[BIOINFO 2025]

From Biomedical Text to Knowledge Graph: A Unified Framework for Entity Recognition, Relation Extraction, and Normalization

Gyujin Son[1], and Mina Rho[1, 2, 3, *]

[1] *Department of Artificial Intelligence, Hanyang University*
[2] *Department of Computer Science, Hanyang University*
[3] *Department of Biomedical Informatics, Hanyang University*
**Corresponding author: minarho@hanyang.ac.kr*

The rapid expansion of biomedical research has resulted in an overwhelming growth of scientific publications containing valuable information on diverse biomedical entities and their interrelations. Efficiently extracting and utilizing such knowledge has become essential but remains challenging, as manual curation cannot keep pace with the scale of available data. Automated natural language processing methods are therefore needed to analyze biomedical texts accurately and systematically. This study introduces an integrated deep learning–based pipeline for biomedical information extraction, consisting of three modules: Named Entity Recognition (NER), Relation Extraction (RE), and Named Entity Normalization (NEN). In the NER stage, sentences are tokenized and processed with pretrained language models to predict IOB tags, enabling the recognition of entities such as diseases, chemicals, and genes. In the RE stage, entity pairs are classified into predefined relation types. To capture contextual information, we adopt a span-based pooling approach for relation representation, followed by multi-class and binary classification strategies. Finally, in the NEN stage, extracted entities are mapped to standardized ontology identifiers by calculating similarity between entity embeddings and ontology dictionary embeddings, with contrastive learning enhancing normalization performance. By combining NER, RE, and NEN in sequence, the proposed framework provides an end-to-end solution for extracting biomedical entities and their relationships, forming the foundation for constructing knowledge graphs. Beyond entity recognition, this pipeline emphasizes relation classification and normalization, allowing for the discovery of novel disease–gene–drug associations and the enrichment of existing ontologies.