Using ChatGPT for (semi-) automatic subject indexing of different document types

Koraljka Golub[1], *Jue Wang*[2], *Johannes Widegren*[1]

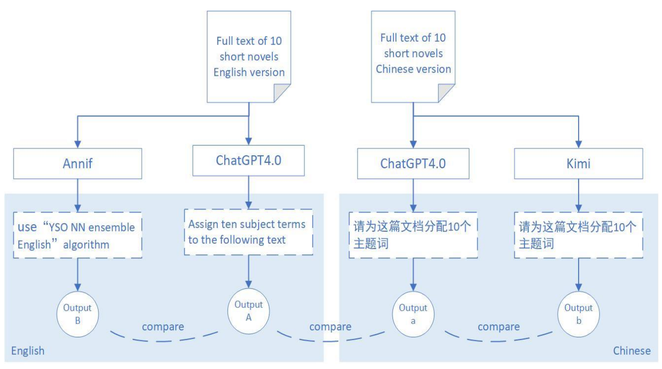
[1]LINNÆUS UNIVERSITY, [2]UNIVERSITY OF CHINESE ACADEMY OF SCIENCE

Introduction

This poster presents a pilot study on the potential use of OpenAI's ChatGPT4 for automatic subject indexing of archival documents in Swedish, Swedish LGBTQ fiction and Chinese fiction. The accuracy of the assigned subject index terms is compared with the output from Annif, an established automatic subject indexing software used in libraries.

O. Henry's short novels in Chinese and English

Data:



Method:

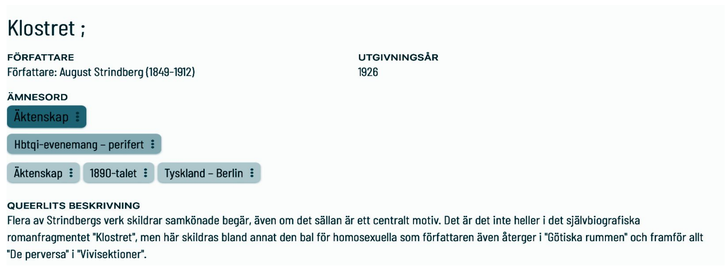
1) Download the English and Chinese versions of the 10 novels
2) ChatGPT4: "Assign 10 subject terms to the following text"
3) Annif: 10 keywords; "YSO NN ensemble English"
4) Kimi:"请为这篇文档分配10个主题词"

Result:

In the comparison between the Chinese and English versions of ChatGPT, an average of 4.7 subject terms with the same or similar meanings were cited. The highest is 7 and the lowest is 2. In the comparison of Kimi with ChatGPT, there are on average 4.4 identical or similar subject terms, Kimi has a more specific description of the original storyline. Overall, **ChatGPT performs well in both Chinese and English, and Kimi focuses more on textual details, while Annif does not reflect the topic of the article well,** but only gives an approximate idea of the domains that the text covers.

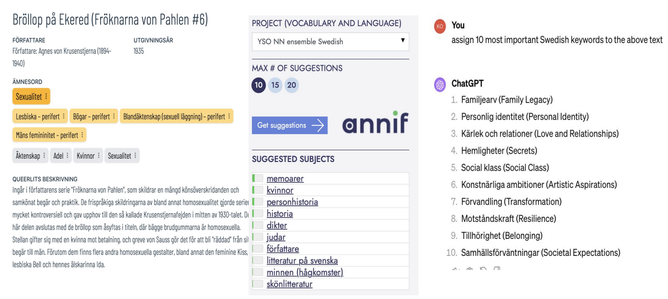
Swedish LGBTQ+ Fiction

Data: 10 books indexed in Queerlit and that were available as full text at Litteraturbanken



Method:

1) First 10,000 words from the novel were selected and pasted into Annif and ChatGPT
2) ChatGPT4: "Assign 10 most important Swedish keywords to the above text"
3) Annif: 10 keywords; "YSO NN ensemble Swedish"



Result: Of the total of 39 LGBTQ+ terms assigned to the 10 works in the sample either as major (n=22) or minor (n=17) index terms, **none of the works were identified as LGBTQ+ by either ChatGPT or Annif.**

Historical Newspaper Articles Pertaining to the Sámi

Data: 10 historical newspaper articles (<1914) featuring the search term *lapp* downloaded from *Svenska Tidningar* (tidningar.kb.se)

Aftonbladet Halvveckoupplagan	1903-09-05	Ett nödrop från lapparna
Skellefteå nya tidning	1893-02-16	Renskötseln och nöden i öfre Norrland
Gefleposten	1904-10-29	Lappfrågan ånyo till k. m:t

Method:

ChatGPT4: "Assign ten subject terms to the following text"

Annif: 10 keywords; "YSO NN ensemble Swedish"

Result: **The terms assigned by ChatGPT were found to correspond well with the subject content of the articles overall.** Annif made more mistakes (3.5 per article) and frequently assigned very generic terms.

<u>Conclusion</u>: ChatGPT4 was very accurate compared to Annif in assigning subject index terms for Chinese fiction novels and historical newspaper articles pertaining to the Sámi. Both failed to identify LGBTQ+ themes in Swedish fiction, however.

The results can be further improved by running the output from ChatGPT through Annif, taking advantage of both ChatGPTs impressive language understanding and Annif's support for controlled vocabularies.