Machine learning and big data are unlocking Europe's archives

December 11 2020, by Horizon Magazine, Fintan Burke

AI models that can be trained to recognise and transcribe historical handwritten documents are helping digitise national and city archives. Credit: pxhere.com/licenced under CC0

From wars to weddings, Europe's history is stored in billions of archival pages across the continent. While many archives try to make their

documents public, finding information in them remains a low-tech affair. Simple page scans do not offer the metadata such as dates, names, locations that often interest researchers. Copying this information for later use is also time-consuming.

These issues are well-known in Amsterdam, which is trying to disclose its entire archives. For the notary records alone 'there's about three and a half kilometres in paper," said Pauline van den Heuvel, an archivist at Amsterdam City Archives in the Netherlands. That's around 11,800 pages of A4 paper laid end-to-end. She says the total collection is about 50km long, equivalent to 170,000 A4 pages. "We know they are really important (documents), but it's really a black hole."

She says that manually recording the names available in these documents usually requires decades of work and funding.

A few years ago, the archive partnered with the READ project and its Transkribus platform, which offers archivists a new way to transcribe and search their historical documents. The online platform allows users to train a computer handwriting recognition model to transcribe historical documents written by hand in a variety of European languages.

Users train a model with 50 to 100 pages of existing transcriptions or ones that are manually transcribed into the system. Once trained, the model uses machine learning to compare the handwriting patterns it now knows with that of the documents the user wants to transcribe. The model automatically transcribes line by line. For it to work, the new documents must be in the same or similar handwriting to what the model has seen before.

So far users have trained more than 7,700 individual models says Dr. Günter Mühlberger of the University of Innsbruck, Austria, who coordinated the project.

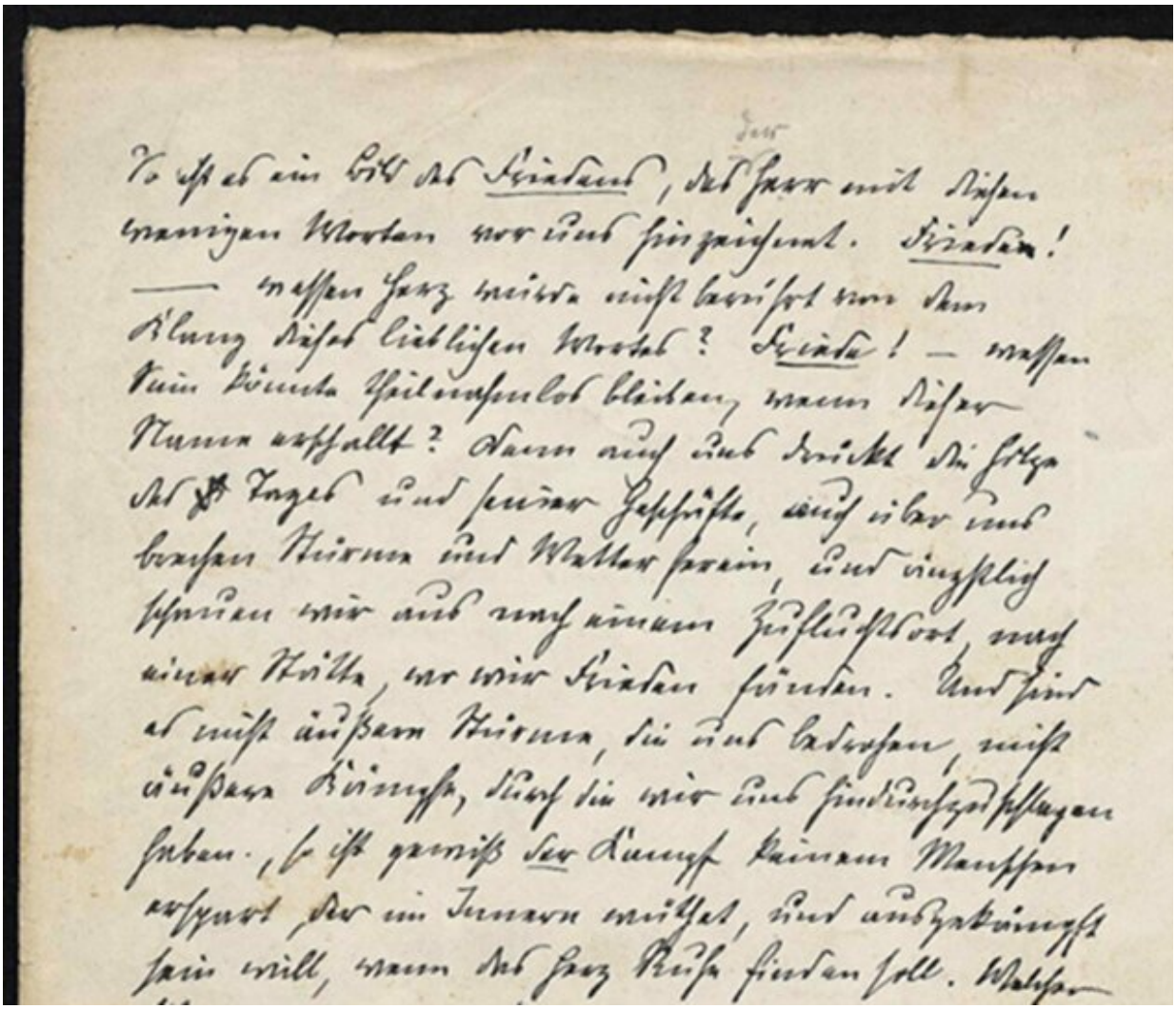
Users can either train their own model or select a pre-existing model. One available model recognises the handwriting style of English philosopher Jeremy Bentham. Another recognises the handwriting styles of 17th century Italian secretaries. A user can use such models as a starting point for their own training.

After Transkribus has done its work, users often just need to proofread to correct any minor errors. While this might seem like a lot of initial work, it can save archivists, historians and scholars hundreds—if not thousands—of hours sitting in front of a computer transcribing the complete set of documents by hand.

Machine learning

Transkribus is the result of the READ project's work to develop new technology to better recognise and automatically transcribe handwritten documents. These transcriptions can then help researchers better search for words or phrases among the billions of pages stored across the continent's archives.

For Transkribus, the project used a 'supervised machine learning' algorithm that collates historical data as it learns. This data can be used to train bigger models.



A handwritten sermon by Heinrich Bassermann from 17 November 1871 is one example of a document that can be digitised with the AI-based software. Credit: Universitätsbibliothek Heidelberg/licenced under CC-BY-SA 4.0

Crucial for the project is 'big data' – enough archival documents that can give the algorithm a complex understanding of handwriting and page layouts. The project cooperated with more than 70 archives, universities and research organisations across Europe, including the Hessian State Archives in Germany and the Archivio Storico Ricordi in Italy. "From

the Middle Ages to the 20th century, we got thousands of pages with different layouts and different (types of) writing," said Dr. Mühlberger.

He says that Transkribus is likely the largest collection of training data for historical handwriting worldwide—more than 700,000 documents.

Their major challenge, says Dr. Mühlberger was to also train the algorithm to recognise what a line of words looks like in a handwritten document. He explains that conventional 'optical character recognition' software used to turn PDFs into text, for example, works well with old, printed documents because the lines and word spaces have a fixed layout.

"If you try to do the same with handwriting," he said, 'you fail completely." It is more or less impossible to isolate single characters in cursive writing, he says.

The project's initial machine learning algorithms could recognise 85% of handwritten text. However, the project soon realised that for archives dealing with thousands of handwritten archival pages this was not good enough.

"Eighty-five percent looks good in a research paper, but not for a user sitting in front of (their) computer," he said.

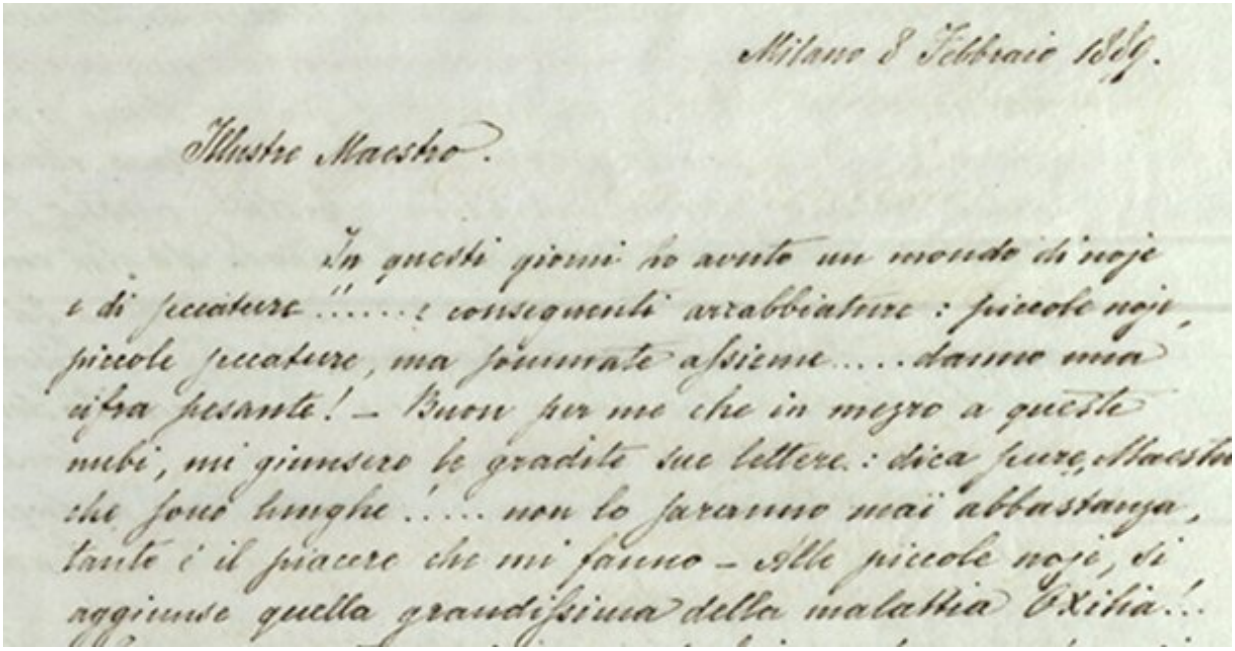
Lines

Researchers then used two methods to increase their program's accuracy. They first reconsidered how their program would recognise lines of text. Rather than look for the entire block area of the text, they trained the algorithm to look for the common 'baseline' on which each word rests, similar to how a line-ruled page teaches children to write evenly on a page. "This was a very important simplification," said Dr. Mühlberger.

More than 100,000 lines were drawn during the project to train the algorithm to recognise what a common line looks like. If Transkribus cannot recognise a line of text users can show the program by drawing a line underneath—a simpler technique that saves hours of time in the long run.

Another change was to how Transkribus recognises languages. Earlier in the project they used dictionaries to help it to recognise whole words in the document. But by switching to recognise only the characters among the training documents the team was able to improve its accuracy by a further 10%. Recognising the letters also means the algorithm is useful for old forms of languages—and is able to deal with abbreviations. A recent addition allows Transkribus to expand abbreviations automatically.

They are looking to further refine how Transkribus works. One method involves merging the different user-trained algorithms to improve Transkribus' text recognition abilities as a whole. Another is adding new features, such as transcribing structured information including tables and forms, and allowing archivists to search and correct keywords en masse. Dr. Mühlberger says that they hope to improve the platform's user experience and layout so that even small-scale family historians can easily use Transkribus to upload and transcribe a scanned copy of a document. Transkribus' cooperative structure means any money earned feeds back into the platform to improve its services.



Once the software has been trained to recognise a particular person's handwriting - such as this letter written in 1889 by Giulio Ricordi, the general manager of the Ricordi publishing house - then it can automatically transcribe other documents from the same author. Credit: Archivio Storico Ricordi, Milan

Archives

Since its launch in 2015, the amount of people using Transkribus has grown substantially. The platform now has more than 45,000 users, including volunteers from the Amsterdam City Archives. Van den Heuvel says that the archive co-opted Transkribus into their work when they realised that indexing the names, places and dates in their 17th and 18th century documents would take decades of work. A trained Transkribus algorithm was able to finish transcribing the project's 18th century documents a year earlier than expected. She says that while volunteers may take months to index 50,000 scanned documents, a model, once trained, takes only a few hours. A team of 300 volunteers

now only needs to double-check the transcriptions, she says.

"It's only the beginning," she said. "Now you can research patterns in big amounts of data, connections between people—it's completely new research." Work is still in progress, though van den Heuvel says that the finished work will be connected to the European Time Machine network of institutions using records to shed light on Europe's social and political evolution over time.

There are other ongoing projects with archives throughout Europe. Finland's national archive is also working to release its national archives and has used Transkribus in its work since 2016. Maria Kallio, senior research officer at the National Archives Service of Finland says that the archive first used Transkribus on a few diary entries they had. After being impressed with the results, they decided on a bigger task.

"We had started transcribing these 19th century court records, which is a huge collection, just the 19th century bit is millions of pages," she said. "To make it easier to do research on the… records we thought it could be a good idea to try the technology on them."

Their work with the READ project has led to the Finnish Archives now releasing around 800,000 transcribed documents to the public, including legal records of deeds, mortgages, and guardianship cases across most of Finland dating back to the 16th century. People can now use these records to research family history and track ownership of property.

There are still limitations with the technology. Van den Heuvel says that a lot of training material is needed for all the varieties of 17th century handwriting to create a general model that could work on such a large, varied collection such as theirs. Collections with a large amount of pages also need to finance the cost of using the Transkribus technology which is free to use for the first 500 pages before needing to buy 'credits' to

transcribe more pages. For example, €18 for the next 120 handwritten pages.

Nonetheless, the technology has been welcomed by researchers. "It's possible to make these kind of research questions to answer wider questions about how things developed," said Kallio. "Now you can actually have a grasp on the whole material, and ask questions that were not possible earlier."

In late September 2020, the READ project and its Transkribus software was named one of the winners of the European Commission's Horizon Impact Award.

Provided by Horizon: The EU Research & Innovation Magazine

Citation: Machine learning and big data are unlocking Europe's archives (2020, December 11) retrieved 20 April 2024 from https://techxplore.com/news/2020-12-machine-big-europe-archives.html