概要

パターン翻訳は, 入力された原言語文に対して, 対訳句と文パターン (原言語文パターンと目的言語文パターンの対の構成) を用いて翻訳文を出力する. パターン翻訳は, 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, 対訳句と文パターンを人手作成するために, 開発にコストがかかる.

そこで, 江木らは, パターンに基づく統計翻訳を提案した. パターンに基づく統計翻訳は, 統計的手法を用いて, 対訳句と文パターンを自動作成して翻訳を行う. しかし, 自動作成された対訳句の問題として, 対応する原言語と目的言語が不自然な対訳句が多く含まれていることがあげられる.

ところで, 人手作成された対訳句には, 鳥バンクと英辞郎がある. 通常, 人手作成された対訳句は, 自動作成された対訳句よりも信頼性が高い. しかし, 対訳句を自動作成する翻訳と比較すると, カバー率が低くなる.

本研究では, 人手作成された対訳句として鳥バンクと英辞郎を用いて, パターンに基づく日英統計翻訳を行い, 翻訳精度とカバー率の調査を行った. また, 対訳句を自動作成するパターンに基づく統計翻訳と対比較評価を行った.

対比較評価の結果, 人手作成された対訳句を用いたパターンに基づく統計翻訳と対訳句を自動作成するパターンに基づく統計翻訳の翻訳精度に大きな差がないことがわかった. しかし, カバー率は, 人手作成された対訳句を用いたパターンに基づく統計翻訳が低かった. よって, 対訳句を自動作成して翻訳するパターンに基づく統計翻訳の有効性を示せた.

目 次

図 目 次

表 目 次

第1章　はじめに

パターン翻訳は [1], 入力された原言語文に対して, 対訳句と文パターン (原言語文パターンと目的言語文パターンの対の構成) を用いて翻訳文を出力する. パターン翻訳は, 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, 対訳句と文パターンを人手作成するために, 開発にコストがかかる.

そこで, 江木らは, パターンに基づく統計翻訳を提案した [2]. パターンに基づく統計翻訳は, 統計的手法を用いて, 対訳句と文パターンを自動作成して翻訳を行う. しかし, 自動作成された対訳句の問題として, 対応する原言語と目的言語が不自然な対訳句が多く含まれていることがあげられる.

ところで, 人手作成された対訳句には, 鳥バンク [3] と英辞郎 [4] がある. 通常, 人手作成された対訳句は, 自動作成された対訳句よりも信頼性が高い. しかし, 対訳句を自動作成する翻訳と比較すると, カバー率が低くなる.

本研究では, 人手作成された対訳句として鳥バンクと英辞郎を用いて, パターンに基づく日英統計翻訳を行い, 翻訳精度とカバー率の調査を行う. また, 対訳句を自動作成するパターンに基づく統計翻訳と対比較評価を行う.

対比較評価の結果, 鳥バンクを用いたパターンに基づく統計翻訳と英辞郎を用いたパターンに基づく統計翻訳ともに, 対訳句を自動作成するパターンに基づく統計翻訳の翻訳精度に大きな差がないことがわかった. また, カバー率においては, 鳥バンクを用いたパターンに基づく統計翻訳と英辞郎を用いたパターンに基づく統計翻訳ともに対訳句を自動作成するパターンに基づく統計翻訳より低かった. よって, 対訳句を自動作成して翻訳するパターンに基づく統計翻訳の有効性を示せた.

第2章　パターン翻訳

2.1　概要

パターン翻訳は, 機械翻訳手法の一種である. 翻訳には, 原言語文と目的言語文の対訳文に対して, 任意の単語や句を変数化した文パターンと単語辞書が必要である. 原言語入力文と原言語文パターンを照合し, 適合する原言語文パターンに対応する目的言語文パターン得る. そして, 文パターンの変数部に対応する単語や句単語辞書を用いて翻訳し, 目的言語翻訳文を出力する.

パターン翻訳は, 適切な文パターンが適合した場合, 翻訳精度の高い翻訳文を得ることができる. しかし, 一般的なパターン翻訳は文パターンを人手作成するため開発に時間がかかる. また, 文パターン辞書に適合しない場合は翻訳ができない. よって, 入力文に対するカバー率が低い.

2.2　手順

一般的な英日パターン翻訳の手順を以下に示す.

手順１　文パターン辞書と単語辞書を用意する.

手順２　英語入力文と英語文パターンを照合する.

手順３　変数部に対応する英単語を単語辞書を用いて日本語単語に翻訳する.

手順４　英語側文パターンに対応する日本語側文パターンの変数部を翻訳した日本語単語に置き換える.

手順５　手順４で得た日本語翻訳文を出力する.

図 2.1 に英日パターン翻訳の手順を示す.

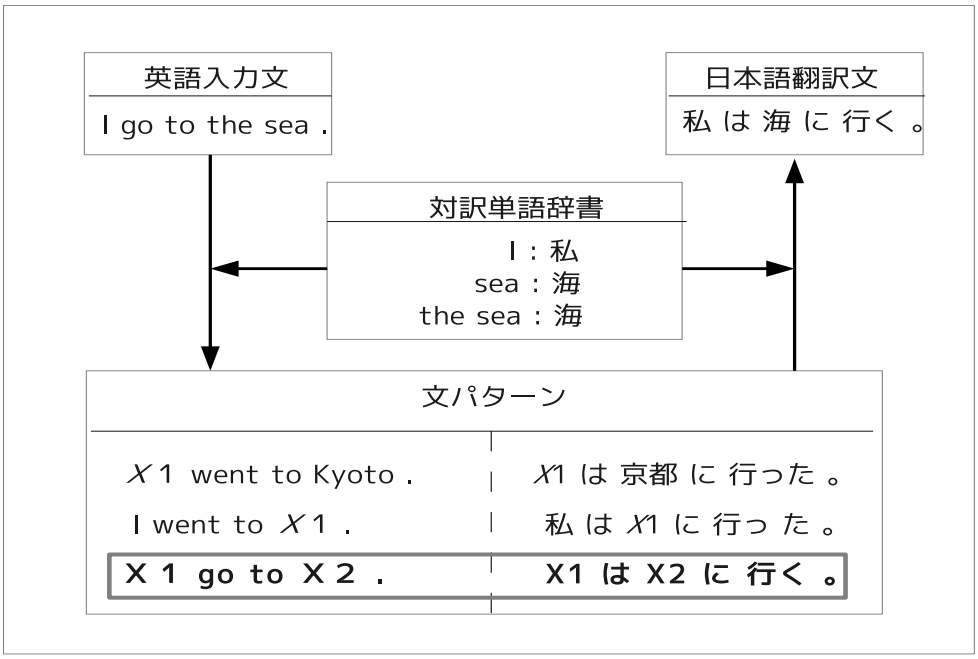


図 2.1: 英日パターン翻訳の手順

2.3 文パターン辞書

文パターン辞書とは, 大量の対訳文から任意の単語や句を変数化して得られる文パターンの集合である. 表 2.1 に英日パターン翻訳の例を示す.

表 2.1: 英日パターン翻訳の例

英語入力文	I go to the sea .
英語文パターン	I go to $X1$.
日本語文パターン	私 は $X1$ に 行く 。
日本語翻訳文	私 は 海 に 行く 。

第3章　統計翻訳

統計翻訳システムを, 原言語 (翻訳の対象となる入力された言語) を日本語, 目的言語 (翻訳された後に出力される言語) を英語とする日英統計翻訳の場合を例として説明する.

3.1　概要

統計翻訳は, 機械翻訳手法の一種である. 日英統計翻訳システムは, 日本語入力文 j が与えられたとき, 全ての組み合わせの中から確率が最大となる英語文 $\hat{e}$ を探索することで翻訳を行う. 以下に基本モデルを示す.

$$\hat{e} = \arg\max_{e} P(e|j) \tag{3.1}$$

$$\approx \arg\max_{e} P(j|e)P(e) \tag{3.2}$$

$P(j|e)$ は翻訳モデル，$P(e)$ は言語モデルと呼ぶ. $\hat{e}$ を探索する翻訳システムをデコーダと呼ぶ. 図 3.1 に日英統計翻訳の流れを示す.

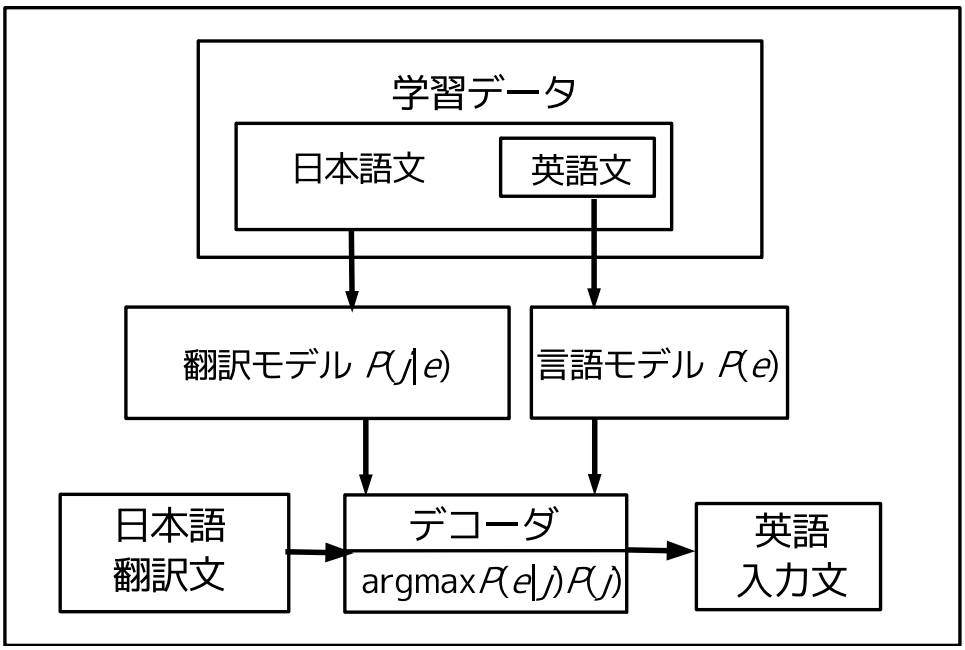


図 3.1: 日英統計翻訳の流れ

3.2 翻訳モデル

翻訳モデルは, 訳語の尤もらしさを規定する統計モデルである. 対訳文 (原言語文と目的言語文) から学習する. 翻訳モデルの代表例として, IBM 翻訳モデル [6] がある.

3.2.1 IBM 翻訳モデル

IBM 翻訳モデルは, 単語に基づく統計翻訳を想定して作成された単語対応の確率モデルである. IBM 翻訳モデルは, 英語文 e , 日本語文 j の翻訳モデル $P(j|e)$ を計算するためにアラインメント a と呼ばれる概念を導入した. 以下に IBM 翻訳モデルの基本式を示す. なお, アラインメントとは, ある日本語単語 j と英単語 e の対応関係のことを示す.

$$P(j \mid e) \;=\; \sum_{a} P(j, a \mid e)$$

IBM 翻訳モデルでは, 日英統計翻訳の場合, 英単語は 1 対 n の対応を持ち, 日本語の単語は 1 つの英単語のみと対応すると仮定する. また, 日本語の単語の対応関係として適切な英単語がなかった場合, 英語文の文頭の特殊文字 e と対応付けを行う.

3.2.2 GIZA++

GIZA++は主に統計翻訳で使用されるツールであり, IBM 翻訳モデル [7] を用いて, 原言語と目的言語の対訳文から対訳単語と単語翻訳確率を自動的に得る.

3.3 言語モデル

言語モデルは, 翻訳候補の文に対して目的言語の文らしさの指標を与えるモデルである. 翻訳モデルでは, 訳語の選択や訳語の位置の選択に対する評価を与えることはできるが, 作られた翻訳候補が目的言語の文としてふさわしいかどうかを判断する評価を与えることはできない. そのため, 言語モデルでは, 日英統計翻訳の場合, より英語らしい文に対して, 高い確率を与えることで, 翻訳モデルで翻訳された訳文候補の中から英語として自然な文を選出する.

・N-gram モデル

統計翻訳では一般的に, N-gram モデルを用いる. N-gram モデルは, “単語列 $w_1^n = w_1, w_2, w_3, ...w_n$ の i 番目の単語 w_i の生起確率 $P(w_i)$ は直前の $(N-1)$ の単語列 $w_{i-(N-1)}$, $w_{i-(N-2)}, w_{i-(N-3)}, ...w_{i-1}$ に依存する” という仮説に基づくモデルである. また, $N = 1$ のモデルを uni-gram, $N = 2$ のモデルを bi-gram, $N = 3$ のモデルを tri-gram と特有の呼びかたをする. $N = 4$ 以上は 4-gram など数値を用いて呼ぶ. N-gram は以下の式で表現される. ここで, w_i^j は i から j 番目までの単語列を表す.

$$
\begin{aligned}
P(w_1^n) &= P(w_1) \times P(w_2|w_1) \times P(w_3|w_1^2)...P(w_n|w_1^{n-1}) && (3.3)\\
&\approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_1^2)...P(w_n|w_{n-(N-1)}^{n-1}) && (3.4)\\
&= \prod_{i=1}^{n} P(w_i|w_{i-(N-1)}^{i-1}) && (3.5)
\end{aligned}
$$

また, $P(w_i|w_{n-(N-1)}^{i-1})$ は以下の式で計算される. ここで, $C(w_1^i)$ は単語列 w_1^i が出現する頻度を表す.

$$
P(w_i|w_{i-(N-1)}^{i-1}) = \frac{C(w_{i-(N-1)}^{i})}{C(w_{i-(N-1)}^{i-1})} \tag{3.6}
$$

たとえば, “I have dogs . ” という単語列に対して $N = 2$ とした bi-gram モデルの言語モデルを適応した場合, 単語列が生成される確率は以下の式で計算される.

$$
P(I\ have\ dogs\ .) \simeq P(I) \times P(have|I) \times P(dogs|have)...P(.|dogs) \tag{3.7}
$$

tri-gram モデルであれば, $P(dogs|I\ have)$, 4-gram モデルであれば $P(.|I\ have\ dogs)$ となる.

(3.6) 式から信頼性の高い値を推定するためには, 単語列 w_1^n が多く出現している必要がある. しかし, 実際には多くの単語列は出現数が 0 となることが多いため信頼できる値を推定できない場合が多い. 低頻度な語彙の場合, $C(w_{i-(N-1)}^{i})$, $C(w_{i-(N-1)}^{i-1})$ の値が小さく, 信頼性が低い. また, 学習データ中に単語列 w_1^i が存在しない場合, この単語列の出現確率は 0 と推定される. そのため, (3.6) 式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, 出現頻度の少ない単語列をモデルの学習から削除 (カットオフ) する方法や, 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-1)-gram の値から推定する.

3.4 デコーダ

デコーダは, 翻訳モデルと言語モデルの全ての組み合わせの中から確率が最大となる出力文を探索して翻訳を行う. 代表的なデコーダに Moses[10] がある.

3.5 句に基づく統計翻訳

句に基づく統計翻訳は, 2000 年代初期に提案された. 句に基づく統計翻訳は, 句の対応を翻訳モデルに用いる. 句を構成する単語の数が, 翻訳する文の句と目的文の句で一致する必要がないため, 単語に基づく統計翻訳の単語対応の問題を解決した. また, 並べ替えにおいても, 単語に基づく統計翻訳よりも優れている. そのため, 近年では句に基づく統計翻訳が主流となっている.

第4章　パターンに基づく統計翻訳

4.1　概要

パターンに基づく統計翻訳は, 統計的手法を用いて, 自動的に対訳句と文パターン辞書を作成して翻訳を行う手法である.

4.2　手順

江木ら [2] によって提案されたパターンに基づく統計翻訳は大きく分けて 5 つのステップで翻訳を行う. 英日翻訳の場合の手順を以下に示す.

4.2.1　単語辞書の作成

対訳文と GIZA++を用いて単語辞書を作成する. まず, GIZA++を用いて英日方向と日英方向の対訳単語と単語翻訳確率を得る. 次に, 英日方向の単語翻訳確率と日英方向の単語翻訳確率を掛け合わせる. 図 4.1 に単語辞書の作成手順と例を示す.

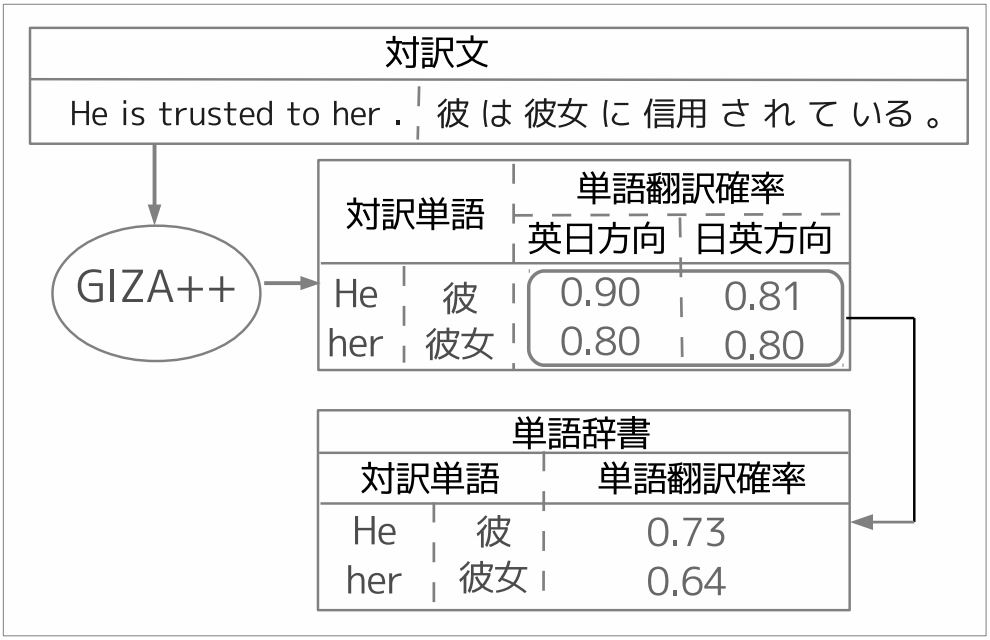


図 4.1: 単語辞書の作成手順

4.2.2 単語に基づく文パターン辞書の作成

対訳文と単語辞書を用いて, 単語に基づく文パターン辞書を作成する.

4.2.2.1 英単語照合

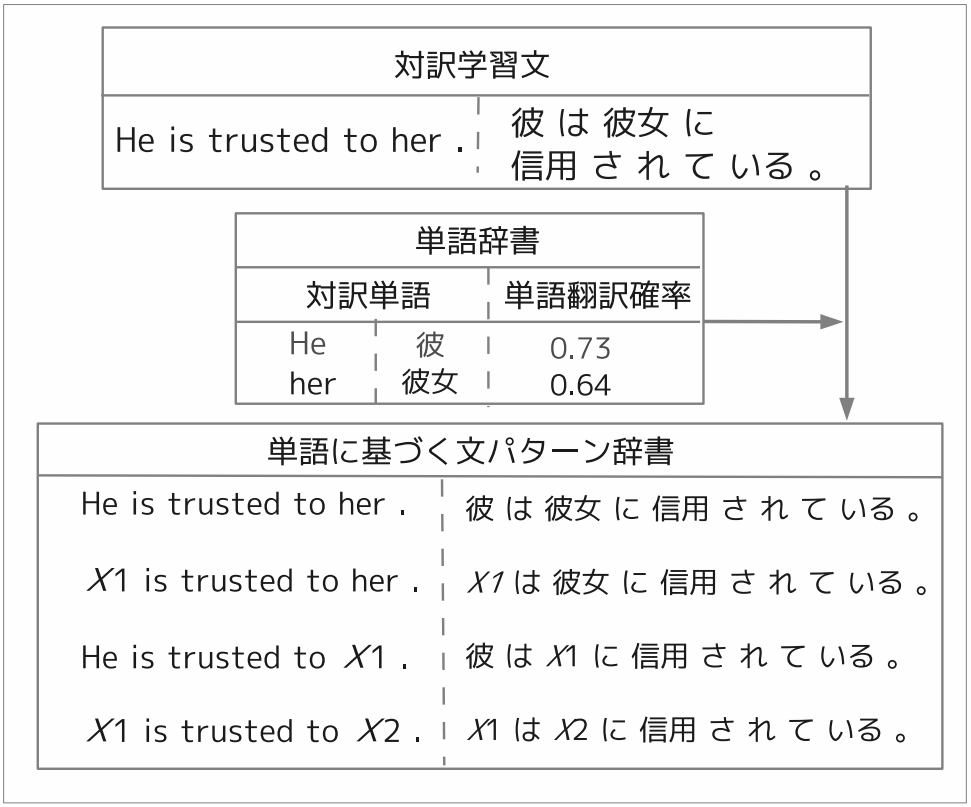
対訳文の英単語と単語辞書の英単語を照合する.

4.2.2.2 日本語単語照合

英単語に対応する日本語単語と対訳文の日本語単語を照合する.

4.2.2.3 変数化

英日両方の単語が照合に成功した場合, 該当箇所を変数化する. 変数化する場合, 変数の組み合わせを考慮し, 可能な限り多くの単語に基づく対訳文パターンを生成する. 図 4.2 に単語に基づく文パターン辞書の作成手順と例を示す.



[1]

図 4.2: 単語に基づく文パターン辞書の作成手順

図 4.2 において変数化される対訳単語は“He | 彼” , “her | 彼女”, である. 2 つの対訳単語が変数化される場合と変数化されない場合の組み合わせを全て考慮し, 2^2=4 通りの単語に基づく文パターンを生成する.

4.2.3 フレーズ辞書の作成

対訳文と単語に基づく文パターン辞書を用いて, 対訳句を抽出する. 次に, 単語翻訳確率を用いて, 対訳句にフレーズ対数確率を付与して, フレーズ辞書を作成する.

4.2.3.1 パターン照合

対訳文と単語に基づく文パターン辞書を照合する.

4.2.3.2 対訳句の抽出

対訳文が単語に基づく文パターンに適合した場合, 単語に基づく文パターンの変数部に対応する対訳句を抽出する. 図 4.3 に対訳句の抽出手順と例を示す.

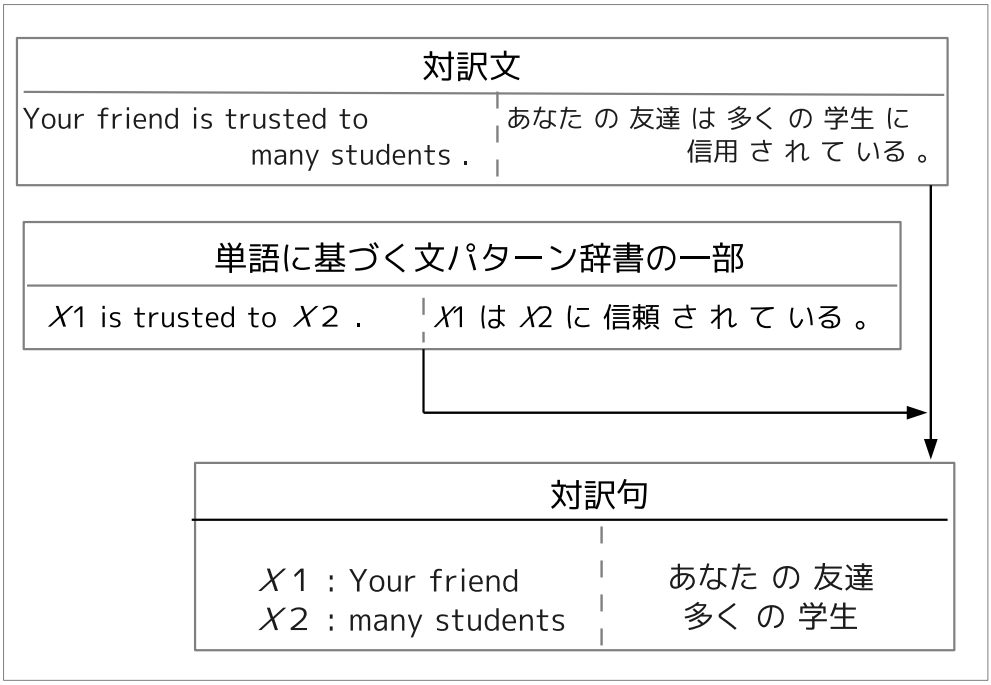


図 4.3: 対訳句の抽出手順

4.2.3.3 フレーズ対数確率の計算

1. 単語の組み合わせの取得

   対訳句において, 英語句の単語と日本語句の単語の全ての組み合わせを得る. 同様に日本語句の単語の組み合わせと英語句の単語の組み合わせも得る.

2. 単語翻訳確率の計算

   英単語に対応する日本語単語の中で, 単語翻訳確率が最大となる確率を得る. 同様に日本語単語に対応する英単語の中で, 単語翻訳確率が最大となる確率を得る.

3. フレーズ対数確率の付与

   得られた確率に対して対数を取り, 英日方向の単語翻訳確率の対数値の総和と日英方向の単語翻訳確率の対数値の総和を求める. 次に, 英日方向の総和と日英方向の総和を足し合わせて, 対訳句のフレーズ対数確率として付与する.

図 4.4 に英日フレーズ対数確率の付与手順を示す.

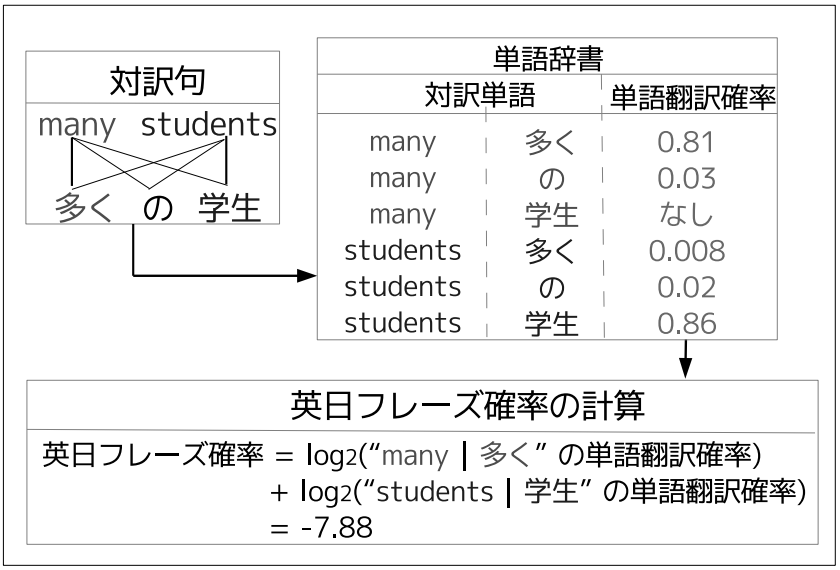


図 4.4: 英日フレーズ対数確率の付与手順

図 4.4 に英日方向の対訳句の例として “many students | 多くの学生” を示す. まず, 英語句の単語と日本語句の単語の全ての組み合わせを得る. 次に, 単語翻訳確率を用いて, 各組み合わせの中から最大となる単語翻訳確率を得る. 図 4.4 では “many | 多く” に付与された確率 “0.81” が最も高いため, 0.81 に対して対数を取る. “students” も同様に単語翻訳確率に対数を取り総和を求める.

4.2.4 句に基づく文パターン辞書の作成

対訳文とフレーズ辞書を用いて, 句に基づく文パターンを作成する. 次に, 単語翻訳確率を用いて文パターン対数確率を付与し, 句に基づく文パターン辞書を作成する. 以下に手順を示す.

4.2.4.1 英語句の照合

対訳文における英語文の各句とフレーズ辞書の英語句を照合する.

4.2.4.2 日本語句の照合

英語句に対応する日本語句と対訳文における日本語文の各句を照合する.

4.2.4.3 変数化

対訳句が照合に成功した場合, 該当箇所を変数化し, 句に基づく文パターンを生成する. 変数化するときに, 変数の組み合わせを考慮して, 可能な限り多くの句に基づく文パターンを生成する. 図 4.5 に句に基づく文パターンの作成手順と例を示す.

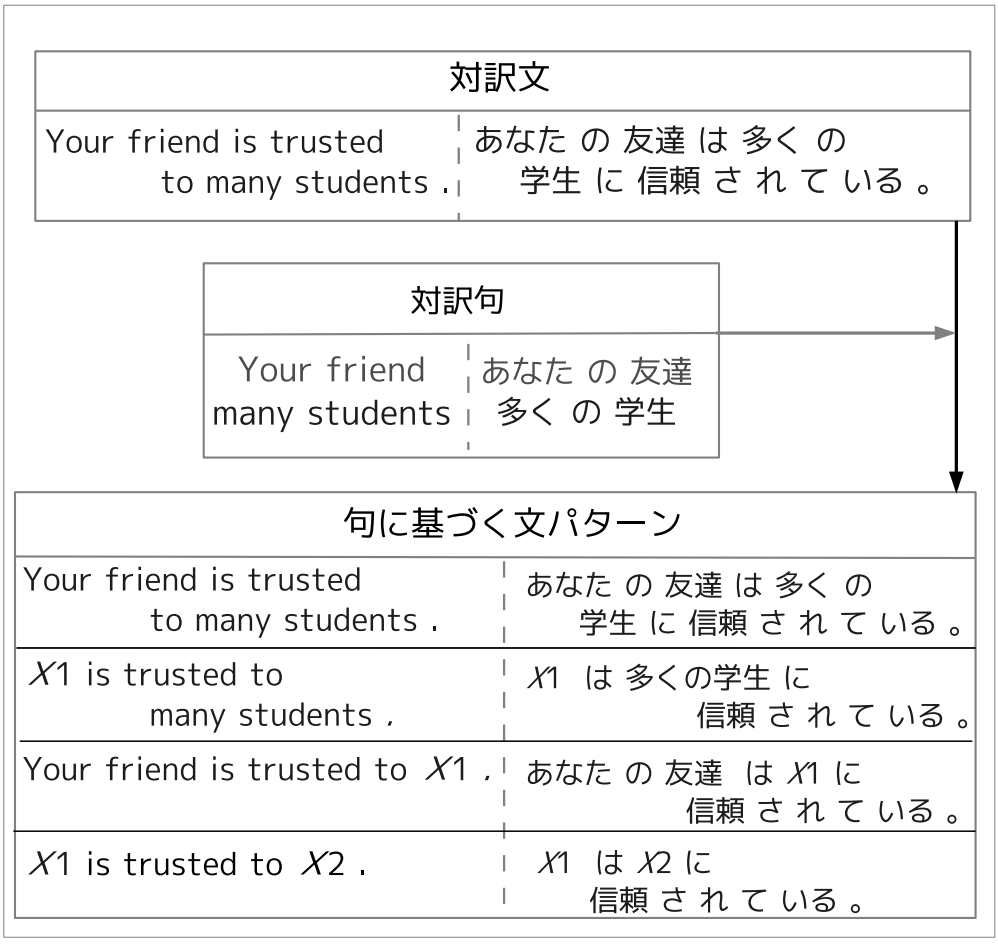


図 4.5: 句に基づく文パターンの作成手順

図 4.5 において変数化されるフレーズ対は“Your friend | あなたの友達” , “many students | 多くの学生” である. この 2 つのフレーズ対が変数化される場合とされない場合の組み合わせを全て考慮し, 2^2=4 通りの句に基づく文パターンを生成する.

4.2.4.4 文パターン対数確率の付与

文パターンの字面と単語翻訳確率を用いて, 文パターン対数確率を付与する. 確率の付与は英日文パターンと日英文パターンに対して行う. また, 手順 4.2.3.3 で説明したフレーズ対数確率の付与と同じ手法を用いる. 図 4.6 に, 英日方向の文パターン対数確率の付与手順と例を示す.

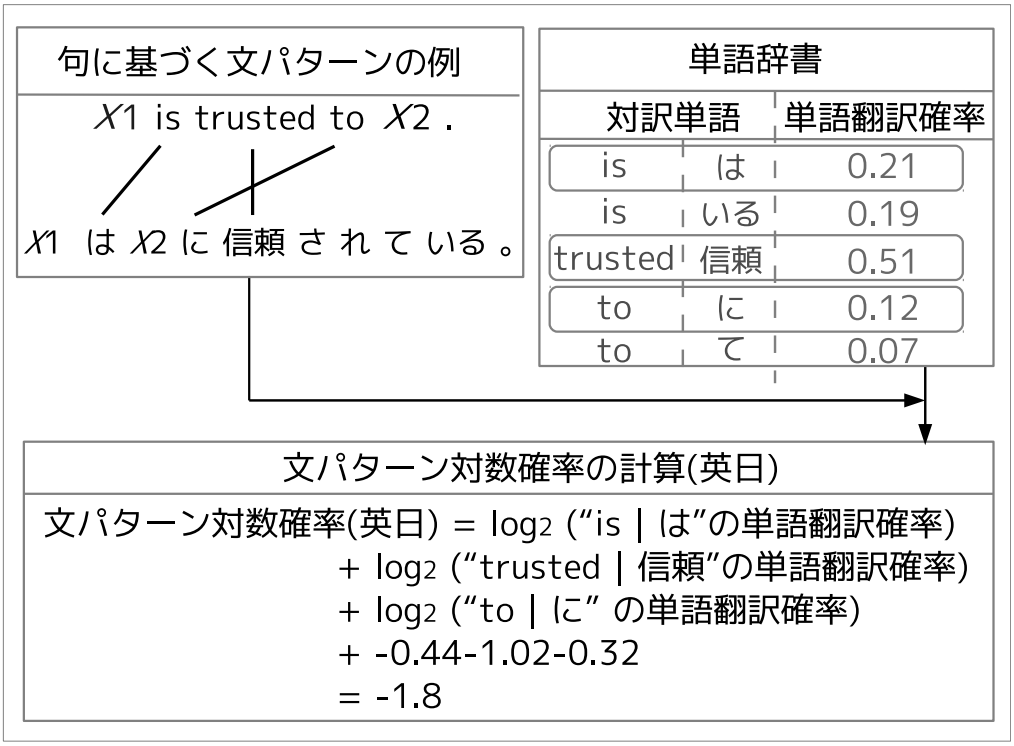


図 4.6: 文パターン対数確率の付与手順

図 4.6 に英日方向の句に基づく文パターンの例として“$X1$ is trusted to $X2$. | $X1$ は $X2$ に 信頼 さ れ て いる 。” を示す. まず, 英語文パターンの単語と日本語文パターンの単語の全ての組み合わせを得る. 次に, 単語翻訳確率を用いて, 各組み合わせの中から最大となる単語翻訳確率を得る. 図 4.6 では“is | は” に付与された確率“0.81” が最も高いため, 0.81 に対して対数を取る. “trusted” , “to” も同様に単語翻訳確率に対して対数を取り総和を求める.

4.2.5 翻訳

フレーズ辞書と句に基づく文パターン辞書を用いて, 日本語翻訳文を出力する. 翻訳精度を向上させるために, 翻訳時に英語入力文と英語文パターンの字面を比較する. そして最も多く字面が一致する英語文パターンから優先して選択する. 日本語翻訳文の絞り込みにはフレーズ対数確率と文パターン対数確率と言語確率 (tri-gram) を用いる. 総和を取り, 確率が最大となる日本語翻訳文を出力する. 図 4.7 に日本語翻訳文を出力するまでの手順を示す. また, 以下に英日パターン翻訳の手順を示す.

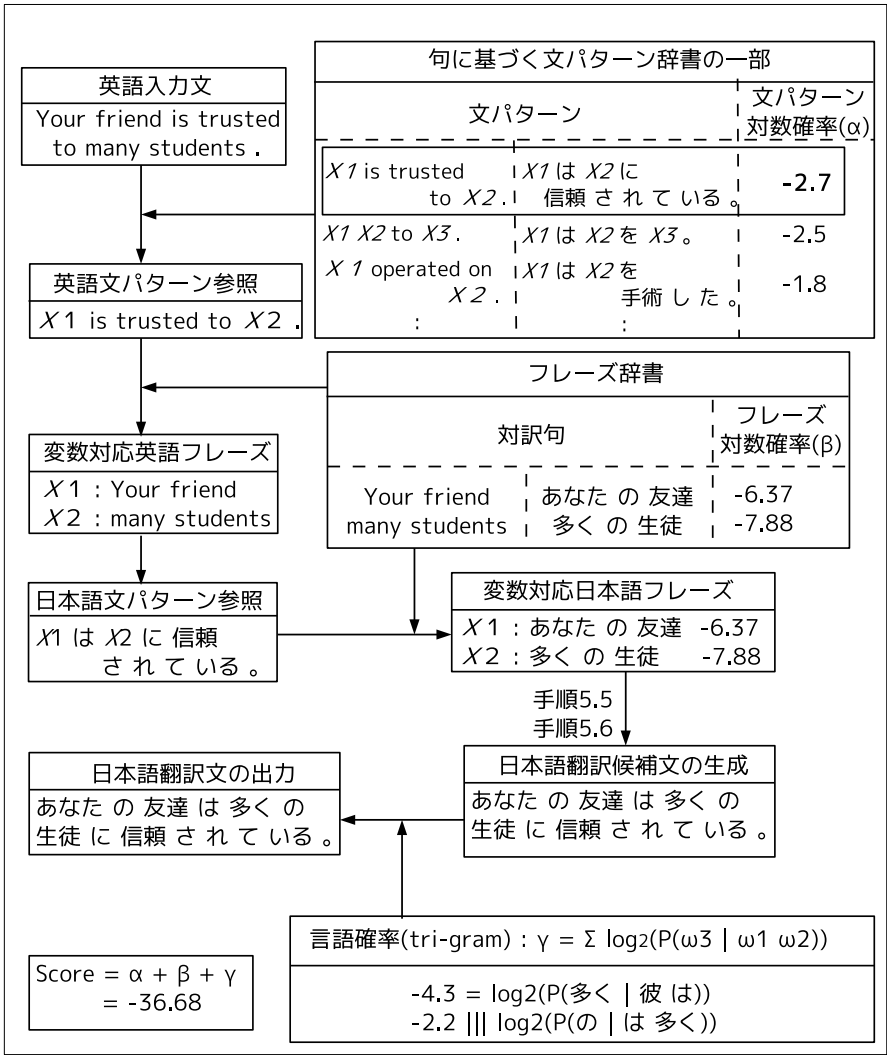


図 4.7: 日本語翻訳文の出力手順

4.2.5.1 英語文パターンの選択

英語文を入力とし, 英語入力文と英語文パターンの字面を比較する. そして最も多く字面が一致する英語文パターンから優先して選択する. 図 4.7 では, “$X1$ is trusted to $X2$. ” がもっとも字面が一致する.

4.2.5.2 英語句の取得

最も多く字面が一致する英語文パターンの変数部に対応する英語句を得る. 図 4.7 では, “$X1$” には “Your friend” , “$X2$” には “many students” が対応する. 次に, “$X1$ is trusted to $X2$. ” に対応する日本語文パターン “$X1$ は $X2$ に 信頼 さ れ て いる 。” と文パターン対数確率 “-2.7” を得る.

4.2.5.3 日本語文パターンの取得

英語文パターンに対応する日本語文パターンと文パターン対数確率を得る. 図 4.7 では,“$X1$ is trusted to $X2$. ” に対応する日本語文パターン “$X1$ は $X2$ に 信頼 さ れ て いる 。” と文パターン対数確率 “-2.7” を得る.

4.2.5.4 日本語句の取得

日本語文パターンの変数部に対応する日本語句とフレーズ対数確率を得る. 図 4.7 では, “$X1$” には “あなた の 友達” , “$X2$” には “多く の 生徒” が対応する. また, フレーズ対数確率は, それぞれ “-6.37” , “-7.88” である.

4.2.5.5 日本語翻訳候補文の生成

日本語文パターンの変数部を取得した日本語句に置き換える. そして, 日本語翻訳候補文として生成する.

4.2.5.6 言語確率 (tri-gram) の算出

日本語翻訳候補文に対して言語確率 (tri-gram) を計算する.

4.2.5.7 日本語翻訳文の選択

フレーズ確率と文パターン確率と言語確率 (tri-gram) の総和を求め, 日本語翻訳候補文に付与する. 最後に総和が最大となる日本語翻訳文を出力する.

第5章　提案手法

パターンに基づく統計翻訳は, 統計的手法を用いて, 対訳句と文パターンを自動作成して翻訳を行う. 自動的に作成された対訳句の問題として, 対応する原言語と目的言語が不自然な対訳句が多く含まれていることがあげられる.

ところで, 対訳句には人手で作成された対訳句がある. 通常, 人手で作成された対訳句は, 自動的に作成された対訳句よりも信頼性が高い. しかし, 対訳句を自動作成する翻訳と比較するとカバー率が低くなると考えられる.

本研究では, 人手で作成された対訳句を用いたパターンに基づく統計翻訳を行う. 本手法は大きく分けて 3 つのステップで翻訳を行う. 以下に日英翻訳の手順を示す.

5.1　フレーズ辞書の作成

人手作成された対訳句には, 確率がない. そこで, 確率を付与する必要がある. 人手作成された対訳句に単語翻訳確率を用いて, フレーズ対数確率を付与し, フレーズ辞書を作成する. なお, フレーズ対数確率を付与する手順は 4.2.3.3 節と同様である.

5.2　句に基づく文パターン辞書の作成

対訳文とフレーズ辞書を用いて, 句に基づく文パターンを作成する. 次に単語翻訳確率を用いて, 句に基づく文パターンに文パターン対数確率を付与し, 句に基づく文パターン辞書を作成する. 句に基づく文パターン辞書の作成手順は 4.2.4 節と同様である. また, 文パターン対数確率の付与手順は 4.2.4.4 節と同様である.

5.3　手順3 英文生成

日本語文を入力とし, フレーズ辞書と句に基づく文パターン辞書と言語確率 (tri-gram) を用いて, 英語翻訳文を出力する.

翻訳精度を向上させるために, 翻訳時に日本語文と日本語文パターンの字面を比較する．そして, 最も多くの字面が一致する日本語文パターンを優先して選択する．英語翻訳文の絞り込みには, フレーズ対数確率と文パターン対数確率と英語翻訳文の言語確率 (tri-gram) を用いる．これらの総和を取り, 確率が最大となる英語翻訳文を出力する．英語翻訳文を出力するまでの手順は 4.2.5 節と同様である．

第6章　実験環境

6.1　日英対訳文

本研究では, 日英対訳文として, 単文コーパス [5] を用いる. なお, 本研究で使用する単文コーパスにおいて, 日本語文は単文である. しかし, 英語文は単文とは限らず, 重文・複文が含まれている. 統計翻訳の前処理として, 日本語文に対して, MeCab[8] を用いて形態素解析を行う. また, 英語文に対して, tokenizer.sed[9] を用いて正規化を行う.

本研究では, 単文コーパスを表 6.1 の内訳で用いる.

表 6.1: 日英対訳文数

対訳文	100,000 文対
テストデータ	100 文

6.2　対訳句

本研究では, 人手作成された対訳句として鳥バンクと英辞郎を用いる.

6.2.1　鳥バンク

鳥バンクは, 日本語の重文と複文を対象として人手作成されたパターン辞書である. 本研究では, このパターン辞書から抽出した対訳句を用いる. 対訳句の例を表 6.2 に示す.

表 6.2: 鳥バンクから抽出した対訳句の例

あなた の お父さん
Your father
息子 の 話
son 's story
移民 政策
immigration policy

6.2.2 英辞郎

英辞郎は, EDP(Electronic Dictionary Project) がアップデートし続けている英和・和英辞書である. 英辞郎のデータには対訳句の他に翻訳例や注釈, 本来の文に出てこない“〜”等の記号が含まれる. 表 6.3 に英辞郎の対訳句の例を示す.

表 6.3: 英辞郎から抽出した対訳句の例

に 心 を 用いる apply the mind to
から 成る consist of
の 結果 として 生じる come out from

6.2.3 対訳句数

本研究で対訳句コーパスとして用いる鳥バンクと英辞郎の対訳句数を表 6.4 に示す.

表 6.4: 対訳句数

鳥バンク	336,161 句対
英辞郎	1,357,047 句対

6.3 翻訳モデルの学習

翻訳モデルの学習には, train-model.perl[10] を用いる.

6.4 言語モデルの学習

言語モデルの学習には, SRILM[11] の ngram-count を用いる. 本研究では, N-gram モデルは 5-gram とする.

6.5 実験内容

本研究では, 鳥バンクと英辞郎を用いて, パターンに基づく統計翻訳を行い, 翻訳精度とカバー率の調査を行う. また, 江木らの手法 (以下, 自動作成) と対比較評価を行う.

6.6 評価方法

本研究では, 翻訳システムによって出力した文の評価に対比較評価法を用いる. 対比較評価は, 二つの文を相対的に比較して, どちらがより正しい文であるかを人手で選択する評価方法である. 2 つの翻訳システムの出力で優劣を判断する場合に有効である.

第7章　実験結果

7.1　カバー率

表 7.1 に鳥バンクを用いたパターンに基づく統計翻訳で得られた句数と文パターン数とカバー率を示す.

表 7.1: 得られたデータ数 (鳥バンク)

フレーズ辞書	336,161 句対
句に基づく文パターン辞書	3,820,745 パターン対
英語翻訳文	42% (42/100 文)

表 7.2 に英辞郎を用いたパターンに基づく統計翻訳で得られた句数と文パターン数とカバー率を示す.

表 7.2: 得られたデータ数 (英辞郎)

フレーズ辞書	1,357,047 句対
句に基づく文パターン辞書	1,179,179 パターン対
英語翻訳文	13% (13/100 文)

表 7.3 に自動作成で得られた句数と文パターン数とカバー率を示す.

表 7.3: 得られたデータ数 (自動作成)

フレーズ辞書	222,102,894 句
句に基づく文パターン辞書	308,566,385 パターン
カバー率	82% (82/100 文)

表 7.1, 表 7.2, 表 7.3 より, 鳥バンクと英辞郎を用いたパターンに基づく統計翻訳が自動作成よりカバー率が低いことがわかる.

7.2 翻訳精度 (対比較評価)

鳥バンクを用いて得られた英語翻訳文 42 文のうち自動作成と出力が重なった 39 文と, 英辞郎を用いて得られた英語翻訳文 13 文のうち自動作成と出力が重なった 12 文に対して, 対比較評価を行った. 表 7.4 に鳥バンクと自動作成の対比較評価結果を示す. また, 表 7.5 に英辞郎と自動作成の対比較評価結果を示す. 表中の表記方法を表 7.4 をもとに説明する.

- “鳥バンク○”: 鳥バンクを用いた翻訳精度が自動作成の翻訳精度より優れている文
- “自動作成○”: 自動作成の翻訳精度が鳥バンクを用いた翻訳精度よりも優れている文
- “差なし”: ２種類の翻訳精度が同程度である文
- “同一出力”: ２種類の出力文が完全に同一な文

表 7.4: 鳥バンクと自動作成の対比較評価結果

鳥バンク○	自動作成○	差なし	同一出力
13	10	16	0

表 7.5: 英辞郎と自動作成の対比較評価結果

英辞郎○	自動作成○	差なし	同一出力
3	6	3	0

7.3 英語翻訳文の例

7.3.1 鳥バンク

表 7.6 に, 表 7.8, 表 7.10 に鳥バンク○の例を示す. 表中の表記方法を表 7.6 をもとに説明する.

- “日本語入力文”：入力された日本語文
- “参照文”：日本語入力文と対になっている英語文
- “日本語文パターン”：英語翻訳文を出力する際に選択された日本語文パターン
- “英語文パターン”：日本語文パターンに対応する英語文パターン
- “英語翻訳文”：出力された英語翻訳文

表 7.6: 鳥バンク○の例 1

	日本語入力文	この 下水 は よく 通る 。
	参照文	The sewer runs well .
鳥バンク	日本語文パターン	この $X1$ は よく $X2$ 。
	英語文パターン	$X1$ $X2$ well .
	英語翻訳文	The sewer pass well .
自動作成	日本語文パターン	この $X1$ は $X2$ 。
	英語文パターン	This $X1$ a $X2$.
	英語翻訳文	This is a good breeze .

表 7.7 に変数部に対応する対訳句を示す.

表 7.7: 変数対応

変数名	鳥バンク	自動作成
$X1$	下水：The sewer	下水：is
$X2$	通る：pass	よく 通る：good breeze

表 7.7 より, 自動作成は変数部に対応する対訳句 (“下水：is) が不自然である. よって, 鳥バンク○と判断した.

表 7.8: 鳥バンク◯の例 2

	日本語入力文	彼 の 考え方 は 極端 すぎる 。
	参照文	His way of thinking goes too far .
鳥バンク	日本語文パターン	$X1X2$ は $X3$ すぎる 。
	英語文パターン	$X1$ $X2$ is $X3$.
	英語翻訳文	His thinking is extremes .
自動作成	日本語文パターン	彼 の 考え方 は $X1$ $X2$ 。
	英語文パターン	His ideas $X1$ ahead of $X2$.
	英語翻訳文	His ideas are ahead of extreme

表 7.9 に変数部に対応する対訳句を示す.

表 7.9: 変数対応

変数名	鳥バンク	自動作成
$X1$	彼 の : His	極端 : extreme
$X2$	考え方 : thinking	すぎる : are
$X3$	極端 : extremes	

表 7.9 より, 自動作成は変数部に対応する対訳句 (“すぎる : are) が不自然である. よって, 鳥バンク◯と判断した.

表 7.10: 鳥バンク◯の例 3

	日本語入力文	私 は 率直 に 話し た 。
	参照文	I spoke plainly .
鳥バンク	日本語文パターン	私 は $X1X2$ た 。
	英語文パターン	$X1$ $X2$.
	英語翻訳文	spoke frankly .
自動作成	日本語文パターン	私 は $X1$ に $X2$ た 。
	英語文パターン	I $X1$ to $X2$.
	英語翻訳文	I spoke to me .

表 7.11 に変数部に対応する対訳句を示す.

表 7.11: 変数対応

変数名	鳥バンク	自動作成
$X1$	話し : spoke	話し : spoke
$X2$	率直 に : frankly	率直 に : me

表 7.11 より, 自動作成は変数部に対応する対訳句 (“率直 に : me) が不自然である. よって, 鳥バンク○と判断した.

表 7.12, 表 7.14, 表 7.16 に自動作成○の例を示す.

表 7.12: 自動作成○の例 1

	日本語入力文	その 都市 の 大半 が 焼失 し た 。
	正解文	Most of the city was burned into cinders .
鳥バンク	日本語文パターン	$X1$ $X2$ の $X3$ が $X4$ た 。
	英語文パターン	$X2$ $X1$ $X3$ $X4$.
	英語翻訳文	city The Most burned .
自動作成	日本語文パターン	$X1$ 大半 が 焼失 し た 。
	英語文パターン	The better part $X1$ was burnt down .
	英語翻訳文	The better part of the city was burnt down .

表 7.13 に変数に対応する対訳句を示す.

表 7.13: 変数対応

変数名	鳥バンク	自動作成
$X1$	その : The	その 都市 の : of the city
$X2$	都市 : city	
$X3$	大半 : Most	
$X4$	焼失 し : burned	

表 7.13, 鳥バンクを用いたパターンに基づく統計翻訳では, 変数に対応する対訳句はそれほど不自然ではない. しかし, 表 7.12 より, 文パターンに問題がある. よって自動作成○と判断した.

表 7.14: 自動作成○の例 2

	日本語入力文	彼女 は 彼 に 失望 し て い た 。
	正解文	She was disappointed in him .
鳥バンク	日本語文パターン	$X1$ は $X2$ に $X2$ て い た 。
	英語文パターン	$X1$ $X2$ with $X3$.
	英語翻訳文	She disappointed with He .
自動作成	日本語文パターン	彼女 は $X1$ に $X2$ て い た 。
	英語文パターン	She was $X2$ in $X1$.
	英語翻訳文	She disappointed in him .

表 7.15 に変数に対応する対訳句を示す.

表 7.15: 変数対応

変数名	鳥バンク	自動作成
$X1$	彼女：She	失望 し：disappointed
$X2$	失望 し：disappointed	彼：him
$X3$	彼：He	

表 7.16: 自動作成○の例 3

	日本語入力文	その 計画 は 成功 の 見込み が 十分 ある 。
	正解文	The plan bids fair to succeed .
鳥バンク	日本語文パターン	$X1$ は $X2$ が $X3$ ある 。
	英語文パターン	$X1$ $X3$ $X2$.
	英語翻訳文	it fully chance of success .
自動作成	日本語文パターン	$X1$ は 成功 の 見込み が $X2$ 。
	英語文パターン	$X1X2$ a chance of success .
	英語翻訳文	The projects has a chance of success .

表 7.17 に変数に対応する対訳句を示す.

表 7.17: 変数対応

変数名	鳥バンク	自動作成
$X1$	その 計画：it	その 計画：the project
$X2$	十分：fully	十分 ある：has
$X3$	成功 の 見込み：chance of success	

表 7.13, 表 7.15, 表 7.17 より, 鳥バンクを用いたパターンに基づく統計翻訳では, 変数に対応する対訳句はそれほど不自然ではない. しかし, 表 7.12, 表 7.14, 表 7.16 より, 文パターンに問題がある. よって自動作成○と判断した. ここで, 変数の数に注目すると, 自動作成は鳥バンクを用いたパターンに基づく統計翻訳より変数が少ない文パターンを選択している. そこで, 変数が少ない文パターンを選択したほうが翻訳精度が良いのではないかと考える.

表 7.18, 表 7.20, 表 7.22 に差なしの例を示す.

表 7.18: 差なしの例 1

	日本語入力文	彼女 に は 文学 の 素養 が ある 。
	正解文	She has learned a good deal of literature .
鳥バンク	日本語文パターン	$X1$ に は 文学 の 素養 が ある 。
	英語文パターン	$X1$ has literary culture .
	英語翻訳文	She has literary culture .
自動作成	日本語文パターン	彼女 に は $X1$ の 素養 が ある 。
	英語文パターン	She is versed in $X1$.
	英語翻訳文	She is versed in Japanese literature .

表 7.19 に変数に対応する対訳句を示す.

表 7.19: 変数対応

変数名	鳥バンク	自動作成
$X1$	彼女：She	文学：Japanese literature

表 7.20: 差なしの例 2

	日本語入力文	川 は その 湖 に 源 を 発し て いる 。
	正解文	The river rises from the lake .
鳥バンク	日本語文パターン	$X1$ は その 湖 に 源 を 発し て いる 。
	英語文パターン	$X1$ rises from the lake .
	英語翻訳文	river rises from the lake .
自動作成	日本語文パターン	彼女 $X1$ その 湖 に 源 を 発し て いる 。
	英語文パターン	$X1$ rises from the lake .
	英語翻訳文	The river rises from the lake .

表 7.21 に変数に対応する対訳句を示す.

表 7.21: 変数対応

変数名	鳥バンク	自動作成
$X1$	川：river	川 は：The river

表 7.22: 差なしの例 3

	日本語入力文	酒 は 米 から 作ら れる 。
	正解文	Sake is made from rice .
鳥バンク	日本語文パターン	$X1$ は $X2$ から $X3$ れる 。
	英語文パターン	$X1$ $X2$ from $X3$.
	英語翻訳文	liquor made from rice .
自動作成	日本語文パターン	$X1$ は $X2$ から 作ら れる 。
	英語文パターン	$X1$ is made from $X2$.
	英語翻訳文	Sake is made from rice .

表 7.23 に変数に対応する対訳句を示す.

表 7.23: 変数対応

変数名	鳥バンク	自動作成
$X1$	酒 : liquor	酒 : Sake
$X2$	米 : rice	米 : rice
$X3$	作ら : made	

表 7.18, 表 7.20, 表 7.22 より, 鳥バンクを用いたパターンに基づく統計翻訳と自動作成ともに, 比較的に変数の数が少ない文パターンを選択していることがわかる. また, 翻訳精度はどちらも高い. よって, 差なしと判断した.

7.3.2 英辞郎

表 7.24, 表 7.26, 表 7.28 に英辞郎○の例を示す.

表 7.24: 英辞郎○の例 1

	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 .
英辞郎	日本語文パターン	$X1$ は $X2X3$ 。
	英語文パターン	$X1$ will $X3X2$.
	英語翻訳文	union will participate in strike .
自動作成	日本語文パターン	$X1$ は $X2$ に $X3$ する 。
	英語文パターン	$X1X3$ in $X2$.
	英語翻訳文	The participants in the strike .

表 7.25 に変数に対応する対訳句を示す.

表 7.25: 変数対応

変数名	英辞郎	自動作成
$X1$	組合 : union	組合 : The
$X2$	ストライキ に : in strike	ストライキ : in the strike
$X3$	参加する : participate	参加 : participants

表 7.24 より, 自動作成では, 文パターンに問題がある. よって, 英辞郎○と判断した.

表 7.26: 英辞郎○の例 2

	日本語入力文	その 講座 は 1月 に 終わる 。
	正解文	The course finishes in January .
英辞郎	日本語文パターン	その $X1$ は $X2X3$ 。
	英語文パターン	The $X1$ $X3X2$.
	英語翻訳文	The lectureship end in January .
自動作成	日本語文パターン	その $X1$ は $X2$ に 終わる 。
	英語文パターン	The $X1$ ends $X2$.
	英語翻訳文	The chair ends of January .

表 7.27 に変数に対応する対訳句を示す.

表 7.27: 変数対応

変数名	英辞郎	自動作成
$X1$	講座：lectureship	講座：chair
$X2$	１月：January	１月：of January
$X3$	に 終わる：end in	

表 7.27 より, 自動作成は変数部に対応する対訳句 (“講座：chair) が不自然である. よって, 英辞郎○と判断した.

表 7.28: 英辞郎○の例 3

	日本語入力文	この 下水 は よく 通る 。
	正解文	The sewer runswell .
英辞郎	日本語文パターン	この $X1$ は $X2X3$ 。
	英語文パターン	This $X1$ is $X2X3$.
	英語翻訳文	This sewage is often pass .
自動作成	日本語文パターン	この $X1$ は $X2$ 。
	英語文パターン	This $X1$ a $X2$.
	英語翻訳文	This is a good breeze .

表 7.29 に変数に対応する対訳句を示す.

表 7.29: 変数対応

変数名	英辞郎	自動作成
$X1$	下水：sewage	下水：is
$X2$	よく：often	よく 通る：good breeze
$X3$	通る：pass	

表 7.29 より, 自動作成は変数部に対応する対訳句 (“下水：is) が不自然である. よって, 英辞郎○と判断した.

表 7.30, 表 7.32, 表 7.34 に自動作成○の例を示す.

表 7.30: 自動作成○の例 1

	日本語入力文	その 都市 の 大半 が 焼失 し た 。
	正解文	Most of the city was burned into cinders .
英辞郎	日本語文パターン	$X1X2$ 大半 が 焼失 し た 。
	英語文パターン	The better part $X2$ the $X1$ was burnt down .
	英語翻訳文	The better part urban the the was burnt down .
自動作成	日本語文パターン	$X1$ 大半 が 焼失 し た 。
	英語文パターン	The better part $X1$ was burnt down .
	英語翻訳文	The better part of the city was burnt down .

表 7.31 に変数に対応する対訳句を示す.

表 7.31: 変数対応

変数名	英辞郎	自動作成
$X1$	その : the	その 都市 の : of the city
$X2$	都市 の : urban	

表 7.32: 自動作成○の例 2

	日本語入力文	その ビル は 倒壊 の 危険 が ある 。
	正解文	The building is in danger of collapsing .
英辞郎	日本語文パターン	その $X1$ は $X2X3X4$ が ある 。
	英語文パターン	The $X1$ has a $X3X4X2$.
	英語翻訳文	The Bill has a of danger collapse .
自動作成	日本語文パターン	その ビル は $X1X2X3$ ある 。
	英語文パターン	The building is $X3X2X1$.
	英語翻訳文	The building is a danger of collapse .

表 7.33 に変数に対応する対訳句を示す.

表 7.33: 変数対応

変数名	英辞郎	自動作成
$X1$	ビル：Bill	倒壊：collapse
$X2$	倒壊：collapse	の：of
$X3$	の：of	危険 が：a danger
$X4$	危険：danger	

表 7.34: 自動作成○の例 3

	日本語入力文	彼女 は 彼 に 失望 し て い た 。
	正解文	She was disappointed in him.
英辞郎	日本語文パターン	その $X1$ は $X2X3X4$ 。
	英語文パターン	The $X1$ $X4X3$ a $X2$.
	英語翻訳文	The girlfriend could disappointment a him .
自動作成	日本語文パターン	彼女 は $X1$ に $X2$ て い た 。
	英語文パターン	She was $X2$ in $X1$.
	英語翻訳文	She disappointed in him .

表 7.35 に変数に対応する対訳句を示す.

表 7.35: 変数対応

変数名	英辞郎	自動作成
$X1$	彼女：girlfriend	彼：him
$X2$	彼 に：him	失望 し：disappointed
$X3$	失望：disappointment	
$X4$	し て い た：could	

表 7.31, 表 7.33, 表 7.35 より, 英辞郎を用いたパターンに基づく統計翻訳では, 鳥バンク用いたパターンに基づく統計翻訳と同様に, 変数に対応する対訳句はそれほど不自然ではない. しかし, 文パターンに問題がある. 変数の数も鳥バンクを用いたパターンに基づく統計翻訳と同様に, 自動作成よりも 変数が多い文パターンを選択している.

表 7.36, 表 7.38, 表 7.40 に差なしの例を示す.

表 7.36: 差なしの例 1

	日本語入力文	豊作 に なり そう だ 。
	正解文	The harvest looks promising .
英辞郎	日本語文パターン	$X1$ に なり そう だ 。
	英語文パターン	It looks like $X1$.
	英語翻訳文	It looks like abundant harvest .
自動作成	日本語文パターン	$X1$ に なり そう だ 。
	英語文パターン	I'm going to have a $X1$.
	英語翻訳文	I'm going to have a great crop .

表 7.37 に変数に対応する対訳句を示す.

表 7.37: 変数対応

変数名	英辞郎	自動作成
$X1$	豊作 : abundant harvest	豊作 : great crop

表 7.38: 差なしの例 2

	日本語入力文	その スーツ は ぴったり 合う 。
	正解文	The suit sets well .
英辞郎	日本語文パターン	その $X1$ は $X2X3$ 。
	英語文パターン	The $X1X3X2$.
	英語翻訳文	The suit conform exactly .
自動作成	日本語文パターン	その $X1$ は $X2$ 。
	英語文パターン	The $X1X2$.
	英語翻訳文	The suit fits me perfectly .

表 7.39 に変数に対応する対訳句を示す.

表 7.39: 変数対応

変数名	英辞郎	自動作成
$X1$	スーツ : suit	スーツ : suit
$X1$	ぴったり : exactly	ぴったり 合う : fits me perfectly
$X1$	合う : conform	

表 7.40: 差なしの例 3

	日本語入力文	殺害 者 に 血 の 復讐 を し た 。
	正解文	He took a bloody vengeance on the murderer .
英辞郎	日本語文パターン	$X1X2X3$ の $X4$ を $X5$ 。
	英語文パターン	$X2X1X5$ the $X4$ of $X3$.
	英語翻訳文	at slayer be the revenge of blood .
自動作成	日本語文パターン	$X1$ 者 に $X2X3$ を し た 。
	英語文パターン	She gave a cold $X3X1X2$.
	英語翻訳文	She gave a cold for him with blood .

表 7.41 に変数に対応する対訳句を示す.

表 7.41: 変数対応

変数名	英辞郎	自動作成
$X1$	殺害 者 : slayer	殺害 : him
$X2$	に : at	血 の : with blood
$X3$	血 : blood	復讐 : for
$X4$	復讐 : revenge	
$X5$	し た : be	

表 7.41 より, 英辞郎を用いたパターンに基づく統計翻訳と自動作成ともに, 変数の数が多い文パターンを選択している. また, 自動作成に至っては, 変数部に対応する対訳句も不自然である. その結果, どちらも翻訳精度が悪い. よって, 差なしと判断した.

7.4 実験結果のまとめ

7.1 節より, 人手作成された対訳句を用いたパターンに基づく統計翻訳 (以下, 人手作成) が自動作成よりカバー率が低いことがわかる. また, 7.2 節より, 人手作成と自動作成は翻訳精度に大きな差がないことがわかる. 人手作成において, 変数に対応する対訳句はそれほど不自然ではない. しかし, 文パターンに問題がある. そこで, 変数の数を比較すると, 自動作成は人手作成より変数の数が少ない文パターンを選択している. そして, 変数の数が少ない文パターンを選択した文の翻訳精度が良い傾向にある. よって, 変数の数が少ない文パターンを選択した方が翻訳精度が良いのではないかと考える.

また, 表 7.4 の“自動作成○ ”(9 文) において, 鳥バンクを用いたパターンに基づく統計翻訳のすべての変数のうち, 不自然な対応をとる対訳句の数を調査した. その結果, 25 個の変数において, 対応が不自然だった対訳句数は 3 句対であった. また, 表 7.5 の“自動作成○ ”(6 文) において, 英辞郎を用いたパターンに基づく統計翻訳のすべての変数のうち, 不自然な対応をとる対訳句の数を調査した. その結果, 19 個の変数において, 対応が不自然だった対訳句数は 2 句対であった. この結果より, 人手作成された対訳句にも多少は不自然な対応を取る対訳句が含まれていることがわかる.

第8章　考察

8.1　カバー率

本研究において, 鳥バンクを用いたパターンに基づく統計翻訳と英辞郎を用いたパターンに基づく統計翻訳ともに, 自動作成よりカバー率が低かった. この原因として, 句に基づく文パターン辞書の文パターン対数の差が考えられる. 表 7.1, 表 7.2, 表 7.3 より, 鳥バンクを用いたパターンに基づく統計翻訳の文パターン対数は, 自動作成の約 100 分の 1 である. また, 英辞郎を用いたパターンに基づく統計翻訳の文パターン対数は, 自動作成の約 250 分の 1 である. この結果より, より多くの文パターンを作成することでカバー率が高くなると考える.

8.2　翻訳精度

本研究において, 人手作成された対訳句を用いたパターンに基づく統計翻訳と対訳句を自動作成するパターンに基づく統計翻訳では, 翻訳精度に大きな差がなかった. 翻訳結果を調査したところ, 鳥バンクを用いたパターンに基づく統計翻訳の“自動作成○ ”において, 10 文中 5 文は自動作成が鳥バンクを用いたパターンに基づく統計翻訳より変数が少ない文パターンを選択していた. 一方, 鳥バンクを用いたパターンに基づく統計翻訳が自動作成より変数が少ない文パターンを選択していた翻訳文は 10 文中 0 文だった. また, 英辞郎を用いたパターンに基づく統計翻訳の“自動作成○ ”では, 6 文中 6 文が自動作成が英辞郎を用いたパターンに基づく統計翻訳より変数が少ない文パターンを選択していた. よって, 自動作成の翻訳精度が良かった要因として, 人手作成より変数が少ない文パターンを選択していたことが挙げられる.

また, 鳥バンクを用いたパターンに基づく統計翻訳の“人手作成○ ”において, 対訳句が不自然なために, 自動作成の翻訳精度が悪かったと判断した翻訳文が 13 文中 11 文であった. そして, 英辞郎を用いたパターンに基づく統計翻訳の“人手作成○ ”において, 対訳句が不自然なために, 自動作成の翻訳精度が悪かったと判断した翻訳文が 3 文中 2 文で

あった.

この結果より, 人手作成された対訳句と自動作成で得られた句に基づく文パターン辞書を用いて翻訳を行うことで, 翻訳精度が向上すると考える.

本研究の類似研究として, 統計翻訳において, 対訳コーパスに対訳句コーパスを追加する手法があげられる. Maja Popović らはセルビア語英語間, スペイン語英語間において, 対訳コーパスに対訳句コーパスを追加し句に基づく統計翻訳を行った [13]. その結果, 自動評価結果が向上したと報告されている.

8.3 鳥バンクと英辞郎の違い

本研究では, 人手作成された対訳句として鳥バンクと英辞郎を用いた. 実験結果より, 鳥バンクと英辞郎ともに, 変数部に対応する対訳句は信頼性が高い. しかし, 変数の数が多い文パターンを選択している翻訳文は翻訳精度が悪くなる傾向にある.

鳥バンクと英辞郎の大きな違いはカバー率である. 鳥バンクを用いたパターンに基づく統計翻訳のカバー率は 42%, 英辞郎を用いたパターンに基づく統計翻訳のカバー率は 13%であった. この原因として, 分野の依存性が考えられる.

第9章　おわりに

パターンに基づく統計翻訳は, 統計的手法を用いて, 対訳句と文パターンを自動作成して翻訳を行う. しかし, 自動作成された対訳句の問題として, 対応する原言語と目的言語が不自然な対訳句が多く含まれていることがあげられる.

そこで, 本研究では, 人手作成された対訳句として鳥バンクと英辞郎を用いて, パターンに基づく統計翻訳を行った. そして, 自動作成との対比較評価により, 翻訳精度とカバー率の調査を行った. 実験の結果, 鳥バンク○が 13 文 (自動作成○ 10 文), 英辞郎○が 3 文 (自動作成○ 6 文) であり, 自動作成と比べて翻訳精度に大きな差はなかった. しかし, カバー率は, 鳥バンクを用いたパターンに基づく統計翻訳は 42%, 英辞郎を用いたパターンに基づく統計翻訳は 13%, 自動作成は 82%であり, 人手作成された対訳句を用いたパターンに基づく統計翻訳が自動作成より低かった.

第10章　追加実験

追加実験として, 人手作成された対訳句と自動作成で得られた句に基づく文パターン辞書を用いてパターンに基づく統計翻訳を行い (以下, 追加実験), 江木らの手法 (以下, 自動作成) および鳥バンクを用いたパターンに基づく統計翻訳と対比較評価を行った.

10.1　実験環境

追加実験では, 単文コーパスとして対訳文とテストデータを用いる. また, 人手で作成された対訳句として鳥バンクを用いる. また, 自動作成で得られた句に基づく文パターン辞書を用いる. 表 10.1 に追加実験で用いるデータ数を示す.

表 10.1: 使用データ数

対訳文	100,000 文対
テストデータ	100 文
鳥バンク	336,161 句対
句に基づく文パターン辞書	308,566,385 パターン対

10.2　対比較評価結果

10.2.1　追加実験と自動作成

追加実験において, 日本語入力文 100 文に対して, 英語翻訳文 71 文を得られた. 得られた英語翻訳文 71 文のうち自動作成と出力が重なった 69 文に対して, 対比較評価を行った. 表 10.2 に追加実験と自動作成の対比較評価結果を示す.

表 10.2: 追加実験と自動作成の対比較評価結果

追加実験○	自動作成○	差なし	同一出力
20	14	35	0

表 10.3, 表 10.5, 表 10.7 に追加実験○の例を示す.

表 10.3: 追加実験○の例 1

	日本語入力文	水 が 腐っ て いる。
	参照文	The water is foul .
追加実験	日本語文パターン	水 が $X1$ て いる 。
	英語文パターン	The water is $X1$.
	英語翻訳文	The water is rotten .
自動作成	日本語文パターン	水 が $X1$ て いる 。
	英語文パターン	The water is at the $X1$.
	英語翻訳文	The water is at the right .

表 10.4 に変数に対応する対訳句を示す.

表 10.4: 変数対応

変数名	追加実験	自動作成
$X1$	腐っ : rotten	腐っ : right

表 10.4 より, 自動作成は変数に対応する対訳句 (“腐っ : right) が不自然である. また, 選択された日本語文パターンと英語文パターンの字面の対応が悪い. よって, 追加実験○と判断した.

表 10.5: 追加実験○の例 2

	日本語入力文	彼 の 考え方 は 極端 すぎる 。
	参照文	His way of thinking goes too far .
追加実験	日本語文パターン	彼 の $X1$ は $X2$ すぎる 。
	英語文パターン	His $X1$ is too $X2$.
	英語翻訳文	His idea is too extremes .
自動作成	日本語文パターン	彼 の 考え方 は $X1$ $X2$ 。
	英語文パターン	His ideas $X1$ ahead of $X2$.
	英語翻訳文	His ideas are ahead of the extreme .

表 10.6 に変数に対応する対訳句を示す.

表 10.6: 変数対応

変数名	追加実験	自動作成
$X1$	考え方 : idea	極端 : the extreme
$X2$	極端 : extremes	すぎる : are

表 10.6 より, 自動作成は変数に対応する対訳句 (“すぎる : are) が不自然である. また, 選択された日本語文パターンと英語文パターンの字面の対応が悪い. よって, 追加実験〇と判断した.

表 10.7: 追加実験〇の例 3

	日本語入力文	ウイスキー を 1 杯 もらお う 。
	参照文	I'll have a whiskey .
追加実験	日本語文パターン	$X1$ もらお う 。
	英語文パターン	I'll have $X1$.
	英語翻訳文	I'll have a glass of whisky .
自動作成	日本語文パターン	$X1$ を $X2X3$ う。
	英語文パターン	$X3$ go for $X2X1$.
	英語翻訳文	I'll go for a glass with whisky .

表 10.8 に変数に対応する対訳句を示す.

表 10.8: 変数対応

変数名	追加実験	自動作成
$X1$	ウイスキー を 1 杯 : a glass of whisky	ウイスキー : with whisky
$X2$		1 杯 : a glass
$X3$		もらお : I'll

表 10.8 より, 自動作成は追加実験より変数の数が多い文パターンを選択しているため翻訳精度が悪くなったと考えられる.

表 10.9, 表 10.11, 表 10.13 に自動作成〇の例を示す.

表 10.9: 自動作成〇の例 1

	日本語入力文	。彼女 は 彼 に 失望 し て いた 。
	参照文	She was desappointed in him .
追加実験	日本語文パターン	彼女 は $X1$ に $X2$ て いた 。
	英語文パターン	She was $X2$ in $X1$.
	英語翻訳文	She was desappointed in his house .
自動作成	日本語文パターン	彼女 は $X1$ に $X2$ て いた 。
	英語文パターン	She was $X2$ in $X1$.
	英語翻訳文	She was desappointed in him .

表 10.10 に変数に対応する対訳句を示す.

表 10.10: 変数対応

変数名	追加実験	自動作成
$X1$	彼：his house	彼：him
$X2$	失望 し：disappointed	失望 し：disappointed

表 10.10 より, 追加実験では, 変数に対応する対訳句 (彼： his house) が不自然である. よって, 自動作成○と判断した.

表 10.11: 自動作成○の例 2

	日本語入力文	繁華 街 で 息子 を 見失っ て しまっ た 。
	参照文	I've lost sight of my son on the busy street .
追加実験	日本語文パターン	$X1$ で $X2$ を $X3$ て しまっ た 。
	英語文パターン	He $X3X2$ with $X1$.
	英語翻訳文	He lost himself with the downtown .
自動作成	日本語文パターン	$X1X2$ を $X3$ て しまっ た 。
	英語文パターン	He has $X3X2X1$.
	英語翻訳文	He has lost his son in a shopping district .

表 10.12 に変数に対応する対訳句を示す.

表 10.12: 変数対応

変数名	追加実験	自動作成
$X1$	繁華 街：the downtown	繁華 街：shopping district
$X2$	息子：himself	で 息子：his son in a
$X3$	見失っ：lost	見失っ：lost

表 10.12 より, 追加実験では, 変数に対応する対訳句 (息子： himself) が不自然である. よって, 自動作成○と判断した.

表 10.13: 自動作成○の例 3

	日本語入力文	酒 は 米 から 作ら れる 。
	参照文	Sake is made from rice .
追加実験	日本語文パターン	$X1$ は $X2$ から 作ら れる 。
	英語文パターン	$X1$ is made from $X2$.
	英語翻訳文	Sake is made from the United States.
自動作成	日本語文パターン	$X1$ は $X2$ から 作ら れる 。
	英語文パターン	$X1$ is made from $X2$.
	英語翻訳文	Sake is made from rice .

表 10.14 に変数に対応する対訳句を示す.

表 10.14: 変数対応

変数名	追加実験	自動作成
$X1$	酒 : sake	酒 : sake
$X2$	米 : the United States	米 : rice

表 10.15, 表 10.17, 表 10.19 に差なしの例を示す.

表 10.15: 差なし○の例 1

	日本語入力文	陽気 な ふるまい を する 。
	参照文	She has a vivacious manner .
追加実験	日本語文パターン	$X1$ な $X2$ を $X3$ 。
	英語文パターン	He $X3$ a $X1$ of $X2$.
	英語翻訳文	He makes a weather of behavior.
自動作成	日本語文パターン	$X1$ な $X2$ する 。
	英語文パターン	$X2$ is $X1$.
	英語翻訳文	He is bright .

表 10.16 に変数に対応する対訳句を示す.

表 10.16: 変数対応

変数名	追加実験	自動作成
$X1$	陽気 : weather	陽気 : bright
$X2$	ふるまい : behavior	ふるまい を : He
$X1$	する : makes	

表 10.17: 差なし○の例 2

	日本語入力文	彼女 に は 文学 の 素養 が ある 。
	参照文	She has learned a good deal of literature .
追加実験	日本語文パターン	$X1$ に は 文学 の 素養 が ある 。
	英語文パターン	$X1$ has literary culture .
	英語翻訳文	The girl has literary culture .
自動作成	日本語文パターン	彼女 に は $X1$ の 素養 が ある 。
	英語文パターン	She is versed in $X1$.
	英語翻訳文	She is versed in Japanese literature .

表 10.18 に変数に対応する対訳句を示す.

表 10.18: 変数対応

変数名	追加実験	自動作成
$X1$	彼女 : The girl	文学 : Japanese literature

表 10.19: 差なし○の例 3

	日本語入力文	子供 たち は 寝室 へ 立ち去っ た。
	参照文	The children disappeared to their bedrooms.
追加実験	日本語文パターン	$X1$ は $X2$ へ $X3$ た 。
	英語文パターン	$X1X3$ to the $X2$.
	英語翻訳文	The boys went away to the bedroom .
自動作成	日本語文パターン	$X1$ たち は $X2$ へ $X3$ た。
	英語文パターン	$X1X3$ to the $X2$.
	英語翻訳文	The children are away to the bedroom.

表 10.20 に変数に対応する対訳句を示す.

表 10.20: 変数対応

変数名	追加実験	自動作成
$X1$	子供 たち : The boys	子供 : The children are
$X2$	寝室 : bedroom	寝室 : bedroom
$X3$	立ち去っ : went away	立ち去っ : away

10.2.2 追加実験と提案手法 (鳥バンク)

追加実験で得られた英語翻訳文 71 文のうち鳥バンクを用いたパターンに基づく統計翻訳と出力が重なった 39 文に対して, 対比較評価を行った. 表 10.21 に追加実験と鳥バンクを用いたパターンに基づく統計翻訳の対比較評価結果を示す.

表 10.21: 追加実験と鳥バンクの対比較評価結果

追加実験○	鳥バンク○	差なし	同一出力
10	7	22	0

表 10.22, 表 10.24, 表 10.26 に追加実験○の例を示す.

表 10.22: 追加実験○の例 1

	日本語入力文	彼 は こんな 要求 に 素直 に うんと 言う 人間 で ない 。
	参照文	He is not a man to give ready consent to such a demand as this .
追加実験	日本語文パターン	彼 は $X1X2$ で ない 。
	英語文パターン	He is not $X2$ tell a $X1$.
	英語翻訳文	He is not the man tell a give ready consent to such a demand as this .
鳥バンク	日本語文パターン	彼 は $X1X2$ で $X3$ 。
	英語文パターン	He $X3X1X2$.
	英語翻訳文	He no give ready consent to such a demand as this human .

表 10.23 に変数に対応する対訳句を示す.

表 10.23: 変数対応

変数名	追加実験	鳥バンク
$X1$	こんな 要求 に 素直 に うんと 言う : give ready consent to such a demand as this	こんな 要求 に 素直 に うんと 言う : give ready consent to such a demand as this
$X2$	人間 : the man	人間 : human
$X3$		ない : no

表 10.23 より, 鳥バンクでは, 選択した文パターンが悪い. よって, 追加実験○と判断した.

表 10.24: 追加実験○の例 2

	日本語入力文	水 が 腐っ て いる 。
	参照文	The water is foul .
追加実験	日本語文パターン	水 が $X1$ て いる 。
	英語文パターン	The water is $X1$.
	英語翻訳文	The water is rotten .
鳥バンク	日本語文パターン	$X1$ が $X2$ て いる 。
	英語文パターン	$X1$ $X2$.
	英語翻訳文	water rotten .

表 10.25 に変数に対応する対訳句を示す.

表 10.25: 変数対応

変数名	追加実験	自動作成
$X1$	腐っ : rotten	水 : water
$X2$		腐っ : rotten

表 10.25 より, 鳥バンクにおいて, 選択した日本語文パターンと英語文パターンの字面の対応が悪い. よって, 追加実験○と判断した.

表 10.26: 追加実験○の例 3

	日本語入力文	子供 たち は 寝室 へ 立ち去っ た 。
	参照文	The children disappeared to their bedrooms .
追加実験	日本語文パターン	$X1$ は $X2$ へ $X3$ た 。
	英語文パターン	$X1X3$ to the $X2$.
	英語翻訳文	The boys went away to the bedroom .
鳥バンク	日本語文パターン	$X1$ は $X2$ へ $X3$ た 。
	英語文パターン	$X1X3X2$.
	英語翻訳文	children left bedroom .

表 10.27 に変数に対応する対訳句を示す.

表 10.27: 変数対応

変数名	追加実験	自動作成
$X1$	子供 たち : The boys	子供 たち : children
$X2$	寝室 : bedroom	寝室 : bedroom
$X3$	立ち去っ : went away	立ち去っ : left

表 10.27 より, 鳥バンクは日本語入力文の翻訳としてふさわしくない. よって, 追加実験○と判断した.

表 10.28, 表 10.30, 表 10.32 に追加実験○の例を示す.

表 10.28: 鳥バンク○の例 1

	日本語入力文	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 .
追加実験	日本語文パターン	$X1$ は たった $X2$ しか 持っ て い ない 。
	英語文パターン	$X1$ has no more than $X2$.
	英語翻訳文	I have no more than one thousand yen .
鳥バンク	日本語文パターン	$X1$ は たった $X2$ しか 持っ て い ない 。
	英語文パターン	$X1$ have no more than $X2$.
	英語翻訳文	He have no more than 1000 yen .

表 10.29 に変数に対応する対訳句を示す.

表 10.29: 変数対応

変数名	追加実験	鳥バンク
$X1$	彼 : I	彼 : He
$X2$	１ ０ ０ ０ 円 : one thousand yen	１ ０ ０ ０ 円 : 1000 yen

表 10.29 より, 追加実験では, 変数に対応する対訳句 (彼 : I) が不自然である. よって, 鳥バンク○と判断した.

表 10.30: 鳥バンク○の例 2

	日本語入力文	酒 は 米 から 作ら れる 。
	参照文	Sake is made from rice .
追加実験	日本語文パターン	$X1$ は $X2$ から 作ら れる 。
	英語文パターン	$X1$ is made from $X2$.
	英語翻訳文	Sake is made from the United States .
鳥バンク	日本語文パターン	$X1$ は $X2$ から $X3$ れる 。
	英語文パターン	$X1$ $X3$ from $X2$.
	英語翻訳文	liquor made from rice .

表 10.31 に変数に対応する対訳句を示す.

表 10.31: 変数対応

変数名	追加実験	鳥バンク
$X1$	酒：sake	酒：liquor
$X2$	米：the United States	米：rice
$X3$		作ら：mede

表 10.31 より, 追加実験は翻訳精度が悪い. よって, 鳥バンク○と判断した.

表 10.32: 鳥バンク○の例 3

	日本語入力文	荷物 は 駅 に 留め て あり ます 。
	参照文	Your packages are being held at the station .
追加実験	日本語文パターン	$X1$ は $X2$ に $X3$ て あり ます 。
	英語文パターン	I have the $X1X3$ at a $X2$.
	英語翻訳文	I have the goods is fixed at a the railway station .
鳥バンク	日本語文パターン	$X1$ は $X2$ に $X3$ て あり ます 。
	英語文パターン	$X1$ is $X3$ in $X2$.
	英語翻訳文	luggage is fasten in station .

表 10.33 に変数に対応する対訳句を示す.

表 10.33: 変数対応

変数名	追加実験	自動作成
$X1$	荷物：goods	荷物：luggage
$X2$	駅：the railway station	駅：station
$X3$	留め：is fixed	留め：fasten

表 10.33 より, 追加実験では, 変数に対応する対訳句 (荷物： goods) が不自然である. よって, 鳥バンク○と判断した.

表 10.27 より, 鳥バンクは日本語入力文の翻訳としてふさわしくない. よって, 追加実験○と判断した.

表 10.34, 表 10.36, 表 10.38 に差なしの例を示す.

表 10.34: 差なしの例 1

	日本語入力文	信仰 は 山 を も 動かす 。
	参照文	Faith can move mountains .
追加実験	日本語文パターン	$X1$ は $X2$ を も 動かす 。
	英語文パターン	$X1$ make the $X2$.
	英語翻訳文	beliefs makes the hill .
鳥バンク	日本語文パターン	$X1$ は $X2$ を $X3X4$.
	英語文パターン	$X1X4X2X3$.
	英語翻訳文	faith move mountain also .

表 10.35 に変数に対応する対訳句を示す.

表 10.35: 変数対応

変数名	追加実験	鳥バンク
$X1$	信仰 : belefs	信仰 : faith
$X2$	山 : hill	山 : mountain
$X3$		も : also
$X4$		動かす : move

表 10.35 より, 追加実験, 鳥バンクともに翻訳精度が悪い. よって, 差なしと判断した.

表 10.36: 差なしの例 2

	日本語入力文	彼女 に は 文学 の 素養 が ある 。
	参照文	She has learned a good deal of literature .
追加実験	日本語文パターン	$X1$ に は 文学 の 素養 が ある 。
	英語文パターン	$X1$ has literary culture .
	英語翻訳文	The girl has literary culture .
鳥バンク	日本語文パターン	$X1$ に は 文学 の 素養 が ある 。
	英語文パターン	$X1$ has literary culture .
	英語翻訳文	She has literary culture .

表 10.37 に変数に対応する対訳句を示す.

表 10.37: 変数対応

変数名	追加実験	鳥バンク
$X1$	彼女 : The girl	彼女 : She

表 10.38: 差なしの例 3

	日本語入力文	新車 の 性能 の 試験 を し た 。
	参照文	The performance test of new cars was held .
追加実験	日本語文パターン	$X1$ の $X2$ の $X3$ を $X4$ 。
	英語文パターン	I $X4X3$ in $X2$ of my $X1$.
	英語翻訳文	I took a test in the performance of my new car .
鳥バンク	日本語文パターン	$X1$ の $X2$ の $X3$ を $X4$ 。
	英語文パターン	He $X4X3X2$ the $X1$.
	英語翻訳文	He made examination performance the new car .

表 10.39 に変数に対応する対訳句を示す.

表 10.39: 変数対応

変数名	追加実験	鳥バンク
$X1$	新車 : new car	新車 : new car
$X2$	性能 : the performance	性能 : performance
$X3$	試験 : a test	試験 : examination
$X4$	し : took	し : made

10.3 追加実験結果のまとめ

10.2.1 節より, 追加実験と自動作成は翻訳精度に大きな差がないことがわかる. また, カバー率も大きな差がないことがわかる. 10.2.2 節より, 追加実験と鳥バンクを用いたパターンに基づく統計翻訳は翻訳精度に大きな差がないことがわかる.

表 10.2 の“自動作成○ ”(14 文) において, 追加実験のすべての変数のうち, 不自然な対応をとる対訳句の数を調査した. その結果, 40 個の変数において, 対応が不自然だった対訳句数は 12 句対であった.

追加実験を通して, 文パターンの数を増やすことでカバー率が高くなることがわかった. また, 人手作成された対訳句にも対応が不自然な対訳句が含まれていることがわかった. 今後は, 誤り解析を行い, 翻訳精度の向上を目指したい.

謝辞

最後に，1 年間に渡りご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工学講座 C 研究室の村上仁一准教授, 徳久雅人講師, 村田真樹教授そして, 春野瑞季さん, 力久剛士さんをはじめ, 計算機工学講座 C 研究室の方々に厚く御礼申し上げます.

また，参考にさせていただいた論文の著者の方々に対して深く感謝申し上げます.

参考文献

[1] Hiroshi Maruyama:“ Pattern-Based Translation: Context-Free Transducer and Its Applications to Practical NLP”, n Proc.of Natural Language Pacific Rim Symposium, pp.232-237, 1993.

[2] 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

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

[4] 英辞郎：http://www.alc.co.jp/

[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] Peter F.Brown, Stephen A.Della Pietra, Vincent J.Della Pietra, Robert L.Mercer, “The mathematics of statistical machine translation:Parameter Estimation” , Computational Linguistics, 1993.

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

[8] MeCab: Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis”, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237, 2004.

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

[10] Moses: Philipp Koehn, Marcello Federico, Brooke Cowan, Richard Zens, Chris Dyer, Ondrej 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.

[11] SRILM: Andreas Stolcke, "SRILM - am Extensible Language Modeling Toolkit", 7th International Conference on Spoken Language Processing, pp.901-904, 2002.

[12] Mert: 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.

[13] Popović Maja, and Ney Hermann "Statistical Machine Translation with a small amount of bilingual training data", 5th LREC SALTMIL Workshop on Minority Languages. 2006.