概要

安場らは機械翻訳のあらたな手法として相対的意味論に基づく変換主導型統計機械翻訳 (TDMT:Transfer Statistical Driven Machine Translation)（以下 TDSMT と訳す）[1] を提案した．TDSMT とは，学習文対と変換テーブルを用いて，原言語文を入力とし，目的言語文を出力する手法である．変換テーブルでは,“ A が B ならば“ Cが D ”で表現する．日英翻訳の例では，A は学習文対の日本語句，B は学習文対の英語句，C は入力文の日本語句，D は出力文の英語句に当たる．しかし，従来変換テーブルの A, B 使用されるのは対訳単語であり、対訳句は利用不可である．

そこで本研究では，対訳句の活用を考える．まず対訳句の代わりに日本語の単語と英語の単語を 2:2 で組み合わせた対訳単語（以下対訳 2 単語連続と訳す）を作成する手法を提案する．また，作成した対訳 2 単語連続を人手評価し，対訳句としての精度調査を行った．

提案手法により 101,354 語の対訳 2 単語連続の抽出が可能であった．また，評価の結果より，対訳 2 単語連続は対訳句として有効な翻訳を行うことが確認できた．

今後は，対訳 2 単語連続の精度向上，対訳 2 単語連続のみでなく日本語の単語と英語の単語を 1:3，2:3，3:1 などに分ける活用などが考えられる．また，対訳 2 単語連続の翻訳での活用を考える必要がある．

目 次

表 目 次

第1章　はじめに

機械翻訳の手法として，パターン翻訳，統計翻訳等が研究されてきた．

パターン翻訳 [2] は，1960 年代に提案された翻訳方法である．人手により作成した，対訳句辞書と対訳文パターン辞書を用いて翻訳を行う．この翻訳方式は入力文が適切な対訳文パターンに適合した場合，翻訳精度の高い出力文が得られる．しかし，対訳句辞書と対訳文パターン辞書の作成は人手で行うため，開発にコストがかかる．そして，入力文が対訳文パターンに適合しない場合は，翻訳ができない．

また，1990 年代に単語に基づく統計翻訳が提案された．原言語文の単語を目的言語文の単語に翻訳する手法である．しかし，翻訳精度が低い．しかし，2000 年代始めに句に基づく統計翻訳が提案された．句に基づく統計翻訳は，単語に基づく統計翻訳よりも翻訳精度が高く，学習データとして，対訳文を与えるだけで翻訳が可能である．そのため翻訳にかかるコストが低い．

安場らは機械翻訳のあらたな手法として TDSMT を提案した．TDSMT とは，学習文対と変換テーブルを用いて，原言語文を入力とし，目的言語文を出力する手法である．変換テーブルでは，“ A が B ならば“ C が D ”で表現する．日英翻訳の例では，A は学習文対の日本語句，B は学習文対の英語句，C は入力文の日本語句，D は出力文の英語句に当たる．しかし，従来では，１単語でしか利用できない．したがって，変換テーブルの A，B 使用されるのは対訳単語であり、対訳句は利用不可である．

そこで本研究では，対訳句の活用を考える．まず対訳句の代わりに対訳２単語連続を作成する手法を提案する．また，作成した対訳２単語連続を人手評価し，対訳句としての精度調査を行った．

本論文の構成を以下に示す．第 2 章では，従来手法について説明し，第 3 章では，提案手法について説明する．第 4 章では，実験概要について示し，第 5 章では，実験結果を示す．第 6 章では，比較実験について示す．最後に第 7 章で，考察について述べる．

第2章　従来手法

従来手法での変換テーブルの作成方法について，日本語と英語の場合を例にして説明する．なお，本章は安場裕人，村上仁一: “変換主導型統計機械翻訳の提案”第 3 章を参考にしている．

2.1　概要

変換テーブルは,“ A が B ならば“ C が D ”で表現する．日英翻訳の例では，A は学習文対の日本語句，B は学習文対の英語句，C は入力文の日本語句，D は出力文の英語句に当たる．変換テーブルの A，B は従来，対訳単語から作成される．対訳単語とは，異なる二言語間において，それぞれ対訳関係にある単語の対である．この対訳単語は学習文対と対訳単語確率（IBM モデル 1）を利用して作成される．

2.2　IBM翻訳モデル

IBM 翻訳モデルを以下に示す．これは，力久ら [2] の抜粋である．統計翻訳の代表的なモデルとして，IBM の Brown らによる仏英翻訳モデルがある．IMB 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．この翻訳モデルは順に複雑な計算を行うモデル 1 から 5 の 5 つのモデルで構成される．

本章では．原言語であるフランス語文を F，目的言語である英語文を E として定義する．

IBM モデルでは，フランス語文 E，英語文 F の翻訳モデル $P(F|E)$ を計算するために，アライメント a を用いる．以下に IBM モデルの基本式を示す．

$$P(F|E) = \sum_{a} P(F, a|E) \tag{2.1}$$

アライメントとは仏単語と英単語の対応を意味している．IBM モデルのアライメントでは，各仏単語 f に対応する英単語 e は 1 つあり，各英単語 e に対応する仏単語は 0 か

ら n 個ある．また仏単語 f において適切な英単語と対応しない場合，英語文の先頭に空単語 e_0 があると仮定し，その仏単語 f と空単語 e_0 を対応づける．

・モデル 1

(2.1) 式は以下の式に分解することができる．m はフランス語文の長さ，a_1^{j-1} はフランス語文における，1 番目から $j-1$ 番目までのアライメント，f_1^{j-1} はフランス語文における，1 番目から $j-1$ 番目まで単語を表している．

$$P(F,a|E) = P(m|E)\prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, E)P(f_j|a_1^{j}, f_1^{j-1}, m, E) \tag{2.2}$$

(2.2) 式ではとても複雑であるので計算が困難である．そこで，モデル 1 では以下の仮定により，パラメータの簡略化を行う．

- フランス語文の長さの確率 ϵ は m, E に依存しない

$$P(m|E) = \epsilon$$

- アライメントの確率は英語文の長さ l に依存する

$$P(a_j|a_1^{j-1}, f_1^{j-1}, m, E) = (l+1)^{-1}$$

- フランス語の翻訳確率 $t(f_j|e_{a_j})$ は，仏単語 f_j に対応する英単語 e_{a_j} に依存する

$$P(f_j|a_1^{j}, f_1^{j-1}, m, e) = t(f_j|e_{a_j})$$

パラメータの簡略化を行うことで，$P(F,a|E)$ と $P(F,E)$ は以下の式で表される．

$$P(F,a|E) = \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.3}$$

$$P(F|E) = \frac{\epsilon}{(l+1)^m}\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.4}$$

$$= \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m}\sum_{i=0}^{l} t(f_j|e_{a_j}) \tag{2.5}$$

モデル 1 では翻訳確率 $t(f|e)$ の初期値が 0 以外の場合，Expectation-Maximization(EM) アルゴリズムを繰り返し行うことで得られる期待値を用いて最適解を推定する．EM アルゴリズムの手順を以下に示す．

手順 1 翻訳確率 $t(f|e)$ の初期値を設定する.

手順 2 仏英対訳対 $(F^{(s)}, E^{(s)})$(但し，$1 \leq s \leq S$) において，仏単語 f と英単語 e が対応する回数の期待値を以下の式により計算する.

$$c(f|e; F, E) = \frac{t(f|e)}{t(f|e_0) + \cdots + t(f|e_l)} \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=0}^{l} \delta(e, e_i) \tag{2.6}$$

$\delta(f, f_j)$ はフランス語文 F 中で仏単語 f が出現する回数，$\delta(e, e_i)$ は英語文 E 中で英単語 e が出現する回数を表している.

手順 3 英語文 $E^{(s)}$ の中で 1 回以上出現する英単語 e に対して，翻訳確率 $t(f|e)$ を計算する.

1. 定数 λ_e を以下の式により計算する.

$$\lambda_e = \sum_{f} \sum_{s=1}^{S} c(f|e; F^{(s)}, E^{(s)}) \tag{2.7}$$

2. (2.7) 式より求めた λ_e を用いて，翻訳確率 $t(f|e)$ を再計算する.

$$\begin{aligned} t(f|e) &= \lambda_e^{-1} \sum_{s=1}^{S} c(f|e; F^{(s)}, E^{(s)}) \\ &= \frac{\sum_{s=1}^{S} c(f|e; F^{(s)}, E^{(s)})}{\sum_{f} \sum_{s=1}^{S} c(f|e; F^{(s)}, E^{(s)})} \end{aligned} \tag{2.8}$$

手順 4 翻訳確率 $t(f|e)$ が収束するまで手順 2 と手順 3 を繰り返す.

・モデル 2

モデル 1 では，全ての単語の対応に対して，英語文の長さ l にのみ依存し，単語対応の確率を一定としている．そこで，モデル 2 では，j 番目の仏単語 f_j と対応する英単語の位置 a_j は英語文の長さ l に加えて，j と，フランス語文の長さ m に依存し，以下のような関係とする.

$$a(a_j|j, m, l) \equiv P(a_j|a_1^{j-1}, f_1^{j-1}, m, l) \tag{2.9}$$

この関係からモデル 1 における (2.4) 式は，以下の式に変換できる．

$$P(F|E) = \epsilon \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) a(a_j|j,m,l) \tag{2.10}$$

$$= \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_{a_j}) a(a_j|j,m,l) \tag{2.11}$$

モデル 2 では，期待値は $c(f|e;F,e)$ と $c(i|j,m,l;F,E)$ の 2 つが存在する．以下の式から求められる．

$$c(f|e;F,E) = \frac{t(f|e)}{t(f|e_0)+\cdots+t(f|e_l)} \sum_{j=1}^{m} \delta(f,f_j) \sum_{i=1}^{l} \delta(e,e_i) \tag{2.12}$$

$$= \sum_{j=1}^{m} \sum_{i=0}^{l} \frac{t(f|e)a(i|j,m,l)\delta(f,f_j)\delta(e,e_i)}{t(f|e_0)a(0|j,m,l)+\cdots+t(f|e_l)a(l|j,m,l)} \tag{2.13}$$

$$c(i|j,m,l;F,E) = \sum_{a} P(a|E,F)\delta(i,a_j) \tag{2.14}$$

$$= \frac{t(f_j|e_i)a(i|j,m,l)}{t(f_j|e_0)a(0|j,m,l)+\cdots+t(f_j|e_l)a(l|j,m,l)} \tag{2.15}$$

$c(f|e;F,E)$ は対訳文中の英単語 e と仏単語 f が対応付けされる回数の期待値，$c(i|j,m,l;F,E)$ は英単語の位置 i が仏単語の位置 j に対応付けされる回数の期待値を表している．

モデル 2 では，EM アルゴリズムで計算すると複数の極大値が算出され，最適解が得られない可能性がある．モデル 1 では $a(i|j,m,l) = (l+1)^{-1}$ となるモデル 2 の特殊な場合であると考えられる．したがって，モデル 1 を用いることで最適解を得ることができる．

・モデル 3

モデル 3 は，モデル 1 とモデル 2 とは異なり，1 つの単語が複数対応する単語の繁殖数や単語の翻訳位置の歪みについて考慮する．またモデル 3 では単語の位置を絶対位置として考える．モデル 3 では以下のパラメータを用いる．

- 翻訳確率 $P(f|e)$
  英単語 e が仏単語 f に翻訳される確率

- 繁殖確率 $n(\phi|e)$
  英単語 e が ϕ 個の仏単語と対応する確率

- 歪み確率 $d(j|i,m,l)$
  英語文の長さ l，フランス語文の長さ m のとき，i 番目の英単語 e_i が j 番目の仏単語 f_j に翻訳される確率

さらに，英単語が仏単語に翻訳されない個数を ϕ_0 とし，その確率 p_0 を以下の式で求める．このとき，歪み確率は $\frac{1}{\phi_0!}$ で，$p_0+p_1=1$ で p_0，p_1 は 0 より大きいとする．

$$P(\phi_0|\phi_1^l,E)=\left(\begin{array}{c}\phi_1+\cdots+\phi_l\\ \phi_0\end{array}\right)p_0^{\phi_1+\cdots+\phi_l-\phi_0}p_1^{\phi_0} \tag{2.16}$$

したがって，モデル 3 は以下の式で求められる．

$$\begin{aligned}P(F|E) &= \sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}P(F,a|E) &(2.17)\\ &= \sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\left(\begin{array}{c}m-\phi_0\\ \phi_0\end{array}\right)p_0^{m-2\phi_0}p_1^{\phi_0}\prod_{i=1}^{l}\phi_i!n(\phi_i|e_i)\\ &\times \prod_{j=1}^{m}t(f_j|e_{a_j})d(j|a_j,m,l) &(2.18)\end{aligned}$$

モデル 3 では，全てのアライメントを計算するため，計算量が膨大となるので期待値を近似により求める．

・モデル 4

モデル 4 では，モデル 3 と異なり，単語の位置を絶対位置ではなく，相対位置で考える．またモデル 3 では考慮されていない各単語の位置，例えば形容詞と名詞の関係を考慮する．モデル 4 では歪み確率 $d(j|i.m,l)$ を 2 つの場合で考える．

- 繁殖数が 1 以上である英単語に対応する仏単語の中で，最も文頭に近い場合

$$P(\Pi_{[i]1}=j|\pi_1^{[i]-1},\tau_0^l,\phi_0^l,E)=d_1(j-\odot_{i-1}|\mathcal{A}(e_{[i-1]}),\mathcal{B}(f_j)) \tag{2.19}$$

  $\odot_{i-1}$ は i-1 番目の英単語が対応する仏単語の位置を表している．

- それ以外の場合

$$P(\Pi_{[i]k}=j|\pi_{[i]1}{}^{k-1},\pi_1^{[i]-1},\tau_0^l,\phi_0^l,E)=d_{>1}(j-\pi_{[i]k-1}|\mathcal{B}(f_j)) \tag{2.20}$$

  $\pi_{[i]k-1}$ は同じ英単語に対応している直前の仏単語を表している．

・モデル 5

モデル 4 では，単語の位置に関して直前の単語以外は考慮されていない．したがって，複数の単語が同じ位置に生じたり，単語の存在しない位置が生成される．モデル 5 では，この問題を避けるために，単語を空白部分に配置するよう改善が施されている．

- 繁殖数が 1 以上である英単語に対応する仏単語の中で，最も文頭に近い場合

$$\begin{aligned} P(\Pi_{[i]1} &= j|\pi_1^{[i]-1}, \tau_0^l, \phi_o^l, E) \\ &= d_1(v_j|\mathcal{B}(f_j), v_{\odot_{i-1}}, v_m - \phi_{[i]} + 1)(1 - \delta(v_j, v_{j-1})) \end{aligned}$$

  v_j は j 番目までの空白数，$\mathcal{A}$ は英語の単語クラス $\mathcal{B}$ はフランス語の単語クラスを表している．

- それ以外の場合

$$\begin{aligned} P(\Pi_{[i]k} &= j|{\pi_{[i]1}}^{k-1}, \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, E) \\ &= d_{>1}(v_j - v_{\pi_{[i]k-1}}|\mathcal{B}(f_j), v_m - v_{\pi_{[i]k-1}} - \phi_{[i]} + k)(1 - \delta(v_j, v_{j-1})) \end{aligned}$$

2.3 GIZA++

GIZA++[?] は，対訳文（原言語文と目的言語文の対）から対訳単語と単語翻訳確率を自動的に得ることができる．単語翻訳確率とは，原言語と目的語における単語の対応関係 (Word Alignment) の確率である．単語翻訳確率を IBM Model1〜5 を用いて計算する．GIZA++を用いることで，以下のファイルを得る．

1. T TABLE(Translation Table)
T TABLE は，IBM Model1〜3 により作成された翻訳確率 $P(f|e)$ のデータである．f は原言語で，e は目的言語である．T TABLE は各行が，目的言語の単語 ID(e_id)，原言語の単語 ID(f_id)，原言語の単語から目的言語の単語へ翻訳する確率 ($P(f_id|e_id)$) で構成される．

2. N TABLE(Fertility Table)
N TABLE は，目的言語の単語における繁殖数を表したデータである．N TABLE は各行が，目的言語の単語 ID(e_id)，繁殖数が 0 である確率 (p^0)，繁殖数が 1 である確率 (p^1)，…，繁殖数が n である確率 (p^n) で構成される．

2.4 翻訳モデルの概要

翻訳モデルは，ある言語の単語列から別の言語の単語列へと確率的に翻訳を行うためのモデルである．

2.5 相対的意味論に基づく変換主導統計機械翻訳 (TDSMT)[?]

“相対的意味論に基づく変換主導型統計機械翻訳 (TDSMT)” とは，安場らが提案した機械翻訳の手法の一種である．TDSMT は，学習文対と，変換テーブルを用いて，原言語文を入力とし，目的言語文を出力する．変換テーブルは “A が B ならば C は D” で表現する．A は学習文対中の原言語句，B は学習文対中の目的言語句，C は入力文中の原言語句，D は出力文中の目的言語句である．

原言語入力文が，学習文対の原言語側と一致するまで，入力文と変換テーブル中の AC を照合する．次に，一致した学習文対の目的言語側を，照合した変換テーブルの BD に従って変換し，目的言語翻訳文を出力する．

TDSMT は適切な学習文対及び，変換テーブルが存在した場合，翻訳精度の高い出力文を得ることができる．しかし，TDSMT は変換テーブルを適用した，入力文が学習文対に完全に一致しない場合は翻訳ができない．従って，問題点として，入力文に対するカバー率が低い．

2.5.1 学習の手順

TDSMT における学習は “変換テーブルの作成” のみである．本節で作成手順を示す．

手順 1　対訳単語の作成

学習文対と対訳単語確率 (IBM Model 1[?]) を利用して，対訳単語を作成する．このとき付与される対訳単語確率を P_w とする．例として，表 2.1 に示す学習文対を使用して，表 2.2 に示す対訳単語を作成する．表 2.2 の値は例であり，実際の数値とは異なる．

表 2.1: 対訳単語作成に用いる学習文対

学習文対 (日本語側)	彼の弟は学生だ。
学習文対 (英語側)	His brother is a student.

表 2.2: 作成される対訳単語

	日本語単語	英語単語	p_w
対訳単語 1	彼	His	0.4
対訳単語 2	弟	brother	0.7
対訳単語 3	学生	student	0.6

手順 2　単語レベル文パターンの作成

学習文対内で対訳単語に当たる部分を変数化し，単語レベル文パターンを作成する．例を表 2.3 に示す.

表 2.3: 単語レベル文パターンの作成例

学習文対 (日本語側)	彼の兄は医者だ。
学習文対 (英語側)	His brother is a doctor.
単語レベル文パターン (日本語側)	*X0* の *X1* は *X2* だ
単語レベル文パターン (英語側)	*X0 X1* is a *X2*

手順 3　変換テーブルの作成

学習文対と単語レベル文パターンを照合する．変数化した対訳単語と，変数に当たる対訳句を変換テーブルとする．表 2.4 では変数 *N2* の部分から変換テーブル “「学生」が「student」ならば「教師」は「teacher」” が得られる.

表 2.4: 変換テーブルの作成例

学習文対 (日本語側)	彼の弟は学生だ。
学習文対 (英語側)	His brother is a student.
単語レベル文パターン (日本語側)	*X0* の *X1* は *X2* だ。
単語レベル文パターン (英語側)	*X0 X1* is a *X2* .
照合する学習文対 (日本語側)	私の母は教師だ。
照合する学習文対 (英語側)	My mother is a teacher.
変換テーブル *(X2)*	*A*:学生　*B*:student *C*:教師　*D*:teacher

手順 4　変換テーブルに確率を付与

対訳単語確率 P_w を利用し，変換テーブルに確率を付与する．この確率を変換

テーブル確率 P_v とする．

1. 変換テーブルの CD に存在する全ての日英単語の組み合わせを確認する．
2. 日本語単語に対応する英語単語の中で，対訳単語確率 P_w の最大値を得る．
3. 各日本語単語について得られた値と，変換テーブルの AB の対訳単語確率 P_w について，対数の総和を求める．

2.6 変換テーブルの作成方法の詳細

対訳文対から対訳単語辞書を作成する．まず，IBM モデル 1 を用い，対訳文の日英方向と英日方向で，対訳単語と単語翻訳確率をそれぞれ得る．そして，両者の単語翻訳確率を掛け合わせ，単語確率と呼ぶ確率値を得る．最後に対訳単語辞書を，対訳単語と単語確率で構成する．図 2.1 に日英方向の対訳単語辞書の作成例を示す．

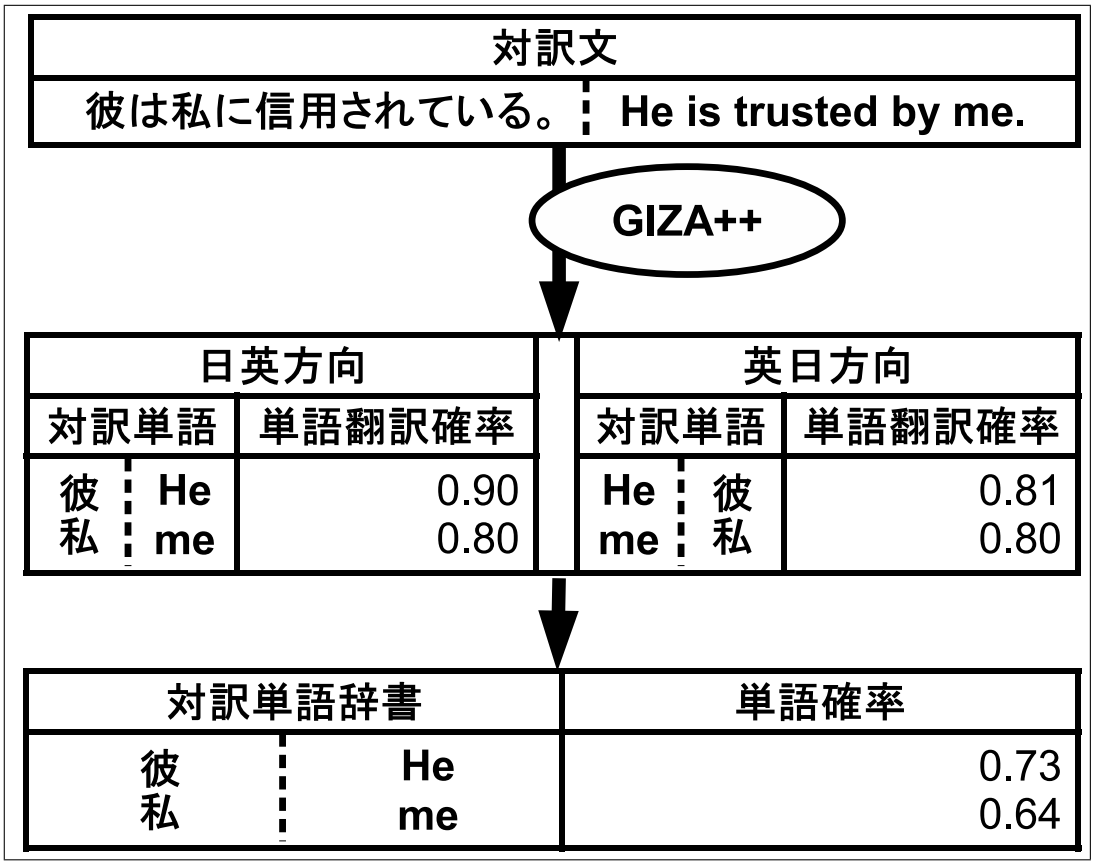

図 2.1: 日英方向の対訳単語辞書の作成

2.7 問題点

従来 Giza++では，１単語でしか利用できない．したがって，変換テーブルの A，B 使用されるのは対訳単語である．よって，対訳句は利用不可である．そこで，誤りを生じてしまう可能性がある．例として、次のような変換テーブルが作成された場合，青りんごが blue apple と訳されてしまう可能性がある．

表 2.5: 変換テーブルの例

変換テーブル 1	A:青　B:green
変換テーブル 2	A:りんご　B:apple

第3章　提案手法

3.1　概要

従来，変換テーブルの A，B に使用されるのは対訳単語のみである．しかし，対訳句は利用不可である．そこで本研究では，対訳句の活用を考える．従来 Giza++では，１単語でしか利用できない．まず対訳句の代わりに対訳２単語連続を作成する手法を提案する．

3.2　抽出手順

対訳２単語連続を自動で抽出するには，以下に示す３つの手順を踏む．

手順 1 対訳文パターンの作成

入力する対訳文の連続２単語の間に“ ＋ ”を入れ１単語とした文を作成する．

手順 2 対訳単語辞書の作成

従来手法と同じ手順で対訳単語辞書を作成し，単語確率を得る．

手順 3 対訳２単語連続

単語確率がしきい値より高い場合（上位８位）対訳２単語連続を抽出する．

第4章　実験概要

4.1　実験目的

対訳２単語連続を対訳句として人手で評価し，提案手法の有効性を検証する．

4.2　実験データ

実験では対訳文対 159,998 文を使用する．

4.3　評価方法

抽出した対訳２単語連続に対し，ランダムな 100 対を以下の評価基準に従い，人出で評価する．評価基準は対訳２単語連続の日本語を基に，英語の対応を焦点としている．

- ◯…翻訳が正しい
- △…翻訳が部分的に正しい
- ×…翻訳が間違っている

第5章　実験結果

5.1　抽出結果

表 5.1 に対訳 2 単語連続の抽出数を示す.

表 5.1: 提案手法の抽出結果

	学習文対	抽出数
提案手法	159,998 対	101,354 語

5.2　評価結果

表 5.2 に対訳 2 単語連続の評価結果を示す.

表 5.2: 提案手法の評価結果

	○	△	×
提案手法	61	20	19

5.2.1　提案手法：評価○の抽出例

表 5.3 に評価○の対訳 2 単語連続の例と順位を示す.

表 5.3: 提案手法：評価○の例

日本語側の抽出例	英語側の抽出例	順位
この+レインコート	This+raincoat	1 位
+朝刊	morning+paper	1 位
産ん+だ	gave+birth	1 位

5.2.2　提案手法：評価△の抽出例

表 5.4 に評価△の対訳 2 単語連続の例と順位を示す.

表 5.4: 提案手法：評価△の例

日本語側の抽出例	英語側の抽出例	順位
風情+も	entertain+you	1 位
回路+が	The+circuit	1 位
に+小説	writing+novels	3 位

5.2.3　提案手法：評価×の抽出例

表 5.5 に評価×の対訳 2 単語連続の例と順位を示す.

表 5.5: 提案手法：評価×の例

日本語側の抽出例	英語側の抽出例	順位
かたくな+な	Gentle+words	2 位
の+だれ	above+all	5 位
努力+も	in+vain	3 位

5.3　実験結果のまとめ

表 5.1，5.2 より，提案手法によって，対訳句の収集が可能であることがわかった.

第6章　比較実験

対訳単語のみを利用して実験を行った．また，提案手法との比較も行った．評価基準は提案手法と同様である．

6.1　実験データ

使用する学習文対は提案手法と同じ 159,998 対である．

6.2　実験結果

表 6.1 に提案手法と従来手法の抽出数を示す．

表 6.1: 提案手法と従来手法の抽出結果

	学習文対	抽出数
従来手法	159,998 対	31,786 語
提案手法	159,998 対	101,354 語

6.3　評価結果

表 6.2 に提案手法と従来手法の評価結果を示す．

表 6.2: 提案手法と従来手法の評価結果

	○	△	×
従来手法	80	6	14
提案手法	61	20	19

6.3.1　従来手法：評価◯の抽出例

表 6.3 に評価◯の対訳単語の例と順位を示す.

表 6.3: 従来手法：評価◯の例

日本語側の抽出例	英語側の抽出例	順位
半値	half-price	7 位
妹	sister	1 位
幸せ	happy	1 位

6.3.2　従来手法：評価△の抽出例

表 6.4 に評価△の対訳単語の例と順位を示す.

表 6.4: 従来手法：評価△の例

日本語側の抽出例	英語側の抽出例	順位
毎朝	morning	1 位
日本銀行	Bank	1 位
女装	disguised	3 位

6.3.3　従来手法：評価×の抽出例

表 6.5 に評価×の対訳単語の例と順位を示す.

表 6.5: 従来手法：評価×の例

日本語側の抽出例	英語側の抽出例	順位
古	newspapers	2 位
採点	papers	1 位
吹い	wind	3 位

6.4　比較実験のまとめ

従来手法と比較すると，対訳単語としてのデータ数は増加しているが，精度は落ちている．統計的に，データ数が増加するほど精度は低下する．さらに，基本的に句の精度は単語の精度より低い.

第7章　考察

7.1　抽出数と人手評価

従来手法では 31,786 語，提案手法では 101,354 語と 3 倍以上の抽出数を得た．また，提案手法で抽出された単語は従来手法と一致していないため，従来手法と組み合わせることによって，さらなる抽出数増加を見込める．

人手評価では，逆に従来手法の方が高い精度を示した．理由としては 2 つ考えられる．1 つは対訳句であるという点．基本的に，対訳単語の方が対訳句より良い精度である．もう 1 つは抽出数が増加している点．統計的に翻訳精度は抽出数と負の相関にある．したがって，従来手法より抽出数の多い提案手法は精度が低くなったと考えられる．

7.2　品詞について

7.2.1　提案手法：評価○の品詞

表 7.1 に評価○の対訳 2 単語連続と品詞を示す.

表 7.1: 提案手法：評価○の品詞

日本語側の抽出例	英語側の抽出例	日本語側の品詞	英語側の品詞
この+レインコート	This+raincoat	代名詞+名詞	代名詞+名詞
+朝刊	morning+paper	文頭+名詞	名詞+名詞
産ん+だ	gave+birth	動詞+助動詞	動詞+名詞

7.2.2　提案手法：評価△の品詞

表 7.2 に評価△の対訳 2 単語連続と品詞を示す.

7.2.3　提案手法：評価×の品詞

表 7.3 に評価×の対訳 2 単語連続と品詞を示す.

表 7.2: 提案手法：評価△の品詞

日本語側の抽出例	英語側の抽出例	日本語側の品詞	英語側の品詞
風情+も	entertain+you	名詞+助詞	名詞+代名詞
回路+が	The+circuit	名詞+助詞	冠詞+名詞
に+小説	writing+novels	助詞+名詞	動詞+名詞

表 7.3: 提案手法：評価×の品詞

日本語側の抽出例	英語側の抽出例	日本語側の品詞	英語側の品詞
かたくな+な	Gentle+words	形容動詞	形容詞+名詞
の+だれ	above+all	助詞+代名詞	前置詞+代名詞
努力+も	in+vain	名詞+助詞	前置詞+名詞

7.2.4 従来手法：評価○の品詞

表 7.4 に評価○の対訳単語と品詞を示す.

表 7.4: 従来手法：評価○の品詞

日本語側の抽出例	英語側の抽出例	日本語側の品詞	英語側の品詞
半値	half-price	名詞	名詞
妹	sister	名詞	名詞
幸せ	happy	名詞	名詞

7.2.5　従来手法：評価△の品詞

表 7.5 に評価△の対訳単語と品詞を示す.

表 7.5: 従来手法：評価△の品詞

日本語側の抽出例	英語側の抽出例	日本語側の品詞	英語側の品詞
毎朝	morning	名詞	名詞
日本銀行	Bank	名詞	名詞
女装	disguised	名詞	動詞

7.2.6　従来手法：評価×の品詞

表 7.6 に評価×の対訳単語と品詞を示す.

表 7.6: 従来手法：評価×の品詞

日本語側の抽出例	英語側の抽出例	日本語側の品詞	英語側の品詞
古	newspapers	形容詞	名詞
採点	papers	名詞	名詞
吹い	wind	動詞	名詞

7.2.7　品詞分類による考察

抽出されている品詞の種類が豊富である点．従来手法では日本語側の品詞も英語側の品詞も「名詞」が多く抽出されていた．しかし，提案手法では様々な品詞の単語が抽出されていた.

しかし，日本語の品詞の「助詞」，英語の品詞の「前置詞」を含む場合，精度が落ちる可能性が高い．様々な品詞を抽出することができるが，日本語の品詞の「助詞」，英語の品詞の「前置詞」は具体的な翻訳がつきにくく，結果として翻訳精度の低下の原因となっている.

第8章　おわりに

従来，変換テーブルの A，B は対訳単語のみである．

そこで提案手法では，対訳句の利用を目指し，対訳２単語連続の作成を行った．実験結果より，対訳句の抽出は可能であることが分かった．また，提案手法は従来手法に比べて抽出数が多かった．しかし，翻訳精度は落ちてしまった．

提案手法の翻訳精度では実用的ではない．今後は翻訳精度を高める方法や，対訳２単語連続の翻訳での活用を考える必要がある．

謝辞

最後に，本研究を遂行するにあたり，研究の進め方や本論文の書き方など, ご指導いただきました鳥取大学工学部電気情報系学科自然言語処理研究室の村上仁一准教授, 村田真樹教授をはじめ, 自然言語処理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，深く感謝致します.

参考文献

[1] 安場裕人, 村上仁一, “変換主導型統計機械翻訳の提案, 言語処理学会”, 第 24 回年次大会 発表論文集, 2018

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

[3] 力久 剛士，“レーベンシュタイン距離を用いた翻訳精度の向上”，平成 26 年度 卒業論文，pp.3-15，February 2015.

[4] 森本智喜, 村上仁一 (2019). 1 変数の対訳文パターンを用いた対訳句の抽出調査 言語処理学会　第 25 回年次大会 発表論文集

付録