた” と “方” が同じフレーズ対で翻訳されている．そのため，フレーズ対により言語間の単語の対応の差を補うことができ，出力文全体の精度が向上した.

このことから，長いフレーズ対，特に文節を区切りに用いたフレーズ対には，助詞や接尾辞といった日本語と英語の文法構造の違いを補う効果があると考えている.

8.1.3 未知語の減少

「未知語の減少」は,「従来手法の結果には未知語があったが，提案手法の結果にはない文」である．提案手法において翻訳精度が向上した 42 文中に，単文の翻訳結果で 2 文，重文複文の翻訳結果で 3 文あった．例を以下に示す.

- **単文の翻訳**

  未知語の減少による翻訳精度向上の例 1:単文

  入力文　　物価 は 依然 暴騰 し て いる 。
  正解文　　Prices are still soaring upward .
  従来手法　Prices are still 暴騰 .
  提案手法　Prices are still skyrocketing .

- **重文複文**

  未知語の減少による翻訳精度向上の例 2:重文複文

  入力文　　人並みの品位を保とうと試みた。
  正解文　　They attempted to maintain ordinary decency .
  従来手法　保と attempted a decent dignity .
  提案手法　I tried to keep ordinary dignity .

学習データの区切りを変えたとき，フレーズテーブルの学習時の単語の対応も変わる．そのため，従来手法において，学習時にフレーズ 対応が “矛盾する” と判断され，フレーズテーブルに生成されなかった語彙が，別の区切りのフレーズテーブルでは “矛盾しない” と判断され，生成できることがある．提案手法では，語彙のカバー範囲が異なる 2 つのフレーズテーブルを併用しているため，語彙のカバー率が向上し，出力文の未知語が減少した.

8.2 提案手法において翻訳精度が低下した文の解析

表 7 において，提案手法で翻訳精度が低下した 18 文 (単文:8 文，重文複文:10 文) に対して，翻訳精度が低下した理由を解析した．解析結果の結果を表 15 に示す．

表 15: 翻訳精度が低下した理由

翻訳精度低下の理由	単文 8 文	重文複文 10 文	合計
単語区切りフレーズ対の問題	1	3	4
文節区切りフレーズ対の問題	2	1	3
組合せの問題	2	3	5
単語区切りフレーズテーブルのみ	3	3	6

表 15 において,「単語区切りフレーズテーブルのみ」は,「文節区切りフレーズテーブルのフレーズ対を用いずに翻訳を行い，精度が低下した文」である．本実験において，従来手法が用いる各モデルの重みと，提案手法が用いる各モデルの重みは異なる．そのため，文節区切りフレーズテーブルのフレーズ対が利用されない場合も翻訳結果が異なる．

「単語区切りフレーズ対の問題」は 8.2.1 項で,「文節区切りフレーズ対の問題」は 8.2.2 項で，「組合せの問題」は 8.2.3 項で，それぞれ述べる．

8.2.1 単語区切りフレーズ対の問題

「単語区切りフレーズ対の問題」は,「提案手法が用いた文節区切りフレーズテーブルのフレーズ対は適切であるが，単語区切りフレーズテーブルのフレーズ対が不適切であるため，翻訳精度が低下した文」である．提案手法において翻訳精度が低下した 18 文中に，単文の翻訳結果で 1 文，重文複文の翻訳結果で 3 文あった．例を以下に示す.

入力文	僕 も それ と 同じ よう な 覚え が ある 。
従来手法	I remember that I have the same type .
提案手法	I like I remember the same as that .

ここで，従来手法の出力文が用いたフレーズ対と提案手法の出力文が用いたフレーズ対を表 16 に示す.

表 16: 出力文が用いたフレーズ対 (単語区切りフレーズ対の問題)

従来手法が用いたフレーズ対	提案手法が用いたフレーズ対
僕 も ||| I have	**僕 も ||| I**
それ と ||| that	それ と 同じ ||| the same as that
同じ よう な ||| the same type	よう な ||| like
覚え が ある ||| I remenber	覚え が ある ||| I remenber
。 ||| .	。 ||| .

提案手法において，適用された文節区切りフレーズテーブルのフレーズ対 “僕も ||| I” は不適切とはいえない．しかし，単語区切りフレーズテーブルのフレーズ対 “よう な ||| like” は，本出力文においては，不適切である．本入力文において，“同じ”，“よう”，“な” は 1 個のフレーズ対で翻訳されることが好ましい．文節区切りフレーズテーブルには “同じ よう な ||| the same” といったフレーズ対が存在するが，本出力文では用いられず，異なるフレーズ対で翻訳され，翻訳精度が低下した．また，係り受けを考えると ‘同じ”，“よう”，“な”，“覚え” が 1 個のフレーズ対で翻訳されることが好ましいが，本手法では学習データから生成されなかった.

このことから，提案手法では語のまとまりとして文節を用いて学習データを統合したが，語の意味や係り受けを考慮した統合を行い学習データとして用いることで，適切なまとまりを持つフレーズ対を学習できると考えている.

8.2.2 文節区切りフレーズ対の問題

「文節区切りフレーズ対の問題」は,「翻訳に用いた文節区切りフレーズテーブルのフレーズ対が不適切であるため，翻訳精度が低下した文」である．提案手法において翻訳精度が低下した 18 文中に，単文の翻訳結果で 2 文，重文複文の翻訳結果で 1 文あった．例を以下に示す．

入力文	米国 は ４つ の 時間 帯 に またがっ て いる 。
従来手法	The United States is straddling the range of four hours .
提案手法	I can The United States is straddling the four hours .

ここで，従来手法の出力文が用いたフレーズ対と提案手法の出力文が用いたフレーズ対を表 17 に示す．

表 17: 出力文が用いたフレーズ対 (文節区切りフレーズ対の問題)

従来手法が用いたフレーズ対	提案手法が用いたフレーズ対
米国 は ||| The United States	米国 は ||| The United States
４つ の ||| of four	４つ の ||| the four
時間 ||| hours	時間 ||| hours
帯 ||| range	**帯 に ||| I can**
に ||| the	またがっ て いる |||is straddling
またがっ て いる |||is straddling	
。 ||| .	

提案手法において，文節区切りフレーズテーブルのフレーズ対 “帯 に ||| I can” は不適切である．

フレーズテーブルの学習において，日本語の単語と英語の単語の対応関係の精度は，単語の出現頻度に依存する．文節区切りの学習データは，助詞や接尾辞を統合しているため，単語の出現頻度は低下する．例えば，統合した単語 “山-へ” と “山-に” は別の単語として扱われる．そのため，出現頻度の少ない単語が多く，学習の精度が低下する可能性がある．

統計翻訳では，フレーズ対の確率はフレーズの出現頻度を考慮していない．本手法の文節区切りの学習データのように各単語の出現頻度が少ない学習データを用いる場合，学習時にフレーズの出現頻度を考慮した確率付けを行う必要があると考えている．

8.2.3 組合せの問題

「組合せの問題」は,「出力文が用いたフレーズ対は適切であるが，うまく並び替えを行うことができず，翻訳精度が低下した文」である．提案手法において翻訳精度が低下した 18 文中に，単文の翻訳結果で 2 文，重文複文の翻訳結果で 3 文あった．例を以下に示す.

入力文	どこ か の 図書館 で 数 か月 懸命 に 勉強 する こと が 必要 だ 。
従来手法	I need some months in the library to study hard .
提案手法	Some few months in the library to study hard necessary .

ここで，従来手法の出力文が用いたフレーズ対と提案手法の出力文が用いたフレーズ対を表 18 に示す.

表 18: 出力文が用いたフレーズ対 (単語区切りフレーズ対の問題)

従来手法が用いたフレーズ対	提案手法が用いたフレーズ対
どこ ||| I	**どこ か の ||| Some**
か の ||| some	図書館 で ||| in the library
図書館 で ||| in the library	数 か月 ||| few months
数 か月 ||| months	懸命 に 勉強 する ||| to study hard
懸命 に 勉強 する ||| to study hard	こと が 必要 だ 。 ||| necessary .
こと が 必要 だ ||| need	
。 ||| .	

従来手法において，フレーズ対 “どこ ||| I” は不適切である．しかし，入力文には主語がなく，“どこ ||| I” を用いて主語を作ることで，出力文の翻訳精度が向上している．

一方で，提案手法では，文節区切りフレーズテーブルのフレーズ対 “どこ か の ||| Some” を用いている．このフレーズ対は適切であるといえるが，他のフレーズ対でも主語が生成されず，翻訳精度が低下する原因となっている．

このことから，主語のない文に対して，任意主語を付与する必要があると考えている．

8.3 単文と重文複文の翻訳精度の傾向の違い

表 6 と表 9 において，単文の翻訳は併用したフレーズテーブルのフレーズ対の数が多いほど，翻訳精度が高くなる傾向があった．しかし．重文複文の翻訳は，フレーズ対の数とは無関係に，併用したフレーズテーブルの学習データの日本語文が文節区切りであった場合，翻訳精度が高い傾向があった．この原因について，次のように考えている．

日英統計翻訳は日本語をフレーズ対を用いて英語に変換し，並び替えにより英語文に翻訳する．単文は文法構造が単純であり，短文が多いため，翻訳に用いるフレーズ対が少なく，並び替えの候補は少ない．そのため，日本語フレーズの長さとは無関係に，フレーズ対の数に比例して翻訳精度が向上する傾向があった．

しかし，重文複文は文法構造が複雑であり，長文が多いため，翻訳に用いるフレーズ対が多く，並び替えの候補が膨大になる．さらも単文と比較して，日本語と英語の文法構造の差の影響が大きい.「日本語文を単語区切り，英語文をフレーズ区切りとした」学習データから生成されたフレーズテーブルは日本語フレーズは短いため，翻訳に用いるフレーズ対を減らす効果が小さく，文法構造の違いも補えない．そのため，翻訳精度の向上は小さかった．しかし,「日本語文を文節区切りとした」学習データから生成されたフレーズテーブルは日本語フレーズが長いため，翻訳に用いるフレーズ対を減らす効果が大きい．また，日本語と英語の文法構造の差を補うこともできる．そのため，翻訳精度の向上が大きかった．

このことから，単文の翻訳には単純にフレーズ対を増やすことが効果的であり，重文複文の翻訳には，長いフレーズ対を増やすことが効果的であると考えている．

8.4 翻訳システムのカバー率と翻訳精度の関係

本研究では，4つの翻訳実験を行った．各翻訳結果に含まれる未知語の数を表19に示す．

表 19: 各翻訳結果の未知語数 (テストデータ:各 9,000 文)

単文		
	BLEU	未知語を含む文の数
日:単語，英:単語 (従来)	0.2015	2,189
従来+日:文節，英:単語	0.2058	2,072
従来+日:単語，英:フレーズ	0.2063	2,030
従来+日:文節，英:フレーズ	0.2045	2,145
4つを併用	0.2086	1,926
重文複文		
	BLEU	未知語を含む文数
日:単語，英:単語 (従来)	0.1746	2,649
従来+日:文節，英:単語	0.1784	2,514
従来+日:単語，英:フレーズ	0.1760	2,457
従来+日:文節，英:フレーズ	0.1783	2,595
4つを併用	0.1797	2,339

表19から，単文の翻訳において，未知語を含む文の数が少ないほど，BLEUスコアは向上している．しかし，重文複文の翻訳において,「日本語文を単語区切り，英語文をフレーズ区切りとした」学習データから生成されたフレーズテーブルを併用した場合，未知語は大きく減少しているが翻訳精度の向上は小さかった．このことから，重文複文の翻訳精度の向上のために，未知語を減らすことより，長いフレーズ対を増やすことが有効であると考えている．

9　おわりに

本研究では，日英統計翻訳において，短いフレーズ対を用いるために並び替えの候補数が増加し，翻訳精度が低下すると考え，長いフレーズ対を増やす手法を提案した．具体的には，学習データの日本語文を文節区切りとし，長いフレーズを多く持つ文節区切りフレーズテーブルを生成した．そして，従来の単語区切りフレーズテーブルと併用し，翻訳実験を行った．

実験の結果，従来手法と比較して，BLEU スコアが単文で 0.43%，重文複文で 0.38%向上した．また，同等の手法を学習データの英語文に対して適用した場合の実験も行った．英語文に適用した場合も，従来手法と比較して，翻訳精度が向上した．さらに，従来手法のフレーズテーブルと提案手法により得られた 3 つのフレーズテーブルを併用した実験も行った．4 つのフレーズテーブルを用いた場合，従来手法と比較して，BLEU スコアが単文の翻訳で 0.71%，重文複文の翻訳で 0.51%向上した．翻訳精度が向上した理由として，長いフレーズ対が，並び替えの候補を減らす効果を持つことに加え，助詞や接尾辞といった日本語と英語の文法構造の違いを補う効果を持つということを考えている．

今後の予定として，文節とフレーズ以外の区切りを用いて，併用するフレーズテーブルを増やすことを考えている．

謝辞

最後に，3 年間に渡って御指導いただきました鳥取大学工学部知能情報工学科計算機 C 研究室の池原悟教授，村上仁一准教授，徳久雅人助教に心から御礼申し上げます．

また，御多忙の中，助言をいただきました菅原一孔教授，川村尚生教授に厚く御礼申し上げます．

その他，参考にさせて頂いた論文の著者の方々に対して深く感謝します．

参考文献

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, “The Mathematics of Statistical Machine Translation, Parameter Estimation”, Computational Linguistics, 19(2), 1993.

[2] 猪澤雅史, 村上仁一, 徳久雅人, 池原悟, “統計翻訳における，単文と重文複文の翻訳精度の評価”, 情報処理学会研究報告, pp.79-84, 2008.

[3] 鏡味良太, 村上仁一, 徳久雅人, 池原悟, “統計翻訳における人手で作成された大規模フレーズテーブルの効果”, 言語処理学会第 14 回年次大会, pp.224-227, 2008.

[4] Jin’ichi Murakami, Masato Tokuhisa, Satoru Ikehara, “Statistical Machine Translation using Large J/E Parallel Corpus and Long Phrase Tables”, International Workshop on Spoken Language Translation 2007, pp.151-155, 2007.

[5] Philipp Koehn, Franz J. Och, and Daniel Marcu, “Statistical phrase-based translation”, In Proceedings of HLT-NAACL 2003, pp.127-133, 2003.

[6] 西山七絵, 村上仁一, 徳久雅人, 池原悟, “単文文型パターン辞書の構築”, 言語処理学会第 11 回年次大会, pp.372-375, 2005.

[7] 村上仁一, 池原悟, 徳久雅人, “日本語英語の文対応の対訳データベースの作成”, 「言語，認識，表現」第 7 回年次研究会, 2002.

[8] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/

[9] CaboCha: Yet Another Japanese Dependency Structure Analyzer, http://chasen.org/ taku/software/cabocha

[10] training-release-1.3.tgz: http://www.statmt.org/wmt06/shared-task/baseline.html

[11] GIZA++: http://www.fjoch.com/GIZA++

[12] SRILM: The SRI Language Modeling Toolkit,
http://www.speech.sri.com/projects/srilm

[13] Moses: moses.2007-05-29.tgz,
http://www.statmt.org/moses/

[14] Franz Josef Och, “Minimum Error Rate Training in Statistical Machine Translation”, In Proceeding of the 41st Annual Meeting of the Association for Computational Linguistics, pp.160-167, 2003.

[15] NIST Open Machine Translation:
http://www.nist.gov/speech/tests/mt

[16] The METEOR Automatic Machine Translation Evaluation System:
http://www.cs.cmu.edu/ alavie/METEOR/

[17] Apple Pie Parser:
http://nlp.cs.nyu.edu/app/

[18] 猪澤雅史, 村上仁一, 徳久雅人, 池原悟, “文節区切りの学習データを用いた，日英統計翻訳の検討”, 言語処理学会 16 回年次大会 (発表予定).