Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Social Computing (Eiji Aramaki (Professor))		
Student ID	2411412	Submission date	2026 / 10 / 27
Name	GAIDUCHENKO SOFIA		
Thesis title	Japanese Idiom WSD with Large Language Models under Surface-Form Variation		

Abstract

Some Japanese idioms can have both literal and idiomatic interpretations and may occur in surface forms that differ from their canonical dictionary representations. This thesis investigates how surface-form variation, prompting strategies, and fine-tuning data composition relate to large language model (LLM) performance on Japanese idiom word sense disambiguation (WSD).

Using the OpenMWE corpus, we examine 69,712 annotated occurrences of 90 ambiguous Japanese idioms. We distinguish preserved and changed surface forms and automatically annotate five families of morphosyntactic operations. Four pretrained models—GPT-4o mini, Qwen2.5-7B-Instruct, Llama-3-ELYZA-JP-8B, and Qwen2.5-0.5B-Instruct—are evaluated under zero-shot and target-idiom four-shot prompting on a common held-out set of 7,724 occurrences. We further fine-tune Qwen2.5-0.5B-Instruct using LoRA under six training conditions across three random seeds.

Prompting effects are model-dependent: target-idiom demonstrations improve macro-F1 for GPT-4o mini (58.86% to 76.37%) and ELYZA-JP-8B (35.48% to 44.69%), but reduce performance for both Qwen models.

Fine-tuning substantially improves the smaller Qwen model, with unrestricted training achieving 86.33% mean macro-F1 compared with 45.93% for the pretrained zero-shot baseline. However, training restricted to morphosyntactically annotated occurrences does not consistently outperform size- and label-matched controls, even on the corresponding evaluation subset, and exhibits substantial variability across seeds. Preserved-only training produces a larger apparent performance gap between preserved and changed forms, but this difference is substantially reduced when idiom and sense contributions are balanced. Furthermore, total per-idiom training frequency shows little association with accuracy, whereas the proportion of idiomatic versus literal training examples is strongly associated with class-specific recall.

These findings show that Japanese idiom WSD performance depends not only on surface-form characteristics but also on model-specific prompting behavior, supervised adaptation, and the distribution of semantic interpretations in training and evaluation data. They highlight the importance of accounting for dataset composition when assessing robustness to linguistic variation.