Proyag Pal

Edinburgh, UK
✉ *proyag.pal@ed.ac.uk*
proyag.github.io
www.linkedin.com/in/proyag-pal
github.com/Proyag

Interests

Natural language processing (NLP), multilingual and document-level machine translation, analysis of neural machine translation models, multi-encoder neural architectures, large-scale high-quality text datasets

Education

2020 – 2024 Edinburgh
Ph.D. in Informatics, *University of Edinburgh (ILCC)*, in progress (expected 2024)
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

Experience

Professional Experience

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. 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, and clean parallel corpora from web crawling data.
- Consolidated and released the ParaCrawl corpus v7.0 and v7.1, comprising hundreds of millions of sentence pairs in many languages.

Feb 2020 – Apr 2020 Lisbon
Junior AI Researcher, *Unbabel*
Machine translation and quality estimation for customer-facing products.
- Built domain-specific machine translation models.
- Built quality estimation models to skip human post-editing for high-quality MT output.

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 retrieve semantically similar content from large collections of text 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 – Present, Edinburgh

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.

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*

Master's Projects

Jun 2017 – Aug 2017

Reward Augmented Maximum Likelihood to Improve Neural Machine Translation Training, *University of Edinburgh*, supervised by Kenneth Heafield

- Used reinforcement learning-inspired task rewards to augment the training objective.
- Improved upon a strong baseline by 1.07 BLEU.
- Re-implemented and integrated into the then Theano-based Nematus framework.

Aug 2015 – May 2016

Permutation Flow Shop Scheduling using Natural Algorithms, *St. Xavier's College, Kolkata*, supervised by Siladitya Mukherjee

- Optimization of makespan in permutation flow shop scheduling, using genetic algorithms.

Programming

Python, with PyTorch, NumPy, sklearn, etc.

C++, Marian toolkit for MT

Bash, Docker, LaTeX