Peerat Limkonchotiwat

LinkedIn: https://www.linkedin.com/in/peerat-limkonchotiwat/ Github: https://github.com/mrpeerat
Email: peerat.l_s19@vistec.ac.th Website: https://mrpeerat.github.io/ Tel. +66 93 789 8114

I am a fifth-year Ph.D. student in information science and technology (IST) at VISTEC, Thailand. I have extensive experience with Natural Language Processing (NLP) and Information Retrieval (IR), for instance, large language models, dense retrievals, semantic understanding, multilingual learning, question answering, and entity linking. I have strong abilities to organize, supervise, and work in a team environment by publishing over 10 papers in top NLP conferences and being the team leader of the NLP group at my university. Moreover, I am interested in applying research in real-world scenarios such as information retrieval (search and ranking) and question-answering systems.

EDUCATION

Vidyasirimedhi Institute of Science and Technology (VISTEC) **Rayong, TH**
Ph.D. in Information Science and Technology (5 years program) *Aug 2019–present*
GPA: 3.81/4.00 *expected graduation 05/24*

- Work under Assoc. Prof. Dr. Sarana Nutanong and Dr. Ekapol Chuangsuwanich at the Natural Language and Representation Learning Lab (NRL).
- Selected courses: pattern recognition, natural language processing, data structure, and algorithms.
- Full scholarship: monthly stipend, annual research grant, and conference travel grant.

Rajamangala University of Technology Lanna Chiang Mai (RMUTL) **Chiang Mai, TH**
Bachelor of Science in Computer Engineering *Sept 2015–Mar 2019*
GPA: 3.25/4.0 (Class rank: 1st)

RESEARCH AND TECHNICAL EXPERIENCES

Student Researcher **Bangkok, TH**
Vidyasirimedhi Institute of Science and Technology (VISTEC) *Aug 2019–Present*

- Research topic: word segmentation, multilingual sentence representation, multilingual sentence and document retrievals, representation learning, and question answering.
- Developed domain adaptation and out-of-domain handling techniques based on a stacked ensemble (deep learning) for word segmentation in 4 languages: Thai, Chinese, Japanese, and Urdu. [EMNLP 2020, ACL 2021]
- Developed a monolingual and multilingual sentence embedding based on a pre-trained language model for semantic textual similarity and text mining tasks. [EMNLP 2021-2022, TACL 2023]
- Developed a novel monolingual and multilingual text embedding for a retrieval question-answering and web-search framework. [NAACL 2022, ACL2023]
- Help supervise and mentor 8 graduated students and 2 research assistants.
- Concentrating on the application of NLP to solve industry and real-world problems.

Applied Science Internship **Cambridge, GBR**
Alexa Knowledge, Amazon. *October 2022– April 2023*

- 6 months internship at Alexa Knowledge, Cambridge, GBR, working with Weiwei Cheng (mentor), Christos Christodoulopoulos, Amir Saffari, Jens Lehmann, and Daniel Masato (PM).
- Implemented a novel multilingual end-to-end entity linking system, mReFinED. The system achieved superior performance and runtime across various benchmark datasets [EMNLP 2023].
- During the internship, I learned about research and development (R&D), Amazon's business model, leadership principles, and software engineering.

Subject Matter Expert in NLP **Bangkok, TH**
Artificial Intelligence Research Institute (AIResearch, Thailand). *Aug 2019– Present*

- Launched two Thai large language models, such as WangchanGLM and WangchanLion. Both models outperform ChatGPT and existing Thai LLMs in Machine Reading Comprehension benchmarks.

- Launched a multi-domain and multi-task Thai instruction dataset, including legal, medical, finance, and retail domains. The dataset consists of 40,000 instructions with high-quality and diverse tasks, such as closeQA, openQA, summarization, creative writing, and brainstorming.
- Developed a direct preference optimization (DPO) for Thai LLMs without any supervised datasets created from human effort. The DPO technique improves Thai LLMs' responses from robot-like to human-like answers.
- Developed a cross-lingual and multilingual search system with LLM called WangchanRetriever to facilitate searching for answers in multiple languages. The system outperforms the state-of-the-art (mE5) on Thai retrieval QA datasets in all cases.

SKILLS

- **Programming languages**: Python (Proficient), C++ (Basic), SQL (Proficient)
- **Deep learning frameworks**: Keras, Tensorflow, PyTorch, and HuggingFace
- **Languages**: Thai (Native), English (Proficient)

SOCIAL OUTREACH

- **AI Builder 2021-2023 Program**. Mentored high-school students in creating AI projects such as Cross-Lingual Data Augmentation For Thai Question-Answering (GenBench@EMNLP'23), machine learning for Alzheimer's detection, and hairstyle recommendation (Website: https://vistec-ai.github.io/ai-builders/).
- **Thai-Sentence-Vector-Benchmark Project**. We formulate the first Thai text embedding benchmark, which consists of four tasks: semantic textual similarity (STS), text classification, text pair classification, and retrieval QA. We also experiment with various baseline and existing models on our benchmark. (Github: https://github.com/mrpeerat/Thai-Sentence-Vector-Benchmark)
- **Reviewers**. ACL (since 2023), EMNLP (since 2023), EACL (since 2023), and AACL (since 2023).
- **Program Committee**: SEALP 2023
- **Talks**. Southeast Asia LLMs: SEA-LION and Wangchan-LION at NLP-OSS@EMNLP'23

SELECTED PUBLICATIONS

- Peerat Limkonchotiwat, Wannaphong Phatthiyaphaibun, Raheem Sarwar, Ekapol Chuangsuwanich, Sarana Nutanong., "**Domain Adaptation of Thai Word Segmentation Models using Stacked Ensemble**", EMNLP 2020
- Peerat Limkonchotiwat, Wannaphong Phatthiyaphaibun, Raheem Sarwar, Ekapol Chuangsuwanich, Sarana Nutanong., "**Handling Cross- and Out-of-Domain Samples in Thai Word Segmentation**" ACL 2021
- Nattapol Trijakwanich, Peerat Limkonchotiwat, Wannaphong Phatthiyaphaibun, Raheem Sarwar, Ekapol Chuangsuwanich, Sarana Nutanong., "**Robust Fragment-Based Framework for Cross-lingual Sentence Retrieval**" EMNLP 2021
- Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, Sarana Nutanong., "**CL-ReLKT: Cross-lingual Language Knowledge Transfer for Multilingual Retrieval Question Answering**" NAACL 2022
- Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Lalita Lowphansirikul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, Sarana Nutanong., "**ConGen: Unsupervised Control and Generalization Distillation For Sentence Representation**" EMNLP 2022
- Panuthep Tasawong, Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, Sarana Nutanong., "**Typo-Robust Representation Learning for Dense Retrieval**" ACL 2023
- Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Lalita Lowphansirikul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, Sarana Nutanong., "**An Efficient Self-Supervised Cross-View Training For Sentence Embedding**" TACL 2023
- Peerat Limkonchotiwat, Weiwei Cheng, Christos Christodoulopoulos, Amir Saffari, Jens Lehmann., "**mReFinED: An Efficient End-to-End Multilingual Entity Linking System**" EMNLP 2023

ACHIEVEMENTS AND CERTIFICATES

- Participated in **"Southeast Asia Machine Learning School (SEA MLS) 2019"** (full scholar) Universitas Indonesia, Depok, Greater Jakarta, Indonesia.
- Completed over 17 certified MOOC courses on machine learning, deep learning, and NLP.