Investigations into Abstract Linguistic Knowledge in Language Models

Name: Miyu Oba

Laboratory's name: Natural Language Processing Laboratory

Supervisor's name: Taro Watanabe

Abstract (should be within 1st page)

Language allows speakers to form and interpret new expressions through novel combinations of familiar linguistic elements. This flexibility is supported by abstract linguistic knowledge that captures regularities across different expressions. Although language models can process a wide variety of expressions, it remains unclear to what extent their behavior relies on such knowledge. Generalization provides a means of investigating this question. However, even successful generalization may be supported by multiple linguistic cues, making it difficult to identify what knowledge models rely on. This dissertation investigates abstract linguistic knowledge through controlled studies of (1) grammatical generalization from different kinds of training evidence and (2) constructional meaning under controlled evaluation conditions.

First, it investigates grammatical generalization from indirect evidence using WIDET, which controls the types of evidence provided through synthetic training data across seven English linguistic phenomena. The results show that direct evidence often improves grammatical judgments, whereas indirect evidence yields more limited generalization, with substantial variation across phenomena. Second, it explores constructional understanding using CxMP, a minimal-pair benchmark covering nine English constructions. The results show that models perform well on grammatical acceptability judgments early in training, whereas their ability to interpret constructional meaning improves more slowly and remains limited for some constructions even in large models. These results reveal limitations in linguistic generalization under certain conditions, highlighting the value of controlled investigations for assessing abstract linguistic knowledge in language models.