Making and Using AI in the Library: Creating a BERT Model at the National Library of Sweden

Chris Haffenden
Research Coordinator
KBLab, National Library of Sweden
chris.haffenden@kb.se

Elena Fano
Data Scientist
KBLab, National Library of Sweden
elena.fano@kb.se

Martin Malmsten
Head Data Scientist
KBLab, National Library of Sweden
martin.malmsten@kb.se

Love Börjeson
Director
KBLab, National Library of Sweden
love.borjeson@kb.se

Accepted for *College & Research Libraries*: December 17, 2021

Anticipated Publication Date: January 2023

Manuscript#: crl - 24880

Abstract
How can novel AI techniques be made and put to use in the library? Combining methods from data and library science, this article focuses on Natural Language Processing technologies in especially national libraries. It explains how the National Library of Sweden's collections enabled the development of a new BERT language model for Swedish. It also outlines specific use cases for the model in the context of academic libraries, detailing strategies for how such a model could make digital collections available for new forms of research: from automated classification to enhanced searchability and improved OCR cohesion. Highlighting the potential for cross-fertilizing AI with libraries, the conclusion suggests that while AI may transform the workings of the library, libraries can also have a key role to play in the future development of AI.

Keywords: AI implementation; NLP; language model; Swedish BERT; national libraries; lesser-resourced languages.

Introduction

Recent developments in machine learning can transform the working practices of the library. The advent of artificial neural networks offers tantalizing possibilities for libraries to be able to classify, organize and make huge digital collections searchable with the help of artificial intelligence (AI). To this end, various academic and national libraries have established data labs as testing sites to explore and harness such potential, with LC Labs at the US Library of Congress one prominent example. Yet remarkably little work has been published on this subject, either theoretically or in terms of practical examples. In contrast to many other fields where studies on the impact of AI have proliferated, a recent survey within information science could still point towards the general "absence of scholarly research on AI-related technologies in libraries".[1]

This article counters this gap by exploring the scope for making and using novel AI techniques in the setting of the library. More precisely, the focus is upon creating and implementing natural language processing (NLP) tools in the context of especially national libraries, with emphasis on the value of AI for medium- and low-resource languages—i.e. for libraries in countries beyond the linguistic resources of the Anglophone world and other major languages.[2] The particular NLP technology examined is the *language model*: e.g. a statistical model that through exposure to vast amounts of text can be used to understand and generate human language.[3] Our principal argument highlights the democratic effects that these libraries can contribute to AI development via such models, given their function as custodians of large volumes of language-specific data. AI may well have the promise to transform the workings of the library, as we will suggest, but libraries also have a potentially significant role to play in the future development of AI.

We consider how AI techniques can be made and put to use at the library via the example of a BERT language model (Bidirectional Encoder Representations from Transformers—elaborated on below) created at the National Library of Sweden (*Kungliga Biblioteket*, hereafter KB).[4] Methodologically, we seek to bridge AI and library science insofar as we write from the perspective of KBLab, KB's lab for data-driven research, where cutting-edge knowledge in data science combines with considerable experience of the library's information systems and working processes.

The first part of this article explains what a BERT model is and describes how we used KB's collections to train such a model for the Swedish language: KB-BERT.[5] The second part outlines specific use cases for a BERT model in academic libraries, detailing strategies for making digital collections available for new forms of research: from *automated classification* to *enhanced searchability* and *improved OCR cohesion*. In showing how the model could be employed to create novel research openings, these use cases suggest the value of AI to the operating practices of libraries more generally. We conclude the article with some broader reflections about the opportunities and risks connected to the cross-fertilizing of AI and libraries—a trend that we expect to grow in the future.

Literature Review

AI Applied, but Not Made, in the Library

There has been surprisingly little research published on the impact of AI techniques in the library. Yet certain exceptions exist that have started to consider how libraries might focus their attention on AI as a means of addressing the distinctive informational challenges posed by digitalization. Ryan Cordell recently offered a panoramic overview of the state of the field in "Machine Learning + Libraries", where he provided a general description of the current applications of machine learning in library settings—from

crowdsourcing and discoverability of collections to library administration and outreach.[6] A similarly broad view can be found in the work of Thomas Padilla and his colleagues in the "Collections as Data" movement, who have produced various reports that, while highlighting the value in applying AI in the library, emphasize the need for libraries to take a responsible approach that mitigates the potential harm of these emerging technologies.[7] There have also been more specific studies that have examined the infrastructural challenges for libraries in supporting data-driven research that seeks to analyse Big Data,[8] as well as the problems that the application of Optical Character Recognition (OCR) technology to historical material has created for both libraries and researchers.[9]

However, a notable characteristic of this body of scholarship is that it has focused upon the library principally as a target site for the application of AI. While understandable, such a focus also risks making libraries an unnecessarily passive agent in this process—as effectively the recipient of black boxed technologies that have been designed and made elsewhere. We wish to nuance the understanding of this relationship between AI and the library by exploring a case study in which novel AI techniques are actually made in the context of the library. Beyond providing a set of practical use cases that detail how a BERT model could be implemented to enhance the research potential of a library, we use this article to provide an account of how the library's collections enabled the production of this model in the first place.[10] We begin therefore with a brief introduction to BERT, framed in terms that are intended to be legible to a non-specialist.

Theoretical Context: Deep Learning and BERT Models

In the following section, we provide the theoretical and practical background to our work in developing KB-BERT at the National Library of Sweden. What is deep learning? What is a BERT model? What is required for a library to make such a language model, and why

bother? We address these questions to provide sufficient contextual knowledge to grasp what is at stake in our subsequent discussion of AI implementation in the setting of the library.

Deep Learning and Natural Language Processing

Deep learning is a subset of machine learning, which in turn is a subset of AI. The main intuition behind deep learning is that machines can learn from being exposed to large amounts of data using algorithms that to some extent resemble biological brains. These types of algorithms are called artificial neural networks.[11] Deep learning is extremely powerful compared to traditional machine learning methods but it requires larger datasets and more computational resources to reach good performance. These are two significant bottlenecks that can make the training of deep learning models a significant challenge for many teams and organizations.

An important milestone in deep learning research has been the appearance of transfer and self-supervised learning.[12] Traditional supervised machine learning techniques learn from labelled datasets where human annotators have marked the properties in the data that they want the model to learn. This is a very time-consuming process and few datasets exist that are large enough to allow deep learning models to reach their full potential. The innovative dimension of transfer learning is to divide the training into two steps: in the first, self-supervised training step, the model is shown a large amount of unlabeled data from which it can extract general patterns; while in the second step, the model is fine-tuned on smaller, annotated datasets to learn how to perform a specific task.

We can take an example from NLP to illustrate how this works in practice. During the pre-training stage, the model is shown a huge amount of natural language text and trained to predict a word given the context in which it occurs, or vice versa. In this way,

the model learns how words co-occur in that language and forms a representation of their meaning. Let's assume we want to train a model to predict whether a movie review is positive, negative or neutral. The number of stars can be considered as the label and the text of the reviews is the training data. We would take our model that we have previously trained on generic language data, and we would train it to specialize in sentiment analysis for movie reviews. The knowledge accumulated during pre-training would make the model much more effective at learning this classification task, since it already has a representation of how language in general works.[13]

Transformers and BERT

The most popular architecture for deep learning in NLP today is the Transformer.[14] The Transformer was originally proposed for machine translation but has since been applied to all kinds of tasks, from text classification to computer vision. Its main strength is a mechanism called "attention", which allows the model to focus on particular parts of a sentence when processing a specific word. For instance, given the sentence "The dog didn't want to play because it was too tired", processing "it" would see the attention mechanism focused upon "the dog" in order to make sense of the pronoun. Transformer models are also popular because of their architecture, which lends itself to efficient parallelization—i.e., the ability to carry out complex tasks simultaneously spread across several processors. This in turn allows researchers to train models that are larger than ever.

The release of the pre-trained Transformer-based model BERT in autumn 2018 marked a significant turning point in NLP research.[15] BERT stands for Bidirectional Encoder Representations from Transformers and applying this architecture to language processing has enabled state-of-the-art performance on many benchmark datasets. Evaluated according to the standard testing framework—GLUE, or General Language

Understanding Evaluation[16]—BERT achieved unprecedented scores on a series of NLP tasks, ranging from question answering (when the model is shown a paragraph of text and then posed a question based upon this) to causal reasoning (i.e. given a sentence, which among four choices is the most obvious continuation?)[17] In short, BERT broadened the horizons of possibility for what a language model could do.

The initial development of this model demanded considerable resources, both computationally and in terms of training data. BERT was trained by Google AI on a corpus of 3.3 billion words that was composed of books and the text of English Wikipedia. The researchers at Google who released the model explained that the training of a medium-sized BERT took 4 days on their specialized processing units called TPUs, which are optimized for machine learning applications.[18] This gives an idea of how much computing power is required to train one such model; it is certainly not something that can be done on an average laptop. However, what makes BERT so attractive from the perspective of AI implementation is that it is freely available for anyone to download and then fine-tune on their own data. As a powerful general-purpose model, it can be adapted to apply cutting-edge language processing to specific use cases at a local level.

The Need for a Swedish BERT

The design and distribution of huge language models such as BERT reflects global hierarchies of power and resources. Whereas Google AI developed dedicated BERTs for English and Chinese, they released a multilingual model for the rest of the world that was trained on Wikipedia articles from 104 different languages: M-BERT. While achieving fairly good performance on many NLP tasks, researchers knew that specialized monolingual models would be able to outperform M-BERT. This led many institutions and universities around the world to train new BERTs for their particular language of interest.[19] Soon most of the major languages like German, French, Spanish, Korean,

Japanese, and Dutch had their own models, the only limitations being the availability of sufficient text data and computing power to produce the model. It was in this context of an expanding array of monolingual models that KBLab at the National Library of Sweden decided to train and publish a BERT for Swedish.

The first dedicated Swedish BERT had already been released by AF-AI, the AI lab at the Swedish Public Employment Agency.[20] AF-AI trained a BERT model using the data from Swedish Wikipedia that consists of about 300 million words, which is just a fraction of the size of the corpus used to train the original English BERT. The developers at AF-AI state that their model was intended as a temporary solution to fill a gap for the Swedish NLP community, while more substantial and better models were in the making.[21] KBLab saw an opportunity to contribute by training a Swedish BERT on a larger and more varied dataset that would enhance performance and produce a model with more robust language understanding.

This undertaking was enabled by KBLab's unique access to otherwise unavailable material. Legal deposit requirements dictate that every publication issued in Swedish must be submitted to KB so a copy can be preserved as a part of future culture heritage. This also applies to digital material since the introduction of legislation for electronic publications from 2015, which means that the library receives an enormous amount of Swedish text every year.[22] The library's collections thus encompass a diverse collection of text genres, ranging from newspapers, magazines and books to scientific journals and governmental reports. Although far from all the physical collections have been digitized, enough exist in digital form to create huge datasets that are orders of magnitude larger than any publicly available, curated collections of Swedish text like Wikipedia. It is the holding of such rich bodies of linguistic material that gives national libraries like KB a key role in the future training and creation of new language models.

Making a BERT from the Library's Collections

To make KB-BERT, we took advantage of the extensive textual resources in the library's digital archives. The model was trained on a corpus of about 3.5 billion words, which is almost exactly the size used by Google AI to train the original English BERT (meaning that KB-BERT could be expected to reach comparable performance levels). Our aim in assembling this specific corpus was to produce a body of text that could be described as being, to a degree, *representative* of the living language of the national community.[23] Here we can point to the distinctive advantages of smaller languages for achieving such data representativity for NLP development, given that KB's collections contain something close to population data for Swedish, whereas this is practically impossible for larger languages like English and Chinese.

We made significant use of the library's newspaper holdings to compile this selection of modern Swedish, extracting over 16 GB of (cleaned) text from OCR'd newspapers in the library's archives for the period ca. 1945-2019. This was supplemented by material derived from Governmental Reports, e-books, social media, and Swedish Wikipedia, with the incorporation of text from a broad range of social domains as a conscious choice to expand the representativeness of the language within the training corpus.[24] The diversity of the social voices represented and the breadth of language usage was strengthened by the presence of quotes and reported speech from a wide variety of actors in the newspaper material, as well as the innovative new forms of Swedish provided by the 31 million words from social media in our corpus.

The variety of registers and styles in the training material was paramount to enhancing the model's performance. During training, BERT is to some extent "learning" the language by trying to extract patterns from the noise. Being exposed to many different types of writing allows the model to identify the underlying principles of syntax and

semantics that are common to all instances of correct (and also slightly incorrect, but intelligible) modern Swedish text. Our central hypothesis was that by training KB-BERT on a broad spectrum of texts, we would be able to produce a model with more flexible and sophisticated understanding of natural language.

That we could test this hypothesis in practice depended upon direct access to KB's collections. While Swedish Wikipedia and social media text, though rather difficult to clean, are publicly available for anyone through different APIs, much of the data used to train KB-BERT is copyrighted and cannot be made available to the scientific community except under very strictly controlled circumstances. Technical security concerns together with legal restrictions of copyright and EU data protection (GDPR), mean it is not currently possible for the digital materials in KB's archives to be shared with researchers outside of KB's internal network. These circumstances give KBLab a strategic, if contingent, role in the future development of NLP resources for Swedish: if language models are not created in-house at KB then the data currently available externally would lead to models of lesser quality and more limited capacities for Swedish AI in general.

Evaluating Performance

Evaluating the quality of a new language model demands comparing its performance with that of existing models. This rests upon the application of a fair and objective means of comparison, which is a challenge for smaller languages like Swedish, since the standard testing benchmarks that exist for NLP in English—i.e., GLUE/ SuperGLUE[25]—are still in the process of being developed. Given the absence of common evaluative frameworks, testing instead involves fine-tuning a model for a so-called downstream task—e.g., a particular NLP task that one might want to use the model for in the future—where results can then be compared with the performance of other models. In the case of KB-BERT, one of the principal evaluation tasks we selected was that of named entity recognition

(NER), which tests the capacity to extract predefined types of named entities such as persons, places, and organizations from a given text.

In order to fine-tune a language model for NER, a dataset annotated with named entities must be available, which is not entirely the case for Swedish. The Stockholm-Umeå Corpus version 3.0— SUC 3.0[26]—has been manually annotated with various NLP characteristics such as part-of-speech tags, morphological analysis and lemma, but the named entities have been automatically tagged using a tool called Sparv.[27] This lack of human annotation means that the dataset cannot be considered a gold standard and that any performance measure should be taken with a pinch of salt. However, despite its shortcomings, SUC 3.0 is still the best NER dataset currently available for Swedish and we used it to evaluate the downstream performance of KB-BERT in relation to other BERTs. While evaluation based on a substandard dataset cannot give us an indication of the absolute performance of any BERT for the task at hand, it does enable reliable comparison between different models that have been evaluated using the same dataset.

To conduct such an evaluation for KB-BERT, we tested our model in relation to the BERT model previously released by the Swedish Public Employment Agency, AF-AI, and Google AI's multilingual model, M-BERT. This evaluation process used standard testing praxis from the field of NLP, whereby the SUC 3.0 dataset was divided into training (70%), development (10%) and testing (20%) subsets.[28] Each model was then fine-tuned using the same training data, before being exposed to a series of NER tasks from the test set. We chose the eight particular categories of named entity for these tasks (person, organization, location, time, measure, work of art, event and object) since these have been annotated in the SUC dataset. The table below indicates how each model performed, with the scores between 0 and 1 signifying the accuracy of the model's predictions of named entities in the test data.

Type of named entity	AF-AI	M-BERT	KB-BERT
Person	0.913	0.945	0.961
Organization	0.780	0.834	0.884
Location	0.913	0.942	0.958
Time	0.655	0.888	0.906
Measure	0.828	0.853	0.890
Work of art	0.596	0.631	0.720
Event	0.716	0.792	0.834
Object	0.710	0.761	0.770
Total average	0.876	0.906	0.927

Table 1: Comparison of Swedish BERT models on NER using SUC 3.0 dataset

These results show that KB-BERT consistently outperforms the other models on this particular evaluation task. More precisely, the model scored 5% higher than the AF-AI BERT trained solely on data from Swedish Wikipedia, and 2% higher than Google's M-BERT that was trained on Wikipedia data from over 100 different languages.[29] KB-BERT performed better on all types of named entities because it was trained on a larger volume of high-quality and more varied data, and thus has a better level of language understanding. This also helps explain the varying performances of the models across these different categories: entities which appear more frequently, and are fairly consistent in both linguistic form and the context in which they are used, are easier to recognize, which is why all three models tend to perform well on "Persons"; yet a more diffuse entity

such as "Work of Art", which is a larger umbrella category with more variable forms, proved harder to recognize. Again, that KB-BERT still performed comparatively better on identifying these more challenging entities is a result of the quality and quantity of KB's Swedish data.

By utilizing the library's textual resources, we were therefore able to create a more powerful and effective model than those trained solely on freely available data. On the one hand, this demonstrates the tautology within NLP regarding the centrality of data volume for improved language understanding: the more data used in training, the better the model performs. On the other hand, it supports our suggestion that using a diverse range of language from a number of social domains would enhance the model's performance; through being trained on a richer corpus of modern Swedish than simply Wikipedia alone, KB-BERT has achieved a more advanced level of understanding. More broadly, this evaluation also highlights the importance of high quality and varied language-specific training data in producing state-of-the-art language models like BERT.

Applying BERT at the Library

The creation of a high-performing general-purpose model for Swedish has opened up many possibilities for the application of NLP techniques. Since being released in February 2020, KB-BERT has been implemented by various organizations and public authorities in Sweden who have been able to take advantage of its capacities for rapidly processing large amounts of language data.[30] Insofar as this implementation contributes to the adoption of more effective procedures for public sector administration, it highlights the broader social benefits that can result from the development of these language models.[31]

But beyond such wider value, what does KB-BERT mean for the library itself? How might this be applied to KB's internal practices, given the centrality of information

management to the library as a cultural heritage institution and an infrastructure for academic research? In the following section, we explain three particular use scenarios where the model could be used to enhance the library's digital collections and the ways that users interact with these to pursue new forms of research. For each specific use case, we provide an overview sketching the area that BERT might be directed towards and how this would work in practice, as well as detailing the advantages and challenges with these applications.

Text Classification

What?

The first scenario involves applying KB-BERT to assist in **classifying incoming digital material** to the library. This is a matter of exploiting one of the principal capacities of BERT: *scalability*, or the ability to quickly and effectively deal with large volumes of text data. By fine-tuning the model to recognize particular categories, it becomes possible to organize—and make more readily available—types of material that have previously remained relatively hidden from the library's users, such as advertising ephemera, which has not conventionally been catalogued at the level of the individual item.

How?

Training BERT to distinguish between different categories requires the creation of an annotated dataset of the specific material that is to be classified. It presumes, in other words, the existence of a series of appropriately categorized examples that can be given to the model to learn from. Just how many examples are necessary depends upon the complexity of the task, but at least a few thousand examples will generally be required. The first step for fine-tuning BERT to classify in the library is therefore to recruit a team of annotators to produce such a set of examples.[32] This could be organized via crowd-

sourcing volunteers among library staff or users at large, the key point is to generate a sufficient and representative range of material that has been tagged.

The next step, once the annotated data has been produced, is using it to teach BERT how to distinguish between different types of material—if, for instance, an advertisement should be classed as relating to, say, sport, technology, or some other category. Here the learning process is made more effective by a machine learning technique called "bootstrapping", whereby the model requires less and less human correction for each round of fine-tuning that is undertaken.[33] The reason a Transformer model is so adept at such training is that text categorization is one of the tasks that it excels at: being exposed to large volumes of data and then asked to categorize based on the contents of documents is one of the principal tasks BERT was created for.

Once the model has achieved sufficiently high-performance levels, it is ready to be applied to classification in practice. How this should be implemented in concrete terms is an open question but given the nature of the task it is conceivable that some form of "human in the loop" implementation would be an appropriate starting point—i.e., a set-up that saw library specialists working in tandem with data scientists, and the language model, to help ensure optimal results. This has the advantage of allowing the model to be refined based on the expertise of the library staff, while at the same time allowing the staff to learn more about—and to trust—the workings of the model.

Figure 1: Making collections of commercial ephemera available for innovative new research (Photo: Ann-Sofie Persson/KB)

Challenges?

While custom-made tools like Prodigy enable the crowdsourcing of annotation,[34] there are still certain challenges with producing the initial annotated dataset needed for this implementation of BERT. On the one hand, it is a process that requires a fair amount of labour to produce sufficient examples, which means having a group of volunteers prepared to undertake the repetitive and (for some) boring task of annotating the data. On the other hand, the cross-checking and approval of completed annotations can be a tricky task in itself, since categories are often fluid and ambivalent—with boundary cases demanding the creation of further distinctions and sub-categories that need to be adjusted iteratively. Given the problem of the "subjectivity of [individual] judgements", having a project leader to oversee and standardize the verification of the tags is a prudent measure.[35]

Why?

Using BERT to classify incoming material is a smart way of making the library's collections more accessible for new forms of research. By training the model to categorize those parts of the legal deposit collections not currently classified as individual objects—i.e., ephemera—this AI implementation would make material that otherwise risks disappearing into the archive more visible and searchable for the library's users. That such classification also provides the material with a structure amenable to machine learning means researchers would be able to apply innovative data-driven approaches in using it.

Enhanced Searchability

What?

The second use case is using KB-BERT to **enrich metadata** for the library's digital collections. A central challenge facing libraries today is how to make their holdings amenable to the type of highly specific, granular search enquiries that users—especially academic researchers—have come to expect from the experience of using the internet and search engines like Google. Of course, a necessary precondition for such searchability is the digitization of the material (born digital collections aside); but the next step demands improved metadata, which is where a BERT model comes in. This is a more complex application than that of classification considered above, but it is also one with greater potential gains for the library's users, and thus for research at large.

How?

The principal NLP technique that BERT performs here is named entity recognition (NER), which —as mentioned previously—is the ability to extract entities such as places and names from texts. To maximize the searchability gains made possible by the model in this regard, there are various components that need to be put in place. This involves

integrating BERT with a number of other tools from data and information science, as we explain below. Here it is worth noting that since NER is a broader, language-wide task, there are more likely to be external resources that can be adapted and used, as opposed to internal library classification where less help is available.

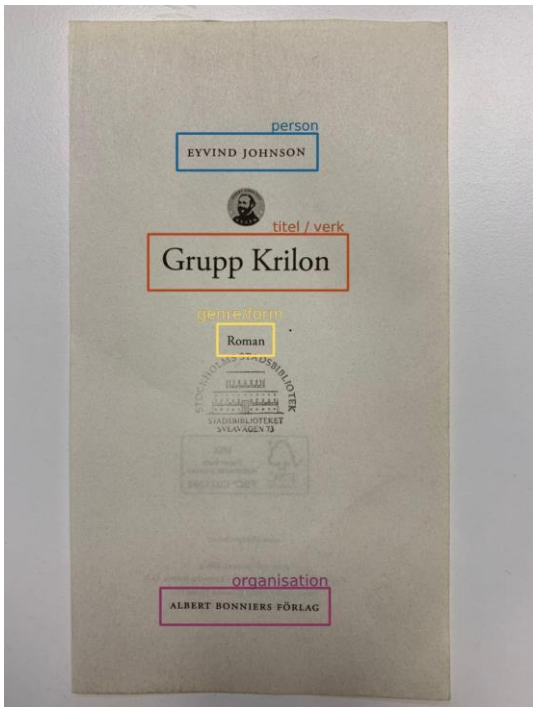
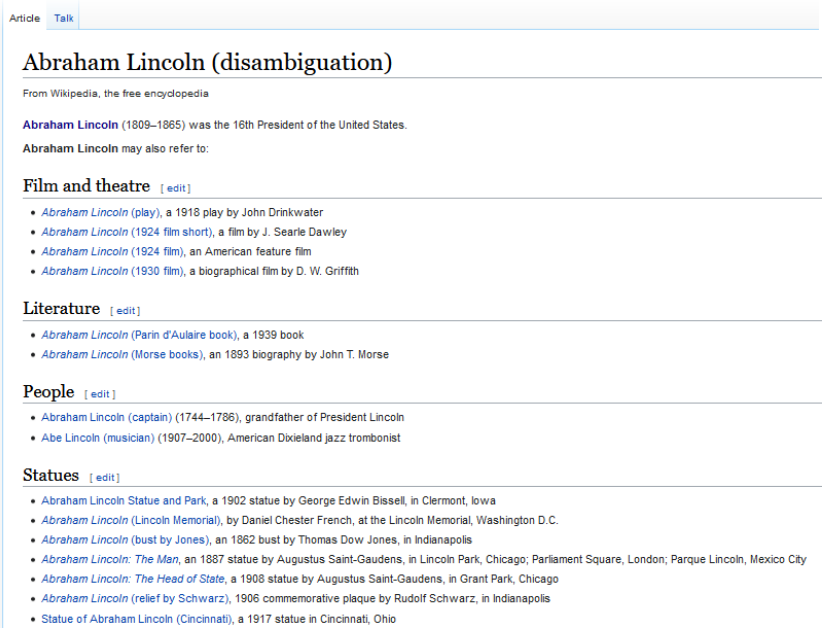


Figure 2: Using BERT to identify named entities in the library's collections (Image: Martin Malmsten/ KB)

Firstly, BERT needs to be fine-tuned to recognize the particular types of entities likely to be of interest to library users when they are searching the collections. The standard types that NER models tend to be trained upon are persons, places, organizations, times and measurement, but this could be expanded for a library ecosystem to include even titles of publications, literary genres and cultural movements. As soon as the particular categories of interest have been decided upon, a set of training data needs to be produced to allow BERT the chance to learn how to identify these entities in large volumes of text. As with the previous example of classification, this requires the production of an annotated dataset compiled of appropriately tagged examples. Beyond its use in training this model, the creation of such a NER dataset is also a significant

contribution to the national NLP community, insofar as it can then be released for other actors to use and develop in future research and development.

Secondly, once trained to find these entities in the library's collections, BERT needs to be connected to a specific model for named entity disambiguation (NED). This means a system with the ability to distinguish between entities with the same name that refer to different things.[36] If, for instance, we were to type "Abraham Lincoln" in Wikipedia's search window, then we arrive at a "disambiguation page" that lists the various different entities this could refer to: from the 16th president of the USA in person to the many works representing his life, and from a list of commemorative statues to other usages within transport (models of trains, etc.) Or if we search for "Lars Andersson" on Swedish Wikipedia, then we receive a similar result: a long list distinguishing the different historical and contemporary figures that share this name (see images below). In practice, there are various ready-made tools that can be adapted for this task of disambiguating different uses of the same name—e.g., Bootleg[37]—but doing so is an essential part of creating an effective search system built upon NER.



Article Talk

Abraham Lincoln (disambiguation)

From Wikipedia, the free encyclopedia

Abraham Lincoln (1809–1865) was the 16th President of the United States.

Abraham Lincoln may also refer to:

Film and theatre [edit]

- *Abraham Lincoln* (play), a 1918 play by John Drinkwater
- *Abraham Lincoln* (1924 film short), a film by J. Searle Dawley
- *Abraham Lincoln* (1924 film), an American feature film
- *Abraham Lincoln* (1930 film), a biographical film by D. W. Griffith

Literature [edit]

- *Abraham Lincoln* (Parin d'Aulaire book), a 1939 book
- *Abraham Lincoln* (Morse books), an 1893 biography by John T. Morse

People [edit]

- Abraham Lincoln (captain) (1744–1786), grandfather of President Lincoln
- Abe Lincoln (musician) (1907–2000), American Dixieland jazz trombonist

Statues [edit]

- Abraham Lincoln Statue and Park, a 1902 statue by George Edwin Bissell, in Clermont, Iowa
- *Abraham Lincoln* (Lincoln Memorial), by Daniel Chester French, at the Lincoln Memorial, Washington D.C.
- *Abraham Lincoln* (bust by Jones), an 1862 bust by Thomas Dow Jones, in Indianapolis
- *Abraham Lincoln: The Man*, an 1887 statue by Augustus Saint-Gaudens, in Lincoln Park, Chicago; Parliament Square, London; Parque Lincoln, Mexico City
- *Abraham Lincoln: The Head of State*, a 1908 statue by Augustus Saint-Gaudens, in Grant Park, Chicago
- *Abraham Lincoln* (relief by Schwarz), 1906 commemorative plaque by Rudolf Schwarz, in Indianapolis
- Statue of Abraham Lincoln (Cincinnati), a 1917 statue in Cincinnati, Ohio

Artikel Diskussion

Lars Andersson

Lars Andersson kan syfta på:

- Lars Andersson Rålamb (1563–1599), fogde
- Lars Andersson (talman), bondeståndets talman år 1734: se listan över bondeståndets talmän
- Lars Andersson i Halmstad (1815–1891), politiker
- Lars Andersson i Utterud (1823–1873), politiker
- Lars Andersson i Nora (1824–1880), politiker
- Lars Anderson i Landa (1831–1913), även kallad *Andersson i Ölme*, politiker
- Lars Andersson i Hedensbyn (1888–1974), politiker i bondeförbundet
- Lars Andersson (konstnär) (1910-2005)
- Lars Andersson (fotbollsspelare) (1927–1992)
- Lars Andersson (militär) (1936–2012)
- Lars Andersson (född 1941), svensk socialdemokratisk riksdagsledamot
- Lars Andersson (professor) (född 1947)
- Lars Andersson (kanotist) (född 1948)
- Lars Andersson (arkeolog) (född 1961)
- Lars Andersson (författare) (född 1954)
- Lars Andersson (ishockeyspelare), (född 1954)
- Lasse Anderson (låtskrivare född 1958), son till Stikkan Anderson
- Lars Andersson (nydemokrat) (född 1959)
- Lars M. Andersson, historiker (född 1961)
- Lasse Andersson (låtskrivare född 1963)
- Lars Andersson (sverigedemokrat), riksdagsledamot (född 1964)

Figures 3 and 4: Entity disambiguation; the diverse referents of "Abraham Lincoln" and "Lars Andersson" (Wikipedia CC BY-SA 3.0)

Thirdly, to produce an end-to-end system for entity recognition that maximizes search potential, BERT should be integrated with a tool for entity linking. This is a matter of connecting the library's internal databases with the wider knowledge base of the semantic web to ensure that the named entities identified by BERT within the collections are interconnected with—and made available as—open linked data.[38] What this linking achieves is the mapping of entities mentioned within the library's collections onto the existing network of information about these items contained in the vast structured data of, say, Wikidata—the database underpinning Wikipedia and related projects. By harnessing the power of BERT to identify entities on a massive scale, and by making this part of the huge informational resources of the semantic web, it becomes possible for users to search the collections with an entirely new level of scope and precision.

To show how this works in practice, we can consider the example of a researcher with a broad set of search parameters—say an interest in murders in West Sweden

between 1850 and 1930. A search system built upon BERT could perform such an enquiry across a large range of material—i.e., all the newspapers from the period—using the contextual knowledge of named entities the model has gained. Given that KB-BERT knows that the city of Gothenburg is situated in West Sweden, for instance, it would be able to include items for the above search terms without the reports explicitly mentioning this wider geographical term. The model thereby provides faster and more effective ways of helping researchers find what they are looking for.

Challenges?

Beyond the difficulties of producing an annotated dataset outlined in the previous example, this use case involves a number of fairly significant challenges. Firstly, developing and operating this NER function using BERT demands substantial computational resources—without sufficient processing power it will simply prove impossible. Secondly, it is demanding in terms of technical expertise: producing an integrated system for entity linking presumes the presence of an in-house team of data scientists to oversee such an implementation. Thirdly, and perhaps most importantly, it requires these experts to manage the integration of the library's internal databases with the open linked data of the web, which is a far from trivial task. In short, this is a complex and resource-intensive form of implementation, but one that offers transformative gains.

Why?

Using BERT as the foundation of an integrated system for named entity recognition and linking enables the library's digital collections to be searched in new, more expansive ways. Whereas conventional searches reveal the presence of named entities only insofar as they are present in traditional metadata—i.e., title, author, date of publication—a BERT model also has the capacity to search among the *contents* of the material in locating such entities. This allows a user to gain an insight of the collections of a vastly different

depth than that which was previously possible. In this context of enhanced searchability, it is no coincidence that models like BERT are already deployed for the online searches we take for granted when using Google. Applying such a technology to the library promises a means of enhancing how researchers are able to interact with, and use, the collections.

Improved OCR Cohesion

What?

The third example for implementing KB-BERT at the library also involves enriching metadata but relates more specifically to **improving the cohesion of digitized collections** to make them more accessible for users, particularly academic researchers. This is especially pertinent for historical material such as newspapers, which have been previously digitized but have lost various contextual markers that we take for granted, as human readers—i.e., the presence of specific articles and sections within the newspaper, rather than simply a collection of blocks of text.[39]

How?

The principal problem that needs to be resolved when trying to reconstruct the structure of a newspaper is identifying where a given article *starts* and *stops*. This is the case since the OCR process used to digitize the material effectively strips the text of such markers in breaking it down into smaller segments (see image below). If a language model like BERT can be taught to recognize the beginning and end of each article, it becomes possible to piece together the newspaper, article by article.

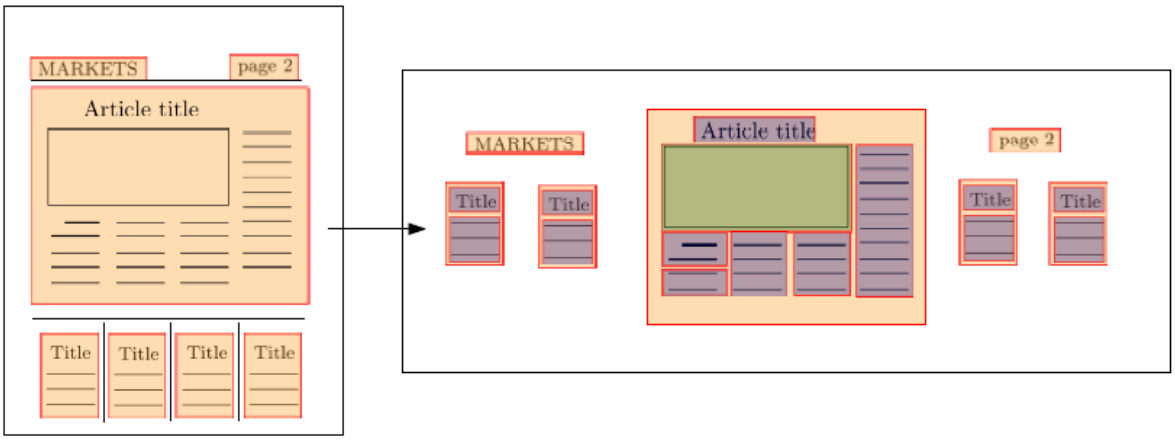


Figure 5: From human-readable newspaper to a set of text blocks.
Using BERT for OCR reconstruction (Image: Faton Rekathati/KB)

There are several different options for how BERT could be used to address this task, depending on the particular circumstances of the library. One alternative is to use a function integral to the model, *next sentence prediction*, to reassemble the OCR boxes into article form. This would work, in part, by BERT being trained to recognize the characteristic linguistic features of the opening and closing sections of an article, and partly through it comparing the likelihood of connecting sentences among adjacent text blocks to establish the correct order of the article's main body.[40]

The advantage of such an approach is that predicting next sentences was one of the original training tasks used to create BERT, meaning that no further fine-tuning is necessary to realize this capacity. However, it presumes the existence of a dataset of coherent newspaper articles that can be used to train BERT to identify how these typically start and finish. If the library is in possession of such data, as is the case at KB through the collection of Swedish newspapers received via electronic legal deposit, then this training process is simple, since it is a form of self-supervised learning where, given a few thousand examples, BERT will be able to teach itself the characteristics of a beginning and an ending. If the library does *not* possess such a material, on the other

hand, this would instead necessitate the manual collection of a set of examples, which is both more time-consuming and resource-intensive.

An alternative approach, and one to consider in the absence of such a readily available training set, is to combine BERT's capacity for language processing with a model for image recognition and use them in tandem.[41] This would involve training this latter model to recognize the visual clues that signify the beginning and end of a particular article—for instance, the presence of a title at the start or of, say, a square at the end. BERT's ability to predict next sentences could then be deployed to link together the sequence of text blocks between the first and the last block of each article. While undoubtedly more complicated, insofar as this involves the use of two different models, such an option has the advantage of using both textual and visual information in the data to reconstruct articles.

Challenges?

In addition to the problems of producing annotated data that might arise here, a number of more specific challenges connected to the particular nature of this implementation also exist. Firstly, treating a material that spans a broad period of time can be complex: if the newspapers to be processed range back over a longer period, there might be different style and visual conventions contained within the material. This, in turn, could mean that the model needs to be further fine-tuned to adapt to such shifts over time. Secondly, there are likely to be a greater number of OCR errors in the digital copies of older newspapers, which can likewise complicate the process of using BERT to understand and make predictions based on the contents of this material.[42] Thirdly, this is an implementation that demands a degree of technical expertise: it is not a straightforward fix, but rather an iterative process that will require systematically testing various parameters to ensure the

best possible results. This presumes, once again, the presence of personnel with a background in data science with experience in working with such questions.

Why?

Taking advantage of BERT's language understanding to reconstruct historical newspapers is another instance of using AI to enhance the accessibility of the library's material. By improving the OCR cohesion of the newspaper archive, this implementation adds a level of metadata to digital collections that is key to unveiling their value for research purposes. Once this structure has been re-established, it becomes possible to navigate the digital archive at the article level—to identify and search all the articles written by a particular author, for instance, which is a far from trivial gain from the users' perspective. As with the previous two use cases, such an application could significantly improve the quality and effectiveness of the library as a research infrastructure.

Concluding Discussion

In this article, we have presented how the textual resources of the National Library of Sweden provided the basis for a powerful new BERT model that outperforms existing models for Swedish. We have also explained three potential use cases for KB-BERT to highlight the relevance of such NLP techniques for the operating practices of the library. More precisely, we showed how the model could be applied to improving access to collections for researchers, by (i) providing an automated form of classification, (ii) enhancing the searchability and (iii) improving the OCR cohesion of digital collections. In each case, we suggest KB-BERT's language processing capacities can be harnessed to add clear value to the library's working processes.

Insofar as we have discussed both how the library can contribute to the future development of AI, and how AI could help transform the future of the library, the article raises broader questions about the opportunities and challenges for cross-fertilizing

libraries with AI. In this conclusion, we therefore delineate the outlines of these questions and explore some of their implications. What are the principal gains of closer interaction between national libraries and AI? What are the potential difficulties, on the other hand, of integrating the AI-related insights of data science with the information practices of the library?

National Libraries as Sites for Ethical AI

The key rationale for locating AI development in the context of national libraries is *democratic* in emphasis. The first part of this argument is essentially positive and concerns maximizing the social good that can be gained from the libraries' collections. Recognizing that the material preserved in the archives constitutes a form of commons—i.e., a shared resource for the community that is publicly funded[43]—then contributing to the making of language models provides a means to distribute the novel forms of value contained within these institutions' collections as widely as possible. Partly a matter of the broad utilitarian benefits that result from more effective and cheaper administration procedures once public authorities implement these AI models, this also pertains to the general value of releasing open source NLP tools for the public to use as they see fit. In exploring new ways to unlock the potential of archival holdings beyond traditional forms of academic research, national libraries can help ensure that society at large derives some benefit from the era of Big Data.

Using the library's resources to participate in a wider societal project of AI development is especially pertinent for lower-resourced languages. As the evaluation results from this article demonstrate, the multilingual model released by Google for languages beyond English and Chinese offers less effective NLP capacities than monolingual BERT models trained for a particular language. Where giant tech companies perceive little incentive to invest in tools specifically for smaller languages, there is a risk

of a chasm emerging between the state of AI in major and lower-resourced languages.[44] In this context, national libraries for this latter group can play a vital role in harnessing their holdings of large volumes of high quality, language-specific data for the making and distribution of state-of-the-art language models.[45] If there are legal restrictions preventing the sharing of such data, establishing in-house data labs at these libraries becomes a necessary work-around. By investing in such projects, national libraries have the chance to underpin the development of a national AI infrastructure, while laying claim to a potent new form of relevance in the process.

The second part of this democratic argument is more critically inclined and relates to probing the problems of an AI future driven purely by private sector actors. One of the key concerns with implementing large-scale language models like BERT is the negative effects of bias, given that the models inevitably reproduce the perspective of the data used to create them. Highlighting the sociopolitical risks of relying upon vast, unaccounted for web materials in training these models, a recent paper warned that such datasets "overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations".[46] That one of the authors of this research was subsequently forced to leave their role at Google amply demonstrates the lack of space to explore these critical issues in the setting of big tech.[47] Developing language models at a national library, by contrast, it is possible not only to scrutinize the workings of data representativity and bias, but also to pursue the type of responsible data documentation that has been proposed as a prerequisite for more accountable forms of NLP.[48] In creating more representative and open tools, national libraries can adopt an ethical approach to AI development that supplements—and in some cases complicates and challenges—the strategies of private tech giants like Google.

Domain-Specific Expertise in Tandem with Data Scientists

Although compelling arguments thus exist for the pursuit of a library-based AI, there are still fairly substantial challenges that must be addressed in order to initiate such an undertaking. The first and clearest obstacle is a question of funding: without significant investment in both computational resources and technical expertise, the type of AI development we have discussed in this article will not prove possible. Presuming such resources have been secured, the second issue that needs to be dealt with is how to organize these experts within the framework of the library to achieve optimal results. This is a generic problem for the introduction of AI-related techniques in an organization: should technical competence be centralized within, say, a lab, or is it better for data scientists to be dispersed and embedded as part of particular groups of the core operation? While there is no "one size fits all" answer, since the particular goals of a given organization will demand specific solutions, it is worth underscoring the need for new forms of collaboration this creates. Developing and implementing AI in a library will require intricate cooperation between the domain-specific expertise of professional librarians and the technical skills of data scientists.

New collaboration will also be needed with a range of external actors, if the maximum potential of this AI development is to be realized. One dimension of this is working alongside, and learning from, researchers who are using the library's collections for innovative forms of data-driven research: in many cases, synergy effects will emerge between the needs and explorations of such projects and the library's AI interests. Another, perhaps more significant, dimension is participating in national and international networks of AI actors with diverse stakeholders from private companies to university departments, who are starting to work together in a rapidly changing field of knowledge. These novel constellations of actors reflect the fact that, in the demanding

space of AI development, it is smarter to pool resources and act collectively than to struggle alone. Given the centrality of high-quality data to the prospects of such enterprises, data labs at libraries—and especially national libraries—can have a significant role to play in the future of AI.

AI should not be regarded as a silver bullet that is capable of providing solutions for all the various complexities of the workings of a library. Neither should the elusive combination of resources, expertise and strategic leadership that is necessary for these libraries to participate in the development of national AI infrastructures be underestimated. Yet as we have sought to demonstrate via the example of a Swedish BERT, there are many good reasons for a closer integration of libraries and data science. In seeking to address the opportunities and challenges created by our era of Big Data, exploring the possibilities of a library-centred AI is certainly a promising place to start.

Acknowledgements

This article has been significantly improved by feedback from our colleagues at KBLab, with particular thanks due to Robin Kurtz, Faton Rekathati, Emma Rende and Fredrik Klingwall. Our work in fine-tuning KB-BERT for NER made use of linguistic resources from Stockholm University, Umeå University and the Swedish Language Bank at Gothenburg University. For non-copyright material, pre-training of the model was supported by Cloud TPUs from Google's TensorFlow Research Cloud (TFRC). To enhance the democratization of the digital resources created at KBLab, our models are hosted on S3 by Hugging Face, where they are freely available for the public.

Notes

---

[1] Amanda Wheatley and Sandy Hervieux, "Artificial Intelligence in Academic Libraries: An Environmental Scan" *Information Services & Use* 39 (2019): 348, https://doi.org/10.3233/ISU-190065.

[2] Alexandre Magueresse, Vincent Carles and Evan Heetderks, "Low-Resource Languages: A Review of Past Work and Future Challenges" *arXiv* (2020), https://arxiv.org/abs/2006.07264.

[3] Dan Jurafksy and James H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition,* 2nd ed. (Upper Saddle River, N.J.: Prentice Hall, 2009).

[4] Jacob Devlin et al, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" *arXiv* (2019), https://arxiv.org/abs/1810.04805v2.

[5] The model, KB-BERT, is available to be downloaded here: https://github.com/Kungbib/swedish-bert-models or here: https://huggingface.co/KB/bert-base-swedish-cased.

[6] Ryan Cordell, "Machine Learning + Libraries: A Report on the State of the Field" LC Labs, Library of Congress (14 July 2020), https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf [accessed 29 March 2021].

[7] For example, Thomas Padilla, *Responsible Operations: Data Science, Machine Learning, and AI in Libraries* (Dublin OH.: OCLC Research, 2019), https://doi.org/10.25333/xk7z-9g97 and Thomas Padilla, Laurie Allen, Hannah Frost, Sarah Potvin, Sarah, Elizabeth Russey Roke and Stewart Varner, "Final Report - Always Already Computational: Collections as Data" (2019), *Zenodo*, https://doi.org/10.5281/zenodo.3152935. See also Thomas Padilla, Hannah Kettler, Stuart Varner et al, *Collections as Data: Part to Whole* (2019), https://collectionsasdata.github.io/part2whole/ [accessed 29 March 2021].

[8] Sarah Ames and Stuart Lewis, "Disrupting the Library: Digital Scholarship and Big Data at the National Library of Scotland" *Big Data & Society* (July 2020), https://doi.org/10.1177/2053951720970576.

[9] David A Smith and Ryan Cordell, "A Research Agenda for Historical and Multilingual Optical Character Recognition," (2018),

https://repository.library.northeastern.edu/files/neu:f1881m035 [accessed 29 March 2021]; D. van Strien, D, K. Beelen, M. Ardanuy, K. Hosseini, B. McGillivray, and G. Colavizza, "Assessing the impact of OCR quality on downstream NLP tasks" *ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, (2020) 1: 484-496, https://doi.org/10.17863/CAM.52068.

[10] Our colleagues at the National Library of Norway have recently engaged in a parallel project of using the library's digital collections to create a Norwegian BERT model. A pre-print detailing their work is now available: Per E. Kummervold, Javier de la Rosa, Freddy Wetjen and Svein Arne Brygfjeld, "Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model" *arXiv* (2021), https://arxiv.org/abs/2104.09617.

[11] Yann LeCun, Yoshua Bengio and Geoffrey Hinton, "Deep Learning" *Nature* 521 (2015): 436–44, https://doi.org/10.1038/nature14539.

[12] S.J. Pan and Q. Yang, "A Survey on Transfer Learning" *IEEE Transactions on Knowledge and Data Engineering* 22:10 (2010): 1345–59, https://doi.org/10.1109/TKDE.2009.191; A. Radford et al, "Improving Language Understanding by Generative Pre-Training" [pre-print] (2018) https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 29 March 2021].

[13] Chuanqi Tan et al, "A Survey on Deep Transfer Learning" *arXiv* (2018), https://arxiv.org/abs/1808.01974.

[14] Ashish Vaswani et al, "Attention Is All You Need" *arXiv* (2017), https://arxiv.org/abs/1706.03762v5.

[15] Devlin, "BERT".

[16] Alex Wang et al, "GLUE: A multi-task benchmark and analysis platform for natural language understanding" *arXiv* (2018), https://arxiv.org/abs/1804.07461.

[17] Devlin, "BERT".

[18] Ibid.

[19] For example, Louis Martin et al, "CamemBERT: a Tasty French Language Model," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), https://arxiv.org/abs/1911.03894.

[20] The Swedish Public Employment Agency's model is available here: https://github.com/af-ai-center/SweBERT.

[21] The provisional character of AF-AI's model was outlined in the documentation connected to its release: https://github.com/af-ai-center/SweBERT [accessed 29 March 2021].

[22] Göran Konstenius, *Plikten under lupp! En studie av pliktlagstiftningens roll, utformning och relevans i förhållande till medielandskapets utveckling*, Kungl. Biblioteket (2017), https://urn.kb.se/resolve?urn=urn:nbn:se:kb:publ-539.

[23] For this notion of a representative language, see Benedict Anderson, *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, rev. ed. (London: Verso, 2016) and Laura M. Ahearn, *Living Language: An Introduction to Linguistic Anthropology*, 2nd ed. (Chichester: John Wiley & Sons, 2017).

[24] For more technical details of the composition and cleaning of data for this training process, see Martin Malmsten, Love Börjeson and Chris Haffenden, "Playing with Words at the National Library of Sweden: Making a Swedish BERT" *arXiv* (2020), https://arxiv.org/abs/2007.01658.

[25] Wang, "GLUE"; Alex Wang et al, "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems" *arXiv* (2019), https://arxiv.org/abs/1905.00537.

[26] This corpus is available here: https://spraakbanken.gu.se/en/resources/suc3.

[27] More details about this annotation tool can be found here: https://spraakbanken.gu.se/en/tools/sparv/annotations.

[28] For more details on evaluating and testing models within NLP, see Jurafksy and Martin, *Speech and Language Processing*, 67–8.

[29] For further details about this evaluation process from a data science perspective, including how we also evaluated the models for part-of-speech (POS) tagging, see Malmsten, "Playing".

[30] For an example of such practical implementation in the public sector, see Annica Wallenbro Stojcevski, "Sigmas utmaning: skapa en lösning för Vinnova att utnyttja AI-teknik för hantering av ansökningar" *Pulse* (2020), https://pulse.microsoft.com/sv-se/work-productivity-sv-se/na/fa1-vinnova-och-sigma-tar-ai-till-hjalp-for-beredning-av-arenden/ [accessed 29 March 2021].

[31] Daniel Castro and Joshua New, "The Promise of Artificial Intelligence" *Centre for Data Innovation* (2016), https://euagenda.eu/upload/publications/untitled-53560-ea.pdf [accessed 29 March 2021].

[32] We have recently tested an annotation project for librarians at the National Library of Sweden, see our blogpost on this for more details: Elena Fano, "The Power of Crowdsourcing" *The KBLab Blog* (8 April 2021), https://kblabb.github.io/posts/2021-04-08-internal-crowdsourcing-for-dataset-creation/ [accessed 11 November 2021].

[33] Cf. Susan Carey, "Bootstrapping & the Origin of Concepts" *Daedalus* 133: 1 (2004): 59–68, https://www.jstor.org/stable/pdf/20027897.pdf.

[34] Marianna Neves and Jurica Ševa, "An Extensive Review of Tools for Manual Annotation of Documents" *Briefings in Bioinformatics* 22: 1 (2021): 146–63, https://doi.org/10.1093/bib/bbz130.

[35] Ron Artstein and Massimo Poesio, "Inter-Coder Agreement for Computational Linguistics." *Computational Linguistics* 34: 4 (2008): 555-96, https://www.aclweb.org/anthology/J08-4004.pdf.

[36] Ikuya Yamada et al, "Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation" *arXiv* (2016), https://arxiv.org/abs/1601.01343.

[37] Laurel Orr et al, "Bootleg: Chasing the Tail with Self-Supervised Named Entity Disambiguation" *arXiv* (2020), https://arxiv.org/abs/2010.10363.

[38] Christian Bizer, Tom Heath and Tim Berners-Lee, "Linked Data: The Story so Far," in *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, ed. Amit P. Sheth (Hershey PA: Information Science Reference, 2011).

[39] Faton Rekathati, "Curating News Sections in a Historical Swedish News Corpus" Independent Master's Thesis. Linköping University, Department of Computer and Information Science (2020), http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-166313.

[40] A recent Master's project at KBLab explored the potential of such an approach, see: Andreas Estmark, "Text Block Prediction and Article Reconstruction Using BERT" Independent Master's Thesis. Uppsala University, Department of Statistics (2021), http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-447248.

[41] For the example of such a multi-modal approach applied to the more specific problem of classifying advertisements in historical newspapers, see our blogpost: Faton Rekathati, "A multimodal approach to advertisement classification in digitized newspapers" *The KBLab Blog* (28 March 2021), https://kblabb.github.io/posts/2021-03-28-ad-classification/ [accessed 11 November 2021].

[42] For problems with OCR quality in historical newspapers, see Myriam C. Traub, Jacco van Ossenbruggen and Lynda Hardman, "Impact Analysis of OCR Quality on Research

Tasks in Digital Archives" in *Research and Advanced Technology for Digital Libraries* ed. S. Kapidakis, C. Mazurek, C. and M. Werla, Springer, Cham (2015), https://doi.org/10.1007/978-3-319-24592-8_19.

[43] David Harvey, "The Future of the Commons" *Radical History Review* 109 (2011): 101–07, https://doi.org/10.1215/01636545-2010-017.

[44] Indeed, the authors of the following study suggest that the current emphasis on major languages creates NLP systems that "dramatically underrepresent the voices of much of the world." Esma Wali et al, "Is Machine Learning Speaking my Language? A Critical Look at the NLP-Pipeline Across 8 Human Languages" *arXiv* (2020), https://arxiv.org/abs/2007.05872.

[45] See also the Norwegian case, which further exemplifies and reinforces this point: Kummervold et al, "Operationalizing a National Digital Library".

[46] Emily M. Bender et al, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proceedings of FAccT* (2021), https://doi.org/10.1145/3442188.3445922; Timnit Gebru, "Race and Gender" in *The Oxford Handbook of Ethics of AI*, ed. Markus D. Dubber, Frank Pasquale and Sunit Das (Oxford: Oxford University Press, 2020), https://doi.org/10.1093/oxfordhb/9780190067397.013.16.

[47] Karen Hao, "We Read the Paper That Forced Timnit Gebru out of Google. Here's What It Says" *MIT Technology Review* (7 Dec 2020), www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/ [accessed 29 March 2021].

[48] Bender, "Dangers".