**Introduction**

Dialects are subsets of a language delineated by geographic/social boundaries, and may or may not be mutually intelligible. Classifying dialects is often difficult and even contentious.

Japanese dialect classification has historically relied on limited components of the language [1, 2]. We aim to create a machine learning model that classifies dialects using more comprehensive measures. Such studies could serve as a starting point towards easing controversy surrounding dialect classification.

**Dataset**

We used the dataset from the “Field Research Project to Analyze the Formation Process of Japanese Dialects” (FPJDS) study done by the National Institute for Japanese Language and Linguistics (NINJAL). Responses to 211 prompts were collected from 554 locations. These prompts assessed how dialects in those regions differed in grammatical structure, pronunciations, words used for common nouns, and so on.

<table>
<thead>
<tr>
<th>Sample prompts</th>
<th>Aspect of interest</th>
<th>Example answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>What do you call a tuber like this?</td>
<td>Noun</td>
<td>Japanese (うなぎの丈)</td>
</tr>
<tr>
<td>When saying, “It’s 10 o’clock and they haven’t come yet,” how would you say “haven’t come yet”?</td>
<td>Grammar (negative conjugation)</td>
<td>Kun (くん)</td>
</tr>
</tbody>
</table>

Table 1: Example prompts from the survey.

A feature vector was created for each prompt response. We included 33 features that correspond to different pronunciations of key vowels and consonants, as well as linguistic features, such as glottal stops.

The feature vectors for each prompt were combined, and its first fifty principal components were used for analysis.

**Supervised Approaches**

![Fig. 2](image)

Fig. 2: Examples labeled using a dialect map created by Japanese linguist Hirayama Tero [2]. Responses from the Hachijo region were not considered because there were only three of them.

![Fig. 3](image)

Fig. 3: Confusion matrices for various classifiers (train/test = 75%/25%). Hyperparameters for each model were chosen using an exhaustive grid search with 5-fold cross validation.

**Unsupervised Approaches**

![Fig. 6](image)

Fig. 6: $k$-means clustering results (top: geographical labels, bottom: clustering in PCA space).

![Fig. 7](image)

Fig. 7: Truncated decision tree (left) and geographical labeling (right) for hierarchical clustering directly on unfactored IPA responses.

Unsupervised learning was able to discern differences in dialects between Sapporo (major city) and the rest of Hokkaido (Fig. 6 and 7).

**Conclusions**

Machine learning models can efficiently synthesize the linguistic richness in dialects, reducing some of the inherent difficulty in dialect classification.

While supervised techniques successfully classified among geographic (island) and political (provincial) boundaries, the unsupervised approaches were able to also pick up subtle linguistic differences, providing evidence that dialect regions are not always associated with geographical or political boundaries.

Exploring more sophisticated features could uncover important aspects that are currently overlooked in dialect classification. While this study only analyzed transcriptions, these methods could also be extended to audio speech samples.

**References**


