修士論文

琉日機械翻訳のための対訳コーパスの自動拡張について

久高優也

主指導教員 白井清昭

北陸先端科学技術大学院大学
先端科学技術研究科
（情報科学）

令和２年３月

Abstract

Several studies on machine translation aim at translation between languages used in the same country, such as a dialog and a standard language of a country. One of them is machine translation between Ryukyu dialect (Okinawa dialect) and Japanese standard language (Japanese). The linguistic features of Ryukyu dialect and Japanese are quite different. Therefore, Ryukyu-Japanese machine translation is required for people who are non-native speakers of Ryukyu dialect to easily understand it.

In recent years, most methods of machine translation are based on statistical machine translation or neural machine translation, which learn translation models from a large amount of a bilingual corpus. The performance of these machine translation methods heavily depends on the amount of a training bilingual corpus. Therefore, the performance of them is not good for low-resource languages such as the Ryukyu dialect.

A method has been proposed to increase the amount of a bilingual corpus by automatically generating translation pairs to improve the performance of machine translation for low-resource languages. An attempt has also been made to apply this method to Ryukyu-Japanese Statistical Machine Translation (Ryukyu-Japanese SMT). However, the previous study didn't consider the quality of automatically generated translation pairs, so the performance of machine translation was not so good. Therefore, this thesis aims at improving performance of Ryukyu-Japanese SMT by expanding a bilingual corpus considering the quality, variety and amount of the translation pairs. Consideration of the quality means to generate natural translation pairs. Consideration of the variety means to construct an expanded bilingual corpus so that it contains not only similar sentences but a wide variety of sentences. Consideration of the amount means to avoid excessive expansion of a bilingual corpus. When too many sentences are added to an expanded bilingual corpus, many unnatural translation pairs are likely to be added too. In this thesis, we implement the proposed method and conduct experiments to translate Ryukyu dialect to Japanese to confirm that the above three ideas can contribute to improve the performance of Ryukyu-Japanese SMT.

Our proposed method to expand a bilingual corpus consists of two steps: generation of translation pair candidates and selection of translation pair candidates.

In the generation of translation pair candidates, new translation pairs are generated to enlarge an initial (small) bilingual corpus. For a given translation pair in an initial bilingual corpus, if a word pair compiled in a Ryukyu-Japanese bilingual lexicon is included in the sentences of both source and target languages, new translation pairs are generated by replacing those words with other words, whose

parts of speech are the same, in the bilingual lexicon.

In the selection of translation pair candidates, appropriate translation pairs are chosen among candidates from a point of views of their quality, variety and amount, then they are added to the initial bilingual corpus to make an expanded dialog corpus. To consider the quality of translation pairs, a score to evaluate fluency of a Japanese sentence in a translation pair is calculated, then translation pairs with high scores are selected. We propose two kinds of the scores: one is the generation probability of the sentence given by the probabilistic language model, the other is the difference of the generation probability between the derived (newly generated) sentence and its original sentence. To consider the variety, the translation pairs are selected so that the same number of translation pairs are generated from all the sentences in the initial bilingual corpus. In this way, words and contexts included in the initial bilingual corpus can be uniformly transferred in the expanded bilingual corpus. To consider the amount of translation pairs, the number of newly generated translation pairs is controlled. It enables us to prevent unnatural translation pairs from being excessively added to the bilingual corpus.

In the experiments, the following methods were evaluated and compared: no expansion (using only the initial bilingual corpus), the method of previous study, random selection, our method considering the quality (with two types of scores), our method considering the quality and variety, and our method combining quality-based selection and random selection. We trained an SMT model using the expanded bilingual corpus constructed by each proposed method or the baseline, translated test sentences of Ryukyu dialect into Japanese, and evaluated their performance using BLEU and RIBES. By considering both the quality and variety, BLEU improved up to 1.24 points and RIBES improved up to 2.54 points comparing with the random selection. By considering the variety, BLEU improved from 7 to 10 points and RIBES improved from 5 to 8 points comparing with the method considering the quality only. In addition, we examined the changes in BLEU and RIBES when the amount of the expanded bilingual corpus was changed. It was found that BLEU and RIBES decreased when more sentences were added to the expanded bilingual corpus. In addition, our method considering both the quality and variety outperformed no expansion method only when the number of expanded translation pairs was 2,000.

From the above results, it was found that the translation performance was improved by expanding the bilingual corpus considering both the quality and variety. It was not necessary to add many expanded translation pairs, but it was important to optimize the amount of the expanded bilingual corpus appropriately. Especially, the method that considers only the quality but not the variety achieved poorer performance than other methods. This may be because sentences with high prob-

ability of the probabilistic language model tended to be short, and the number of words in the expanded bilingual corpus became small. Therefore, the obtained expanded bilingual corpus might not contain sufficient words and contexts. We can conclude that the basic idea of the proposed method to keep the variety of the expanded bilingual corpus is effective.

BLEU and RIBES of the proposed methods were improved comparing to the baseline, but the difference was small. Furthermore, when the amount of translation pairs in the expanded bilingual corpus was increased, BLEU and RIBES decreased. It may be caused by a naive method to generate translation pair candidates, where they are generated by randomly replacing the words. Most of generated translation pairs are unnatural, and only the small number of natural translation pairs are expanded. In the future, instead of using the naive method by word replacement, we will explore a method to generate translation pair candidates by paraphrase sentences with sophisticated natural language techniques as a better method of expansion of the bilingual corpus.

概要

機械翻訳に関する研究には，方言から標準語など，同じ国で使用される言語間の翻訳を対象とした研究がある．琉球方言（沖縄方言）と日本語標準語の間の機械翻訳もその一つである．琉球方言の言語的特徴は，日本語標準語とはかなり異なる．そのため，琉球方言に馴染みのない人がそれを手軽に理解するために，琉球方言と標準語の機械翻訳が求められている．

近年の機械翻訳の研究は，大量の対訳コーパスから翻訳モデルを学習する統計的機械翻訳やニューラル機械翻訳が主流になっている．これらの機械翻訳方式の翻訳精度は学習に用いる対訳コーパスの量に大きく依存する．したがって，琉球方言などの低言語資源の言語をこれらの方式で機械翻訳する場合，翻訳の性能が低くなることが知られている．

低言語資源の言語を対象とした機械翻訳の性能を向上させる研究に，対訳文の自動生成により対訳コーパスの量を増やす手法が提案されている．また，この手法を琉日統計的機械翻訳に適用した研究もある．しかし，その先行研究は自動生成した文の品質を考慮しておらず，評価実験においても琉日機械翻訳の精度はそれほど良くなかった．したがって，本研究では，品質，多様性，量を考慮して対訳コーパスを拡張することで，琉日統計的機械翻訳の精度を向上させることを目指す．品質についての考慮とは，対訳文を生成するときにできるだけ自然な文を生成することを指す．多様性についての考慮とは，拡張後の対訳コーパスが似たような文だけで構成されることなく，様々な文を含むようにすることを指す．量についての考慮とは，あまりに多くの対訳文を拡張対訳コーパスに含めてしまうと，不自然な文が多く含まれる可能性が高くなるため，対訳コーパスを過度に拡張しないための工夫を指す．提案手法を実装し，琉球方言の文を日本語標準語に機械翻訳する実験を行い，上記の 3 つの工夫によって琉日機械翻訳の性能が向上することを確認する．

本研究で提案する対訳コーパス拡張手法は，対訳文候補生成処理と対訳文候補選択処理の 2 つのステップからなる．

対訳文候補生成処理は，対訳コーパスを拡大するために，新しい対訳文を生成する処理である．初期の対訳コーパスにおいて，琉日対訳辞書に登録されている単語が原言語文と目標言語文の両方に出現するときに，それらの単語を，対訳辞書に登録されていて，かつ品詞が同じである別の単語に置き換えることで，新しい対訳文の候補を生成する．

対訳文候補選択処理は，品質，多様性，量を考慮して対訳文候補の中から適切なものを選択し，拡張対訳コーパスを作成する処理である．まず，対訳文の品質を考慮するために，対訳文における標準語文の自然さを評価するスコアを計算し，これが高い対訳文を選択する．スコアの計算方法として，確率言語モデルによる文の生成確率をスコアとする手法と，自動生成する前の元の文と自動生成された後の文の生成確率の差をスコアとする手法を提案する．次に，対訳文の多様性を

考慮するために，結果的に初期の対訳コーパスの全ての文から同じ数の対訳文が生成されるように対訳文を選択する．これにより，初期の対訳コーパスに含まれる語彙や文脈を偏りなく拡張対訳コーパスへ含めることができる．最後に，量を考慮した対訳文の選択として，不自然な対訳文が過度に対訳コーパスに含まれるのを防ぐために，拡張対訳コーパスの文の数を調整する．

実験では，拡張なし（初期の対訳コーパスのみを用いる手法），先行研究の拡張手法，ランダム選択，品質のみを考慮した拡張手法（2 種類のスコアによる），品質と多様性を考慮した拡張手法，品質を考慮した選択とランダム選択を組み合わせた拡張手法を評価した．各提案手法で構築した拡張対訳コーパスを用いて統計的機械翻訳モデルを学習し，琉球方言テスト文を標準語に翻訳し，その正確性を BLEU と RIBES を指標として評価した．その結果，品質と多様性の両方を考慮することによって，ランダム選択と比較して BLEU が最大 1.24 ポイント，RIBES が最大 2.54 ポイント向上した．多様性を考慮することで，多様性を考慮せずに品質のみを考慮した手法よりも BLEU が 7〜10 ポイント，RIBES が 5〜8 ポイント向上した．さらに，拡張対訳コーパスの量を変化させ，それによる BLEU と RIBES の変化を調べた．文の数が多いほど BLEU もしくは RIBES が低下した．拡張なしの手法より評価指標が高くなったのは文の数が 2,000 文のときだけであった．

これらの結果から，品質と多様性の両方を考慮して対訳コーパスの拡張を行うことで翻訳精度が向上することがわかった．また，拡張する文の量はただ多ければよいものではなく，適切な量に調整することが重要であることがわかった．特に，多様性を考慮せずに品質のみを考慮した拡張手法は，他の手法と比べて評価指標が低かった．これは，確率言語モデルの確率が高い文は短い文である傾向があるため，拡張対訳コーパスの単語数が少なくなり，十分な語彙や文脈を含む対訳コーパスが得られなかったためと考えられる．このことから，拡張対訳コーパスの多様性を確保する提案手法のアプローチは有効であることが確認された．

本研究の提案手法は，ベースラインと比べて BLEU や RIBES が改善したが，その差は小さかった．また，拡張対訳コーパスの文の量を変化させたときに，文の数を多くしていくほど BLEU や RIBES が低下していった．これらの原因として，単語のランダムな置換により対訳文候補を生成したことで，不自然な対訳文が多く生成され，自然な対訳文が拡張対訳コーパスにあまり多く追加されなかったためであると考えられる．今後は，単語置換によるナイーブな手法ではなく，言い換え技術などを用いて元の文を自然な文に置き換えることで対訳文候補を生成し，対訳コーパスを拡張する手法を探究したい．

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

機械翻訳とは，ある言語で書かれた文に対して，その文が表す意味を保持したまま別の言語の文へと変換する翻訳と呼ばれる処理をコンピュータで実現する技術である．機械翻訳に関する研究として，日本語から英語など，異なる国の言語間での翻訳だけでなく，方言から標準語など，同じ国で使用されている言語間での翻訳についても研究が行われている．琉球方言（沖縄方言）は，同じ日本で使用されている言語であるが，母音の数，文法，アクセントなど，日本語標準語とは言語的特徴がかなり異なる．そのため，琉球方言にあまり馴染みのない人，特に本土出身の人にとっては，琉球方言を理解することは難しい．したがって，琉球方言に馴染みのない人がこれを手軽に理解するためには，琉球方言と標準語の機械翻訳が求められる．

ここで，本研究で対象としている琉球方言の詳細を説明する．沖縄には多くの集落が存在し，それぞれの集落ごとに独自の言葉が育まれていた．そのため，琉球方言は，図 1.1 に示すように様々な種類に分類されている [12]．これらはお互いに語彙，発音，アクセントなどに違いがある．特に，北グループ（奄美沖縄方言群）と南グループ（宮古八重山方言群）では話が通じないほど大きく異なっている．しかし，その中でも沖縄中南部方言は，琉球王国の文芸や芸能の中心であった首里（那覇）の方言であり，琉球列島全体でも比較的通じやすい言語である [11]．そのため，本研究では沖縄中南部方言を研究の対象とする．本論文では，これ以降,「琉球方言」は沖縄中南部方言を指すものとする．

機械翻訳方式には多様な種類があり，ルールベース機械翻訳，中間言語方式，トランスファー方式，用例に基づく機械翻訳などが存在する．近年では，大量の対訳コーパスから翻訳モデルを学習する統計的機械翻訳（SMT:Statistical Machine Translation）やニューラル機械翻訳が主流になっている．しかし，これらの翻訳方式の翻訳精度は対訳コーパスの量に大きく依存する．したがって，低言語資源の言語をこれらの方式で機械翻訳すると，翻訳の性能が低くなることが知られている．「低言語資源の言語」とは，ここでは，使用者の少ない言語や方言など，対訳コーパスの量を十分に確保できない言語を指す．そのため，低言語資源の言語を対象に機械翻訳の性能を向上させる研究も行われている．

低言語資源の言語を対象とした機械翻訳に関する研究として，対訳文を自動的に生成することで対訳コーパスの量を拡充する手法が提案されている．また，こ

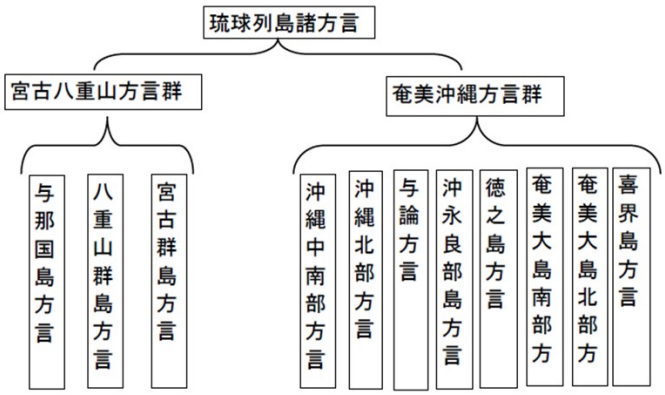

図 1.1: 琉球方言の分類

の手法を琉日統計的機械翻訳に応用した研究例も存在する．しかし，その先行研究では，自動生成した対訳文候補の文の品質については考慮していないため，不自然な対訳文が対訳コーパスに含まれる可能性が高いという問題があった．論文で報告されている実験でも，琉日機械翻訳の精度は決して高いとは言えなかった．

1.2 目的

本研究は，低言語資源の言語である琉球方言（沖縄方言）を機械翻訳の対象とし，琉球方言から日本語への統計的機械翻訳の精度を向上させる手法を提案する．具体的には，少量の琉日対訳コーパスから，新しい対訳文を自動生成し，対訳コーパスの量を増やした上で，SMT のモデルを学習する．対訳コーパスを拡張する際，自動生成した対訳文の (1) 品質，(2) 多様性，(3) 量を考慮して，対訳コーパスに追加する対訳文を選別する．品質についての考慮とは，対訳文を生成する際，できるだけ自然な文を生成することを指す．多様性についての考慮とは，拡張後の対訳コーパスが同じような文だけで構成されることなく，様々な文を含むようにすることを指す．量についての考慮とは，あまりに多くの対訳文を生成すると不適切なものが含まれる可能性が高くなるため，対訳コーパスを過度に拡張しないための工夫を指す．提案手法を実装し，琉球方言を日本語標準語に機械翻訳する実験を行い，上記の 3 つの工夫が機械翻訳の性能向上にどれだけ寄与するかを確認する．

1.3 論文の構成

本論文の構成は以下の通りである．2 章では，本研究の関連研究である統計的機械翻訳の説明と，低言語資源の言語を対象とした機械翻訳の研究，自動対訳コーパス生成手法の研究を紹介する．3 章では，本研究の提案手法である対訳コーパス

拡張に基づく琉日統計的機械翻訳について述べる．4章では，提案手法の評価実験について述べる．最後に5章では，本論文のまとめと今後の課題について述べる．

第2章　関連研究

本章では本研究の関連研究について述べる．本研究は統計的機械翻訳に関する研究であるため，2.1 節では統計的機械翻訳を紹介する．2.2 節では，低言語資源の言語を対象とした機械翻訳の研究例を紹介する．2.3 節では，対訳コーパスの自動拡張手法を用いた機械翻訳の研究を紹介する．最後に，2.4 節では，2.3 節で紹介した先行研究との比較も交えつつ，本研究の特色について述べる．

2.1　統計的機械翻訳

統計的機械翻訳（SMT）とは，大量の対訳文を集積した対訳コーパスを用いて翻訳モデルを統計的に学習し，翻訳を行う手法である．一般的な SMT の概略図を図 2.1 に示す．この図は，英語から日本語への SMT を例に，入力文である「He loved me.」を出力文「彼はあなたを好きだった。」に翻訳するときの処理の流れを示している．まず，対訳文の組を大量に集めた対訳コーパスを用意する．この例では英語文とその日本語訳の組を集める．次に，得られた対訳コーパスから，ある英語文が別の日本語文に翻訳される確率を与えるモデルを学習する．「He loved me.」という英語文が与えられたとき，翻訳後の確率が最大となるような日本語文（この場合は「彼はあなたを好きだった。」）を生成し，翻訳結果として出力する．図中の「デコーダ」とは，翻訳確率が最大となるような出力文を効率的に探索するモジュールである．

SMT は一般的に式 (2.1) によって定式化される．この式は，翻訳元言語文（原言語の文）f が与えられたとき，翻訳先言語文（目標言語の文）e の候補の中から，翻訳確率が最大になる目標言語文 $\hat{e}$ を選択することを示している．ここで $P(e|f)$ は，f が e に翻訳される翻訳確率である．これは，翻訳モデル確率 $P(f|e)$ と言語モデル確率 $P(e)$ の積で計算される．

$$\hat{e} = \mathrm{argmax}_e P(e|f) = \mathrm{argmax}_e P(f|e)P(e) \tag{2.1}$$

言語モデル $P(e)$ は，翻訳された目標言語の文 e が文としてどれだけ自然であるかを評価する確率モデルである．良い言語モデルは，流暢性の高い文（自然な文）に高い確率を与える．一般に，言語モデルは N-gram によって計算されることが多い．N-gram モデルは式 (2.2) のように定義される．

図 2.1: 一般的な SMT の概略図

$$P(e) = \prod_{i=1}^{n} P(w_i | w_{i-N+1}, w_{i-N+2}, \cdots, w_{i-1}) \tag{2.2}$$

w_i は文 e における i 番目の単語を表す．文 e の生成確率は単語 w_i の生成確率の積として計算される．単語 w_i の生成確率は，その直前に出現した $N-1$ 個の単語列 $(w_{i-N+1}, w_{i-N+2}, \cdots, w_{i-1})$ の次に w_i が出現する条件付き確率として計算される．式 (2.3) に「みな 様 はじめまして」という 3 単語の文に対する 2-gram モデル (N=2) の計算例を示す．</s> は文末記号を意味する．

$$\begin{aligned} P(\text{みな 様 はじめまして}) = {} & P(w_1 = \text{“みな”}) \\ & * P(w_2 = \text{“様”} | w_1 = \text{“みな”}) \\ & * P(w_3 = \text{“はじめまして”} | w_2 = \text{“様”}) \\ & * P(w_4 = \text{“</s>”} | w_3 = \text{“はじめまして”}) \end{aligned} \tag{2.3}$$

翻訳モデルは，目標言語の文 e が原言語の文 f の意味をどれだけ保持しているかを評価する確率モデルである．基本的に，f 中の多くの単語が e の中で正しい単語に翻訳されているとき，この確率は高くなる．翻訳モデルの確率 $P(f|e)$ は式 (2.4) により求められる．

$$P(f|e) = \sum_{a} P(f, a|e) = \sum_{a} P(f|e, a) P(a|e) \tag{2.4}$$

ここで，a はアラインメント（単語の対応関係）を表す．アラインメントの例を図 2.2 に示す．この例では,「ぐすーよー」が「みな」,「はじみてぃ」が「はじめまして」,「うぃがなびら」が「様」と対応していることを表している．

SMT において，翻訳モデル確率とアラインメントはフレーズテーブルと呼ばれる表として管理される [14]．フレーズテーブルの例を表 2.1 に示す．この例は琉日機械翻訳におけるフレーズテーブルで，原言語が琉球方言，目標言語が日本語（標準語）である．表の要素は，左から順に「琉球フレーズ w_r」,「日本語フレーズ w_j」,「日琉方向フレーズ翻訳確率 $\phi(w_r|w_j)$」,「日琉方向の単語翻訳確率の積」,「琉日方向フレーズ翻訳確率 $\phi(w_j|w_r)$」,「琉日方向の単語翻訳確率の積」を表している．

図 2.2: アラインメントの例

表 2.1: フレーズテーブルの例

ぐすーよー	みな	1	0.5	1	0.6
はじみてぃ	はじめまして	1	0.3	0.4	0.5
うぅがなびら	様	0.35	0.2	0.4	0.3

フレーズベースの SMT では，原言語の文 f が I 個のフレーズ $(f_1, \cdots, f_I)$ に分割される．各フレーズ f_i が目標言語のフレーズ e_i に翻訳されるとき，翻訳モデル確率 $P(f|e)$ は，式 (2.5) に示すように，フレーズ翻訳確率 $\phi(f_i|e_i)$ と相対的なフレーズ歪みスコア $d(a_i - b_{i-1})$ の積で近似される [9]．フレーズ翻訳確率 $\phi(f_i|e_i)$ は，ある目標言語フレーズ e_i が複数の原言語フレーズ f_i' と対応付けられているときに，f_i から e_i へと翻訳される確率であり，式 (2.6) により定義される．$count(f, e)$ は対訳コーパスの中で f と e が対応付けられている回数を表す．フレーズ歪みスコア $d(a_i - b_{i-1})$ は，翻訳によりフレーズの位置が大きく異なる場合に大きいペナルティを与えるものであり，式 (2.7) で定義される．α は翻訳前後のフレーズの位置の違いに対するペナルティの強さを調整するパラメータであり，任意に設定される．a_i は i 番目の目標言語フレーズ e_i に翻訳された原言語フレーズ f_i の開始位置，b_{i-1} は $(i-1)$ 番目の目標言語フレーズ e_{i-1} に翻訳された原言語フレーズ f_{i-1} の終了位置を表す．この値は，左端は 0，右端はフレーズ分割数 I となる．

$$P(f|e) = \prod_{i=1}^{I} \phi(f_i|e_i) d(a_i - b_{i-1}) \tag{2.5}$$

$$\phi(f_i|e_i) = \frac{count(f_i, e_i)}{\sum_{f_i'} count(f_i', e_i)} \tag{2.6}$$

$$d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1}, -1|} \tag{2.7}$$

式 (2.8) は，図 2.2 の例文について，式 (2.5) にしたがって計算された翻訳モデルの確率である．パラメータは表 2.1 の値を用いている．

$$
\begin{aligned}
&P(\text{“ぐすーよー はじみてぃ うぅがなびら”}|\text{“みな 様 はじめまして”})\\
&\quad = \phi(\text{“ぐすーよー”}|\text{“みな”})d(0-0)\\
&\quad * \phi(\text{“うぅがなびら”}|\text{“様”})d(2-1)\\
&\quad * \phi(\text{“はじみてぃ”}|\text{“はじめまして”})d(1-3)\\
&\quad = 1 * 2.718^{1} * 0.35 * 2.718^{0} * 1 * 2.718^{3}\\
&\quad = 19.101\cdots
\end{aligned} \tag{2.8}
$$

SMT は，自然言語の文法やルールに関する情報を明示的に保持せず，これらを対訳コーパスから得られる統計的モデルによって表現するため，これらの知識がなくとも大量の対訳コーパスがあれば翻訳を行うことが可能であるといった利点がある [14]．しかし，翻訳精度は対訳コーパスの規模に依存するため，方言などの言語資源の乏しい言語を対象とした機械翻訳では翻訳の精度が低いという欠点がある．

2.2 低言語資源の言語を対象とした機械翻訳の研究

本節では，低言語資源の言語を対象とした機械翻訳の研究について述べる．特に方言を対象とした研究を多く紹介する．

- 山形方言を対象とした研究

  柴田らは，山形の一つの地域である村山地域の方言（以下，村山方言とする）と日本語の間の双方向での統計的機械翻訳システムを構築した [14]．そのシステムでは，彼らが以前に構築したルールベースの機械翻訳システムを用いて村山方言と日本語の対訳コーパスを構築し，これを用いて SMT モデルを学習している．

  村山方言は詳細な文法や語彙に関する文献がほぼ皆無であり，ルールベースの機械翻訳を行うために必要な包括的なルールと語彙辞書を作成することが難しい．そこで，小規模なルールと辞書で最小限のルールベースの機械翻訳システムを最初に構築し，これを用いてある程度の文の誤りを許した共通語-方言対訳コーパスを作成し，フレーズベースの SMT により翻訳システムを学習するアプローチを採用している．これにより，ルールや文法作成の時間的コストを減らし，方言に関する知識や文献が少なくても村山方言と日本語の間の翻訳が可能になった．翻訳精度については，ルールベースの機械翻訳システムと同程度以上であったと報告している．

- ベトナム語を対象とした研究

Nguyen らは，言語構造の大きく異なる言語対である日本語とベトナム語の間で統計的機械翻訳を行うシステムを構築した [10]．近年，ヨーロッパ言語や英語-中国語などの言語対における機械翻訳の精度は向上しているが，言語構造の大きく異なる言語対に関しては発展途上であり，その一つである日本語-ベトナム語の機械翻訳を研究の対象としている．構文トランスファ方式をベースとし，構文解析により日本語の構文をベトナム語の構文に変換することで機械翻訳を実現している．統計的機械翻訳ツール Moses[21] による機械翻訳（ベースライン）と，提案手法である構文トランスファ方式による機械翻訳を評価した結果，提案手法の方がベースラインの手法よりも精度が向上したと報告している．また，翻訳性能を向上させるために，より多くの量の対訳コーパスを作成することを今後の課題として挙げている．

- ドイツ語方言を対象とした研究

  Honnet らは，ドイツ語方言とドイツ語標準語の間で統計的機械翻訳を行うシステムを構築した [4]．ドイツ語方言の未知語に対して正規化処理を行うことで，未知語を翻訳可能な単語へと変換する手法を提案している．提案された正規化処理は，スペル変換規則の利用，発音表記の利用，文字単位で翻訳を行う CBNMT(Character-based Neural Machine Translation) の利用の 3 種類であった．実験の結果，CBNMT による正規化処理が最も良い手法であると結論付けている．

- インド言語を対象とした研究

  Irvine と Callison-Burch は，コンパラブルコーパス（対訳関係にはないが同じトピックに関する異なる言語のテキストの組を集めたコーパス）と，原言語ならびに目標言語の単言語コーパスから，バイリンガル辞書を自動構築して SMT モデルを補完する手法を提案し，低言語資源の言語である 6 種類のインド言語（Tamil，Telugu，Bengali，Malayalam，Hindi，Urdu）と英語の間で統計的機械翻訳を行うシステムを構築した [5]．自動構築されたバイリンガル辞書におけるスコアが上位 k 個の単語をフレーズテーブルに追加し，未知語を SMT のモデルに組み込むことでフレーズテーブルにおける未知語の割合を低減することを狙う方法 (+OOV Trans.) と，コンパラブルコーパスを用いて時間的素性・文脈的素性・トピック素性・正字法の素性・頻度の類似性の素性を抽出し，それらの素性をフレーズテーブルに組み込むことで精度を改善する方法 (+Features) を評価した．その結果，(+OOV Trans.) と (+Features) の両方を適用することで，6 つの言語のうち 5 つについて，いずれかの手法を単独で適用するよりも優れた翻訳が得られた．

2.3 対訳コーパスの自動拡張手法を用いた機械翻訳の研究

本節では対訳コーパスの自動拡張手法に関する研究について述べる．対訳コーパスの自動拡張とは，低言語資源の言語の文を対象としたときなど，対訳コーパスの量が十分にないときに，対訳文を自動的に生成し，対訳コーパスの量を増やす手法を指す．

- 意味役割付与を用いた対訳コーパス自動拡張手法

  Gao と Vogel は，意味役割付与（SRL:Semantic Role Labeling）を利用して対訳コーパスを自動生成する手法を提案し，中国語-英語間の統計的機械翻訳に適用した [2]．まず，初期のコーパスに対して単語アラインメントとフレーズ抽出を行い，原言語または目標言語のどちらかの文に SRL ラベラー（SRL のツール）を用いて意味役割を付与する．次に，意味フレーム・意味役割・対応する原言語フレーズと目標言語フレーズのセットである SRL 置換ルール（SSR:SRL Substitution Rules）を抽出する．そして，初期のコーパスの文のフレーズペアが同じ意味フレームと意味役割を持つ場合，そのフレーズペアを SSR で置き換えて新しい対訳文を生成する．実験の結果，5 つの異なるテストセットにおいて，対訳コーパスの自動拡張によって翻訳性能が改善し，その精度は人手で作成された大量の対訳コーパスで学習されたシステムに匹敵した．

- 言い換えによる対訳文自動生成手法

  藤原らは，対訳文を自動的に生成する手法である ACG(Automatic Corpora Generation) を提案し，日英統計的機械翻訳に適用した [1]．ACG は類似候補文生成処理と候補識別処理の 2 つから構成されている．類似候補文生成処理では，初期の日英対訳コーパスの原言語（日本語）の文に対して，WordNet や PPDB などの言語資源から構築した言換え表現のデータベースを用いて別の文に言い換える処理を行い，類似候補文を生成する．候補識別処理では，意味的・文法的に破綻した類似候補文を除くために，確率言語モデル（N-gram モデル）による生成確率の大きい文を選別する．最後に，選別された類似候補文を元の目標言語（英語）と組み合わせて，新しい日英対訳文の組を作成し，これを対訳コーパスに追加する．また，システム利用者が出力文に対して品質評価を行い，その品質が低い場合に訳文の選択・修正を行うフィードバック処理を行う．すなわち，この手法は完全に自動的に対訳コーパスを拡張するのではなく，人手作業も必要とするが，その人的コストを抑えながら翻訳性能の向上を図っている．

  実験の結果，初期のコーパスを用いて SMT を学習したときと比較して，類似候補文を加えた対訳コーパスを用いて SMT を学習することで精度が大き

く向上した．また，人によるフィードバックにより追加される文を増やすことで，自動生成した類似候補文のみを使う場合と比較して，フィードバック後の対訳コーパスの文量が 1/3∼1/2 であっても，翻訳の精度が向上していることを確認している．

- 単語置換による対訳文候補生成と doc2vec を用いた対訳文候補選択手法

  久高と金城は，ACG を琉日統計的機械翻訳に適用した [7]．藤原らの提案した ACG とは異なり，初期の対訳コーパスの文に対し，琉日対訳辞書を用いた単語置換により対訳文候補を生成する．さらに，doc2vec[8] を利用して自動生成した対訳文候補の文間類似度を考慮して候補文を選択した．この対訳文候補選択のステップでは，文間類似度の低い候補文を選択することで，互いに似ていない候補文から構成される多様性のある対訳コーパスを構築することを狙っている．多様な文から構成される対訳コーパスから学習された SMT のシステムは，様々な文の翻訳に対応することが可能になるため，翻訳精度の向上に繋がると考えたためである．

  実験の結果，SMT モデルの学習に使用した対訳コーパスの文をテスト文として利用するクローズドテストにおいては翻訳精度が向上したが，SMT モデルの学習に使用していない対訳文をテスト文として利用するオープンテストにおいては翻訳精度が向上しなかった．この論文では，計算量の問題から，自動生成した対訳文候補のすべてに対して文間類似度を計算せず，ランダムに選択された少量の対訳文候補に対して文間類似度を計算しているが，提案した翻訳システムの翻訳性能が十分に高くないことから，自動生成した候補文全体から満遍なく文を選択できるような処理を考案することを今後の課題として挙げている．

2.4 本研究の特色

藤原らの研究 [1] では日英間の統計的機械翻訳において ACG を適用していたが，本研究では低言語資源の言語である琉球方言の統計的機械翻訳に対して ACG を適用する．その際，対訳コーパスをどのように拡張するかについて，品質・多様性・量の観点から検討する．品質について，藤原らの研究と同様に，確率言語モデルを用いて標準語の文の品質評価を行うことで自然な対訳文を選別するが，これを琉日機械翻訳に適用した研究例は本研究が初めてとなる．多様性について，久高と金城の研究 [7] では，doc2vec により計算した文間類似度の低い文を選択することで対訳コーパスの多様性を確保することを実現していたが，本研究ではそれとは異なる方法を採用する．具体的には，対訳文候補の生成元となる文に偏りがないように対訳コーパスに追加する対訳文を選別することで対訳コーパスの多様性を確保する．量について，ACG によって自動生成される対訳文の量と翻訳精度の関係を明らかにする試みは本研究が初めてとなる．

第3章　提案手法

本研究の提案手法の概要を図 3.1 に示す．まず，初期の琉日対訳コーパスと琉日対訳辞書を用意する．初期の琉日対訳コーパスの量は少量と仮定する．対訳文候補生成のステップでは，新しい琉球方言と標準語訳の組を自動生成する（図 3.1 の①）．対訳文の評価のステップでは，生成した対訳文候補の中から品質・多様性・量を考慮して最適なものを選択する（図 3.1 の②）．このように自動生成した対訳コーパスと初期の対訳コーパスを合わせて，拡張琉日対訳コーパスを構築する（図 3.1 の③）．最後に，作成した拡張琉日対訳コーパスから統計的機械翻訳のモデルを学習する（図 3.1 の④）．

以下の節では，それぞれの処理について詳細を述べる．3.1 節では，図 3.1 の①②③に相当する対訳コーパスの拡張について述べる．3.2 節では，図 3.1 の④に相当する SMT モデルの学習について述べる．

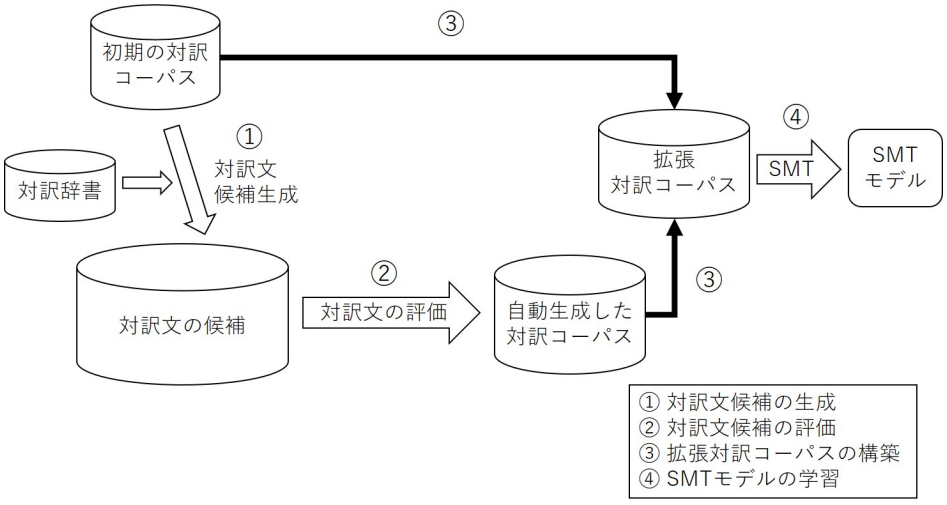

図 3.1: 提案手法の概要

3.1　対訳コーパスの拡張

本節では，初期の対訳コーパスから拡張対訳コーパスを構築する手法について述べる．以下，初期の対訳コーパスを式 (3.1) のように表記する．

$$I = \{(s_1, t_1), \cdots, (s_n, t_n)\} \tag{3.1}$$

s_i は原言語（琉球方言）文，t_i は目標言語（標準語）の文を表す．また，n は初期の対訳コーパスに含まれる対訳文の数を表す．初期の対訳コーパスの例を図 3.2 に示す．この例は，4 章の実験で使用する初期の対訳コーパスの琉球方言および標準語の文集合のうち，それぞれの先頭から 5 文を表示したものである．左端の数値は行番号であり，同じ行番号を持つ琉球方言と標準語の文が対訳関係にある．

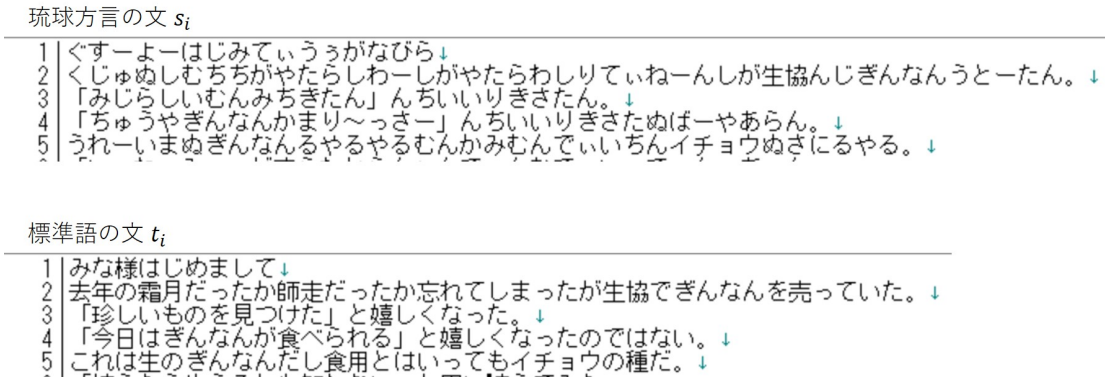

図 3.2: 初期の対訳コーパスの例

一方，琉日対訳辞書を式 (3.2) のように表記する．

$$D = \{(w_1^s, w_1^t, p_1), \cdots, (w_d^s, w_d^t, p_d)\} \tag{3.2}$$

対訳辞書は原言語の単語，目標言語の単語，品詞の 3 つ組から構成される．w_i^s は原言語の単語，w_i^t はその目標言語の対訳，p_i は w_i^s と w_i^t の品詞を表す．また，d は琉日対訳辞書に含まれる単語の数を表す．琉日対訳辞書の例を図 3.3 に示す．この例は，4 章の実験で使用する琉日対訳辞書の先頭から 15 単語を抽出したものである．各行の左から，標準語の単語，琉球方言の単語，品詞が記載されており，品詞別に琉球方言の 50 音順で並んでいる．

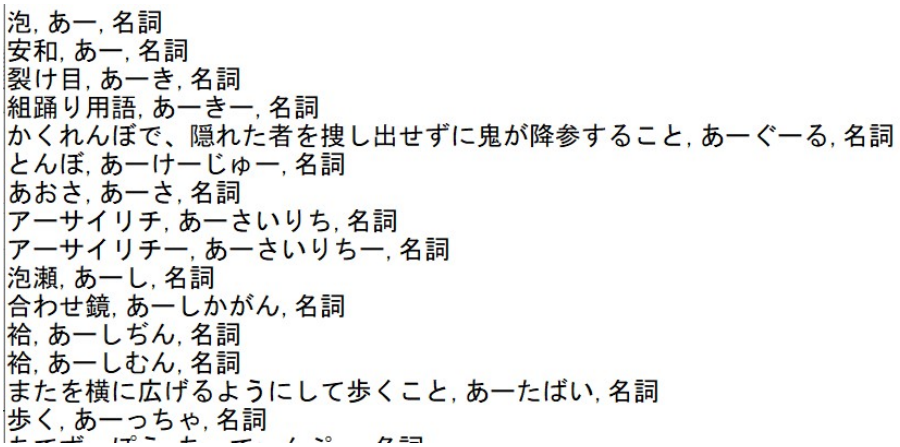

図 3.3: 琉日対訳辞書の例

3.1.1 対訳文候補生成

対訳文候補生成処理は，対訳文の候補を新しく生成する処理である．本研究では，先行研究 [7] と同じ手法で対訳文候補を生成する．初期対訳コーパス I の対訳文において，琉日対訳辞書 D に登録されている単語が原言語文 s_i と目標言語文 t_i の両方に出現するとき，それらの単語を，琉日対訳辞書に登録されている品詞が同じ別の単語に置き換えて，新しい原言語文 s'_i と目標言語の文 t'_i を生成する．図 3.4 に対訳文候補生成処理の例を示す．色付き部分の単語が置換された単語である．「はじめまして，はじみてぃううがなびら」の単語ペアが琉日対訳辞書に登録されているため，同じ感動詞の品詞である「こんにちは，ちゅーうがまびら」と「さよなら，んじちゃーびら」の単語ペアを用いて新しい対訳文候補が生成される．なお,「みなさん，ぐすーよー」と「おいしい，まーさん」の単語ペアは品詞が異なるために新しい対訳文候補の生成に使われていないことに注意していただきたい．

初期対訳コーパス I における対訳文

t_i	みな様　はじめまして	s_i	ぐすーよー　はじみてぃううがなびら

自動生成した対訳文の候補　　単語ペアの置換

t'_i	みな様　こんにちは	s'_i	ぐすーよー　ちゅーうがまびら
t'_i	みな様　さよなら	s'_i	ぐすーよー　んじちゃーびら

琉日対訳辞書 D

日本語 w_i^t	琉球方言 w_i^s	品詞 p_i
はじめまして	はじみてぃううがなびら	感動詞
こんにちは	ちゅーうがまびら	感動詞
さよなら	んじちゃーびら	感動詞
みなさん	ぐすーよー	名詞
おいしい	まーさん	形容詞

図 3.4: 対訳文候補の生成例

琉日対訳辞書中の単語が 1 つの文に複数回出現する場合は，それぞれの単語について単語置換を行い，新しい対訳文を生成する．この際，同時に複数の単語を置換せず，1 つの単語のみを置換する．この処理の例を図 3.5 に示す．この例では，「母，あんまー」という単語ペアが 1 文中に 2 箇所あり，それぞれが単語置換の対象となる．1 つ目の単語ペア（緑色で示した部分）の置換により対訳文 (s'_i, t'_i) を生成し，2 つ目の単語ペア（黄色で示した部分）の置換により対訳文 (s''_i, t''_i) を生成する．

上記の対訳文候補生成のアルゴリズムの疑似コードを Algorithm 1 に示す．疑似コードの詳細は以下の通りである．

初期対訳コーパス I における対訳文

t_i	私の 母 と あなたの 母	s_i	わったーぬ あんまー とぅ やーぬ あんまー

自動生成した対訳文の候補 ⇩ 単語ペアの置換

t'_i	私の 祖父 と あなたの 母	s'_i	わったーぬ たんめー とぅ やーぬ あんまー
t'_i	私の 祖母 と あなたの 母	s'_i	わったーぬ はーめー とぅ やーぬ あんまー
t''_i	私の 母 と あなたの 祖父	s''_i	わったーぬ あんまー とぅ やーぬ たんめー
t''_i	私の 母 と あなたの 祖母	s''_i	わったーぬ あんまー とぅ やーぬ はーめー

琉日対訳辞書 D

日本語 w_i^t	琉球方言 w_i^s	品詞 p_i
母	あんまー	名詞
祖父	たんめー	名詞
祖母	はーめー	名詞

図 3.5: 対訳文候補の生成例（対訳辞書の単語が 1 文中に複数回出現する場合）

- 1 行目は，初期の対訳コーパス I の i 番目の対訳文 (s_i, t_i) と対訳辞書 D が入力として与えられることを表す．
- C_i は i 番目の対訳文から生成される対訳文候補の集合を表す．2 行目は，この初期値を空集合としている．
- 3 行目は，対訳辞書 D に含まれる単語対 (w_j^s, w_j^t, p_j) それぞれに対して，4 行目からの処理を行うことを表す．
- 4 行目は，w_j^s が s_i に含まれ，かつ w_j^t が t_i に含まれているとき，5 行目からの処理を行うことを表す．
- 5 行目は，対訳辞書に含まれている品詞 p_j が同じで (w_j^s, w_j^t, p_j) 以外の単語対 (s_k, t_k, p_k) に対して，6~8 行目の処理を行うことを表す．
- 6 行目は，対訳文の原言語側 s_i 中に含まれる単語 w_j^s を w_k^s に置き換えて，対訳文候補の原言語側の文 s' を新しく生成する処理を表す．
- 7 行目は，対訳文の目標言語側 t_i 中に含まれる単語 w_j^t を w_k^t に置き換えて，対訳文候補の目標言語側の文 t' を新しく生成する処理を表す．
- 8 行目は，新しく生成した対訳文候補 (s', t') を C_i に追加することを表す．
- 12 行目は，対訳文候補生成処理の結果として C_i を出力することを表す．

Algorithm 1 対訳文候補の生成

procedure GENERATEPARALLELSENTENCE(s_i,t_i,D)
  $C_i \leftarrow \emptyset$
  for $(w_j^s, w_j^t, p_j) \in D$ **do**
    if $w_j^s \in s_i \wedge w_j^t \in t_i$ **then**
      for $(w_k^s, w_k^t, p_j) \in D$ s.t. $k \neq j$ **do**
        $s' =$ replace w_j^s with w_k^s in s_i
        $t' =$ replace w_j^t with w_k^t in t_i
        $C_i \leftarrow C_i \cup \{(s', t')\}$
      end for
    end if
  end for
  return C_i
end procedure

3.1.2 対訳文候補選択

前節で生成した対訳文の候補の中から対訳コーパスに追加するべき対訳文を選択する．既に述べたように，選択の際には，最終的に構築される拡張対訳コーパスの品質，多様性，量を考慮する．3.1.2.1，3.1.2.2，3.1.2.3 では，それぞれ品質，多様性，量を考慮した対訳文候補の選択について説明する．3.1.2.4 ではこれら全てを考慮した対訳文候補選択アルゴリズムを示す．

3.1.2.1 品質を考慮した対訳文候補の選択

単語置換により自動生成した対訳文候補には不自然な文も含まれているため，確率言語モデルによって文の品質を評価しスコア付けすることで，良い品質の対訳文候補を選択する．対訳文 (s, t) の品質のスコア $score_{LM}(s, t)$ を式 (3.3) のように定義する．

$$score_{LM}(s, t) = \log P(t) \tag{3.3}$$

$P(t)$ は目標言語側の対訳文候補 t の生成確率であり，式 (2.2) の N-gram モデルにより求める．本研究では，N=5 とし，毎日新聞の 10 年分（1994∼1997 年，2001 年，2007∼2011 年）の記事からバックオフスムーシング法により推定する．確率言語モデルの学習には SRILM[22] を用いる．確率言語モデルの学習の際には，コーパスに対する前処理として，全角を半角に変換する処理，形態素解析器 MeCab[20] を用いた単語分割処理，数字を特殊な記号 <N> に置換する処理を行う．前処理後の毎日新聞記事のコーパスの規模は，7,351,312 文，302,523,136 単語である．なお，対訳文のスコアを計算する際にも，目標言語の文 t に対して同様の処理を行う．

$score_{LM}$ は目標言語文の生成確率 $P(t)$ の対数をスコアとしており，単に目標言語文全体が自然な単語の並びであるかを評価している．したがって，このスコアは単語置換された部分以外の単語並びの自然さも考慮される．しかしながら，置換された単語の前後の単語並びの自然さのみを考慮した方が，新しく生成した対訳文の品質を適切に評価できる可能性もある．そこで，単語置換された部分以外の単語の並びにスコアが影響されないように，単語置換前後の文の確率言語モデルの差をスコアとする方法も提案する．このスコアの定義を式 (3.4) に示す．

$$score_{dif}(s,t) = \log P(t) - \log P(t^o) \tag{3.4}$$

$P(t^o)$ は対訳文候補 t の生成元となった文（初期対訳コーパスの文）t^o の生成確率を表し，$P(t)$ と同様に，式 (2.2) の N-gram モデルにより求める．単語置換前後の文の生成確率の対数の差をスコアと定義することで，単語置換された単語とその前後の単語との繋がりが自然なものであるかを評価することができる．

対訳文候補の品質評価をするためには，本来は原言語の文の自然さと目標言語の文の自然さの両方を考慮してスコアを定義するべきである．しかし，上記で定義した 2 つのスコアは原言語文 s について全く考慮していない．これは，原言語（琉球方言）の確率言語モデルの学習に必要充分な量の琉球方言の単言語コーパスを確保できなかったためである．

本研究では，琉球方言側の対訳文候補の品質評価について，確率言語モデルによる品質評価方法の代替案を検討した．具体的には，初期の対訳コーパスの文と対訳文候補に出現する単語 n-gram の重なりの度合を測る方法や，検索エンジンでの対訳文候補の単語 n-gram のヒット件数を使用する方法を検討した．しかし，初期の対訳コーパスでも，ウェブにおいても，琉球方言のデータが全体的に少なく，単語 n-gram の重なりの頻度が少なかったり，検索エンジンによるヒット件数が少なかったりしたため，どちらの方法もうまくいかなかった．したがって，本研究では原言語（琉球方言）の品質評価は行わず，目標言語（標準語）の品質評価のみ行うこととする．

3.1.2.2　多様性を考慮した対訳文候補の選択

3.1.2.1 で述べた方法にしたがって，単にスコアの高い（生成確率の高い）対訳候補文を選択すると，似たような文のみが選択される可能性がある．生成確率の高い文は，使用頻度の高い一般的な単語を多く含む文であると考えられるため，そのような文ばかりを集めてしまうと拡張後の対訳コーパスの多様性が低くなると予想される．また，N-gram モデルでは短い文に高い生成確率を与える傾向があり，短い文ばかりが選択されやすい．このとき，拡張後の対訳コーパス全体の単語数が少なくなり，これもまた対訳コーパスの多様性の低下につながる．一方，統計的機械翻訳のモデルを学習する際には，学習データとする対訳コーパスに多様な

文が含まれていた方が，様々な文を正確に翻訳できる汎用性の高いモデルが学習されやすい．

拡張後の対訳コーパスが多様な文から構成されるようにするために，本研究では，初期の対訳コーパス I 中のすべての対訳文から同じ数の対訳文を生成することで拡張後の対訳コーパスの多様性を確保する．言い方を変えると，I における i 番目の対訳文から生成された対訳文候補 C_i の中から，スコアが大きい文を同じ数だけ選択する．これにより，初期の対訳コーパスにおける全ての文について，それを元にして生成された候補文が対訳コーパスに含まれるようになり，初期の対訳コーパスが持つ文脈情報の全てを SMT モデルの学習に用いることができる．一方，単に確率言語モデルのスコアの高い対訳文のみを選択するときは，初期の対訳コーパスのうち短い文から生成された対訳文しか拡張対訳コーパスに含まれない可能性がある．

3.1.2.3　量を考慮した対訳文候補の選択

拡張後の対訳コーパスの量も機械翻訳の性能に影響を与える．自動生成する文の数が多いほど SMT の訓練データ量を増やすことができるが，その分，不自然な対訳文が含まれる可能性も上がる．ここでは，拡張対訳コーパスの文の数 m を最適化することを考える．最適化とは，m を変化させて SMT モデルを学習し，その翻訳性能を開発データの対訳コーパスを使用して測定し，翻訳性能が最も高くなる m を選択することである．ただし，本研究の実験においては，初期の対訳コーパスが十分になかったため，m を最適化するための開発データを用意することができなかった．したがって，本研究では m の最適化は行わず，m の値を変えたときの機械翻訳の自動評価指標の値の変化を調査する．

3.1.2.4　対訳文候補選択のアルゴリズム

拡張対訳コーパス作成のアルゴリズムの疑似コードを Algorithm 2 に示す．このアルゴリズムの詳細は以下の通りである．

- 1 行目は，I（初期の対訳コーパス），D（対訳辞書），m（拡張対訳コーパス E の文の数）を入力とすることを表す．
- 2 行目は，初期の対訳コーパスの文量 $|I|$ を n として定義している．
- 3 行目は，1 つの対訳文候補の集合から選択する対訳文の数 m' を定義している．最終的に構築する対訳コーパスは，初期の対訳コーパス内の対訳文と自動生成した対訳文から構成されるため，自動生成するべき対訳文の数は $m-n$ である．また，初期の対訳コーパス内の全ての対訳文から同じ数だけ新しい対訳文を生成するため，1 つの対訳文から生成するべき新しい対訳文の数は $(m-n)/n$ となる．

- 4 行目は，初期の対訳コーパスの 1 番目から n 番目の文，すなわちすべての文について，5〜7 行目の処理を行うことを表す．
- 5 行目は，Algorithm 1 で示した対訳文候補生成処理を対訳文 (s_i, t_i) に適用し，(s_i, t_i) から新たに生成された対訳文候補の集合 C_i を得ることを表す．
- 6 行目は，C_i 中のすべての候補文についてスコアを計算することを表す．このスコアは式 (3.3) または式 (3.4) で計算される．
- 7 行目は，スコアの高い上位 m' 個の候補文の集合 $\widehat{C}_i$ を選択することを表す．
- 9 行目は，初期の対訳コーパスの各文から生成された $\widehat{C}_i$ と初期の対訳コーパス I を合わせて，拡張対訳コーパス E を作成することを表す．
- 10 行目は，拡張対訳コーパス作成処理の結果として E を出力することを表す．

Algorithm 2 拡張対訳コーパスの作成

1: **procedure** ExpandParallelCorpus(I, D, m)
2: $n \leftarrow |I|$
3: $m' \leftarrow \frac{m-n}{n}$
4: **for** $i = 1$ to n **do**
5: $C_i \leftarrow$ GenerateParallelSentence(s_i, t_i, D)
6: Compute $score(s, t)$ for all $(s, t) \in C_i$
7: $\widehat{C}_i \leftarrow$ most highly scored m' pairs in C_i
8: **end for**
9: $E \leftarrow I \cup \left(\bigcup_i \widehat{C}_i \right)$
10: **return** E
11: **end procedure**

3.2 琉日 SMT

本節では，前節で作成した拡張対訳コーパスを用いて琉日統計的機械翻訳（琉日 SMT）のモデルを学習し，琉球方言を日本語標準語に翻訳する手続きについて述べる．

本研究で構築した琉日 SMT のフローチャートを図 3.6 に示す．使用したツールを表 3.1 に示す．また，琉日 SMT の翻訳処理手順を以下の 1.〜4. に述べる．この手順は図 3.6 の①〜④と対応している．

[翻訳処理手順]

1. 対訳コーパスの前処理

   KyTea[19] を用いて対訳文の単語分割を行う．図 3.7 は単語分割の例である．元の文に対し，単語間にスペースを挿入した文字列が出力として得られる．KyTea は日本語の形態素解析ツールであり，これを琉球方言の文に用いても単語分割が正しくできない．ただし，KyTea では，単語分割された単言語コーパスを用いて単語分割モデルを学習し，それを利用することで日本語以外の言語の文の単語分割が可能である．本研究では，琉球方言の単語分割を行うために，初期の対訳コーパスの琉球方言の文を人手分割したコーパスを用いて琉球方言の単語分割モデルを学習している．

   日本語標準語以外の場合，文を単語に正しく分割するためには，単語分割モデルの学習に大量の単言語コーパスが必要となる．しかし，単語分割済みの琉球方言の単言語コーパスを大量に用意することは難しく，本研究では少量の初期の対訳コーパスだけを学習に用いている．したがって，琉球方言は標準語と比べて KyTea による単語分割の誤りが多いことが予想される．そのため，4 章の実験では，琉球方言のテスト文について，KyTea で機械的に単語分割を行う自動分割と，あらかじめ手動で単語分割を行う人手分割の 2 通りで実験を行い，両者の違いを比較する．

2. 言語モデルの作成

   拡張琉日対訳コーパスの標準語の文から，単語 5-gram モデルを KenLM[18] を用いて学習する．このとき，modified Kneser-Ney smoothing における discount パラメータをデフォルトの設定値で与えるオプション「--discount_fallback」を指定する．

3. 翻訳モデルの作成

   フレーズ翻訳モデルを Moses 付属のスクリプトにより学習する．アラインメントツールとして GIZA++[17] を用いる．このときオプションとして，「--alignment grow-diag-final-and」,「--reordering msd-bidirectional-fe」を指定する．また，オプション「-external-bin-dir」で使用するアラインメントツールのディレクトリを指定する．スクリプト実行後に生成されたパラメータ設定ファイル moses.ini について，以下の記述を変更する．

   - フレーズ並べ替え範囲のパラメータである [distortion-limit] を，デフォルト値 6 から −1（制限なし）に変更する．
   - 使用する言語モデル作成ツールの指定をデフォルト「SRILM」から「KenLM」へと変更する．

4. 入力琉文の日文への翻訳

   手順 2. と 3. で作成した言語モデルと翻訳モデルを用い，入力琉球方言文を標準語文へ翻訳する．この際のデコーダとして Moses を用いる．

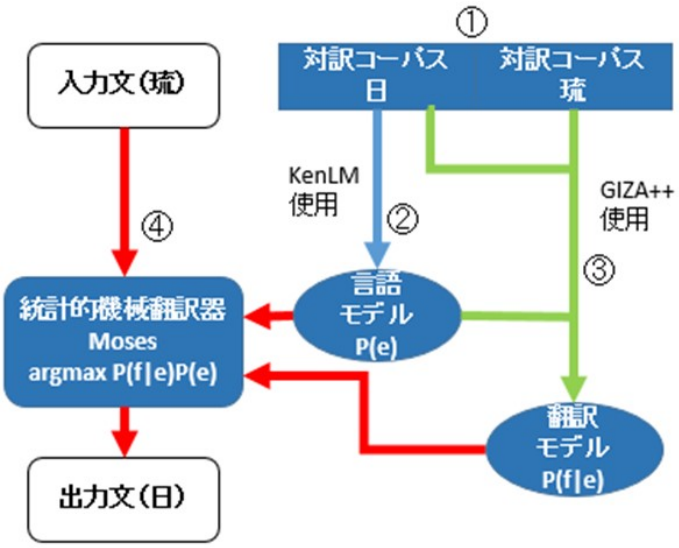

図 3.6: 琉日 SMT フローチャート

表 3.1: 琉日 SMT で使用するツール

名称	役割
Moses	統計的機械翻訳器（デコーダ）
KyTea（京都テキスト解析ツールキット）	形態素解析ツール
GIZA++	アラインメントツール
KenLM	言語モデル作成ツール

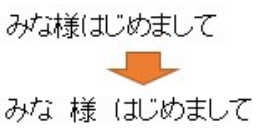

図 3.7: 単語分割の例

第4章　評価実験

4.1　使用データ

琉日対訳コーパスとして，ウェブサイト「沖縄方言であれこれ」[15] から取得した琉球方言文と標準語文の対訳 1,102 組を用いた．これらのうち，ランダムに選択した 100 組の対訳文をテストデータとし，残りの 1,002 組を初期の琉日対訳コーパス I とした．

対訳文候補の生成に用いる琉日対訳辞書として,「琉球語音声データベース」[16] に記載されている 17,499 個の単語対と品詞のセットを用いた．同データベースの単語対は，名詞，動詞，自動詞，他動詞，形容詞，副詞，連体詞，接続詞，感動詞，助詞，接頭辞，接尾辞，句，連詞の 14 種類に品詞が分類されている．品詞別の単語数を表 4.1 に示す．

表 4.1: 琉日対訳辞書の品詞別単語数

品詞	単語数
名詞	13,617
動詞	37
自動詞	877
他動詞	1,177
形容詞	367

品詞	単語数
副詞	794
連体詞	55
接続詞	11
感動詞	160
助詞	63

品詞	単語数
接頭辞	62
接尾辞	190
句	82
連詞	7

4.2　評価尺度

本実験では，提案手法やそれと比較するベースライン手法など，様々な手法で SMT のモデルを学習し，それを用いてテスト文の琉球方言を標準語に翻訳する．そして，対訳コーパスの標準語の文を参照訳（正解の翻訳）として，得られた翻訳文（システム翻訳）の品質を自動評価する．評価指標として，BLEU(BiLingual Evaluation Understudy)[13] と RIBES (Rank-based Intuitive Bilingual Evaluation Score)[3] の 2 つを用いる．以下，それぞれの指標について説明する．

BLEU は，文の類似度を測定する指標であり，0〜100%のスコアで表される．この値が大きいほど良い評価を表す．機械翻訳の評価の場合，システム翻訳と参照訳の類似度を測る．BLEU スコアは式 (4.1) により求められる．p_n は n-gram 適合率であり，翻訳結果中の単語 n-gram の総数のうち，複数の参照訳中のいずれかに含まれるものの割合を表している．基本的に，システム翻訳と参照訳で重複する単語 n-gram（n 個の単語の列）が多いほど，BLEU は高い値を取る．一方，w_n は n-gram の n に対する重みである．一般に，n が大きいほど w_n が高くなるように設定される．BP(Brevity Penalty) は，短い翻訳文が高評価点にならないように補正するパラメータであり，式 (4.2) によって計算される．BLEU スコアの計算は Moses 付属スクリプトを用いて行い，w_n や N などのパラメータはデフォルト値を使用する．

$$\mathrm{BLEU} = \mathrm{BP} * \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4.1}$$

$$\mathrm{BP} = \min\left(1, \exp\left(1 - \frac{\text{参照訳文集合の語数}}{\text{システム翻訳文集合の語数}}\right)\right) \tag{4.2}$$

RIBES は，訳語の違いや語順を考慮した自動評価法であり，参照訳とシステム翻訳との間で一致して出現する単語の出現順の近さに基づいて評価を行う．RIBES は式 (4.3) で定義される．H はシステム翻訳文集合，R は参照訳文集合である．τ は Kendall の順位相関係数 [6] である．それを [0,1] の値に正規化したものが式 (4.4) の NKT（Normalized Kendall's τ）である．$P(h_i, r_i)$ は単語正解率であり，式 (4.5) で定義される．これは，システム翻訳 h_i の単語のうちアラインメントをとることができた単語数の割合を表す．式 (4.6) で定義される BP_s は式 (4.2) とほぼ同じ考えに基づいているが，BLEU では文集合全体で単語数を計算していたものを，RIBES においては文単位で計算する．α は単語適合率の重みであり，大きいほど訳語の違いに敏感になる．β は BP の重みであり，大きいほど訳文の長さに敏感になる．本研究では，この重みは公開されている RIBES 計算ツール[1]のデフォルト値である $\alpha = 0.25$, $\beta = 0.1$ に設定している．

$$\mathrm{RIBES}(H, R) = \frac{\sum_{h_i \in H} \max_{r_j \in R_i} \left\{\mathrm{NKT}(h_i, r_j) \cdot P(h_i, r_j)^{\alpha} \cdot \mathrm{BP}_s(h_i, r_j)^{\beta}\right\}}{|H|} \tag{4.3}$$

$$\mathrm{NKT} = \frac{\tau + 1}{2} \tag{4.4}$$

$$P(h_i, r_i) = \frac{\text{アラインメントをとることができた語数}}{h_i\text{の語数}} \tag{4.5}$$

$$\mathrm{BP}_s(h_i, r_i) = \min\left(1, \exp\left(1 - \frac{r_i\text{の語数}}{h_i\text{の語数}}\right)\right) \tag{4.6}$$

[1] http://www.kecl.ntt.co.jp/icl/lirg/ribes/index-j.html

4.3 実験条件

対訳文候補生成処理によって生成した候補文に対して，以下の手法で m 個の対訳文候補を選択し，対訳コーパスを拡張する．それぞれの対訳コーパスを用いて琉日 SMT モデルを学習し，テスト文 100 文の翻訳精度を測定する．また，比較のため，ACG を行わない場合と，先行研究 [7] による文の類似度（doc2vec）を用いた ACG 手法も評価する．今回の実験で比較する手法は以下の通りである．

(1) ACG なし (no-ACG)
　初期の琉日対訳コーパスのみで SMT モデルを学習する手法である．

(2) 文の類似度を考慮して対訳文を拡張する手法 (ACG-doc2vec)
　先行研究 [7] の手法．doc2vec で文の分散表現（ベクトル）を学習し，文間の類似度をそのベクトルのコサイン類似度で測る．対訳コーパスを拡張する際に，自動生成した対訳文の文間類似度を計算し，互いに似ていない対訳文の部分集合を求め，これを拡張対訳コーパスとする．互いの類似度が低い対訳文を選択することにより，多様な文が対訳コーパスに含まれることが期待される．本実験では，メモリ不足のため，自動生成したすべての対訳文の候補を使うことは困難だったため，この中からランダムに 1,882,176 組の対訳文を選択し，これらの標準語の文の集合から文の分散表現（ベクトル）を学習する．候補文の中から互いに文間類似度の低い 50 万文を選択し，拡張対訳コーパスとする．

(3) ランダム (ACG-random)
　自動生成した対訳文候補の中からランダムに対訳文を選択し，拡張対訳コーパスを得る．拡張対訳コーパスの文の数 m は，50 万，30 万，20 万，10 万，5 万，2 万 5 千，1 万，5 千，2 千とする．また，文をランダムに選択することによる結果のばらつきを考慮し，拡張対訳コーパスの作成と SMT モデルの学習の試行を 5 回繰り返し，BLEU または RIBES の平均値を測る．

(4) 提案手法 1 (OurACG-LM)
　自動生成した対訳文候補の中から確率言語モデル (Language Model: LM) のスコアの高い文を選択する手法である．すなわち，式 (3.3) のスコア $score_{LM}$ によって対訳文を選別する．拡張対訳コーパスの文の数 m は 50 万とする．

(5) 提案手法 2 (OurACG-LM-diverse)
　OurACG-LM と同様に式 (3.3) のスコア $score_{LM}$ によって対訳文を選別するが，拡張対訳コーパスの多様性を確保するために，初期の対訳コーパス I のすべての対訳文について，それから生成された対訳文候補から同じ数だけ対訳文を選択し，拡張対訳コーパスを構築する．拡張対訳コーパスの文の数 m は，50 万，30 万，20 万，10 万，5 万，2 万 5 千，1 万，5 千，2 千とする．

(6) 提案手法 3 (OurACG-Dif)

自動生成した対訳文から，それを生成した元の文と自動生成した文の確率言語モデルの差が大きいものを選択する手法である．すなわち，式 (3.4) のスコア $score_{dif}$ によって対訳文を選別する．拡張対訳コーパスの文の数 m は 50 万とする．

(7) 提案手法 4 (OurACG-Dif-diverse)

OurACG-Dif と同様に式 (3.4) のスコア $score_{dif}$ によって対訳文を選別するが，拡張対訳コーパスの多様性を確保するために，初期の対訳コーパス I のすべての対訳文について，それから生成された対訳文候補から同じ数だけ対訳文を選択し，拡張対訳コーパスを構築する．拡張対訳コーパスの文の数 m は，50 万，30 万，20 万，10 万，5 万，2 万 5 千，1 万，5 千，2 千とする．

(8) 提案手法 5 (OurACG-LM-random)

ACG-random と OurACG-LM を組み合わせた手法である．まず，生成した候補文からランダムに 50 万文を選択し，その中から式 (3.3) のスコア $score_{LM}$ が設定した閾値以上になる候補文を選別し，拡張対訳コーパスを構築する．閾値は -100，-200，-300 のいずれかとする．閾値が高いほど拡張対訳コーパスの文量は少なくなる．拡張対訳コーパスの作成と SMT モデルの学習を 5 回試行し，BLEU および RIBES スコアの平均値で評価する．確率言語モデルのスコアの高い文を選択する前にランダムに文を選択するのは，単に確率言語モデルのスコアが高い文を選ぶと短い文が選ばれやすくなるため，最初に文をランダムに選択することにより長い文が選ばれやすくなるようにするためである．拡張対訳コーパスの多様性を確保する手法の一つといえる．

実験では琉球方言の文の単語分割について，KyTea で機械的に単語分割を行う自動分割と，あらかじめ手動で単語分割を行う人手分割の 2 通りの方法を採用した．後者では琉球方言の文の単語分割の誤りがないという理想的な条件の下で機械翻訳の手法を比較する．また，人手分割のとき，テスト文だけでなく，初期の対訳コーパスの琉球方言の文もすべて人手により単語分割を行った．なお，人手分割を適用したのは琉球方言の文だけであり，標準語の文については KyTea によって自動的に文を単語に分割した．

4.4 実験結果と考察

初期の対訳コーパス 1,002 組の対訳文から，3.1.1 項の手法で生成された対訳文候補の数は 16,049,071 組であった．これらの候補文に対して，4.3 節で示した各手法で拡張した対訳コーパスを用いて SMT モデルを学習した．テストデータにおける 2 つの文に対するシステム翻訳の例を図 4.1 に示す．これは手法 no-ACG による翻訳結

果であり,「Translating」の行が入力文（琉球方言の文）,「BEST TRANSLATION」の行がそのシステム翻訳文（標準語の文）を表す.

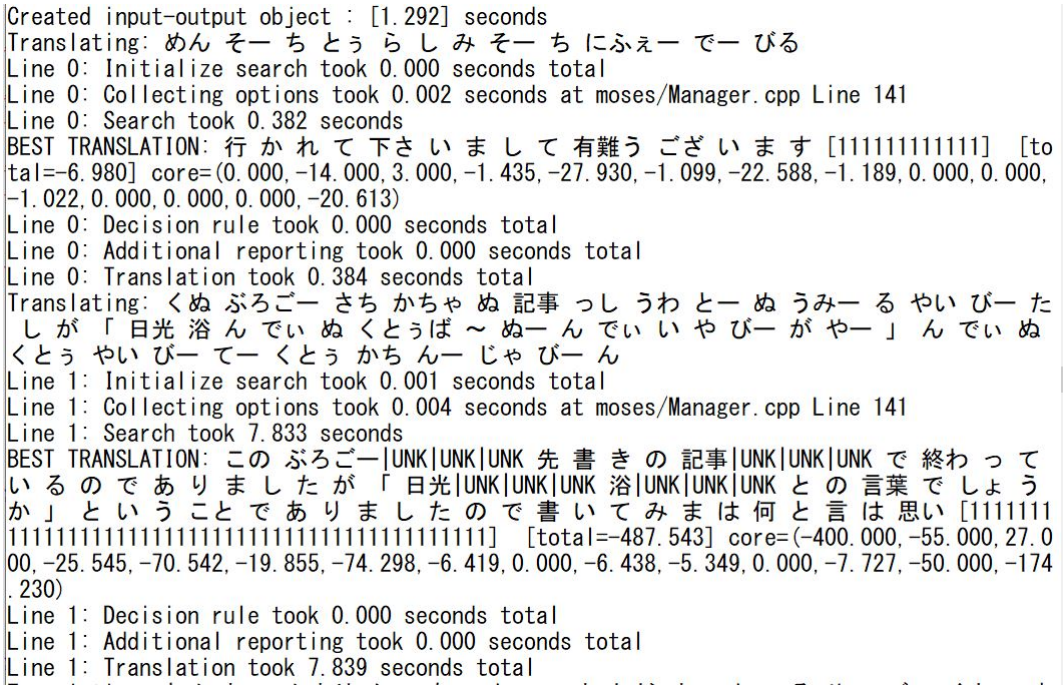

図 4.1: 翻訳結果の出力例

各手法のシステム翻訳の BLEU，RIBES による評価結果を表 4.2 に示す.「自動」は対訳コーパスの琉球方言の文の単語分割を形態素解析器 KyTea を用いて自動分割したとき,「人手」は人手で単語分割をしたときの結果を表す．次に，各手法のテスト文における未知語の数と割合を表 4.3 に示す．SMT では，訓練コーパス中に出現しない単語（未知語）が多いときに翻訳の性能が低くなることが知られているため，未知語数と未知語割合を調べた．未知語数は，図 4.1 に示したような翻訳出力文のうち，UNK タグの付いている単語数をカウントしたものである．未知語割合は，琉球方言のテストコーパスに含まれる未知語の割合を表し，式 (4.7) で定義される．

$$\text{未知語割合} = \frac{\text{未知語数}}{\text{テストコーパスの単語数}} \tag{4.7}$$

学習に用いた拡張対訳コーパス（no-ACG の場合は初期の対訳コーパス）の規模や個々の文の長さを調べるため，その単語数と平均文長を表 4.4 に示す．表中の「単語数」は拡張対訳コーパスに含まれる単語の数を表す.「平均文長」は拡張対訳コーパスの 1 文あたりの単語数を表し，式 (4.8) で定義される．

$$\text{平均文長} = \frac{\text{拡張対訳コーパスの単語数}}{m} \tag{4.8}$$

なお，テストコーパスの単語数は，自動分割のときは 2,468，人手分割のときは 2,528 であった．

表 4.2: 機械翻訳の評価結果

手法		m	BLEU		RIBES	
			自動	人手	自動	人手
no-ACG		1,002	35.93	37.92	77.62	80.53
ACG-doc2vec		500,000	28.08	29.82	69.95	71.00
ACG-random		2,000	35.32	36.85	76.94	78.30
		5,000	33.54	35.02	75.44	76.94
		10,000	32.29	33.72	74.75	76.40
		25,000	30.89	31.84	73.90	75.15
		50,000	30.68	32.01	74.10	74.14
		100,000	29.28	30.58	72.04	73.71
		200,000	28.12	29.69	70.96	72.87
		300,000	27.16	28.15	70.94	71.91
		500,000	28.85	29.60	69.22	68.94
OurACG-LM		500,000	20.84	21.77	64.75	66.40
OurACG-LM-diverse		2,201	36.53	37.85	77.36	80.16
		5,032	34.46	34.80	75.68	75.98
		11,032	33.55	33.96	74.75	76.50
		25,965	32.32	33.55	74.34	74.47
		50,960	31.30	32.79	71.72	74.62
		101,022	30.84	31.51	74.32	72.96
		201,256	28.51	30.04	70.96	72.36
		303,432	29.09	30.55	72.84	72.53
		502,054	27.73	29.71	69.76	71.35
OurACG-Dif		500,000	18.26	20.18	64.02	63.63
OurACG-Dif-diverse		2,201	36.07	38.09	76.98	80.84
		5,032	34.36	35.34	76.57	77.20
		11,032	33.57	34.64	77.53	78.55
		25,965	31.42	33.32	74.15	75.60
		50,960	31.07	32.68	72.94	73.18
		101,022	29.14	31.22	69.71	73.75
		206,422	30.97	32.57	74.44	75.76
		303,432	28.30	29.39	70.03	71.54
		502,053	27.64	30.13	69.10	71.65
OurACG	閾値 −100	137,753	25.48	25.82	70.28	71.87
-LM-random	閾値 −200	281,335	29.11	30.07	70.39	71.47
	閾値 −300	394,721	29.55	29.84	70.36	70.08

表 4.3: テストデータにおける未知語数・未知語割合

手法		m	未知語数		未知語割合	
			自動	人手	自動	人手
no-ACG		1,002	253	256	10.25	10.13
ACG-doc2vec		500,000	128	116	5.19	4.59
ACG-random		2,000	241	239	9.77	9.45
		5,000	234	222	9.47	8.79
		10,000	232	217	9.40	8.57
		25,000	213	197	8.65	7.78
		50,000	202	179	7.00	7.07
		100,000	200	173	8.10	6.84
		200,000	197	171	7.97	6.76
		300,000	190	167	7.69	6.60
		500,000	130	120	5.28	4.74
OurACG-LM		500,000	163	149	6.61	5.89
OurACG-LM-diverse		2,201	239	233	9.68	9.22
		5,032	239	239	9.68	9.45
		11,032	237	230	9.60	9.10
		25,965	239	228	9.68	9.02
		50,960	227	221	9.20	8.74
		101,022	215	206	8.71	8.15
		201,256	200	191	8.10	7.56
		303,432	194	175	7.86	6.92
		502,054	193	171	7.82	6.76
OurACG-Dif		500,000	156	141	6.32	5.58
OurACG-Dif-diverse		2,201	237	236	9.60	9.34
		5,032	244	239	9.89	9.45
		11,032	226	225	9.16	8.90
		25,965	222	209	9.00	8.27
		50,960	213	205	8.63	8.11
		101,022	211	196	8.55	7.75
		206,422	187	168	7.58	6.65
		303,432	197	176	7.98	6.96
		502,053	198	170	8.02	6.73
OurACG	閾値 −100	137,753	162	152	6.55	6.01
-LM-random	閾値 −200	281,335	141	132	5.72	5.23
	閾値 −300	394,721	135	125	5.48	4.93

表 4.4: 対訳コーパスの単語数・平均文長

手法		m	標準語		琉球方言	
			単語数	文長	単語数	文長
no-ACG		1,002	19,986	19.95	16,024	15.99
ACG-doc2vec		500,000	26,570,608	53.14	20,621,914	41.24
ACG-random		2,000	59,197	29.60	47,883	23.94
		5,000	176,600	35.32	143,411	28.68
		10,000	373,829	37.38	303,683	30.37
		25,000	963,031	38.52	782,992	31.32
		50,000	1,939,148	38.78	1,577,211	31.54
		100,000	3,911,761	39.12	3,180,966	31.81
		200,000	7,842,058	39.21	6,377,270	31.89
		300,000	11,760,954	39.20	9,561,346	31.87
		500,000	20,766,999	41.53	16,559,281	33.12
OurACG-LM		500,000	3,271,917	6.54	3,171,170	6.34
OurACG-LM-diverse		2,201	62,937	28.60	47,679	21.66
		5,032	186,290	37.02	141,038	28.03
		11,032	436,237	39.54	333,241	30.21
		25,965	1,049,205	40.41	812,384	31.29
		50,960	2,066,220	40.55	1,612,304	31.64
		101,022	4,074,692	40.34	3,213,105	31.81
		201,256	8,072,775	40.11	6,415,686	31.88
		303,432	12,145,957	40.03	9,679,494	31.90
		502,054	20,059,333	39.96	16,025,206	31.92
OurACG-Dif		500,000	4,095,036	8.19	3,822,281	7.64
OurACG-Dif-diverse		2,201	58,890	26.76	47,679	21.66
		5,032	172,840	34.35	141,038	28.03
		11,032	408,344	37.02	333,241	30.21
		25,965	996,832	38.39	812,384	31.29
		50,960	1,980,210	38.86	1,612,304	31.64
		101,022	3,948,798	39.09	3,213,104	31.81
		206,422	8,089,745	39.19	6,580,283	31.88
		303,432	11,900,892	39.22	9,679,500	31.90
		502,053	19,701,993	39.24	16,025,197	31.92
OurACG	閾値 −100	137,753	2,600,078	18.88	2,059,195	14.95
-LM-random	閾値 −200	281,335	7,429,534	26.41	5,965,273	21.20
	閾値 −300	394,721	12,858,549	32.58	10,176,570	25.78

まず，全体的に結果を見ると，単語を自動で分割した場合よりも人手で分割した方が，評価スコアは高くなり，未知語数・未知語割合は低下することがわかった．このことから，琉球方言の単語分割の誤りが機械翻訳の性能に影響を与えることがわかる．

次に，ACG ありと ACG なしの手法を比較する．提案手法のうち評価指標の値が良かった OurACG-LM-diverse と OurACG-Dif-diverse を ACG ありの手法として，ACG なしの手法 (no-ACG) と比較する．表 4.5 は，比較のためにこれらの手法の評価指標の値を抜粋したものである．この表では，それぞれの提案手法において最も評価指標の値が高くなったときの m を選択している．OurACG-LM-diverse は no-ACG と比べて，自動分割のときの BLEU が 0.6 ポイント向上したが，その他の評価指標のスコアについては低下した．また，OurACG-Dif-diverse は no-ACG と比べて，自動分割のときの RIBES だけが低下し，その他の評価指標のスコアは 3 つとも向上した．しかし，そのスコアの差はいずれも 1 ポイント以下であった．以上のことから，ACG を適用することで初期の対訳コーパスから文の量を増やすことができたが，翻訳性能の改善の度合は小さいことがわかった．

表 4.5: ACG の有無による比較

手法	m	BLEU		RIBES	
		自動	人手	自動	人手
no-ACG	1,002	35.93	37.92	77.62	80.53
OurACG-LM-diverse	2,201	36.53	37.85	77.36	80.16
OurACG-Dif-diverse	2,201	36.07	38.09	76.98	80.84

次に，実験結果を拡張対訳コーパスの品質，多様性，量の観点から考察する．

品質に関する考察

拡張対訳コーパスを構築する際に，確率言語モデルを用いた文の品質評価をするモデルとしないモデルとで翻訳性能の違いを比較する．まずは，品質評価を行っていない手法 (ACG-random) と品質評価を行っている手法 (ACG-LM-diverse，ACG-Dif-diverse) において，最も評価指標のスコアが高い m のときの結果を抜粋したものを表 4.6 に示す．ACG-LM-diverse，ACG-Dif-diverse は ACG-random と比べて，自動分割の場合は，BLEU が最大 1.21 ポイント，RIBES が最大 0.42 ポイント向上し，人手分割の場合は，BLEU が最大 1.24 ポイント，RIBES が最大 2.54 ポイント向上した．このことから，品質を考慮することによって，差は大きくないものの，翻訳性能が向上することがわかった．

次に，ランダム選択と確率言語モデルによる対訳文の品質評価を組み合わせた手法の効果を検証する．具体的には，ACG-random と OurACG-LM-random を比較

表 4.6: 対訳候補文の品質評価の有無による比較

手法	m	BLEU		RIBES	
		自動	人手	自動	人手
ACG-random	2,000	35.32	36.85	76.94	78.30
OurACG-LM-diverse	2,201	36.53	37.85	77.36	80.16
OurACG-Dif-diverse	2,201	36.07	38.09	76.98	80.84

する．表 4.7 はこの 2 つのシステムの評価結果の抜粋である．OurACG-LM-random の BLEU は，閾値 -200 では m がほぼ同数である ACG-random($m = 300,000$) よりも 2 ポイント程度高く，閾値 -300 のときも ACG-random($m = 500,000$) よりも 1 ポイント程度高いことがわかる．しかし，閾値 -100 のときは ACG-random($m = 100,000$) と比べて 4 ポイント程度低くなっている．BLEU が低くなった原因として，表 4.4 からわかるように，OurACG-LM-random(閾値 -100) の対訳コーパスは平均文長が他の手法と比べて短い．確率言語モデルでは短い文に高い生成確率を与える傾向があるため，短い文ばかりが選択されたことで，対訳コーパスの品質は向上しても多様性が失われ，翻訳性能が低下したと考えられる．以上のことから，単にランダムに対訳候補文を選択するよりも，文の品質のスコアが悪い文（閾値 -300 未満）を除くことで BLEU が向上するといえる．一方，m がほぼ同じときで RIBES を比較すると，ACG-random の方が全般的に良い結果が得られている．ただし，$m = 500,000$ の ACG-random と $m = 394,721$ の OurACG-LM-random の比較では，提案手法の方が RIBES の値が高い．

表 4.7: ランダム選択と対訳文の品質評価の組み合わせの評価

手法		m	BLEU		RIBES	
			自動	人手	自動	人手
ACG-random		100,000	29.28	30.58	72.04	73.71
		300,000	27.16	28.15	70.94	71.91
		500,000	28.85	29.60	69.22	68.94
OurACG	閾値 -100	137,753	25.48	25.82	70.28	71.87
-LM-random	閾値 -200	281,335	29.11	30.07	70.39	71.47
	閾値 -300	394,721	29.55	29.84	70.36	70.08

多様性に関する考察

拡張対訳コーパスを構築する際に対訳文の多様性を考慮する手法について翻訳

性能を比較する．まず，先行研究の提案手法 (ACG-doc2vec) と本研究の提案手法 (OurACG-LM-diverse，OurACG-Dif-diverse) を比較する．文の数が同じとき（$m = 500,000$）のこれらの手法の結果の抜粋を表 4.8 に示す．人手分割のとき，OurACG-Dif-diverse は BLEU，RIBES ともに ACG-doc2vec を上回ったことがわかる．これは文の品質と多様性を同時に考慮したことが翻訳性能の向上につながったと考えられる．

表 4.8: 対訳候補文の多様性を考慮する手法の比較

手法	m	BLEU		RIBES	
		自動	人手	自動	人手
ACG-doc2vec	500,000	28.08	29.82	69.95	71.00
OurACG-LM-diverse	502,054	27.73	29.71	69.76	71.35
OurACG-Dif-diverse	502,053	27.64	30.13	69.10	71.65

次に，多様性を考慮する手法としない手法とで翻訳性能を比較する．多様性を考慮せず品質のみを考慮した手法 (OurACG-LM，OurACG-Dif) と，品質と多様性の両方を考慮した手法 (OurACG-LM-diverse，OurACG-Dif-diverse) について，文の数が同じとき（$m = 500,000$）の結果を抜粋したものを表 4.9 に示す．OurACG-LM-diverse，OurACG-Dif-diverse は，OurACG-LM，OurACG-Dif よりも，BLEU は 7∼10 ポイント，RIBES は 5∼8 ポイント向上したことがわかった．このように評価指標のスコアが向上したのは，初期の対訳コーパスが持つ自然な文脈や単語を偏りなく学習に用いることができ，誤訳が少なくなったためであると考えられる．

表 4.9: 対訳候補文選択時の多様性の考慮の有無による比較

手法	m	BLEU		RIBES	
		自動	人手	自動	人手
OurACG-LM	500,000	20.84	21.77	64.75	66.40
OurACG-LM-diverse	502,054	27.73	29.71	69.76	71.35
OurACG-Dif	500,000	18.26	20.18	64.02	63.63
OurACG-Dif-diverse	502,053	27.64	30.13	69.10	71.65

具体例として，図 4.2 に OurACG-Dif と OurACG-Dif-diverse のテスト文の翻訳結果および正解文を示す．OurACG-Dif においては「？」マークが「でしょう」に誤訳され，OurACG-Dif-diverse では「？」のまま正しく翻訳されている．これは，OurACG-Dif で拡張した対訳コーパスには「？」が「でしょう」に対応している特殊な対訳文が多く含まれており，この対応関係が SMT モデルで強く学習されたた

めに，このような誤訳が発生したと考えられる．OurACG-Dif-diverse では拡張対訳コーパスの多様性を確保することで,「？」と「でしょう」の対応関係が一般的でないことを学習でき,「？」は「？」のまま正しく出力されたと考えられる．

別の具体例を図 4.3 に示す.「ちゅらじんちち」という文が OurACG-Dif では「美しい着物聞いて」に誤訳され，OurACG-Dif-diverse では「綺麗な服を着て」に正しく翻訳されている.「ちち」という琉球方言の単語は通常「着て」や「聞いて」といった標準語の単語に翻訳されるが,「着物」という単語の後に「聞いて」が続くのは不自然であり，この文においては「着て」に訳されるのが正しい．拡張対訳コーパスの多様性を確保することで，このような自然な単語の並びが学習されたといえる．

以上の結果より，拡張後の対訳コーパスの多様性を確保する提案手法のアプローチは有効であるといえる．

入力文　わん ねー んかし ん ちゅ ！ ？

OurACG-Dif　出力文　私 は 昔 の 人 ！|UNK|UNK|UNK で しょ う

OurACG-Dif-diverse　出力文　私 は 昔 の 人 ！|UNK|UNK|UNK ？

参照訳（正解訳）　私 は 昔 の 人 ！ ？

図 4.2: OurACG-Dif と OurACG-Dif-diverse の翻訳結果の例 (1)

入力文　ちゅら じん ちち あま はい くま はい さびー ん てー なー

OurACG-Dif　出力文　美し い 着物 聞 い て あっち 馳せ こっち 馳せ す わ よ

OurACG-Dif-diverse　出力文　綺麗 な 服 を 着 て あっち 馳せ こっち 馳せ し て い ま し た

参照訳（正解訳）　綺麗 な 服 を 着 て あっち 馳せ こっち 馳せ し ま す よ もう

図 4.3: OurACG-Dif と OurACG-Dif-diverse の翻訳結果の例 (2)

量に関する考察

対訳コーパスの量を変化させたときの翻訳性能の変化について考察する．ここでは，比較する手法として，no-ACG，ACG-random，OurACG-LM-diverse，OurACG-Dif-diverse を比較する．これらの結果の抜粋を表 4.10 に示す．また，各手法での対訳コーパスの量を変化させたときの評価指標の値の変化を折れ線グラフで表した図を，BLEU については図 4.4，RIBES については図 4.5 に示す．全ての手法において，$m \leq 100,000$ においては，文量を増加させると評価スコアは減少する傾向にあることがわかった．また，no-ACG の評価スコアが他の ACG 手法と比べて比較的高く，対訳コーパスの自動拡張によって評価スコアが向上したのは，OurACG-LM-diverse と OurACG-Dif-diverse の $m = 2,201$ のときのみであった．表 4.3 から，対訳コーパスの文の数を増やすことにより未知語は減少していることは確認できたが，翻訳性能の向上にはつながらなかった．この原因として，本研究では単語置換により対訳文候補を生成しているが，このように生成された候補文には自然なものが少ないため，選択する文の数を多くしても翻訳性能が向上しなかったと考えられる．このことから，SMT の学習に用いる対訳コーパスの量はただ多ければ良いものではなく，それよりも不自然な対訳文候補を取り除くことが重要であるといえる．

表 4.10: 対訳コーパスの量を変化させたときの評価指標の値の変化

手法	m	BLEU		RIBES	
		自動	人手	自動	人手
no-ACG	1,002	35.93	37.92	77.62	80.53
ACG-random	2,000	35.32	36.85	76.94	78.30
	5,000	33.54	35.02	75.44	76.94
	10,000	32.29	33.72	74.75	76.40
	25,000	30.89	31.84	73.90	75.15
	50,000	30.68	32.01	74.10	74.14
	100,000	29.28	30.58	72.04	73.71
	200,000	28.12	29.69	70.96	72.87
	300,000	27.16	28.15	70.94	71.91
	500,000	28.85	29.60	69.22	68.94
OurACG-LM-diverse	2,201	36.53	37.85	77.36	80.16
	5,032	34.46	34.80	75.68	75.98
	11,032	33.55	33.96	74.75	76.50
	25,965	32.32	33.55	74.34	74.47
	50,960	31.30	32.79	71.72	74.62
	101,022	30.84	31.51	74.32	72.96
	201,256	28.51	30.04	70.96	72.36
	303,432	29.09	30.55	72.84	72.53
	502,054	27.73	29.71	69.76	71.35
OurACG-Dif-diverse	2,201	36.07	38.09	76.98	80.84
	5,032	34.36	35.34	76.57	77.20
	11,032	33.57	34.64	77.53	78.55
	25,965	31.42	33.32	74.15	75.60
	50,960	31.07	32.68	72.94	73.18
	101,022	29.14	31.22	69.71	73.75
	206,422	30.97	32.57	74.44	75.76
	303,432	28.30	29.39	70.03	71.54
	502,053	27.64	30.13	69.10	71.65

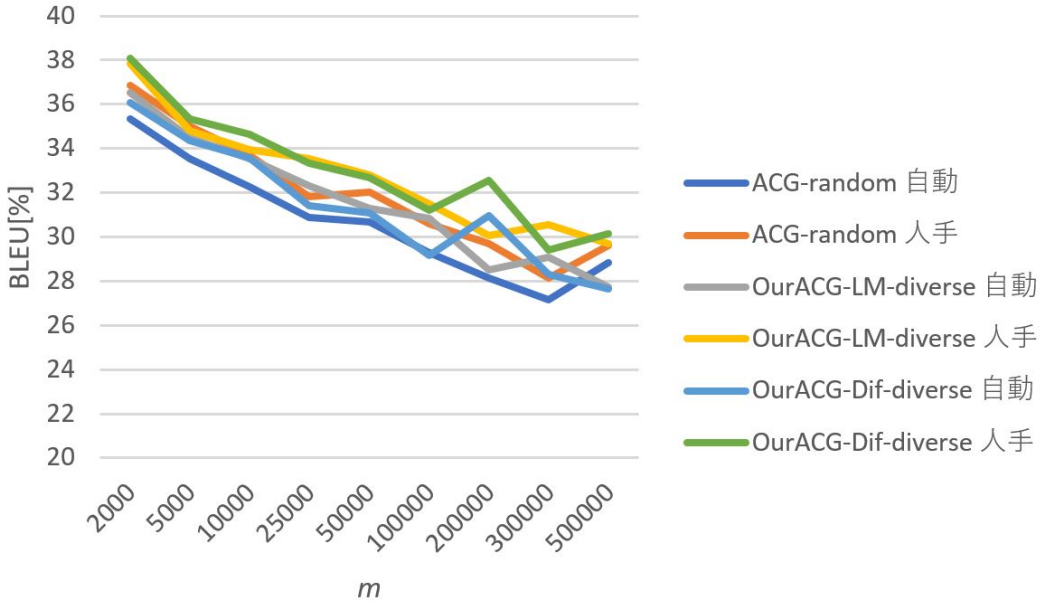

図 4.4: 対訳コーパスの量を変化させたときの BLEU の値の変化

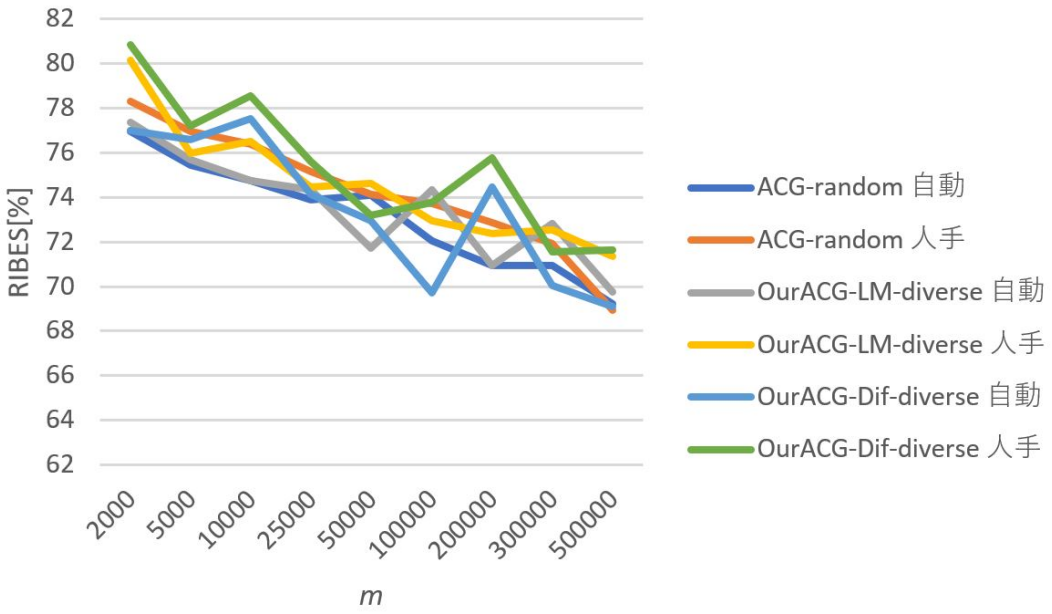

図 4.5: 対訳コーパスの量を変化させたときの RIBES の値の変化

第5章　おわりに

5.1　まとめ

本研究では，低言語資源の言語である琉球方言を機械翻訳の対象とし，琉球方言から日本語への統計的機械翻訳（琉日 SMT）における翻訳精度を向上させるため，対訳コーパスを自動的に拡張する手法を提案した．初期の対訳コーパスから，新しい対訳文を自動的に生成し，その中から (1) 品質，(2) 多様性，(3) 量の 3 つの観点を考慮して適切な対訳文を選別し，初期の対訳コーパスよりもサイズの大きい拡張対訳コーパスを構築した．この拡張対訳コーパスを用いて SMT のモデルを学習した．

品質を考慮した対訳文の選択とは，確率言語モデルにより対訳文（琉球方言と標準語の文の組）における標準語の文の生成確率を計算し，それが大きい対訳文を選択することである．また，自動生成する前のオリジナルの文と自動生成した後の文の生成確率の差によって対訳文の品質を評価する手法も提案した．評価実験では，ランダムに対訳文を選択する手法と比べて，品質を考慮することによって BLEU が最大 1.24 ポイント，RIBES が最大 2.54 ポイント向上した．また，単にランダムに対訳文を選択する場合と，ランダムに対訳文を選択して候補文の数を絞り込んでから確率言語モデルのスコアが閾値以上の対訳文を選択する手法を比べたところ，後者の BLEU が 2 ポイント程度向上した．しかし，閾値をあまりに高く設定して高品質の文だけを残すと翻訳精度が低下することがわかった．これは，確率言語モデルのスコアが高い文は短い文である傾向があるため，拡張対訳コーパスの単語数が少なくなり，一般的な機械翻訳モデルを学習するのに十分な語彙や文脈を含む対訳コーパスが得られなかったためと考えられる．

多様性を考慮した対訳文の選択とは，拡張対訳コーパスに様々な翻訳事例が含まれるように対訳文を選択する処理を指す．具体的には，初期の対訳コーパス中のすべての対訳文から同じ数の対訳文を生成することで，拡張対訳コーパスの多様性を確保する手法を提案した．これにより，初期の対訳コーパスに含まれる文脈や単語を偏りなく拡張対訳コーパスへ含めることができる．実験の結果，多様性を考慮しない手法と比べて，BLEU が 7~10 ポイント，RIBES は 5~8 ポイント向上した．

量を考慮した対訳文の選択とは，不自然な対訳文が過度に対訳コーパスに含まれるのを防ぐために，拡張対訳コーパスの文の数を調整する手法を指す．評価実験では，拡張対訳コーパスの文の数を 2,000~500,000 の範囲で変更し，翻訳精度を測定

した．その結果，文の数が多いほど BLEU もしくは RIBES は低下し，ACG を行わない手法より評価指標が高くなったのは文の数が 2,000 文のときだけであった．

結論として，本論文で提案する品質ならびに多様性を考慮した対訳コーパスの自動拡張手法は翻訳性能の向上に貢献すること，自動拡張後の対訳コーパスの量は適切に決定する必要があることが確認された．

5.2 今後の課題

本研究で提案した対訳コーパスの拡張手法は，ベースラインと比べて BLEU や RIBES が改善したものの，その差はあまり大きくなく，効果は限定的であった．この原因として，本研究では単語のランダムな置換により対訳文候補を生成しているため，不自然な候補文が多く，自然な候補文が選択されにくかったことが考えられる．ナイーブな方法で対訳文候補を大量に生成してから自然な候補文を選別するよりも，候補文を生成する時点で自然な文を生成してコーパスを拡張する方が，翻訳精度の向上につながると予想される．そのため，今後は，単語置換によるナイーブな手法ではなく，言い換え技術などを用いて元の文を自然な文に置き換えた上で対訳文候補を生成し，対訳コーパスを拡張する手法を探究したい．

参考文献

[1] 藤原菜々美，今出昌宏，山内真樹，内山将夫，隅田英一郎．自動コーパス生成とユーザフィードバックによる機械翻訳，言語処理学会第 23 回年次大会，pp.569–572 (2017)

[2] Qin Gao and Stephan Vogel. Corpus Expansion for Statistical Machine Translation with Semantic Role Label Substitution Rules, *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics(ACL): shortpapers*, pp.294–298 (2011)

[3] 平尾努，磯崎秀樹，須藤克仁，Kevin Duh，塚田元，永田昌明．語順の相関に基づく機械翻訳の自動評価法，自然言語処理 Vol.21, No.3，pp421–444 (2014)

[4] Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German, *In Proceedings of LREC*, pp.3781–3788 (2018)

[5] Ann Irvine and Chris Callison-Burch. Combining Bilingual and Comparable Corpora for Low Resource Machine Translation, *In Proceedings of the Eighth Workshop on Statistical Machine Translation of the Association for Computational Linguistics(ACL)*, pp.262–270 (2013)

[6] Kendall, M. G. *Rank Correlation Methods*, Charles Griffin (1975)

[7] 久高優也，金城伊智子．琉日間 SMT における doc2vec を用いた ACG 手法の検討，電子情報通信学会総合大会，D-5-3 (2018)

[8] Quoc V. Le and Tomas Mikolov. Distributed Representations of Sentences and Documents, *In Proceedings of the 31st International Conference on Machine Learning*, pp.1188–1196 (2014)

[9] 永田昌明，渡辺太郎，塚田元．機械翻訳最新事情：(上) 統計的機械翻訳入門，情報処理 Vol.49, No.1，pp.89–95 (2008)

[10] Nguyen Le An, 松本啓之亮，森直樹．日本語–ベトナム語翻訳における考察，言語処理学会第 21 回年次大会，pp.808–811 (2015)

[11] 西岡敏，仲原穣，伊狩典子，中島由美．沖縄語の入門 たのしいウチナーグチ，白水社 (2000)

[12] 沖縄県立図書館．琉球方言を調べる 沖縄県立図書館調べ方案内 No.7.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation, *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics(ACL)*, pp.311–318 (2002)

[14] 柴田直由，横山晶一，井上雅史．統計的手法を用いた双方向方言機械翻訳システム，言語処理学会 第 19 回年次大会 発表論文集，pp.311–318 (2013)

[15] 沖縄方言であれこれ，http://buutoria.blog110.fc2.com/

[16] 琉球語音声データベース，http://ryukyu-lang.lib.u-ryukyu.ac.jp/

[17] GIZA++, http://www.statmt.org/moses/giza/GIZA++.html

[18] KenLM, https://kheafield.com/code/kenlm/

[19] KyTea, http://www.phontron.com/kytea/index-ja.html

[20] MeCab, http://taku910.github.io/mecab/

[21] MOSES, http://www.statmt.org/moses/

[22] SRILM, http://www.speech.sri.com/projects/srilm/