ニューラルネットを用いた多方言の翻訳と類型分析

阿部 香央莉 [1]　松林 優一郎 [1]　岡崎 直観 [2]　乾 健太郎 [1, 3]
[1] 東北大学　[2] 東京工業大学　[3] 理研 AIP
{abe-k,y-matsu,inui}@ecei.tohoku.ac.jp, okazaki@c.titech.ac.jp

1 はじめに

日本語方言を共通語へ翻訳するための試みは，言語処理分野の中で度々行われているが，その多くは方言に関する言語資源が少ないという問題に直面する．一般に，機械翻訳では大量の対訳コーパスが必要である．そこで，小規模な方言・共通語対訳コーパスから変換ルールを事前に抽出しておき，方言文と共通語文の大規模な疑似対訳コーパスを獲得する研究が発表されている [3, 6, 9]．しかし，これらの研究では特定の地域の方言のみに限定して対訳コーパスを作成しており，それ以外の地域の方言に対応するためには新たな対訳コーパスを作成する必要がある．方言コーパスの収集手法としては，各方言の特徴語を検索クエリとして用い，方言を含んだ文書を Web から検索・抽出する研究 [7] があるが，翻訳システムを構築するためには共通語文の対訳が必要であり，それを人手で作るとなると高いコストが必要である．

本研究では，一般的な文字レベルのニューラル機械翻訳（NMT）システムに対して 2 つの工夫を加えることで，小規模な方言対訳コーパスを利用する場合でも方言翻訳の精度を向上できることを示す．第 1 に，我々は複数方言の対訳コーパスを同時に学習に利用し，多方言を翻訳する単一の NMT モデルを構築する．一般に，方言間には語彙や音韻変化の特徴が共通するものがあることが知られている．複数方言間で共有された翻訳モデルを用いることで，方言間の共通性が自動的に考慮され，データ量不足の問題が軽減される．さらに，多言語 NMT の手法 [1] に基づき，方言が話される地域を表す地域ラベルを入力系列の先頭に追加することで，方言ごとの特徴をとらえた翻訳を目指す．第 2 に，文単位の翻訳の代わりに文節単位での逐語翻訳を行うことで，言語モデルの学習におけるデータ疎問題を軽減し，翻訳精度を向上させる．本研究では，日本語の方言翻訳では方言文と共通語文の間で語順変化は起こらないと仮定し，1 文節ごとに翻訳する．実験では，これらのアイデアを導入することで，ベースラインとなる NMT モデルよりも翻訳精度が向上することを示す．

また，複数の方言を単一のモデルで翻訳するモデルを構築することの恩恵の一つとして，第 1 の手法で導入した地域ラベルの埋め込み表現（ベクトル）を比較することで，方言の類型分析など方言研究への応用の可能性を示す．提案モデルによって得られた地域ラベルの埋め込み表現を可視化した結果，各地域に対応するベクトルの位置関係は概ね東北・九州などの地理的な地方区分にまとまっており，また東北地方と九州地方の方言が近接するなど，既存の方言分布に関する様々な知見との興味深い一致がみられた．

方言文：
「ニャ　ソエガラ　アレー　オスロネ　エデ　スキー　ヨグ　ノレスタデバー　タイカイマデ　ヤエスタデバ。」

共通語文：
「いや　それから　あれ　お城［＝弘前城］に　行って　スキー［を］　よく　乗った［＝した］ではないですか　大会まで　やったではないですか。」

ひらがな化

方言文：
「にゃ　そえがら　あれー　おすろね　えで　すきー　よぐ　のれすたでばー　たいかいまで　やえすたでば。」

共通語文：
「いや　それから　あれ　おしろに　いって　すきー　よく　のったではないですか　たいかいまで　やったではないですか。」

図 1　全国方言談話データベースオリジナルの例文, および前処理後の例文

2 全国方言談話データベース

本研究では，国立国語研究所が刊行する「全国方言談話データベース ふるさとことば集成」に収録された，2 人の方言話者間の対話を書き起こしたテキストを学習・評価用コーパスとして利用する．このデータには，48 地域（47 都道府県，沖縄県のみ本土と離島の 2 地域）の方言による対話とその共通語訳が収録されている．

本コーパスでは，各地域の方言研究者が方言話者の発話を表音的カタカナ表記[*1]で文字に書き起こした文（便宜上，方言文と呼ぶ）と，その共通語訳文（漢字かな混じり文）が追加されている（図 1 上）．方言・共通語訳文はそれぞれ分かち書きがされているが，この分かち書きは便宜的なものであり，厳密な単語・形態素区切りではない．また，できるだけ発話に忠実に書き起こされているため，方言文は感嘆詞や終助詞を多く含む．

[*1] 長音は「ー」，助詞「は」「を」「へ」は「ワ」「オ」「エ」と表される．東北方言等におけるか行の鼻濁音は「カ゜」のように表される．また，「時には」を「トキニァ」と表記するなど．「ァ」や「ェ」等の文字も使用されている．

このコーパスは，もともと方言学研究のために作成されたため，可読性を高くするための様々な注釈が括弧書きで付与されている[*2]．本研究で学習・評価用データとして用いる際は，これらの注釈を削除し，実際の発話部分のみを採用した．

3　ニューラル多方言翻訳器

3.1　ひらがな文からひらがな文への方言翻訳

全国方言談話データベースのように，方言研究では発話音声に忠実な表記が重要視されるため，発話の記録にカナ表記を利用することが多い．日本語への機械翻訳タスクは漢字表記への翻訳として取り組むことが一般的であるが，方言のコーパスをそのまま学習に用いると，かな表記の方言から漢字表記の標準語に翻訳するというタスクになる．これは，かな表記から漢字への変換や，かな表記による語義曖昧性の解消など，方言翻訳以外の問題を持ち込んでしまう．そこで，本研究では方言から共通語への翻訳のタスクを，ひらがな表記の方言文からひらがな表記の共通語文への翻訳タスクとして定義する．このように定義することで，方言・共通語間の音韻的な特徴をモデル化，分析できるといった利点もある．

本研究で利用するコーパスは，方言文がカタカナ，共通語訳文が漢字かな混じり文で記述されているため，これらをひらがな表記へと変換する（図 1）．方言文をひらがなに変換する際は日本語のかな文字変換モジュール cnvk[*3]を用い，共通語訳文をひらがなに変換する際は形態素解析器 MeCab を用いた．

3.2　文節単位の文字レベル方言翻訳

本研究では，方言・共通語間の音韻的な変化をとらえることを念頭に，一般的な NMT システム [2] に対して，1 文字を系列の 1 要素とし，1 文字ずつ単語を翻訳する文字レベル翻訳を採用する．このアプローチは，Twitter 上の崩れた日本語表記を正規化する研究など，音韻の変化を捉えるモデルとして成功が報告されている [3]．

通常，NMT では 1 文を入力系列とする（図 2（a））．しかし，英日翻訳など語順や文化圏が異なる言語間の翻訳に比べ，方言翻訳は入出力系列の間で大きな語順変化が生じることは少なく，訳し分けの問題が発生する事例も少ないと推測される．そこで，本研究では翻訳時に入力文全体の文脈を考慮する必要がないと考え，文全体を入力系列とするかわりに，文を文節ごとに区切り，1 文節を 1 入力系列として文節単位での翻訳を繰り返すことで文全体の翻訳を行う（図 2（b））．

また，方言間には語彙や音韻変化の特徴が共通しているものがある（例えば，東北圏のズーズー弁，関西圏の助動詞「や」など）．そこで，各地域ごとに翻訳モデルを構築するのではなく，全地域の翻訳を一つのモデルで同時に学習し，日本語としての共通の特徴を捉えることを試みる．本研究では，多言語翻訳における Johnson らの手法 [1] を参考に，翻訳している方言の地域を区別するためのラベルを入力系列と出力系列の先頭に追加することで，地域ごとの方言の特徴をモデルに学習させる[*4]．

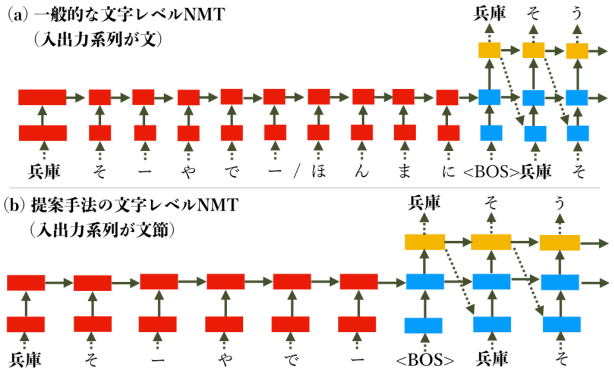

図 2　地域ラベルを追加したニューラル多方言翻訳器
（attention 機構は省略している.）

4　実験

実験では，文節単位で翻訳することの効果と地域ラベルの効果を定量的，定性的に検証する．また一般的な統計的機械翻訳（SMT）システムとの性能の比較を行う．比較のため，入力系列を 1 文とした NMT モデルと，入力系列を 1 文節とした NMT モデル，地域ラベルを入れない場合の NMT モデルおよび入力系列を 1 文節とした SMT モデルを作成した．

4.1　実験設定

「全国方言談話データベース ふるさとことば集成」の全地域の書き起こし文に対して前処理を行い，得られた計 34,117 文（116,928 文節）を実験に使用した．ここで，文節の単位にはコーパスであらかじめ与えられる分かち書きの区切りを利用した[*5]．このコーパスを 8 : 1 : 1 の割合で分割し，訓練・開発・評価セットを作成した．NMT システムには OpenNMT-py[*6]をデフォルト設定で使用し，SMT システムには Moses[*7]を使用した．Moses における共通語訳言語モデルは KenLM で学習したものを用い，distortion limit は 0 に設定した．

4.2　定量評価

表 1 に，評価セット上の各モデルの BLEU スコアを示す．文節単位で翻訳することの効果に着目すると，文節を入力系列とした NMT は，文を入力系列とした NMT

[*2] 発話中に出現する固有名詞，方言語彙等に対する解説など．
[*3] https://github.com/yuka2py/cnvk
[*4] Johnson らは出力の先頭に地域ラベルを追加していないが，本研究では予備実験において精度の向上を確認したため，出力系列の先頭にも地域ラベルを追加した．
[*5] この分かち書きは，正確にはコーパス上では文節とは呼ばれていないが，ここでは便宜上これを文節の区切りとして利用する．
[*6] https://github.com/OpenNMT/OpenNMT-py
[*7] http://www.statmt.org/moses/

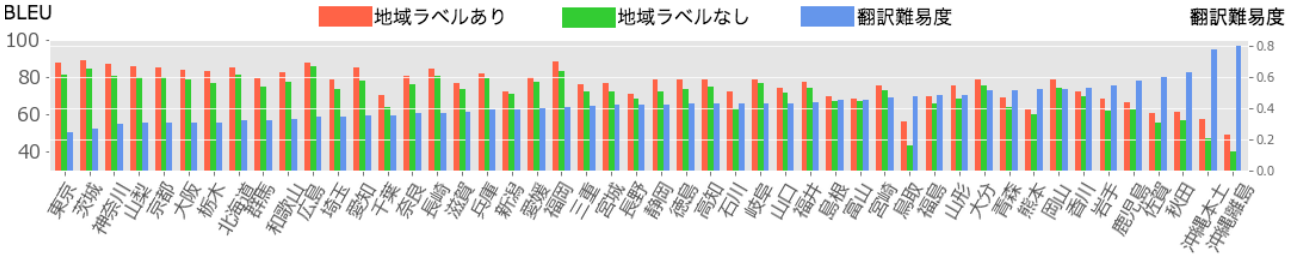

図 3　文節を入力系列とした NMT における各地域の BLEU スコアと翻訳難易度．地域は左から翻訳難易度の昇順で列挙．

表 1　各方言翻訳器の BLEU スコア

地域ラベル	入力系列	MT の種類	BLEU
なし	文節	NMT	72.66
あり	文	NMT	67.77
あり	文節	NMT	**77.10**
なし	文節	SMT	74.74

よりも，9.33 ポイント BLEU スコアが高かった．このことから，方言から共通語への翻訳など，大きい語順変化がなく基礎的な語彙を共有する言語間の翻訳では，文を入力系列とした翻訳よりも文節を入力系列とした逐語翻訳のほうが高い性能を実現することが分かった．

次に，地域ラベルの効果に着目する．文節を入力系列とした NMT モデルにおいて，地域ラベルを先頭に追加したモデルは追加しないモデルに比べて BLEU スコアが 4.44 ポイント向上した．地域ラベルの導入により，方言間のデータを共有した単一の多言語翻訳モデルにおいても効果的な学習が行われていることが分かる．

図 3 には各地域の方言ごとの翻訳難易度と BLEU スコアを示した．ここで，翻訳難易度は，ある地域の方言文と共通語文の間の正規化レーベンシュタイン距離の平均と定義した．各地域の BLEU スコアは翻訳難易度と強い負の相関 (相関係数 −0.85) があり，岩手県や沖縄県など，訛りが強い地域ほど方言翻訳を行うことが難しいことが定量的に読み取れる．また，地域ラベルを追加することで全地域で BLEU スコアの増加が確認できた．さらに，地域ラベルを追加しなかった場合の BLEU スコアと，地域ラベルを追加したことによる BLEU スコアの差分にも負の相関（相関係数 −0.49）があり，翻訳精度が低かった地域ほど，地域ラベルを追加したことによってより精度が向上した．

一般に，学習コーパスが小規模の場合には NMT より SMT の方が優れた性能になると報告されている [4]．そこで，提案手法と SMT の比較も行った．文節のみを入力系列とした SMT と比べて，同条件の NMT は性能が劣るものの，SMT は地域ラベルに相当する情報を導入する手段が自明ではなく，結果として地域の情報を利用できる文節単位の NMT が最高性能を示す結果となった．

4.3　定性評価

表 2 に各モデルによる実際の翻訳例を示す．例 1，2，4 はそれぞれ兵庫方言文，青森方言文，沖縄本土方言文の翻訳例で，参照文と非常に近い翻訳文が出力された例である．単一のモデルで多言語の翻訳に対応する提案モデルは，関西弁・東北弁・琉球弁という特徴が異なる 3 種類の方言文を正しく翻訳できた．

青森方言文について各モデルの特徴を詳しく見ると，地域ラベルと文を入力系列とした NMT では，「つだりすて」を「ついて」と訳してしまい，「ついたりして」という話し言葉特有の表現をうまく翻訳できていない．地域ラベルなしの文節を入力系列とした NMT では，「ついたり」が母音の欠落，濁音化によって変化した「つだり」を翻訳できていない．SMT では青森地域の方言特有である名詞に「こ」という接尾辞がつく傾向を学習できていない[*8]．しかし，地域ラベル＋文節を入力系列とした NMT はこれらの問題を全て解決し，完全に参照文と一致した翻訳文を出力した．

沖縄方言は共通語と大きく異なる語彙を使用するなど，非常に翻訳が難しい方言の 1 つである（翻訳難易度は 0.8 に近い）．ところが，例 4 では提案手法の NMT が完全に参照文と同じ内容を出力している．地域ラベルを追加したことにより，他の方言と区別して沖縄特有の表現を学習することができ，他の方言の翻訳精度を落とすことなく沖縄方言の翻訳精度を大幅に向上させたと考えられる．

例 3 に，提案手法で正しく翻訳できなかった例を示す．この岩手方言文の例では，提案モデルが東北弁の特徴の一つである濁音化や，「い」→「え」のような音韻変化を正しくとらえていることが観測できるものの，「かか」が「妻」という意味を表す単語であることは捉えきれていない．今回の学習データの規模では，このような語彙的な知識を十分に学習できなかったと考えられる．一方で，「えってる」に対応する出力「いっている」のように，正解文の「はなしている」と意味的に一致しているものの，表層的な文字列の違いから BLEU スコアが低下してしまうといった事例も観測された．このような

[*8] 方言の傾向に関しては，使用したコーパス「全国方言談話データベース」における「青森県弘前市 1979 解説」ページを参照．

表 2　実際の翻訳例（都道府県名に続く括弧はコーパスの対話収録地点.（N）は NMT,（S）は SMT を指す）

	例 1　兵庫県（相生）方言	例 3　岩手県（遠野）方言
方言文	むかしの/ひとは/えーんやけんど/いまわ/もー	がが/とった/どきぁ/えま/えってる
参照文	むかしの/ひとは/いいのだけど/いまは/もう	つま/もらった/ときは/いま/はなしている
地域 + 文節 (N)	むかしの/ひとは/いいんだけれど/いまは/もう	かか/とった/ときは/いま/いっている
	例 2　青森県（津軽）方言	例 4　沖縄本土（今帰仁）方言
方言文	あげぁ/いろこ/つだり/すて/それか゜/ほら	わったー/いー
参照文	あかい/いろ/ついたり/して/それが/ほら	わたしたち/おまえたち
地域 + 文節（N）	あかい/いろ/ついたり/して/それが/ほら	わたしたち/おまえたち
文節（N）	あかい/いろ/つだり/して/それが/ほら	わたしたち/いい
地域 + 文（N）	あかい/いろ/ついて/それが/ほら	わたしたち/は
文節（S）	あかい/いろこ/ついたり/して/それが/ほら	わたしたち/いい

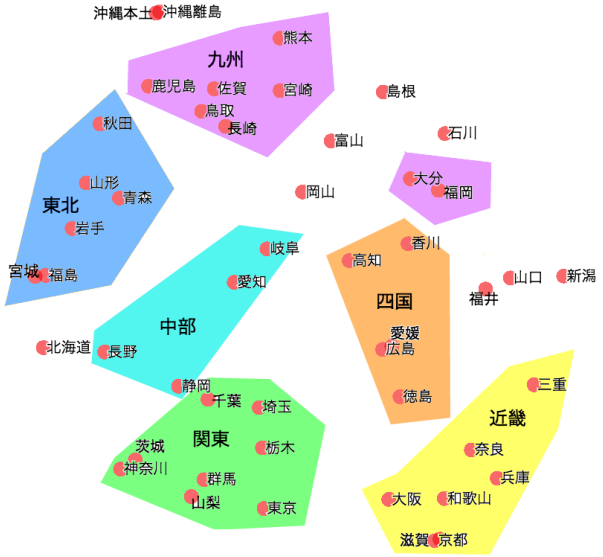

図 4　地域ベクトルの可視化

表層の異なりによる BLEU スコアの低下はほぼ全地域で生じており，意味的に一致している翻訳結果をどのように評価するかという点は今後の課題である．

我々の翻訳モデルは，複数地域の方言文を同時に学習に用いるモデルである．このモデルが地域ごとの共通性をとらえながら学習を行っているかを分析するために，地域ラベルに対応する埋め込み表現（地域ベクトルと呼ぶ）を t-SNE を用いて二次元空間に写像したものを図 4 に示す．我々の手法では，学習時に地域の地理的な情報を一切埋め込んでいないにも関わらず，各地域ベクトルの位置関係は概ね東北・九州などの地理的な地方区分に対応することが読み取れる．また，興味深いことに，東北地方と九州地方の方言の距離が近いという，周圏分布 [5] のような特徴もみられる．さらに，秋田・石川・富山・鳥取などの地域が九州地方の方言に近いといった特徴は，周圏分布だけでなく，九州地方の方言が交易関係にある地域に飛火的に伝播したとする遠隔地分布のパターン [8] を支持する結果であり，既存の方言分布に関する様々な知見との興味深い一致が見られる．

5　おわりに

国立国語研究所の「全国方言談話データベース」を学習データとして，複数地域の方言を共通語に翻訳する文字ベースのニューラル多方言翻訳器を作成した．この翻訳器では，1 地域ごとの方言共通語対訳コーパスが小規模でも，複数地域のコーパスを合わせて学習することで，方言間で共通している音韻変化を学習することができ，さらに翻訳元の文頭に地域ラベルを追加することで各方言の特徴を考慮して翻訳できることを示した．また，学習によって得られる地域ベクトルを分析することにより，方言の類型分析へ応用できる可能性を示した．

一方で，現状のコーパス規模では低頻度の方言語彙知識を獲得することが難しいことも確認された．今後は，方言語彙を含む辞書を訓練データに取り入れるなどして方言翻訳のさらなる精度向上を目指したい．

謝辞

本研究で使用したコーパスのデータを快くご提供くださった東北大学方言研究センターの皆様，および国立国語研究所の大槻知世様に深謝いたします．また，本研究は JSPS 科研費 JP15K16045，JP15H05318，およびトヨタ自動車株式会社の助成を受けたものです．

参考文献

[1] M. Johnson, M. Schuster, Quoc V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *TACL*, Vol. 5, pp. 339–351, 2016.
[2] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*, pp. 1–6.
[3] I. Saito, J. Suzuki, K. Nishida, and K. Sadamitsu. Improving Neural Text Normalization with Data Augmentation at Character- and Morphological Levels. In *IJCNLP*, pp. 257–262, 2017.
[4] B. Zoph, D. Yuret, J. May, and K. Knight. Transfer Learning for Low-Resource Neural Machine Translation. In *EMNLP*, pp. 1568–1575, Austin, Texas, November 2016.
[5] 柳田国男. 蝸牛考. 岩波書店, 1980.
[6] 長谷川駿, 田中駿, 山本悠二, 高村大也, 奥村学. 事前学習と汎化タグにおける方言翻訳の性能向上. 情報処理学会研究報告, Vol. 2017-NL-23, No. 12, pp. 3–8, 2017.
[7] 廣田壮一郎, 笹野遼平, 高村大也, 奥村学. 方言コーパス収集システムの構築. 2013 年度人工知能学会全国大会（第 27 回）, pp. 1–4, 2013.
[8] 木部暢子, 竹田晃子, 田中ゆかり, 日高水穂, 三井はるみ. 方言学入門, pp. 26–27. 三省堂, 2014.
[9] 柴田直由, 横山昌一, 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. 126–129, 2013.