Kai **Zhang**

3400 N. Charles St., Baltimore, MD 21218

kzhan118@jhu.edu | +1 (443) 226 3132 | Leadlegend | kaizhang-jhu

Education

MS, Computer Science *2024 - Now*
Johns Hopkins University *Baltimore, MD*

BS, Computer Science (Turing Class) *2019 - 2023*
Peking University *Beijing, China*

- Thesis: Exploring Few-Shot Learning of Large Language Models on Document-level Relation Extraction
- GPA: 3.61/4.00 | Major Rank: 34 / 97 | Main Honor: The Third Prize of Peking University Scholarship

Professional Experience

Tencent AI Lab *Apr. 2021 - Jan. 2022*
Machine Learning Engineer Intern *Beijing, China*

- **Entity Linking**: Implemented and optimized an entity linking model based on knowledge graph Topbase via distributed parallel development on multi-GPU, using data parallel and gradient parallel to improve contrastive learning effectiveness and inference accuracy.
- **Data Efficient Fine-grained NER**: Designed a domain-specific (Sports & Education) semi-supervised NER model based on contrastive learning paradigm Self-Tuning.
  To solve the problem of inefficient Chinese domain-specific data, we introduced training signal annealing, in-domain pretraining and knowledge distillation. The model achieved around 0.70 F-1 accuracy in both domains with low demand for annotated data.
- **Commercial Text Generation**: Designed a controlled text generation model for Tencent Online Reading Platform based on Chinese GPT-2, UER-py framework and mention flags.
  The project has been applied to Tencent's advertisement business and received "Tencent Monthly Innovation Award".
- **Large Knowledge Graph**: Participated in construction and maintenance of multi-lingual universal-domain knowledge graph Topbase.

Knowledge Graph | Information Extraction | LLMs

University of Washington *Apr. 2022 - Aug. 2022*
Research Assistant *Seattle, WA*

- **Paper Implementation**: Reproduced DrugCell, a canonical interpretable model for drug response prediction on cancer cell-line and optimized the model's inference efficiency and prediction accuracy.
- **interpretability of Biomedical Deep Learning**: Investigated the interpretability of neural networks, a critical problem in BioNLP, especially the way of encoding feature and information among neurons in models and how to comprehend it.
  Designed a new interpretable model architecture for drug response prediction: Readable Neural Networks, which extracted contextual text embeddings of Gene Ontology terms from PubMed literatures through distant supervision.

BioNLP | Interpretable DL

Selected Projects

Exploring Few-Shot Learning of Large LMs on Document-level Relation Extraction *Jan. 2023 - Jun. 2023*
Supervisor: Associate Prof. Yansong Feng, Wangxuan Institute of Computer Technology, Peking University *Beijing, China*

- Reviewed the few-shot learning (FSL) performance of large language models (LLMs) on mainstream NLP tasks, and investigated key factors contributing to models' generalization ability, especially their pre-training phases such as instruction tuning and prompt learning.
- Studied the limitations of document-level relation extraction (DocRE) on supervised learning settings, and explored the challenges and benefits of conducting DocRE task on FSL setting.
- Explored the influences of LLMs' DocRE generalization ability by FSL ablation experiments on scientific LLMs suite Pythia, especially the number of samples and model parameter amount.
- Validated the facilitating effect of positive correlation between pre-training corpus and inference data on DocRE task, and conducted experiments to check the correlation saliency for different models.

Development of Commonsense-based Question Generation Models *Jun. 2020 - Oct. 2020*
Supervisor: Associate Prof. Yunfang Wu, Institute of Computational Linguistics, Peking University *Beijing, China*

- Independently designed and implemented a seq-to-seq question generation model, leveraging prior knowledge from knowledge graph to enhance model performance and the quality of generated output.
- Reviewed development of pre-trained NLG methods (BERTsum, BART, ProphetNet, etc.), especially focusing on text summarization, and designed feasible ways to introduce pretraining paradigm into question generation task.

Skills

Languages Python | C/C++ | Java | HTML/css | Bash | SQL | Rust
Developer Tools Docker | Git | Cloud Platform | Hadoop