論 文 の 内 容 の 要 旨

論文題目　Evaluating Natural Language Understanding
in Machine Reading Comprehension
（機械読解における自然言語理解の評価）

氏　名　　菅原 朔

Building machines that can understand human language is one of the long-standing challenges in natural language processing. This thesis tackles how to evaluate natural language understanding in machine reading comprehension—a task in which computer systems answer questions about given texts. Machine reading comprehension is an important testbed for jointly evaluating various aspects and components of language understanding. There are large-scale, various datasets presented recently on some of which proposed systems achieved human-level performences. However, we raise two major issues in machine reading comprehension. The first issue is about evaluation metrics. Because systems are evaluated with simple accuracy in most existing datasets, we cannot obtain fine-grained information about a system's capability of reading comprehension. Therefore, we cannot explain what the system achieved in terms of language understanding, which prevent us from improving the development of systems. The latter issue is about the quality of questions. Even if questions seem to require human-level understanding of given texts, they may be solved only by simple matching word patterns between a question and given texts. In this situation, we cannot conclude that the system achieved human-level language understanding even if it exhibits the performance comparable with humans.

In this thesis, we discuss how we can design a dataset of machine reading comprehension for precisely and correctly evaluating the capability of language understanding. This thesis consists of seven chapters. In Chapter 1, we introduce current issues in machine reading comprehension and our motivation. In Chapter 2, we overview machine reading comprehension datasets, systems, and related language understanding tasks. In Chapter 3, we consider evaluation metrics, namely, how to evaluate the performance of machines beyond simple accuracy. We propose new metrics comprised of requisite skills and text readability to highlight systems' abilities in detail. In Chapter 4, we address how to investigate the quality of questions so that they can correctly evaluate intended language understanding. We propose analysis methods to look into question difficulty and requisite skills. In Chapter 5, we present a methodology for automatically assessing the benchmarking capacity of machine reading comprehension datasets from language understanding skills. We combine our proposed skills and analysis methods and reveal what kind of skills are required for answering questions. In Chapter 6, we discuss the explainability of machine reading comprehension and provide theoretical foundations for reading comprehension and its evaluation. We inspect current machine reading comprehension using these foundations and list requirements for the explainability. In Chapter 7, we summarize conclusions and mention the future of machine reading comprehension.