人工知能を活用した津軽弁と共通語双方向の音声・文字変換システムの開発

Development of Bidirectional Voice and Character Translation System between Tsugaru Dialect and Common Language utilizing Artificial Intelligence

今井　雅
Masashi Imai

弘前大学大学院理工学研究科
教授
Professor,
Graduate School of
Science and Technology,
Hirosaki University

研究の目的、背景
Purpose and Background of the Research

青森県民と県外や国外からの転勤居住者や観光客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。また、津軽地方は中南津軽、北五津軽、西津軽、東青津軽の４つに分けることができ、それぞれ使用している津軽弁が少しずつ異なっている。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、あるいは理解することは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。そこで、我々は人工知能（AI：Artificial Intelligence）を活用した方言と共通語双方向の音声・文字変換システム（図１）を実現することを目的として、「弘大×AI×津軽弁プロジェクト」を行っている。学部横断的なチームを組み、さまざまな学問・文化領域における津軽弁を広く収集してAIの学習に使用するとともに、収集した津軽弁を体系的に整理し、津軽弁を中心とした津軽文化を保存し、次世代に活用できる基盤を整備することも目的としている。

Tsugaru-ben, which is a dialect particular to the region, can be an obstacle to communication between Aomori residents and residents who have transferred here for work and tourists from outside the prefecture and abroad. The Tsugaru region includes four areas: Chunan-Tsugaru, Kitago-Tsugaru, Nishi-Tsugaru, and Tosei-Tsugaru, and the Tsugaru-ben used in each area is slightly different. It is often seen that even young people in the Tsugaru region either cannot understand the Tsugaru-ben spoken by older people, or the young people can understand it but not use it themselves, resulting in the disappearance of the old Tsugaru culture. Therefore, we are conducting the “Hirodai x AI x Tsugaru-ben Project” with the aim of developing a bidirectional voice and character translation system between Tsugaru-ben and standard Japanese utilizing Artificial Intelligence (AI) (Fig.1). Tsugaru-ben in various academic and cultural fields is widely collected by a cross-faculty team and used for AI learning. The collected Tsugaru-ben is also systematically organized to establish a data infrastructure that can be used by the next generation.

研究成果
Research Results

研究分担者の協力のもと、各担当領域における津軽弁関連情報及び津軽弁音声情報の収集を行った。その結果、各種文献や津軽弁の使用されている各種メディアを収集することができ、目標とする3万例の文例収集に近づくことができた。音声情報の収集に関しては、対面での音声情報の収集が難しい中、可能な範囲で情報の収集を行い、約500種のデータを収集することができた。また、弘前地域関係者の協力のもと、50時間を越える会話音声データも収集することができた。

これまでの研究活動により収集した津軽弁は、約１万語の津軽語辞書として、Ｗｅｂ上で公開している（ http://tgrb.jp/dic/ ）。また、津軽語辞書と文例のメンテナンスシステムの構築を開発し、辞書の拡張を随時行っている。

本研究で実現する文字・音声変換システムは、津軽弁音声を津軽弁文字列に変換するAIと津軽弁文字列を共通語文字列に変換するAIがある。これまでは後者のシステム開発を中心に研究を進めており、AIで文字列変換を行うための品詞分解ツールMeCabに対して、28,700語の津軽語を含むライブラリを構築し、品詞分解の精度を28%から62%に向上させることができた（図２）。

We collected Tsugaru-ben-related information and Tsugaru-ben voices in each area of responsibility with the cooperation of the research team members. As a result, we have collected various documents and various media in which Tsugaru-ben is used and achieved our first target, which was about 30,000 examples. While it was difficult to collect voice information face-to-face, we have collected about 500 types of data. In addition, we have collected more than 50 hours of conversational voice data with the cooperation of Hirosaki area collaborators.

The Tsugaru-ben collected through research activities is published on the Web as a Tsugaru-ben dictionary of about 10,000 words (http://tgrb.jp/dic/). We have developed a maintenance system for the Tsugaru-ben dictionary and sentences. The database is expanded with new vocabulary and expressions whenever they are discovered.

In this research, the bidirectional voice and character translation system contains two AI systems; 1. AI that converts a Tsugaru-ben voice into a Tsugaru-ben sentence, 2. AI that

converts a Tsugaru-ben sentence into a standard Japanese sentence. So far, we have developed focusing on the latter system and built a library containing 28,700 Tsugaru words for the morphological analysis tool “MeCab” (Fig.2). As a result, we can improve the accuracy of morphological analysis from 28% to 62%.

今後の展望
Future Prospects

品詞分解の精度はまだ低く、形態素解析ツールで正確な品詞分解を行う手法について今後も検討していく。また、不正確な品詞分解や偽物の津軽弁があったとしても、うまくAI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文字列を共通語文字列に変換するAIを実現する。

津軽弁音声認識を行うAIに関して、収集した音声情報の中にはAIの学習に利用する前に加工しなければならないものもあるため、今後これらのデータを精査・整理し、精度高く津軽弁音声を文字列に変換するAIを実現するとともに、アーカイブ化して津軽文化を継承するための基盤整備を行う。

We will continue to study methods for accurate morphological analysis using MeCab since its accuracy is still low. Even if there is an inaccurate morphological analysis or a fake Tsugaru-ben, it is possible to convert it into the correct standard Japanese by appropriately learned AI. Thus, we try to develop several AIs that convert Tsugaru-ben sentences into standard Japanese sentences by preparing various learning data.

Some of the collected voice information must be processed before it can be used for AI learning, so we will scrutinize and organize these data. And then, we will develop an AI which correctly converts Tsugaru-ben voices into Tsugaru-ben sentences using these data In addition, we will develop a data infrastructure to inherit the Tsugaru culture.

主な研究資金（直接経費）
Main Research Funding (Direct Costs)

・弘前大学次世代機関研究/2020年度～2021年度/4,000,000円

・Hirosaki University Institutional Research Grant for Future Innovation FY2020-2021　4,000,000 Yen

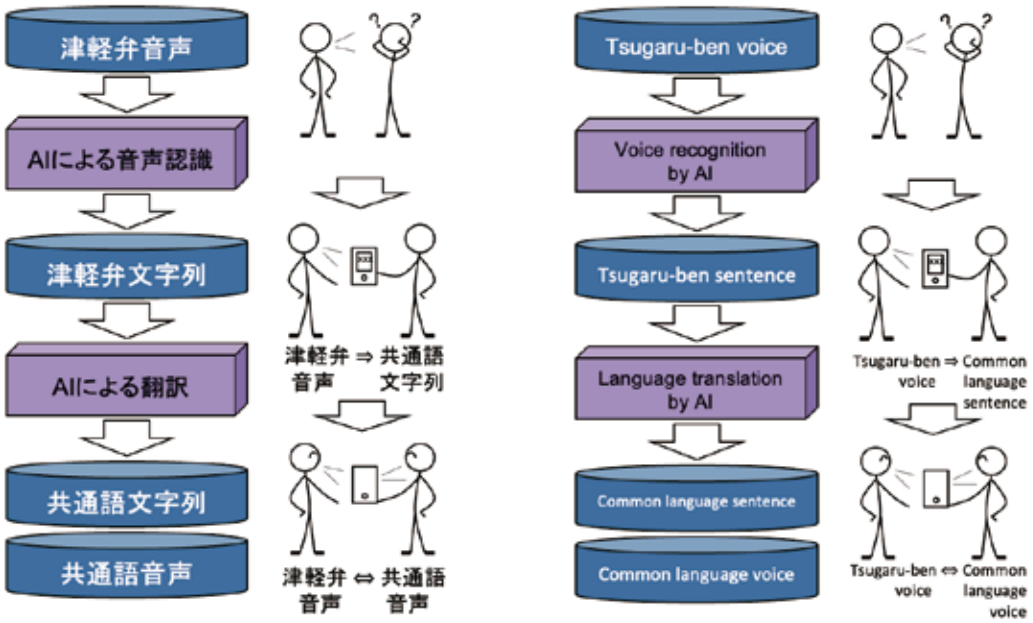

図１：変換システム概要
Fig.1: Translation system overview.

図２：津軽弁文字列から共通語文字列への変換AI
Fig.2: Text translation AI from Tsugaru-ben to standard Japanese.