パターンに基づく統計翻訳とその未知語処理

川原 宰　村上仁一
鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
{s122019,murakami,tokuhisa}@ike.tottori-u.ac.jp

1　はじめに

パターン翻訳とは，機械翻訳における一種の翻訳手法であり，句辞書と文パターン辞書を用いて翻訳を行う．入力文が適切な文パターンと照合した場合，翻訳精度の高い文が得られる．しかし，人手で辞書を作成するためコストがかかる．

この問題を解決するため，江木らは，統計的手法を用いて自動的に辞書を作成して翻訳を行う “パターンに基づく統計翻訳 [1]”(以下，従来手法) を提案した．従来手法により，辞書作成のコストを削減することに成功した．しかし，従来手法の問題点として，訳語が単語辞書にあるにも関わらず未知語が出力されることが挙げられる．

そこで，本研究では，未知語処理として単語辞書を用いる方法を 2 手法提案する．提案手法 A では，従来手法の出力文から未知語を抽出し，単語辞書で検索を行い，翻訳確率が最大の訳語を元の未知語部分に置換して翻訳する．提案手法 B では，句辞書を適用する段階で単語辞書を加えて翻訳する．これらの手法により未知語を削減し，翻訳精度向上を試みる．

2　従来手法 [1]

2.1　全体の流れと具体的な手順

従来手法では 6 つのステップを用いて翻訳を行う．日英統計翻訳における従来手法の全体の流れを図 1 に示す．

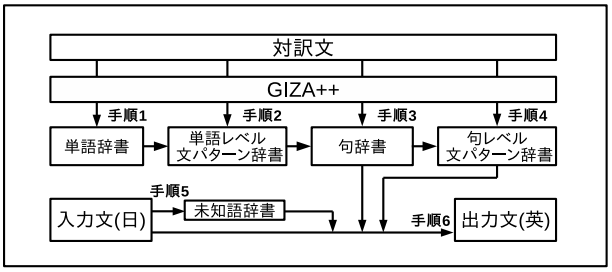

図 1　日英統計翻訳における従来手法の全体の流れ

また，具体的な手順を以下に示す．

手順 1　対訳文と GIZA++[2] を用いて単語辞書を作成

手順 2　対訳文と単語辞書を用いて単語レベル文パターンを作成

手順 3　対訳文と単語レベル文パターンを用いて句辞書を作成

手順 4　対訳文と句辞書を用いて句レベル文パターンを作成

手順 5　入力文を用いて未知語辞書を作成

手順 6　句レベル文パターン辞書と句辞書/未知語辞書を用いて翻訳

2.2　未知語辞書

従来手法では，未知語を出力する際に入力文から作成した未知語辞書を用いる．具体的には，入力文において考えられる全ての日本語句の組み合わせを全ての英語句の未知語として辞書に登録する．未知語辞書作成の流れを図 2 に示す．

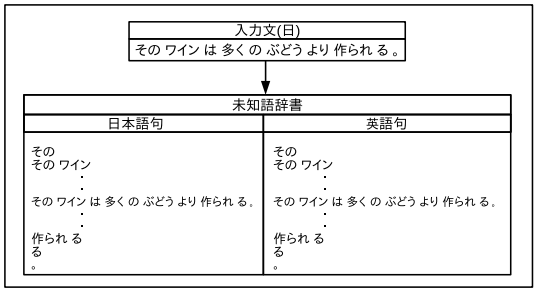

図 2　未知語辞書辞書作成の流れ

2.3　翻訳

従来手法では，入力文と句レベル文パターン辞書を照合して文パターンを決定した後，各変数部に対して句辞書を適用して翻訳する．この時，句辞書に変数化した日本語句に対応する英語句が存在しない場合は，未知語辞書を適用し未知語として出力する．翻訳の流れを図 3 に示す．

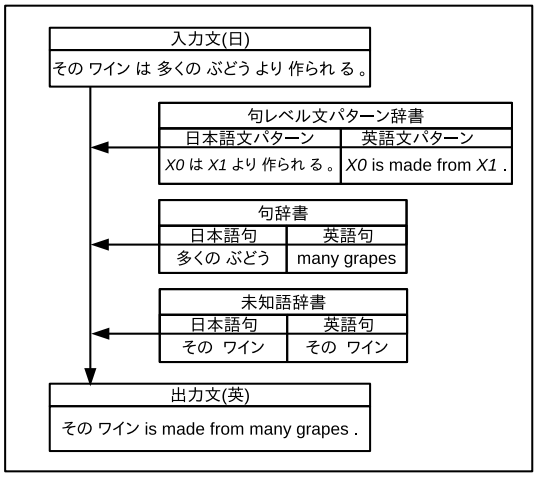

図 3　従来手法における翻訳の流れ

3 提案手法 A

提案手法 A では，従来手法の出力文から未知語を抽出した後，単語辞書で訳語検索を行い，翻訳確率が最大の単語を元の未知語部分に置換して出力する．日英統計翻訳における提案手法 A の流れを図 4 に示す．

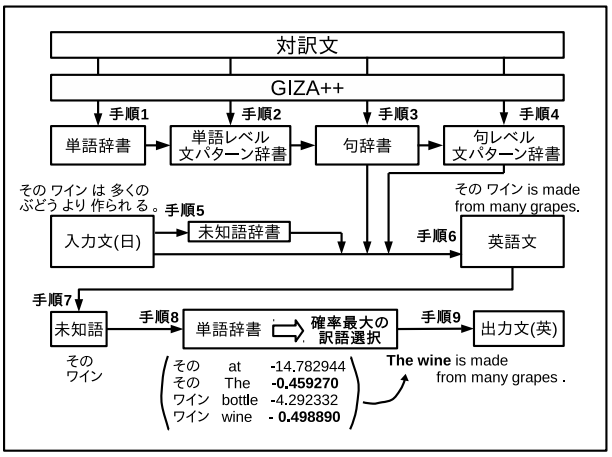

図 4　日英統計翻訳における提案手法 A の流れ

また，具体的な手順を以下に示す．

手順 1〜6　従来手法と同様

手順 7　手順 6 の英語文において未知語が存在する場合は未知語を抽出

手順 8　手順 1 で作成した単語辞書において，手順 7 で抽出した未知語を検索し，翻訳確率最大の訳語を選択

手順 9　手順 8 で訳語が見つかった場合はその訳語を元の未知語部分に置換して出力

4 提案手法 B

提案手法 B では，従来手法における手順 6(もしくは図 3) の段階で単語辞書を加えて翻訳する．日英統計翻訳における提案手法 B の流れを図 5 に示す．

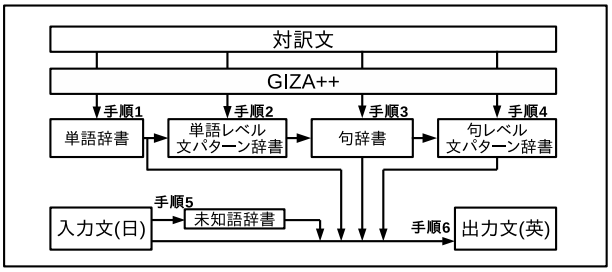

図 5　日英統計翻訳における提案手法 B の流れ

また，具体的な手順を以下に示す．

手順 1〜5　従来手法と同様

手順 6　句レベル文パターン辞書と句辞書/単語辞書を用いて翻訳

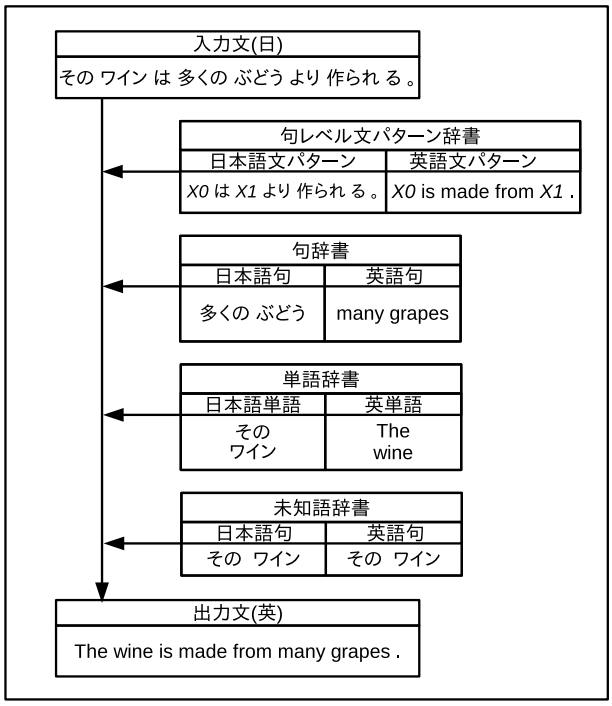

図 6　提案手法 B における翻訳の流れ

5 実験環境

5.1 実験データ

本実験には，電子辞書などの例文より抽出した単文コーパス [3] を用いる．使用するデータの内訳を表 1 に示す．

表 1　実験データ

学習文	100,000 文
テスト文	1,000 文

5.2 コーパスの前処理

本実験では，単文コーパスに対して前処理を行う．具体的には，日本語文に対して “MeCab”による形態素解析を行い，英語文に対して “tokenizer.perl”による分かち書きを行う．前処理を行った単文コーパスの例を表 2 に示す．

表 2　前処理を行った単文コーパスの例

空き巣 に 入ら れ た 。
Our house was robbed while we were away .

5.3 評価方法

本研究では，未知語数の比較評価と，出力文の翻訳精度の比較評価を行う．具体的には，従来手法の出力文と提案手法 A と B の出力文を比較する．また，出力文の翻訳精度の評価として人手評価と自動評価を行う．人手評価として対比較評価を行う．なお，自動評価には BLEU, METEOR，RIBES，TER および WER を用いる．

6 実験結果

6.1 未知語処理結果

提案手法 A と B における未知語数および未知語の翻訳精度の結果を以下に示す．

6.1.1 未知語数

従来手法と提案手法 A と B において，未知語を含む文数の調査を行った．調査結果を表 3 に示す．

表 3　未知語を含む文数 (1,000 文中)

翻訳手法	未知語を含む文数
従来手法	516 文
提案手法 A	121 文
提案手法 B	325 文

表 3 より，提案手法 A において未知語を含む文を大幅に削減することができた．

6.1.2 提案手法 A における未知語の翻訳精度

従来手法からランダムに抽出した未知語を含む文 100 文において未知語は 213 単語存在した．提案手法 A の未知語処理結果を表 4 に結果を示す．

表 4　提案手法 A における未知語の翻訳結果 (213 単語中)

翻訳成功	正しい翻訳	誤った翻訳
186 単語	96/186 単語	90/186

また，表 4 において，正しい翻訳結果の一例を表 5 に示す．誤った翻訳結果の一例を表 6 に示す．

表 5　正しく翻訳できた未知語の一例（96 単語）

草花	plants
運輸	transport
警棒	truncheon
カギ	key

表 6　正しく翻訳できなかった未知語の一例（90 単語）

共闘	joint
差し上げ	you
生傷	scamp
根深い	palpable

6.1.3 提案手法 B における未知語の翻訳精度

提案手法 B において，未知語処理を行った単語は約 210 単語[*1]存在した．未知語処理結果を表 7 に結果を示す．

表 7　提案手法 B における未知語の翻訳結果 (約 210 単語中)

翻訳成功	正しい翻訳	誤った翻訳
60 単語	20/60 単語	40/60

6.2 出力文の翻訳結果

提案手法 A と B における出力文の翻訳精度を従来手法の出力文と比較した．比較方法として，人手評価と自動評価を行った．以下に評価結果を示す．

6.2.1 提案手法 A における人手評価結果

従来手法と提案手法 A 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．評価の基準を以下に示す．評価結果を表 8 に示す．

- 提案手法 A ○　：提案手法 A の方が良い
- 提案手法 A ×　：提案手法 A の方が悪い
- 差なし　：翻訳精度に明確な差がない
- 同一出力　：完全に同一の出力

表 8　従来手法 VS 提案手法 A の評価結果 (100 文中)

提案手法 A ○	提案手法 A ×	差なし	同一出力
4 文	2 文	79 文	15 文

また，提案手法 A ○と提案手法 A ×の場合の出力例を以下に示す．

表 9 において，従来手法では “炒っ”という動詞が未知語となり文全体を通して意味が通じないが，提案手法では未知語が翻訳され意味が通じるようになったため，提案手法 A ○とした．

表 9　提案手法 A ○の出力例

入力文	彼 が コーヒー 豆 を 炒っ た 。
参照文	He roasted the coffee beans .
従来手法	He 炒っ the coffee beans .
提案手法 A ○	He roasted the coffee beans .

表 10 において，“ちょうだい”が “wedding”に翻訳され不適切な対応をとっており，なおかつ，文全体の日本語訳が “私は完全な挙式をする”という誤った意味で通じてしまうため，提案手法 A ×とした．

表 10　提案手法 A ×の出力例

入力文	十分 に ちょうだい し ま し た 。
参照文	I have had enough .
従来手法	I have ちょうだい fully .
提案手法 A ×	I have wedding fully .

6.2.2 提案手法 B における人手評価結果

従来手法と提案手法 B 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．評価の基準は提案手法 A の場合と同様である．評価結果を表 11 に示す．

表 11　従来手法 VS 提案手法 B の評価結果 (100 文中)

提案手法 B ○	提案手法 B ×	差なし	同一出力
4 文	3 文	60 文	33 文

また，提案手法 B ○と提案手法 B ×の場合の出力例を以下に示す．

表 12 において，従来手法では “中身”という名詞が未知語となり文全体を通して意味が通じないが，提案手法では未知語が正しく翻訳され意味が通じるようになったため，提案手法 B ○とした．

[*1] 実験の都合上推定値としている

表 12　提案手法 B ○の出力例

入力文	問題 は その 中身 で ある 。
参照文	What is important is the content of the budget .
従来手法	The problem is 中身 that .
提案手法 B ○	The issue is the contents .

表 13 において，“免疫”が “immune”に正しく翻訳されているが，文全体の日本語訳が “私はインフルエンザの免疫がない”という意味で通じてしまうため，提案手法 B ×とした．

表 13　提案手法 B ×の出力例

入力文	わたし は インフルエンザ に は 免疫 が ある 。
参照文	I am immune to the flu .
従来手法	I have a 免疫 the flu .
提案手法 B ×	I have not immune to the flu .

6.2.3　自動評価結果

テスト文 1,000 文を入力として翻訳実験を行い，出力文に対して自動評価を行った．表 14 に，それぞれの手法における自動評価の結果を示す．

表 14　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	WER
従来手法	0.0917	0.3485	0.6963	0.7075	0.7209
提案手法 A	0.0945	**0.3718**	**0.7075**	**0.6906**	**0.7081**
提案手法 B	**0.0982**	0.3633	0.6941	0.6993	0.7135

6.3　実験結果のまとめ

表 3 より，提案手法 A と B の両方において未知語の削減に成功した．特に提案手法 A においては大幅に未知語を削減できた．しかし，表 8 と 11 と 14 より，従来手法と比べて翻訳精度に差がないことが確認できた．したがって，提案手法 A において大幅な未知語の削減が翻訳精度向上にほとんど影響しないことが分かった．

7　考察

7.1　未知語処理による翻訳精度への影響

第 6.3 節で述べた結果の理由として以下の 2 点が挙げられる．

7.1.1　従来手法の翻訳精度

提案手法 A では，従来手法の出力文に対して未知語処理を行うため従来手法の翻訳精度が大きく影響する．表 8 における “差なし”の 79 文中 54 文は，従来手法の翻訳精度が低いため “差なし”としている．例えば表 15 の場合，未知語は正しい意味に翻訳できているが，従来手法における出力文の段階で翻訳精度が低い．そのため，提案手法 A において翻訳精度が向上しなかった．

表 15　従来手法の翻訳精度が低いため “差なし”とした例

入力文	ダイアナ 妃 は お忍び で ディスコ に 現れ た 。
参照文	Princess Diana showed up at the disco incognito .
従来手法	Princess appeared on a お忍び Princess disco .
提案手法	Princess appeared on a incognito Princess disco .

7.1.2　句レベルの未知語

従来手法の出力文において，句レベルの未知語を含む文が約 2 割存在した．表 16 に句レベルの未知語を含む文の例を示す．この場合，“上陸”以外の未知語はそれぞれ翻訳に成功しているが，句レベルで見た場合には翻訳に成功していない．さらに，句レベルの未知語の傾向として “A の B”というパターンが多く存在した．このようなパターンの場合，基本的には “B of A”という形に翻訳する必要があるが，提案手法 A では “A of B”という形にしか翻訳できない．したがって提案手法 A では，句レベルの未知語を含む文において翻訳精度が向上しなかった．

表 16　句レベルの未知語を含む文の例

入力文	このヨーロッパ から の 移民 の 上陸 地点 は この 島 に あっ た 。
参照文	The landing place of immigrants from Europe was on this island .
従来手法	from Europe 移民 の 上陸 地点 This was on the islands .
提案手法	from Europe immigrants of shore point This was on the islands .

7.2　提案手法 A と B の性能差

表 8 と 11 より，提案手法 A と B では出力文の翻訳精度にほとんど差がないと言える．一方で，表 4 と 7 より，提案手法 B における未知語の翻訳精度は約 30% であるのに対して，提案手法 A における未知語の翻訳精度は約 50% であり，約 20% の差が生じている．したがって，提案手法 A の方が提案手法 B よりも僅かながら有効であると考える．

8　おわりに

本研究では，従来手法に単語辞書を用いた未知語処理を加えた 2 手法を提案した．提案手法 A では，従来手法の出力文から未知語を抽出し，単語辞書で検索を行い，翻訳確率が最大の訳語を元の未知語部分に置換して翻訳を行った．提案手法 B では，句辞書を適用する段階で単語辞書を加えて翻訳を行った．これらの結果，未知語を大幅に削減することができた．しかし，翻訳精度向上は僅かであった．また，提案手法 A の方が提案手法 B よりも僅かながら有効であるという結果になった．今後は，更なる未知語処理の手法を検討したい．

参考文献

[1] 江木孝史 “句に基づく対訳句パターンの自動作成と統計的手法を用いた英日パターン翻訳”，言語処理学会第 20 回年次大会, A6-2, pp.951-954, 2014.

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

[3] 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.