Research Statement

Yixin Cao
School of Computing and Information Systems,
Singapore Management University
Tel: (65) 6826-4871; Email: yxcao@smu.edu.sg
16/12/2023

Background

Natural Language Processing (NLP) is a key subfield of Artificial Intelligence. Although Large Language Models (LLMs) have achieved great success and are becoming more and more important in many areas, they still suffer from several issues towards Artificial General Intelligence (AGI), such as unclear working mechanism, hallucinations, poor reasoning. My objective is to develop efficient algorithms and systems for the automated analysis and understanding of knowledge within LLMs, and leverage knowledge engineering to improve LLMs' trustworthiness and reasoning skills. I am exploring the following key scientific questions:

- How to automatically evaluate and interpret large language models to enhance their capabilities?
- How to manage knowledge as an external memory to verify and correct factual errors (i.e., hallucinations) in predictions?
- How to utilize symbolic knowledge to assist large language models in conducting interpretable and rigorous reasoning?

My expertise in NLP and knowledge engineering, together with my critical and creative thinking, has successfully bolstered this new research direction and yielded notable research outcomes. Along this direction, I have published over 40 papers in top venues, accrued more than 5,000 citations.

Research Areas

To do so, I propose to develop a knowledgeable language agent. In particular, I focus on the following three research areas: Understanding LLMs, Knowledge-patched LLMs, and neural-symbolic reasoning.

A. Understanding LLMs

How to understand and assess the abilities of LLMs are becoming more and more critical. On the one hand, there is a practical need in training and evaluating current LLMs. On the other hand, this will help us to design more advanced LLMs for improvements. However, the current evaluation pipeline heavily relies on human efforts in curating the datasets and evaluating the outputs. This is not only costly, but also may raise data-leakage issues --- pretraining data may include the testing data, leading to an over-estimation of the performance [WZZ23, L23, ZZC23].

Therefore, we have conducted extensive experiments to assess both open-sourced and closed-sourced LLMs [SCW23, XDC23, MCH23], and propose a LLM-as-an-Examiner framework to automate the evaluation pipeline including dataset curation and an evaluator [BYC23]. In specific, for a more comprehensive and equitable

evaluation, we devise three strategies: (1) we instruct the LM examiner to generate questions across a multitude of domains to probe for a broad acquisition, and raise follow-up questions to engage in a more in-depth assessment. (2) Upon evaluation,

the examiner combines both scoring and ranking measurements, providing a reliable result as it aligns closely with human annotations. (3) We additionally propose a decentralized Peer-examination method to address the biases in a single examiner.

B. Knowledge-patched LLMs

This research direction focuses on knowledge acquisition and application, as well as pre-trained language models. My major contributions revolve around the following themes, including:

- Proposing a comprehensive new methodology for knowledge fusion, extraction, and completion.
- Patching LLMs with prior knowledge guidance.

In particular, I propose the following techniques.

Low-resource Information Extraction (IE). My early works focus on entity-centric IE. The goal is to identify entities and their relations from texts, involving three sub-tasks of Named Entity Recognition (NER), Relation Extraction (RE), and Entity Linking (EL). However, the annotations are not always available or very expensive due to the background requirement for annotators. Therefore, we focus on automatically labeled data, with two main challenges of noise and long-tail distribution. For NER, we propose Partial-CRFs with non-entity sampling model [CHC19] that can well use the partial labeled sentences — not all the words in the sentence have corresponding labels. And it achieves a good balance between effectiveness and efficiency via pre-training on massive noisy data while finetuning on relative high-quality annotations. Furthermore, we differentiate the miscellaneous semantics of non-entity words via prototypical network and clustering for few-shot NER [TWB21]. Then, to link entities to KG (EL), we are the first to incorporate Graph Convolutional Network (GCN) towards consistent semantics for global EL [CHL18]. We disambiguate those entities based on not only the textual contexts, but also their neighbor entities in KG. Finally, for RE, targeting the noise issue, we systematically investigate attention's working mechanism to denoise [HCH21]. We found that attention can effectively filter out noise but also harm the robustness. So, when the noise issue is severe, we propose to remove attention and introduce entity features to improve RE robust to noisy sentences. We have built a complete IE pipeline, and, without any human efforts, our NER system achieves 6.6% F1 gains on five low-resource food types as well as multi-lingual low resource settings.

Recently, we also proposed a general few-shot IE framework [MWC23], which achieved the current state-of-the-art, and explored the abilities of LLMs in IE and propose a combination framework of LLMs and small-sized LMs, to achieve a good balance between performance and cost [MCH23].

Pushing IE into Event-centric and Document-level. Since events have better expressiveness than entities while usually scattered across the entire document, my recent work tends to push conventional entity-centric IE into event-centric and document-level [YHC23, HYL23]. There are two subtasks: event detection and event relation identification. The main challenges here, compared with entity-centric IE, are the complex event ontology and long-text processing. For event detection, we first rethink and reformulate the recognition of event trigger word as a word sense disambiguation problem, which has many mature techniques. Then, to extract arguments for the event trigger, we propose PAIE [MWC22] that takes the best advantage of PLMs to alleviate the annotation burden via prompt tuning. Besides, we design multi-argument prompt template that can predict all arguments together to capture their interactions for better performance and efficiency. For event relation identification, we build an event relational graph [CCD22, CCZ23] for cross-sentence reasoning at the document-level, where each node denotes an event, or a pair of events and each edge denotes possible interactions. Thus, via graph reasoning and node classification, we improve the F1 score around 10% on average.

Multi-Modal Event KG. Based on the above techniques, we have built the first large-scale multi-modal event KG [MWL22]. It not only has powerful expressive ability, e.g., procedure knowledge, but also can naturally bridge multiple modalities of data for deep understanding and comprehensive reasoning. It contains over 990,123 concept events and 863 million instance events. These events are connected via 644 types of 934 million relations. Besides, the byproduct of IE system has shown state-of-the-art performance on several public available datasets.

Self-guided Entity Alignment (EA) aims to merge different KGs by finding equivalent entities between graphs. Conventional methods usually encode the graph via GNN or knowledge representation learning to preserve the structure information, and then push the embeddings of equivalent entities together based on a set of pre-matched seed pairs. However, the structures of the same entities may be very different between KGs. And, the training of GNN is not efficient due

to the issue of recursive neighbor expansion — to obtain the final representation of a single node, it needs all node embeddings in its L-hop neighborhood.
We thus utilize the two KGs as the guidance of each other, so that the proposed model will complete each KG towards more consistent structures and be more robust. We first utilized cross-graph attention and regularized the GNN with relations [LCH20]. Second, we proposed to infer a universal rule set from both KGs, which serves as a pivot to transfer meta-knowledge between each other [CLL19]. Third, we introduce entity attributes and their values as weak matching signals [LCP20]. More recently, to improve the efficiency of GNN, our work provides both theoretical foundation and empirical evidence of training-free graph matching framework [LCF22]. This work improves the alignment accuracy by 5.1% on cross-lingual settings and to a 98% score on mono-lingual settings. As a result of these efforts, we have successfully constructed a large-scale cross-lingual KG with balanced number of English entities and Chinese entities, which also doubled the inter-language links as compared with the biggest Multi-lingual KG Wikidata. We have also constructed a cross-domain KG targeting wellness analytics. It has successfully included 4 domains of food, lifestyle, nutrition, and disease with over 1 million facts.

KG-guided NLP includes the following techniques.

KG-guided IE. One of the main challenges in IE is the expertise requirements of annotators. Thus, my research proposal is to introduce existing knowledge to automatically obtain either explicit or implicit weak supervisions. Learn new knowledge based on what we have learnt. As mentioned in the last section, our IE system has already used entity features to provide background knowledge. To further improve RE performance on long-tail relation types, we propose to introduce structured knowledge for better reasoning by learning relation prototypes [KCZ20, CKG21], which can capture relation proximity based on their corresponding entity pairs in KG, so as to facilitate knowledge transfer from relation types with sufficient annotations to those long-tail relations. Besides, for event-centric IE, we leverage lexicon knowledge to supervise the event detection model via a teacher student network [TWC20].

KG-guided Pre-train Language Model (PLM). Natural language is a key intermediary of human knowledge. Therefore, I am interested in injecting knowledge into the fundamental text encoder --- PLMs. I am one of the earliest researchers that propose a jointly model for KG and texts. I aim at learning a cross-lingual common semantic space, which is a very interesting idea to embed similar words and entities, no matter in which languages, closely in the vector space [CHJ17, CHL18]. Moreover, inspired by our KG-guided IE works, we extend distant supervision to cross-lingual settings, so that it doesn’t require parallel corpora, but relies on automatically generated comparable sentences for training. The idea of cross-lingual common semantic space has been successfully used by many IE systems, such as the RPI system that ranked within top3 for 13 languages entity discover and linking challenge in TAC-KBP2017.

More recently, I have tried to design multiple knowledge-driven task from KG, which will be used for pre-training large language models. However, I realize that it is against the objectives of PLMs --- easier tasks, more parameters, more data. Therefore, my recent research focuses on combining KG reasoning and PLM via prompt tuning, instead of finetuning. [LCH22] has developed two modules to incorporate commonsense knowledge. One is for generating commonsense prompts, and the other aims at predicting next event by taking the prompts. The basic idea is that it is the intermediate character psychology (e.g., intents and feelings) of current event that pushes the story forward for next events. [LLC22] systematically investigates the issues of using PLMs for knowledge reasoning and proposes a new benchmark. In the future, we are particularly interested in this direction and our ongoing work is to prompt scalable and trustable knowledge reasoning (next section) into large PLMs without training or finetuning.

C. Neural-symbolic Reasoning

Symbolic knowledge reasoning. Given a KG, we aim at inferring new knowledge that is not only correct, but also inferred in a trustworthy reasoning process. Knowledge reasoning is typically evaluated via the task of KGC, which learns entity/event/relation embeddings by preserving the structures and capturing latent inference pattern from UKG. Formally, given an incomplete triplet (h, r, ?) or (?, r, t), KGC is to predict the missing entity based on their embeddings.

Although there are many KGC models, only a handful of works focus on reasoning trustworthiness. We are one of the earliest works that systematically analyze the issue of KG reasoning reliability and contribute two datasets for quantitatively evaluation [LCH21, CJL21], which still shows a large gap to human's results and provides valuable findings to address the issue. Based on these insights, our recent work thus proposes to capture high-order relations among relations (i.e., rules) to enhance model's reasoning ability [CCF22]. It achieves the state-of-the-art performance and shows a great generalization ability in sparse settings. [XHC23] further focuses on modeling knowledge uncertainty via normalization flow and provides theoretical proof. In the future, we are very interested in the bias and generalization of knowledge reasoning towards trustworthiness.

Integration of LLMs and symbolic reasoning. Neurons and symbols are building blocks of two different AI paradigms, namely, neural networks (connectionism) and symbolicism. They are complementary: neural networks can learn from big data and predict based on latent space inference, but remains black box systems. Conversely, symbolicism strictly follows the pre-defined ontology, conducting reasoning rigorously and interpretably, but it falls short in generalization ability. These two directions will ideally meet somewhere in the middle and will lead to representations that can act as a bridge for novel neural computing.

To push forward this direction, we have explored multi-agent intelligence to improve the reasoning ability of LLMs (taking language as symbols) [XDC23], and host a workshop at Coling2024.

Selected Publications and Outputs

[WZZ23] Wei, T., Zhao, L., Zhang, L., Zhu, B., Wang, L., Yang, H., ... & Zhou, Y. (2023). Skywork: A more open bilingual foundation model. arXiv preprint arXiv:2310.19341.

[L23] Li, Y. (2023). An open source data contamination report for llama series models. arXiv preprint arXiv:2310.17589.

[ZZC23] Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., ... & Han, J. (2023). Don't Make Your LLM an Evaluation Benchmark Cheater. arXiv preprint arXiv:2311.01964.

My outputs

[SCW23] Shui, R., **Cao, Y.**, Wang, X., & Chua, T. S. (2023). A Comprehensive Evaluation of Large Language Models on Legal Judgment Prediction. arXiv preprint arXiv:2310.11761.

[XDC23] Xiong, K., Ding, X., **Cao, Y.,** Liu, T., & Qin, B. (2023). Examining the Inter-Consistency of Large Language Models: An In-depth Analysis via Debate. arXiv e-prints, arXiv-2305.

[MCH23] Ma, Y., Cao, Y., Hong, Y., & Sun, A. (2023). Large language model is not a good few-shot information extractor, but a good reranker for hard samples!. arXiv preprint arXiv:2303.08559.

[BYC23] Bai, Y., Ying, J., **Cao, Y**., Lv, X., He, Y., Wang, X., ... & Hou, L. (2023). Benchmarking Foundation Models with Language-Model-as-an-Examiner. arXiv preprint arXiv:2306.04181.

[CHC19] **Yixin Cao**, Zikun Hu, Tat-seng Chua, Zhiyuan Liu and Heng Ji. Low-Resource Name Tagging Learned with Weakly Labeled Data. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.

[TWB21] Meihan Tong, Shuai Wang, Bin Xu, **Yixin Cao**, Minghui Liu, Lei Hou and Juanzi Li. Learning from Miscellaneous Other-Class Words for Few-shot Named Entity Recognition. Annual Meeting of the Association for Computational Linguistics (ACL), 2021.

[CHL18] **Yixin Cao**, Lei Hou, Juanzi Li and Zhiyuan Liu. Neural Collective Entity Linking. International Conference on Computational Linguistics (COLING), 2018.

[HCH21] Zikun Hu, **Yixin Cao**, Lifu Huang and Tat-Seng Chua. How does Knowledge Graph and Attention Help? A Qualitative Analysis into Bag-level Relation Extraction. Annual Meeting of the Association for Computational Linguistics (ACL), 2021.

[HYL23] Huang, H., Yuan, C., Liu, Q., & Cao, Y. (2023). Document-level Relation Extraction via Separate Relation Representation and Logical Reasoning. ACM Transactions on Information Systems, 42(1), 1-24.

[YHC23] Yuan, C., Huang, H., Cao, Y., & Wen, Y. (2023). Discriminative reasoning with sparse event representation for document-level event-event relation extraction. ACL.

[CCZ23] Chen, M., Cao, Y., Zhang, Y., & Liu, Z. (2023). CHEER: Centrality-aware high-order event reasoning network for document-level event causality identification.
[MWC23] Ma, Y., Wang, Z., Cao, Y., & Sun, A. (2023). Few-shot Event Detection: An Empirical Study and a Unified View. arXiv preprint arXiv:2305.01901.
[MWC22] Yubo Ma† , Zehao Wang† , **Yixin Cao***, Mukai Li, Meiqi Chen, Kun Wang, Jing Shao. Prompt for Extraction? PAIE: Prompting Argument Interaction for Event Argument Extraction. ACL2022.
[CCD22] Meiqi Chen, **Yixin Cao**, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, Yan Zhang. ERGO: Event Relational Graph Transformer for Document-level Event Causality Identification. COLING2022
[MWL22] Yubo Ma†, Zehao Wang†, Mukai Li†, **Yixin Cao†***, Meiqi Chen, Xinze Li, Wenqi Sun, Kunquan Deng, Kun Wang, Aixin Sun, Jing Shao. MMEKG: Multi-modal Event Knowledge Graph towards Universal Representation across Modalities. ACL2022 Demo
[LCH20] Chengjiang Li, **Yixin Cao**, Lei Hou, Jiaxin Shi, Juanzi Li and Tat-Seng Chua. Semi-supervised Entity Alignment via Joint Knowledge Embedding Model and Cross-graph Model. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.
[CLL19] **Yixin Cao**, Zhiyuan Liu, Chengjiang Li, Zhiyuan Liu, Juanzi Li and Tat-Seng Chua. Multi Channel Graph Neural Network for Entity Alignment. Annual Meeting of the Association for Computational Linguistics (ACL), 2019.
[LCP20] Zhiyuan Liu, **Yixin Cao**, Liangming Pan, Juanzi Li, Zhiyuan Liu and Tat-Seng Chua. Exploring and Evaluating Attributes, Values, and Structure for Entity Alignment. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.
[LCF22] Zhiyuan Liu, **Yixin Cao**, Fuli Feng, Xiang Wang, Jie Tang, Kenji Kawaguchi and Tat-Seng Chua. Training Free Graph Neural Networks for Graph Matching. arXiv preprint arXiv:2201.05349 (2022).
[KCZ20] Jun Kuang, **Yixin Cao**, Jianbing Zheng, Xiangnan He, Ming Gao and Aoying Zhou. Improving Neural Relation Extraction with Implicit Mutual Relations. The annual IEEE International Conference on Data Engineering (ICDE), 2020.
[CKG21] **Yixin Cao**, Kuang Jun, Ming Gao, Aoying Zhou, Yonggang Wen and Tat-Seng Chua. Learning Relation Prototype from Unlabeled Texts for Long-tail Relation Extraction. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2021.
[TWC20] Meihan Tong, Shuai Wang, **Yixin Cao**, Bin Xu, Lei Hou, Juanzi Li and Jun Xie. Improving Event Detection via Open-domain Event Trigger Knowledge. Annual Meeting of the Association for Computational Linguistics (ACL), 2020.
[CHJ17] **Yixin Cao**, Lifu Huang, Heng Ji, Xu Chen, Juanzi Li. Bridge text and knowledge by learning multi-prototype entity mention embedding. Annual Meeting of the Association for Computational Linguistics (ACL), 2017.
[CHL18] **Yixin Cao**, Lei Hou, Juanzi Li, Zhiyuan Liu, Chengjiang Li, Xu Chen and Tiansi Dong. Joint Representation Learning of Cross-lingual Words and Entities via Attentive Distant Supervision. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.

[LCH22] Lin Li, **Yixin Cao***, Lifu Huang, Shu'ang Li, Lijie Wen. What Makes The Story Forward? Inferring Commonsense Explanations as Prompts for Future Event Generation. SIGIR2022.

[LLC22] Xin Lv, Yankai Lin, **Yixin Cao**, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, Jie Zhou. Do Pre-trained Models Benefit Knowledge Graph Completion? A Reliable Evaluation and a Reasonable Approach. ACL2022 Finding.

[LCH21] Xin Lv, **Yixin Cao**, Lei Hou, Juanzi Li, Zhiyuan Liu, Yichi Zhang and Zelin Dai. Is Multi-Hop Reasoning Really Explainable? Towards Benchmarking Reasoning Interpretability. EMNLP2021.

[CJL21] **Yixin Cao**, Xiang Ji, Xin Lv, Juanzi Li, Yonggang Wen, and Hanwang, Zhang. Are Missing Links Predictable? An Inferential Benchmark for Knowledge Graph Completion. ACL2021.

[CCF22] Weijian Chen, **Yixin Cao**, Fuli Feng, Xiangnan He, and Yongdong Zhang. Explainable Sparse Knowledge Graph Completion via High-order Graph Reasoning Network. arXiv preprint arXiv:2207.07503.

[XHC23] Changyi Xiao, Xiangnan He, **Yixin Cao**. Knowledge Graph Embedding by Normalizing Flows. AAAI2023.