論　文　の　要　旨

題　目　Sequence Studies for Text Data Augmentation and Molecular Generation
（テキストデータの拡張と化合物分子の人工生成のためのデータベースシーケンスの研究）

氏　名　Huidong Tang (唐　惠東)

Abstract

This thesis studies sequence modeling for text data augmentation and molecular generation. Sequence modeling is a widely studied research area, including stock price and weather data prediction, gene sequencing, Natural Language Processing (NLP), and molecular generation. These research topics cover economics, environment, biology, linguistics, and chemistry, which are tightly linked to the development of society.

Among these foundations, natural language is the cornerstone of civilization, facilitating communication and storing knowledge. To solve language-related problems with the help of computational techniques, NLP, which consists of various tasks such as text classification, entity extraction, text summarization, and text generation, is being developed. Text classification is one of the fundamental NLP tasks, that aims to categorize a text into one or more classes. However, the robustness of such a task remains a concern, as the predictions can be manipulated by adding perturbations. To address this concern, we propose three data augmentation methods based on word substitution, combining synonyms, antonyms, and sentimentally related words for the robustness enhancement of the text classification task. We attempted to generate samples that differed in semantics from the training data to improve the robustness. We evaluated our methods on four publicly available datasets using text adversarial attack techniques, and the experimental results validated the robustness enhancement.

On the other hand, we also studied molecular generation for drug discovery using Generative Adversarial Networks (GANs). Modeling molecular generation using the Simplified Molecular Input Line Entry System (SMILES) as a token-level sequence generation is straightforward. However, naively adopting the cumulative reward for the token-level sequence generation is time-consuming and incompatible with the SMILES nature. To address these limitations, we introduced an efficient reward function that combines moment and global rewards, along with the information entropy maximization, as an alternative to the cumulative reward. The combination of moment and global rewards reduces the training time and ensures molecular consistency, and the information entropy maximization allows for diverse explorations and avoids the mode collapse problem common in GANs. Our Enhanced Actor-critic Reinforcement Learning (RL) agent-driven GAN, EarlGAN, can generate molecules with a highly balanced performance. Our extensive evaluation experiments validated the effectiveness of our model.