

Klim Zaporojets

Machine learning researcher — natural language processing, information extraction and knowledge graphs.

Open to research and applied-ML roles in industry and academia. Based in Aarhus, Denmark; EU work authorisation.

klimzaporojets@gmail.com · +45 22 32 84 63 · klimzaporojets.github.io · github.com/klimzaporojets · linkedin.com/in/klim-zaporojets-9102b0a · Google Scholar

SUMMARY

I build machine learning systems that turn text into structured knowledge: information extraction, entity linking and knowledge graphs, in general and biomedical domains. My work has produced widely used open datasets and benchmarks — DWIE, TempEL, BioDEX — alongside 24 peer-reviewed publications at ACL, NeurIPS, VLDB and CIKM. Before returning to research I spent seven years building software in industry, including production machine learning for fraud prevention at Latin America's largest e-commerce platform.

TECHNICAL SKILLS

Programming	Python (10+ yrs) · Java (7+ yrs, Sun Certified Programmer) · SQL (4+) · Scala (2+) · Groovy (2+) · JavaScript
Machine learning	PyTorch (6+ yrs) · Pandas (6+) · scikit-learn (5+) · NLTK (5+) · spaCy (4+) · Matplotlib (5+) · PyTorch Geometric · NetworkX · TensorFlow
Data and retrieval	Spark · Lucene · UIMA · Tableau · R · Matlab/Octave
Databases	Oracle (4+ yrs) · MySQL · Neo4j
Infrastructure	Linux (10+ yrs) · Git (10+) · Jenkins (5+) · Docker · Nginx · Flask · JUnit

EXPERIENCE

Research Scientist (Marie Skłodowska-Curie Fellow) — Aarhus University, Denmark

2024 – 2026

- Won and ran an independently funded **€230,000** research programme on keeping knowledge graphs correct as their source text changes.
- Designed and released **EMERGE**, a benchmark for updating knowledge graphs from emerging text, with public code and data.
- Co-supervised PhD and MSc researchers across six institutions, producing publications at ACL (oral, **SAC Highlights Award**), CIKM and ESWC.

Research Scientist — Aarhus University, Denmark

2023

- Built entity-summarization and knowledge-graph/text integration components for a **clinical decision-support system**, matching patient information to procedural guidelines.
- Delivered inside an industry partnership with **Systematic** (health IT) and clinicians at Aalborg University Hospital.

Doctoral Researcher — Ghent University, Belgium

2018 – 2022

- Led development of **DWIE** and **TempEL**, document-level information extraction and temporal entity-linking benchmarks now adopted as standards (118 and 28 citations; 51 and 10 GitHub stars).
- Built entity linking and information extraction over news for the EU Horizon 2020 **CPN** project, fusing Wikipedia text with Wikidata.
- Python, PyTorch, spaCy, NLTK, Flask.

Computational Linguist — Actonomy, Belgium

2015 – 2018

- Improved semantic matching between **CVs and job vacancies** for a commercial HR product, using pointwise mutual information and neural models for HR knowledge-graph completion.
- Shipped Tableau and Matplotlib dashboards surfacing labour-market skill-demand trends to end users.
- Python, scikit-learn, TensorFlow, Java, Scala, UIMA, Jenkins.

Research Assistant — University of Massachusetts Amherst, USA

2014 – 2015

- Conditional random fields for section and reference extraction from scientific articles, with **Prof. Andrew McCallum**.
- Knowledge-graph construction over materials-science literature with **AI2**, the **Materials Project** and **MIT**.
- Python, Scala, Spark.

Software Engineer — MercadoLibre, Inc., Argentina

2009 – 2013

- Shipped **production machine learning for fraud prevention** on the checkout path of Latin America's largest e-commerce platform (NASDAQ: MELI).
- Trained, tested and deployed highly scalable models — decision trees, logistic regression, neural networks — reaching production throughput via feature selection, model pruning, tuned SQL/PL-SQL, and Nginx load balancing and caching.
- Java, Python, R, Oracle 11g, Lucene, Jenkins, Nginx; Scrum.

SELECTED PUBLICATIONS

401 citations, h-index 11 (Google Scholar). Full list: klimzaporojets.github.io

- **EMERGE: A Benchmark for Updating Knowledge Graphs with Emerging Textual Knowledge.** *In submission, NeurIPS 2026.* First author.
- **Influences on LLM Calibration.** *ACL 2025* — oral presentation, **SAC Highlights Award.**
- **BioDEX: Large-Scale Biomedical Adverse Drug Event Extraction for Real-World Pharmacovigilance.** *Findings of EMNLP 2023.* 65 GitHub stars.
- **TempEL: Linking Dynamically Evolving and Newly Emerging Entities.** *NeurIPS 2022, Datasets and Benchmarks Track.* First author; 28 citations.
- **DWIE: An entity-centric dataset for multi-task document-level information extraction.** *Information Processing & Management*, 2021. First author; **118 citations.**

EDUCATION

PhD in Computer Science <i>Ghent University, Belgium</i>	02/2018 – 12/2022
MS in Computer Science <i>University of Massachusetts Amherst, USA</i>	09/2013 – 05/2015
Systems Engineer <i>Instituto Universitario Aeronáutico, Córdoba, Argentina</i>	05/2002 – 12/2008

LANGUAGES

Spanish (native) · Russian (native) · English (proficient) · Ukrainian (intermediate) · Dutch (intermediate) · Danish (beginner)