Machine Translation and Scholarly Communication

Lynne Bowker (lbowker@uottawa.ca)

School of Translation & Interpretation and School of Information Studies, University of Ottawa, Canada

NAWA Visiting Researcher, Scholarly Communication Research Group, Adam Mickiewicz University, Poland

Machine translation

- Long history, linguistic approaches
  - Slow, low quality, not widely used
- Post 1990s
  - Fast, powerful, cheaper computers
  - Digital document production
  - Internet
- Data-driven approaches to MT
  - Vast improvement in translation quality, but still not problem-free

(Image credit: Pixabay.com)

Data-driven approaches

- Need **HUGE** quantities of data

  But also

- The **right kind** of data
  - Generative, **derivative**
  - New content based entirely on **examples** in training corpus

- **High-resource** and **low(er)-resource** situations

- Think of the “hands” problem with AI-generated images
  - Fewer examples, many variations

(Image credit: Pixabay.com)

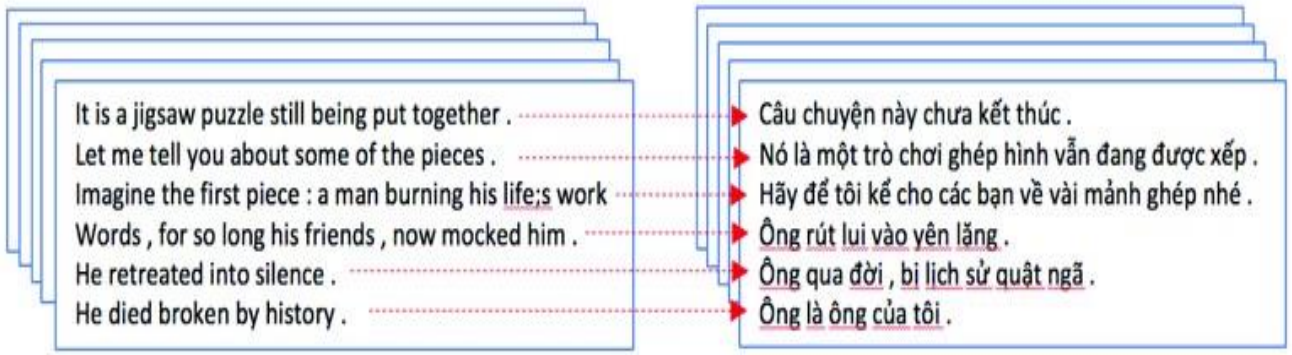
High- and low-resource situations

	High-resource language	Lower-resource language
High-resource domain/text type	• English computer user manuals • French recipes	• Polish administrative texts • Finnish texts on telecommunications
Lower-resource domain/text type	• English texts on reindeer husbandry • French texts on cricket	• Polish nuclear reactor maintenance manuals • Finnish texts on dengue fever

Translation requires a bilingual parallel corpus

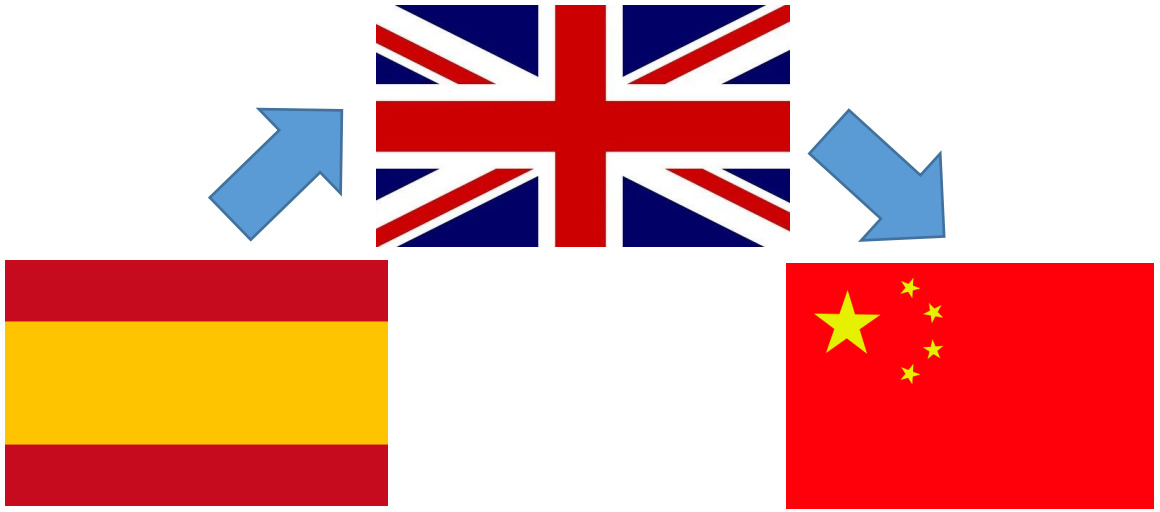
- Need a lot of translation activity between a given **pair** of languages
- High-resource languages
  - **Russian** and **Bengali** both have ~250 million speakers

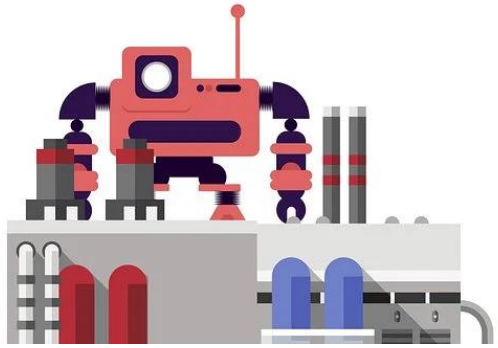
  BUT
- Relatively little translation activity **between** them = **low-resource language *pair***
- Language **varieties**
  - Less dominant varieties get further marginalized



Approaches to overcoming low-resource situations

- **Pivot** through another language
  - Twice as many opportunities to make errors
  - Errors in step 1 carried forward into step 2
- Make **synthetic** data
  - Using MT to translate and feeding the results back into the training corpus
  - Lower quality input

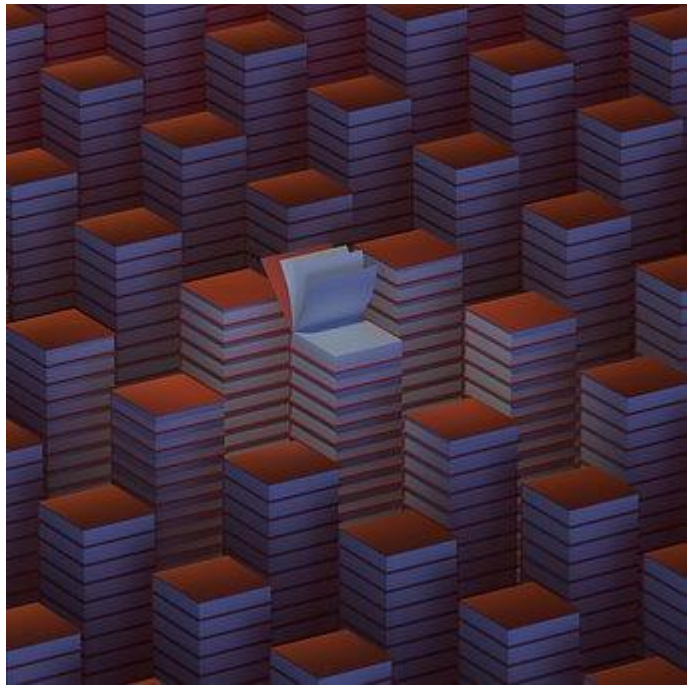




(Image credits: Pixabay.com)

Machine translation & scholarly communication

- Challenge
  - Many languages and domains are low(er) resourced
- What can **authors** do?
  - Produce **accessible** resources in other languages
    - **Open access**
    - Multilingual **abstracts**
    - **Science communication** (plain language summaries)
  - Improve their **MT literacy**
    - Understanding how data-driven MT works
    - Translation friendly writing



(Image credit: Pixabay.com)

Translation friendly writing

- Applies to **any** language
  - Specific techniques may be language dependent
- Quality of the **input** text has a major impact on **translation quality**
  - Authors of the original text can make it easier for others to read their work through an additional language or via MT
- "Garbage in, garbage out" (**GIGO**)
  - Write clearly and unambiguously

(Image credit: Pixabay.com)

Writing for (machine) translation

- Skilled human translators can **compensate**
  - **Fix** spelling, grammar or punctuation errors
  - Make **educated** guesses
  - Refer to other parts of the text for **context**
- MT tools **cannot** compensate
  - Process the character string **as presented**
  - Have **no real-world knowledge**
  - Process each sentence as a **separate unit**

(Image credit: Pixabay.com)

Sample tips (not language specific)

- Use spelling and grammar checkers
- Ensure sentences are not too long… or too short (~10-25 words)
- Use terminology consistently
- Minimize acronyms, abbreviations
- Minimize pronouns
- Avoid using all caps
- Avoid unnecessary line breaks

(Image credit: Pixabay.com)

Using other AI tools to support MT

- ChatGPT-4 (using a LLM for MT)
  - **Underperforms** for low-resource languages, distant pairs
  - **Prompting** is important
    - "You are a machine translation system"
    - Task- and domain-specific prompts
  - **Adjusting "temperature"**
    - High (1) = more variety, more potential for errors (responses to the same prompt vary)
    - Low (0) = more conservative, predictable (get the same response given the same prompt)

(Image credit: OpenAI.com)

Other language-related tasks for AI

- REMEMBER: Translation is a hugely **complex** task

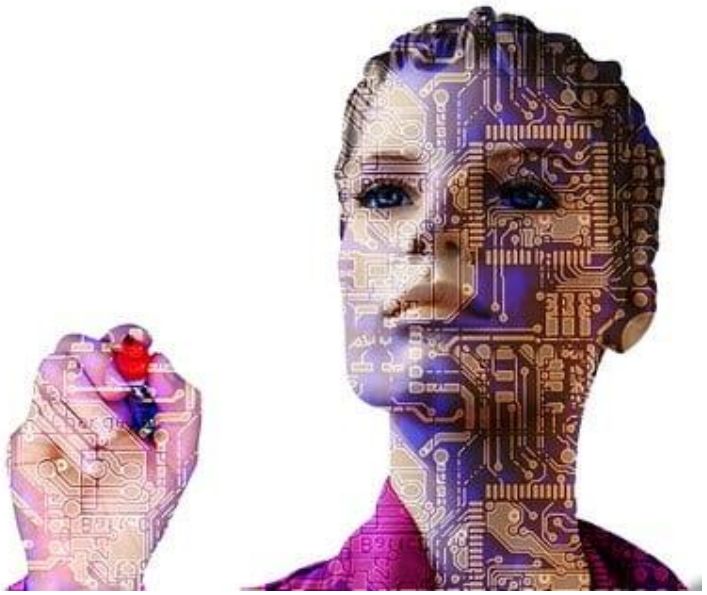
- Other AI tools can help you to get your text in shape for MT
  - **Paraphrasing**, e.g. ParaphraseTool
  - **Simplification**, e.g. Simplif.ai
  - **Summarization**, e.g. Genei

- Other tools
  - **Plain language** checkers (e.g. Hemingway)
  - **Readability** checkers (e.g. in MS Word)
  - **Read-aloud** features (e.g. in MS Word)

(Image credit: Pixabay.com)

Other issues?

- Ethical considerations
  - **Evolving** landscape of attitudes and practices
  - Distinction between **content generation** and **linguistic support**
  - Need for more **nuanced** policies, not simple bans



(Image credit: Pixabay.com)

Beyond authors?

- How else can MT help?
- Knowledge **discovery**
  - Using MT to translate metadata for cross-language information retrieval
- Knowledge **assimilation**
  - Translation-friendly writing + MT into your dominant language: your linguistic and domain knowledge should help you to compensate for some translation issues
  - Multilingual **subtitles** (speech generation + MT)
- Writing in other languages
  - Combined with other tools, MT can be used iteratively as a 2nd-language **writing aid**

(Image credit: Pixabay.com)

Comments or questions

lbowker@uottawa.ca

(Image credit: Pixabay.com)

Find out more