Grounding and Auditability for Large Multimodal Models, From Robot Perception to Low-Resource Knowledge Graphs

Name: Atuhurra Jesse

Laboratory’ s name: Natural Language Processing

Supervisor’ s name: Professor Taro Watanabe

Abstract (should be within 1st page)

Vision language models and large language models increasingly power real systems, yet they often fail in the same underlying way, they produce fluent outputs without reliable grounding, and they are difficult to audit under noisy, low-resource conditions. This thesis develops a unified, evidence-first approach to grounding and auditability across two domains, robot perception from Japanese human robot interactions, and information extraction for Swahili knowledge graph construction.

First, we study attribute-grounded robot scene understanding with VLMs using J-ORA, a dataset of Japanese dialogue scenes that require object identification, reference resolution, and next-action prediction. We show that attribute-rich grounding improves end to end multimodal perception and supports affordance-oriented reasoning, while also revealing systematic gaps between open and proprietary models under realistic ambiguity and distractors. Second, we construct an evidence-anchored Swahili knowledge graph from one million sentences, emphasizing provenance and verification. We introduce DR-Prune for relation pruning, CalREL for canonicalization and low-resource entity linking, and MAC-VF for multi-agent consensus and verification, and we quantify precision recall tradeoffs and local-entity coverage using gold-set evaluation.

Bringing these projects together, the thesis argues that robust grounding requires explicit intermediate structure, attributes, canonical forms, and evidence spans, paired with evaluation protocols that measure not only task accuracy but also faithfulness and auditability. We synthesize shared failure modes and propose cross-project extensions that transfer verification, provenance, and calibration ideas between multimodal perception and text-based extraction. Overall, this work provides practical methods and committee-ready evidence that grounded, auditable LLM and VLM systems are achievable even in low-resource and deployment-realistic settings.