Quzhe Huang

Homepage : https://andrewzhe.github.io/

Email : huangquzhe@pku.edu.cn
Google Scholar: quzhe's scholar

Education

Peking University, Wangxuan Institute of Computer Technology — Beijing, China
Ph.D. in Computer Science. Advisor: Prof. Yansong Feng and Prof. Dongyan Zhao — *Sep 2019 -*

Peking University, Department of Computer Science — Beijing, China
B.S. in Computer Science — *Sep 2015 - Jun 2019*

Research

- **Document Information Extraction**
  Document Information Extraction (IE) aims to extract structure knowledge, e.g., entity, relation, and events, from a whole paragraph or a passage. Due to the longer context, document IE faces two new challenges compared with sentence-level IE. On the one hand, the useful information for one knowledge instance might scatter on a very large scale rather than only in a short sentence. On the other hand, a document will contain more knowledge instances, thus we need to model the dependency between different instances. To alleviate the above challenges, I have done the following work:

  1. **Filtering useful information with heuristic rules:** Empirically showed that determining the relation between entities in long texts only requires limited evidence, and proposed a method based on co-reference and multi-hop reasoning to select evidence sentences for document-level RE. Equipped with our filtering method, a simple sequence model could perform better than fancy graph neural network based methods[6]. The proposed methods are also useful in alleviating the burden of humans for annotation. We re-annotate the valid dataset of DocRED, and find that DocRED has severe false negative problems and bias, which is due to overly relying on distant supervision[5].

  2. **Modeling the dependency between different instances:** Some instances in a document can help the extraction of the others in the same document. For example, the trigger words of already extracted events are more likely to trigger the same kind of events. However, such a rationale is not perfect and sometimes may be misleading. I propose two auxiliary loss functions to make better use of such imperfect rationale[7]. Considering some instances are harder to extract than others, I also design a simple-to-complex extraction framework, first estimating the difficulty of each instance and then processing from the most simple one[3].

- **Adpating LLM to Specific Domain / Modality**
  In the era of large language models, I try to adapt the general LLM to a specific domain or modality.

  1. **Adapting the general LLM to the legal domain[1]:** We take three steps to adapt the general LLM to the legal domain: injecting domain knowledge, learning reasoning skills, and augmenting the model with legal article retrieval. Specifically, we continue-train the general LLM with a newly collected Chinese legal corpus to help the model learn legal knowledge. Then a small amount of expert-annotated data is used to supervise the model to learn how to solve practical problems. Considering China is adopting civil law where every judgment is based on legal articles, we also incorporate our model with a legal article retrieval module. We propose a framework to adapt a general LLM to a specific domain, and our open-source project has earned about 600 stars on GitHub.

  2. **Constructing Unified Language-Vision Pretraining framework[2]:** Prevailing approaches primarily regard visual input as the prompt and focus exclusively on optimizing the text generation process conditioned upon vision content by a frozen LLM. Such an inequitable treatment of vision and language heavily constrains the model's potential. We break through this limitation by representing both vision and language in a unified representation. To this end, we craft a visual tokenizer that translates the non-linguistic image into a sequence of discrete tokens like a foreign language that LLM can read. Pre-trained on the web-scale image-text corpus, our model LaVIT outperforms the existing models by a large margin on massive vision-language tasks.

Projects

- **Train a model with long context length from scratch (Team Leader)**
  We establish a pipeline encompassing data cleaning, training framework adaptation, model evaluation, and dynamic training strategy adjustment. Using the framework, we train a 7B bilingual model from scratch with an 8K context length and then expand it to a context length of 128K. In the project, I mainly take responses for model training and evaluation. I also implemented the context length expansion.

- **Continue train 70B llama to learn Chinese**
  Considering the outstanding performance of the 70B Llama2 model in English, we were eager to explore its capabilities in the Chinese language. We expand the original vocabulary with Chinese words and continue train the model with a mixture of selected Chinese and English corpora. We find that, even with a very small amount of Chinese data, Llama2

70B demonstrated results close to the state-of-the-art on Chinese evaluation datasets such as CMMLU. After training with billions of tokens, the model is now being used to support downstream applications.

Publication

1. **Lawyer LLaMA: Enhancing LLMs with Legal Knowledge** Under Review
   Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu and Yansong Feng
2. **Unified Language-Vision Pretraining in LLM with Dynamic Discrete Visual Tokenization** Under Review
   Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, **Quzhe Huang**, et, al.
3. **A Progressive Framework for Document-level Informative Argument Extraction** Findings of EMNLP 23
   Quzhe Huang, Yanxi Zhang, Dongyan Zhao
4. **More than Classification: A Unified Framework for Event Temporal Relation Extraction** ACL 23
   Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, Dongyan Zhao
5. **Does Recommend-Revise Produce Reliable Annotations?** ACL 22
   Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, Dongyan Zhao
6. **Three sentences are all you need: Local path enhanced document relation extraction** ACL 21
   Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, Dongyan Zhao
7. **Exploring distantly-labeled rationales in neural network models** ACL 21
   Quzhe Huang, Shengqi Zhu, Yansong Feng, Dongyan Zhao
8. **Relation-Aware Question Answering for Heterogeneous Knowledge Graphs** Findings of EMNLP 23
   Haowei Du, **Quzhe Huang**, Chen Li, Chen Zhang, Yang Li, Dongyan Zhao
9. **Customizing Small Language Model for Dynamic Token Pruning** Findings of EMNLP 23
   Chang Liu, Chongyang Tao, Jianxin Liang, Jiazhan Feng, Tao Shen, **Quzhe Huang**, Dongyan Zhao
10. **Do Charge Prediction Models Learn Legal Theory?** Findings of EMNLP 22
    Zhenwei An*, **Quzhe Huang***, Cong Jiang, Yansong Feng, Dongyan Zhao
11. **Rethinking Task-Specific Knowledge Distillation: Contextualized Corpus as Better Textbook** EMNLP 22
    Chang Liu, Chongyang Tao, Jianxin Liang, Tao Shen, Jiazhan Feng, **Quzhe Huang**, Dongyan Zhao
12. **Knowledge-enhanced Iterative Instruction Generation and Reasoning for Knowledge Base Question Answering** NLPCC 22 (Best paper candidate)
    Haowei Du, **Quzhe Huang**, Chen Zhang, Dongyan Zhao
13. **Why Machine Reading Comprehension Models Learn Shortcuts?** ACL 21
    Yuxuan Lai, Chen Zhang, Yansong Feng, **Quzhe Huang**, Dongyan Zhao
14. **Towards context-aware code comment generation** Findings of EMNLP 20
    Xiaohan Yu, **Quzhe Huang**, Zheng Wang, Yansong Feng, Dongyan Zhao

Service

Area Chair of ARR Oct, 2023
Reviewer of ACL Rollings since 2021.9; ACL 2022-2023, EMNLP 2022-2023; COLING 2022-2023; EACL 2023; AAAI 2023

Honors and Awards

- **President Scholarship (Top 5%)**, Peking University, 2022
- **Uniqlo Scholarship**, Peking University, 2017
- **Panasonic Scholarship**, Peking University, 2016