# Rishu Kumar

✉ rishu_kumar@live.com   |   ⌂ rish.uk   |   ⌂ pyRis   |   in pyRis   |   📊 Semantic Scholar

## Education ─────────────────────────────────

**Charles University & Saarland University**                    *Prague, CR | Saarbrücken, DE*

EUROPEAN MASTER'S PROGRAMME IN LANGUAGE & COMMUNICATION TECHNOLOGIES                    *2021 - 2025*

- **Thesis:** Does multilingual training improve the performance for low-resource language summarization?
- **Projects:**
    - Summarization System for context retention in generating theatre script with customized GPT-2 beyond 1024 tokens for THEaiTRE project
    - BART-based summarization system for processing Shakespeare's plays in English and German

**Bhagalpur College of Engineering**                    *Bhagalpur, IN*

B.TECH IN COMPUTER SCIENCE ENGINEERING                    *2014 - 2018*

- **Major Project:** *e-Dustbin: A fully automated dustbin using Arduino to optimize route for waste pickup vehicles*
- **Minor Project:** *Finding weakest node in a network for attack and maximum damage using graph theory*

## Publications ─────────────────────────────────

**GrEmLIn: A Repository of Green Baseline Embeddings for 87 Low-Resource Languages Injected with Multilingual Graph Knowledge**                    *NAACL'25 Findings*

*Daniil Gurgurov, **Rishu Kumar**, Simon Ostermann*

**UM-DFKI Maltese Speech Translation systems**                    *IWSLT'24*

***Kumar Rishu**, Aiden Williams, Claudia Borg, Simon Ostermann*

**FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN**                    *IWSLT'23*

*Multiple authors*

**UM-DFKI Maltese Speech Translation**                    *IWSLT'23*

*Aiden Williams, Kurt Abela, **Rishu Kumar**, Martin Bär, Hannah Billinghurst, Kurt Micallef, Ahnaf Mozib Samin, Andrea DeMarco, Lonneke van der Plas, Claudia Borg*

**Team ÚFAL @ CreativeSumm: BART and SamSum based few-shot approach for creative Summarization**                    *CreativeSumm'22 (COLING Workshop)*

***Rishu Kumar**, Rudolf Rosa*

**Short-Term Word-Learning in a Dynamically Changing Environment**                    *arxiv*

*Christian Huber, **Rishu Kumar**, Ondřej Bojar, Alexander Waibel*

**Team ÚFAL at CMCL 2022 Shared Task: Figuring out the correct recipe for predicting Eye-Tracking features using Pretrained Language Models**                    *CMCL (ACL 2022)*

*Sunit Bhattacharya, **Rishu Kumar**, Ondřej Bojar*

**ELITR multilingual live subtitling: Demo and strategy**                    *EACL 2021*

*Multiple Authors*

**Operating a Complex SLT System with Speakers and Human Interpreters**                    *MTSummit 2021*

*Ondřej Bojar, Vojtěch Srdečný, **Rishu Kumar**, Otakar Smrž, Felix Schneider, Barry Haddow, Phil Williams, Chiara Canton*

**SEntFiN 1.0: Entity-Aware Sentiment Analysis for Financial News**                    *JASIST*

*Ankur Sinha, Satishwar Kedas, **Rishu Kumar**, Pekka Malo*

## Work Experience ─────────────────────────────────

**IRIT**                    *Information retrieval and Information synthesis*

INTERN                    *Apr 2025 - Now*

- Exploring the performance difference of LLMs with a knowledge base in comparison to LoRA fine-tuning
- Evaluating the performance between random context for inference time learning w.r.t topic-augmented context
- Working on model merging library to select the best merging strategy and parameters
- Participation in TREC-iKAT for intent classification and response generation

**Legaleey Inc.**

Research Consultant                                                                                    *July 2024*

- Researched LLM-based parsing of PDFs and pdftotext + heuristics for knowledge base creation
- Incorporated parsing pipelines for clean and semantically relevant output in Elastic Search
- Expanded legal data extraction points to create a minimal viable product based on CUAD
- Moved the tech stack from Generative models to extractive models for efficient and trustworthy results

**DFKI**                                                                    *Efficient and Explainable NLP Group*
                                                                                                    *(MLT)*

Research Assistant / Jr. Researcher                                                    *Nov 2022 - June 2024*

- Worked on creating intentionally erroneous data to check automated evaluation for summarization
- Co-organized Dialectical and Low-Resource SLT shared task in IWSLT ('23, '24)
- Created baseline system for evaluation and ran evaluation campaign
- Participation in IWSLT shared tasks for low-resource languages ('23, '24)

**Saarland University**                                                        *Language Science & Technology*

Tutor                                                                                          *Oct 2022 - Feb 2024*

- Tutor for Neural Networks Theory and Implementations in WS'22 & Statistical NLP in SS'23
- Proposed and trialled a course, "Basic tools for NLP," which covers bash, git, latex, etc., in SS'23; Confirmed as a regular course in WS'23

**European Live Translator (Charles University)**                            *Institute of Formal and Applied*
                                                                                                *Linguistics*

Research Intern                                                                              *Aug 2020 - Mar 2022*

- Prepared pipeline for continuous evaluation of ELITR-testset
- Improved SLTev a tool to evaluate ASR, MT and SLT performance
- Domain adaptation/retraining of ASR models (English and Czech) to improve quality and subsequently produce a better translation
- Maintained and improved bash pipelines to connect ASR, MT, and publishing workers
- Stress testing of systems to ensure reliability before using in research demo at multiple venues
- Created pipelines for automation of real-time meeting summarization of meeting in collaboration with alfaview and PerVoice

**Indian Institute of Management Ahmedabad**                        *Production & Quantitative Methods*

Research Assistant                                                                          *Mar. 2018 - Jul. 2020*

- Developed lexicon-based language models capable of extracting entity-based sentiments; Average news headline length $\sim$ 8 words
- Released an annotated dataset of 12,000+ financial news headlines
- Compared GBMs, SVMs, and MLPs for classification; Contrasted feature-engineered input vectors vs GloVe and Word2Vec; Achieved $\sim$ 91% accuracy and $\sim$ 93% F1-score using Roberta
- Validated the effect of sentiment information contained in news flow on the over-night aggregate market (NSE 500 index) returns
- Audited a PhD level course on Non-Linear Optimization

# Academic Misc

- Program Committee member for **IWSLT**'24, '25
- Student Council Member for Language Science & Technology programme for 2022-23
- Emergency reviewer for Workshop on Information Extraction from Scientific Publications @ AACL-IJCNLP 2022
- Student volunteer at COLING'22
- Student Representative for European Master's Programme in Language Science and Technologies for 2021-22
- Secured a rank of 3,978 in Graduate Aptitude Test in Engineering'19 among 100,000+ candidates
- Started a coding club and organized events to promote pair-programming
- Attended an Undergraduate Summer School at Computer Science and Automation Dept., Indian Institute of Science, Bangalore
- All India Rank 419 in IndiaHacks'16, an algorithm challenge on HackerEarth; College Rank 1

# Technical Knowledge

| | |
|---|---|
| **Programming** | Python, C, C++, bash, Matlab, HTML, TS, Rust |
| **Frameworks** | PyTorch, TensorFlow, Keras, Scikit-learn, MySQL, Marian |
| **Miscellaneous** | LaTeX, SLURM, SGE, *nix |

# Languages

| Fluency | Native | C2 | B1 | A2 | A1 |
|---|---|---|---|---|---|
| **Languages** | Hindi, Angika | English | Gujarati, Bengali, Bhojpuri, Maithili | German | Czech, French |