Vol.2016-NL-228 No.3
2016/9/29

レシピに対する日英機械翻訳の誤り分析

佐藤 貴之[1,2,a)]　原島 純[2,b)]　小町 守[1,c)]

概要：インターネット上で取得可能なレシピが増加するにつれ，レシピの解析や検索，要約，推薦など，レシピを対象とした研究も増加している．一方，レシピを対象として，これまでに研究されていないトピックとして機械翻訳がある．食文化の多様化とともにレシピの需要も国際的に拡大しており，機械翻訳はその需要に応える手段の一つであると考えられる．そこで，本研究ではレシピの機械翻訳に取り組み，その誤りを分析した．具体的には，多くのレシピを抱えるクックパッドの日本語のレシピを，多くの使用者がいる英語に翻訳することを試みた．そして，翻訳結果における誤りを分類・分析し，それらの誤りにどのように対処すべきかを検討した．

1. はじめに

近年，インターネット上で取得可能なレシピが増加している．例えば，日本の料理レシピサービスであるクックパッド[*1]では 245 万品以上のレシピが取得できる（数字は 2016 年 9 月のもの）．同様に，アメリカの料理レシピサービスである Yummly[*2]でも 100 万品以上のレシピが取得できる．

取得可能なレシピが増加するにつれ，これらに関する研究も増加している．これまでに研究されてきたトピックとしては，例えば，レシピの解析 [1] や検索 [2]，要約 [3]，推薦 [4] などがある．レシピは文法が単純であるが，レシピ特有の単語や表現によって，その解析が難しいという問題がある．そのため，専用の辞書構築 [5] や特有のアノテーション [6] なども研究されている．

レシピに関する研究が増加する中，これまでに研究されていないトピックとして機械翻訳がある．食文化の多様化とともにレシピの需要も国際的に拡大している．特に，日本食は健康にも良いことから，日本以外でも需要が大きい．機械翻訳で日本語のレシピを他言語に翻訳することで，多くの人がそれらを利用できるようになると思われる．

そこで，本研究ではレシピ翻訳の現状と課題を確認するため，機械翻訳でレシピを翻訳し，その誤りを分析する．翻訳対象には 16,283 レシピから構成される日英対訳コーパスを使用する．翻訳手法としてフレーズベース統計的機械翻訳 [7] とニューラル機械翻訳 [8] を使用し，日本語のレシピを英語に翻訳する．翻訳誤りは，QTLaunchPad[*3]の Multidimensional Quality Metrics（以下、MQM）[9] を参考に分類する．最後に，分類された誤りを分析し，それらの誤りにどのように対処すべきかを検討する．

[1] 首都大学東京
[2] クックパッド株式会社
[a)] sato-takayuki@ed.tmu.ac.jp
[b)] jun-harashima@cookpad.com
[c)] komachi@tmu.ac.jp

[*1] https://cookpad.com
[*2] https://www.yummly.com

2. レシピ対訳コーパス

前節で述べた通り，本研究では 16,283 レシピから構成される日英対訳コーパスを使用する．このコーパスは，クックパッドが海外向けサービスを開発する過程で構築されたものである．クックパッドのレシピの一例を図 1 に示す．クックパッドのレシピは主にタイトルや材料，手順などのフィールドから構成されている．以下は図 1 のレシピのタイトルの対訳である．

簡単シンプル！ふわふわ卵のオムライス
Easy and Simple Fluffy Omurice

以下は材料の対訳の一例である．材料は名前と分量から構成されている．

ご飯（冷ご飯でも可）
Rice (or cold rice)

2 杯分
2 rice bowl's worth

以下は手順の対訳の一例である．一般的な対訳コーパスと違って，一つの対訳が一つの文とは限らない．この例では一つの対訳が二つの文となっている．

[*3] http://www.qt21.eu/launchpad

図 1 クックパッドのレシピ

ケチャップとソースを混ぜあわせます．味見しながら比率は調節してください．
Mix the ketchup and Japanese Worcestershire-style sauce. Taste and adjust the ratio.

これらの対訳は，初訳と修正という二つの作業を通して収集された．まず，日本語ネイティブ 1 名がレシピを英語に初訳した．ただし，日本語ネイティブは海外在住の日本人や，配偶者が英語ネイティブの日本人であった．次に，英語ネイティブ 2 名が初訳結果を確認して，必要があれば，これを修正した．なお，日本語ネイティブと英語ネイティブはともに料理に精通するものであった．

最終的に構築されたコーパスはタイトル 16,283 文と材料 139,477 文，手順 118,002 個（≠ 文）から構成されている．なお，手順 118,002 個を構成する文の数は日本語側で 209,291 文，英語側で 190,111 文であった．ただし，文の数は，日本語側は句点で，英語側はピリオドで分割することで計数した．

タイトルと材料，手順の長さは大きく異なっている．参考のため，各フィールドの単語の総数を表 1 に示す．なお，日本語の単語数は各フィールドを MeCab（＋ IPADIC）で分割して計数したものである．また，英語の単語数は Moses [10] の添付スクリプトでトークナイズして計数したものである．

表 1 各フィールドの単語の総数

言語	タイトル	材料	手順	全て
日本語	116,827	361,498	2,756,242	3,234,567
英語	101,033	402,039	2,940,816	3,443,888

3. 機械翻訳手法

3.1 フレーズベース統計的機械翻訳

フレーズベース統計的機械翻訳（以下，PBSMT）は対訳コーパスから言語モデルと翻訳モデルを構築する [7]．言語モデルは，翻訳結果が文としてどれだけ自然かを確率的に表すモデルである．翻訳モデルは主に二つの構成要素からなる．以下，フレーズは 1 単語から複数単語で構成される単語列を指す．一つ目は両言語のフレーズの対応を表す単語アライメントである．二つ目はあるフレーズがどのフレーズに翻訳されるかを表す翻訳確率である．これらのモデルをもとに，理論的には原言語の文 f から目的言語の文 e が出力される条件付き確率が最大となるフレーズの組み合わせを選出する．実際には，条件付き確率を直接対数線形モデルによってモデル化し，これを最大化するような e を出力する．

PBSMT は，英語とドイツ語のような語順が似ている言語間の翻訳で高い精度を達成している [7]．一方，日本語と英語のように語順が大きく違う言語間の翻訳ではこの限りでない．これは，組み合わせの探索空間が広くなり，並び替える距離を制限する必要があるためである．また，構文情報を考慮していないため，文法的に誤った訳出が多く見られるという欠点もある．

3.2 ニューラル機械翻訳

ニューラル機械翻訳（以下、NMT）は入力された単語列をベクトルに変換し，これをもとに単語列を出力することで翻訳を行う [11]．一般的に，Encoder と Decoder と呼ばれる二つのリカレントニューラルネットワークから構成される Encoder-Decoder モデルが使用される．前述のベクトルへの変換は Encoder，単語列の出力は Decoder の役割によるものである．また，このモデルを拡張した注意型ネットワークを用いたモデルも提案されている [8]．これは，翻訳時に Encoder のどの隠れ層の情報をどれだけ使用するか（注意度）を動的に決定するモデルである．注意度は確率値で与えられるため，[11] のモデルと比較すると分析がし易い．そのため，本研究では NMT のモデルとして注意型ネットワークモデルを採用した．以降，本稿で NMT と称した場合，注意型ネットワークモデルを指すものとする．

NMT は構文情報を利用していないにもかかわらず，自然な文を生成する．一方，NMT では出力可能な語彙の数を制限する必要がある．これは，一般的な NMT では，出力層においてソフトマックス演算を行なっているためであ

る．ソフトマックス演算には，出力可能な語彙の数に比例して計算量が増加する．ゆえに，多くのフレーズ候補を確保しておける PBSMT と比較すると，NMT では低頻度語の翻訳が難しい [11]．また，原言語側のどの単語にも対応しない単語を出力しやすいという欠点もある [12]．

4. 機械翻訳の誤り体系

本研究では，PBSMT と NMT の訳出に対してブラックボックス分析を行う．ブラックボックス分析とは，訳出の導出過程を考慮せずに出力のみを分析するものである．本研究では翻訳前に必要な過程（単語分割やアライメント獲得）を無視する．ブラックボックス分析に用いる誤り体系は，MQM ANNOTATION DECISION TREE [9] を参考とした．これは誤りを決定木で分類するものである．各誤りは優先度を持っており，より高い優先度を持つ誤りに分類された場合，優先度の低い誤りに分類されるかどうかは考慮しない．それぞれの誤りに対し，Yes/No で答えられるような問いがあり，Yes ならばその誤りに分類される．同様の作業を，最も優先度の低い誤りまで繰り返す．MQM ANNOTATION DECISION TREE を用いることで，一貫性を保って誤りを分類できる．

MQM における誤り体系は妥当性と流暢性の二種類に大別される．妥当性は入力文と翻訳結果の整合性の度合いを測る分析の観点であり，流暢性は翻訳結果の語法や文法の正しさを測る分析の観点である．以下より，妥当性と流暢性の細分類と分類・分析方法について述べる．

4.1 妥当性

MQM における妥当性に関する誤りを以下に示す．

(1) 消失
(2) 未翻訳
(3) 挿入
(4) 術語
(5) 誤翻訳
(6) 妥当性一般

誤翻訳は，異なる意味の単語・フレーズに翻訳している置換誤りと，適切な位置に訳出できてないために意味が異なる位置誤りを含む．以降，フレーズは日本語における一つ以上の文節，英語における句もしくは節を指す．

本研究でも，MQM ANNOTATION DECISION TREE に類似した方法で誤りを分類する．すなわち，出力文に誤りがある場合，上記の順にどの誤りに該当するかを決定する．ある誤りに分類された単語・フレーズは後続する各誤りには該当しないとする．

一方，本研究では，通常の MQM ANNOTATION DECISION TREE にもとづく方法と異なる点が三つある．一つ目に，誤翻訳における上記の置換誤りと位置誤りを別々の誤りとして考える．これは，各翻訳手法で置換誤りと位置誤りの傾向が大きく異なり，その差を反映させるためである．二つ目は，誤りの分類が MQM の決定木の順番でなく，置換誤りと位置誤りを優先誤りとした点である．これは，置換誤りなのか、消失＋挿入なのかという分類を容易にするためである．また，NMT では消失や挿入が多いという点もこの変更の理由の一つである．三つ目は，術語誤りを除いた点である．術語誤りはドメインの差によって起きる語義の選択誤りである．本研究で用いたコーパスは一つの分野に限定したものであり，術語誤りはほとんど見られない．以上を踏まえて，本研究では，以下の優先度の細分類を採用する．

(1) 置換誤り
(2) 位置誤り
(3) 消失
(4) 未翻訳
(5) 挿入
(6) 妥当性一般

誤りを分類する際の具体的な流れは以下の通りである．

(1) 単語・フレーズで対応の取れている箇所を，原言語文の文頭から順に主観によって判定する．（この時，単語・フレーズの位置の正誤は問わない）
(2) 対応の取れた単語・フレーズを正しいものとし，周辺単語に対し，品詞の一致などの情報から置換誤りを決定する．
(3) 置換誤りを決定した後，新しく完成したフレーズがあればそれも含めて，位置誤りに該当するかを決定する．
(4) 置換誤りと位置誤りに分類されなかった単語・フレーズに対して，残りの誤り体系を考える．

本研究では，原言語に日本語を用いているため，文節単位での誤翻訳一つにつき一つの誤りとする．以下より，各誤り体系について例とともに説明する．

4.1.1 置換誤り

原言語文のある単語の意味が，置換の誤りによって目的言語文のある単語において別の意味に変わっている場合の誤りである．以下の例では，‘Heat’ は「割る」の置換誤りとして分類される．‘Heat’ は動詞であるため，「割る」との品詞の一致がとれる．加えて「卵 を」の訳出である ‘an egg’ を目的語としているので，「割る」が ‘Heat’ に翻訳されたものとして扱う．

卵 を 割る .
Heat an egg .

4.1.2 位置誤り

原言語文のフレーズが不適切な位置へ出力することで別の意味に変わっている場合の誤りである．以下の例では ‘from step 1’ が「1 の」の位置誤りとして分類される．誤り数は一つである．

1 の 器 に レタス を 入れる .

Add the lettuce from step 1 into a bowl .

4.1.3 消失

原言語文に存在し，かつ，省略されてはいけない単語の意味が目的言語文で表されていない場合の誤りである．以下の例では，「はちみつ」に対応する単語が消失している．

はちみつ 生地 は 1 次 発酵 まで 済ませる .
Make the dough until the first rising .

4.1.4 未翻訳

原言語文の単語がそのままの形で目的言語文に出現している場合の誤りである．そのままの形で出現している単語一つにつき一つの誤りとする．本研究で用いた NMT は原言語文の単語をそのまま出現するようなモデルではない．よって，この未翻訳誤りは PBSMT における誤り分析でのみの分類となる．以下の例では，原言語文の「狭い」をそのまま出力している．

長 さ を 整え ， 幅 の 狭い ほう で カット する .
Adjust the length , and cut the 狭い into it .

4.1.5 挿入

原言語文に存在しない情報が目的言語文で表されている場合の誤りである．本研究では，英語側に出現した単語を日本語に翻訳し，1 文節につき一つの誤りとする．以下の例では，‘red’ は「赤い」，‘into a pot’ は「鍋に」と翻訳されたとして，二つの誤りとする．

ソース を 加える .
Add the red sauce into a pot .

4.1.6 妥当性一般

上記のどの誤りにも分類が難しい場合，この「妥当性一般」に分類する．誤り個数は原言語文の文節の個数とする．以下の例では四つの誤りとする．

出来上がっ た 時に 倒れ ない ため です .
It will be hard to cover the cake .

4.2 流暢性

MQM における流暢性に関する誤りを以下に示す．

(1) 並べ替え
(2) 語形
(3) 機能語
(4) 文法誤り一般
(5) 理解困難

並べ替え，語形，機能語，文法誤り一般は文法的に不適切な場合に分類される誤りである．理解困難は，文法的には適切だが語義を考慮すると不適切な場合に分類される誤りである．分類方法は妥当性の時に従ったものから，原言語文と対応をとる過程を除いたものになる．文法誤りは文法的にみて誤りを含む単語・句・節に適用され，理解困難は文法的には正しいが意味をとれない箇所に適用される．本研究では，タイトルと材料に対しては，名詞句のみの出力でも誤りとしない．手順は，主語と動詞を含んだものを文として正しいとする．つまり，手順で名詞句のみの出力ならば誤りとする．以下より，各誤り体系について例とともに説明する．

4.2.1 並べ替え

不適切な位置に単語・フレーズが出現している場合の誤りである．複数の誤り候補が考えられる場合には，全体の誤り個数が最小となるような候補に適用する．目的言語側のフレーズ単位で並べ替えが必要とされる際には，そのフレーズに含まれる内容語の数だけ誤り数を加算した．以下の例では，‘Parts of the face’ の場所が不適切であり，正しくは ‘place’ と ‘on’ の間にあるべきである．よって，対象フレーズに含まれる内容語は ‘Parts’ と ‘face’ であり，誤り数は二つとなる．

Parts of the face , place on a baking sheet .

4.2.2 語形

主語との不一致，または時制の不一致の場合の誤りである．動詞の個数だけ誤りを加算する．以下の例では，‘uses’ が不適切であり誤り数は一つである．

I uses the dough for step 4 .

4.2.3 機能語

前置詞，限定詞，助動詞，関係詞の誤用の場合の誤りである．不要な機能語の挿入，必要な機能語の消失，機能語の使用誤りが該当する．以下の例では，不要な ‘to’ が挿入されているため，誤り数は一つである．

It ’s finished to .

4.2.4 文法誤り一般

上記三つの誤りに該当しない場合の誤りである．主に，不要な内容語の挿入や必要な内容語の消失が該当する．以下の例では，動詞が欠落しているため，誤り数は一つである．

The honey dough for the first rising .

4.2.5 理解困難

文法的には正しいが意味が取れない場合の誤りである．文の冒頭にある単語・フレーズは正しいとし，意味が取れなくなる箇所から内容語の数だけその誤りを加算する．以下の例では，‘I was going to be taken’ までは正しく，以後の ‘from the cake’ と ‘in the future’ が誤っているとする．各フレーズに含まれる内容語はそれぞれ ‘cake’ と ‘future’ なので，誤り数は二つである．

I was going to be taken from the cake in the future .

表 2　各フィールドの単語の総数（前処理後）

言語	タイトル	材料	手順	全て
日本語	115,336	322,529	1,830,209	2,268,074
英語	100,796	361,931	1,932,636	2,395,363

5. 実験

5.1 実験データ

実験データには 2 節で述べた対訳コーパスを使用した．2 節で述べた通り，手順には複数の文が含まれることがある．そこで，本研究ではコーパスに対して以下の前処理を行なった．まず，手順において，日本語側を句点ごとに，英語側をピリオドごとに分割した．そして，日本語側を分割して得られる日本語 n 文に対し，英語側を分割して得られる対訳 m 文が一致してない（$n \neq m$）場合，その手順は実験データから除いた．また，タイトルと材料，前処理済みの手順に対し，括弧表現が原言語文と目的言語文のどちらか一方にしか使われていないものは実験データから除いた．

さらに，各テキストを正規化した．まず，日本語側で出現していて，英語側に出現していない特殊記号 (「♡」,「♪」など) を削除した．また，日本語側の全角英数字，全角記号に対して半角のものに変換する処理を行なった．

上記の前処理の結果，タイトルと材料，手順の文数はそれぞれ 16,170 文，131,938 文，124,771 文となった．各フィールドの単語の総数を表 2 に示す．単語の異なり数は日本語側で 23,519, 英語側で 17,307 であった．このうち，レシピ単位で 100 レシピずつランダムにサンプリングしたものをそれぞれ dev セット（1,706 文），test セット（1,647 文）とした．[*4][*5]

誤り分析は，test セットからランダムにサンプリングした 25 レシピ（タイトル 25 文，材料 222 文，手順 195 文）に対して行った．また、前述の 100 レシピに対して，BLEU [13] と RIBES [14] による自動評価も行なった．RIBES の単語適合率に対する重み α は 0.25 とした．また，（出力文長 ÷ 参照訳の長さ）で与えられるへのペナルティ（以下、Brevity Penalty）に対する重み β は 0.10 とした．なお、BLEU は Moses [10] の添付スクリプトを用い，RIBES はバージョン 1.03.1[*6]を用いた．

5.2 手法の設定

PBSMT には最も代表的な PBSMT のツールである Moses（ver2.1.1）[10] を用いた．単語分割には MeCab [15] を用い，辞書は IPADIC（ver2.7.0）とした．単語アライメントは Giza++[*7]により獲得し，単言語コーパスとして対訳コーパスのうち英語側全文を用い言語モデルを学習した．フレーズテーブルサイズは約 300 万対であった．各素性については dev セットで MERT [16] によるチューニングを行い，重みを決定した．

NMT には Bahdanau らの手法 [8] を再実装したものを用いた．ただし，モデルを構成するユニットは Long short-term memory [17] を採用した．NMT の埋め込み層と隠れ層の次元数はともに 512 で，隠れ層は 1 層とした．入出力可能な語彙は制限せず，未知語に対する特定の記号への置換は行なっていない．最適化手法には学習率の初期値を 0.01 とした Adagrad [18] を用いた．また，原言語と目的言語の埋め込み層の初期値は word2vec[*8]のデフォルト設定で学習したものを用いた．原言語の埋め込み層の初期値は対訳コーパスとは別に用意した手順約 1,300 万文から学習した．目的言語の埋め込み層の初期値は対訳コーパスの英語文のうち手順約 12 万文から学習した．[*9][*10]バッチサイズは 64 とした．エポック数は 10 で，各エポックのモデルのうち dev セットで最も高い BLEU を示すモデルを選択した．

6. 結果と考察

6.1 誤り分析

6.1.1 妥当性

各手法の妥当性の誤り数を表 3 に示す．表から，NMT と比較すると，PBSMT は位置誤りが多いことがわかる．一般的に，PBSMT は語順が離れた言語対に対し，並べ替えが困難となり，翻訳精度が落ちる．本研究で用いたコーパスは単語数が少ない文が多数を占める．最も単語数の多い手順のみを考慮しても，平均単語数は日本語が 14.0，英語が 15.0 であった．単語数が少ない場合，並べ替えの最大距離も小さくなるため，PBSMT での翻訳は容易になる．しかし，手順の多くは英語側で命令文となっている．そのため，単語数が比較的短いときでも，長い距離での並べ替えが頻繁に起き，位置誤りが生じたと考えられる．以下の例は PBSMT の翻訳結果である．名詞を複数列挙するような文の一部であり，日本語側と同じ語順で訳出してしまっている．

> 4 の 鍋 に 1 の ブリ & 3 の 大根 & しいたけ & 生姜 を 入れ ,
> Amberjack and daikon radish and shiitake mushrooms , and add the ginger from step 1 to the pan from step 3

[*4] 1 レシピあたり，一つのタイトル，複数の材料，複数の手順から構成される
[*5] 前処理によって，レシピから一部の材料と手順は削除されうる
[*6] http://www.kecl.ntt.co.jp/icl/lirg/ribes/index-j.html
[*7] http://github.com/moses-smt/giza-pp
[*8] https://radimrehurek.com/gensim/models/word2vec.html
[*9] タイトルを学習データから除いたのは，タイトルが自由な文体で書かれているため，学習を妨げると考えたためである．
[*10] 材料を学習データから除いたのは，平均単語数が少ないため，窓幅による文脈を考慮できず，学習を妨げると考えたためである．

表 3　妥当性の誤り個数

手法	置換誤り	位置誤り	消失	未翻訳	挿入	妥当性一般	総数
PBSMT	49	98	139	23	95	43	447
NMT	102	20	176	0	114	119	531

レシピの手順では複数の材料を列挙することが多くあり，このような文が多く見られる．複数の名詞が並ぶことによって，日本語側の動詞「入れ」と英語側の動詞 'add' の並べ替えの距離が大きくなり、このような訳出になったと考えられる．また，「数詞+の」に対応する前置詞句を正しい場所に訳出できていない誤りがある．これは，言語モデルから得られる確率には，数詞を含む前置詞句が訳出候補内で誤った位置に訳出されていても不確かさがないためである．「〜へ」や「〜の」のような他の前置詞句で表現されるものについても，同様の誤りが多く見られた．これらの誤りに対しては，句構造や依存構造を考慮するなどの構文情報を組み込んだ翻訳システムが対応可能であると考えられる．

一方，PBSMT と比較すると，NMT は置換誤りが多いことがわかる．置換誤りには，意味が近い単語を出力している誤りから，品詞のみが一致している単語を出力している誤りまで，様々なものがあった．例えば，前者では，「炒める」に対して 'Heat' を出力する誤りがあった．後者では，「キャベツ」に対して 'sweet potato' を出力する誤りがあった．頻度が少ない単語でも，翻訳候補の揺れが少なければ，PBSMT は正しく翻訳できる傾向がある．以下に例を示す．

入力: クリーム ツイスト
PBSMT: Twisted cream
NMT: Cream cream
（参照訳: Twisted cream bread）

これは，低頻度のフレーズに対して PBSMT が SMT より有効にはたらいた例である．

消失と挿入はどちらの手法にも多くあった．特に消失はどちらの手法においても最も大きい割合を占めた．以下に消失と挿入が PBSMT と NMT の両方で起きている例を示す．

入力: ホーム ベーカリー の 生地 作り コース で 生地を 作る .
PBSMT: Make the dough in the bread maker to make the dough .
NMT: Make the dough using the dough setting .
（参照訳: Use the bread dough function on the bread maker to make the bread dough .）

消失や挿入が起きている文は，誤り箇所は異なるものの PBSMT と NMT で同じ文である傾向があった．PBSMT では「生地作り コースで」の消失と 'to make' や'the dough' の挿入が見られる．一方，NMT では「ホーム ベーカリーの」の消失が見られる．このように，それぞれの手法で翻訳困難な文はある程度共通しているのではないかと考えられる．

また，挿入誤りは，日本語側で目的語が省略されている文に見られた．レシピの手順の日本語側では，同一レシピ内で一度出現した単語を省略することがある．そのような文の訳出で，省略された目的語の位置に何かしらの単語が挿入されることがあった．以下に例を示す．

入力: 紙 に 包ん で
NMT: Wrap the cake in the cake paper
（参照訳: Wrap the cakes in parchment paper）

この例では，'the cake' にあたる原言語文の単語は存在しないが，このフレーズが訳出されている．これは訓練時の参照訳の省略度合いによるものだと考えられる．この例でも，参照訳は 'the cakes' を補完している．しかし，文によっては，他動詞でありながら補完していないものも多くある．従って，省略するか補完するかは外部から何かしらの形で情報を与える必要があると考えられる．

最後に，未翻訳は PBSMT でのみ考慮する誤りであるが，その割合は最も少ないことがわかる．今回用いたコーパスは語彙数が小さいため，訓練時に出現する語彙が test セットのほとんどの語彙を含んだ．従って，test セットに含まれる未知語の割合がわずかで，このような結果になったと考えられる．

6.1.2 流暢性

各手法の流暢性の誤り数を表 4 に示す．並べ替えの誤りは，妥当性の位置誤りの時と同様の原因で起きていると考えられる．ただし，妥当性での位置誤りと違って，流暢性における並べ替え誤りは日本語側の意味を考慮しない．よって，妥当性の位置誤りで誤りに分類されたものも流暢性の並べ替え誤りには該当しないため，誤り数は少なくなる．以下の例は，妥当性の位置誤りの例として示したものだが，誤りとなるのは 'add' のみとなる．

Amberjack and daikon radish and shiitake mushrooms , and add the ginger from step 1 to the pan from step 3

機能語の誤りは PBSMT で多く見られた．主な誤りは不要な前置詞の挿入であった．これは，フレーズ抽出で得られた前置詞について，適切な挿入場所が存在しなかったためであると考えられる．つまり，フレーズ抽出の時点で

表 4　流暢性の誤り個数

手法	並べ替え	語形	機能語	文法一般	理解困難	総数
PBSMT	18	2	24	73	12	129
NMT	4	1	6	17	55	83

表 5　自動評価の結果

評価尺度	手法	タイトル	材料	手順	全文
BLEU	PBSMT	**22.15**	**56.10**	25.37	**28.09**
	NMT	19.68	55.75	**25.68**	28.01
RIBES	PBSMT	**61.85**	**90.03**	74.98	81.72
	NMT	61.49	89.70	**77.84**	**82.79**

誤っていたと考えられる．以下の例では，‘in’ が不要であるとした．

> Remove the sinew from the chicken tenders and fold in lightly .

文法一般についての主な誤りは基本的に内容語の誤りであった．特に，動詞や名詞の消失や挿入が多くの出力文で見られた．これも，機能語での誤りと同じ理由で起きていると考えられる．以下の例は動詞が消失したものである．

> Basic chiffon cake milk to make the dough .

NMT には理解困難な文が非常に多く見られた．並べ替えや機能語，文法一般などの文法的な誤りはなくても，同じ単語・フレーズの繰り返しや，ある動詞に対して意味的整合性の取れない目的語が見られた．以下の例では，‘and open the pot’ が繰り返されている．

> leave to steam for about 2 hours , and open the pot , and open the pot

語形については，各手法においてほとんど誤りが見られなかった．タイトルや材料は名詞句であり，手順の多くは命令文で表される．命令文における接続詞節での時制は現在形で表される．その時の主語はほとんどが材料を指す名詞であり，三人称単数である．ゆえに，時制の不一致や，主語と動詞の不一致が起きなかったと考えられる．

6.2　自動評価

test セットにおける BLEU と RIBES での評価結果を表 5 に示す．まず，タイトルについて議論する．タイトルには自由な語彙や意訳が多く見られる．言い換えれば，比較的低頻度な形態で書かれている．また，タイトルが占める割合は表 2 からわかるように非常に小さい．以上から，タイトルの翻訳は材料や手順の翻訳より困難であった．表 5 のタイトルの項目を見ると，PBSMT が NMT に対して BLEU でも RIBES でも良い結果を示している．PBSMT は数単語からなる単語列をフレーズとして翻訳するため，自由な語彙や意訳で記述されているタイトルでも，部分的に正しく翻訳できる．一方，NMT にこのようなテキストを入力すると，原言語文のどの単語も訳せてなかったり，極端に短い出力となってしまい，BLEU が低くなってしまった．

次に，材料の評価結果について議論する．材料は 3 単語程度と短い文であり，かつ，単語ごとに翻訳候補が少ない．そのため，PBSMT と NMT ともに非常に高い結果が得られた．このように，辞書引きのような翻訳が要求される文には PBSMT が優位であると考えられる．そのため，わずかではあるが，どちらの評価尺度においても PBSMT が上回る結果となった．

手順では，4.1 節の一つめの事例のような複数の名詞を列挙する文が見られる．そして，目的言語文の文体は命令文であることが多く，並べ替えの距離が大きくなってしまう．このような場合，NMT の方が誤りが少ない．また、原言語文において省略が起き，目的言語文でその補完をしなければならない場合がある．PBSMT も NMT も，どの単語を補完するべきかという情報を明示的に与えていない以上，正しい単語を訳出するのは難しい．ただし、NMT では，何かしらの単語で補完する傾向が見られた．

最後に，RIBES について補足する．RIBES は NMT に有利な尺度となっている可能性がある．RIBES は，単語適合率に対する重み α と Brevity Penalty に対する重み β をハイパーパラメータとして決定する．一方で，BLEU は Brevity Penalty のみ考慮し，かつ，重みは決定しない．NMT では，参照訳に対して短い文を訳出することが多くあるが，この β によってその問題が無視されうる．語順は正しいことが多いため，β が低い際には高いスコアが出やすい．PBSMT は原言語文の単語・フレーズをもとに，目的言語側とのフレーズ対応を獲得し，それを並べ替えることで訳出する．そのため，NMT ほど極端に短い文を訳出することはほとんどない．しかし，並べ替える候補が増えるほど，正しい語順にして訳出するのは難しくなり，RIBES の高いスコアを得るのは難しくなる．以上より，RIBES はハイパーパラメータ次第で NMT に有利となっている可能性がある．

7.　関連研究

7.1　レシピ言語処理

インターネット上で取得可能なレシピが増加するにつれ，これらに関する研究も増加している．以下は，レシピの解析にフォーカスを当てた研究である．Kiddon らは，調理行動をノード，それらの関係をエッジとするグラフでレシピを表現する手法を提案している [19]．一方，Jermsurawong

らは，材料を終端のノード，調理行動を内部のノードとする木構造でレシピを表現している [20]．Maeta らは，材料や調理器具，調理行動をノード, それらの関係をエッジとするグラフでレシピを表現している．Nanba らはレシピ解析に利用するため，料理用語に関するオントロジーを構築している [5]．これらの研究は基礎解析にフォーカスを当てたもので，応用システムにフォーカスを当てた本研究とは異なる．

一方，応用システムにフォーカスを当てた研究としては以下のものがある．Yasukawa らは第 11 回 NTCIR ワークショップでレシピ検索のタスクを開催しており, タスクには国内の四つの研究グループが参加している [2]．Yamakata らは，複数のレシピに共通するグラフ構造を検出することで，レシピを要約する手法を提案している [3]．Forbes らは，レシピの推薦における Matrix Factorization の有効性を検証している [4]．Wang らは，中国語のレシピに対して，類似するレシピを検索する手法を提案している [21]．これらの研究はレシピの検索や要約，推薦にフォーカスを当てたもので，レシピの翻訳にフォーカスを当てた本研究とは異なる．

7.2 機械翻訳

機械翻訳の誤りを分析した研究として Vilar ら [22] や星野ら [23]，赤部ら [24] のものがある．

Vilar らは独自の機械翻訳システムを対象とし，英西翻訳と中英翻訳における誤りを分析した．前者では議会の内容を英語からスペイン語に，後者ではニュースの内容を中国語から英語に翻訳している．誤り体系は大分類と小分類に分かれている．具体的には，大分類として単語の消失や挿入，並べ替えなどがあり，小分類として各大分類を細分化したものがある．例えば，単語の消失の小分類として内容語の消失と機能語の消失がある．誤り分析は誤り体系に従ったブラックボックス分析で，自動評価には BLEU と WER [25]，PER [26]，NIST [27] を用いている．

星野らは Google 翻訳[*11]や Bing Translator[*12]のようなインターネット上の機械翻訳システムを対象とし，英日翻訳における誤りを分析した．星野らが用いたデータは新聞記事を英語から日本語に翻訳したものである．Vilar らの誤り体系にもとづいてブラックボックス分析を行なっている．自動評価には BLEU と WER を用いている．また，誤りの分類だけでなく，人手による 5 段階評価を行なっている．

赤部らは六つの機械翻訳システムに対し，日英翻訳における誤りを分析した．三つはオープンソースソフトで残りの三つは商用システムであった．オープンソースソフトは，PBSMT と原言語の構文情報を考慮した翻訳システム，原言語と目的言語の両方の構文情報を考慮した翻訳システムであった．商用システムは一つがルールベース機械翻訳システムで，残りの二つが PBSMT であった．訓練データとして，様々なドメインのコーパスを用いており，評価セットとして書き言葉の均衡コーパスを用いている．赤部らも Vilar らの誤り体系にもとづいてブラックボックス分析を行なっている．自動評価は行なっていないが，訳出の導出過程を分析対象としたグラスボックス分析を行なっている．

本研究は，これまでに分析されていないドメインであるレシピを対象とした機械翻訳の誤りを分析している．また，誤り体系は MQM ANNOTATION DECISION TREE の誤り体系にもとづいている。加えて，NMT の訳出を誤り分析対象としている．このように，本研究はドメインと誤り体系，翻訳手法で先行研究と異なる．

8. おわりに

本研究では，レシピに対する日英機械翻訳の誤り分析を行なった．翻訳手法には PBSMT と NMT を用い，誤り体系には MQM ANNOTATION DECISION TREE を拡張したものを用いた．誤りを分類したところ，各誤りの傾向はそれぞれの手法において大きく異なることがわかった．PBSMT は NMT と比較すると文法的な誤りが多かった．一方で，NMT は PBSMT より置換誤りが多く，別の語義の単語を出力する傾向が見られた．また，NMT は文法的に正しいが，意味がとれない訳出も多かった．そして，どちらの手法でも消失と挿入が多く見られた．

レシピを構成する 3 種類の文では，それぞれにおいて特徴が見られた．タイトルは分量が少ない割に語彙が多かったため，学習が難しかった．そのため、NMT によるタイトルの訳出には，原言語文をほとんど訳せていない，訳出が短いなどの問題が起きていた．一方，PBSMT では，タイトル全体を訳せていなくても，フレーズ対によって部分的に訳せた．材料はタイトルや手順と比較すると非常に平易な文体であり，どちらの手法でも高い精度が得られた．これは，どちらの手法でも辞書引きのような翻訳が可能であることを示している．最後に，手順では PBSMT と NMT で異なった誤りの傾向が見られた．PBMT は NMT と比較すると，多くの位置誤りが見られた．手順では複数の名詞を列挙した後，動詞が続く文がある．かつ，手順では目的言語文の多くが命令文で書かれている．そのため，原言語文の動詞と目的言語文の動詞で並べ替えの距離が大きくなり，位置誤りを起こしたと考えられる．NMT は文法的に正しい文を生成するため，PBSMT に比べ位置誤りは少ない．RIBES においても，NMT が約 1 ポイント上回る結果となった．

今後は，グラスボックス分析を行ない，システムの導出過程からどのような現象が誤りにつながるのかを把握する．次に，手順における目的語省略問題に対し，文をまた

[*11] https://translate.google.co.jp
[*12] https://www.bing.com/translator

いだ情報を明示的に与えることを試みる．

参考文献

[1] Maeta, H., Sasada, T. and Mori, S.: A Framework for Procedural Text Understanding, *Proceedings of the 14th International Conference on Parsing Technologies (IWPT 2015)*, pp. 50–60 (2015).

[2] Yasukawa, M., Diaz, F., Druck, G. and Tsukada, N.: Overview of the NTCIR-11 Cooking Recipe Search Task, *Proceedings of the 11th NTCIR Conference (NTCIR-11)*, pp. 483–496 (2014).

[3] Yamakata, Y., Imahori, S., Sugiyama, Y., Mori, S. and Tanaka, K.: Feature Extraction and Summarization of Recipes using Flow Graph, *Proceedings of the 5th International Conference on Social Informatics (SocInfo 2013)*, pp. 241–254 (2013).

[4] Forbes, P. and Zhu, M.: Content-boosted Matrix Factorization for Recommender Systems: Experiments with Recipe Recommendation, *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys 2011)*, pp. 261–264 (2011).

[5] Nanba, H., Doi, Y., Tsujita, M., Takezawa, T. and Sumiya, K.: Construction of a Cooking Ontology from Cooking Recipes and Patents, *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication (UbiComp 2014 Adjunct)*, pp. 507–516 (2014).

[6] Mori, S., Maeta, H., Yamakata, Y. and Sasada, T.: Flow Graph Corpus from Recipe Texts, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 2370–2377 (2014).

[7] Koehn, P., Och, F. J. and Maruc, D.: Statistical Phrase-based Translation, *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2003)*, pp. 48–54 (2003).

[8] Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *5th International Conference on Learning Representations (ICLR 2015)* (2015).

[9] Burchardt, A. and Lommel, A.: Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality, Technical report, QTLaunchPad (2014).

[10] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C. and Zens, R.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180 (2007).

[11] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, *In Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pp. 3104–3112 (2014).

[12] Tu, Z., Lu, Z., Liu, Y., Liu, X. and Li, H.: Modeling Coverage for Neural Machine Translation, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 177–180 (2016).

[13] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40st Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pp. 138–145 (2002).

[14] Isozaki, H., Hirao, T., Duh, K., Sudoh, K. and Tsukada, H.: Automatic Evaluation of Translation Quality for Distant Language Pairs, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pp. 944–952 (2010).

[15] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 230–237 (2004).

[16] Och, F. J.: Minimum Error Rate Training in Statistical Machine Translation, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pp. 160–167 (2003).

[17] Hochreiter, S. and Schmidhuber, J.: LONG SHORT-TERM MEMORY, *Neural Computation 9*, pp. 1735–1780 (1997).

[18] Duchi, J., Hazan, E. and Singer, Y.: Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *Journal of Machine Learning Research 12*, pp. 2121–2159 (2011).

[19] Kiddon, C., Ponnuraj, G. T., Zettlemoyer, L. and Choi, Y.: Mise en Place: Unsupervised Interpretation of Instructional Recipes, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pp. 982–992 (2015).

[20] Jermsurawong, J. and Habash, N.: Predicting the Structure of Cooking Recipes, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pp. 781–786 (2015).

[21] Wang, L., Li, Q., Li, N., Dong, G. and Yang, Y.: Substructure Similarity Measurement in Chinese Recipes, *Proceedings of the 17th International World Wide Web Conference (WWW 2008)*, pp. 979–988 (2008).

[22] Vilar, D., Xu, J., Luis FernandoD' Haro, Ney, H.：Error Analysis of Machine Translation Output, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 79–84 (2006).

[23] 星野翔，建石由佳：インターネット上の英日統計的機械翻訳サービスの誤り分析，情報処理学会研究報告 (2011-NL-201)，pp. 1–6 (2011).

[24] 赤部晃一，Neubig, G.，　工藤拓，Richardson, J.，中澤敏明，　星野翔：Project Next における機械翻訳の誤り分析，言語処理学会第 21 回年次大会ワークショップ (2015) (2015).

[25] SonjaNie β en，Josef Och, F.，Leusch, G.，Ney, H.：An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)* (2000).

[26] Franz, O. J., Ueffing, N. and Ney, H.: An Efficient A* Search Algorithm for Statistical Machine Translation, *Proceedings of the ACL-2001 Workshop onData-Driven Machine Translation*, pp. 55–62 (2001).

[27] Lin, C.-Y. and Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics, *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, pp. 71–78 (2003).