Digital transformation and the changing role of news media in the 21st Century

IFLA Newspapers/GENLOC Pre-Conference Satellite Meeting
Geneva, Switzerland, 13-14 August 2014

Large-scale refinement of digital historic newspapers with named entity recognition

Clemens Neudecker, Lotte Wilms, Willem Jan Faber, Theo van Veen
@ KB National Library of the Netherlands

Abstract

Within the Europeana Newspapers project (www.europeana-newspapers.eu), full-text will be produced for over 10 million pages of digitised historical newspapers by applying Optical Character Recognition (OCR) and Optical Layout Recognition (OLR). In order to further increase the usability of the full-text, Named Entity Recognition (NER) is also applied to materials in Dutch, German and French language. The main aim of NER is to identify and classify entities such as persons, locations and organisations in the full-text in order to enhance the searchability, and to subsequently link them to online resource descriptions and authority files (e.g. DBPedia, VIAF). By also cross-linking the named entities, a whole scale of possibilities in terms of analyzing large multilingual corpora can be unlocked. Therefore, the KB National Library of the Netherlands has been adapting an open source tool from Stanford University for training a state-of-the-art NER system specifically for Europeana Newspapers. Some of the main considerations in producing the software were scalability and the support for the standards and formats that are most widely used in newspaper digitisation such as METS and ALTO. Another important requirement was to retain information about the exact location of the named entities on a page throughout the refinement process, so they can for example be highlighted in a viewer. We will introduce the overall workflow for NER as implemented in Europeana Newspapers, discuss some design considerations as well as technical issues and

lessons learned while building the software, and present first results from applying the system to historical newspaper content from three different languages.

Introduction

A group of 18 European partner institutions have joined forces in the "Europeana Newspapers" project and will, over a period of three years, provide more than 18 million newspaper pages to the Europeana service.

Europeana is the online collection of European cultural heritage, from book to photo and from painting to artifact. All items have a thumbnail with a link back to the holding institution. There are currently over 30 million objects available from more than 2300 institutions and 36 countries. The portal saw 3.5 million visits in 2012, from January to September alone. The actual number is probably even higher, as all content is also made available via an API[1].

The Europeana Newspapers Project (funded under the European Commission's Competitiveness and Innovation Framework Programme 2007 – 2013) started in February 2012 and aims at the aggregation and refinement of newspapers through The European Library. In addition, the project addresses challenges particularly linked with digitised newspapers:

- use of refinement methods for OCR, OLR/article segmentation, and named entity recognition (NER) to enhance search and presentation functionalities for Europeana customers,
- quality evaluation for automatic refinement technologies,
- transformation of local metadata to the Europeana Data Model (EDM),
- metadata standardization in close collaboration with stakeholders from the public and private sector.

This paper focusses on the task of adding named entities (people, places and organisations) to a selection of the newspapers that are enriched in the project. Named entities offer the user a new way of browsing the material, which will heighten the experience of the portal that is currently being developed. Ultimately, the NEs will provide users with a quick insight into the text and with identification, disambiguation and linking, the named entities will open up possibilities of searching across languages and in the linked data network.

We will discuss our approach in detecting the named entities in the digitised newspapers, what challenges we faced when doing so and the results we achieved with the current set-up.

Approach

The KB already gained some experience with various technologies for Named Entities Recognition during the STITCH+ project[2], which formed part of the larger CATCH/CATCH+ (Continuous Access To Cultural Heritage) Programme[3] of the Netherlands Organisation for Scientific Research (NWO), and also in the EU Project IMPACT[4] (IMProving ACcess to Text).

For the implementation of our Named Entities Recognition system for Europeana Newspapers, we decided to use the Stanford NER tagger[5] and adapt it for the purposes of the project, mainly for these reasons:

- We wanted a statistical/machine-learning tool rather than a fully rule-based system, since we did not have linguistic experts but the resources to manually create large sets for training
- Prior testing in IMPACT confirmed the competitive results that could be achieved with the Stanford-NER-Tagger
- As part of the IMPACT project, the Institute for Dutch Lexicology (INL) developed some modules to improve NER for documents with OCR errors and historical spelling variants[6] – we wanted to test the effectiveness of these modules
- The Stanford system is thread-safe and can thus significantly increase throughput when run on a multi-core system
- The Stanford software is open source under the GPL license and supported by a large international and active community
- The CRF algorithm that is used internally by the Stanford-NER-Tagger is quite robust against noise (OCR errors, etc.)

Accordingly, a workflow has been set up to support the creation of annotated training materials and classifiers for digitised historical newspaper content in the languages Dutch, German and French. In a first step, libraries select suitable

newspaper titles taking into consideration user demand, age and condition of the newspaper as well as intellectual property rights that might hinder the re-distribution of a given newspaper titles full-text. The newspaper content is OCRed by either the University of Innsbruck or Content Conversion Specialists (CCS). The resulting full-text in ALTO format is then loaded into a MySQL database and transformed into raw text with metadata. The database is connected to an online Attestation Tool in which the libraries have to mark named entities manually. Any annotations that the user is adding via his browser are internally stored in the database.

Figure 1: The INL attestation tool

In the Attestation Tool, a user can mark named entities of the categories Person, Location, Organisation and Other. It is also possible to indicate a named entity that contains an OCR error – in this way, such entities can be filtered out and

corrected in a post-processing step, so they can also feed into the training of the NER. Typically, around 100 pages of OCR are annotated in this way by the content holders – split into two sets. Around one quarter of the manually annotated data is used as ground truth for evaluation, which is done twice, after the first 50 pages and then again after the full 100 pages (for more, see chapter Evaluation).

Once all 100 pages have been fully annotated, the data is exported from the database, and again transformed into the ALTO format, but now with additional tags containing the information about the named entities that were added by the user. Using a Python-script, these ALTO files have to be transformed yet once more, into the BIO format[1], a simple plain text format that is used for training the Stanford NER system. As part of the transformation to the BIO format, an additional cleaning step can be applied: a filter is used to select only sentences that contain more than 20 words and at least one named entity. With this additional processing step, we found that we could significantly reduce the size of the resulting classifier (and thus also the processing speed and throughput of our NER system) while at the same time maintaining the performance of the detection of named entities.

Finally, the post-processed BIO files form the basis for the training of the Stanford NER classifier, which is described in detail in the FAQ of the software[2]. In this final step of the creation of the classifier there is also the option to add gazetteers, i.e. lists of known names of persons and places that the classifier can train itself to recognize. These can simply be plain text lists the location of which has to be indicated in the .props properties file of the Stanford tool. Experiments showed that the recognition performance could be boosted considerably if appropriate gazetteers are used (meaning they should match well with the content that the classifier is trained for).

The BIO files together with the gazetteers and the properties file are all that is needed to proceed with the (automatic) training of the Stanford NER system. This

[1] http://ifarm.nl/erikt/papers/eacl99.pdf
[2] http://nlp.stanford.edu/software/crf-faq.shtml#a

step can take a considerable amount of time and requires large amounts of memory, but it is a one-time thing. As soon as the training has finished and a classifier has been created, this classifier can be directly applied to recognize named entities in texts.

Due to some financial windfalls in the project, we had the opportunity to not only work on adding named entities to the text, but also to look into linking these to relevant databases. By doing so, the named entities can provide even more valuable information for the user, as the entity can be disambiguated and by linking the user can use certain background information that is connected to the entity. To achieve this, each named entity will be searched in DBpedia titles. The completeness of the match (e.g. first name and surname versus only surname), the frequency of the DBpedia title occurring as a link in DBpedia, and the amount of matches between other named entities in the text and in the abstract of the DBpedia article are used to estimate the probability that a specific DBpedia title corresponds to the named entity in the text. The DBpedia link and the estimate for the probability is stored and can then be used when presenting a newspaper article and for indexing relevant DBpedia data for the resources containing the named entity. This way, the entity is first checked against the database to ascertain if the entity is indeed a certain person, location or organisation and can then be connected to more information about the relevant entity.

All results will be shown in the historic newspaper browser that is the end product of the Europeana Newspapers project and is built by The European Library. This browser combines all content from the 12 partners that work together to publish the 10 million digitised newspapers pages for the European public. While writing this paper, the prototype of the browser has been released and will be updated regularly until the end of the project in January 2015. The named entities will be integrated in the browser during the second half of 2014.

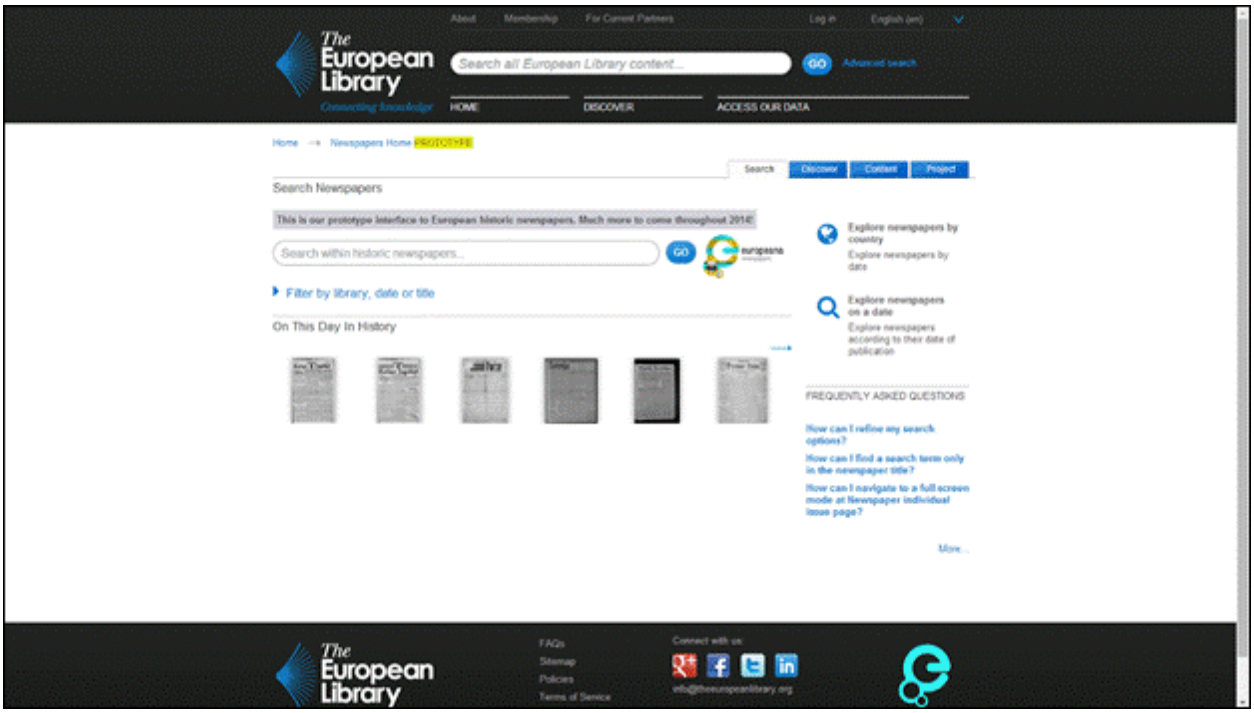

Figure 2: The Europeana Newspapers browser

Challenges

Due to the nature of the content in the Europeana Newspapers project – a diverse set of 10 million pages of mainly historical newspapers drawn from the digital collections of 12 different libraries in Europe – we are facing a number of particular challenges.

At first, there is the obvious issue of bad OCR quality. Due to the typical characteristics of newspapers (multiple columns, mix of articles with graphical elements like advertisements, small characters, paper deterioration, etc.) the OCR that is derived with automated processing is of average quality, with word accuracy rates sometimes being lower than 50 percent even. While creating the different models for the different languages, we tried an approach with a filter step in between the export from attestation, before generating a classifier. We took a basic tokenizer and looked at sentences with at least 20 words and at least one entity. Doing this yielded a smaller model in MB that was also able to perform better than without this noise reduction step.

Secondly, there is the problem of historical spelling variations. As the content comprises newspapers as old as from the 17th century, there occur not only words but also person or in particular place names that are spelled differently than today. We attempted to tackle this problem by using a spelling variation module that was developed by the Institute of Dutch Lexicology (INL) within the IMPACT project. However, this module was made for an older version of the Stanford tool and adapting it for the new version proved to be quite difficult with the limited resources that were available for this task from the INL.

Also, Europeana Newspapers is an international project and thus deals with content in various languages. This means that choices had to be made when dealing with the selection of newspapers for the enrichment, as training material is needed for each language to ensure a usable quality of NEs. We opted for those languages where we could make use of existing material, which means that only newspapers in French, German and Dutch will be processed. There is also more than enough material for English, but there is no partner that contributes a large selection of English content to the project and was therefore not followed up. Further material to adapt the software to the historical material had to be created, which was done by the libraries themselves with the attestation tool.

Evaluation

In order to evaluate the success rate of the Europeana Newspapers NER system, 20 pages are split off from the manually annotated 100 pages for each language. In a first step, a k-fold cross-evaluation is performed, i.e. the classifier is trained on the remaining 80 pages and the results of the automatic processing are then compared to the manually added named entities on the pages set aside for evaluation. The evaluation is performed in the standard way for named entities using Precision and Recall. A distinction is being made between true positives (named entities found by the automatic system and also in the manual annotations), true negatives (tokens that have not been identified as named entities either by the automatic system or in the manual annotations), false positives (named entities that were identified by the automatic classification but

not in the manual annotations) and false negatives (named entities that were classified in the manual annotations but not recognized by the automatic system).

Based on this, it is then possible to calculate the score for Precision, which gives an account of the number of named entities that have been correctly found by the software that are really named entities (of the correct type) and also Recall, which indicates how many of the total number of named entities present in the text have been recognized by the software. Precision and Recall can then be combined to a normalized score between 0 and 1, which is called the F-measure.

Here are the preliminary figures we derived from the evaluation of the available Dutch and French training sets:

Results for Dutch:

	Persons	Locations	Organizations
Precision	0.940	0.950	0.942
Recall	0.588	0.760	0.559
F-measure	0.689	0.838	0.671

Results for French (preliminary):

	Persons	Locations
Precision	0.529	0.548
Recall	0.834	0.216
F-measure	0.622	0.310

Note that a score for organisations could not be computed due to the lack of sufficient amount of organisations in the French training set.

Conclusion

Europeana Newspapers is working very hard to deliver the optimal digitised newspaper browser for European newspapers. Detecting and tagging named entities in a selection of these newspaper is part of this undertaking. After careful testing, the workflow that we have opted for is that of firstly selecting the content and using some 100 pages to make a training set with a specific tool. This is then converted to the BIO format and used to train the classifier of the Stanford NLP tool. In this step gazetteers are also added into the mix to ensure the best possible matches and finally the material is processed. After the named entities are recognised they can be used to link to other existing online resources.

The work that is done to achieve as many as possible tagged names, places and organisations in the Dutch, French and German newspapers that are being contributed to the project is very valuable for the project and ultimately the users of the historic newspaper browser. Research into user behaviour, such as that of Paul Gooding[3], has shown time and again that users of digital newspaper portals mainly search for names of persons and places, and identifying and disambiguating named entities provides the users with even more context that will help them using these to do e.g. genealogy research. Also by tagging the named entities, the researchers can browse through the material in new ways and discover parts of the collection that they have not seen before. The project also gives the partners the opportunity to enrich their material in ways that were not possible without the help of the consortium.

Finally, all technical resources that are being created in the project for the purpose of named entity recognition (raw training data + binary classifiers) will be made available for reuse as open source, so other interested parties can easily benefit from or even further extend the named entities relate work of Europeana Newspapers.

For the future, it would be very interesting to explore the linking more further to identify as many NEs as possible. The main challenge in linking the NEs lies in the

[3] http://dharchive.org/paper/DH2014/Paper-310.xml

disambiguation and identification of variant names. However, if this step does succeed, the researchers will have the best possible options when browsing and searching the 10 million enriched newspaper pages in the collection of Europeana Newspapers.

References

- John Lafferty, Andrew McCallum, Fernando C.N. Pereira: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), pp. 282-289, http://repository.upenn.edu/cis_papers/159/.
- Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke and Magdalena Luszczynska: Comparison of Named Entity Recognition tools for raw OCR text. Proceedings of KONVENS 2012 (LThist 2012 workshop), pp. 410-414, September 2012, http://www.researchgate.net/publication/230898508_Comparison_of_Named_Entity_Recognition_tools_for_raw_OCR_text/file/d912f505ec8128d617.pdf.
- Rosa Stern, Benoit Sagôt and Frédéric Béchet: A Joint Named Entity Recognition and Entity Linking System. EACL 2012 Workshop on Innovative hybrid approaches to the processing of textual data (2012), http://hal.archives-ouvertes.fr/docs/00/69/92/95/PDF/eacl12hybrid.pdf.
- Thomas Packer, Joshua Lutes, Aaron Stewart, David Embley, Eric Ringger and Kevin Seppi: Extracting Person Names from Diverse and Noisy OCR Text. AND '10 Proceedings of the fourth workshop on Analytics for noisy unstructured text data, pp. 19-26, http://www.deg.byu.edu/papers/Ancestry_NAACL_HLT_Paper.pdf.
- Claire Grover, Sharon Givon, Richard Tobin and Julian Ball: Named Entity Recognition for Digitised Historical Texts. International Conference on Language Resources and Evaluation 2008, http://www.ltg.ed.ac.uk/np/publications/ltg/papers/bopcris-lrec.pdf.
- Mārcis Pinnis: Latvian and Lithuanian Named Entity Recognition with TildeNER. LREC, pp. 1258-1265, European Language Resources Association ELRA (2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/948_Paper.pdf.

- Marco Dinarelli and Sophie Rosset: Tree-Structured Named Entity Recognition on OCR Data: Analysis, Processing and Results. LREC, pp. 1266-1272, European Language Resources Association ELRA (2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/1046_Paper.pdf.
- Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum and Ludovic Quintard: Extended Named Entity Annotation on OCRed Documents: From Corpus Constitution to Evaluation Campaign. LREC, pp. 3126-3131, European Language Resources Association ELRA (2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/343_Paper.pdf.
- Frank Landsbergen: Named Entity Work in IMPACT. IMPACT Final Conference 2011, 24-25 October, London, UK, http://www.digitisation.eu/fileadmin/user_upload/Deliverables/IMPACT_D-EE2.6_NE_work_in_IMPACT.pdf.

Author biographies

Clemens Neudecker, M.A. Philosophy, Computer Science, Political Science, Technical Coordinator Research in the Innovation and Development department of the KB. He has been involved in numerous digitisation projects over more than a decade, previously at the Bavarian State Library. He has a particular interest in OCR and scalable digitisation workflows.

Willem Jan Faber, Research Programmer at the Innovation and Development department of the KB. Early adaptor of information technology, hacker and open source/Linux evangelist. He holds a community degree in information and communications technology and a Bachelor degree in the field of Computer Human Interaction design. In the past he also worked for internet companies XS4ALL and FOX-IT.

Theo van Veen, Senior Researcher at the Innovation and Development department of the KB. After getting his degree in physics he has been involved in IT and library automation and contributed in various European projects. His main focus is now on service integration and identifying and linking named entities.

Lotte Wilms, KB Project Leader for the Europeana Newspapers Project. She has a BA in English Language and Culture and an MA in Medieval Studies from the University of Utrecht. She has worked at the KB since 2008 on various projects, such as the IMPACT project, the Short-Title Catalogue Netherlands and the digitisation projects Staten-Generaal Digitaal and Early Dutch Books Online.

---

[1] For more information about Europeana, see http://europeana.eu/portal/aboutus.html
[2] http://stitchplus.kbresearch.nl/
[3] http://www.catchplus.nl/
[4] http://www.impact-project.eu/
[5] http://nlp.stanford.edu/software/CRF-NER.shtml
[6] http://www.digitisation.eu/tools/browse/toolbox-for-lexicon-building/named-entities-recognition-tool-nert/