Informatics PhD projects at King's College London, AY 24-25 — Natural Language Processing

The PhD project proposals listed below will be considered for 2024/25 studentships available in the Department of Informatics to start 1 October 2024 or later during the 2024/25 academic year. Please note that this list is not inclusive and potential applicants can alternatively identify and contact appropriate supervisors outlining their background and research interests or proposing their own project ideas.

The PhD projects are listed in two groups. In the first group are the projects with allocated studentships: each project in this group has one allocated studentship. The remaining studentships will be considered for the projects listed in the second group. The number of those remaining studentships is smaller than the number of the projects in the second group. The allocation of studentships will be based on the merits of individual applications. Applications for PhD studies in the Department of Informatics, for all listed projects as well as for other projects agreed with supervisors, are also welcome from students applying for other funding (within other studentship schemes) and from self-funded students. See also this list of funding opportunities available at King's for post-graduate research in Computer Science.

- Scholarship Allocated
- Scholarship Not Allocated

Scholarship Allocated

(Back to Top)

- Causality Representation Learning based Language Models
- Towards Robust Reasoning in Large Language Models
- Computational argumentation
- Improving Understandability of Automatically Generated Test Cases using Text-to-Text Transformer Models
- Estimating the ground truth of LLMs in softare engineering Tasks

Causality Representation Learning based Language Models

Supervisor: Supervisor

Areas: Natural Language Processing (NLP)

(Back to Scholarship Allocated)

Project Description

The fields of machine learning and graphical causality have historically evolved independently, with limited interaction. However, recent developments have showcased the potential benefits of merging these two disciplines. This project aims to establish the groundwork for harnessing the power of causal representation learning within the realm of Natural Language Processing (NLP), particularly focusing on Large Language Models (LLMs). The primary objective is to investigate how causal representation learning can enhance NLP, utilising state-of-the-art LLMs, by addressing the following key aspects: 1) Transfer and Generalisation: We will explore the application of causal inference to overcome challenges in NLP, such as transfer learning and generalisation. By incorporating causal relationships into LLMs, we aim to improve model performance across a wide range of tasks and domains. 2) Causal Representation Learning: A fundamental challenge for both artificial intelligence and causality research is the discovery of high-level causal variables from low-level observations. We propose to develop techniques for automating the extraction of causal variables from textual data, which can subsequently be integrated into LLMs for enhanced understanding and decision-making.

References

1. Hanqi Yan*, Lingjing Kong*, Lin Gui, Yuejie Chi, Eric Xing, Yulan He, Kun Zhang. Counterfactual Generation with Identifiability Guarantees. The 37th Annual Conference on Neural Information Processing Systems (NeurIPS), New Orleans, US, 2023.
2. Hanqi Yan, Lin Gui, and Yulan He. Hierarchical Interpretation of Neural Text Classification. Computational Linguistics (CL), 2022.
3. Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, Yoshua Bengio: Toward Causal Representation Learning. Proc. IEEE 109(5): 612-634 (2021)
4. Toan Nguyen, Kien Do, Duc Thanh Nguyen, Bao Duong, Thin Nguyen: Causal Inference via Style Transfer for Out-of-distribution Generalisation. KDD 2023: 1746-1757

Towards Robust Reasoning in Large Language Models

Supervisor: Yulan He

Areas: Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP)

(Back to Scholarship Allocated)

Project Description

Context Reasoning is a core aspect of human intelligence, playing a crucial role in tasks such as critical thinking, evaluation and decision-making. With the development of large language models (LLMs), we have witnessed their impressive performance in various natural language processing tasks that involve reasoning processes. For an intelligent system to succeed, it must effectively analyse key information within a given context and provide accurate responses by drawing upon its internal knowledge and available resources. Achieving this is a complex process as LLMs must stay updated with the latest information, remain robust in noisy contexts, and be capable of utilising external tools for verification when necessary.

Project: Despite the advancements in reasoning capabilities of LLMs, there remains uncertainty regarding the extent to which LLMs can engage in reasoning beyond mere memorisation. Recent empirical studies have highlighted their susceptibility to challenges posed by noisy contexts, new information, and novel tasks. Consequently, our objective is to establish a robust reasoning framework that empowers LLMs to engage in reasoning effectively when presented with new and unfamiliar inputs. To accomplish this goal, example tasks include:

- Enhancing reasoning through tool augmentation based on a neuro-symbolic approach. LLMs could benefit from neuro-symbolic reasoning facilitated by external interpreters, particularly in complex tasks.
- Facilitating model adaptation to reason with the most recent knowledge. This involves model editing and fine-tuning the model with new information while retaining its capacity for reasoning in tasks that it has encountered before.
- Promoting collaboration among multiple agents to facilitate reasoning across diverse domains. When faced with an input from an unfamiliar domain, integrating knowledge from multiple trained agents based on its relevance to the specific input could be advantageous.

References

References:

- Jie H, Kevin Chen-Chuan C. 2023. Towards Reasoning in Large Language Models: A Survey. [pdf]
- Collin B, Haotian Y, Dan K, Jacob S. 2022. Discovering Latent Knowledge In Language Models Without Supervision. [pdf]
- Almog G, Elad V, Colin R, Noam S, Yoav K, Leshem C. 2023. Knowledge is a Region in Weight Space for Fine-tuned Language Models. [pdf]
- Luyu G, Aman M, Shuyan Z, Uri A, Pengfei L, Yiming Y, Jamie C, Graham Ng. 2023. PAL: Program-aided Language Models. [pdf]
- Marco F, Florian W, Luca Z, Alessandro A, Emanuele R, Stefano S, Bernhard S, Francesco L. 2023. Leveraging sparse and shared feature activations for disentangled representation learning. [pdf]
- Jonas P. Sebastian R. Ivan V.. Edoardo M. P∗.2023. Modular Deep Learning. [pdf]

Computational argumentation

Supervisor: Oana Cocarascu

Areas: Natural Language Processing (NLP)

(Back to Scholarship Allocated)

Project Description

Computational argumentation is a research area in natural language processing which encompasses several tasks such as argument mining (the automatic identification of natural language arguments and their relations from text) and argument quality assessment, amongst others (see [1] for a survey). Argument mining has been applied to several areas: persuasive essays, scientific articles, Wikipedia articles, news articles, online debates, product reviews, social media, legal documents, and political debates. The project will focus on one of the following topics:

- identifying and analysing arguments
- argument quality assessment
- argument generation
- multimodal argument mining
- applications of argument mining
- robustness of argument mining models

[1] https://aclanthology.org/2022.tacl-1.80.pdf

Improving Understandability of Automatically Generated Test Cases using Text-to-Text Transformer Models

Supervisor: Gunel Jahangirova

Areas: Systems (SE, programming, autonomous systems, robotics, ...), Artificial Intelligence (AI), Natural Language Processing (NLP), Machine Learning (ML)

(Back to Scholarship Allocated)

Project Description

The costs associated with software testing activities make their full automation an important research topic. The existing automated test case generation tools (ATGTs) have made significant progress in achieving high coverage, high fault detection rate and input diversity. However, the research in software testing is still far from fulfilling its dream of full automation because multiple studies demonstrate that developers find automatically generated test cases hard to read and understand. This project proposes three directions to tackle the problem of the understandability of automatically generated test cases. The first direction is based on the insight that developer-written test suites capture the information about what testing the given class looks like when performed by the developer and therefore contains features that make the test cases more understandable. We aim to extract the available understandability-related information from developer-written test suites and transfer it into the automatically generated test cases. Our second direction aims to make the understandability of the test case part of the test case generation process such that it favours the test cases with higher understandability. For this, we want to collect a large dataset with human-annotated understandability scores and train a learning model that can predict the understandability score for a candidate test case. The last direction aims to take advantage of the increasing success of text-to-text transformer models. We plan to collect a large dataset of pairs of automatically generated and developer-written test cases that test similar behaviour and train a transformer model that takes an automatically generated test case and transforms it into a version that looks like developer-written. The expected results from the project are in two directions. The first one is the deepened comprehension of the understandability problem. The second one is the set of automated software testing tools that will produce an output that is more understandable by the developers leading to wider adoption of such tools in industrial settings. Moreover, we plan to conduct large studies involving human participants to evaluate the understandability, which will hopefully provide the software engineering research community with examples of well-designed studies evaluating the qualitative properties of test cases.

References

Related Work:

1. E. Daka, J. M. Rojas, and G. Fraser, “Generating unit tests with descriptive names or: Would you name your children thing1 and thing2?” in Proceedings of the 26th ACM SIGSOFT International Symposium on Software Testing and Analysis, 2017, pp. 57–67.
2. G. Fraser, M. Staats, P. McMinn, A. Arcuri, and F. Padberg, “Does automated white-box test generation really help software testers?” in Proceedings of the 2013 International Symposium on Software Testing and Analysis, 2013, pp. 291–301.
3. J. M. Rojas, G. Fraser, and A. Arcuri, “Automated unit test generation during software development: A controlled experiment and think-aloud observations,” in Proceedings of the 2015 international symposium on software testing and analysis, 2015, pp. 338–349.
4. E. Daka, J. Campos, G. Fraser, J. Dorn, and W. Weimer, “Modeling readability to improve unit tests,” in Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, 2015, pp. 107–118.

Estimating the ground truth of LLMs in softare engineering Tasks

Supervisor: Jie M. Zhang

Areas: Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP), Systems (SE, programming, autonomous systems, robotics, ...)

(Back to Scholarship Allocated)

Project Description

When using LLMs for software engineering tasks such as code generation, it is important to understand how reliable the generated outputs are. Most of the time the ground truth is unknown. Thus, it is important to estimate the confidence and accuracy of LLMs so as to improve their usability and help users judge whether to adopt the provided solutions. This proposal aims to explore different methods to estimate the confidence of LLMs in generating solutions, in particular to software engineering-related tasks.

References

https://arxiv.org/pdf/2310.03533.pdf
https://openreview.net/forum?id=gjeQKFxFpZ

Scholarship Not Allocated

(Back to Top)

- Requirements formalisation using machine learning
- Understanding the Complexity of Negotiations
- Co-generation Enhancement for Knowledge Extraction and Language Modelling
- Indexing text data: practical and (near)-optimal schemes.

Requirements formalisation using machine learning

Supervisor: Kevin Lano

Areas: Systems (SE, programming, autonomous systems, robotics, ...), Machine Learning (ML), Natural Language Processing (NLP)

(Back to Scholarship Not Allocated)

Project Description

Formalisation of natural language requirements in software modelling languages such as UML is an essential activity in software development. Various heuristic and machine learning approaches have been applied to this problem over the last 10 years. This research proposal will investigate the application of deep learning approaches and in particular large language models (LLMs) to the formalisation of software requirements.

References

"On the assessment of ChatGPT for modeling tasks"
https://link.springer.com/article/10.1007/s10270-023-01105-5

Understanding the Complexity of Negotiations

Supervisor: Dr Alfie Abdul-Rahman & Dr Rita Borgo

Areas: Visualisation, Human-centred computing, Natural Language Processing (NLP), Social computing, Machine Learning (ML)

(Back to Scholarship Not Allocated)

Project Description

A negotiated text is the product of a formal decision-making process where a text has been negotiated and drafted over a period of time. Many of the foundational texts of the modern world have not been written by individuals, by negotiated by groups of people in formal settings. For example, treaties between states such as the Universal Declaration of Human Rights or the Treaty of Versailles; or constitutions, such as the one negotiated by the American states in the Constitutional Convention of 1787. During such negotiations, it is important for us to keep track of the delegations and their involvements to grasp their influence on the negotiation process either using techniques such as close reading, distance reading, or machine learning. Even relatively short historical documents written collectively in this way have been the product of thousands of specific proposals and decisions. This project will apply a visual analytics approach towards the understanding of the complexity of a negotiation and the influence of the delegations during a negotiation process. Possible research questions: a. Developing new static and interactive visualization to assist with data discovery and insight generation in large datasets of events within interacting timelines. b. Developing new approaches to show the evolution of complicated, technical documents over the period of months or years. c. Developing new approaches for indexing the datasets related to the negotiation of documents, and more intuitive displays of the results. d. Developing natural-language-based approaches to relate information captured in 'informal' archives (such as private diaries, letters, social media feeds etc.) to the formal records of a negotiation. This project will work closely with the Quill Project, based at Oxford University:

References

1. Exploring Interpersonal Relationships in Historical Voting Records. G. D. Cantareira, Y. Xing, N. Cole, R. Borgo, and A. Abdul-Rahman. In Computer Graphics Forum, 42(3):211-221, 2023.
2. Quill: A Framework for Constructing Negotiated Texts — with a Case Study on the US Constitutional Convention of 1787. N. Cole, A. Abdul-Rahman, and G. Mallon. In ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017.

Co-generation Enhancement for Knowledge Extraction and Language Modelling

Supervisor: Lin Gui

Areas: Natural Language Processing (NLP)

(Back to Scholarship Not Allocated)

Project Description

1. Introduction: In the rapidly evolving landscape of natural language processing (NLP) and artificial intelligence, the demand for advanced language models and knowledge extraction techniques is on the rise. This proposal outlines a research initiative focused on co-generation enhancement to elevate the capabilities of language modeling and knowledge extraction systems. 2. Objectives: The primary objectives of this research project are as follows: a. Enhance Co-generation Techniques: Investigate and develop advanced co-generation techniques that facilitate the simultaneous generation of diverse and contextually relevant outputs in language modeling tasks. b. Optimise Knowledge Extraction: Improve existing knowledge extraction methods by integrating co-generation capabilities, enabling more comprehensive and accurate extraction of information from diverse textual sources. c. Interactive Language Generation: Explore and advance interactive language generation methods that leverage human queries to enhance performance by summarising content and shifting focus. This approach will also facilitate the seamless injection and editing of external knowledge bases. d. Evaluate Model Performance: Rigorously evaluate the performance of the enhanced language models and knowledge extraction systems using benchmark datasets and real-world applications.

References

1. Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, Antoine Bosselut: PeaCoK: Persona Commonsense Knowledge for Consistent and Engaging Narratives. ACL (1) 2023: 6569-6591.
2. Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, Yulia Tsvetkov: Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. ACL (1) 2023: 13960-13980.
3. Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, Ziwei Liu: Octopus: Embodied Vision-Language Programmer from Environmental Feedback. CoRR abs/2310.08588 (2023).
4. Yujie Wang, Hu Zhang, Jiye Liang, Ru Li: Dynamic Heterogeneous-Graph Reasoning with Language Models and Knowledge Representation Learning for Commonsense Question Answering. ACL (1) 2023: 14048-14063.

5. Runcong Zhao, Wenjia Zhang, Jiazheng Li, Lixing Zhu, Yanran Li, Yulan He, Lin Gui: NarrativePlay: Interactive Narrative Understanding. CoRR abs/2310.01459 (2023).
6. Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, Gao Huang: Avalon's Game of Thoughts: Battle Against Deception through Recursive Contemplation. CoRR abs/2310.01320 (2023).
7. Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, Yang Liu: Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. CoRR abs/2309.04658 (2023).
8. Lixing Zhu, Runcong Zhao, Lin Gui, Yulan He: Are NLP Models Good at Tracing Thoughts: An Overview of Narrative Understanding. CoRR abs/2310.18783 (2023).

Indexing text data: practical and (near)-optimal schemes.

Supervisor: Grigorios Loukides

Areas: Foundations of computing, Data science, Computing Applications, Natural Language Processing (NLP), Machine Learning (ML)

(Back to Scholarship Not Allocated)

Project Description

In many real-world database systems, a large fraction of the data is represented by strings: sequences of letters over some alphabet. This is because strings can easily encode data arising from different sources. It is often crucial to represent such string datasets in a compact form but also to simultaneously enable fast pattern matching queries. This is the classic text indexing problem. Unfortunately, however, most (if not all) widely-used indexes (e.g., suffix tree, suffix array, or their compressed counterparts) are not optimized for all four measures (index space, construction space, query time, construction time) simultaneously, as it is difficult to have the best of all four worlds. The topic seeks to take an important step towards designing new indexes that offer good performance in all four measures. One promising direction to do this is to explore specific application-driven special cases of the problem, such as when we have at hand a lower bound ℓ on the length of the queried patterns or when we have extra knowledge about them or the text (e.g., given by a machine learning model or text properties such as the fact that it is repetitive). The candidates should have strong knowledge in algorithms and programming (C++).

References

Ayad et al. Text Indexing for Long Patterns: Anchors are All you Need. Proceedings of the VLDB Endowment (PVLDB) 2023. https://www.vldb.org/pvldb/vol16/p2117-loukides.pdf
Loukides et al. Bidirectional String Anchors for Improved Text Indexing and Top-K Similarity Search. IEEE TKDE 2023. https://ieeexplore.ieee.org/document/10018284