句に基づく統計翻訳における未知語処理の１手法

藤原 勇　　村上 仁一　　徳久 雅人
鳥取大学大学院 工学研究科
{s072046, murakami, tokuhisa}@ike.tottori-u.ac.jp

1　はじめに

近年，機械翻訳において統計翻訳が注目され，盛んに研究が行われている．統計翻訳は，パラレルコーパスから自動的に翻訳規則を生成し，翻訳を行う手法である．統計翻訳において，翻訳されない単語は未知語として出力される．

この未知語を減少させるため，様々な試みが行われている．代表的な手法として，単語辞書などの対訳辞書データをパラレルコーパスに追加する手法がある [1]．しかし，この手法では，単語辞書などのパラレルコーパス以外のリソースが必要となる．

そこで本研究では，パラレルコーパスのみを利用して，未知語を削減する方法を提案する．統計翻訳では，翻訳の確からしさを表すモデルとして翻訳モデルを用いている．翻訳モデルは，フレーズテーブルと呼ばれる表で管理される．フレーズテーブルは単語対応からヒューリスティックスを用いて作成される．一般的に用いられるヒューリスティックス “grow-diag-final-and”では長いフレーズが作成され，短いフレーズが作成されない傾向にある．そして，翻訳において長いフレーズが優先的に利用されるため，短い単語列，特に 1 単語が未知語として出力される場合がある．

一方，ヒューリスティックスの一つである “intersection（単語対応の積集合）”を用いたフレーズテーブルには，未知語として出力された単語に対応するフレーズが存在する場合がある．しかし，“intersection”を用いた翻訳では，フレーズの候補が膨大になるため，枝刈り探索の問題から翻訳効率および翻訳精度が低下する．そこで，“grow-diag-final-and”のフレーズテーブルと，未知語として出力される単語に対応する “intersection”のフレーズテーブルを併用することで，未知語が軽減できる可能性がある．

本研究では，“grow-diag-final-and”と “intersection”を併用することで，未知語の軽減と翻訳精度の改善を目指す．

2　句に基づく統計翻訳の概略

現在統計翻訳には単語に基づく統計翻訳，句に基づく統計翻訳，階層型統計翻訳などがある．本研究では，句に基づく統計翻訳を用いる．

句に基づく日英統計翻訳は，与えられた日本語文 j について，翻訳モデルと言語モデルの組合せの中から確率値が最大となる英語文 $\hat{e}$ を探索することにより翻訳を行う．

$$\begin{aligned}\hat{e} &= argmax_e P(e|j) \\ &\approx argmax_e P(j|e)P(e)\end{aligned}$$

$P(\ j|e)$ は翻訳モデル，$P(e)$ は言語モデルと呼ばれる．また，$\hat{e}$ を探索するシステムはデコーダ [2] と呼ばれる．

2.1　翻訳モデル

翻訳モデルは，原言語の単語列から目的言語の単語列へ確率的に翻訳を行うためのモデルである．翻訳モデルは，フレーズテーブルと呼ばれる表により管理されている．表 1 にフレーズテーブルの例を示す．

表 1　フレーズテーブルの例

明日	Tomorrow	0.25 0.18 0.06 0.05
いくつ か	Some	0.05 0.01 0.01 0.04
うそ	lie	0.33 0.14 1 0.16

左から，日本語フレーズ，英語フレーズ，フレーズの日英翻訳確率 $P(j|e)$，単語の日英翻訳確率の積，フレーズの英日翻訳確率 $P(e|j)$，単語の英日翻訳確率の積である．本研究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．

2.1.1　フレーズテーブルの作成方法

フレーズテーブルは単語対応からヒューリステックスを利用して作成される．まず，GIZA++[3] により IBM 翻訳モデルを推定することで最尤な単語 alignment を得る．これを英日，日英の両方向に対して行う．なお，IBM モデルは単語を基本単位とした翻訳モデルである．そして両方向の alignment から，両方向に 1 対多の対応を認めた単語 alignment を計算する．この単語 alignment は基本的に両方向の単語対応の積集合と和集合の中間をヒューリスティックスで求める．各ヒューリスティックスの概要を以下に示す．

intersection(積集合)　日英方向および英日方向の両方向ともに単語対応が存在している場合を対応点とする．アライメントの精度を重視

union(和集合)　少なくとも片方向に単語対応が存在している場合を対応点とする．アライメントのカバー率を重視

grow-diag-final-and　積集合から始まり，和集合にしかない単語対応が妥当であるかを判断しながら，単語対応を追加する

3 未知語

統計翻訳において，学習データが十分にあっても，未知語が出力される．フレーズテーブル作成に grow-diag-final-and を利用した翻訳において，未知語が出力された例を表 2 に示す．また，翻訳の際に利用されたアライメントを表 3 に示す．

表 2 未知語の出力例

入力文:ライオン が 調教 師 に 歯向かっ た 。
参照文 : The lion rose up against its trainer .
出力文 : The lion 調教 turned against his teacher .

表 3 baseline のアライメント

ライオン が	The lion
調教	調教
に 歯向かっ	turned against
師	his teacher
た 。	.

表 2 において，“調教”が翻訳されず，未知語として出力されている．表 2 の翻訳で用いたフレーズテーブルの一部を表 4 に示す．

表 4 フレーズテーブル (grow-dial-final-and) の一部

ライオン は 調教 師 に かみつい た 。
||| The lion bit his trainer the arm .
調教 師 に かみつい た
||| bit his trainer the arm
調教 師 は
||| The trainer
調教 師 は ライオン を 飼い 馴らし た
||| The trainer tamed the lion .

表 4 より，“調教”のアライメントは存在するが，grow-diag-fianl-and のヒューリステックを利用しているため，このフレーズテーブルに “調教” 1 単語に対応するフレーズは存在しない．一方，intersection を利用して作成したフレーズテーブルでは，“調教”の 1 単語に対応するフレーズが存在する．intersection を用いたフレーズテーブルの例を表 5 に示す．

表 5 フレーズテーブル (intersection) の一部

調教 ||| The trainer
調教 ||| bit his trainer
調教 ||| his trainer
調教 ||| trainer

4 提案手法

前節で示した通り，grow-diag-final-and で作成したフレーズテーブルに存在せず，intersection で作成したフレーズテーブルに存在するフレーズ対がある．そこで本研究では，grow-diag-final-and を用いた翻訳において，未知語として出力された単語（ 1 単語）を日本語側に含む intersection のフレーズテーブル（表 5）を，grow-diag-final-and のフレーズテーブル（表 4）に追加する手法を提案する．

5 実験環境

5.1 実験データ

提案手法において，辞書より抽出した単文および重文・複文 [4] を用いた実験を行う．使用したデータの内訳を以下に示す．

表 6 実験データ（単文）

train	単文 150,000 文
test	単文 10,000 文
dev	単文 1,000 文

表 7 実験データ（重文・複文）

train	単文 100,000 文＋重文・複文 100,000 文
test	重文・複文 10,000 文
dev	重文・複文 1,000 文

5.2 評価手法

出力文の評価において，自動評価と人手評価を行う．自動評価法は BLEU[5]，METEOR[6]，RIBES[7] を用いる．また，人手による評価として対比較評価を行う．

6 実験結果

6.1 提案手法の効果

表 2 の例文における提案手法の出力例を表 8 に示す．表中のベースラインはフレーズテーブル作成に grow-diag-final-and を用いた場合の結果である．また，表 8 における提案手法のアライメントを表 9 に示す．

表 8 提案手法の出力例

入力文	ライオン が 調教 師 に 歯向かっ た 。
参照文	The lion rose up against its trainer .
ベースライン	The lion 調教 turned against his teacher .
提案手法	The lion turned against his trainer .

表 9 提案手法のアライメント

ライオン が	The lion
に 歯向かっ	turned against
師	his
調教	trainer
。	.

表 8 より，intersection のフレーズ対の追加による未知語の改善が確認できる．

6.2 未知語

6.2.1 未知語の改善

本章では，未知語が改善されている文を調査する．ベースライン（grow-diag-final-and）の出力において，未知語を含む文からランダムに 100 文を抽出し，提案手法の出力との比較を行った．そして，提案手法において

未知語が改善されている文数を調査した．結果を表 10 に示す．また，例を表 11 に例を示す．

表 10　未知語の改善文数

単文	重文・複文
61/100	68/100

表 11　未知語改善の例

入力文	桃太郎 は 鬼 ども を 退治 し た 。
参照文	Momotaro subdued the ogres .
ベースライン	We finish off the 桃太郎 is so .
提案手法	We finish off the Momotaro is so .
入力文	彼 は 純情 な 男 で だまさ れ 易い 。
参照文	He is naive and easily cheated .
ベースライン	He was a man 純情 .
提案手法	He is a man was innocent .

表 11 の例では “桃太郎” と “純情” の未知語が改善されている．しかし，全体の翻訳品質に影響はない．

6.2.2　未知語改善による翻訳品質への影響

未知語が改善した文において，翻訳品質が向上した文数の調査を行った．結果を表 12 に示す．また，表 13 に例を示す．

表 12　翻訳品質が向上した文数

単文	重文・複文
21/61	12/68

翻訳品質が向上した場合以外の文（単文 40 文，重文・複文 46 文）において，ベースラインと提案手法の翻訳品質はほぼ同等であり，提案手法の翻訳品質が低下した例は存在しなかった．

表 13　未知語改善による翻訳品質向上の例

入力文	憂鬱 な 天気 が 続い て いる 。
参照文	The weather has been gloomy .
ベースライン	The weather has been a 憂鬱 .
提案手法	The weather has been a dark .
入力文	彼ら は 勝ち誇っ て 帰っ て 来 た 。
参照文	They came home in triumph .
ベースライン	They 勝ち誇っ came home .
提案手法	They came back with a triumphant air .

表 13 では “憂鬱” や “勝ち誇っ” が翻訳されることで翻訳品質が向上している．したがって，未知語に対する提案手法の有効性が認められる．

6.3　システム全体に対する提案手法の影響

6.3.1　全出力文中の未知語数

ベースラインと提案手法それぞれの出力文 10,000 文において，未知語として出力された単語数を表 14 に示す．

結果より，単文および重文・複文それぞれの実験において，未知語として出力された単語の減少が確認できる．したがって，出力文全体においても提案手法の効果が確認できる．

表 14　未知語数

	単文	重文・複文
ベースライン	2,486	2,400
提案手法	974	950

6.3.2　自動評価

表 15 に翻訳実験の自動評価結果を示す．表中のベースラインとはフレーズテーブル作成に “grow-diag-final-and” を用いた結果である．

表 15　自動評価結果

	BLEU	METEOR	RIBES
単文			
ベースライン	0.1618	0.5333	0.7275
提案手法	0.1649	0.5331	0.7328
重文・複文			
ベースライン	0.1626	0.4896	0.6940
提案手法	0.1627	0.4824	0.6931

結果より，単文実験において，提案手法の効果がわずかに認められる．しかし，重文・複文実験においては，提案手法の効果は認められない．

6.3.3　人手評価結果

提案手法の出力文 10,000 文からランダムに抽出した 100 文に対して，ベースラインとの対比較評価を行う．判断基準を以下に示す．

提案手法○　提案手法の出力結果がベースラインの出力結果よりも優れている
提案手法×　提案手法の出力結果がベースラインの出力結果よりも劣っている
差なし　文質に明確な差がない
同一出力　完全に同一の出力

- 人手評価結果
  表 16 に人手による対比較評価の結果を示す．

表 16　人手評価結果

提案手法○	提案手法×	差なし	同一出力
単文			
7	4	45	44
重文・複文			
5	2	53	40

- 対比較評価出力例
  対比較調査におけるそれぞれの評価の文例を表 17 に示す．

表 17 対比較調査文例

提案手法○（単文）	
入力文	国王 は 退位 さ せ られ た 。
参照文	The King has been deposed .
ベースライン	The king was sent 退位 .
提案手法	The king abdicated the throne .
提案手法×（単文）	
入力文	あなた は 姉さん に 非常 に よく 似 て いる 。
参照文	You bear a strong likeness to your sister .
ベースライン	You are very much alike in the older sister .
提案手法	You are very much alike the girl .
提案手法○（重文・複文）	
入力文	この 仕事 は どうしても 引受ける 気 に なれ ぬ 。
参照文	I can not bring myself to undertake the work .
ベースライン	This work is , I can not bring myself to 引受ける .
提案手法	I can not bring myself to undertake the work by all means .
提案手法×（重文・複文）	
入力文	彼 は 私 に それ を 盗め と はっきり 言っ た 。
参照文	He told me expressly to steal it .
ベースライン	He advised me you steal the it .
提案手法	He did it to you steal the .

7 解析

7.1 intersection との比較

フレーズテーブル作成に intersection を用いた結果と提案手法の比較を行った．自動評価結果を表 18 に示す．

表 18 自動評価結果

	BLEU	METEOR	RIBES
単文			
intersection	0.1552	0.5102	0.724
提案手法	0.1649	0.5331	0.7328
重文・複文			
intersection	0.1531	0.4575	0.6863
提案手法	0.1627	0.4824	0.6931

さらに，表 16 の人手による対比較調査において，提案手法○と判断された文，単文 7 文例，重文・複文 5 文例において，intersection の出力文との比較を行った．結果を表 19 に示す．また，それぞれの実験における提案手法○の例を表 20 に示す．

結果より，intersection と比較においても，提案手法の有効性が確認できる．

表 19 人手評価結果

提案手法○	提案手法×	差なし	同一出力
単文			
4	0	2	1
重文・複文			
5	0	0	0

表 20 対比較評価文例

提案手法○（単文）	
入力文	何とか 金 の 工面 が つい た 。
参照文	I managed to raise the money .
ベースライン	I 工面 of money .
intersection	I somehow manage of money
提案手法	I somehow could manage of money .
提案手法○（重文・複文）	
入力文	この 仕事 は どうしても 引受ける 気 に なれ ぬ 。
参照文	I can not bring myself to undertake the work .
ベースライン	This work is , I can not bring myself to 引受ける .
intersection	This work , but I can not bring myself to undertake it by all means .
提案手法	I can not bring myself to undertake the work by all means .

8 まとめ

本論文では，統計翻訳における未知語を減少させる手法を提案した．提案手法の特徴として，フレーズテーブルの作成におけるヒューリスティックスの併用を行うことで，対訳辞書データなどの外部リソースを必要としない点が挙げられる．ヒューリスティックスの併用として，“grow-diag-final-and”のフレーズテーブルと未知語として出力される単語に対応する “intersection”のフレーズテーブルを併用する．実験の結果，出力文全体の自動評価値に影響はなかったが，未知語の減少に大きな効果が認められた．ベースラインの出力において未知語を含む文 100 文中，単文を用いた実験では 61 文，重文・複文を用いた実験では 68 文の未知語が改善した．さらに，未知語が改善した文のうち，単文実験において 61 文中 21 文，重文・複文実験において 68 文中 12 文の翻訳品質が向上した．したがって，未知語問題に対して，提案手法の有効性が認められる．

参考文献

[1] 日野聡子, 村上仁一, 徳久雅人, 村田真樹: “鳥バンクと英辞郎を日英対訳文に追加した統計翻訳の調査”, 言語処理学会第 18 回年次大会発表論文集, pp.491-494, 2012.

[2] Philipp Koehn, Marcello Federico, Brooke Cowan, Richard Zens, Chris Dyer, Alexan dra Constantin, Evan Herbst: “Moses: Open Source Toolkit for Statistical Machin e Translation”, Proc. of the ACL, pp.177-180, 2007.

[3] Franz Josef Och, Hermann Ney: ”A Systematic Comparison of Various Statistical Alignment Models”, Computational Linguistics, pp.19-51, 2003.

[4] 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.

[5] Papineni Kishore, Salim Roukos, Todd Ward, Wei-Jing Zhu: “BLEU: A Method for Automatic Evaluation of Machine Translation”, Proc. of the ACL, pp.311-318, 2002.

[6] Satanjeev Banerjee, Alon Lavie: “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”, Proc. of the ACL, pp.65-72, 2005.

[7] 平尾努, 磯崎秀樹, Kevin Duh, 須藤克仁, 塚田元, 永田昌明: “RIBES: 順位相関に基づく翻訳の自動評価法”, 言語処理学会第 17 年次大会発表論文集, pp.1111-1114, 2011.