

Proyag Pal

Edinburgh, UK

✉ proyag.pal@ed.ac.uk

📁 proyag.github.io

in www.linkedin.com/in/proyag-pal

🐙 github.com/Proyag

Interests

Analysis of neural machine translation models, low-resource and multilingual machine translation, multi-encoder neural architectures, natural language processing

Education

2020 – 2023 Edinburgh	Ph.D. in Informatics , <i>University of Edinburgh (ILCC)</i> , in progress (estimated 2023) Ph.D. research in machine translation. Supervised by Dr. Kenneth Heafield.
2016 – 2017 Edinburgh	M.Sc. in Informatics , <i>University of Edinburgh</i> , with Distinction <i>Selected Courses</i> : Machine Translation, Accelerated Natural Language Processing
2014 – 2016 Kolkata	M.Sc. in Computer Science , <i>St. Xavier's College</i> , GPA: 8.7/10 <i>Selected Courses</i> : Artificial Intelligence, Data Mining & Warehousing, Computer Architecture
2011 – 2014 Kolkata	B.Sc. in Computer Science , <i>St. Xavier's College</i> , GPA: 8.26/10

Experience

Academic Research Experience

Nov 2020 – Present Edinburgh	Ph.D. Student , <i>University of Edinburgh (ILCC)</i> , School of Informatics Research in machine translation, focusing on analysis. Supervised by Dr. Kenneth Heafield. <ul style="list-style-type: none">Working on using multi-encoder models to provide additional context to neural machine translation models to analyse and improve them.Research interests mainly in analysis of machine translation models, low-resource and multilingual machine translation.
Sep 2017 – Dec 2017 Edinburgh	Research Assistant , <i>University of Edinburgh (ILCC)</i> , School of Informatics Low-resource domain-specific machine translation research on the MeMaT project. Supervised by Dr. Kenneth Heafield and Dr. Alexandra Birch. <ul style="list-style-type: none">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.

Professional Experience

Nov 2022 – Feb 2023 Santa Clara	Applied Scientist Intern , <i>Amazon Web Services (AWS)</i> Four-month internship working on isochronic machine translation for automatic dubbing. Co-organised the dubbing track at IWSLT 2023.
Jun 2020 – Oct 2020 Amsterdam	Data Engineer , <i>TAUS</i> Worked on the EU-funded ParaCrawl project to collect parallel corpora from large-scale web crawls. <ul style="list-style-type: none">Optimised, maintained, and ran a highly scalable processing pipeline to extract, translate, align, and clean parallel corpora obtained through web crawling.Consolidated and released the ParaCrawl corpus v7.0 and v7.1, comprising hundreds of millions of sentence pairs in many languages.

- Feb 2020 – **Junior AI Researcher**, *Unbabel*, Applied AI
- Apr 2020
Lisbon
- 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 – **Fellow in Neural Machine Translation**, *World Intellectual Property Organization (WIPO)*,
Jan 2020
Geneva
- 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 similar content from large collections of text using sentence embeddings and Faiss indexes.
 - Assisted in the adoption of neural MT at IMF, OECD, WTO, IAEA, and KIPO.

Publications

- NAACL 2022 **Cheat Codes to Quantify Missing Source Information in Neural Machine Translation**, *Proyag Pal and Kenneth Heafield* [Link]
- WMT21 at
EMNLP 2021 **The University of Edinburgh's Bengali-Hindi Submissions to the WMT21 News Translation Task**, *Proyag Pal, Alham Fikri Aji, Pinzhen Chen, and Sukanta Sen* [Link]

Master's Projects

- Jun 2017 – **Reward Augmented Maximum Likelihood to Improve Neural Machine Translation Training**, *University of Edinburgh*, supervised by Dr. Kenneth Heafield
Aug 2017
- 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 legacy Theano-based Nematus framework.
- Aug 2015 – **Permutation Flow Shop Scheduling using Natural Algorithms**, *St. Xavier's College, Kolkata*, supervised by Prof. Siladitya Mukherjee
May 2016
- Optimization of makespan in permutation flow shop scheduling, using genetic algorithms.

Programming

Python, *advanced*, with PyTorch, NumPy, sklearn, etc.
C++, *intermediate*, Marian toolkit for MT
Julia, Perl, Bash, Docker, L^AT_EX

Languages

English, Bengali, *Native/Bilingual*
French, *Conversational*

Chinese (Mandarin), *Basic*
Hindi, *Fluent*