Assessment of Text Level Using Deep Learning Methods

Jeff Winchell, Mariko Itoh Henstock, Pirinka Georgiev, Peter V. Henstock

Reading authentic materials is important for not only language instruction but for empowering language students to pursue their interests. However, given the difficulty in learning the Japanese language, identifying materials at the right level is challenging. We propose a deep learning method that is an artificial intelligence (AI) approach to classify texts based on the language level of the students.

Artificial intelligence is a subfield of computer science that officially began in 1956. It can be defined as the field of computing that solves problems that we generally associate with human intelligence such as recognizing images, understanding speech, writing text, machine translation and playing games. It has historically had its periods of success and failures (called AI winters). The AI subfield of machine learning uses methods that make decisions based on numeric features. These have been successfully used across many fields for decades. The standard approach is for field experts to construct sets of features such as linguistic characteristics and use the machine learning to discovery patterns or classify the data based on training sets. Such approaches have been used to determine which language is being spoken, for example. In the last 5 years, a new class of machine learning methods has emerged called deep learning. They use neural networks that were inspired by the way neurons work in the brain. Given enough data, deep learning methods have demonstrated that they can recognize images and sounds as well as humans. Furthermore, the deep learning approach doesn't require the extraction of features but can work on the raw data (i.e. language text).

This project aims to determine whether a deep learning approach can correctly classify texts as appropriate for a 1st year, 2nd year, or 3rd year Japanese language student. The approach is to train a deep learning system on Japanese text extracted from the standard language textbooks such as Genki, Nakama, Tobira, etc. Unlike traditional machine learning methods and previous statistical methods that require the extraction of kanji, grammar terms, and other linguistic features, the deep learning will use just the sequences of Japanese text and the textbook level. The approach could also be used to identify issues when switching textbooks across a program since one book may have gaps or redundancies when going from one level to the next.