Proyag Pal

Edinburgh, UK
✉ *proyag.pal@gmail.com*
proyag.github.io
`www.linkedin.com/in/proyag-pal`
github.com/Proyag

Interests

Natural language processing (NLP), large language models, financial AI, large-scale and high-quality text datasets, neural machine translation

Experience

Professional Experience

Aug 2024 – Present
Edinburgh

Senior AI Engineer, *Aveni*
Building FinLLM – LLMs for NLP applications in the financial services industry.

- Working across the whole pipeline – data filtering, pre-training, fine-tuning, pruning and distillation, domain adaptation, evaluation, retrieval-augmented generation – to build the FinLLM suite of language models.
- Developing Agent Assure for automated safety and compliance assurance of AI agents in the financial services domain.

Dec 2023 – Apr 2024
Edinburgh

Deep Learning Engineer, *Efficient Translation Limited*, part-time
Corpus extraction and efficient low-resource machine translation.

- Trained efficient machine translation and corpus cleaning models for low-resource language pairs.
- Ran and optimised an efficient scalable parallel corpus extraction pipeline on web-scale data.
- Delivered datasets and models to customers on time and meeting requirements.

Nov 2022 – Feb 2023
Santa Clara

Applied Scientist Intern, *Amazon AWS AI*, internship
Four-month internship working on improving isochronous machine translation for automatic dubbing.

- Improved translation and timing accuracy of automatically dubbed videos. Published at InterSpeech 2023 as an oral presentation.
- Co-organised the automatic dubbing track at IWSLT 2023.

Jun 2020 – Oct 2020
Amsterdam

Data Engineer, *TAUS*
Worked on the EU-funded ParaCrawl project to collect parallel corpora from large-scale web crawls.

- Optimised, maintained, and ran a highly scalable processing pipeline to extract, translate, align, clean, and release parallel corpora from web crawling data.

Feb 2020 – Apr 2020
Lisbon

Junior AI Researcher, *Unbabel*
Machine translation and quality estimation for customer-facing products.

- Built domain-specific production machine translation models and quality estimation models.

Feb 2018 – Jan 2020
Geneva

Fellow in Neural Machine Translation, *World Intellectual Property Organization (WIPO)*, Advanced Technology Applications Center
Development and maintenance of WIPO Translate and related NLP tools and technologies.

- *WIPO Translate*: Built, improved, evaluated and deployed domain-specific neural and statistical machine translation models using the Marian and Moses toolkits.
- *IPCCAT*: Developed neural text classification systems for patent categorisation.
- Developed a system to efficiently retrieve semantically similar patents from large collections using sentence embeddings and Faiss indexes.
- Instrumental in the training and deployment of neural MT systems at several other international organisations and patent offices including IMF, OECD, WTO, IAEA, and KIPO.

Academic Research Experience

Nov 2020 – Dec 2024

Ph.D. Student, *University of Edinburgh (ILCC)*, School of Informatics
Doctoral research in machine translation. Supervised by Kenneth Heafield and Alexandra Birch.

- Research on analysing and incorporating extra information required by neural machine translation models in addition to source text to produce accurate translations.
- Introduced "cheat codes" – providing compressed target-side information to models – as a method to analyse additional information required by the models.
- Created large-scale document-level translation corpora in several language pairs based on ParaCrawl and built and analysed context-aware translation models.
- General research interests mainly in analysis of machine translation models, multilingual and document-level machine translation.

Mar 2023 – May 2023
Zurich

Visiting Researcher, *University of Zurich*, Department of Computational Linguistics
Three-month visit, conducting research on detection and analysis of underspecification of the source sentence in machine translation. Supervised by Rico Sennrich.

Sep 2017 – Dec 2017
Edinburgh

Research Assistant, *University of Edinburgh (ILCC)*, School of Informatics
Low-resource domain-specific machine translation research on the MeMaT project. Supervised by Kenneth Heafield and Alexandra Birch.

- Worked on developing isiXhosa-English medical-domain machine translation to facilitate doctor-patient communication in health centres in South Africa.
- Collected corpora released as a public resource.

Education

2020 – 2024
Edinburgh

Ph.D. in Informatics, *University of Edinburgh (ILCC)*
Ph.D. research in machine translation. Supervised by Kenneth Heafield and Alexandra Birch.

2016 – 2017
Edinburgh

M.Sc. in Informatics, *University of Edinburgh*, with Distinction
Selected Courses: Machine Translation, Accelerated Natural Language Processing

2011 – 2016
Kolkata

B.Sc. & M.Sc. in Computer Science, *St. Xavier's College*
Selected Courses: Artificial Intelligence, Data Mining & Warehousing, Computer Architecture

Selected Publications

Full list of publications at `https://proyag.github.io/publications`

ACL 2024

Document-Level Machine Translation with Large-Scale Public Parallel Corpora, ***Proyag Pal****, Alexandra Birch, and Kenneth Heafield*

Interspeech 2023

Improving Isochronous Machine Translation with Target Factors and Auxiliary Counters, ***Proyag Pal****, Brian Thompson, Yogesh Virkar, Prashant Mathur, Alexandra Chronopoulou, and Marcello Federico*

EACL 2023 (Findings)

Cheating to Identify Hard Problems for Neural Machine Translation, ***Proyag Pal*** *and Kenneth Heafield*

NAACL 2022

Cheat Codes to Quantify Missing Source Information in Neural Machine Translation, ***Proyag Pal*** *and Kenneth Heafield*

Programming

Python, with PyTorch, NumPy, sklearn, etc.

C++, Marian toolkit for MT

Bash, Docker, LaTeX