Martin Malmsten, Viktoria Lundborg, Elena Fano, Chris Haffenden, Fredrik Klingwall, Robin Kurtz, Niklas Lindström, Faton Rekathati and Love Börjeson

13 Without Heading? Automatic Creation of a Linked Subject System

Abstract: Can problems with library subject headings systems designed and operated by humans be mitigated by machine learning? This chapter discusses work undertaken at *Kungliga biblioteket*/National Library of Sweden (KB) to explore the use of an automated system for identifying subjects that would sidestep the need for fixed headings from a controlled vocabulary. The new approach taken combined the natural language processing (NLP) capacities of large language models (LLM) with topic modelling, to cluster and order texts according to topics derived from the material, rather than any pre-defined terms reflecting existing biases. The chapter explains how BERTopic was tested on a text corpus compiled from KB's digital collections and presents and analyses the results. The value of such critical exploration is to illuminate the shortcomings of existing systems that might otherwise remain unnoticed.

Keywords: Subject headings; Knowledge organization in subject areas; Machine learning; Linked data

Introduction

How can libraries create order and provide effective access to information in an era of Big Data? While traditional means of knowledge organisation and document arrangement used by libraries such as subject headings and classification systems are useful, they demand considerable resources to maintain and are prone to bias given their rigid nature in an ever-changing context. A more dynamic option stems from recent developments within machine learning (ML), where the emergence of transformer models has enabled a subject system that is content-based, can be created and re-created at a moment's notice, and allows its biases to be measured. Can such an automated approach complement, enhance, or even replace existing systems?

This chapter describes a project exploring the possibility of fully automating the creation of a subject heading system, albeit without the actual headings. To achieve the desired outcome, novel methods from natural language processing (NLP) were combined with topic modelling techniques to create vector-based clusters derived

https://doi.org/10.1515/9783111336435-014

from the content of the collections under consideration. Subjects could thereby be defined without the constraint of using words or phrases to denote a pre-existing and inherently biased concept. While opening up new forms of granular and empirically-driven indexing, the new approach retains the functional requirements of a traditional subject headings system, including the ability to expose the system as linked data and to provide a useful search tool for users. Allowing ML methods to create particular topics based on the specifics of the material under consideration suggests one way in which the arbitrary role of human bias in existing systems of subject indexing might be countered.

This chapter provides a vision of how an automated linked subject system without predetermined headings might work. The first part offers a brief contextualising discussion of subject analysis and topic modelling, pointing to various ways in which the problems of the former could be mitigated by the advantages of the latter. The consideration of both manual and automated approaches stems from the multidisciplinary character of the project group, which combined expertise in manual indexing systems on the one hand with innovative perspectives of AI and data science on the other. The second part of the chapter explains how a sentence BERT language model trained at KBLab, the digital research infrastructure at *Kungliga biblioteket*/National Library of Sweden (KB) (Börjeson et al. 2023) as the basis for a new transformer-based approach to topic modelling called BERTopic (Grootendorst 2022). A description of how the approach was tested on a text corpus compiled from the library's digital collections is provided before a discussion and evaluation of the results. The chapter concludes with broader reflections on the value of such experimentation in illuminating the shortcomings of existing systems that otherwise tend to be normalised to the point that problems are no longer noticed.

Subject Analysis of Library Collections

The history of libraries readily demonstrates the suggestion that "to classify is human" (Bowker and Star 1999, 1). Since at least as far back as the ancient Babylonian library of Ashurbanipal (Finkel 2009) some form of subject analysis has been used in libraries. The concept of aboutness has received considerable attention in the literature (Beghtol 1986; Hutchins 1978; Yablo 2014). In this chapter, the terms *subject*, *theme*, *aspect*, *topic* and *cluster* are all used in relation to aboutness. *Topic* and *cluster* are used to denote topics extracted using topic modelling techniques, whereas *subject* is used to refer to existing subjects in a subject heading system. The terms are not entirely equivalent, but in this context and in actual usage they are

similar enough. Any attempt to distinguish them further could lead to an ontological rabbit hole from which it would be difficult to escape. Similarly, *theme* or *aspect* is used to designate the actual subject matter of a particular topic or subject.

Catalogues and subject systems have been utilized by both library professionals and users to organise, navigate and search large collections. Such tools greatly reduce the time needed to query a collection for resources in a given category, and together with classification schema have played a crucial role in creating order. The principal value of using words and controlled vocabularies to embody a subject is to convey its boundaries to a user or, with the addition of a scope note, a librarian, based on the assumption that the receiver of the information will use language and general knowledge to quickly grasp the intended scope and meaning of the subject. It depends, in short, on a common understanding of the world to communicate, receive and understand information in a highly compressed form.

A subject heading system is useful when exposing library data as linked data, making it machine readable (Malmsten 2009). Knowledge organisation systems for subjects are easily modelled and can map relationships and connections to other subjects, indicating if a subject is broader, narrower, or somehow related. Many subject headings systems with controlled vocabularies have been developed throughout the world, the most well-known being the Library of Congress Subject Headings (LCSH). Specialist subject headings lists or thesauri exist in specific subject areas, for example Medical Subject Headings (MESH). The various systems can be linked. For example, the *Svenska ämnesord*/Swedish subject headings (SAO) terms can be connected with LCSH.

Yet subject heading systems are not without their problems. The most obvious is the lack of shared understanding. As soon as more than one person is involved in classifying a subject area or searching for information, it becomes apparent that people have slightly, or wildly, differing ideas about the nature of a subject, both when choosing a heading within the system to apply to an item, or when using it to search. “Indexers do not agree with each other nor themselves” (de Keyser 2012, 47). One could say the same for people in general. And the problem is only exacerbated when linking subject heading systems from differing origins. Moreover, there is the ever-looming question of bias. People building subject systems are biased; people applying subject terms are biased; and the people using the systems are biased (Knowlton 2005; Olson 1998). One might argue that knowledge organisation formalises bias, based on the mistaken assumption that everyone shares the same biases. Martin notes that bias can have significant consequences: “Naming controls what can and cannot be easily talked about, grappled with, and faced” (Martin 2021, 282).

Beyond Machine Indexing

In Bloomfield's cutting terms, "Machine indexing is noted as 'rotten' and human indexing as 'capricious'" (2002). Even if human indexing is unreliable, can machine indexing or automated subject indexing still be dismissed? The answer is not necessarily. While techniques for automatic clustering and indexing of text have existed for some time, recent developments in ML offer potentially transformative new possibilities. Topic modelling is an example. It refers to a technique where topics are extracted automatically from a corpus of text fragments (Blei 2012). Each text fragment is then assigned to one or more of the topics. Topics deemed close to one another can be clustered to create broader topics.

> Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time… [T]he topics emerge from the analysis of the original texts. Topic modeling enables us to organize and summarize electronic archives at a scale that would be impossible by human annotation (Blei 2012, 77–8).

Topic modelling produces a structure similar to that of a subject headings system that can be exposed in a machine-readable way. It previously required a large amount of pre-processing and tended to be regarded as a qualitatively inflected method demanding substantial prior knowledge of the corpus to apply it correctly. However, the emergence of large, transformer-based language models like BERT has changed the situation dramatically (Devlin et al. 2019). Using the new models significantly reduces the need for pre-processing and specific corpus knowledge. In short, topic modelling has been made far easier since the incorporation of BERT (Fano and Haffenden 2022).

The project hypothesised that new AI techniques could be leveraged to achieve benefits similar to those of existing subject headings systems but without some of the drawbacks. The lack of commensurability that besets systems created and applied by humans could be mitigated by a fully automated approach in which a single neural network provides a type of cohesive universal understanding. By exploring such an alternative, the value of hand-crafted subject heading systems when used by end users would also be questioned. Experience has shown that users deploy the headings unknowingly when searching and do not browse the subjects *per se* (Antell and Huang 2008). Users rarely understand the systems they are using which suggests that if the second order effects of a subject heading system can be replicated, the system itself does not need to be legible to a user.

In using BERTopic, the approach differed from other efforts to automate subject indexing within predetermined frameworks (Golub 2021). Systems like Annif in Finland (Suominen 2019) and more recently Kratt in Estonia (Asula et

al. 2021) produce a form of semi-automated subject indexing that assigns documents to given subject headings or controlled vocabularies. In essence, they replace manual indexing with automation but reproduce existing practices and forms from established knowledge organisation systems. By contrast, it is precisely these practices themselves that could be flawed, and rather than emulating them they should be replaced by an alternative, ML-based system that can be applied in an unsupervised manner. The Swedish project critically probed the suggestion that "algorithms are really not able to entirely replace the intellectual work of subject indexing professionals" (Golub 2021, 703). The framework used by professionals may be part of the problem and any solution must be found outside the box.

Subject Analysis in Sweden

KB maintains the primary subject heading system used in Sweden: *Svenska ämnesord*/Swedish subject headings (SAO). SAO is used in the Swedish union catalogue Libris by all types of libraries including academic, public and special libraries and other cultural institutions like museums and broadcasting companies for analysing materials ranging from books to videogames. SAO is a subject heading system that is based on international principles (Chan 1990), including:

- Current usage; the preferred terms should mirror what a subject is usually called
- Literary warrant; there is at least one resource in Libris for each heading
- Uniform heading; one heading per subject
- Unique heading; one subject per heading
- Specific and direct entry; the most precise term naming the subject
- Stability; a heading is changed only when authoritative sources consistently use a new term
- Consistency; regarding form and structure for similar heading, and
- References and relationships; synonyms are recorded as variants and the subject headings are placed in hierarchies with relationships to related terms.

The indexing guidelines follow international principles such as depth, covering the main points, and assigning headings only for topics that comprise at least 20% of the work, and specificity, assigning headings that precisely represent the subject content.

The purpose of using controlled subject headings is to provide standard subject access to resources in a given collection. By using a controlled vocabulary based on the principles outlined, there is a predictability regarding the indexing that should

enable the user to find everything about a given subject. Indexing in Libris using the SAO system is undertaken by human cataloguers. Some are librarians who work with cataloguing on a full-time basis; some work part-time for only a few hours a week; and others conduct indexing as part of their jobs.

In practice, indexers use subject heading systems in slightly different ways, which tends to lead to inconsistency. Not all resources about the same topic are analysed in the same way. There are many reasons for the different approaches:

- Not all indexers have studied the subject heading manual or read the individual scope notes connected to specific subject headings
- The subject heading system allows for subjects and related subjects to be described in more than one way. For example, a resource about gender inequality in the job market may be indexed with some or all of these headings: Women—Employment; Sex discrimination; Equality; Women; Men; Men and women; Labour market; Labour market—Gender aspects
- The tools that cataloguers use to find appropriate subject headings are not optimised. There are challenges in indicating specific narrower terms and understanding scope notes when choosing headings
- The search options for finding related bibliographic records on the same subject might be limited which is also an issue for users, and
- Indexers are human individuals with diverse life and education experiences.

In commenting previously on inconsistency in human indexing, the wording "tends to" was used. There is no specific research regarding the quality of the manual subject indexing in Libris, but more general studies on indexing such as de Keyser's have noted the issues (de Keyser 2012, 40–47).

The SAO system was not constructed from scratch using the above-mentioned principles. Rather, it was launched as a subject heading list in 2000 based on a Swedish classification system's subject index and has since grown into a more thesaurus-like form. Hierarchies and LCSH-mapping have been added in piecemeal fashion over the years, and there are still some parts of the system that consist of single terms with no relationship to other terms. The isolated terms are at higher risk of not being found and used by cataloguers. Another challenge is that maintaining a manual subject heading system is a costly enterprise. SAO is maintained by one full-time member of staff, but even so keeping the 36,000 subject headings and 2000 genre/form terms up-to-date can prove a challenge.

Topic Modelling with BERTopic

Topic modelling has already been introduced in this chapter. It is a text mining technique used to determine which topics are represented in a collection of documents without manually reading them. The most popular algorithm used has been Latent Dirichlet Allocation (LDA) where topics are represented as a set of words with different probabilities, and documents are understood as being generated by a distribution of different topics (Blei 2012). When a topic model is fitted to a collection of texts, it denotes the words relevant for a given topic and indicates the documents in the collection most representative of that topic.

BERTopic is a modelling technique based on clustering document embeddings generated by a transformer model. "BERTopic generates document embedding with pre-trained transformer-based language models, clusters these embeddings, and finally, generates topic representations with the class-based TF-IDF procedure" (Grootendorst 2022). It therefore leverages the language understanding of the language model to find similar documents and build topics.

The first step when building a model with BERTopic is to run a SentenceBERT model on the corpus to obtain embeddings for each document. The BERTopic library allows different kinds of backend models to generate representations for clustering, but SentenceBERT is the recommended method since it is a transformer model that can embed entire documents at once. The project used a Swedish Sentence-BERT trained in-house at KBLab and available through Huggingface (Rekathati 2023).

Once the document embeddings are ready, their dimensionality must be reduced for the clustering algorithm to work properly. BERTopic uses a dimensionality-reducing algorithm called UMAP that preserves both global and local information to an adequate degree. The representations of the documents are then clustered using an algorithm called hierarchical density-based spatial clustering of applications with noise (HDBSCAN) which is a density-based clustering method that is good at detecting outliers and is particularly suited to noisy data like natural language text.

The result of the clustering algorithm is a number of clusters that correspond to topics, but they are still virtually uninterpretable at this stage. The next step is to use an adapted version of term frequency, inverse document frequency (TF-IDF), to extract relevant words from each cluster and label the topics. For the purpose of class-based TF-IDF, all documents in the cluster are treated as a single document, and then TF-IDF is applied to the whole text to find words that are important for that cluster. A ranking algorithm called maximal marginal relevance (MMR) can then be applied to diversify the words that represent each topic and make sure that they form a coherent but non-overlapping representation.

Topic modelling is regarded as a tool to gain general insights about a corpus to be followed by a deeper dive as required. It previously involved many modelling choices which can have a significant impact on the end results. Documents must be pre-processed to maximise the information content of the words to be modelled with a range of parameters in place to control the modelling behaviour. BERTopic greatly simplifies both aspects of the process. Since document representations are generated through a pre-trained language model, the need for pre-processing is greatly reduced. Transformer models leverage transfer learning to output high-quality representations of text data, in this case accepting whole paragraphs as input. They take into account both semantics and syntax while generating representations, whereas regular topic models usually operate on a bag-of-words basis. The traditional components of an NLP pipeline like tokenising, part-of-speech tagging and lemmatising (Tech Gumptions 2023) are not required, as single words are not used as features in this kind of modelling.

In terms of parameter settings, BERTopic comes with a number of default parameters that are automatically optimised and yield consistent modelling outcomes. For instance, one of the main challenges in traditional topic models is determining the number of topics. This has to be done manually and is usually based on some coherence measures and/or knowledge of the subject matter. BERTopic provides an automatic setting to let HDBSCAN determine the number of topics based on optimal distance between the clusters. The alpha and beta parameters to control topic word density and document topic density are absent in BERTopic, as the clustering is handled by HDBSCAN instead of latent Dirichlet allocation.

In summary, BERTopic offers a higher degree of automation compared to traditional topic modelling, while also allowing for some flexibility by letting the user override the default settings.

Embeddings

To understand the mechanics of the project approach, the role of a BERT model in enabling the automated clustering that underpins it requires further explanation. A class of models called sentence transformers, s-BERT, is used to convert text sequences into exact numerical vector representations which are referred to as embeddings. The term sentence is used loosely to refer to a sequence of text of arbitrary length, rather than its traditional linguistic meaning.

Using neural networks to compute continuous numerical vector representations of words was introduced in word2vec (Mikolov et al. 2013). Depending on configuration, the networks were challenged to predict either the current word given its surrounding context, or the surrounding context of words given the

current word. The idea and the expectation were that words frequently co-occurring within some context length N would also be semantically related to each other, which would manifest itself in the form of vector representations of similar words moving more closely together during training. The concept of word embeddings would later be extended to documents in another NLP tool doc2vec (Le and Mikolov 2014; Shperber 2017).

While being effective, word2vec produced only a static word embedding for each word. Static embeddings could not account for words having several different definitions nor for changes in their meaning based on surrounding context, such as the presence or absence of negations and other modifiers. Transformer models addressed the shortcomings of static embeddings through the use of another set of network layers with the purpose of contextualising the embeddings (Vaswani et al. 2017). In these models, word embeddings representing the same word start out the same in the input stage, but end up with different numeric representations depending on the influence of the surrounding context.

As already noted, BERT (Devlin et al. 2019) is a transformer architecture that incorporates both previous and future context in a sequence in computing output embeddings. Prior architectures focused either on machine translation or language modelling tasks (Radford et al. 2018), gaining a general language understanding through challenging the models to predict the next word in a sequence while only considering the context of previous words in the sequence. BERT modified the pre-training objective to a masked language modelling (MLM) task (Briggs 2021). As opposed to predicting the next word, its approach relied on trying to fill in and predict a proportion of words in the sequence that had been masked out while being able to consider the full context of the sequence. Considering the full context of a sentence, paragraph, or sequence when making a prediction proved an advantage on many downstream tasks such as extractive question answering and sentence similarity.

The manner in which BERT initially was fine-tuned and adapted for sentence similarity tasks meant a sequence pair would be passed to it as input and it would then be tasked with producing a similarity score for the pair. In the training setup, the two sequences would be passed to the network together, and contextual information from both sequences would be used in producing the output embeddings needed to determine the similarity score. While incorporating information from both sequences yielded strong results, it unfortunately suffered from an impractical inference procedure. Involving the neural network in the computation of every single similarity score between a sentence pair was computationally expensive.

Ideally, performing inference on a set of documents would use a model that outputs a meaningful document embedding which can be later used for similarity computations or clustering algorithms. Unfortunately, the original BERT train-

ing setup did not produce a representative sentence embedding, since during its training it always cross-encoded information from two sequences. Sentence-BERT (Reimers and Gurevych 2019) modified the training setup to pass sequences independently and separately to the network. By isolating the two inputs in a bi-encoder setup, the BERT model could learn to output meaningful sentence embeddings.

Practical Details of the Project

The techniques outlined above were used to create topics from the corpus. Most of the heavy lifting was carried out by the sentence-BERT language model in producing embeddings. Once the embeddings were produced, the standard algorithms already mentioned, UMAP and HDBSCAN, could be applied to the embeddings. The tools are further packaged in BERTopic making the process easy to experiment with through testing different parameters. The parameters included the minimum distance between topics to determine whether closely related topics should be merged or not. Since the aim was the exposure of the created topic system as linked data between topics rather than a balanced list of terms, having many similar topics was not a problem.

Sentence-BERT was used to make the model for the embeddings and BERTopic was used to fit the model to the data to find clusters of paragraphs. The resulting clusters became the topics. The language model does not operate on a word level, but uses embeddings which means that the model can handle synonyms, and even though a particular word might be chosen to represent the cluster it does not have to be present in the paragraphs. It is likewise important to note that each paragraph is not only clustered into a single topic, but also is distributed over all topics with an attached score. Paragraphs that deal with many themes receive a high score for multiple topics.

The Corpus

To create and evaluate topics, a corpus was extracted from 954 titles from the Swedish publishing foundation Natur och Kultur. Fiction and biographies were filtered out. The remaining titles comprised a corpus of roughly 65,000 paragraphs. Metadata from the Swedish union catalogue, Libris, was used for filtering and to find the attached subject headings from the existing system.

A SPARQL-query (Prud'hommeaux and Seaborne 2008) was used to obtain a list of title, ISBN, topics and genre/form for all monographic texts matching the

criteria. An overwhelming percentage had been catalogued with any of the three description levels: Full Level, Abbreviated Level, and Minimal Level, where Abbreviated Level does not require subjects or genre/form. Only controlled terms were used for the project and non-internationalised resource identifier (IRI) types were filtered out, which also excluded post-coordinated complex subjects. Corpus details are contained in Table 13.1.

Table 13.1: Corpus disposition

Titles	954
Paragraphs	65165
Subject headings from SAO	522

Project Results and Findings

The result with most parameters set to the defaults was a system with 1211 topics covering 50% of the paragraphs. The seemingly high number of unclassified paragraphs is in line with previous experiments. Given a particular text, the model provided output values or topic loadings that represent the distribution topics for that text. Since there is no gold standard for comparison, it is not feasible to evaluate the automatically created subject system using quantitative metrics such as precision and recall and the F1 score, the harmonic mean of the precision and recall. Instead, qualitative evaluation and proxy measurements were used. Topics were assigned per paragraph while subjects were assigned per work, which made the task of comparison more complicated.

Optimally, both systems would be evaluated against real-world use-cases with actual users. The effects of the systems could thereby be measured rather than examining similarities. However, this is unfortunately beyond the scope of this chapter. To create a sample for evaluation, topics were first chosen at random, and paragraphs then sampled from the topics. A small number of works was chosen. The resulting sample contained twenty-five topics, 125 paragraphs and ten works.

Representing Topics as Strings

It is difficult for human beings to recognise what a generated topic represents. It would be possible to read all of the paragraphs in a topic and then come up with a term, that is, a subject heading. But since the project's aim was sidestepping human

involvement to counter the problem of divergent judgement in existing systems, this strategy was not deployed. Another approach to the problem is to find words in each topic that are disproportionately common compared to the rest of the corpus. Traditionally this has been done using TF-IDF where the frequency of a term in relation to a topic is divided by the frequency of the term globally. Taking the top ten words for each topic and concatenating them creates a string that gives some idea of what the topic is about. It is important to note that the string is not a subject heading in that it does not necessarily represent the topic. The string provides a sense of what the topic encompasses, as shown in Table 13.2.

Table 13.2: Words selected using TF-IDF for sample clusters

Topic	#0	#34	#112	#710
Words	religious, god, religion, humanity, spiritual, christian, gods, faith	stage, tour, driver, armstrong, anquetil, kilometer, france, desgrange	lens, light, eye, mirror, figure, reflection, ray, rays	russian, rolf, women, men, woman, petersburg, russia, kalle, sune, romantic

As an initial, fail fast attempt to evaluate the results, the words chosen are examined for apparent affinity. At first sight, the results are inconclusive. While topics often have words that are clearly related, some are a mix of seemingly random words as shown in Table 13.2. However, on further inspection it becomes apparent that the seeming randomness is an effect of the method. In retrospect, the class of words should have been considered, since names, for example, are often disproportionately common in a text while not being connected to the actual topic.

Topic Cohesion

The degree to which a topic collects paragraphs that are similar indicates the extent of clarity of topic definition. If the paragraphs seem to have no connection to each other, the cohesion is low, whereas if the paragraphs have a clear common denominator, then the topic cohesion is high. From sampling and evaluating paragraphs, three distinct categories of topics were identified:

1. Topics that mapped well to existing subject headings such as religion or alcoholism (Table 13.3).
2. Topics that mapped to observable themes or aspects in the text, but where the work spanned multiple seemingly disparate subjects, for example client, parents, and time. These topics can show up in any work that has such a theme.

3. Topics for paragraphs that have a particular form and content such as introductory texts, instructions and acknowledgements. These paragraphs do not relate to the subject matter and can therefore show up in a work regardless of the subject (Table 13.4, especially topic #27).

An option for differentiating the three categories is to examine how closely they map to subjects using entropy (Stevens et al. 2012). Entropy is low when a topic maps cleanly to a single or a few subjects and high if it maps equally to many or all topics. If the goal is to find topics that map to subjects, then using an entropy threshold is effective. It was not appropriate for the project because it was not concerned with reifying existing subject terms.

After manual evaluation of the sample, topic cohesion in all three categories was deemed to be high. Even though topics might not cleanly map to subjects, there was often something observable that indicated why paragraphs had been clustered together.

Table 13.3: Topics with an apparent subject

#	Words	Comment
104	Drugs, metabolize, effect, side effect, liver, concentration, dosage, uses, oral, anesthesia	Paragraphs all deal with the effect of drugs
400	Algorithm, dataism, data, human, google, decision, data processing	Computer science
302	God, my, mine, me, father, gods, mercy, felt	Paragraphs in this topic contain multiple themes, but they all also deal with religion

Table 13. 4: Examples of topics that do not align with traditional subject headings or themes

#	Words	Comment
27	Thank, thank you, book, my, helped, editor, me, contribute	This cluster very precisely contains acknowledgements where the author thanks family members, and the editor
94	Psychologist, psychotherapist, university, psychology, clinical, dr, karolinska, professor, institute	Introductory texts and description of authors. For example, “Adam Smith is therapist at the Center for Mental Health, he has […]” is a good example of when the selected words do not give a good indication of the topic. The cluster shows high topic cohesion
165	Copy, cid, questions, students, read, whiteboard	Paragraphs contain task instructions for students

Mapping Topics to Subjects

The existence of the first category of topics, those with low entropy mapping to subjects, indicates that the method operates similarly to human creation of a system manually. If there had been no correlation the appropriateness of the method might be in question. Such correlation could be used to map generated topics for new works to existing subjects, in a similar vein to existing automated or semi-automated approaches to subject headings. While this aspect was neither within the scope nor the intention of the work with BERTopic, it can still be useful as an aid to cataloguers.

Aggregating Topics for Whole Works

Given the entire set of paragraphs for a single work, all topics can be aggregated to obtain a work-level topic distribution. The distribution is more stable than that from a single paragraph since any noise from paragraphs that do not deal with the topic(s) are fewer by comparison, assuming that one or more distinguishable topics exist in the work.

Mapping from topics to subject can be used to determine the aggregated subject heading. The approach is more effective than using a paragraphs distribution, because it deals with the topic of one paragraph only. For example, a book with the subject heading “chemistry” might contain a paragraph about physics which will result in a topic connected to physics even though the broader subject of the work is chemistry. However, when aggregating the topic distribution of all paragraphs, one single paragraph will not influence the total in any meaningful way (Figure 13.1). There is evidence that topics at the aggregate level can be used to propose, or automatically add, subject headings if required.

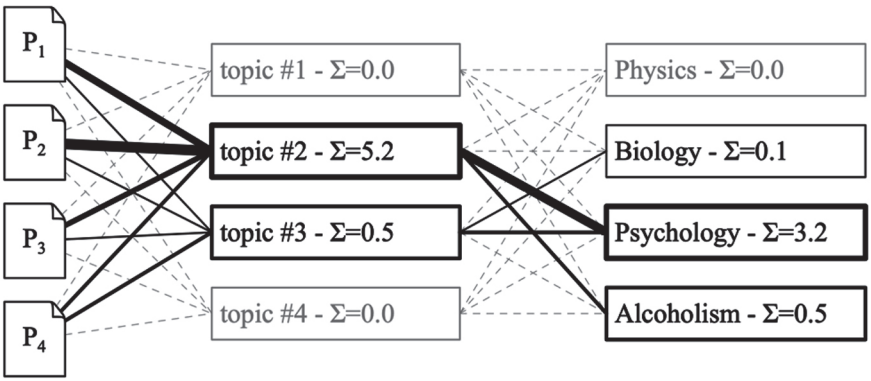

Figure 13.1: Example of aggregated topic distribution for multiple paragraphs and optional mapping back to subjects

Linked Data

Following the development of topics and mapping through paragraphs to works, both the topic system and its relationships can be easily exposed as linked data (Bizer et al. 2008). Because HDBSCAN is a hierarchical clustering algorithm, topics are grouped into increasingly broader topics until there is only one overarching topic. The process creates a tree of topics that can be exposed as linked data. Dublin core (DC) was used for links between topics and works and Simple Knowledge Organization System (SKOS) was also applied using sKOS:Concept for topics and skos:broader for links (Figure 13.2).

The resulting Resource Description Framework (RDF) data can be imported into Libris and examined further by users and cataloguers. This is possible since the infrastructure underpinning the Swedish union catalogue, Libris XL, is based on linked data (Wennerlund and Berggren 2017) facilitating data sharing among libraries and interfaces using Libris.

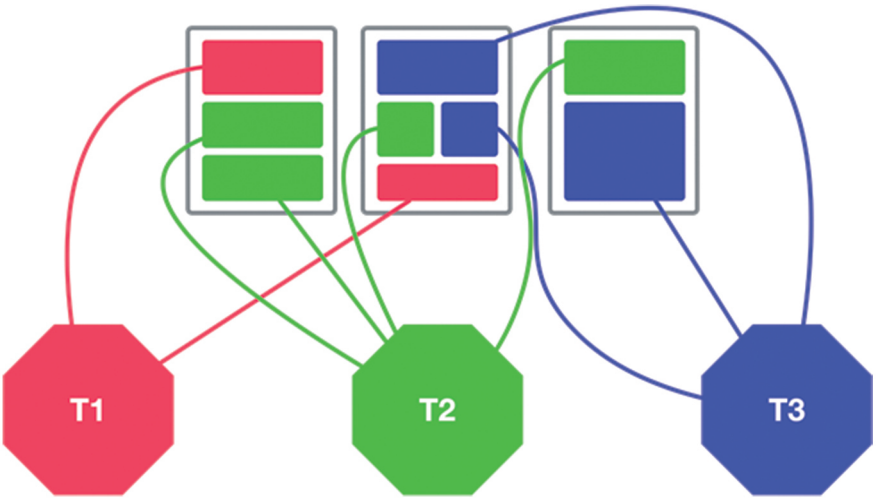

Figure 13.2: A visual representation of paragraphs and their connected topics, with the different colours showing the paragraphs where the respective topics can be found in the works

Other Findings

A surprising side-effect of evaluating the BERTopic approach was the insights into the workings and application of the existing subject headings system that it provided. Comparing topics to subjects in paragraphs led to questions about why a particular subject had been used. The answer could be custom and habit, or that the subject heading system had been co-opted for another purpose, making it a general-purpose solution for tagging content. An example is the use of the subject "Swedish as a second language" to tag resources that are suitable for readers with Swedish as a second language, rather than being about Swedish as a second language as might be expected. On the other hand, BERTopic would probably not have found the particular aspect or generated a topic for it. In this particular instance, it would be advantageous to target the specific intended use and analyse the text to determine its suitability for people with limited language understanding, rather than focus on the topic/subject system.

Conclusion

The project undertaken at KB and described in this chapter examined whether problems with library subject headings systems designed and operated by humans might be overcome by the use of machine learning. The work undertaken showed that the fully automated creation of a linked subject system using BERTopic is possible. It also suggested that such a topic modelling approach to indexing need not reproduce the headings of prevailing systems for knowledge organisation, with their inherent layers of bias. It is possible to develop a machine-learning based approach that enables users to search and navigate according to their interests without relying on fixed headings from a controlled vocabulary. Whether such a system is used to replace, extend or support existing practices will ultimately depend on the usage of the current subject heading system at any given institution.

During the evaluation of the topic modelling approach, multiple problems with existing approaches were discovered. Working through the data provided insights into the manual subject heading system. There was evidence that subject headings are sometimes seen as a golden hammer, an available tool used for purposes for which it was not designed. If all one has are subject headings, everything will look like a subject. The problem was highlighted with the example of the "Swedish as a second language" heading which had been liberally applied to refer to the style of content rather than its subject matter.

Another broad outcome of the project was the reinforcement of the importance of digitisation in libraries, not only as a means of preserving and representing physical objects and improving access. Digitising books facilitates the creation of metadata and subject analytical data created using the text itself. The findings in relation to the advantages of digital data point to a "digitise first and ask questions later" approach. Since the new system does not rely or expand on the current one, the experiment carried out in Sweden with BERTopic means that a valuable additional tool has been gained to examine both the current system and its application. Critical exploration using new methods and techniques helps to illuminate shortcomings of existing systems that might otherwise remain hidden.

Next Steps

The pilot project undertaken at KB investigated the practical and technical feasibility of creating an automated subject system without fixed headings using BERTopic. To build on the work and capitalise on the experience gained, there are three directions in which future efforts could usefully be directed:

- Incorporate a user experience (UX) perspective and determine how the system might work from a user perspective
- Test at scale by examining how the topic modelling approach might function with a larger corpus than considered in the pilot, and
- Conduct qualitative interpretation and assessment.

While the lack of benchmarks makes assessment of performance a challenging issue, further attention must be directed towards comparing the effectiveness and efficiency of manual and automated systems used to analyse the content of library collections.

Acknowledgements

The authors would like to thank the National Library of Sweden (KB) for providing metadata, the KBLab for providing resources, *Natur & Kultur* for providing the full texts for the corpus and Maarten Grootendorst for creating BERTopic. Part of the development work for this project was carried out within Huminfra, the Swedish national infrastructure for digital and experimental work in the humanities.

References

Antell, Karen, and Jie Huang. 2008. "Subject Searching Success: Transaction Logs, Patron Perceptions, and Implications for Library Instruction." *Reference & User Services Quarterly* 48, no. 1: 68–76. https://www.jstor.org/stable/20864994.

Asula, Marit, Jane Makke, Linda Freienthal, Hele-Andra Kuulmets, and Raul Sirel. 2021. "Kratt: Developing an Automatic Subject Indexing Tool for the National Library of Estonia." *Cataloging & Classification Quarterly* 59, no. 8: 775–93. https://doi.org/10.1080/01639374.2021.1998283.

Beghtol, Clare. 1986. "Bibliographic Classification Theory and Text Linguistics: Aboutness Analysis, Intertextuality and the Cognitive Act of Classifying Documents." *Journal of Documentation* 42, no. 2: 84–113. https://doi.org/10.1108/eb026788. Available at https://edisciplinas.usp.br/pluginfile.php/7880387/mod_resource/content/0/BEGHTOL_1986_Bibliographic%20classification%20theory%20and%20text%20linguistics....pdf.

Bizer, Christian, Tom Heath, and Tim Berners-Lee. 2011. "Linked Data: The Story so Far." In *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, edited by Amit Sheth, 205–27. Hershey, PA: IGI Global. Reprinted in *Linking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web*, edited by Oshani Seneviratne and James Hendler, 115–143. New York, NY: Association for Computing Machinery, 2023. https://doi.org/10.1145/3591366.3591378.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55, no. 4: 77–84. https://doi.org/10.1145/2133806.2133826.

Bloomfield, Masse. 2002. “Indexing – Neglected and Poorly Understood.” *Cataloging & Classification Quarterly* 33, no. 1: 63–75. https://doi.org/10.1300/J104v33n01_07.

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out : Classification and Its Consequences*. Cambridge, Mass.: MIT Press.

Börjeson, Love, Chris Haffenden, Martin Malmsten, Fredrik Klingwall, Emma Rende, Robin Kurtz, Faton Rekathati, Hillevi Hägglöf, and Justyna Sikora. 2023. “Transfiguring the Library as Digital Research Infrastructure: Making KBLab at the National Library of Sweden.” *SocArXiv Papers*. Accepted for publication in *College & Research Libraries.* December 8, 2023. Anticipated publication date 2025. https://osf.io/preprints/socarxiv/w48rf.

Briggs, James. 2021. “Masked-Language Modeling With BERT.” *Medium*, May 20, 2021. https://towards-datascience.com/masked-language-modelling-with-bert-7d49793e5d2c.

Chan, Lois Mai. 1990. *Library of Congress Subject Headings : Principles of Structure and Policies for Application*. Annotated version. Washington, D.C.: Cataloging Distribution Service, Library of Congress. Available at https://babel.hathitrust.org/cgi/pt?id=mdp.39015057586136&seq=9.

de Keyser, Pierre. 2012. *Indexing : From Thesauri to the Semantic Web.* Oxford: Chandos Publishing. Available at https://www.comecso.com/wp-content/uploads/2019/05/Indexing-from-thesauri-to-the-Semantic-Web.pdf.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), 4171–86. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Fano, Elena, and Chris Haffenden. 2022. ‘The KBLab Blog: BERTopic for Swedish: Topic Modeling Made Easier via KB-BERT’. https://kb-labb.github.io/posts/2022-06-14-bertopic/.

Finkel, Irving. 2019. “Assurbanipal’s Library: An Overview.” In *Libraries Before Alexandria: Ancient Near Eastern Traditions* edited by Kim Ryholt and Gojko Barjamovic, 367–89. Oxford: Oxford University Press.

Golub, Koraljka. 2021. “Automated Subject Indexing: An Overview.” *Cataloging & Classification Quarterly* 59, no. 8: 702–19. https://doi.org/10.1080/01639374.2021.2012311.

Grootendorst, Maarten. 2022. “BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure.” *arXiv*:2203. 05794. https://doi.org/10.48550/arXiv.2203.05794.

Hutchins, William J. 1978. “The Concept of ‘Aboutness’ in Subject Indexing.” *Aslib Proceedings* 30, no. 5: 172–181. https://doi.org/10.1108/eb050629. Available at https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ddaeb42ad5889edf6e3eba808133ee0066c4fea2

Knowlton, Steven A. 2005. “Three Decades Since Prejudices and Antipathies: A Study of Changes in the Library of Congress Subject Headings.” *Cataloging & Classification Quarterly* 40, no. 2 123–145. Available at https://steven-knowlton.scholar.princeton.edu/sites/g/files/toruqf3746/files/steven.a.knowlton/files/knowlton_three_decades.pdf.

Le, Quoc, and Tomas Mikolov. 2014. “Distributed Representations of Sentences and Documents.” *PMLR*: *Proceedings of Machine Learning Research* 32:1188–96. Proceedings of the 31st Machine Learning Research, Beijing, China, 22–24 Jun 2014. https://proceedings.mlr.press/v32/le14.html.

Malmsten, Martin. 2009. “Exposing Library Data as Linked Data.” Paper presented at the IFLA Satellite Preconference sponsored by the Information Technology Section. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=2e0791d88a65cb2517e284c2bfca02b7c6660f30.

Martin, Jennifer M. 2021. “Records, Responsibility, and Power: An Overview of Cataloguing Ethics.” *Cataloging & Classification Quarterly* 59, no. 2–3: 281–304. https://doi.org/10.1080/016393

74.2020.1871458. Available at https://api.mdsoar.org/server/api/core/bitstreams/c204d913-b186-46b6-8ecd-409ef215adc4/content.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv*:1301.3781 (cs.).Submitted January 16, 2013 (this version), latest version 7 Sep 2013 (v3)]. https://arxiv.org/abs/1301.3781.

Olson, Hope A. 1998. "Mapping Beyond Dewey's Boundaries: Constructing Classificatory Space for Marginalized Knowledge Domains." *Library Trends* 47, no. 2: 233–254. https://www.proquest.com/docview/220452238?sourcetype=Scholarly%20Journals. Available at https://core.ac.uk/download/pdf/4817546.pdf.

Prud'hommeaux, Eric, and Andy Seaborne. 2008. "SPARQL Query Language for RDF." *W3C Recommendation*, January 15, 2008. https://www.w3.org/TR/rdf-sparql-query/. New Version Available: SPARQL 1.1 (Document Status Update, 26 March 2013) https://www.w3.org/TR/sparql11-overview/.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. 'Improving Language Understanding by Generative Pre-Training'. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." https://arxiv.org/abs/1908.10084.

Rekathati, Faton. 2023. "Swedish Sentence Transformer 2.0." *The KBLab Blog*, January 16, 2023. https://kb-labb.github.io/posts/2023-01-16-sentence-transformer-20 /.

Shperber, Gidi. 2017. "A Gentle Introduction to Doc2Vec." *Medium*, July 26, 2017. https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e.

Stevens, Keith, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. "Exploring Topic Coherence over Many Models and Many Topics." In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea. 952–61. Association for Computational Linguistics, https://aclanthology.org/D12-1087/.

Suominen, Osma. 2019. "Annif: DIY Automated Subject Indexing Using Multiple Algorithms." *Cataloging & Classification Quarterly* 29, no. 1: 1–25. https://doi.org/10.18352/lq.10285.

Tech Gumptions. 2023. "Natural Language Processing (NLP) Pipeline." *Medium*, October 15, 2023. https://medium.com/@tech-gumptions/natural-language-processing-nlp-pipeline-e766d832a1e5.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 30. [Presented at] 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wennerlund, Bodil, and Anna Berggren. 2019. "Leaving Comfort Behind: A National Union Catalogue Transition to Linked Data." Paper presented at IFLA WLIC 2019, Athens, Greece. Libraries: dialogue for change in Session S15 - Big Data. In: Session S15 Data Intelligence in Libraries: The Actual and Artificial Perspectives, 22-23 August 2019, Frankfurt, Germany. https://library.ifla.org/id/eprint/2745/.

Yablo, Stephen. 2014. *Aboutness*. Princeton, NJ: Princeton University Press.