Current Research Projects

Below is a List of Research Projects:

Latest Announcements

We are thrilled to share our latest updates at The Lee Language Lab; remember with every success comes a lot of hard work, dedication, and learning from failures.

Winter 2025 [Updated March 18 2025]

- Our paper, "A Multilingual Dataset (MultiMWP) and Benchmark for Math Word Problem Generation", has been accepted for publication in IEEE/ACM Transactions on Audio, Speech, and Language Processing.
- Our journal paper has been accepted for publication “Benchmarking Interpretability in Healthcare Using Pattern Discovery and Disentanglement”
- Two poster presentations at SIGCSE this weekend, York University’s ML Certificate, University of Toronto’s AI Concentration for Professional Master’s
- We continue our momentum with four accepted papers for NAACL:
    - Research Papers: **WorldCuisine** [Arxiv] | [LinkedIn], **IrokoBench**[Arxiv] | [LinkedIn], **ProxyLM** [Arxiv] | [LinkedIn], and **AlignFreeze** [Arxiv];
    - Demos: **AiTaigi Hokkien Learning App** [LinkedIn] [Application] and **LangLearn FlipApp**
- Professor Lee received the **ARIA Spotlight Award** at the University of Toronto’s Master of Science in Applied Computing's ARIA Ceremony [LinkedIn]

Fall 2024:

- **AfriInstruct** presented at **EMNLP Findings 2024,** *Funded* by the Fields Institute. [LinkedIn | [EMNLP 2024 Findings] | [Video] | [Slides] [Paper][Field’sAnnouncement]
- **URIEL+ World Language Database** accepted at **COLING 2025,** *Funded* by the Fields Institute. [Paper | LinkedIn | PiPy | Video | Field’s Announcement]
- Undergrad Vincent Shuai won **the Student Engagement Award in the Department of Computer Science at the University of Toronto** [LinkedIn] [App]
- As well as engagement through **Canadian Tech at the Scale** Conference, Osler/RadicalVenture’s **Women in AI** night, and **Celebration of Women in Computing** Conference.
- We finished our Fall Invited Speaker Series, featuring TWO CIFAR AI Chairs from the Vector Institute.

Summer 2024

- **Best Audience Award** at Teaching NLP on **Empowering Multlinguality,** *Funded* by NSERC USRA and the Fields Institute LinkedIn | [Paper] | Video | [Slides] | [OTU Announcement] [Field’sAnnouncement]
- **STEM Gender Equity at Small Institution** presented at **SIGCSE Virtual 2024;** *Funded* by Women in Research Council and the Fields Institute [LinkedIn] | [SIGSCE Virtual 2024] | [Video] | [Slides] | [Teaching Symposium] [Field’sAnnouncement]

- **September 20th 2024:** Congratulations on the acceptance of EMNLP Findings short paper @Kosei Uemura, @Chika Maduabuchi, @Mahe Chen, @Alex Pejovic, @Yifei Sun. The conference will take place early November.
- **August 28, 2024:** Faculty of Science Assistant Professor’s paper wins Audience Award at TeachNLP workshop.
    - Assistant Professor Dr. Annie (En-Shiun) Lee, Faculty of Science, and her team from the Lee Language Lab, including Kosei Uemura, Mekael Wasti, and Mason

Shipton won the Audience Award from the TeachNLP workshop at an Association for Computational Linguistics event in Bangkok, Thailand.

- Their paper was titled ‘Empowering the Future with Multilinguality and Language Diversity’. Watch: Teach NLP Presentation - Empowering the Future with Multilinguality and Language Diversity

Projects that are currently recruiting students

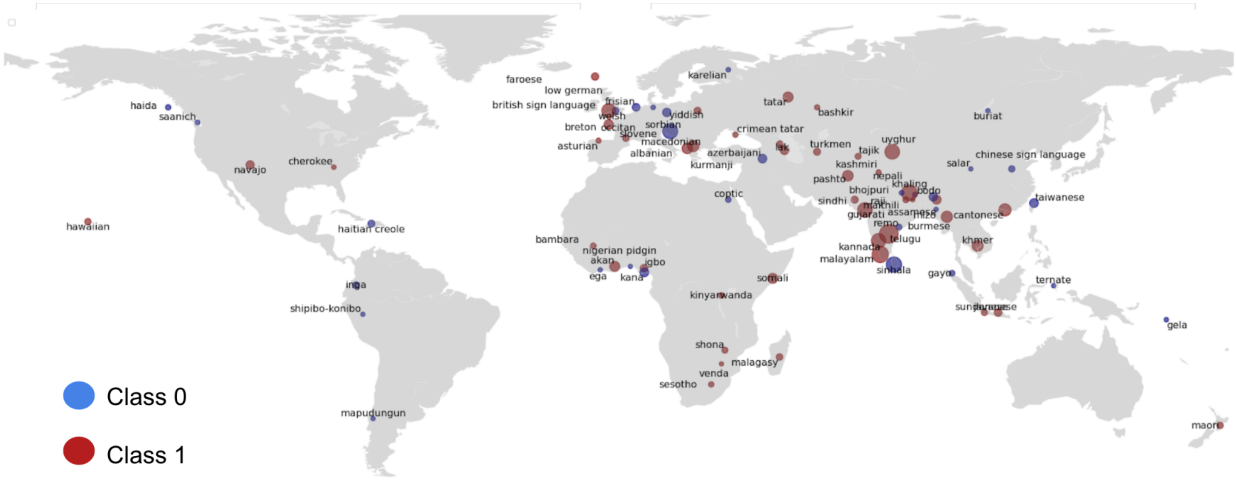
See Lee Language Lab Challenge Questions

Low-Resource Language

Translating Low-Resource Language

Paper, cite, slides, ACL video, video1, tutorial1, tutorial2

Team Leads: Sarubi Thillainathan, Shravan Nayak, Menan Velayuthan, Prof Surangika Ranathunga



Summary

Language is how we connect, share ideas, build communities. However, for the more than 7,000 languages in the world, only a handful have sufficient language technology supporting them. The remaining low-resource languages have limited resources in terms of data (unlabelled and labelled), computational resources, available language toolkits, and supporting community (speaking-wise as well as technology and research). There are billions of people speaking thousands of these low-resource languages that cannot access online content in their native language.

Neural Machine Translation (NMT) has seen a tremendous spurt of growth in the past decade and has entered maturity. While considered the most widely used solution for Machine Translation, its performance on low-resource language pairs still remains sub-optimal compared to the high-resource counterparts, due to the unavailability of large parallel corpora. Our survey presents details of research advancements, along with a quantitative analysis identifying the most popular solutions, guidelines to select the possible technique for a given data set, and a holistic view of the research landscape with a list of recommendations.

Next, we studied whether pre-trained multilingual sequence-to-sequence models like mBART can contribute to translating low-resource languages. We conduct a thorough empirical experiment in 10 languages to ascertain this, considering five factors: (1) the amount of fine-tuning data, (2) the noise in the fine-tuning data, (3) the amount of pre-training data in the model, (4) the impact of domain mismatch, and (5) language typology. In addition to yielding several heuristics, the experiments form a framework for evaluating the data sensitivities of machine translation systems. While mBART is robust to domain differences, its translations for unseen and typologically distant languages remain below 3.0 BLEU. In answer to our title's question, mBART is not a low-resource panacea; we therefore encourage shifting the emphasis from new models to new data.

References

- Lee, En-Shiun, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya D. McCarthy. "Pre-Trained Multilingual Sequence-to-Sequence Models: A Hope for Low-Resource Language Translation?." In Findings of the Association for Computational Linguistics: ACL 2022, pp. 58-67. 2022.
- Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, Rishemjit Kaur. "Neural Machine Translation for Low-Resource Languages: A Survey". ACM Survey [2023]
- Surangika Ranathunga, En-Shiun Annie Lee, Rishemjit Kaur, Marjana Prifti Skenduli, Sarubi Thillainathan University of Moratuwa, University of Toronto, "Neural Machine Translation for LowResource Languages" ICON Tutorial 2021
- Workshop at Toronto Machine Learning Summit - Microsummit in NLP Recording
- Talk at Toronto Machine Learning Summit - Microsummit in NLP

Data Collection for Low-Resource Languages

Educational Projects

Language Education

Computer Science Education

- Assistant Professor Dr. Annie (En-Shiun) Lee, Faculty of Science, and her team from the Lee Language Lab, including Kosei Uemura, Mekael Wasti, and Mason Shipton won the Audience Award from the TeachNLP workshop at an Association for Computational Linguistics event in Bangkok, Thailand.

- Their paper was titled ‘Empowering the Future with Multilinguality and Language Diversity’. Watch: Teach NLP Presentation - Empowering the Future with Multilinguality and Language Diversity

Previous Projects

Interpretability Project

Pattern Discovery and Disentanglement in Clinical Text

Nature Paper Video

Team Members: Malikeh Ehghaghi, Wendy Chen, Peiyuan Zhou, Sahar Rajabi, Sophy Sun, David Basil, Chih-hao Kuo, Selina Sha, April Liu
Team Lead: Malikeh Ehghaghi

Summary:
For artificial intelligence (AI) systems to be adopted in high stake human-oriented applications, they must be able to make complex decisions in an understandable and interpretable manner. While AI systems today have grown leaps and bounds in predictive power using larger datasets with more complex architectures, existing models remain ineffective at generating interpretable insights in the clinical setting. In this paper, we address the challenge of discovering interpretable insights from the clinical text for disease prediction. For this purpose, we apply the clinical notes from the electronic health records (EHR) available in the Medical Information Mart of Intensive Care III (MIMIC-III) dataset, which are labeled with the international classification of diseases (ICD9) codes. Our proposed algorithm combines interpretable text-based features with a novel pattern discovery and disentanglement algorithm. Specifically, our approach encompasses the following: (1) uncovering strong association patterns between clinical notes and diseases, (2) surpassing baseline clustering algorithms in effectively distinguishing between disease clusters, and (3) demonstrating comparable performance to baseline supervised methods in predicting diseases. Our results validate the model's capability to strike a balance between interpretability and outcome prediction accuracy. By unveiling insightful patterns between clinical notes and diseases, our approach upholds a reasonable level of diagnostic accuracy.

LRL Projects

Research Project 1. Creating a User Interface for Data Collection [02/2024]

- *Creating a web or mobile user interface using visualisation software toolkits (e.g., d3) and front-end development (e.g., React)*
- *Conduct user studies (surveys and interviews) and software usability testing, creating research plan, recruiting participants, conducting the study (aka collecting data), analyzing data, and reporting*
- *Oversee research ethics board approval and data management plan*

Research Project 2. Expanding Large Language Models for Additional Languages [02/2024]

- *Explore and benchmark latest techniques for adding languages into language models*
- *Require knowledge of python with deep learning (i.e. Pytorch, tensorflow, keras), huggingface*
- *Reproduce experiments and code from paper with code, github, and data repository*

Research Project 3. Text and Language Model interpretability for Domain-Specifc Data or Low-Resource Language [02/2024]

- *Work with industry and client on patented technology and proprietary data*
- *Foundational understanding of data mining and machine learning methods*
- *Strong in implementation and datasets*

Undergraduate Project 1: Agricultural Data Science (India) [02/2024]

***Slides**, video (last 10-15 minutes)*

Undergraduate Project 2. CS Equity Diversity Inclusion at OntarioTech [02/2024]

(volunteer: tasks in Feb/March, paper deadline August, CANCWIC)

Current Projects [March 15th 2024]

- ***Multilingual Modelling Projects***
  - *AlignFreeze with Dr Félix Gaschi from France and Prof Jeremy Bradbury*
  - *ProxyModel using Indonesian Languages with Dr Genta Winata from Bloomburg/HKUST*
  - *AfricanLlama2+ALMA with Prof David Adelani from CIFAR chair from MILA/McGill and Saarland*
- ***Mathematical Projects***
  - *URIEL/Lang2Vec Reproducibility with Prof A. Seza Doğruöz*
  - *Language Similarity at the Fields Institute*
- ***Applied HCI Projects***
  - *Language Education App with Professor Richard Tsai*
  - *HCI Translation Interface*

Old: User-Aware MultiLingual Text Simplification

Paper, cite, slides, video
Team Leads: Eric Haonan Gao, Prof Ravi Shekhar, Prof Surangika Ranathunga (Xiao Sun, Murphy Tian)

Summary

Currently, most of the text simplification datasets are available for English only, and also most of these datasets ignore the user's education (i.e. the user is University Graduate or Highschool student). Our aim is to create a dataset such that it covers multiple diverse languages and also

targets different education levels of the user. Therefore we create a multilingual dataset for text simplification by keeping the user's Education level in mind.

Collaborators
Ravi Shekhar (Hindi/Thai)- University of Essex
Surangika Ranathunga (Sinhala) - University of Moratuwa, Sri Lanka
En-Shiun Annie Lee (English, Chinese) - University of Toronto
Marjana Prifti Skunduli (Albanian) - University of New York Tirana, Albania
Mehreen Alam (Urdu) - NUCES, Islamabad
Vukosi Marivate (Setswana/Xitsonga) - University of Pretoria, South Africa
Sarveswaran K (Tamil) - University of Jaffna, Sri Lanka

Old: African Medical Intent Classification

Team Leads: Arjiana Jung, Jingwen Ji, Prof David Adelani

Summary

Masakhane Lucane Grant group is working on a project that aims to expand the African-language dataset for NLP projects. Aviro Health is an African company working on an HIV self-testing application that needs several HIV testing instructions translated into many different African languages. This project aims to support these two stakeholders by providing datasets and a model trained in African language(s), where both are based on the medical domain.

Human-Centric Active Interpretability - Survey and Study

Paper, cite, slides, video

***Team Leads**: Dr. Yilun Zhou (MIT, Amazon), Chih-Hao Kuo, and Haoyu Du*

Summary

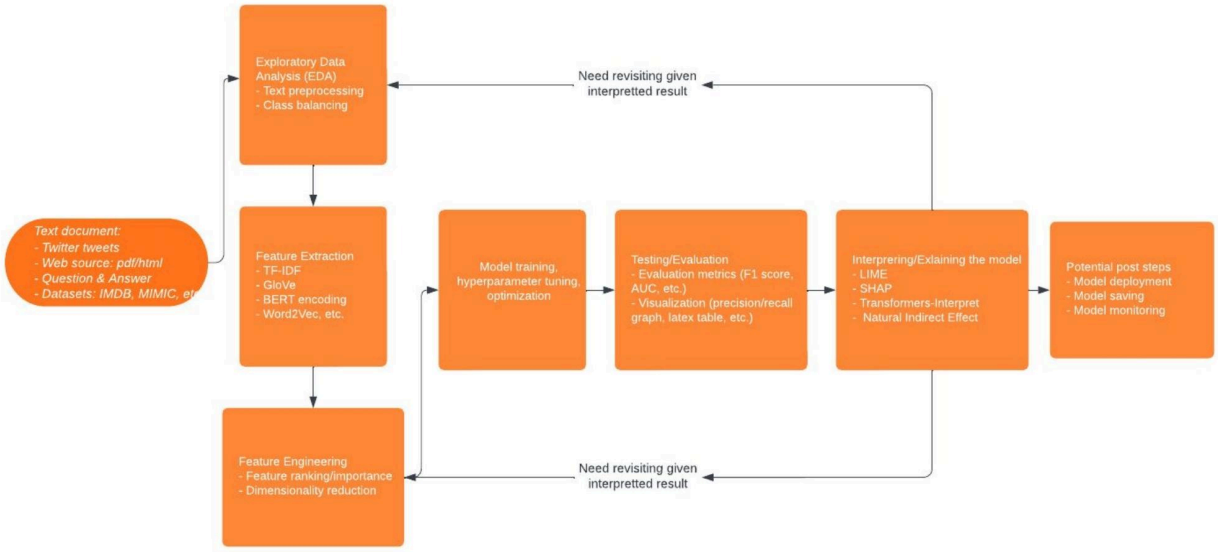
What good does improved interpretability do for us? It has been found that in many machine learning modelling contexts, even though sometimes interpretable methods come at an expense of accuracy, the penalty is limited, giving more interpretable models a leg up especially in scientific research, such as in medical contexts. When we think of ML-based NLP models, we often need to critically look at their applications in sensitive fields such as clinical decision-making. When human lives are directly involved, we want to be as transparent as possible whenever we give a diagnosis, which requires a more interpretable and accurate model. Ultimately, our goal is to help improve the quality of interpretability methods. We propose to do this by codifying the evaluation metric of interpretability. Through extensive literature

research, we hope to provide insights for researchers to holistically evaluate interpretability in both model-centric (faithfulness) and human-centric (understandability) ways.

Natural Language Processing Pipeline and Demo

Paper: Draft demo paper
Cite: -
slides: Orientation Presentation
Video: -

***Student Leads**: Mariia Ponomarenko, Arya Sighn*



Summary

Text classifiers are widely used to organize, structure, and categorize various types of text. Common approaches use supervised learning to classify texts. However, categorizing unlabeled data involves manual effort, which is time-consuming and expensive. In addition, many existing classification algorithms may have ineffective interpretability. Therefore, in this study, we compare different feature extraction techniques (TF-IDF, BERT, Word2Vec, Doc2Vec, LDA) for various text classification tasks and datasets and analyze how the number of selected features can influence the results of the models. Moreover we look at the problem of text interpretability for both supervised and unsupervised learning. As a solution, we propose a generic pipeline including 1) preprocessing; 2) Feature engineering (i.e., feature extraction, feature importance), 3) modelling, and 4) interpreting.

Aspect-Based Sentiment Analysis at the Fields Institute 2017

Other works include: Named Entity Recognition, Entity-Aspect Linking,