AI in language work

February 2025

What is AI?

'Artificial Intelligence' or 'AI' describes using machines that are given information and some basic rules, and are tasked with learning to identify patterns in that information that they can then apply to new information to make decisions about that information (this process is also called 'machine learning'). Current approaches to AI in language work use machine learning over large amounts of information (mostly written language) in order to understand or produce language-related outputs.

How might AI help?

There are several kinds of AI that are proving to be useful in language work. These include:

- AI can help recognize text in your PDFs and make it available for further use through '**Optical Character Recognition**'.
- AI can help go through a lot of text and identify the different words and word parts in your language, and figure out how they are put together through '**Parsing, Tokenization, Labeling and Annotation**'.
- AI can help match audio materials to transcriptions, and then use what it learns that way to create new draft transcriptions of audio materials through '**Automatic Speech Recognition**'.
- AI can help match text in your language to English (or other language) translations, and use what it learns to produce draft translations through '**Machine Translation**'.
- AI can help search through materials in a set of documents to find out what information is in these documents through '**Information Extraction**'.

How might AI hurt?

- All types of AI require very large amounts of 'training data'. Most major commercial AIs are trained on huge amounts of written language which is mostly in English, and which has been scraped from the Internet. Since humans' use of language online occasionally reflects human biases, those biases can be replicated or amplified when that data is used to train an AI system. Therefore, commercial AI products are highly **English-centric and biased**.
- All types of AI require large or very large amounts of processing power to train and to use. This processing power is provided by special kinds of computer hardware (i.e. 'GPUs', 'TPUs', or 'NPUs') that are usually housed in data centers around the world. This processing power is expensive, and it uses very large amounts of energy to run. The data centers also use tremendous amounts of water to keep this hardware cool. Therefore, commercial AI products, and even local AI projects, are **costly and environmentally harmful.**
- Working with AI may result in your own data being made available to others in ways that you did not agree to, and do not want. Issues of **intellectual property rights and data sovereignty are** not yet worked out in a way that protects your rights.
- Machines do not learn as well as humans do. Therefore, AI products never produce 100% accurate results. Some types of AI are known to 'hallucinate' - that is to make up information that looks plausible but is wrong. **AI products produce inaccurate results**.

How can you use AI (safely and effectively)?

If you are interested in pursuing AI for language work, you can mitigate the risks and take advantage of the benefits of AI by investing in:

- **People** who can write relevant computer code and are trained in **machine learning**. Most AI requires coders who are proficient in the computer language **python**.
- **Computers** that can handle AI development, which is very resource-intensive. This means either having your own hardware that has **high-speed processors** and graphic processing units (**GPU**s), or having access to a cloud-based or remote high-speed computing resource.
- **Examples** that can be given to the computer so it can detect patterns. Computers need a lot of examples to learn from, and the more examples you have, the better they do. Examples include **hand-typed** versions of PDFs, **hand-annotated** text files showing words, word parts with relevant labels, **hand-transcribed** audio or video materials, and **hand-translated** materials.

How can you learn more?

- Reach out to friends and relatives who are already using AI in language work! If you're not sure who is using these systems, ask about NLP or AI for language work.
- Reach out to your local computer scientists or electrical engineers! If you don't have a computer science or engineering department, reach out to peer Colleges and Universities to find scholars working in NLP. You can also access the Advancing Indigenous Language Technologies (AILT) network at https://ailt.arizona.edu.
- Reach out to scholarly societies! The Computational Linguistics Society has a Special Interest Group on NLP for Endangered Languages (their terminology, not ours): https://acl-sigel.github.io/.

But always remember….

No outside scholar, engineer, or language activist knows your community's interests and needs better than you do. Always make sure that you communicate clearly about your views on the ethics, norms, and expectations of outside collaborations. Look for partners who are more interested in listening than in talking, and talk with trusted others about their experiences before agreeing to work on projects, spend resources, or share information with outside individuals or groups.

Resources

Barness, S, J. Cummins, M. Fernandez, A. James, K. Pierce Farrier, J. Pringle, S. Russo Carroll, R. Taitingfong & A. Wieker. 2023. CARE Data Principles, Indigenous data, Data related to Indigenous Peoples and Interest. https://github.com/DataCurationNetwork/data-primers.

First Nations Technology Council. 2022. Our Work. https://www.technologycouncil.ca/our-work/.

Native Nations Institute. 2024. Indigenous Data Sovereignty & Governance Publications. https://nni.arizona.edu/publications/indigenous-data-sovereignty-governance-publications