概要

近年，機械翻訳において，統計翻訳が注目されている．統計翻訳では，対訳学習文から自動的に翻訳規則を獲得し，翻訳を行うため，翻訳精度は対訳学習文の量に大きく依存する．対訳学習文の量が少ない場合，翻訳されない単語が出力される．本研究では，そのような単語を未知語と定義する．代表的な未知語の対策として，対訳学習文を追加する方法が挙げられる．しかし，対訳学習文を追加するにはコストがかかる．

そこで，本研究では，対訳学習文を追加せずに未知語処理を行う新たな手法を提案する．具体的には，出現した未知語を抽出し，文字単位化した後，文字単位化した未知語を入力として再度翻訳を行う．この手法を用いて未知語の削減と翻訳精度の向上を試みた．実験の結果，ベースラインでは未知語を含む文が 3,146 文出力されていたが，提案手法により 236 文まで削減することに成功した．更に，人手による対比較評価を行ったところ，翻訳精度の向上も認められた．

目 次

図 目 次

表 目 次

1　はじめに

機械翻訳において，人手で翻訳規則を定義し，翻訳を行うルールベース翻訳が一般的であった．しかし，人手で翻訳規則を定義するには，莫大なコストがかかる．また，言語毎に文法規則が異なるため，多言語への拡張が困難であった．そのため，近年では，統計翻訳が主流となっている．統計翻訳では，対訳学習文から自動的に翻訳規則を獲得し，翻訳を行うため，翻訳精度は対訳学習文の量に大きく依存する．対訳学習文の量が少ない場合，翻訳されない単語が出力される．本研究では，そのような単語を未知語と定義する．代表的な未知語の対策として，対訳学習文を追加する方法が挙げられる．しかし，対訳学習文を追加するにはコストがかかる．この問題を解決するために，藤原ら [10] は，日英翻訳において，対訳学習文を追加せずに，フレーズテーブル作成時のヒューリスティックスを併用することで，未知語の削減と翻訳精度の改善を試みた．その結果，未知語の削減には成功したが，翻訳精度はほとんど向上しなかった．

そこで，本研究では，対訳学習文を追加せずに未知語処理を行う新たな手法を提案する．具体的には，出現した未知語を抽出し，文字単位化した後，文字単位化した未知語を入力として再度翻訳を行う．この手法を用いて未知語を削減し，翻訳精度の向上を試みた．この結果，大幅な未知語の削減に成功し，翻訳精度の向上が確認できた．

本論文の構成を以下に示す．第 2 章で日英統計翻訳システムについて説明する．第 3 章では類似研究について説明し，第 4 章で提案手法のシステムについて説明する．そして，第 5 章では実験環境を，第 6 章で実験結果を示し，第 7 章で本研究の考察を述べる．

2　日英統計翻訳システム

2.1　概要

統計翻訳において,「単語に基づく統計翻訳」と,「句に基づく統計翻訳」がある．初期の統計翻訳は，単語に基づく統計翻訳であった．しかし，近年提案された句に基づく統計翻訳 [1] は，語順の並び替えや文脈における訳語の選択や翻訳精度において，単語に基づく統計翻訳よりも優れている．そのため，現在は句に基づく統計翻訳が主流となっている．したがって，本研究では，統計翻訳システムにおいて，句に基づく統計翻訳を用いる．日英統計翻訳では，日本語文 j を入力文とした場合，翻訳モデル $P(j|e)$ と言語モデル $P(e)$ の全ての組み合わせから，確率が最大となる英語文 $\hat{e}$ を探索し，出力文とする．E を探索する翻訳システムをデコーダと呼ぶ．以下に基本的なモデルを示す．また，日英統計翻訳の流れを図 1 に示す.

$$E = arg\max_{e} P(e|j) \tag{1}$$

$$\simeq arg\max_{e} P(j|e)P(e) \tag{2}$$

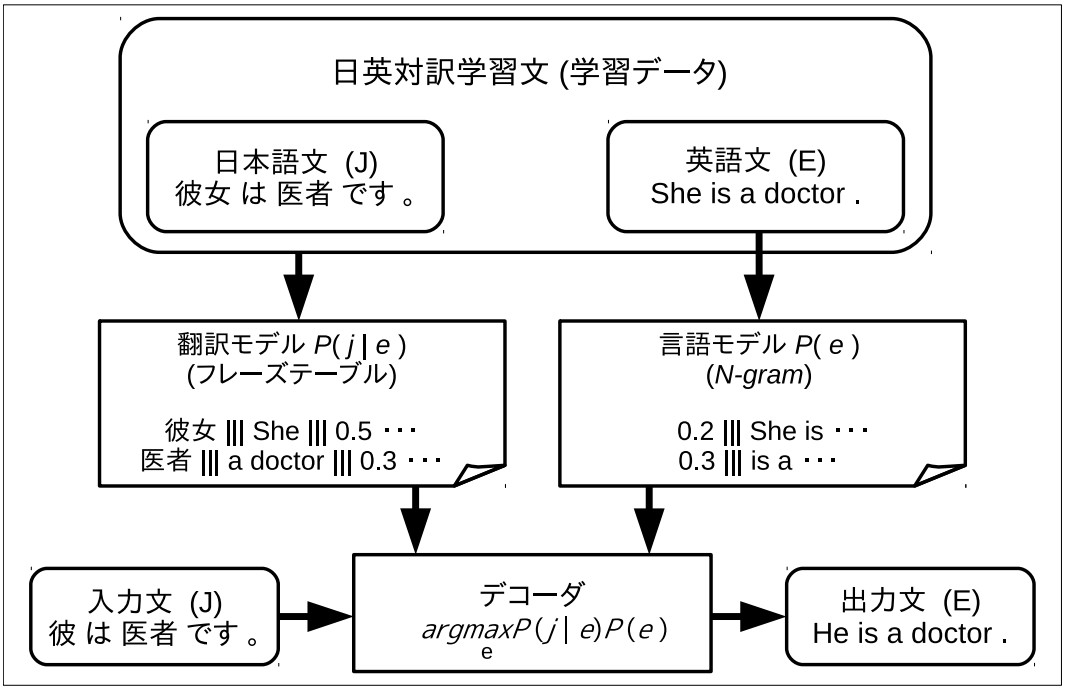

図 1: 日英統計翻訳の流れ

2.2 言語モデル

言語モデルは，単語列の生成確率を付与するモデルである．日英翻訳では，翻訳モデルを用いて生成された翻訳候補から，英語として自然な文を選出するために用いる．統計翻訳では一般的に，N-gram モデルを用いる．

2.2.1 N-gram モデル

N-gram モデルとは “単語列 $P(W_1^n) = w_1^n = w_1, w_2, w_3, ...w_n$ 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．計算式を以下に示す．

$$
\begin{aligned}
P(W_1^n) &= P(w_1) \times P(w_2|w_1) \times P(w_3|w_1^2)...P(w_n|w_1^{n-1}) && (3)\\
&\approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_1^2)...P(w_n|w_{n-(N-1)}^{n-1}) && (4)\\
&= \prod_{i=1}^{n} P(w_i|w_{i-(N-1)}^{i-1}) && (5)
\end{aligned}
$$

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．

$$
P(w_i|w_{i-(N-1)}^{i-1}) = \frac{C(w_{i-(N-1)}^{i})}{C(w_{i-(N-1)}^{i-1})} \tag{6}
$$

2.3 単語に基づく翻訳モデル

統計翻訳における単語対応を獲得するため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 [2] がある．IBM 翻訳モデルは，model1 から model5 までの 5 つのモデルから構成されている．各モデルの概要を以下に示す.

model1 目的言語のある単語が原言語の単語に訳される確率を用いる

model2 model1 に加えて，目的言語のある単語に対応する原言語の単語の原言語文中での位置の確率（以下，permutation 確率と呼ぶ）を用いる（絶対位置）

model3 model2 に加えて，目的言語のある単語が原言語の何単語に対応するかの確率を用いる

model4 model3 の permutation 確率を改良（相対位置）

model5 model4 の permutation 確率を更に改良

IBM 翻訳モデルは仏英翻訳を前提としているが，本研究では日英翻訳を扱っているため，日英翻訳を前提に説明する．なお，以下の説明は藤原ら [10] の論文より引用した．

原言語の日本語文を J，目的言語の英語文を E として定義する．IBM 翻訳モデルにおいて，日本語文 J と英語文 E の翻訳モデル $P(J|E)$ を計算するため，アライメント a を用いる．以下に IBM モデルの基本的な計算式を示す.

$$P(J|E) = \sum_{a} P(J, a|E) \tag{7}$$

ここで，アライメント a は，J と E の単語の対応を意味している．IBM 翻訳モデルにおいて，各日単語に対応する英単語は 1 つであるのに対して，各英単語に対応する日単語は 0 から n 個あると仮定する．また，日単語と適切な英単語が対応しない場合，英語文の先頭に e_0 という空単語があると仮定し，日単語と対応させる．

2.3.1 model1

式 (3) は以下の式に置き換えられる．

$$P(j,a|E) = P(m|E)\prod_{j=1}^{m} P(a_j|a_1^{j-1}, j_1^{j-1}, m, E)P(j_j|a_1^{j}, j_1^{j-1}, m, E) \tag{8}$$

m は日本語文の文長を示す．また，a_1^{j-1} は日本語文の 1 単語目から j-1 単語目までのアライメントである．そして j_1^{j-1} は日本語文の 1 番目から j-1 番目までの単語を示す．ここで，Model1 では以下を仮定している．

- 日本語文の長さの確率 ϵ は，m と E に依存しない
  $\epsilon \equiv P(m|E)$
- アライメントの確率は英語文の長さ l にのみ依存する
  $P(a_j|a_1^{j-1}, j_1^{j-1}, m, E) \equiv (l+1)^{-1}$
- 日本語の翻訳確率 $t(j_j|e_{a_j})$ は，日単語に対応する英単語にのみ依存する
  $P(j_j|a_1^{j}, j_1^{j-1}, m, E) \equiv t(j_j|e_{a_j})$

以上の仮定を用いて，式 (4) は簡略化することができる．以下に式を示す．

$$P(J,a|E) = \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m} t(j_j|e_{a_j}) \tag{9}$$

$$P(J|E) = \frac{\epsilon}{(l+1)^m}\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\prod_{j=1}^{m} t(j_j|e_{a_j}) \tag{10}$$

$$= \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m}\sum_{i=0}^{l} t(j_j|e_i) \tag{11}$$

model1 において，翻訳確率 $t(j|e)$ の初期値が 0 でない場合， EM アルゴリズム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 $t(j|e)$ に初期値を設定する．

手順 2 日本語と英語の対訳文 $(J^{(s)}, E^{(s)})(1 \le s \le S)$ において，日単語 j と英単語 e が対応付けられる回数の期待値を求める．ここで $\delta(j, j_j)$ は日本語文 J において日単語 j が出現する回数を表す．そして $\delta(e, e_i)$ は英語文 E において英単語 e が出現する回数を表す．

$$c(j|e;J,E) = \frac{t(j|e)}{t(j|e_0)+\cdots+t(j|e_l)}\sum_{j=1}^{m}\delta(j,j_j)\sum_{i=0}^{l}\delta(e,e_i) \tag{12}$$

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

- 定数 λ_e を以下の式で計算する

$$\lambda_e = \sum_j \sum_{s=1}^{S} c(j|e; J^{(s)}, E^{(s)}) \tag{13}$$

- 上式で求めた定数 λ_e を用いて $t(j|e)$ を以下の式で再計算する

$$t(j|e) = \lambda_e^{-1} \sum_{s=1}^{S} c(j|e; J^{(s)}, E^{(s)}) \tag{14}$$

$$= \frac{\sum_{s=1}^{S} c(j|e; J^{(s)}, E^{(s)})}{\sum_j \sum_{s=1}^{S} c(j|e; J^{(s)}, E^{(s)})} \tag{15}$$

手順 4 $t(j|e)$ が収束するまで，手順 2 と手順 3 を繰り返す．

2.3.2 model2

model1 において，アライメントの確率は英語文の長さ l にのみ依存する．そこで model2 では，英語文の長さ l に加え，j 単語目のアライメント a_j，日本語文の長さ m に依存するとし，以下の式で表す．

$$a(a_j|j, m, l) \equiv P(a_j|a_1^{j-1}, j_1^{j-1}, m, l) \tag{16}$$

よって，model1 の式 (6) は以下のように置き換えられる．

$$P(J|E) = \epsilon \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(j_j|e_{a_j}) a(a_j|j, m, l) \tag{17}$$

$$= \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} t(j_j|e_i) a(i|j, m, l) \tag{18}$$

model2 において，対訳文中の英単語 e と日単語 j が対応付けされる回数の期待値である $c(j|e; J^{(s)}, E^{(s)})$ と，日単語の位置 j と英単語の位置 i が対応付けられる回数の期待値 $c(i|j, m, l; J^{(s)}, E^{(s)})$ が存在する．以下に，期待値 $c(j|e; J^{(s)}, E^{(s)})$ と $c(i|j, m, l; J^{(s)}, E^{(s)})$ を求める式を示す．

$$c(j|e; J^{(s)}, E^{(s)}) = \sum_{j=1}^{m} \sum_{i=0}^{l} \frac{t(j|e) a(i|j, m, l) \delta(j, j_j) \delta(e, e_i)}{t(j|e_0) a(0|j, m, l) + \cdots + t(j|e_l) a(l|j, m, l)} \tag{19}$$

$$c(i|j, m, l; J^{(s)}, E^{(s)}) = \frac{t(j_j|e_i) a(i|j, m, l)}{t(j_j|e_0) a(0|j, m, l) + \cdots + t(j_j|e_l) a(l|j, m, l)} \tag{20}$$

model2 においても，最適解を推定するために EM アルゴリズムを用いる．しかし，計算によって複数の極大値が算出され，最適解が得られない場合が存在する．model2 の特殊な場合に，$a(i|j,m,l) = (l+1)^{-1}$ が挙げられるが，これは model1 として考えることができる．また，最適解が保証されている model1 で求められた値を初期値として用いることで，最適解を求めることができる．

2.3.3 model3

model1 および model2 において，日単語と英単語の対応は 1 対 1 の場合のみを考慮していた．しかし，model3 では，1 つの単語が複数の単語に対応する場合や，単語の翻訳位置の距離についても考慮する．また，モデル 3 では単語の位置を絶対位置として考えている．モデル 3 では以下のパラメータを用いる．

- $P(j|e)$
  英単語 e が日単語 j に翻訳される確率
- $n(\phi|e)$
  英単語 e が ϕ 個の日単語と対応する確率
- $d(j|i,m,l)$
  英語文の長さ l，日本語文の長さ m のとき，i 番目の英単語 e_i が j 番目の日単語 j_j に翻訳される確率

さらに，英単語に翻訳されない日本語の単語数を ϕ_0 として，そのような単語が発生する確率 p_0 を以下の式に表す．

$$P(\phi_0|\phi_1^l, e) = \left(\begin{array}{c} \phi_1 + \cdots + \phi_l \\ \phi_0 \end{array} \right) p_0^{\phi_1+\cdots+\phi_l-\phi_0} p_1^{\phi_0} \tag{21}$$

したがって，model3 は以下の式によって表される．

$$\begin{aligned} P(j|e) &= \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} P(j,a|e) && (22) \\ &= \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \left(\begin{array}{c} m-\phi_0 \\ \phi_0 \end{array} \right) p_0^{m-2\phi_0} p_1^{\phi_0} \prod_{i=1}^{l} \phi_i! n(\phi_i|e_i) \times \\ &\qquad \prod_{j=1}^{m} t(j_j|e_{a_j}) d(j|a_j,m,l) && (23) \end{aligned}$$

モデル 3 では，全ての単語対応を考慮して計算するため，計算量が膨大となる．そのため，期待値は近似によって求められる．

2.3.4 model4

model3 と model4 の違いは，単語の位置の考慮の仕方である．model3 において，単語の位置は絶対位置で考慮していた．それに対して，model4 では単語の位置を相対位置で考慮する．また，各単語ごとの位置も考慮している．model4 では，単語位置の歪みの確率である $d(j|i,m,l)$ を以下の 2 通りで考慮する．

- 英単語に対応する日単語が 1 以上あるときに，その中で最も文頭に近い場合

$$P(\Pi_{[i]1}=j|\pi_1^{[i]-1},\tau_0^l,\phi_0^l,E)=d_1(j-\odot_{i-1}|\mathcal{A}(e_{[i-1]}),\mathcal{B}(j_j)) \tag{24}$$

- それ以外の場合

$$P(\Pi_{[i]k}=j|{\pi_{[i]1}}^{k-1},\pi_1^{[i]-1},\tau_0^l,\phi_0^l,E)=d_{>1}(j-\pi_{[i]k-1}|\mathcal{B}(j_j)) \tag{25}$$

2.3.5 model5

モデル 4 では，単語の位置に関して直前の単語のみを考慮している．そのため，複数の単語が同じ位置に生じたり，単語が存在しない位置に生成されるという問題がある．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に制約が施されている．

2.4 GIZA++

GIZA++[3] とは，日英方向と英日方向の対訳文から最尤な単語対応を得るための計算を行うツールである．IBM 翻訳モデルの model1 から model5 に基づいて，単語の対応関係の確率値を計算する．GIZA++を用いた場合，以下の 2 つのファイルが出力される．

1. **T TABLE (Translation Table)** T TABLE は，Model1 から Model3 により作成された翻訳確率 $P(f|e)$ のデータである．f は翻訳する言語で，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)，翻訳する言語の単語 ID(f_id)，翻訳する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 ($p0$)，繁殖数が 1 である確率 ($p1$)，…，繁殖数が n である確率 (pn) で構成される．

2.5 句に基づく翻訳モデル

句に基づく翻訳モデルとは，確率的に日本語から英語の単語列へ翻訳を行うためのモデルである．統計翻訳において，句に基づく翻訳モデルとして，一般的にはフレーズテーブルが用いられている．フレーズテーブルは以下の手順で作成される．

手順 1 IBM モデルを用いて，単語の対応を得る

手順 2 ヒューリスティックスなルールを用いて句に基づく対応を得る

手順 3 手順 2 で求めた句対応から，フレーズテーブルを作成する

詳しい作成手順については，2.6 節にて説明する．また，表 2.1 にフレーズテーブルの例を示す．

表 2.1: フレーズテーブルの例

突然 天気 が ||| Suddenly , the weather ||| 0.5 0.00217118 1 3.39949e-05 2.718 ||| 0-0 0-1 2-2 1-3 ||| 2 1 1
突然 天気 が 変わっ た ||| Suddenly , the weather changed ||| 0.5 9.13961e-05 0.5 4.2075e-06 2.718 ||| 0-0 0-1 2-2 1-3 3-4 4-4 ||| 2 2 1
突然 天気 が 変わっ た 。 ||| Suddenly , the weather changed . ||| 0.5 9.13961e-05 0.5 4.20734e-06 2.718 ||| 0-0 0-1 2-2 1-3 3-4 4-4 5-5 ||| 2 2 1

左から順に，日本語フレーズ，英語フレーズ，日英方向の翻訳確率 $P(j|e)$，日英方向の単語の翻訳確率の積，英日方向の翻訳確率 $P(e|j)$，英日方向の単語の翻訳確率の積，フレーズペナルティ，フレーズ内単語対応（日英方向）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．

2.6 フレーズテーブル作成法

IBM モデルは，方向のある 1 対多の単語アライメントである．よって，句レベルであるフレーズテーブルを得るには，両方向の 1 対多のアライメントを求める必要がある．

まず，GIZA++を用いて，学習文から日英方向と英日方向の対訳文において最尤な単語アライメントを得る．例として，日本語文 “風で火が消えた” と，その対訳英語文 “The wind blew out the fire” を挙げる．図 2 に日英方向の単語対応を示す．また，図 3 に英日方向の単語対応を示す．なお，図 2 と図 3 において，●は対応点を示す．

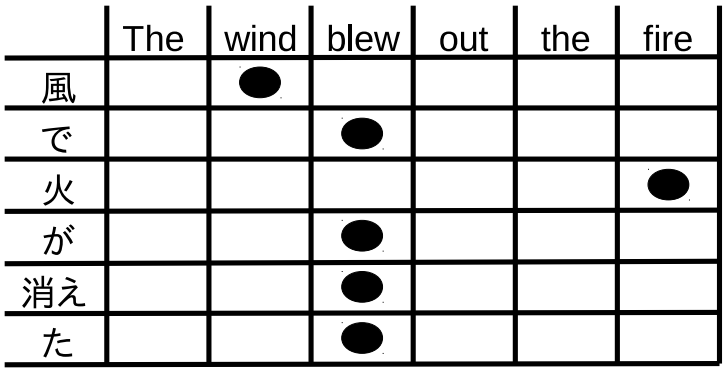

図 2: 日英方向の単語対応

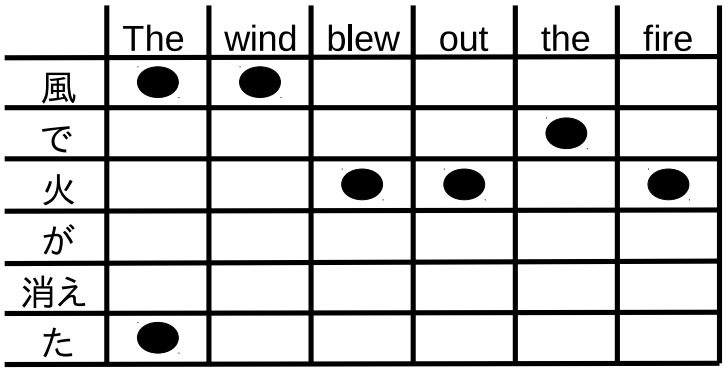

図 3: 英日方向の単語対応

次に，両方向のアライメントから，両方向に 1 対多の対応を認めた単語アライメントをヒューリスティックスなルールにより計算する．基本的なヒューリスティックスとして，“intersection”，“union”，“grow”，そして “grow-diag” がある．

2.6.1 intersection

intersection（積集合）は，日英方向と英日方向の両方に単語対応が存在する場合，その単語対応を “対応点” とする方法である．intersection の例を図 4 に示す.

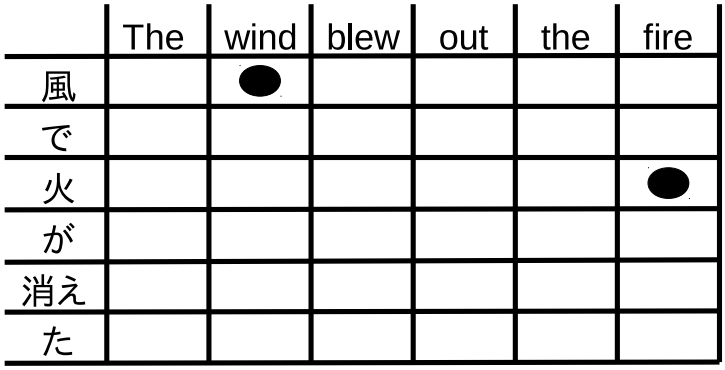

図 4: intersection の例

2.6.2 union

union（和集合）は，日英方向と英日方向のどちらか一方に単語対応が存在する場合，その単語対応を “対応点” とする方法である．union の例を図 5 に示す.

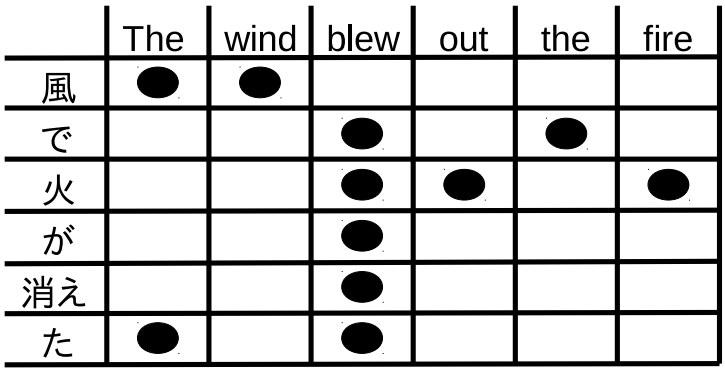

図 5: union の例

2.6.3 grow と grow-diag

grow, grow-diag は intersection と union の中間である．intersection からスタートし，既に採用した対応点の周りに union の対応点を加えていく．grow では縦と横の方向に，grow-diag では縦と横と対角の方向に union の対応点がある場合に，その対応点を用いる．図 6 に grow の例を，図 7 に grow-diag の例を示す.

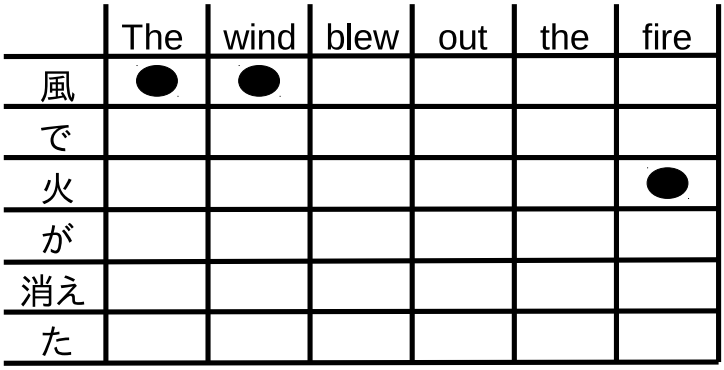

図 6: grow の例

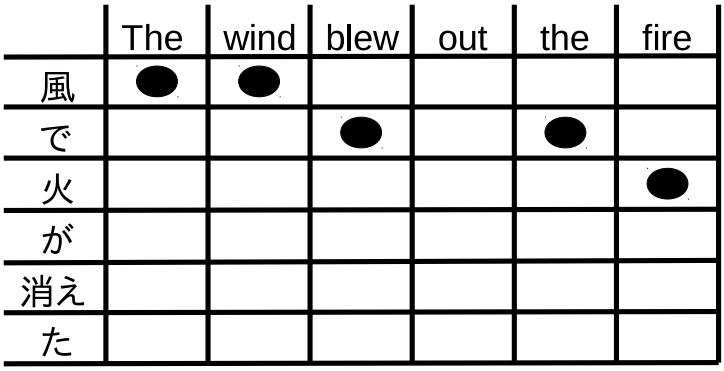

図 7: grow-diag の例

2.6.4 final と final-and

最終処理のヒューリスティックスとして，“final”と “final-and”を用いる．final は，少なくとも片方の言語における単語の単語対応がない場合に，union の単語対応を追加する．また，final-and は，両側言語における単語の単語対応がない場合に，union の候補対応点を追加する．図 8 に grow-diag-final の例を，図 9 に grow-diag-final-and の例を示す．

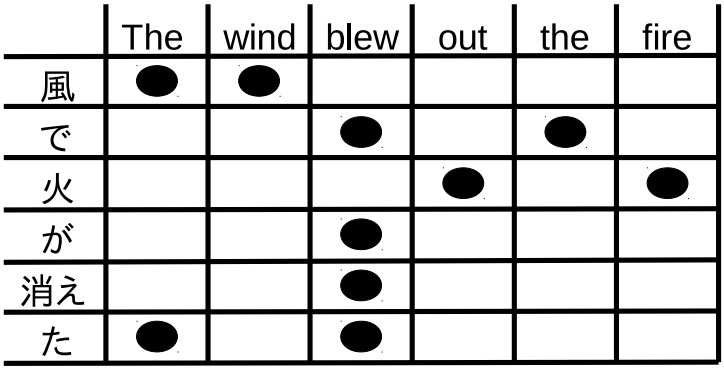

図 8: grow-diag-final の例

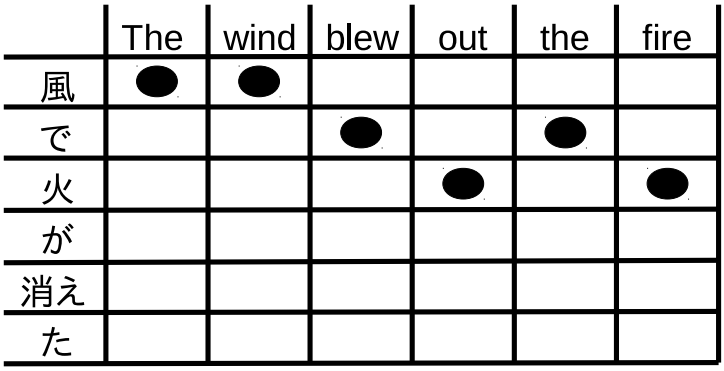

図 9: grow-diag-final-and の例

得られた単語アライメントから，全ての矛盾しないフレーズ対を得る．このとき，そのフレーズ対に対して翻訳確率を計算し，フレーズ対に確率値を付与することで，フレーズテーブルを作成する．

2.7 デコーダ

デコーダは言語モデルと翻訳モデルの全ての組み合わせから，確率が最大となる翻訳候補を探索し，出力する．入力文として,「彼は医者です。」が入力されたときの翻訳例を図 10 に示す.

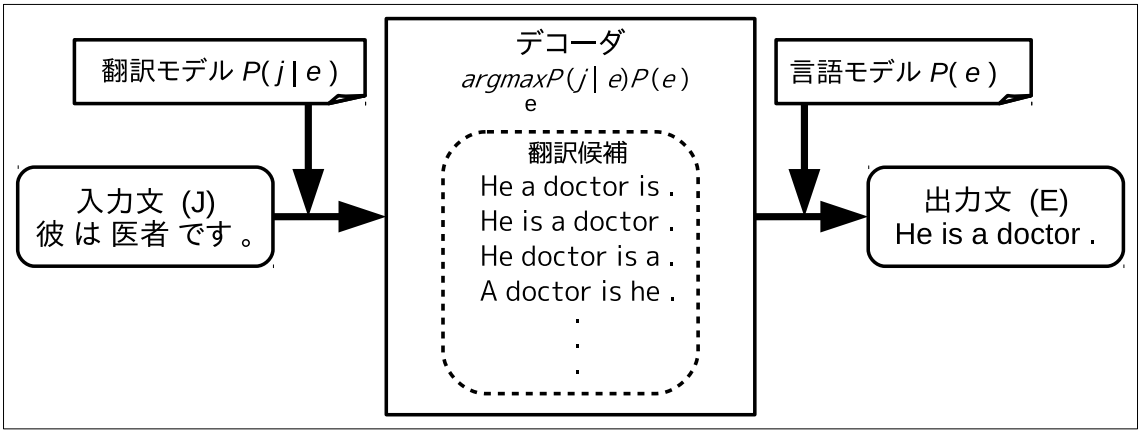

図 10: デコーダの動作例

デコーダは，日英統計翻訳において，$arg\max_{e} P(j|e)P(e)$ の確率が最大となる英語文を出力するために，適切な順序で日本語と英語の単語対応を選択する必要がある．しかし，適切な英語文を決定するためには，莫大な計算量が必要となる．そこで，計算量を削減するための手法として，ビームサーチ法が存在する.

2.8 パラメータチューニング

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ると，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．各モデルに与える重みをパラメータと呼ぶ．このパラメータを最適化するために，MERT(Minimum Error Rate Training)[4] という手法を用いる．MERT は，後述する自動評価法 BLEU[5] のスコアが最大となる翻訳結果を出力するようにパラメータ $\hat{\lambda_1^n}$ の調整を行う．n 個のパラメータ $\hat{\lambda_1^n}$ の最適化に用いる式を以下に示す．

$$\hat{\lambda_1^n} = arg \max_{\lambda_1^n} BLEU(smt(\lambda_1^n), e_{ref}) \tag{26}$$

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．なお，パラメータチューニングにおける入力文として，ディベロップメント文と呼ばれるデータを用いる．ディベロップメント文を用いて試し翻訳を行い，各文に対して上位 N 個の翻訳候補を出力する．そして N 個の中から，より自動評価値が高い翻訳候補が上位に来るようにパラメータに $\hat{\lambda_1^n}$ 最適化する．試し翻訳とパラメータの調整を繰り返すことで，パラメータチューニングを行う．

3　類似研究

3.1　未知語処理における類似研究の概要

未知語処理における類似研究として，藤原ら [10] の提案手法（以下，先行手法と呼ぶ）が挙げられる．藤原らは，日英翻訳において，対訳学習文を追加せずに，フレーズテーブル作成時のヒューリスティックスを併用することで，未知語の削減と翻訳精度の改善を試みた．その結果，未知語の削減には成功したが，翻訳精度はほとんど向上しなかった．本研究との違いは，フレーズテーブル作成時のヒューリスティックスに着目している点である．

3.2　先行手法の手順

以下に先行手法の具体的な手順を示す．また，先行手法の流れを図 11 に示す．

準備　英語学習文と日本語学習文を準備する．

手順 1 英語学習文を用いて言語モデルを作成する．

手順 2 英語学習文と日本語学習文を用いて翻訳モデルを作成する. また，ヒューリスティックスとして “grow-diag-final-and” を用いる．

手順 3 手順 2 と同様にして翻訳モデルを作成する. また，ヒューリスティックスとして “intersection” を用いる．

手順 4 手順 3 で作成されたフレーズテーブルから未知語が含まれるフレーズ対を抽出し，手順 2 で作成されたフレーズテーブルに直接追加する．このフレーズテーブルを用いて翻訳を行う．

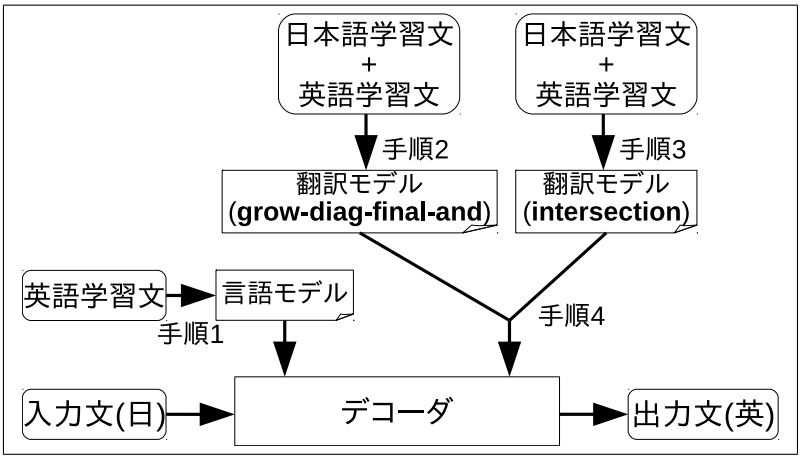

図 11: 日英統計翻訳における先行手法の流れ

4　提案手法

4.1　提案手法の概要

本研究では，対訳学習文を追加せずに未知語処理を行う新たな手法を提案する．具体的には，出現した未知語を抽出し，文字単位化した後，文字単位化した未知語を入力として再度翻訳を行う．

4.2　提案手法の手順

以下に具体的な手順を示す．また，提案手法の流れを図 12 に示す．

準備 英語学習文と単語単位の日本語学習文および文字単位の日本語学習文を準備する．

手順 1 英語学習文を用いて言語モデルを作成する．

手順 2 英語学習文と日本語学習文を用いて翻訳モデルを作成する．

手順 3 手順 1，手順 2 で作成したモデルを用いて一回目の翻訳を行い，出力された英語文から未知語を抽出する．

“He has 画才 .” → “画才” (例 1)

手順 4 手順 3 で抽出した未知語を文字単位化し，二回目の翻訳の入力とする．

“画才” → “画 才” (例 2)

手順 5 英語学習文と文字単位化した日本語学習文を用いて翻訳モデルを作成する．

手順 6 手順 1, 手順 5 で作成したモデルを用いて二回目の翻訳を行う．

“画 才” → “artistic talent” (例 3)

手順 7 手順 6 で出力された英語を，一回目の翻訳結果における未知語部分に置換して，英語文を出力する．

“He has 画才 .” → “He has artistic talent” (例 4)

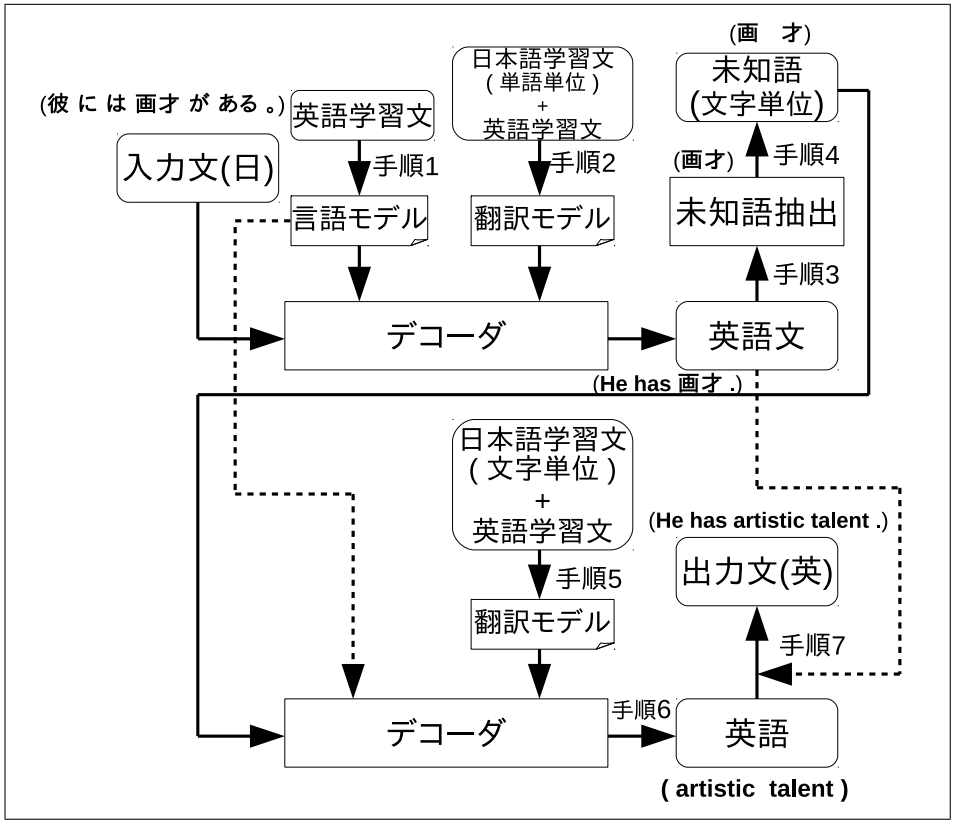

図 12: 日英統計翻訳における提案手法の流れ

5　実験環境

5.1　言語モデル

言語モデルの学習には，“SRILM[7]”の “ngram-count”を用いる．本研究では，N-gram モデルに 5-gram を用いる．

5.2　翻訳モデル

翻訳モデルの学習には，“GIZA++[3]” 用いる “train-factored-phrase-model.perl[6]”を用いる．なお，本研究では，ヒューリスティックスとして，“grow-diag-final-and”を用いる．

5.3　デコーダのパラメータ

デコーダには，“moses[6]”を用いる．また，moses の各パラメータは “mert-moses.pl[6]” を用いて最適化する．しかし，“ttable-limit”と “distortion-limit”についてはパラメータチューニングでは変更されない．“ttable-limit”とは，１つの日本語のフレーズに対して考慮する，目的言語のフレーズ数の制限である．また，“distortion-limit”とは，フレーズの並び替えの範囲の制限である．本研究では，“ttable-limit”の値を 10，また “distortion-limit” の値を-1（無制限）とする．

5.4 実験データ

本研究では，実験に単文のみを用いる．単文の本来の意味は，主語と述語の関係が 1 回のみ成り立つ文である．しかし，本研究で用いる単文は，形態素解析器を用いて形態素解析した際に動詞が 1 つの文を抽出したものである．例えば,「彼は生き返った。」という文は，本来ならば単文であるが，形態素解析において,「彼/は/生き/返っ/た/。」と解析された場合には,「生き返る」という動詞ではなく,「生きる」と「返る」の 2 つの動詞が含まれているとみなして，本研究には用いない．以下に，本研究で用いる単文コーパスの例を示す.

表 5.1: 単文コーパスの例

日本語句	水 が 腐っ て いる 。
英語句	The water is foul .
日本語句	素行 を 改め なさい 。
英語句	You should mend your ways .
日本語句	彼 は 最後 の 断 を 下し た 。
英語句	He made a final decision .

本研究では，電子辞書などの例文より抽出した単文コーパス [8] を用いる．使用するデータの内訳を表 5.2 に示す． 統計翻訳の前処理として，各コーパスの日本語文に対し

表 5.2: 実験データ

日本語学習文	100,000 文
英語学習文	100,000 文
ディベロップメント文	1,000 文
テスト文	10,000 文

て，“MeCab[9]”を用いて形態素解析を行う．また，英語文に対して “tokenizer.perl[6]”を用いて分かち書きを行う.

5.5 評価方法

本研究では，未知語の数を比較して評価を行う．具体的には，提案手法の一回目の翻訳における出力文をベースラインとし，ベースラインと提案手法および先行手法における未知語の数を比較する．また，文全体の翻訳精度の評価として，人手評価を行う．人手評価には，対比較評価を用いる．ベースライン VS 提案手法の対比較評価では，“入力文”，“正解文”，“ベースライン出力文”，“提案手法出力文”が与えられ，ベースライン出力文と提案手法出力文の比較を行う．先行手法 VS 提案手法の対比較評価では，“入力文”，“正解文”，“先行手法出力文”，“提案手法出力文”が与えられ，先行手法出力文と提案手法出力文の比較を行う．

6　実験結果

6.1　未知語を含む文数と未知語の単語数

テスト文 10,000 文を入力文として翻訳実験を行い，ベースライン，提案手法および先行手法において，未知語の数の調査を行った．調査結果を表 6.1 に示す．

表 6.1 中の “ベースライン” とは，提案手法の一回目の翻訳における結果である．また，“提案手法” とは，本研究で提案した手法の結果であり，“先行手法” とは，藤原らが提案した手法の結果である．

表 6.1: 未知語の調査結果 (10,000 文中)

翻訳手法	未知語を含む文数	未知語の単語数
ベースライン	3,170 文	3,915 単語
提案手法	236 文	253 単語
先行手法	1,281 文	1,458 単語

表 6.1 の結果より，提案手法において，未知語を大幅に削減できたことが確認できる．

6.2 未知語の翻訳品質

文字単位化した未知語をランダムに 100 単語抽出し，翻訳できた未知語数を調査した．この結果，翻訳できた未知語は 93 単語存在し，翻訳できなかった未知語は 7 単語存在した．この 93 単語の内，正しく翻訳できていた未知語は 21 単語存在し，正しく翻訳できなかった未知語は 72 単語存在した．正しく翻訳できた未知語の一例を表 6.2 に，正しく翻訳できなかった未知語の一例を表 6.3 に，翻訳できなかった未知語の一例を表 6.4 に示す．また，正しく翻訳できた未知語は，ひらがな，カタカナよりも漢字の方が多かった．

表 6.2: 正しく翻訳できた未知語の一例（21 単語）

翻訳前	翻訳後
壇	altar
誤 差	error
輝 き	sheen
画 用 紙	drawing paper
オ ー バ ー ホ ー ル	overhaul

表 6.3: 正しく翻訳できなかった未知語の一例（72 単語）

翻訳前	翻訳後
無 私	I without
豊 浜	beach diverse
名 器	arms name
か も め	to much
イ ベ ン ト	Best image of

表 6.4: 翻訳できなかった未知語の一例（7 単語）

翻訳前	翻訳後
槍	槍
都	郡
梶 山	梶 mountain
楊 枝	楊 branch
僧 侶	僧 侶

6.3 提案手法におけるシステム全体の翻訳精度

提案手法におけるシステム全体の翻訳精度を調べるために，ベースラインと提案手法，先行手法と提案手法における未知語処理後の文の翻訳品質を比較した．比較方法として，人手評価と自動評価を行った．なお，人手評価には対比較評価を用いる．

6.3.1 人手評価結果

- ベースライン VS 提案手法

表 6.1 中のベースラインにおいて未知語を含む 3,170 文から，ランダムに抽出した 100 文を用いて，人手による対比較評価を行った．判断基準を表 6.5 に示す．また，評価結果を表 6.6 に示す．

表 6.5: ベースライン VS 提案手法における判断基準

提案手法○	提案手法の方がベースライン手法よりも良い
ベースライン○	ベースライン手法の方が提案手法よりも良い
差なし	双方とも意味がわからない，または，意味に差がない
同一出力	完全に同一の出力

表 6.6: ベースライン VS 提案手法の対比較評価結果 (100 文中)

提案手法○	ベースライン○	差なし	同一出力
15 文	0 文	84 文	1 文

表 6.6 の結果より，人手評価において，提案手法による翻訳精度の向上が確認できた．

- ベースライン VS 提案手法の出力例

ベースライン VS 提案手法における，提案手法○の出力例を表 6.7〜表 6.9，差なし場合の出力例を表 6.10〜表 6.12，同一出力の出力例を表 6.13 に示す.

(a) 提案手法○の出力例

表 6.7 において，ベースラインの “ていねい”が，提案手法では “respectful” となり意味が分かるようになったため，提案手法○とした.

表 6.7: 提案手法○の出力例

入力文	ていねい な 礼状 を 受け取っ た 。
正解文	I received a gracious letter of acknowledgment .
ベースライン	I received a ていねい message of thanks .
提案手法	I received a respectful message of thanks .

表 6.8 において，ベースラインの “恩恵”が，提案手法では “benefits” となり意味が分かるようになったため，提案手法○とした.

表 6.8: 提案手法○の出力例

入力文	万 人 が その 恩恵 に 浴 し て いる 。
正解文	The blessing is shared by all .
ベースライン	million people are with the 恩恵 .
提案手法	million people are with the benefits .

表 6.9 において，ベースラインの “草案”が，提案手法では “draft” となり意味が分かるようになったため，提案手法○とした.

表 6.9: 提案手法○の出力例

入力文	彼ら は 憲法 の 草案 を 作っ た 。
正解文	They prepared a draft of the constitution .
ベースライン	They made a 草案 of the constitution .
提案手法	They made a draft of the constitution .

(b) 差なしの出力例

表 6.10 において，ベースラインと提案手法の双方とも “ツアーガイドについていく”という意味が読み取れないため，差なしとした.

表 6.10: 差なしの出力例

入力文	彼女 は ツアー ガイド に ぴったり つい て い た 。
正解文	She fastened herself to the tour guide .
ベースライン	She was ツアー guide perfectly .
提案手法	She was tour guide perfectly .

表 6.11 において，ベースラインと提案手法の双方とも意味が分からないため，差なしとした.

表 6.11: 差なしの出力例

入力文	今 それ を 事細か に 述べる 暇 が ない 。
正解文	I have no time to go into its details .
ベースライン	There will be 事細か it now .
提案手法	There will be minutely it now .

表 6.12 において，ベースラインと提案手法の双方とも “笑いこけた”という意味が読み取れないため，差なしとした.

表 6.12: 差なしの出力例

入力文	彼ら は 笑いこけ た 。
正解文	They bent over with laughter .
ベースライン	They were 笑いこけ .
提案手法	They were The joke as .

(c) 同一出力の出力例

表 6.13 において，ベースラインと提案手法の出力文が完全に同一であったため，同一出力とした．

表 6.13: 同一出力の出力例

入力文	蝋 は たいてい の 物 の 表面 に くっつく 。
正解文	Wax adheres to most surfaces .
ベースライン	Most pill on the surface of the 蝋 through electrostatic forces .
提案手法	Most pill on the surface of the 蝋 through electrostatic forces .

- 先行手法 VS 提案手法

先行手法と提案手法の出力文から，それぞれランダムに抽出した 100 文を用いて，人手による対比較評価を行った．評価の基準を表 6.14 に示す．また，評価結果を表 6.15 に示す．

表 6.14: 先行手法 VS 提案手法における判断基準

提案手法〇	提案手法の方が先行手法よりも良い
先行手法〇	先行手法の方が提案手法よりも良い
差なし	双方とも意味がわからない，または，意味に差がない
同一出力	完全に同一の出力

表 6.15: 先行手法 VS 提案手法の対比較評価結果 (100 文中)

提案手法〇	先行手法〇	差なし	同一出力
5 文	10 文	62 文	23 文

表 6.15 より，人手評価における翻訳精度は，提案手法が先行手法よりも劣る結果となった．

- 先行手法 VS 提案手法における出力例

先行手法 VS 提案手法における，提案手法○の出力例を表 6.16〜表 6.18，先行手法○の出力例を表 6.19〜表 6.21，差なし場合の出力例を表 6.22〜表 6.24，同一出力の出力例を表 6.25〜表 6.27 に示す.

(a) 提案手法○の出力例

表 6.16 において，提案手法では “コーラスを聞いた” という事実が読み取れるが，先行手法では読み取れないため，提案手法○とした.

表 6.16: 提案手法○の出力例

入力文	わたし たち は コーラス の 美しい ハーモニー に 聞きほれ た 。
参照文	We were charmed by the beautiful harmony of the chorus .
提案手法	We listened in at the beautiful chorus .
先行手法	We ハーモニー in a chorus of the piano’s beautiful .

表 6.17 において，提案手法では “10 人の仕事ができる” という事実が読み取れるが，先行手法では読み取りにくいため，提案手法○とした.

表 6.17: 提案手法○の出力例

入力文	この ロボット は １ ０ 人 分 の 仕事 が できる 。
参照文	This robot can do the work of ten men .
提案手法	This robot can work of ten persons .
先行手法	The work of this robot is ten persons .

表 6.18 において，提案手法では “この物質が神経に影響を及ぼす” という事実が読み取れるが，先行手法では読み取れないため，提案手法○とした.

表 6.18: 提案手法○の出力例

入力文	この 物質 は 神経 を 冒す 。
参照文	The substance affects the nerves .
提案手法	This material affects the nerves .
先行手法	This substance nervous .

(b) 先行手法○の出力例

表 6.19 において，提案手法には動詞が無く意味が不適切であるが，先行手法には動詞があり意味が理解できるため，先行手法○とした.

表 6.19: 先行手法○の出力例

入力文	彼 は 怒り で 荒れ狂っ た 。
参照文	He raged with anger .
提案手法	He off his rough in anger .
先行手法	He raged in anger .

表 6.20 において，先行手法では “午後の日差しが部屋を満たす” という意味が読み取れるが，提案手法では読み取れないため，先行手法○とした.

表 6.20: 先行手法○の出力例

入力文	室内 に 午後 の 日差し が 満ち あふれ た 。
参照文	The light of the afternoon sun flooded into the room .
提案手法	The sun is full of in the room was full of this afternoon .
先行手法	The afternoon sun filled the room .

表 6.21 において，先行手法と提案手法の双方とも “昼食を取った” という意味が読み取れるが，先行手法では “昼食をたっぷり取った” ということまで分かるため，先行手法○とした.

表 6.21: 先行手法○の出力例

入力文	昼食 を たっぷり 取っ た 。
参照文	I had a big lunch .
提案手法	I took a lunch .
先行手法	I took plenty for lunch .

(c) 差なしの出力例

表 6.22 において，先行手法と提案手法の双方とも意味がわからないため，差なしとした．

表 6.22: 差なしの出力例

入力文	青信号 で 渡り なさい 。
参照文	Cross the road when the light is green .
提案手法	Migratory in redesign .
先行手法	Cross in Kadett .

表 6.23 において，先行手法と提案手法の双方とも意味がわからないため，差なしとした．

表 6.23: 差なしの出力例

入力文	便所 は 今 空い て いる 。
参照文	The washroom is empty now .
提案手法	now is vacant me where he could wash his hands .
先行手法	me where he could wash his hands now is vacant .

表 6.24 において，先行手法と提案手法の双方とも意味がわからないため，差なしとした．

表 6.24: 差なしの出力例

入力文	信仰 は 山 を も 動かす 。
参照文	Faith can move mountains .
提案手法	Work a mountain faith .
先行手法	faith the mountains .

(d) 同一出力の出力例

表 6.25 において，先行手法と提案手法の出力文が完全に同一であったため，同一出力とした．

表 6.25: 同一出力の出力例

入力文	彼 は ちょっと 考え込ん だ 。
参照文	He bethought himself a moment .
提案手法	He was thoughtful for a moment .
先行手法	He was thoughtful for a moment .

表 6.26 において，先行手法と提案手法の出力文が完全に同一であったため，同一出力とした．

表 6.26: 同一出力の出力例

入力文	この 製品 は 品質 が 落ち た 。
参照文	The quality of this product has gone down .
提案手法	The quality of this product .
先行手法	The quality of this product .

表 6.27 において，先行手法と提案手法の出力文が完全に同一であったため，同一出力とした．

表 6.27: 同一出力の出力例

入力文	私 は 都会 に 出 た 。
参照文	I arrived in the city .
提案手法	I went out into the city .
先行手法	I went out into the city .

6.3.2 自動評価結果

- ベースライン VS 提案手法

ベースラインと提案手法の出力文（10,000 文）に対して自動評価を行った．表 6.28 に，自動評価結果を示す．

表 6.28: ベースライン VS 提案手法の自動評価結果．精度が高い方を太字で示す

翻訳手法	BLEU	METEOR	RIBES	TER
ベースライン	**0.1394**	**0.4140**	**0.7138**	**0.6934**
提案手法	0.1391	0.4107	0.7116	0.7143

表 6.28 より，自動評価において，提案手法がベースラインよりも劣る結果となった．

- 先行手法 VS 提案手法

先行手法と提案手法の出力文（10,000 文）に対して自動評価を行った．表 6.29 に，自動評価結果を示す．

表 6.29: 先行手法 VS 提案手法の自動評価結果．精度が高い方を太字で示す

翻訳手法	BLEU	METEOR	RIBES	TER
先行手法	**0.1434**	**0.4203**	**0.7166**	**0.6978**
提案手法	0.1391	0.4107	0.7116	0.7143

表 6.29 より，自動評価において，提案手法が先行手法よりも劣る結果となった．

7　考察

7.1　二段階翻訳の効果

表 6.7〜表 6.9 より，翻訳品質が向上する文は，ベースラインの時点で文の構造がある程度良い，という特徴がある．また，表 6.10〜表 6.12 より，翻訳品質が向上しない文は，ベースラインの時点で文の構造が悪い，という特徴があることが分かった．したがって，ベースラインの翻訳品質がある程度良い場合においては，二段階翻訳は有効性があると考えられる．

7.2　評価方法の考察

本研究では，未知語の数の評価と，文全体の翻訳精度の評価として人手による対比較評価を行った．その結果，提案手法の有効性が確認できた．一方で，未知語をローマ字に変換すれば全ての未知語を削減できる．しかし，未知語処理後の文の翻訳品質が下がることが考えられる．例えば，“The temperature of 東京 is high .”という未知語を含む文の場合，固有名詞である “東京”をローマ字に変換し “TOKYO”とすると，“The temperature of TOKYO is high .”となり効果があると言える．一方で，“This part 壊れる a lot .”という未知語を含む文の場合，動詞である “壊れる”をローマ字に変換し “KOWARERU”としても，“This part KOWARERU a lot .”となり効果があるとは言えない．したがって，実際に，これらの事例がどれだけ存在しているかを調査した上で，提案手法と未知語をローマ字に変換する手法の比較を行い，評価する必要があると考える．

7.3 先行手法と提案手法の併用

本研究では，追加実験として，先行手法と提案手法を併用した手法（以下，追加手法と呼ぶ）の実験を行った．以下に追加手法の具体的な手順を示す．また，追加手法の流れを図 13 に示す．

準備 英語学習文と単語単位の日本語学習文および文字単位の日本語学習文を準備する．

手順 1 英語学習文を用いて言語モデルを作成する．

手順 2 英語学習文と日本語学習文を用いて翻訳モデルを作成する. また，ヒューリスティックスとして “grow-diag-final-and” を用いる．

手順 3 手順 2 と同様にして翻訳モデルを作成する. また，ヒューリスティックスとして “intersection” を用いる．

手順 4 手順 3 で作成されたフレーズテーブルから未知語が含まれるフレーズ対を抽出し，手順 2 で作成されたフレーズテーブルに直接追加する．このフレーズテーブルを用いて翻訳を行う．

手順 5 手順 4 の翻訳結果から未知語を抽出する．

“He has 画才 .” → “画才” (例 5)

手順 6 手順 3 で抽出した未知語を文字単位化し，次の入力とする．

“画才” → “画 才” (例 6)

手順 7 英語学習文と文字単位化した日本語学習文を用いて翻訳モデルを作成する．

手順 8 手順 1, 手順 7 で作成したモデルを用いて二回目の翻訳を行う．

“画 才” → “artistic talent” (例 7)

手順 9 手順 8 で出力された英語を，一回目の翻訳結果における未知語部分に置換して，英語文を出力する．

“He has 画才 .” → “He has artistic talent” (例 8)

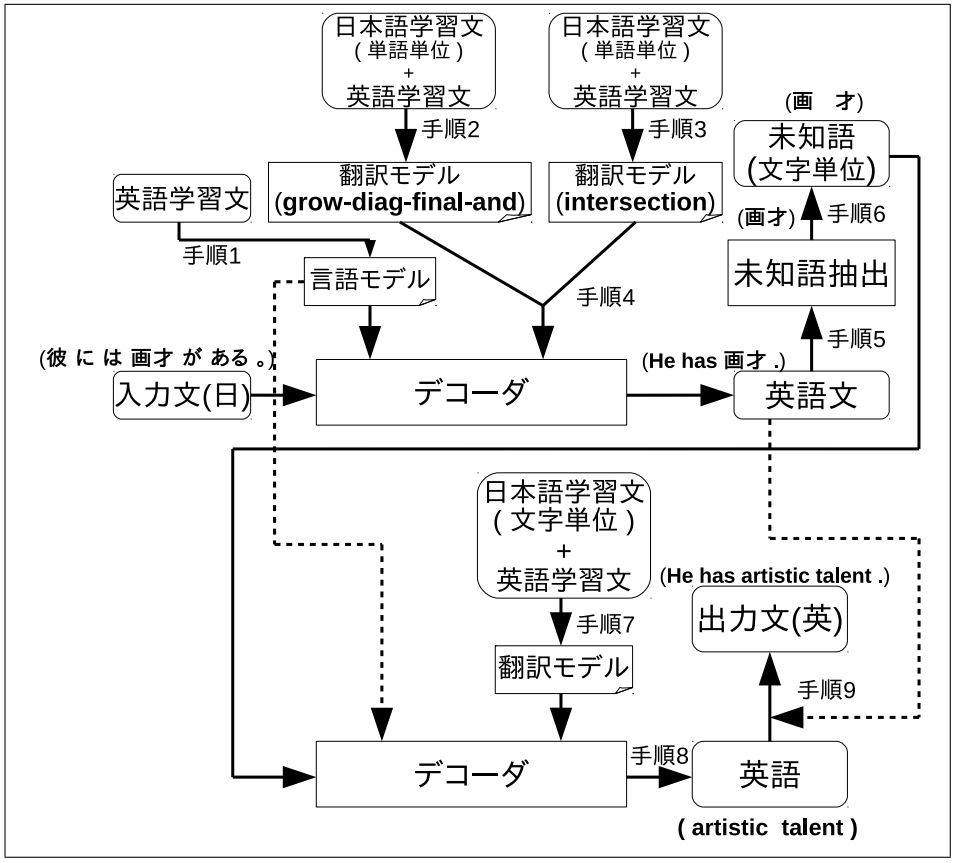

図 13: 日英統計翻訳における追加手法の流れ

7.3.1 追加手法の実験結果

- 未知語を含む文数

追加手法の出力文において未知語を含む文数を調査した結果を表 7.1 に示す.

表 7.1: 未知語を含む文数 (10,000 文中)

翻訳手法	未知語を含む文数	未知語の単語数
ベースライン	3,170 文	3,915 単語
提案手法	236 文	253 単語
先行手法	1,256 文	1,458 単語
追加手法	118 文	129 単語

表 7.1 より，先行手法と提案手法を組み合わせることにより，未知語を最も削減することができた.

- 未知語の翻訳品質

文字単位化した未知語をランダムに 100 単語抽出し，翻訳できた未知語数を調査した．この結果，翻訳できた未知語は 95 単語存在し，翻訳できなかった単語は 5 単語存在した．そして，この 95 単語の内，正しく翻訳できていた未知語は 6 単語存在した．正しく翻訳できた未知語の一例を表 7.2 に，正しく翻訳できなかった未知語の一例を表 7.3 に，翻訳できなかった未知語の一例を表 7.4 に示す.

表 7.2: 正しく翻訳できた未知語の一例（6 単語）

翻訳前	翻訳後
昨 春	last spring
失 火	fire lost
外 為 法	Foreign Exchange Law

表 7.3: 正しく翻訳できなかった未知語の一例（89 単語）

翻訳前	翻訳後
美 女	her beauty
分 厚 い	Heavy banks
マ ス ク	The public

表 7.4: 翻訳できなかった未知語の一例（5 単語）

翻訳前	翻訳後
嵎	嵎
冥 福	冥 Fukushima
遷 移	遷 to

- 追加手法におけるシステム全体の翻訳精度

追加手法におけるシステム全体の翻訳精度を調べるために，先行手法と追加手法における未知語処理後の文の翻訳品質を比較した．比較方法として，人手評価と自動評価を行った．人手評価では，表 6.1 中の先行手法において未知語を含む 1,281 文から，ランダムに抽出した 100 文を用いて，対比較評価を行った．判断基準を表 7.5 に示す．また，評価結果を表 7.6 に示す．

表 7.5: 先行手法 VS 追加手法における判断基準

追加手法○	追加手法の方が先行手法よりも良い
先行手法○	先行手法の方が追加手法よりも良い
差なし	双方とも意味がわからない，または，意味に差がない
同一出力	完全に同一の出力

表 7.6: 先行手法 VS 追加手法の対比較評価結果 (100 文中)

追加手法○	先行手法○	差なし	同一出力
1 文	1 文	98 文	0 文

表 7.6 の結果より，人手評価における追加手法の翻訳精度は向上しなかった．また，追加手法○の出力例を表 7.7，先行手法○の出力例を表 7.8，差なしの出力例を表 7.9 に示す．

表 7.7 において，先行手法の “一部分”が，追加手法では “partial” となり意味が分かるようになったため，追加手法○とした．

表 7.7: 追加手法○の出力例

入力文	それ は 理由 の 一部分 に すぎ ない 。
正解文	It is only one of the reasons .
先行手法	It is a mere 一部分 reason .
追加手法	It is a mere partial reason .

表 7.8 において，先行手法の “モーリーン”は人の名前であるが，追加手法では “the motor”となり人の名前でないため，反って意味が分からなくなった．そのため先行手法○とした．

表 7.8: 先行手法○の出力例

入力文	モーリーン が 新しい 支店 長 に 指名 さ れ た 。
正解文	Maureen was tapped as the new branch manager .
先行手法	The new manager was appointed to モーリーン .
追加手法	The new manager was appointed to the motor .

表 7.9 において，先行手法と追加手法において，双方とも意味が分からないため，差なしとした．

表 7.9: 差なしの出力例

入力文	学問 が 細分 化 する 。
正解文	The science gets more specialized .
先行手法	The scholarship 細分 .
追加手法	The scholarship into thin .

- 自動評価結果

先行手法と追加手法の出力文（10,000 文）に対して自動評価を行った．表 7.10 に，自動評価結果を示す.

表 7.10: 先行手法 VS 追加手法の自動評価結果. 精度が高い方を太字で示す

翻訳手法	BLEU	METEOR	RIBES	TER
先行手法	**0.1434**	**0.4203**	**0.7166**	**0.6978**
追加手法	0.1431	0.4169	0.7151	0.7100

表 7.10 より，自動評価において，追加手法が先行手法よりも劣る結果となった.

8　おわりに

統計翻訳では，対訳学習文から自動的に翻訳規則を獲得し，翻訳を行うため，翻訳精度は対訳学習文の量に大きく依存する．そのため，対訳学習文の量が少ない場合，未知語が出力される．代表的な未知語の対策として，対訳学習文を追加する方法が挙げられる．しかし，対訳学習文を追加するにはコストがかかる．この問題を解決するために，藤原らは，日英翻訳において，対訳学習文を追加せずに，フレーズテーブル作成時のヒューリスティックスを併用することで，未知語の削減と翻訳精度の改善を試みた．その結果，未知語の削減には成功したが，翻訳精度はほとんど向上しなかった．

本研究では，対訳学習文を追加せずに未知語処理を行う新たな手法を提案した．具体的には，出現した未知語を抽出し，文字単位化した後，文字単位化した未知語を入力として再度翻訳を行う．この手法を用いて未知語の削減と翻訳精度の向上を試みた．実験の結果，ベースラインでは未知語を含む文が 3,146 文出力されていたが，提案手法により 236 文まで削減することに成功した．更に，翻訳精度の向上も認められた．また，先行手法を併用した結果，未知語を含む文を 118 文まで削減することに成功した．

今後は，未知語をローマ字に変換する手法と提案手法の比較を行い，評価することを考えている．

謝辞

最後に，一年間に渡り，本研究のご指導を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 C 講座研究室の村上仁一准教授，村田真樹教授に深く感謝すると共に，厚く御礼申し上げます．そして，日常の議論を通じて多くの知識や示唆を頂いた同研究室の皆様に深謝いたします．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] Franz Josef Och, Hermann Ney: ”A Systematic Comparison of Various Statistical Alignment Models”, Computational Linguistics, volume 29, number 1, pp.19-51, March 2003.

[2] 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.

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

[4] 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.

[5] Papineni Kishore, Salim Roukos, Todd Ward, Wei-Jing Zhu: “BLEU: a method for automatic evaluation of machine translation”, 40th Annual meeting of the Association for Computational Linguistics pp. 311-318, 2002.

[6] Philipp Koehn, Marcello Federico, Brooke Cowan, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation”, Proceedings of the ACL 2007 Demo and Poster Sessions, pages 177-180, June 2007.

[7] SRILM(The SRI Language Modeling Toolkit) : srilm.tgz
http://www.speech.sri.com/projects/srilm/.

[8] 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.

[9] Mecab：mecab-0.97.tar.gz，mecab-ipadic-2.7.0-20070801.tar.gz
http://mecab.sourceforge.net/.

[10] 藤原勇: “パターン翻訳を用いた学習データ増加手法の検討”，修士論文，pp.43-59，2013.