テトゥン語を対象としたニューラル機械翻訳の研究

松元 航太郎　　速水 悟　　田村 哲嗣
岐阜大学 大学院 自然科学技術研究科
{matsumoto, hayamizu, tamura}@asr.info.gifu-u.ac.jp

1 はじめに

訪日外国人の増加などを背景に，共通言語を用いてコミュニケーションがとれない機会が増加している．このような状況では，ある言語の文書を別の言語の文書に自動的に翻訳する機械翻訳が有用である．特に，翻訳モデルにニューラルネットワークを用いるニューラル機械翻訳は，その翻訳精度に注目が集まっている．ニューラル機械翻訳モデルの構築には，同一内容を記したパラレルデータを大量に用意する必要がある．日英翻訳のような母語話者の多い言語同士では，大量のパラレルデータを用意し精度の高いモデルを学習できる．しかし，母語話者の少ない言語を対象とした場合，利用できるデータが少なく，そのままではニューラル機械翻訳モデルの構築は難しい．

本研究は，東南アジアの小国である東ティモールで用いられるテトゥン語を対象とする．本稿では，ノンパラレルコーパスから学習データを抽出する手法と，英葡翻訳を用いた事前学習の 2 手法により，テトゥン語-英語翻訳の精度向上を図る．また同条件における英仏翻訳と精度比較を通じて，学習手法の妥当性評価を行う．

2 テトゥン語について

2.1 特徴

テトゥン語は東ティモールで話されている公用語の一つである．東ティモールとはオーストラリア大陸の北，ティモール島の東に位置する国であり，人口は 120 万人ほどである．この言語は東ティモールの特有の言語であるが，同国がポルトガルの植民地であったことから，ポルトガル語の影響を受けている．これは東ティモールがポルトガルにより植民地化された後，ポルトガル語が借用語としてテトゥン語に追加されたという時代背景によるものである．テトゥン語は首都ディリ周辺で話されている dili 方言と西ティモール国境周辺で話されている prasa 方言が存在する．本研究では東ティモールで幅広く話されており，話者人口が最も多い dili 方言を対象とする．この母語話者数は約 40 万人である．

テトゥン語と英語の訳例を以下に示す．

(テトゥン語)Nia servisu iha kompania.
↑↓
(英語) He works at a company.

テトゥン語の特徴として以下のようなものが挙げられる．

- 基本的な語順はテトゥン語と英語の語順は同じ［主語 + 動詞 +目的語］である
- 動詞が活用しない
- 時制が存在しない
- 英語と同じアルファベットが用いられている

2.2 学習データ

テトゥン語はマイナー言語であるため，一般にニューラル機械翻訳に必要なパラレルコーパスが不足しているという問題がある．JICA や UNICEF は同じ内容をテトゥン語と英語両方で記述している文書 [1] [2] を公開しているが，文同士で対応が取れていないノンパラレルコーパスであるため，同期を取る必要がある．表 1，表 2 にテトゥン語と英語翻訳において利用可能な学習データとデータ数の詳細を示す．

表 1: 利用したパラレルコーパス

データ名	文数
新約聖書 (トレーニング用)	6369 文対
新約聖書 (テスト用)	1000 文対
人手によるテ-英ペア (テスト用)	999 文対

表 2: 利用したノンパラレルコーパス

データ名	テトゥン語文数	英語文数
UNICEF	4959 文	4579 文
CADEFEST	644 文	586 文
データ名	テトゥン語文書数	英語文書数
卒業論文 *1	84 文書	84 文書
Facebook*2	164 文書	164 文書

*1 東ティモール国立大学の学生によって執筆された卒業論文の概要部分のテトゥン語と英語の文書単位のペア

*2 JICA Timor-Leste[1] によって Facebook に投稿されたテトゥン語と英語の文書単位のペア

3 提案手法

ノンパラレルコーパスから学習データを抽出する手法を説明する.

3.1 DTW を用いたノンパラレルコーパスのアライメント手法

本研究では声質変換などで用いられる DTW(Dynamic Time Warping)[3] により, 文書から文と文の同期を取る手法を提案する. 手順を以下に示す.

1. 文書をテキスト化し, 文末表現 (ピリオドかエクスクラメーションマーク) で分割を行う. その際に文末表現+(スペース) とすることで"Mr."のような略字表現や目次 (.....3) のような文末でないピリオドを除外することができる.

2. 文の文長をデータを表す特徴量として変換する. その際, 空文字は削除する.

3. 文の文長をリストに変換する. これにより, テトゥン語と英語の文書は 2 つの系列データで表される (図 1). 縦軸は文長, 横軸は文番号を示す. この 2 つの系列データを用いて DTW の計算を行う. テトゥン語と英語をそれぞれ m 次元,n 次元のリストで表す. x_i はテトゥン語文書の i 番目の文の文長, y_j は英語文書の j 番目の文の文長である.

$$x = [x_0, x_1, ..., x_i, ..., x_m] \tag{1}$$

[1]https://ja-jp.facebook.com/JICATimorLeste/

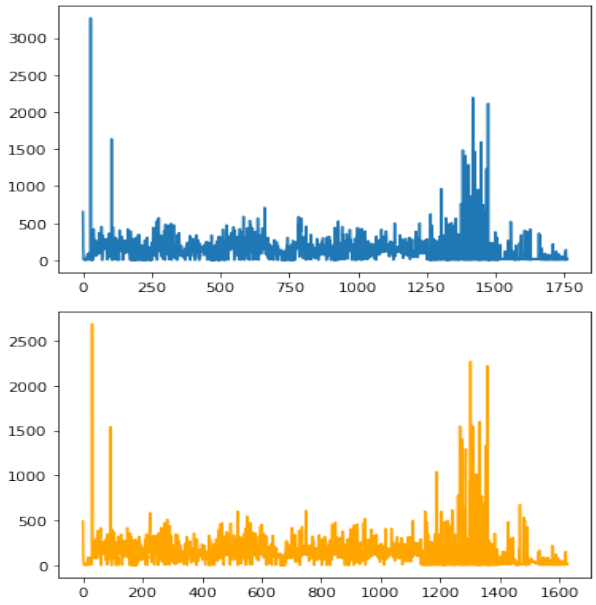

図 1: UNICEF 文書を系列データに変換した結果 (上図:テトゥン語, 下図:英語)

$$y = [y_0, y_1, ..., y_j, ..., y_n] \tag{2}$$

次に $m \times n$ の大きさの距離行列 DTW を作成する. 本研究では $d|x_i, y_j|$ は 2 乗誤差とした.

$$DTW(x_i, y_j) = d|x_i, y_j| \tag{3}$$

動的計画法により $DTW[0][0]$ から $DTW[m-1][n-1]$ への最短のパスを計算する (図 2). 得られたパス同士の番号に紐付いた文同士をパラレルコーパスに追加する. この手法を表 2 で示した 4 種類のノンパラレルコーパスに適用する. ノンパラレルコーパスのうち卒業論文と Facebook のデータは文書数が一致しているため, 文書ごとに同期をとった後, 文書順に結合する

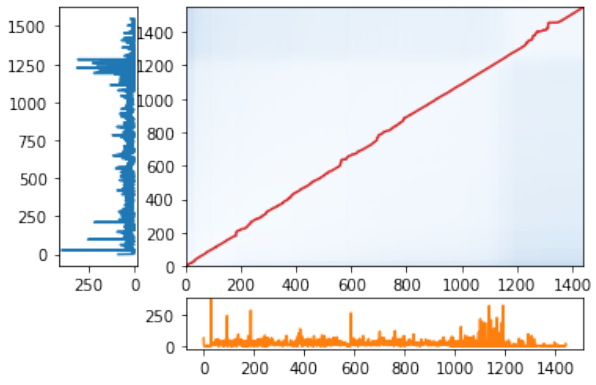

図 2: DTW により得られた文同士のパス

ことで学習データを作成する. 得られた学習データ数を表 3 に示す.

表 3: DTW により得られた学習データ数

データ名	データ数 (文数)
UNICEF	1822
CADEFEST	232
卒業論文 (概要)	773
Facebook	884

3.2 予備実験:ポルトガル語彙の割合の調査

テトゥン語が借用している言語の中でポルトガル語は対英語の学習データが多く存在する言語である．テトゥン語の文にポルトガル語由来の単語が多く含まれれば，事前学習が効果があるという仮定のもと，以下の手順でポルトガル語彙の割合を計算した.

1. ポルトガル語コーパス [4] をテキスト化し word2vec で学習する.

2. 学習したモデルから語彙を取得する.

3. 語彙をポルトガル語リストに変換する.

4. テトゥン語聖書データに含まれる単語のうち，ポルトガル語彙リストに含まれる単語を異なり語数でカウントする.

数字や固有名詞などポルトガル語に限らない単語までカウントしてしまうが，本実験では考慮しない．実験で得られた分析結果を表 4 に示す．テトゥン語がポルトガル語に影響されていることが確認できた.

表 4: テトゥン語文書におけるポルトガル語単語の出現回数

言語名	単語数 (割合)	単語例
ポルトガル語	238 単語 (13%)	mata,ida,sei
その他	1609 単語 (87%)	karik,atu,faru

4 評価実験

4.1 学習データ増強による効果

本実験では，提案手法の有効性を確かめるため学習データを増やさないベースモデルとの比較を行った．モデルは Open-NMT [5] の seq2seq モデルを用いた．学習パラメータは表 5 に示される．事前学習として埋め込み層は word2vec で得られる単語ベクトル 100 次元を用いた．注意機構はグローバルアテンション，評価指標として BLEU スコアを用いた．open テストデータとして学習に用いない聖書 1000 文 (インドメイン) と一般文 1000 文 (アウトドメイン) の 2 つを用意し，翻訳を行った．比較対象として WMT'14 [4]europarl コーパスの英仏の学習データを 10000,50000,100000,200000 文と変化させたときの英仏翻訳と精度比較を行った.

4.2 事前学習による効果

ポルトガル語-英語の学習データを事前に学習することによる翻訳精度向上の有効性を検証するため，事前学習を行わないベースモデルと事前学習を行う手法で比較を行った.

WMT'14 [4] のタスクにおいて europarl コーパスのポルトガル語-英語のデータからテトゥン語内に存在するポルトガル語を 1 語でも含むパラレルコーパスを 30000 文抽出し事前学習を行った．事前学習と本学習ではテトゥン語とポルトガル語を合わせて学習した word2vec の埋め込みベクトルを共有する．学習回数は 30000 とした．その他のパラメータは 4.1 節と同様である.

表 5: 使用した学習パラメータ

隠れ層	400 次元
テトゥン語語彙数	3534 単語
英語語彙数	12191 単語
学習回数	10000
ビーム幅	10

4.3 結果と考察

データ増強をしていない場合とした場合の 2 通りと事前学習をしていない場合とした場合の 2 通りの組み合わせ 4 通りの学習条件において表 6，表 7 にインドメインとアウトドメインの翻訳結果を示す．両ドメインともに学習データを増やすと翻訳精度が改善された.

図 3 にアウトドメインで最も精度の良かった各文の BLEU スコアの変化を示す．翻訳精度が上昇したデータは 668 件存在したが，ベースライン手法，提案手法

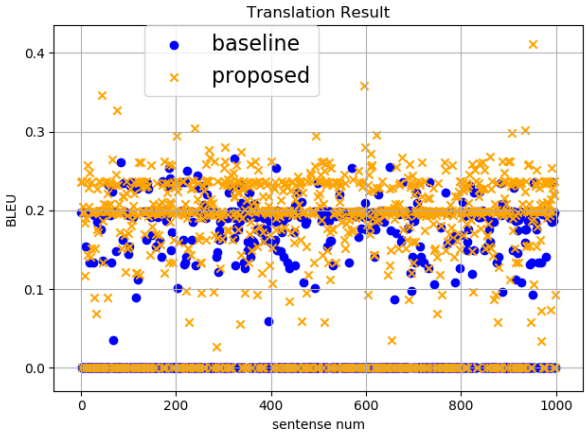

図 3: 翻訳結果の比較 (青 (丸印):ベースライン, 橙 (ばつ印):提案手法 (データ拡張+事前学習))

表 6: インドメインにおける翻訳精度 (BLEU)

手法	事前学習なし	事前学習あり
ベースライン	0.1328	0.1755
学習データの増強	0.1662	0.1843

表 7: アウトドメインにおける翻訳精度 (BLEU)

手法	事前学習なし	事前学習あり
ベースライン	0.0736	0.1170
学習データの増強	0.0824	0.1246

ともに BLEU スコアが 0 のテストデータが 164 件存在した．そのため，subword による未知語への対策が考えられる．

表 8 にアウトドメインで学習データを増強し，事前学習を行った翻訳例を示す．一致する単語は増えたが，翻訳結果としてはまだ難がある．

英仏機械翻訳のタスクとの比較を表 9 に示す．本研究のインドメインでの翻訳はでは学習データ数 6000 と少数であるが英仏翻訳の学習データ 50000 件と 100000 件の間に相当することを確認した．アウトドメインでの翻訳では同様の学習データ数である英仏翻訳と同等の翻訳精度であることを確認した．

5 おわりに

本研究では学習データが少ないマイナー言語と英語機械翻訳のタスクにおいて DTW を用いたデータ増強

表 8: アウトドメインにおける翻訳出力の比較

入力文	favor le manual ne'e ho huidadu
正解文	**please** read the **manual** carefully
ベースライン	the sower soweth the word
提案手法	**please** enjoy **manual** ne'e and **manual manual**

表 9: 英仏翻訳 (アウトドメイン) における学習データ (文数) と翻訳精度の比較

学習データ数	BLEU
10000	0.07980
50000	0.16401
100000	0.19401
200000	0.19665

手法を提案した．そしてポルトガル語-英語の学習データによる事前学習により BLEU スコアが改善することを確認した．しかし，メジャー言語と比較すると翻訳精度に未だ問題があることが分かった．

今後はポルトガル語以外のインドネシア語やマレー語借用語の情報を利用した事前学習により，さらなる翻訳精度向上を目指す．

謝辞

テストデータ作成に協力して頂いた岐阜大学工学部深井英和先生，Fernao 氏に感謝する．データの提供元に感謝する．

参考文献

[1] National Action Plan for Children in Timor-Leste 2016-2020 https://www.unicef.org/timorleste/reports/national-action-plan-children-timor-leste-2016-2020

[2] 独立行政法人 国際協力機構 事業協力 第 4 号「Cadefest Newletter(英文)」 https://www.jica.go.jp/project/easttimor/007/newsletter/index.html

[3] H. Sakoe and S. Chiba "A dynamic programming approach to continuous speech recognition". In Proc. 7th International Congress on Acoustics, pp. 6569, 1971.

[4] ACL 2014 NINTH WORKSHOP ON STATISTICAL MACHINE TRANSLATION http://www.statmt.org/wmt14/translation-task.html

[5] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A.M. Rush " OpenNMT: Open-source toolkit for neural machine translation " ArXiv e-prints, p.1,2017.