ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名：
A Knowledge-aware Multi-tasks-based Disease Network Construction on Biomedical Literature

実施時期：2025
所属機関名：National Institute of Advanced Industrial Science and Technology
代表者氏名：Mohammad Golam Sohrab

成果概要：
We introduce Bio-016 CAT5, a biomedical domain-specific text-to-text transfer transformer that integrates concept related knowledge from the unified medical language system (UMLS) during pre-training.

成果のポイント：
- BioCAT5 leverages concept mention-aware span masking pre-training by leveraging the concept mentions from the unified medical language system (UMLS) that facilitates over a million of biomedical concept unique identifiers (CUI).
- For BioCAT5 data creation, to the best of our knowledge, we are the first to perform a large-scale concept or concept unique identifier's (CUI) name mapping that grafts medical knowledge during pre-training.
- Publicly available datasets are used and compared with the state-of-the-art (SOTA) biomedical models where the BioCAT5 outperformed four out of five NER datasets.
- BioCAT5-Base Model: For the base model pretraining was carried out with ABCI 3.0 where NVIDIA 865 RTX 6000 ada with 48GB GDDR6 memory 866 is used. The BioCAT5 model is trained on 13.6B tokens, where we employ a batch size of 61,440 tokens with a maximum step of 221,600 steps.
- BioCAT5-3B Model: BioCAT5-3B is a large language model (LLM) and carried out with ABCI 3.0 where NVIDIA H200 SXM 141GB GPU for the continual pretraining of BioCAT5-3B. During continual pretraining of BioCAT5, 16 GPUs are assigned where the batch for each GPU is set to 10 with a maximum sequence length of 256 to reduce the computational complexity. BioCAT5-3B model is trained on 1.06B tokens, where we employ a batch size of 40,960 tokens with a maximum step of 26,000 steps.
- Performance comparison of BioCAT5 with the encoder-only models including BERT, BioBERT, SciBERT, ClinicalBERT, Blue-BERT, and PubMEdBERT and Performance comparison over the encoder decoder-based models including T5base, T5large, SciFIvebase, and SciFivelarge. Our BioCAT5 base model shows comparable performances among all the encoder-only models. In contrast, the BioCAT5-3B parameter model outperforms all the encoder-only and encoder-decoder approaches over the NCBI-Disease, BC5CDR-Disease, BC2GM, and JNLPBA datasets.

成果についてより詳細な情報を提供しているWebページ、発表論文などの情報：
- Accepted Paper in JSAI 2026.
- Paper Under Review in ACL Rolling Review.