

Proyag Pal

Edinburgh, UK

☎ (+44)07513660849

✉ proyag.pal@ed.ac.uk

📁 proyag.github.io

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

Experience

Academic Research Experience

- Nov 2020 – **Ph.D. Student**, *University of Edinburgh (ILCC)*, School of Informatics
Present Edinburgh Research in machine translation, focusing on analysis. Supervised by Dr. Kenneth Heafield.
- 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 – **Research Assistant**, *University of Edinburgh (ILCC)*, School of Informatics
Dec 2017 Edinburgh Low-resource domain-specific machine translation research on the MeMaT project. Supervised by Dr. Kenneth Heafield and Dr. 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.

Professional Experience

- Jun 2020 – **Data Engineer**, *TAUS*
Oct 2020 Amsterdam 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 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 **Advanced Technology Applications Center**
 Geneva 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**, **LaTeX**

Languages

English, Bengali , <i>Native/Bilingual</i>	Chinese (Mandarin) , <i>Basic</i>
French , <i>Conversational</i>	Hindi , <i>Fluent</i>