

Statement of Research Interest and Bibliography: LLMs and Endangered Language Revitalization

Jared Coleman

Introduction

This is an informal, ever-changing collection of interesting papers related to current limitations and potential in Large Language Models (LLMs) for low- and no-resource language tasks. LLMs have been shown to exhibit remarkable capability for a wide variety of natural language tasks including machine translation, summarizing, question-and-answering, auto-completion, dialog, and more [2]. LLMs are trained on vast amounts of natural language data from the internet [19] and, as a result, do not perform as well for low- or no-resource languages [4, 27]. We refer to languages with very little publicly available bilingual or monolingual corpora as “low-resource” languages and those with *no* publicly available corpora as “no-resource” languages.

Research Questions

In exploring how LLMs might be used for endangered language preservation and revitalization, we have identified the following research questions as some of the most interesting and important:

- How do models “know” language? This is important for understanding how they might be taught new languages from scratch. By taught, I don’t mean fine-tuned or trained (in the ML sense of the word “train”). Rather, I mean *taught* like a human is taught language: through dialog, question and answering, context, and experience.
 - Black-box experimentation: The past few decades have seen many advances in linguistics through creative black-box experiments.
 - Linguistic Probing: We can perform experiments and “brain-scan” models to see which parts of the underlying network activate to better understand how they work! Interestingly, this has only relatively recently become possible (still with extreme limitations) for humans (via MRI).
 - We care less about whether or not LLMs learn *like* humans and more about understanding how LLMs learn so that we can leverage the knowledge to build useful tools for low/no-resource languages.
- How can we use popular LLM tool-building techniques to create tools for the documentation, preservation, and revitalization of endangered languages?
 - In the context window: few-shot learning, prompt engineering, function calling. We proposed a new approach for low/no-resource language machine translation using a combination of these techniques [6].
 - Tokenization: Can adding tokens for target language words help with natural language tasks?
 - Finetuning: with limited data, fine-tuning is difficult.
- How can LLMs be used for foreign language education
 - Ultimately, the goal of endangered language revitalization is to create new *human* speakers.
 - How can LLMs be used effectively in language education?

Useful Tools Enabled by Research

Pursuing the above research questions will guide and enable the development of many practically useful tools for endangered language revitalization. Some of these tools include:

- Parsing linguistic literature for grammar, vocabulary, etc.
- Summarizing/explaining content for language learners
- Grammar induction
- Auto-completion
- Data sanitization/standardization
- Adaptive data collection: using an LLM to help adjust the questions or queries made to native speakers during data collection to gather the most relevant and useful information.

Special Concerns for Indigenous Communities

When working with indigenous communities in language revitalization, history and context matter. Genocide and forced assimilation [15] have led to the endangerment of many indigenous cultures and languages throughout the United States. In Boarding Schools, indigenous children were forced to abandon their languages and cultures in favor of English and Christianity [14].

Even more modern efforts in language documentation and revitalization can be harmful. My tribe, for example, prohibits the telling of traditional stories except in the winter months. To document these stories and make them publicly available would violate this important tradition. Different indigenous communities have different boundaries and rules for what is appropriate to share and what is not. It is important to respect these boundaries and to work with communities to ensure that the work being done is culturally appropriate and respectful.

Finally, it is imperative that indigenous communities reap the benefits of the work being done to document and revitalize their languages. This means that the tools and resources developed should be made available to the communities in a way that is accessible and useful to them. Another personal example: my great-grandmother was a fluent speaker of our language and so was the subject of a study on our language by a University of California, San Diego Ph.D. student studying linguistics. His thesis "A Grammar Sketch And Comparative Study Of Eastern Mono" [18] is locked behind a Proquest academic paywall and almost impossible for my family and other tribal members to access.

Bibliography

The following bibliography is organized into different categories. Some papers apply to more than one category and therefore appear multiple times.

Our Work

- [6] Jared Coleman et al. *LLM-Assisted Rule Based Machine Translation for Low/No-Resource Languages*. 2024. DOI: [10.48550/arXiv.2405.08997](https://doi.org/10.48550/arXiv.2405.08997).
- [33] Sheng Yu, Jared Coleman, and Bhaskar Krishnamachari. “Chatlang: A Two-Window Approach to Chatbots for Language Learning”. In: (2023). URL: <https://anrg.usc.edu/www/papers/chatlang.pdf>.

Work on LLMs (Large Language Models)

- [2] Sébastien Bubeck et al. “Sparks of Artificial General Intelligence: Early experiments with GPT-4”. In: (2023). DOI: [10.48550/arXiv.2303.08774](https://doi.org/10.48550/arXiv.2303.08774).
- [4] Aakanksha Chowdhery et al. “PaLM: Scaling Language Modeling with Pathways”. In: (2022). DOI: [10.48550/arXiv.2204.02311](https://doi.org/10.48550/arXiv.2204.02311).
- [10] Amr Hendy et al. “How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation”. In: (2023). DOI: [10.48550/arXiv.2302.09210](https://doi.org/10.48550/arXiv.2302.09210).
- [12] Séamus Lankford, Haithem Afli, and Andy Way. “adaptMLLM: Fine-Tuning Multilingual Language Models on Low-Resource Languages with Integrated LLM Playgrounds”. In: *Inf.* 14.12 (2023), p. 638. DOI: [10.3390/INF014120638](https://doi.org/10.3390/INF014120638).
- [13] Patrick S. H. Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle et al. 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>.
- [19] OpenAI. “GPT-4 Technical Report”. In: (2023). DOI: [10.48550/arXiv.2303.08774](https://doi.org/10.48550/arXiv.2303.08774).
- [27] Nathaniel R. Robinson et al. “ChatGPT MT: Competitive for High- (but not Low-) Resource Languages”. In: (2023). DOI: [10.48550/arXiv.2309.07423](https://doi.org/10.48550/arXiv.2309.07423).

Work on Low-Resource Languages

- [1] Jessie Little Doe Baird. “Wopanaak language reclamation program: bringing the language home”. *Journal of Global Indigeneity*, 2(2). 2016. URL: <https://ro.uow.edu.au/jgi/vol2/iss2/7>.
- [7] Serafín M. Coronel-Molina and Teresa L. McCarty. “Indigenous Language Revitalization in the Americas”. In: 2016. URL: <https://api.semanticscholar.org/CorpusID:217243422>.
- [9] Barry Haddow et al. “Survey of Low-Resource Machine Translation”. In: *Computational Linguistics* 48.3 (Sept. 2022), pp. 673–732. ISSN: 0891-2017. DOI: [10.1162/coli_a_00446](https://doi.org/10.1162/coli_a_00446). eprint: https://direct.mit.edu/coli/article-pdf/48/3/673/2040361/coli_a_00446.pdf.
- [12] Séamus Lankford, Haithem Afli, and Andy Way. “adaptMLLM: Fine-Tuning Multilingual Language Models on Low-Resource Languages with Integrated LLM Playgrounds”. In: *Inf.* 14.12 (2023), p. 638. DOI: [10.3390/INF014120638](https://doi.org/10.3390/INF014120638).
- [23] Surangika Ranathunga et al. “Neural Machine Translation for Low-resource Languages: A Survey”. In: *ACM Comput. Surv.* 55.11 (2023), 229:1–229:37. DOI: [10.1145/3567592](https://doi.org/10.1145/3567592).
- [27] Nathaniel R. Robinson et al. “ChatGPT MT: Competitive for High- (but not Low-) Resource Languages”. In: (2023). DOI: [10.48550/arXiv.2309.07423](https://doi.org/10.48550/arXiv.2309.07423).
- [28] Anton Schäfer et al. *Language Imbalance Can Boost Cross-lingual Generalisation*. 2024. DOI: [10.48550/arXiv.1906.04341](https://doi.org/10.48550/arXiv.1906.04341).

- [31] Joshua Taylor and Timothy Kochem. “Access and empowerment in digital language learning, maintenance, and revival: a critical literature review”. In: *Diaspora, Indigenous, and Minority Education* 16.4 (2022), pp. 234–245.
- [32] Daniel Torregrosa et al. “Leveraging Rule-Based Machine Translation Knowledge for Under-Resourced Neural Machine Translation Models”. In: *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks, MTSummit 2019, Dublin, Ireland, August 19-23, 2019*. Ed. by Mikel L. Forcada et al. European Association for Machine Translation, 2019, pp. 125–133. URL: <https://aclanthology.org/W19-6725/>.

Work on RBMT (Rule-Based Machine Translation)

- [11] Tanmai Khanna et al. “Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages”. In: *Machine Translation* 35.4 (Dec. 2021), pp. 475–502. ISSN: 1573-0573. DOI: [10.1007/s10590-021-09260-6](https://doi.org/10.1007/s10590-021-09260-6).
- [22] Tommi A Pirinen. “Workflows for kickstarting RBMT in virtually No-Resource Situation”. In: *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*. Ed. by Alina Karakanta et al. Dublin, Ireland: European Association for Machine Translation, Aug. 2019, pp. 11–16. URL: <https://aclanthology.org/W19-6803>.
- [32] Daniel Torregrosa et al. “Leveraging Rule-Based Machine Translation Knowledge for Under-Resourced Neural Machine Translation Models”. In: *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks, MTSummit 2019, Dublin, Ireland, August 19-23, 2019*. Ed. by Mikel L. Forcada et al. European Association for Machine Translation, 2019, pp. 125–133. URL: <https://aclanthology.org/W19-6725/>.

Work on RAG (Retrieval Augmented Generation)

- [13] Patrick S. H. Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle et al. 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>.

Work on Embeddings & Semantic Similarity

- [3] Jialun Cao et al. “SemMT: A Semantic-Based Testing Approach for Machine Translation Systems”. In: *ACM Trans. Softw. Eng. Methodol.* 31.2 (Apr. 2022). ISSN: 1049-331X. DOI: [10.1145/3490488](https://doi.org/10.1145/3490488).
- [17] Niklas Muennighoff et al. “MTEB: Massive Text Embedding Benchmark”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*. Ed. by Andreas Vlachos and Isabelle Augenstein. Association for Computational Linguistics, 2023, pp. 2006–2029. DOI: [10.18653/V1/2023.EACL-MAIN.148](https://doi.org/10.18653/V1/2023.EACL-MAIN.148).
- [25] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. DOI: [10.48550/arXiv.1908.10084](https://doi.org/10.48550/arXiv.1908.10084).
- [30] Yurun Song, Junchen Zhao, and Lucia Specia. “SentSim: Crosslingual Semantic Evaluation of Machine Translation”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Ed. by Kristina Toutanova et al. Association for Computational Linguistics, 2021, pp. 3143–3156. DOI: [10.18653/V1/2021.NAACL-MAIN.252](https://doi.org/10.18653/V1/2021.NAACL-MAIN.252).

Work on Linguistic Probing

- [5] Kevin Clark et al. “What Does BERT Look at? An Analysis of BERT’s Attention”. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*. Ed. by Tal Linzen et al. Association for Computational Linguistics, 2019, pp. 276–286. DOI: [10.18653/V1/W19-4828](https://doi.org/10.18653/V1/W19-4828).

Other References

- [14] K Tsianina Lomawaima and Teresa L McCarty. *"To remain an Indian": Lessons in democracy from a century of Native American education*. Teachers College Press, 2006.
- [15] Benjamin Madley. *An American Genocide: The United States and the California Indian Catastrophe, 1846-1873*. Yale University Press, 2016.
- [16] Christopher Moseley. *Atlas of the World’s Languages in Danger*. Unesco, 2010. ISBN: 978-92-3-104096-2.
- [29] SIL International. *639 Identifier Documentation: mnr*. Accessed: 11 Mar 2024. 2024. URL: <https://iso639-3.sil.org/code/mnr>.