文字ベース翻訳による未知語処理

川原 宰 [*1]　村上仁一 [*2]　徳久雅人 [*2]
[*1] 鳥取大学 工学部 知能情報工学科
[*2] 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

統計翻訳において，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．一般的な未知語の対策として，学習データを追加する方法が挙げられる．しかし，学習データを追加するにはコストがかかる．そこで，学習データを追加せずに未知語処理を行う手法を藤原らが提案した [1].

本研究では，学習データ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，文字単位にした後，文字単位にした未知語を入力として再度翻訳を行う手法である．この手法を用いて未知語の削減を試みる．

2　先行手法 [1]

藤原らの手法では，学習データを追加せずに未知語処理を行う．具体的には，2 種類のヒューリスティックスを用いたフレーズテーブルを併用して翻訳を行う．

以下に具体的な手順を示す．また，先行手法の流れを図 1 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．

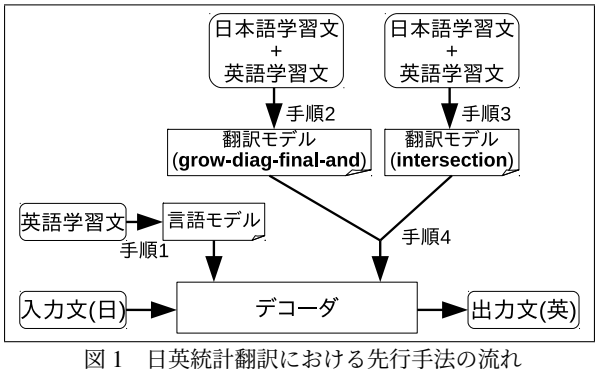

図 1　日英統計翻訳における先行手法の流れ

3　提案手法

本研究では，学習データを追加せずに未知語処理を行う．具体的には，出現した未知語を抽出し，文字単位にした後，文字単位にした未知語に対して再度翻訳を行う．

以下に具体的な手順を示す．また，提案手法の流れを図 2 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にし，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に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 skill”　(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 skill”　(例 4)

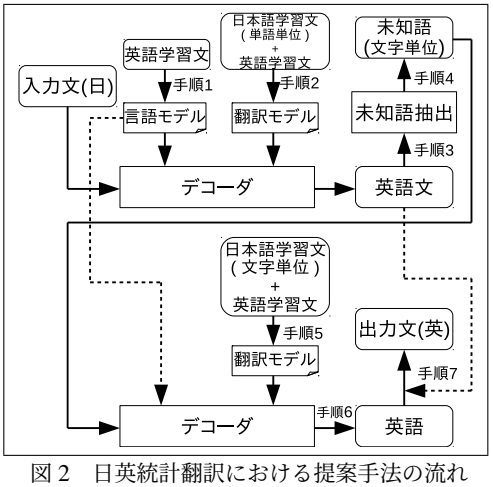

図 2　日英統計翻訳における提案手法の流れ

4 実験環境

4.1 言語モデルの学習

言語モデルの学習には，“SRILM[2]”の “ngram-count” を用いる．本研究では，*N*-gram モデルに 5-gram を用いる．

4.2 翻訳モデルの学習

翻訳モデルの学習には, “train-model.perl[3]”を用いる．

4.3 デコーダ

本研究では, デコーダとして, “moses[3]”を用いる．また, moses の各パラメータは “mert-moses.pl[3]”を用いて最適化する．また，“ttable-limit”の値を 10，“distortion-limit”の値を-1 として実験を行う．

4.4 実験データ

4.4.1 単文コーパス

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 [4] 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

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データ	1,000 文
テストデータ	10,000 文

統計翻訳の前処理として，各コーパスの日本語文に対して，“MeCab[5]”を用いて形態素解析を行う．また，英語文に対して “tokenizer.perl[3]”を用いて分かち書きを行う．

4.5 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，提案手法，ベースライン，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，人手評価と自動評価を行う．人手評価では，対比較評価を行う．なお，自動評価には BLEU, METEOR，RIBES および TER を用いる．

5 実験結果

5.1 未知語を含む文数

ベースライン，提案手法，先行手法において，未知語を含む文数の調査を行った．調査結果を表 2 に示す．

表 2　未知語を含む文数 (10000 文中)

翻訳手法	未知語を含む文数
ベースライン	3170 文
提案手法	236 文
先行手法	1256 文

表 2 より，提案手法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．また，先行手法よりも未知語を含む文を削減できたことが分かる．

5.2 未知語の翻訳品質

文字単位に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 単語存在した．正しく翻訳できた未知語の一例を表 3 に示す．正しく翻訳できなかった未知語の一例を表 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 3　正しく翻訳できた未知語の一例（21 単語）

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

表 4　正しく翻訳できなかった未知語の一例（72 単語）

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

5.3 未知語処理後の文の翻訳品質

未知語処理後の文の翻訳品質を調べるために，5.2 節において正しく翻訳できた 21 単語を，一回目の翻訳結果の未知語部分に置換した．その結果，翻訳品質が向上した文は 4 文存在し，翻訳品質が向上しなかった文は 16 文存在した．翻訳品質が向上した例を表 5 に示す．また，翻訳品質が向上しなかった例を表 6 に示す．

表 5　翻訳品質が向上した例（4 文）

入力文	彼女 の 目 から 輝き が 消え た 。
参照文	The light died out of her eyes .
ベースライン	The 輝き away from her eyes .
提案手法	The sheen away from her eyes .
入力文	彼 は 彼女 に 恋い焦がれ て いる 。
参照文	He is desperately in love with her .
ベースライン	He is 恋い焦がれ to her .
提案手法	He is eagerly love to her .

表 6　翻訳品質が向上しなかった例（16 文）

入力文	多少 の 誤差 は 気 に する 必要 は ない 。
参照文	There is no need to get nervous about errors to some degree .
ベースライン	There is no need to in some of 誤差 .
提案手法	There is no need to in some of error .
入力文	少年 の 快活 さ が 私 たち の 悲しみ を やわらげ た 。
参照文	The cheerfulness of the boy eased our sorrow .
ベースライン	The 快活 of the boy into our grief .
提案手法	The cheerfulness of the boy into our grief .

6 先行手法と提案手法の比較

提案手法の有効性を調べるために，未知語処理後の文の翻訳品質を先行手法と比較した．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．

6.1 人手評価

人手評価として，対比較評価を行った．

6.1.1 対比較評価結果

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．評価の基準を以下に示す．また，評価結果を表 7 に示す．

- 提案手法○　：提案手法の方が良い
- 先行手法○　：先行手法の方が良い
- 差なし　：翻訳品質に明確な差がない
- 同一出力　：完全に同一の出力

表 7　対比較評価 (100 文中)

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

表 7 より，人手評価において，提案手法が先行手法よりも劣る結果となった．

6.1.2 出力例

提案手法○と先行手法○の場合の出力例を以下に示す．

表 8 において，“コーラスを聞いた”という事実を提案手法では読み取れるが，先行手法では読み取れないため，提案手法○とした．

表 8　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 9 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 9　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 .

6.2 自動評価

テスト文 10,000 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 10 に，それぞれの手法における自動評価の結果を示す．

表 10　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
提案手法	0.1391	0.4107	0.7116	0.7143
先行手法	**0.1434**	**0.4203**	**0.7166**	**0.6978**

表 10 より，翻訳精度においては提案手法が先行手法よりも劣る結果となった．

7 追加実験

追加実験として，先行手法と提案手法を組み合わせた実験を行った．以下に具体的な手順を示す．また，追加実験の流れを図 3 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にし，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に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 skill”　(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 skill”　(例 8)

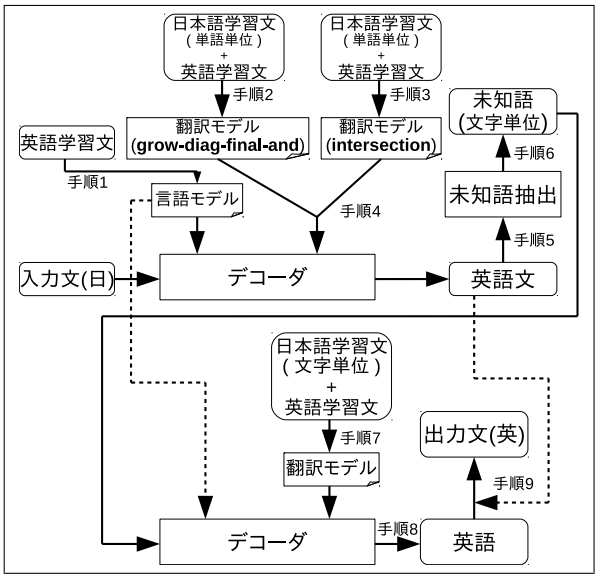

図 3　追加実験の流れ

7.1 実験結果

7.1.1 未知語を含む文数

追加実験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 11 に示す.

表 11 未知語を含む文数 (10000 文中)

翻訳手法	未知語を含む文数
ベースライン	3170 文
提案手法	236 文
先行手法	1256 文
先行&提案	118 文

表 11 より，先行手法と提案手法を組み合わせることにより，未知語を含む文数を最も少ない数まで削減することができた.

7.1.2 未知語の翻訳品質

文字単位に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 12 に示す．正しく翻訳できなかった未知語の一例を表 13 に示す.

表 12 正しく翻訳できた未知語の一例（6 単語）

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

表 13 正しく翻訳できなかった未知語の一例（89 単語）

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

7.1.3 未知語処理後の文の翻訳品質

未知語処理後の文の翻訳品質を調べるために，7.1.2 節において正しく翻訳できた 6 単語を，一回目の翻訳結果の未知語部分に置換した．その結果，翻訳品質が向上した文は 1 文存在し，翻訳品質が向上しなかった文は 5 文存在した．翻訳品質が向上した例を表 14 に示す．また，翻訳品質が向上しなかった例を表 15 に示す.

表 14 翻訳品質が向上した例（1 文）

入力文	紙切れ に 何 か 書い た 。
参照文	He wrote something on a slip of paper .
先行手法	I wrote something to 紙切れ .
追加手法	I wrote something to of paper .

表 15 翻訳品質が向上しなかった例（5 文）

入力文	そ の ぼや は 失火 と 推定 さ れる 。
参照文	The blaze is thought to have been caused by carelessness with fire .
先行手法	The is estimated at the 失火 broke out .
追加手法	The is estimated at the fire lost broke out .

7.1.4 自動評価

追加実験の出力文に対して自動評価を行った．評価結果を表 16 に示す.

表 16 追加実験の自動評価結果. 精度が高い方を太字で示す

翻訳手法	BLEU	METEOR	RIBES	TER
提案手法	0.1391	0.4107	0.7116	0.7143
先行手法	**0.1434**	**0.4203**	**0.7166**	**0.6978**
先行&提案	0.1431	0.4169	0.7151	0.7100

8 考察

8.1 二段階翻訳の効果

表 5 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 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.

8.2 評価方法の考察

本研究では，未知語の数で評価を行い，提案手法の有効性を確認した．一方で，未知語をローマ字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．今後は，提案手法と未知語をローマ字変換する手法の比較を行い，評価する必要がある.

9 おわりに

本研究では，学習データを追加しない新たな未知語処理の手法として，出現した未知語を抽出し文字単位にした後，文字単位にした未知語を入力として再度翻訳を行うという手法を提案した.

その結果，未知語を大幅に削減することができた．さらに，先行手法と組み合わせることで，より未知語を削減することに成功した．ただし，文の翻訳品質はあまり向上しなかった．今後は，先行手法と提案手法を組み合わせた手法の更なる未知語の調査を検討する.

参考文献

[1] 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.

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

[3] Moses：http://www.statmt.org/moses/.

[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] MeCab：http://mecab.sourceforge.net/.