Automated Glossing for Language Learning: Multiword Expression Identification and Dictionary Definition Generation

Name: Yusuke Ide

Laboratory's name: Natural Language Processing Laboratory

Supervisor's name: Taro Watanabe

Abstract (should be within 1st page)

Glossing—the provision of explanations, such as definitions, for unfamiliar lexical items in a text—can effectively support second-language learners' reading. This dissertation addresses two core tasks required for automated lexicon-based glossing: multiword expression identification (MWEI), which identifies the multiword lexical units to be glossed, and learner's dictionary definition generation (LDDG), which generates lexically simple dictionary definitions for them. For MWEI, we introduce CoAM, a high-quality dataset comprising 1.3K English sentences. CoAM was created through a multi-step quality control pipeline combining human annotation, human review, and automated consistency checks. Using CoAM, we evaluate large language model (LLM)-based methods and show that fine-tuned LLMs outperform the previous state-of-the-art encoder-based system. For LDDG, we propose a reliable LLM-as-a-judge evaluation framework based on novel criteria. To support this evaluation, we construct D3J, a dataset of Japanese dictionary definitions that serves as reference material. Validation shows that the proposed framework aligns with human judgments at a level comparable to that of inter-annotator agreement. We further present an LLM-based LDDG method with iterative simplification. Experiments show that it produces high-scoring definitions under the proposed criteria while maintaining lexical simplicity. Together, these results establish practical methods for building robust lexicon-based glossing systems.