Yong Cao

Huazhong University of Science and Technology, Wuhan, China.
+45 5273-0171 yongcao2018@gmail.com yongcaoplus.github.io update 26th Mar 2024

Research Area and Interests

NLP, Cultural Adaptation in LLMs, Multilinguality, Dialogue System, Semantic Representation, KBQA

Education

University of Copenhagen **Nov 2022 – Apr 2024**
Visiting Ph.D. student, Supervisor: Daniel Hershcovich. *Copenhagen, Denmark*

Huazhong University of Science and Technology **Sep 2018 – Jun 2024**
Ph.D. student of Computer Science, Supervisor: Min Chen. *Wuhan, China*

Chinese University of Hong Kong **Jul 2020 – Nov 2020**
Visiting Ph.D student, Supervisor: Kai Hwang. *Shenzhen, China*

Sichuan University **Sep 2014 – Jun 2018**
Bachelor of Telecommunication Engineering, Rank: 1/60 (1%). *Sichuan, China*

Experience

Ali-DAMO-AI Company **Jun 2022 – Oct 2022**
Intern of NLP Algorithm Researcher *Hangzhou, China*

Xiaomi-AI Company **Nov 2021 – Jun 2022**
Intern of NLP Algorithm Researcher *Wuhan, China*

Deepwisdom-AI Company **Jan 2021 – Jun 2021**
Intern of NLP Algorithm Engineering *Shenzhen, China*

Published Paper

1. **Yong Cao**, Ruixue Ding, Boli Chen, Xianzhi Li, Min Chen, Daniel Hershcovich, Pengjun Xie, Fei Huang. "Geo-Encoder: A Chunk-Argument Bi-Encoder Framework for Chinese Geographic Re-Ranking", EACL 2024 main.
2. **Yong Cao**, Min Chen, Daniel Hershcovich, "Bridging Cultural Nuances in Dialogue Agents through Cultural Value Surveys", findings of EACL 2024.
3. **Yong Cao**, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, Daniel Hershcovich. Cultural Adaptation of Recipes. Transactions of the Association for Computational Linguistics (TACL).
4. **Yong Cao**, Xianzhi Li, Huiwen Liu, Wen Dai, Shuai Chen, Bin Wang, Min Chen, Daniel Hershcovich, "Pay More Attention to Relation Exploration for Knowledge Base Question Answering.", In Findings of ACL 2023.
5. **Yong Cao**, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, Daniel Hershcovich. "Assessing cross-cultural alignment between chatgpt and human societies: An empirical study", In Proceedings of the Workshop on Cross-Cultural Considerations in NLP, EACL 2023.
6. **Yong Cao**, Wei Li, Xianzhi Li, Min Chen, Guangyong Chen, Long Hu, Zhengdao Li, Hwang Kai, "Explore More Guidance: A Task-aware Instruction Network for Sign Language Translation Enhanced with Data Augmentation", In Findings of NAACL 2022.
7. **Yong Cao**, R. Wang, M. Chen, A. Barnawi, "AI Agent in Software-defined Network: Agent-based Network Service Prediction and Wireless Resource Scheduling Optimization", IEEE Internet of Things Journal, DOI: 10.1109/JIOT.2019.2950730, 2019.
8. Zhengdao Li, **Yong Cao**, Kefan Shuai, Yiming Miao, Kai Hwang, "Rethinking the Effectiveness of Graph Classification Datasets in Benchmarks for Assessing GNNs", IJCAI 2024 under review.
9. Andrea Morales-Garzón, **Yong Cao**, Daniel Hershcovich, "Consistency Evaluation of Recipe Cultural Adaptation", LREC-COLING 2024 under review.
10. M. Chen, **Yong Cao**, R. Wang, Y. Li, D. Wu, Z. Liu, "DeepFocus: Deep Encoding Brainwaves and Emotions with Multi-scenario Behavior Analytics for Human Attention Enhancement", IEEE Network, Vol. 33, No. 6, 2019.
11. Rui Wang, **Yong Cao**, Adeeb Noor, Thamer A.Alamoudi, Redhwan Nour. "Agent-enabled task offloading in UAV-aided mobile edge computing", Computer Communications 149, 324-331.
12. M. Chen, Y. Jiang, **Yong Cao**, A. Y. Zomaya, "CreativeBioMan: Brain and Body Wearable Computing based Creative Gaming System", IEEE Systems, Man, and Cybernetics Magazine, Vol. 6, No. 1, pp. 14-22, Jan. 2020.

13. Li Zhou, Wenyu Chen, **Yong Cao**, etc, "MLPs Compass: What is learned when MLPs are combined with PLMs?", ICASSP 2024.
14. Tarik Alfif, Bander Alzahrani, **Yong Cao**, Reem Alotaibi, Ahmed Barnawi, Min Chen, "Generative Adversarial Network BasedAbnormal Behavior Detection in Massive CrowdVideos: A Hajj Case Study".Journal of Ambient Intelligence and Humanized Computing, 2021: 1-12.
15. Tianshu Hao, Jianfeng Zhan, Kai Hwang, **Yong Cao**, "Edge AiBench:Scenario-Based AI Benchmarking for Cloud/Edge/Device Computing", IEEE Transactions on Computers, 2023.
16. Jun Yang, Jiayi Lu, Yiming Miao, Lu Wang, Yiting Zhao, **Yong Cao**, "The Effective Recycling of Crashed Drone Based on Machine Intelligence", 14th International Wireless Communications & Mobile Computing Conference (IWCMC), 2018.

Academic Activities

Program Committee: AAAI 2022, 2023, 2024, EACL 2023, ARR 2024, CoLM 2024.
Co-Organizer: Cross-Cultural Considerations in NLP workshop @ ACL 2024.
Media: "Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study" was picked up by Politiken, Børsen, Ekstra Bladet, P1 Morgen, TV 2 (Denmark) and Science et Avenir (France).

Talk

2024.03, "Cultural Considerations in Large Language Models", Max Planck Institute for Human Development.
2024.02, "Cultural Considerations in Dialogue Systems", University of Marburg.
2023.10, "Cultural Adaptation of Large Language Models", University of Copenhagen.

Projects

Cultural Adaption in LLMs | *Copenhagen University* **Nov 2022 - Apr 2024**

- Consider the cultural difference in large language models (LLMs), and study the cultural alignment between LLMs and human societies based on human society surveys.
- Construct cultural adaption benchmark datasets, and propose solutions to the cultural adaptation on specific domains (e.g. recipe) and general domains (e.g. dialogue).

Modeling of Abnormal Behavior of Large-Scale Crowd | *Collaboration with KAUST* **Dec 2019 - Dec 2022**

- Collect large-scale crowd abnormal behavior benchmark dataset on hajj scenario.
- Design abnormal behavior classification algorithm based on the optical flow algorithm and GAN model.

Product of Sequence Tagging Based on AutoML | *Deepwisdom* **Mar 2021 - May 2021**

- Develop sequence tagging product based on AutoML, and automatically construct the task pipeline in four stages: Data EDA, offline training, testing and online prediction. The algorithm is implemented based on PyTorch and Keras framework and is deployed through Docker.
- Expand the basic operators of the model to 30+, and evaluate the product performance based on 10+ benchmark.
- Seven basic NLP SaaS services are developed based on this product.

Resume Parsing System for Multi-Source Unstructured Data | *Deepwisdom* **May 2021 - Jun 2021**

- Build PDF parsing and word parsing operator based on the company's resume data, extract information based on entity extraction algorithm and rule matching, and establish a resume parsing prototype system.
- Optimize specific rules for specific field extraction (eg. educational background and work experience), and merge the algorithm into the company's platform.

Early Warning System and Intervention Strategy for Depression | *National key R&D plan* **Nov 2018 - Jun 2021**

- Develop collection and storage scheme of multi-modal dataset from real depression patients and normal volunteers.
- The depression diagnosis model was established based on EEG signal, near-infrared signal, video and audio data, and the depression diagnosis result was realized based on prediction fusion.
- Develop intelligent follow-up system, and construct the development of psychological counseling and intervention robot.

Honors and Awards

- 2023 DAAD AInet Fellow on Human-centered AI.
- 2023 Outstanding PhD Scholarship of HUST.
- 2021 International Youth Talent Fund by Zhejiang Lab, Hangzhou, China.
- 2020 Zhixing Scholarship of HUST.
- 2019 Outstanding Student of HUST.
- 2018 Outstanding Graduates of SCU.
- 2017,2016 National Scholarship.
- 2017 Outstanding student cadre of SCU.
- 2016 Excellent Paper Award in the National Mathematical Modeling Challenge.
- 2015 Outstanding Students of SCU.

NON-NUMERIC VARIABLE EXPLORATION AND TRANSFER LEARNING IN CLINICAL SUBGROUP DISCOVERY

Clinical subgroup discovery involves using data mining techniques to explore meaningful relationships among cases within a dataset, focusing on a specific property, often linked to a medical condition. With dataset Y as the basis, the objective is to identify subsets of N_y patients who either share common variables, represented by D_y, or display specific patterns related to the property or disease X being studied. Clinical subgroup discovery is crucial for optimizing patient care, enabling personalized medicine, improving diagnostic accuracy, advancing medical research, and efficiently allocating healthcare resources by uncovering meaningful relationships and patterns within clinical datasets.

OBJECTIVES.

Due to the challenge of limited supervised data and non-numeric variables in clinical data, the overall goal of this area is to develop robust machine learning algorithms capable of effectively leveraging diverse sources of information, including unstructured data and non-numeric variables, to enhance the credibility and acceptance in clinical subgroup discovery. To achieve this, I set the following two sub-objectives:

- Discover an effective method, differing from current research, for incorporating non-numeric variables (mainly textual) from clinical data into established algorithms.
- Develop an elegant and effective method to improve the accuracy, reliability, and generalizability of subgroup identification in smaller dataset by learning from larger dataset.

During my time pursuing a PhD at the CoAStaL NLP Group at the University of Copenhagen in Denmark and at Huazhong University of Science and Technology in China, fortunately, I had the opportunity to contribute to some relevant projects in this field, such as depression diagnosis and natural language understanding. Moving forward, I'll explore the cutting-edge advancements and provide a detailed overview of the research plans for this proposal.

STATE OF THE ART

Identifying specific patient subgroups that would derive significant benefits from particular treatments is increasingly emerging as a crucial objective in precision medicine. The concept of subgroup discovery was initially defined by Klösgen and Wrobel. Clinical subgroup discovery can be essential in many application scenarios, such as the detection of coronary heart disease, brain ischaemia, COVID-19 clustering, etc.

Traditional subgroup discovery (SGD) typically involves three main steps: candidate generation, pruning, and post-processing for ranking. Candidate generation methods predominantly include Exhaustive Search, Beam Search, and Genetic Algorithm Helal (2016). Moreover, recent research has shifted focus towards approaches based on clustering, statistical methods, latent profile analyses, and a combination of clustering and subgroup discovery Cooper et al. (2021). Except for that, more clinical rules based methods are also proposed, such as APRIORI-SD, CN2-SD, DSSD, PRIM, SSD++, and MCluster-VAEs Vagliano et al. (2023).

However, there are still two primary challenges encountered by existing research: (1) Current research fails to utilize non-numerical variables, such as medical record text, patient self-reports, prescriptions, and consultation records, for identifying patient subgroups. (2) The transfer of subgroup discovery paradigms from large-scale datasets to small-scale ones remains an unexplored area.

RESEARCH PLANS

In this section, I delve into the detailed plan for my research roadmap based on existing works and my research experience. Objective (1) will be essential to ensure the fusion of more effective variables, enabling the implementation and deployment of current powerful large language models in practice. Secondly, it will enable domain adaptation and transfer learning for objective (2), to support adapting subset-specific knowledge and patterns from existing models to novel domains or specialized tasks. Specifically, I will aim to achieve the objectives by following the steps.

To achieve (1) Data: I will focus on user datasets related to psychological disorders such as depression and anxiety. One available dataset is the Depression Diagnostic dataset, which includes interview transcripts, numerical features, and depression labels Mao et al. (2024). Methodology: As a case study, I will investigate the process of encoding textual data, extracting essential information, and employing semantic matching algorithms to delineate subgroups within patient populations. Additionally, I will explore the potential of multimodal fusion strategies to augment subgroup identification by integrating both numerical and non-numerical variables. Furthermore, the scalability of this approach to broader contexts will be examined, thus contributing to its applicability across diverse scenarios.

To achieve (2) Data: I will mainly use large datasets such as Census Income, Social Media Datasets, and Biometric Datasets, as learning sources. And I will also collect or construct a small set as an adapted target. Methodology: I will try to explore transfer learning methods by transferring the model parameters, representations, or both from a large dataset to a small dataset, distributions aligning methods of different datasets to improve generalization performance, semi-supervised learning techniques to label small dataset by learning from a large dataset, etc. Moreover, feature transformation proves to be effective in this task Zhuang et al. (2020), such as feature augmentation, feature reduction, and feature alignment techniques. Also, instance weighting such as estimation method, heuristic method, etc, could have the potential to alleviate this problem.

Expected Outcomes

This project will set three milestones: (1) Explore the effectiveness of integrating numerical and non-numerical variables using multimodal fusion techniques to augment subgroup identification within patient populations, with a focus on depression and anxiety disorders. (2) Delve into the efficacy of transfer learning methods in adapting models from large datasets to small ones, focusing on aligning distributions to enhance performance. (3) Explore semi-supervised learning techniques aiming for labeling small datasets by harnessing knowledge from larger datasets.

All datasets and models will be publicly released as open-source resources for the community. Each milestone will result in a research paper to be submitted to a top conference or journal.

Clinical Application

Clinical subgroup discovery is valuable in medical resource allocation, medical diagnosis etc. With the development of this project, it is beneficial to identify patients suffering from depression and anxiety. Clinicians could utilize the insights gained from the research to better understand the heterogeneous nature of these disorders and personalize interventions based on subgroup characteristics. Additionally, for objective 2, it will be crucial to derive more valuable guidance, clues, or diagnostic foundations from large scale datasets.

Related activities and timeline

My research timeline plan entails: (1) dedicating 2 months to collecting, organizing, and acquainting myself with existing datasets comprising numerical and non-numerical variables, (2) allocating 4 months to devising novel algorithms leveraging current large language models and numerical subgroup discovery algorithms, (3) aiming to publish the first paper within 1 month, (4) spending 3 months transitioning to new application scenarios or mental illness, such as autism, etc, (5) investing 4 months in further exploring transfer learning methods through distribution alignment algorithms, (6) devoting 2 months to structuring the second paper, (7) conducting 4 months of research on semi-supervised learning techniques, (8) arranging the third paper over a period of 2 months, and (9) setting aside 3 months for flexible tasks such as data collection or other adjustments.

References

Aidan Cooper, Orla Doyle, and Alison Bourke. 2021. Supervised clustering for subgroup discovery: An application to covid-19 symptomatology. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 408–422. Springer.

Sumyea Helal. 2016. Subgroup discovery algorithms: a survey and empirical evaluation. *Journal of computer science and technology*, 31:561–576.

Kaining Mao, Deborah Baofeng Wang, Tiansheng Zheng, Rongqi Jiao, Yanhui Zhu, Bin Wu, Lei Qian, Wei Lyu, Jie Chen, and Minjie Ye. 2024. Kangning dataset of clinical interview for depression.

I Vagliano, MY Kingma, DA Dongelmans, DW de Lange, NF de Keizer, MC Schut, MS Arbous, DP Verbiest, LF te Velde, EM van Driel, et al. 2023. Automated identification of patient subgroups: A case-study on mortality of covid-19 patients admitted to the icu. *Computers in Biology and Medicine*, 163:107146.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.