From Sparse Frequencies to Semantic Knowledge: A Progressive Framework for High-Quality Knowledge Graph Construction

Name
Xincan Feng

Laboratory's name
Natural Language Processing Lab

Supervisor's name
Taro Watanabe

Abstract
Knowledge graphs face a semantic sparsity paradox: surface sparsity coexists with unexploited semantic redundancy. This dissertation presents a progressive framework from frequency estimation to semantic expansion.
Three contributions: (1) Model-based Subsampling (MBS) replaces counting with embedding predictions for frequency estimation, improving MRR by 2-5% across FB15k-237, WN18RR, and YAGO3-10. (2) Triplet Adaptive Negative Sampling (TANS) unifies probability smoothing, theoretically proving SANS smooths $p(y|x)$ while subsampling smooths $p(x)$, simultaneously optimizing both components through temperature-controlled distributions. (3) LLMKG+ addresses triple-level semantic redundancy in LLM-based KG expansion through retrieval-augmented generation and hierarchical verification combining BERT similarity with LLM reasoning, achieving 20.47%-73.71% QCI improvement on UMLS.
The framework progresses from counting to prediction (MBS), unifies probability smoothing theory (TANS), and extends from statistical to semantic redundancy control (LLMKG+). By combining embedding methods' structural reasoning with LLMs' semantic understanding, we achieve balanced quality-coverage optimization for knowledge graph construction.