特別講演「あなたの津軽弁を共通語に ――弘大×AI×津軽弁の取り組み――」

Translate your Tsugaru-ben into common language: Hirodai × AI × Tsugaru-ben project

今井 雅

Masashi IMAI

弘前大学大学院理工学研究科

Hirosaki University Graduate School of Science and Technology

Ⅰ．はじめに

青森県民と県外や国外からの転勤居住者とのコミュニケーションにおいて、地域固有の方言である津軽弁はその妨げになることがある。例えば、医療現場における津軽地方出身の患者と県外出身の医師との間のやりとりで、患者が「にやにやする」「うじゃめぐ」と症状を訴えても、医師側がそれらの意味を「（お腹に）不快感がある」「悪寒がする」という意味だと理解していなければ、適切な対応を行うことができない。また、津軽地方は中南津軽、北五津軽、西津軽、東青津軽の 4 つに分けられ、使用している津軽弁が異なる場合がある。さらに、年配の方の話す津軽弁を、津軽地方出身の若者であっても理解できない、あるいは理解できてもその言葉遣いを若者自身は使用しないということが多くの津軽弁で見られ、古くからある津軽文化の消滅が懸念されている。

われわれは人工知能（AI）を活用した津軽弁と共通語双方向の音声・文字情報変換システムを開発することを目標として、研究開発を開始した。学部横断的なチームを組み、それぞれの学問・文化領域における津軽弁を広く収集するとともに、Web システム等を用いて広く一般からも津軽弁の文例や音声情報の収集を行うなどの活動を行ってきた。それらの活動を通して得られたデータを人工知能の学習に使用するとともに、体系的に整理し、津軽弁を含む津軽文化を次世代に活用するための基盤整備を行っている。本稿では、これまでのわれわれの取り組みを述べるとともに、得られた知見及び今後の課題・展望を紹介する。

Ⅱ．これまでの取り組みと成果

1.　2017～2018 年度までの共同研究

弘前大学では、2017～2018 年度の 2 年間、東北電力および株式会社エーアイスクエアとの共同研究において、東北電力のサポートセンターで収録された津軽弁を含む音声情報を文字情報に変換することに取り組んだ。本共同研究では人工知能（音声認識エンジン）として Nuance 社製のテキスト化エンジンを使用した[1)]。このエンジン内に津軽弁の語彙はないため、単語としてではなく、音節単位での推定を行うものとして扱ったほか、学

doi: 10.24680/rnsj.100104

習用のドメイン言語モデルを追加登録して利用することにも取り組んだ。

音声認識の難しさとしては、津軽弁は「シ（チ）」と「ス（ツ）」や、「ジ（ヂ）」と「ズ（ヅ）」の区別が曖昧なこと、「イ」と「エ」の中間音があること、「カキクケコ」「タチツテト」が濁音になりやすいこと、アクセントの位置を認識しなければならないこと、鼻濁音があること、「私」が「わ」、「あなた」が「な」、「おいしい」が「め」など、言葉が短縮されやすいことなどが挙げられる。音声認識エンジンとしては、これらの共通語との違いを理解した上で、正しく分割した音節ごとに、それぞれ文字列に変換する必要がある。

最終的に、音声認識・テキスト変換の能力に関して、高いものでは 95% を達成したが、ドメイン言語モデルでは決まったフレーズやパターンでの登録とならざるを得ず、無数にある会話パターンに対応することは困難であることが認識された。また、音節単位での推定では単語として最適な単位での置き換えができない問題があることも確認された。

これらの結果より、より精度の高い音声認識を行うためには、言語モデルや音響モデルの詳細なチューニングができる、日本語に最適化された音声認識エンジンを用いることが一つの解決策であることが示唆された。

2.　2019 年からの実証研究

1）概要

既存の音声認識・変換システムからの実用化には困難な点があることが認識されたため、2019 年度からは独自システムの開発に着手することとした。また、より広く一般から津軽弁を収集するため、tgrb.jp ドメインを新たに取得し、Web サーバ（http://tgrb.jp/）を構築した。作成した Web ページでは、地方ごと、性別ごと、年代ごとの津軽弁アーカイブを構築することを目的として、津軽弁とそれに対応する共通語のほか、出身地や性別、年齢を収集できるようにした。これらにより、2019 年度末時点で単語を含む約 1 万の例文を収集することができた。しかしながら、ひとつの言語の音声認識・翻訳を精度よく行うためには約 20 万程度の例文が必要とも言われており、さらなる例文の収集が課題となっている。

津軽弁テキストから共通語テキストへの変換を行う独自システムは、図 1 に示すように、津軽弁テキストを入力として形態素解析を行うシステムと、人工知能により翻訳を行うシステムにより構成している。

2）形態素解析

形態素とは、ある言語においてそれ以上分割したら意味をなさなくなるところまで分割して抽出された、音素のまとまりを指す。自然言語のテキストデータから、文法や辞書情報等に基づいて形態素の列に分割し、形態素の品詞等を判別することを形態素解析と呼ぶ。日本語に対応した形態素解析ツールとして、京都大学情報学研究科と NTT コミュニケーション科学基礎研究所による共同研究ユニットプロジェクトを通して開発されたオープンソース形態素解析エンジン MeCab（和布蕪）がある[2)]。

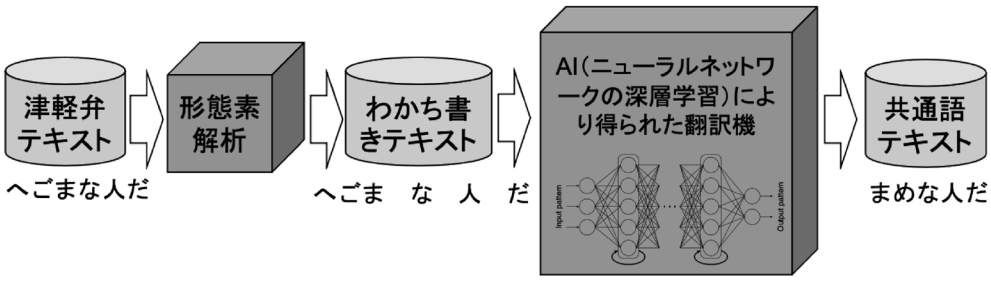

図 1．津軽弁テキストから共通語テキストへの変換システム

MeCab では、独自のライブラリを追加することができるため、収集した津軽弁情報を利用し、活用形を含む約 28,700 語の津軽弁ライブラリを作成した。このライブラリを用いることで、形態素の分割精度が 28% から 62% に向上し、登録した語彙を含む多くの津軽弁テキストを正しい分割（分かち書き）にできることを確認した。しかしながら、作成した津軽弁ライブラリを用いたとしても正しく分割できない例もあり、さらなる精度の向上が課題である。

3）人工知能による翻訳機

津軽弁テキストを共通語テキストに変換する仕組みは、ディープラーニングフレームワーク Chainer[3] を利用して Python で実装した。ニューラルネットワークのモデルは時系列データを扱うことのできるリカレントニューラルネットワークモデルを基本として、いくつかの拡張手法を適用している。

これまでに収集した津軽弁を作成したシステムに学習させることにより、学習した津軽弁及び類似の文章は正しく共通語に変換できることを確認した。しかしながら、学習に利用していない文章や登録されていない語彙の変換では、正しい共通語を出力することができていない。そのため、人工知能による翻訳の精度を高めるためには、より多くの文例による学習、類義語のクラスタリングなどの処理が必要となっている。

III．成果の公表と今後の課題・展望

これまでに収集した津軽弁の情報は、約 10,000 語の津軽語辞書として Web ページ（http://tgrb.jp/dic/）で公開している。未登録の情報もあるため、津軽語辞書のメンテナンスシステムを構築し、随時更新を行う予定である。

形態素解析では、分割精度は実用的なレベルとしてはまだ低いため、正確な形態素分割を行う手法について今後も検討していく。また、不正確な分割や偽物の津軽弁があったとしても、うまく人工知能に学習させることで正しい共通語へ変換することも可能であるため、さまざまな学習用データを用意して、より精度高く津軽弁テキストを共通語テキストに変換する人工知能の構築を狙う。

最終的に、図 2 に示すように、津軽弁音声を共通語音声に双方向で変換できるようなシステム開発を目指してこれからも研究を行っていく。

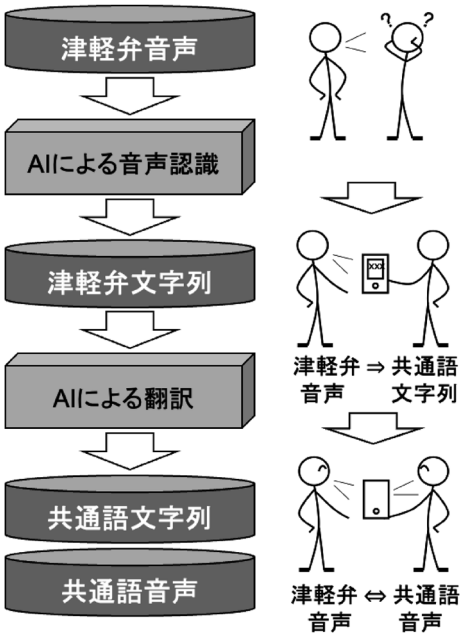

図 2．方言と共通語双方向の音声・文字情報変換システム

Ⅳ．まとめ

1. 人工知能（AI）を活用した津軽弁と共通語双方向の音声・文字情報変換システムを開発することにより、地域社会での課題解決に貢献するべく、弘前大学では「弘大×AI×津軽弁プロジェクト」を実施している。
2. 津軽弁を中心とした津軽文化を未来に継承するべく、さまざまな学問・文化領域の津軽弁文例・音声を収集し、津軽弁アーカイブを構築するとともに、約 10,000 語の津軽語辞書として Web サイトで公開している。

研究助成

本研究は、2017～2018 年度東北電力共同研究、2020～2021 年度弘前大学次世代機関研究の助成を受けて実施した。

利益相反

本研究における利益相反は存在しない。

引用文献

1) Nuance 社音声認識エンジン．https://www.nuance.com/（検索日：2022 年 3 月 10 日）.
2) 形態素解析エンジン MeCab. https://taku910.github.io/mecab/（検索日：2022 年 3 月 10 日）.
3) ディープラーニングフレームワーク Chainer. https://chainer.org/（検索日：2022 年 3 月 10 日）.