Language technologies for supporting multilingual scholarly communication

Lynne Bowker (lbowker@uottawa.ca)

School of Translation & Interpretation

Helsinki Initiative 5th Anniversary, 3 April 2024

Major changes in technological landscape

- 1940s to 1990s
  - Computers were brand new (enormous, limited power, speed, storage)
  - Few people had one
  - Stand-alone
- MT = linguistic, rule-based

- 1990s to present
  - Computers are ubiquitous (small, fast, powerful, lots of storage)
  - Everyone creates digital text
  - Networked (internet, www, intranets), easy to access/share
- MT = **data-driven**

What is data-driven? Machine learning?

Word embeddings?

Transformers?

Vectors?

Data!

(fuel)

Image credit: Pixabay.com

Training data = examples (for machine learning)

1) Identify a **task** for the AI tool
   - e.g. image classification, translation
2) Show the tool **examples** of what you want it to learn
   - e.g. photos of two different types of animals, previously translated texts
3) Give the tool some **feedback** (e.g. confirm correct answers)
4) Test the tool on **new** data that it hasn't seen before

Narrow vs **general** tasks

MANY, MANY examples

- (*Enough* fuel)

High quality examples

- (The *right kind* of fuel)

AI is *not smart*

- It can **process** data, but it doesn't **understand** it

Data-driven approaches have strengths

- Free versions available
- Convenience (24/7)
- Fluent (sounds good)
- Able to learn (patterns)

- Work well for **high-resource** languages, domains and text types

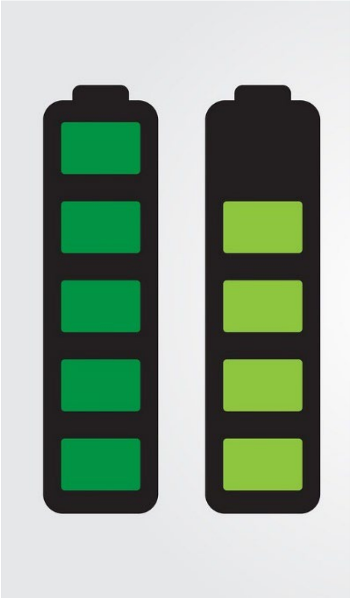


Image credit: Pixabay.com

But also limitations

- Hallucinations
- Only does pattern matching + counting (no understanding)
- Data-driven = data-sensitive (e.g. bias, including lang variety)
- Perform less well for low-resource languages/pairs, domains and text types
  - Google Translate = 134 languages (/7000+)
  - "No Language Left Behind" = 200 languages
  - Overwhelming use of EN for scholarly communication means some languages don't have well developed scientific terminology

  - Specialized content = lower volume
  - Paywalled content = lower volume
    - Open access is important!

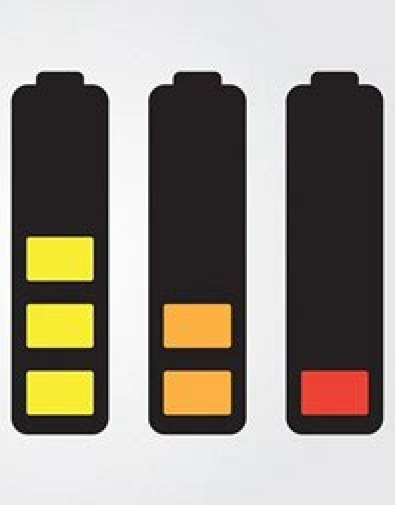


Image credit:

Pixabay.com

Other costs

- Extremely computationally intensive
  - Only large corporations can afford to develop, train and fine tune very large-scale models
    - Determines who can and cannot participate
- May not remain freely accessible forever
- Not environmentally friendly
  - Training one model = carbon footprint of 6 cars
- We have work to do to develop more efficient algorithms, SML, etc.

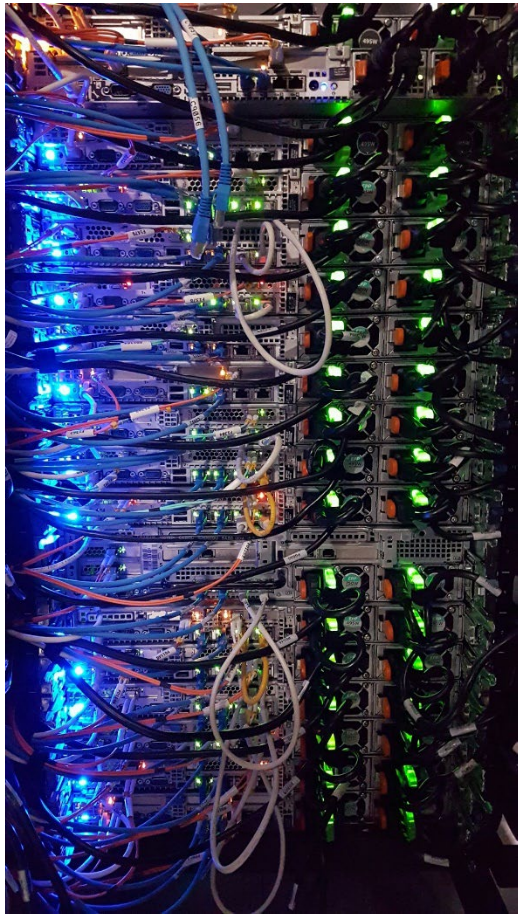


Image credit: Pixabay.com

NMT vs LLM

NMT

- Dedicated to translation
- Parallel corpora (equal amount of data in both languages)
- *May* pivot through EN or another language, but not as often
- *Usually* a better choice when the task is translation (esp. for low-resource situations)

LLM

- Multi-use tool
  - Q&A, summarization, paraphrasing, translation
- Often unbalanced language resources (e.g. 90% of ChatGPT's corpus is in EN, remaining 10% covers all other languages)
- Often pivots through EN behind the scenes
- EN language also equals EN (US) worldview

Garbage in, garbage out!

- Quality of input text affects quality of translated text
    - Well-written input (plain language) is more translatable
- We can **ALL** work to craft clearer input (reader- and translation-friendly writing, intralingual translation)
- Plain language summaries
- *NOTE: post-editing will likely still be necessary

- **FREE** resources on Machine Translation Literacy Project site (>>Teaching Resources)
- https://sites.google.com/view/machinetranslationliteracy/

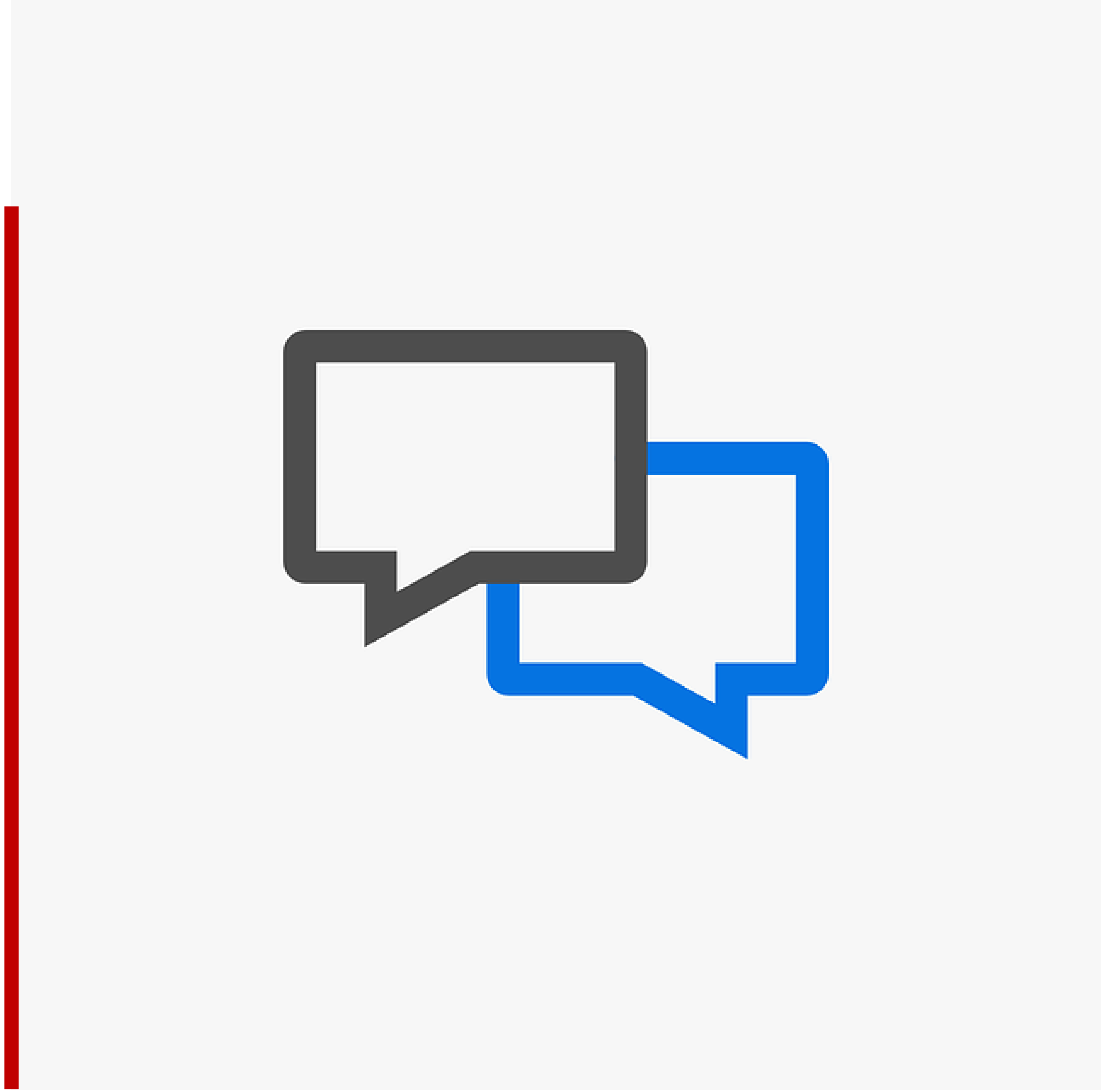
MT can help… but MT *alone* is NOT sufficient

- Policies to value and promote multilingual publishing
- Multilingual metadata to support discovery
  - MT better suited to support reading work in other languages, rather than writing it
- Human-computer interaction
  - OPERAS, CLF
- Beyond published articles (slides, posters, presentations)?

Bowker, Ayeni & Kulczycki (2023)

- Systematic review of literature at the intersection of translation technology and scholarly communication

https://doi.org/10.20381/858s-q632



lbowker@uottawa.ca

(Image credits: Pixabay.com)