RESEARCH STATEMENT

Ye Liu (liuyer@mail.ustc.edu.cn)

My research interests encompass a wide range of subjects within the fields of *Knowledge-aware Natural Language Processing (NLP)*, focusing on two main areas: 1) *Knowledge Acquisition* and 2) *Knowledge Application*. I am dedicated to uncovering innovative and efficient solutions by employing statistical methods, machine learning and deep learning algorithms. These methods and algorithms play a critical role in various text understanding and application tasks, such as document categorization, sentiment analysis and chatbot development. By enhancing the ability of machines to comprehend and generate natural language, they enable more sophisticated and reliable interactions between humans and AI. This not only advances the field of natural language processing but also has significant implications for numerous industries and societal applications. With the advent of Large Language Models (LLMs), the importance of mining the knowledge embedded in text and ensuring the reliability of generated text has become increasingly crucial. This shift has heightened the requirements for advanced knowledge-aware natural language processing techniques.

I have been actively exploring the creation of general and efficient methods for knowledge acquisition and application, aiming to develop more powerful and reliable NLP systems. My approach emphasizes reducing the data dependence in the knowledge acquisition phase and focusing more on the generalization ability and robustness during the knowledge application phase. This results in a more versatile and transferable solution for natural language understanding and processing techniques. The overview of my research work is illustrated in Figure 1.

Knowledge Acquisition

In the field of Knowledge-aware NLP, obtaining accurate and effective knowledge is always a key issue. One crucial challenge is the lack of effective labeled data. As this task mainly aims to extract knowledge from these recently published documents, models trained on existing labeled data may be invalid due to domain or semantic gap. Previous research tends to utilize rule-based methods or transfer learning models to mitigate the reliance on labeled data. However, these methods fall short in the use of semantics and knowledge inherent in documents, resulting in poor performance in terms of accuracy or data usage. My research aims to develop a series of solutions that require little to no labeled data to achieve the goal of extracting effective knowledge from unstructured documents. It focuses on two aspects including *Knowledge Concept Extraction* and *Knowledge Relation Extraction*.

1. Knowledge Concept Extraction

The first step of our research is to recognize the knowledge concepts contained in various unstructured documents. To achieve label-free and efficient knowledge concept recognition, I introduce a Hierarchical Multi-aspect Concept Extractor (HMCE) [1][2]. It takes account of the semantic characteristics of knowledge concepts and the potential structure of given documents, such as multi-aspect semantics and multi-level relevances. I further devise a graph-based concept ranking algorithm to assess the quality of candidate concepts, thus achieving robust and accurate knowledge concept recognition. This framework also exhibits superior efficiency, as verified by extensive runtime analysis. This work has been published in the proceedings of the *IEEE International Conference on Data Mining (ICDM-2020)* and *ACM Transactions on Knowledge Discovery from Data (ACM TKDD-2023).* It has received positive comments from the community.

2. Knowledge Relation Extraction

With the extracted knowledge concept, the next target is to determine the relation categories between different concepts. Traditional deep learning methods are significantly hampered by a lack of necessary prior knowledge, while large language models fall short in their task-specific capabilities for relation extraction. To address these shortcomings, I propose a Dual-System Augmented Relation Extractor (DSARE) [3], which synergistically combines traditional methods with LLMs. Specifically, DSARE imparts the prior knowledge inherent in LLMs to the traditional models while simultaneously transferring the traditional model's understanding of relation extraction to LLMs, facilitating the creation of a mutually reinforcing model system. In practice, DSARE demonstrates the optimal performance, especially in extremely low-resource scenarios. This work has been published in the proceedings of the *International Conference on Database Systems for Advanced Applications (DASFAA)* in 2024.

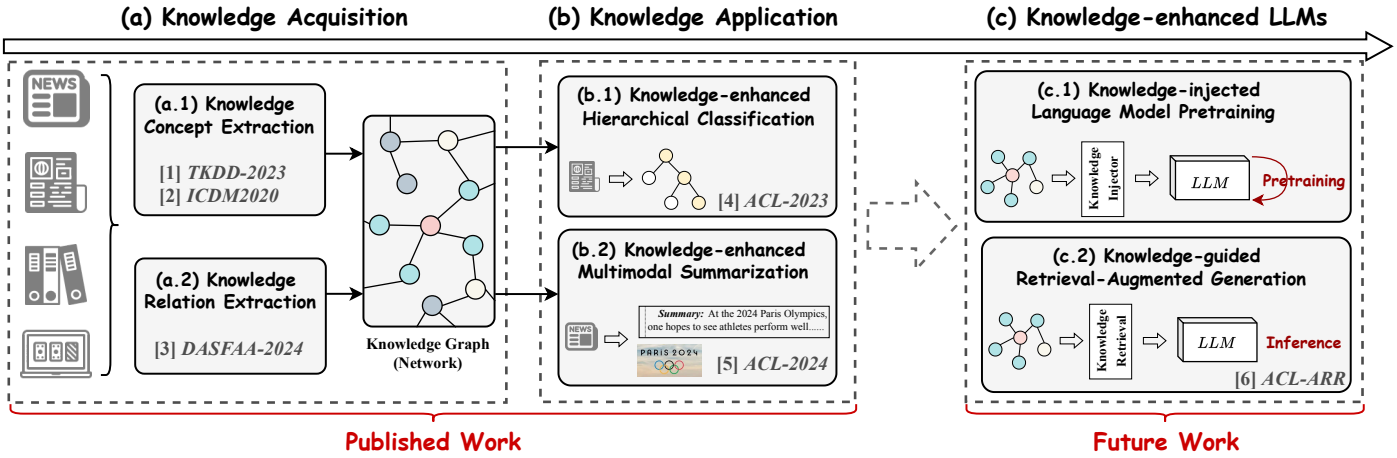

Figure 1: The Architecture of My Research

Knowledge Application

With the constructed knowledge network (knowledge graph), researchers can easily retrieve effective and relevant knowledge to solve various downstream tasks, particularly those that are domain-specific and challenging. Previous works focused on combining the representation of the original module with external knowledge. In my research, I am dedicated to exploring closer and finer interactions between external knowledge and specific downstream tasks, thereby making fuller use of the knowledge and achieving more competitive performance.

1. **Knowledge-enhanced Hierarchical Classification Framework**

   Hierarchical Text Classification (HTC) is an essential and challenging subtask of multi-label text classification with a taxonomic hierarchy. Traditional methods have significant limitations due to the lack of domain knowledge, particularly for the classification at lower levels of the hierarchy. In my research [4], I attempt to incorporate the extracted knowledge into the hierarchical classification process. Specifically, I innovatively integrate knowledge into both the text representation and hierarchical label learning process, addressing the knowledge limitations of traditional methods. Additionally, a novel knowledge-aware contrastive learning strategy is proposed to further exploit the information inherent in the data. Extensive experiments demonstrate the efficacy of our proposed method, and further indicate the necessity of incorporating external knowledge graphs in HTC tasks, especially for the classification at deeper and more difficult levels. Our work was published in the proceedings of the *Findings of The Annual Meeting of The Association for Computational Linguistics (ACL-Findings-2023)*.

2. **Knowledge-enhanced Multimodal Summarization**

   Multimodal Summarization with Multimodal Output (MSMO) aims to produce a multimodal summary with a textual abstract alongside a pertinent image. Traditional approaches typically adopt a holistic perspective on coarse image-text data or individual visual objects, overlooking the essential connections between objects and the entities they represent. To address this, we propose an Entity-Guided Multimodal Summarization model (EGMS) [5]. EGMS utilizes dual multimodal encoders with shared weights to process text-image and entity-image information concurrently. Further, a gating mechanism combines visual data for knowledge entity-enhanced summary generation. Experiments validate its superiority and indicate the necessity of integrating knowledge entity information. This work has been published in the *Findings of The Annual Meeting of The Association for Computational Linguistics (ACL-Findings)* in 2024.

Industry Application Experience

As a researcher with a passion for practical applications of NLP technology in the industry, I have had the honor of interning at Huawei Cloud & AI (Jan. 2019 – Aug. 2019) and ByteDance AI Lab (Feb. 2023 – Aug. 2023). At Huawei, I focused on developing an entity-guided question generation model. It employs the given entities in text as anchors to guide the question generation through a sequence-to-sequence (seq2seq) paradigm, overcoming the shortcomings of traditional question generation methods, such as irrelevant questions. This model has been deployed in the Huawei's annotation tool, effectively reducing the workload of annotators and delivering remarkable outcomes. At ByteDance, I explored the preference learning of LLMs. Specifically, I trained a reward model with approximately 30,000 ranking pairs labeled by annotators. Then, using

the RLHF (Reinforcement Learning from Human Feedback) algorithm, we optimized a finetuned generation model (13B size). This model has been deployed to generate summary descriptions of entities in the online encyclopedia knowledge graphs. Subsequently, with this KG, we developed a knowledge enhanced question answering (QA) system, which significantly mitigated the hallucination problem of LLMs. This QA system has also been deployed in ByteDance's internal system.

Future Research Direction

In future, I will continue exploring the knowledge application in various scenarios, with a focus on the Knowledge-enhanced Large Language Models, as illustrated in Figure 1 (c). Although LLMs exhibit rich prior knowledge and powerful reasoning abilities, the hallucination problem and the lack of long-tail knowledge remain significant challenges that impede the boarder application of LLMs. To address these issues, I plan to incorporate external knowledge from various sources, such as knowledge graphs and search engines, into the LLMs' training/inference processes. This will help mitigate the hallucination drawbacks and elicit the complex reasoning abilities of LLMs.

1. **Knowledge-injected Language Model Pretraining**

   Most existing LLMs, such as GPT-4 and Llama-3, are trained on various text corpora, including books and documents. This paradigm relies on LLMs to learn knowledge from unstructured documents and store it in the model parameters. However, due to the capacity limitations of LLMs and the forgetting phenomenon during the learning process, these LLMs often fail when faced with many domain-specific cases and situations requiring complex reasoning abilities. I will attempt to inject structured knowledge, including but not limited to domain knowledge triples, commonsense, long-tail and long-context knowledge, into the pretraining process of LLMs. This approach aims to enhance the complex reasoning abilities of LLMs, resulting in more powerful and reliable models.

2. **Knowledge-guided Retrieval-Augmented Generation (RAG)**

   The training process of LLMs requires powerful computing sources, which are often beyond the reach of most researchers worldwide. Fortunately, the powerful reasoning abilities of LLMs offer another way to integrate external knowledge: Retrieval-Augmented Generation (RAG). I have already made attempts in this direction, such as fake news detection [6]. Along this line, a key issue is how to ensure the correctness of retrieved information, especially in this age of disinformation. I will propose a knowledge-guided RAG system, using the concrete knowledge from knowledge graphs to judge and select retrieved information and address conflicts between different sources. This approach aims to construct a more reliable and robust RAG system, overcoming hallucination and knowledge limitations.

References

[1] Ye Liu, Han Wu, Zhenya Huang, Hao Wang, Yuting Ning, Jianhui Ma, Qi Liu, Enhong Chen. TechPat: Technical Phrase Extraction for Patent Mining, *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*, volume 17, number 9, pages 1-31, 2023.

[2] Ye Liu, Han Wu, Zhenya Huang, Hao Wang, Jianhui Ma, Qi Liu, Enhong Chen, Hanqing Tao and Ke Rui. Technical Phrase Extraction for Patent Mining: A Multi-level Approach, *IEEE International Conference on Data Mining (ICDM)*, pages 1142-1147, 2020.

[3] Ye Liu, Kai Zhang, Aoran Gan, Linan Yue, Feng Hu, Qi Liu, Enhong Chen. Empowering Few-Shot Relation Extraction with The Integration of Traditional RE Methods and Large Language Models, *The International Conference on Database Systems for Advanced Applications (DASFAA)*, 2024.

[4] Ye Liu, Kai Zhang, Zhenya Huang, Kehang Wang, Yanghai Zhang, Qi Liu, Enhong Chen. Enhancing Hierarchical Text Classification through Knowledge Graph Integration, *Findings of The Annual Meeting of The Association for Computational Linguistics (ACL-Findings)*, pages 5797-5810, 2023.

[5] Yanghai Zhang, Ye Liu, Shiwei Wu, Kai Zhang, Xukai Liu, Qi Liu, Enhong Chen. Leveraging Entity Information for Cross-Modality Correlation Learning: The Entity-Guided Multimodal Summarization, *Findings of The Annual Meeting of The Association for Computational Linguistics (ACL-Findings)*, 2024.

[6] Ye Liu, Jiajun Zhu, Kai Zhang, Haoyu Tang, Yanghai Zhang, Xukai Liu, Qi Liu, Enhong Chen. Detect, Investigate, Judge and Determine: A Novel LLM-based Framework for Few-shot Fake News Detection, Submitted to *The Annual Meeting of The Association for Computational Linguistics (ACL) ARR 2024 June*.