Humans, Machines, and Meaning: Rethinking Subject Indexing in the Age of AI

Clara Wan Ching Ho

FID Linguistik | University Library JCS Frankfurt

Hi, here’s a bit about me

- Computational linguist
- University Library JCS at Goethe University Frankfurt
- FID Linguistik Project

  FID - Fachinformationsdienst / Specialised Information Services

  - The Bibliography of Linguistic Literature (BLL)
  - Ontology of linguistic subjects from the BLL
  - The Lin|gu|is|tik portal

Who are you?

Who are you?

Cataloguers /
Indexers?

Who are you?

Cataloguers / Indexers?

Researchers?

Who are you?

Cataloguers / Indexers?

Others?

Researchers?

Transitional era of AI integration:
What do we trust to delegate to machines / “AI”?

If AI could help you with indexing,
what do you want?

Indexers’ dream tools

1. Translation tool

Indexers’ dream tools

1. Translation tool
2. Key topic statistics

Indexers’ dream tools

1. Translation tool
2. Key topic statistics
3. Rare/new terms identifier

From the perspective of indexing for a specialized bibliography

Challenges in indexing

- Balancing time in reading and information obtained
- Looking up relevant keywords to the ones they can come up with
- Weighing the importance and rarity of a term
- Deciding whether a publication fits into the profile of the specialized bibliography

Does end-to-end subject indexing/suggestion solve these problems?

Does end-to-end subject indexing/suggestion solve these problems?

Depends on what the system offers.

Advantages of using GenAI/LLM

1. Scalability: Shorter time, quicker decisions
   - Diminish language barrier
   - Machine reading full text instead of humans skimming and scanning
   - All the *most probable* keyword options are in front of you

Advantages of using GenAI/LLM

1. Scalability: Shorter time, quicker decisions
   - Diminish language barrier
   - Machine reading instead of human skimming and scanning
   - All the *most probable* keyword options are in front of you
2. Consistency: One machine = One brain
   - Inter-indexer consistency: 38-49%; 5-16 terms/doc (Medelyan and Witten, 2006)
   - Number of subjects assigned
   - Importance of subjects based on the same criteria prescribed

Limitations of using GenAI/LLM

1. Explainability and accountability
   - “Where is it coming from?”
   - No guarantee it follows the guidelines in using the controlled vocabulary consistently
2. Data integrity and consistency
   - Hallucination
   - Understanding bias based on training data

Limitations of using GenAI/LLM

3. Keeping the controlled vocabulary updated
   - Does not support the creation of new keywords in vocabularies
4. Technical issues
   - Processing books requires bigger LLMs
   - Less attention to earlier tokens as context length increase
5. Will AI prime humans?
6. Limitations in physical books

Does end-to-end subject indexing solve these problems?

Challenges in indexing

- Balancing time in reading and information obtained

  Maybe - but they still need to read enough to verify

- Looking up relevant keywords to the ones they can come up with

  Yes - if more subject terms are suggested

- Weighing the importance and rarity of a term

  No - topics mentioned in smaller sections could be overlooked

If decision making should be done
only by indexers, what do machines do?

Division of labour between AI & Indexers

AI

- Help navigate through the documents

- Flag publications with potential new terms, store in system & <u>create statistics</u>

Indexers

→ Decide on where to read and what is important

→ Decide on whether new subjects are needed based on usage

→ Decide on the terms' position in the controlled vocabularies

Viable options for human-machine collaboration

- Named Entity Recognition (NER)
- Term Frequency–Inverse Document Frequency (TFIDF)
- Relevant keywords finder
- Retrieval Augmented Generation (RAG)

Viable options for human-machine collaboration

- Named Entity Recognition (NER) and Stats

  → Help navigation through documents

- Term Frequency–Inverse Document Frequency (TFIDF) and Stats

  → Identify rare terms in the publication

- Relevant keywords finder
- Retrieval Augmented Generation (RAG)

Viable options for human-machine collaboration

- Named Entity Recognition (NER)
- Term Frequency–Inverse Document Frequency (TFIDF)
- Relevant keywords finder

  → Expose indexers with frequent collocations of the subjects used

- Retrieval Augmented Generation (RAG)

Viable options for human-machine collaboration

- Named Entity Recognition (NER)
- Term Frequency–Inverse Document Frequency (TFIDF)
- Relevant keywords finder
- Retrieval Augmented Generation (RAG)

→ Possibility to query with directly the passages to focus on

Before we try to give the whole task to “AI”, there are many ways we can support our indexers with it.

Thank you for your attention!
Questions and comments are welcome.

Contact:

Clara Wan Ching Ho

Universitätsbibliothek Johann Christian Senckenberg

c.ho@ub.uni-frankfurt.de