Lester James V. Miranda | CV

ljvmiranda@gmail.com • ljvmiranda921.github.io
ljvmiranda921 • Last updated: March 21, 2024

SUMMARY: Lj Miranda specializes in natural language processing with over five years of experience in consulting, open-source software development, and research.

Experience

Allen Institute for AI **Seattle, US**
Predoctoral Young Investigator, AllenNLP Team *Oct 2023 – present*

- As a predoctoral researcher, conducts research broadly in large language model adaptation (preference data collection, reward model evaluation, etc.).

ExplosionAI GmbH **Berlin, DE**
Machine Learning Engineer, spaCy Team *Oct 2021 – July 2023*

- Authored spaCy's first technical paper, *Multi-hash embeddings in spaCy*, that benchmarks the library's hash-embedding trick (first-author).
- Developed annotation workflows for a data annotation product, Prodigy, that integrates large language models (LLM) like GPT-3.5/4 to common natural language processing tasks.
- Improved spaCy's sequence labeling component, Span Categorizer, by adding new features, writing documentation, performing benchmark experiments, and optimizing performance.
- Co-developed several open-source software NLP libraries and developer tools including spacy-llm (production LLM pipelines), vscode-prodigy (Visual Studio Code extension for data annotation), and spaCy projects (end-to-end NLP workflows for production).

Thinking Machines Data Science, Inc. **Metro Manila, PH**
Machine Learning Researcher, Machine Learning Team *Oct 2018 – Jul 2021*

- Developed several production-grade natural language processing applications for a major investment firm in Singapore, ranging from in-house search engines to document processing tools.
- As Tech Lead, led a project team to deliver a large-scale digitization project of all the local governments' financial statements across the country for The World Bank.
- Automated building detection from aerial images for one of the largest telecommunications companies in the Philippines using computer vision techniques.

Internships

Preferred Networks, Inc. **Tokyo, JP**
Research Intern, ChainerRL Team *Aug 2018 – Sep 2018*

- Developed a reinforcement learning parallelization framework based on batch Proximal Policy Optimization (PPO) for the open-source ChainerRL library.

Education

Waseda University **Fukuoka, JP**
M.Eng., Major in Information Architecture *Sep 2016 – Sep 2018*
Thesis: Autoencoder-based Feature Extraction Techniques for Protein Function Prediction

Ateneo de Manila University **Metro Manila, PH**
B.S., Electronics & Communications Engineering, Cum Laude *Jun 2011 – Jun 2016*
Thesis: Appliance Recognition using Hall-Effect Current Sensors for Power Management Systems
Minor in Philosophy

Fellowships

RIKEN-Advanced Institute for Computational Sciences **Kobe, JP**
Fellow, RIKEN International School for Data Assimilation *Jan 2018*
Studied data assimilation techniques (3DVar, Kalman Filters, etc.) for real-time numerical simulations.

Institut Catholique d'Arts et Métiers **Lille, FR**
Exchange Student, Fall Semester *Sep 2015 – Jan 2016*
Took courses in control systems and software development

Open-source Software

I've maintained several open-source projects in the scientific tooling space. You can also visit my Github profile for more information.

calamanCy
ljvmiranda921/calamanCy *2023*
A natural language processing toolkit for building Tagalog pipelines based on spaCy and written on Python.

spaCy
explosion/spaCy *2021*
An industrial-strength natural language processing (NLP) software. I'm one of the core contributors as part of the spaCy team. I also contributed to related software such as spacy-llm and spaCy projects.

PySwarms
ljvmiranda921/pyswarms *2017*
A Python-based framework for implementing swarm optimization algorithms. Software paper was published in the *Journal of Open Source Software* (JOSS).

Awards and Certifications

Professional Certifications

Google Cloud Professional Data Engineer (Certification ID: enjfUz) *2018*

Scholarships

Monbugakusho (MEXT) Japanese Government Scholarship *2016*

French Ministry of Foreign and European Affairs Grant *2015*

Department of Science & Technology SEI Merit Scholarship *2011*

Ateneo College Scholarship *2011*

Publications

You can also check my Google Scholar profile. Note: an asterisk (*) denotes equal contributions.

[1] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating Reward Models. *arXiv*, abs/2403.13787, March 2024.

[2] Lester James Validad Miranda. Allen Institute for AI @ SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 151–159, St Julian's, Malta, March 2024. Association for Computational Linguistics.

[3] Lester James V. Miranda. calamanCy: A Tagalog Natural Language Processing Toolkit. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 1–7, Singapore, Singapore, December 2023. Empirical Methods in Natural Language Processing.

[4] Lester James V. Miranda. Developing a Named Entity Recognition Dataset for Tagalog. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 13–20, Nusa Dua, Bali, Indonesia, November 2023. Association for Computational Linguistics.

[5] Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F. Karlsson, Peiqin Lin, Nikola Ljubešić, LJ Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. Universal NER: A Gold-Standard Multilingual Named Entity Recognition Benchmark. *arXiv*, abs/2311.09122, November 2023.

[6] Lester James V. Miranda*, Ákos Kádár*, Adriane Boyd, Sofie Van Landeghem, Anders Søgaard, and Matthew Honnibal. Multi hash embeddings in spaCy. *arXiv*, abs/2212.09255, December 2022.

[7] Lester James V. Miranda and Jinglu Hu. Feature Extraction using a Mutually-Competitive Autoencoder for Protein Function Prediction. In *Proceedings of the IEEE International Conference on System, Man, and Cybernetics (SMC)*. IEEE, October 2018. doi: 10.1109/SMC.2018.00234.

[8] Lester James V. Miranda and Jinglu Hu. A Deep Learning Approach based on Stacked Denoising Autoencoders for Protein Function Prediction. In *Proceedings of the 42nd IEEE Computer Society Signature Conference on Computers, Software, and Applications (COMPSAC)*. IEEE, July 2018. doi: 10.1109/COMPSAC.2018.00074.

[9] Lester James V. Miranda. PySwarms, a research-toolkit for Particle Swarm Optimization in Python. *Journal of Open Source Software (JOSS)*, 3(433), 2018. doi: 10.21105/joss.00433.

[10] Lester James V. Miranda*, Marian Joice Gutierrez*, Samuel Matthew Dumlao, and Rosula Reyes. Appliance Recognition using Hall-Effect Sensors and k-Nearest Neighbors for Power Management Systems. In *Proceedings of the 2016 IEEE Region 10 Conference 2016 (TENCON)*. IEEE, November 2016. doi: 10.1109/TENCON.2016.7847947.