Machine learning modeling for “AI HANASHIKOTOBA Checker”

Sosuke KAWAGOE（川越 颯亮）

Reports and dissertations must be written using academic representations (hereafter referred to as “KAKIKOTOBA”). However, many students find it difficult to use KAKIKOTOBA while writing reports, and specialists in Japanese language education believe that employing HANASHIKOTOBA in sentences is particularly hard. In response to this, our research team has been studying and developing a rule-based sentence correction system for first-year university students, “HANASHIKOTOBA Checker” (hereinafter referred to as “this system”) using a database that collects information on HANASHIKOTOBA.

Some ambiguous expressions are likely to be interpreted differently by various teachers, making it challenging to provide consistent instruction. One of the reasons is that some representations change depending on the context in which they are used and on whether they are HANASHIKOTOBA or KAKIKOTOBA. Because the system detects HANASHIKOTOBA based on a rule-based algorithm, it sometimes provides incorrect detection results. In particular, some representations are identical to HANASHIKOTOBA but cannot be distinguished in meaning from KAKIKOTOBA in different contexts, which we define as "gray representations".

Due to the growth of the Internet, texts and photos are now freely accessible datasets that anybody can utilize. These datasets can be used by anybody to create machine-learning models. In the system, machine learning is also considered effective for identifying gray representations that require contextual judgment. However, building a machine-learning model that can gather gray representation instances is challenging owing to the absence of a publicly accessible corpus of HANASHIKOTOBA.

This project aimed to build an AI-powered machine-learning model that can between HANASIKOTOBA and KAKIKOTOBA in context-dependent gray representations. For developing such a machine-learning model, a significant amount of training data is required. Therefore, in this study, we assumed that among sentences published on the Internet, subjective sentences include HANASHIKOTOBA, objective sentences include KAKIKOTOBA, and sentences can be classified as being written from either a subjective or objective viewpoint. Based on this assumption, a dataset was created by compiling words with gray representations from various sources, and a gray-representation discrimination model was created by tweaking a pretrained Japanese BERT model. The model was then confirmed and assessed using the Teshimau dataset created by Yamashita of our research team, which contains gray representations.

The validation results demonstrated that BERT fine-tuning was able to discriminate between subjective and objective sentences with 77.0% accuracy, using Amazon review sentences as subjective sentences and peer-reviewed papers from the Japan Society for Educational Technology as objective sentences. As a result, it is proposed that the model can be utilized to create an “AI

HANASHIKOTOBA Checker.” The subjective or objective nature of the statement will determine whether the gray representations are HANASHIKOTOBA or KAKIKOTOBA. However, it was confirmed once again that it is challenging to identify gray representations because the viewpoints of subjectivity and objectivity differ depending on the field in which the text was submitted and the reviewer or other correctors.

Future research should identify gray representations other than Teshimau and further enhance the gray-representation discrimination accuracy of the model constructed in this study. The model should then be introduced into the system, and a flexible speech-detection system built for rule-based and AI-based gray-representation discrimination using machine-learning models should be developed.

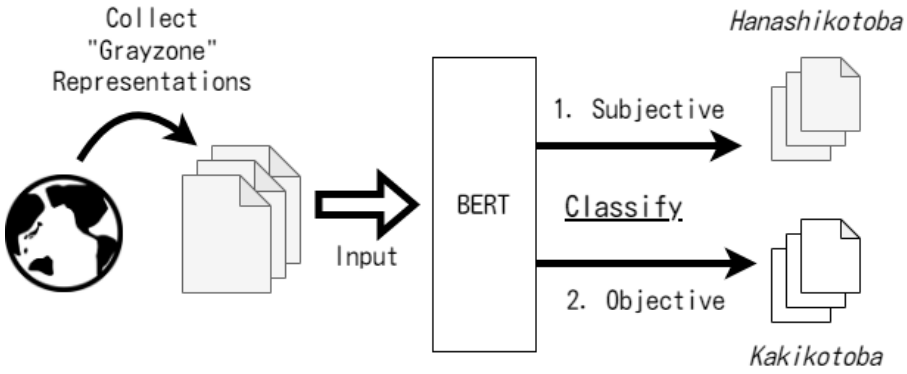

Image: "Grayzone" classification model