Discovery of Linguistic Patterns in Japanese News Articles

Wyatt Bramblett
College of Computing and
Software Engineering
wbrambl1@students.kennesaw.edu

***Abstract*—This project applies unsupervised data mining techniques to a large corpus of Japanese news articles collected between 2005 and 2021. The goal is to discover latent structure in real world Japanese text without relying on predefined labels or prediction tasks. By using document clustering and association rule mining, this project aims to uncover emergent thematic groupings and recurring linguistic patterns that characterize Japanese news writing across different sources and time periods.**

I. Introduction

Large scale text collections such as news archives contain rich linguistic and semantic structure that is not always explicitly labeled. In the context of Japanese-language data, this structure is further complicated by the lack of whitespace between words, frequent use of kanji compounds, and stylistic variation across news domains. This project focuses on discovering patterns that naturally emerge from the data itself.

There are two main motivations for doing this project. First, it provides an opportunity to apply data mining techniques to non-English text, which introduces unique preprocessing and representation challenges. Second, the use of Japanese news articles aligns with my interest in the Japanese language, allowing discovered patterns to be interpreted both computationally and linguistically.

II. Dataset Description

A. Dataset Source

The dataset used in this project is the *Japanese Newspapers 2005–2021* dataset, publicly available on Kaggle [1]. It contains Japanese-language news articles collected from multiple online newspaper sources, including but not limited to (nikkei.com, tomamin.co.jp, and hokkaido-np.co.jp), over a sixteen-year period.

B. Dataset Structure

The dataset is provided in CSV format, where each row corresponds to a single news article. The primary attributes used in this project are summarized in Table I.

The dataset contains tens of thousands of articles and several million characters of text. Article lengths vary widely, ranging from short announcements to full-length reports.

Field	Description
source	Newspaper website or domain name
date	Article publication date
text	Full Japanese article content

TABLE I
Dataset schema

C. Data Characteristics and Quality Considerations

The dataset does not include topic labels or metadata describing article categories, making it particularly suitable for unsupervised discovery tasks. However, several preprocessing challenges must be addressed. Japanese text does not use whitespace to separate words, requiring morphological analysis to identify meaningful tokens. Additionally, articles may contain punctuation, numbers, or stylistic markers that do not contribute to semantic analysis.

Despite these challenges, the dataset represents authentic, real-world language use across a wide range of topics and writing styles. Each article is treated as an independent document for analysis purposes.

III. Discovery Questions

This project investigates the following discovery-oriented research questions.

A. Q1: Document Clustering

Are there natural clusters of Japanese news articles based on linguistic similarity, and what types of themes or writing styles emerge within these clusters?

This question explores whether articles naturally group together based on vocabulary usage and linguistic structure, without relying on predefined categories. By clustering documents using unsupervised techniques, the project seeks to uncover latent groupings that may correspond to broad themes such as politics, economics, culture, or local events, as well as stylistic differences between sources.

The value of this question lies in understanding how thematic structure emerges organically from language usage, rather than being imposed by human labeling.

B. Q2: Linguistic Associations

Which Japanese words or phrases frequently co-occur within news articles, and what semantic patterns do these associations reveal?

This question focuses on identifying frequently co occurring words or phrases that reflect common expressions, domain specific terminology, or recurring narrative patterns in Japanese news writing. Discovering such associations can reveal how concepts are linguistically linked in practice, offering insight into both journalistic conventions and real world language use.

IV. Planned Techniques

A. Text Preprocessing

Prior to analysis, the text data will undergo several preprocessing steps. These include morphological tokenization using a Japanese tokenizer, removal of stopwords and punctuation, and normalization of text where appropriate. Each article will then be represented in a numerical format suitable for data mining techniques.

B. Clustering Methods

To address Q1, clustering techniques such as K-Means or Hierarchical Clustering will be applied to document representations such as term-frequency or TF-IDF vectors. The resulting clusters will be analyzed by examining representative documents, frequently occurring terms, and source distributions within each cluster. This analysis will focus on interpretation rather than cluster accuracy.

C. Association Rule Mining

To address Q2, association rule mining techniques such as Apriori or FP-Growth will be applied. Each document will be treated as a transaction containing a set of words. Frequent itemsets and strong association rules will be analyzed to identify meaningful word relationships and recurring linguistic patterns.

D. Interpretation and Validation

Because this project emphasizes discovery, results will be evaluated qualitatively. Cluster coherence and association rules will be interpreted through inspection of example articles and linguistic analysis, rather than predictive metrics.

V. Preliminary Timeline

A. Milestone Plan

- **M2:** Data preprocessing, tokenization, and exploratory analysis
- **M3:** Implementation of clustering and association rule mining techniques
- **M4:** Interpretation of discovered patterns and preparation of final report

B. Anticipated Challenges

Anticipated challenges include effective Japanese tokenization and interpretation of unsupervised results. These challenges will be mitigated through careful preprocessing and qualitative analysis of discovered patterns.

VI. Conclusion

This project applies discovery focused data mining techniques to a large corpus of Japanese news articles. By uncovering latent document clusters and linguistic associations, the project aims to provide insight into thematic structure and language usage patterns without relying on prediction or labeled data. The findings are expected to contribute both to computational understanding of text data and to linguistic insight into Japanese news writing.

VII. GitHub Repository

A GitHub repository has been created to manage the project code, documentation, and data organization throughout the semester [2].

The repository includes a `README.md` file that provides an overview of the project, including the project title, a brief description of the research goals, the dataset source link, and author information. This ensures the project is clearly documented and reproducible.

The repository is organized using the following structure:

- `data/` - Contains the dataset or placeholders with a
- `notebooks/` - Will contain Jupyter notebooks used for preprocessing, clustering, and association rule mining in later milestones.
- `docs/` - Will contain all documents related to this research, such as this proposal.
- `README.md` - Provides a high-level description of the project and dataset.

This repository will be continuously updated throughout Milestones M2, M3, and M4 as additional analyses and results are produced.

References

[1] V. Huholl, "Japanese Newspapers 2005–2021," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/vyhuholl/japanese-newspapers-20052021

[2] W. Bramblett, "Discovery of Linguistic Patterns in Japanese News Articles (Code Repository)," GitHub, 2026. [Online]. Available: https://github.com/WyattBram/CS4412-DataMiningProject