An Open Corpus for Named Entity Recognition in Historic Newspapers

Clemens Neudecker
Staatsbibliothek zu Berlin
Potsdamer Straße 33, 10785 Berlin
clemens.neudecker@sbb.spk-berlin.de

Abstract

The availability of openly available textual datasets ("corpora") with highly accurate manual annotations ("gold standard") of named entities (e.g. persons, locations, organizations, etc.) is crucial in the training and evaluation of named entity recognition systems. Currently there are only few such datasets available on the web, and even less for texts containing historical spelling variation. The production and subsequent release into the public domain of four such datasets with 100 pages each for the languages Dutch, French, German (including Austrian) as part of the Europeana Newspapers project is expected to contribute to the further development and improvement of named entity recognition systems with a focus on historical content. This paper describes how these datasets were produced, what challenges were encountered in their creation and informs about their final quality and availability.

Keywords: Named Entity Recognition, Historic Newspapers, Optical Character Recognition

1. Introduction

Europeana Newspapers [1] is a collaborative European research and digitization project co-funded by the European Commission that brought together 18 partners from 12 countries to aggregate full text and scanned images of approximately 12 million pages of historic European newspapers and make them available via an online service provided by The European Library[2].

The entire collection comprises more than 1,000 digitized newspaper titles from 23 libraries across Europe, spanning news articles from four centuries (from 1618 up to 1990) in more than 40 different languages. Based on scans of the newspaper pages, full texts were produced by processing the images with optical character recognition (OCR) using state-of-the-art software tools. A conservative estimate of 2,000 words per full newspaper page leads to a total amount of approximately 25 billion tokens in the overall corpus that can be searched.

Figure 1: Example scan of a title page from the French newspaper "Le Temps"

Named entity recognition (NER) is the process of automatic extraction of named entities by means of recognition (finding the entities in a given text) and their classification (assigning a type). The application of named entity recognition to the full text collection derived by means of OCR can dramatically improve the usability of the collection, since it allows users to find specific individuals or places in large text collections. It can also be extremely useful for automatic indexing, automatic summarisation, topic detection, trend analysis and numerous data analysis and visualisation applications. Furthermore, it enables additional refinements such as the cross-linking of the newspaper content with other online information resources like authority files and the linked (open) data cloud. This is particularly relevant in the case of newspaper content. An analysis of the web log files of the digital newspaper collection from the National Library of Wales (Gooding, 2014) indicated that up to 9 out of 10 user queries are either person or place names, rather than standard keywords. This is a specific feat of digital newspaper collections, where users often search for news items relating either to their own family history or a location that they live(d) in or have some other personal relation to, and which are of particular interest to genealogists. Accordingly, providing users of such or similar digital newspaper collections with the ability to directly search and browse names and places that have been extracted from the newspaper articles has been a frequently requested feature.

2. Approach

For a subset of the total collection in the languages Dutch, French and German, technologies and resources for the extraction of named entities from the electronic full texts have been developed. The decision to focus on those three languages was based on the available resources in the project, and the fact that close to 50 percent of the 12 million pages full text in the collection was in one of these three languages. While the project team concerned with

[1] http://www.europeana-newspapers.eu/.
[2] The European Library: Historic Newspapers, http://www.theeuropeanlibrary.org/tel4/newspapers.

NER had prior experience in the development of resources for Dutch, German and English, for the French language newspapers from the Bibliotèque nationale de France it was decided to cooperate with the ACASA LIP6 Group of the Université Pierre et Marie Curie in Paris, which took a slightly different approach (Mosallam, 2014). Nevertheless, training resources and classifiers were produced in such a way that they could always be used interchangeably with the materials being developed for the other languages, and with the same open licensing conditions applied.

Following up on the prior work and experience obtained in the CATCH+ project[3], the project team decided to make use of the Stanford Named Entity Recognizer for the given task. The Stanford software provides a general implementation of (arbitrary order) linear chain Conditional Random Field (CRF) sequence models (Finkel et al., 2005). Next to the availability of prior experience with the tool, there were further reasons for choosing the Stanford NER software:

- The probabilistic approach has benefits over rule-based systems in terms of scalability and ease of adaptation to different languages
- The tool is mature, widely used and supported by a large community of researchers and practitioners
- The tool is available as open source under the GNU General Public License (GPL)
- The tool is platform-independent as the source code is in Java
- The program code is robust and thread-safe, allowing parallelization for increased throughput

In order to make the tool fit for the project workflow, the software had to be slightly adapted to allow the parsing and processing of some domain-specific file formats such as METS and ALTO which are common XML-based formats used to encode OCR results and related structural information in a digital library environment.

3. Methodology

Since the Stanford NER software uses a statistical rather than rule-based approach, it first requires training on a manually annotated corpus to deliver the optimal recognition results for a given text collection. Unfortunately there were no corpora available at the time that could cover the requirements of the project, i.e. historic newspaper content, texts in the languages Dutch, French and German, and carrying sufficiently open licenses that would allow for the adaptation, extension and redistribution of such corpora. Accordingly, the project set out to create such corpora, and to release them into the public domain for take-up in the community.

In a first step, a selection of 100 pages representative of the overall newspaper titles available for each language was defined. Next to the selections made for Dutch and French, there were two distinct sets produced for German - one consisting of German language newspapers from the Dr Friedrich Teßmann library in South-Tyrol, Italy, and another one with newspapers from the Austrian National Library. An especially important criterion in the selection of suitable newspaper pages for the training data was the quality of the OCR: a threshold of 80 percent word accuracy was specified for those texts that were to be manually annotated.

Following a small-scale evaluation of software tools available for the manual annotation of texts with semantic markup such as named entities (Yimam et al., 2014; Stenetorp et al., 2012), the project decided to make use of a tool created at the Institute for Dutch Lexicology (INL), the INL Attestation Tool[4]. The main reason was that tagging texts with this tool has proven to be more time-efficient than using any of the other tools available. In addition, there was already ongoing collaboration with INL in other projects.

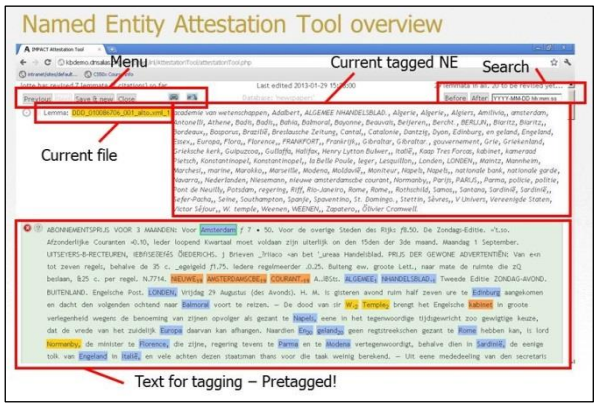

Figure 2: Screenshot of the INL Attestation Tool

The INL Attestation Tool is a LAMP-based web application with a graphical user interface that can be used in the production of computational lexica and gold standard data for named entity tagging. It can also be used to manually correct automatically executed tagging of named entities in texts. The tool internally uses a SQL database to store the text and its annotations.

The pages in German were pre-processed with the Stanford software using the readily available binary classifier for German[5]. The results with the derived initial tagging were then ingested to the annotation tool. As no classifiers were available for Dutch and French, the raw OCR texts without any tags were directly loaded into the annotation tool. After that, multiple users started adding annotations for the entity categories person, location and organization via the web-interface of the tool.

[3] http://www.catchplus.nl/

[4] Available from: https://github.com/INL/AttestationTool.

[5] http://www.nlpado.de/~sebastian/software/ner_german.shtml.

Some technical issues were encountered in deriving a valid representation of the texts from the ALTO files (including pre-tagging with NER) in the internal data model of the Attestation Tool in case there were already annotations present. These issues were eventually resolved with the help of INL.

	# tokens	# PER	# LOC	# ORG
French	207,000	5,672	5,614	2,574
Dutch	182,483	4,492	4,448	1,160
German	96,735	7,914	6,143	2,784

Table 1: Annotation statistics per language

The observation that significantly less entities of type "organization" were contained in the texts than "person" or "location" is no big surprise here: due to the historic nature of the material, simply not so many organisations were known at the time. However, above statistics need to be taken with a grain of salt: a deeper inspection of the annotated files during the writing of this paper revealed that sometimes not all entities occurring in the full text were indeed also annotated. This can be due to the annotators accidentally missing an entity (only 1 run was made per page) or, some of the existing annotations having been scraped during post-processing.

Once the manual tagging had been completed, the texts including their annotations were exported from the annotation tool in the ALTO format. Following discussions with the ALTO editorial board, a new feature was introduced in the ALTO standard version 2.1 that supports the valid mark-up and labelling of named entities directly in the ALTO files[6].

This is achieved by the use of "Tags", and more specifically, the "NamedEntityTag" element. The "NamedEntityTag" element can have the following attributes: "ID", "LABEL", "TYPE", "DESCRIPTION" and "URI". An important benefit of having the named entity information available directly in the ALTO files is that they are linked with the string coordinates of the scanned image, which makes retrieving their position much easier in presentation and retrieval contexts (especially in the case of newspaper pages with large image dimensions).

With the help of a Python script[7], the ALTO files were then transformed into the BIO file format (Tjong Kim Sang & Veenstra, 1999) required for the training of the Stanford software. Based on the BIO files, and, for the German set, suitable gazetteers, classifiers for the Stanford NER Recognizer have been trained and evaluated in a 4-fold cross-evaluation. The resulting BIO files, trained binary classifiers that were derived, as well as the source ALTO files have since been released under a Creative Commons Zero (CC0) license on the KB Research Lab[8], and also on GitHub[9].

4. Challenges

Next to the fact that three different languages had to be tackled within a common framework, there were two more key challenges that needed to be addressed:

1. The texts were produced using automated optical character recognition (OCR) technology which tends to produce a rather high degree of errors in the recognition of (especially historical) words.
2. The texts are of historic origin (i.e. mainly newspaper articles published before 1900), which means that considerable spelling variations compared to modern orthographic and grammar rules will occur.

To compensate for the sometimes poor OCR accuracy, an approach was tested where named entities were extracted from the OCR text, and spelling errors deliberately introduced (in the same way they are found in poor OCR results) into the training data. The intention was to cancel out the spelling errors found in the original text. This approach did not produce beneficial results, because the named entity recognizer that was trained on the corrupted training data did produce results that were slightly inferior to those that were obtained without this extra step.

Another experiment to compensate for OCR errors was carried out by including a spelling variation module that was developed by INL as part of the IMPACT project[10]. The assumption, upon which the spelling variation module is based, is that the Stanford NER module uses contextual text features such as capitalization, preceding and following words, and position in the sentence. For example, the sentences "I come from London", "I come fro London" and "I come frcm London" all have different words preceding the location "London" for the tool, although they are all (made up) variants of the word "from". Thus, the tool would benefit if these variations would be diminished, and this is what the spelling variation reduction module intends to do.

The module tries to reduce spelling variations on the input data by matching potential variants, creating internal rewrite rules and by executing these rewrite rules before the tool actually uses the input. The actual output remains unchanged. In the above example, it would identify the words "from", "fro" and "frc" as variants and create the rewrite rules "fro" → "from" and "frc" → "from". These

[6] See the summary of changes on the ALTO XML GitHub: https://github.com/altoxml/documentation/wiki/v2.1-Changes and the use case examples in the documentation http://altoxml.github.io/alto_tags_use_cases/ALTO_tags_usecases.html#named_entity_tagging.

[7] Available on the Europeana Newspapers GitHub: https://github.com/EuropeanaNewspapers/europeananp-ner/tree/master/alto_tools.

[8] http://lab.kbresearch.nl/static/html/eunews.html.

[9] https://github.com/EuropeanaNewspapers/ner-corpora.

[10] http://www.impact-project.eu/

rewrite rules are applied to the input data, the tool is run, and, in the case of extraction, the original text is used for output.

However, this approach had to be abandoned at an early stage as it turned out to be too complex to implement and it also required changes to some of the core classes of the Stanford software which turned out difficult to keep in sync with the main software development. Nevertheless, a fork that integrates the spelling variation functionality with the core Stanford NER tool was produced for testing purposes[11]. Tests on Dutch historical data have shown that the module is indeed capable of improving the scores by a few percent. However, having the proper rewrite rules is essential. More rules do not necessarily lead to better performance, due to the fact that they also lead to more incorrect variant matches.

5. Evaluation

Based on the BIO files, and (where available) suitable gazetteers, classifiers for the Stanford NER Recognizer were trained and subsequently evaluated in a 4-fold cross-evaluation.

Figure 3: Evaluation results of Dutch NER

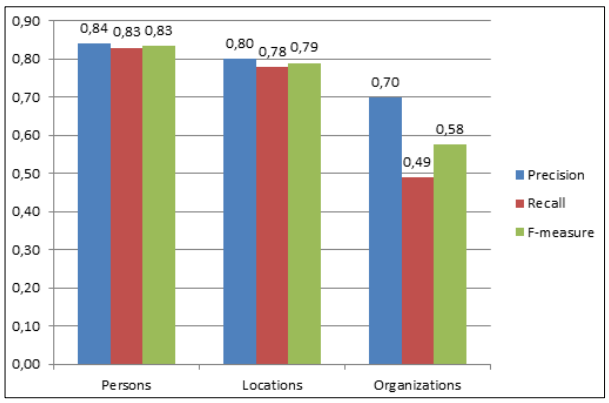

Figure 4: Evaluation results of French NER

[11] https://github.com/KBNLresearch/europeananp-ner/tree/inl.

Unfortunately, due to data corruption in the Attestation Tool's database, no exact evaluation figures could yet be derived for the German/Austrian set[12].

With precision scores for Dutch and French in the 80-90 percent range and recall at around 60-70 percent, the evaluation results do confirm that the Stanford NER system has a slightly "conservative" bias. This is in fact beneficial in this use case, where entities that are recognized shall subsequently be disambiguated and, in a semi-supervised step, linked to unique resource descriptions (e.g. authority files, DBpedia/Wikidata identifiers). The entities that are recognized have a high precision and can thus be mapped to their corresponding information resource more easily.

Further improvements to the recall rate could be achieved by the inclusion of additional features in the training, such as for example distributional similarity (Clark, 2003) or semantic generalization (Faruqui & Padò, 2010). General performance improvements can also be expected from the use of larger amounts of training material.

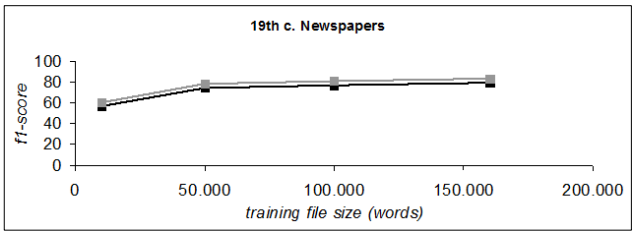

Figure 5: Effect of the size of the training material on recognition performance (f1-score)

6. Next Steps

Some additional works is still required to raise the quality and thus the usefulness of the derived corpus for scientific purposes. This includes some further clean-up and refinement of named entity tags, but also the enrichment of the dataset with more granular metadata (such as e.g. the OCR accuracy, date of publication of a newspaper text and URL to its digital facsimile).

Part of this work will be carried out in a small-scale pilot in the framework of the EUDAT project[13], where the semi-automatic disambiguation and matching of the extracted named entities to DBpedia identifiers will be tested. Thanks to the multilingual nature of DBpedia data, this would also allow for the cross-lingual linking of named entities (e.g. "Londres" → "London").

So far, there has been some experimental usage of the dataset for Dutch at the University of Utrecht, the KU

[12] A small-scale evaluation based on 11 out of 200 pages indicated precision across all entity categories in the range of 0,76-0,79 and recall between 0,66-0,69.

[13] https://www.eudat.eu/communities/enriching-europeana-newspapers.

Leuven and the KB National Library of the Netherlands (van Veen et al., 2015). The annotated data from the Dr Friedrich Teßmann Library has been used in the OPATCH project[14] at the European Academy, Bolzano. Further use, contributions and enhancements to the corpus from the library, digital humanities and computational linguistics communities are expected.

7. Conclusion

Alongside the large-scale OCR refinement of scanned European historic newspapers, the Europeana Newspapers project produced a corpus of 400 newspaper pages in Dutch/French/German manually selected and annotated with named entities. In addition, a number of software tools and technical resources were produced and/or enhanced for the purpose of having a starting ground for anyone working on the development or application of NER for historic newspaper content. Both, software and training resources were released under fully open licenses and are available to the community without any restrictions.

It is envisaged that due to their permissive licensing, these NER resources produced by Europeana Newspapers will alleviates the absence of training data for historical language material in Dutch, French and German, find applications in the NLP community, and will be further enhanced and extended to support the development and evaluation of named entity recognition systems for historical texts.

8. Acknowledgements

This work has been partially funded through the EU Competitiveness and Innovation Framework Programme grant Europeana Newspapers (Ref. 297380).

9. Bibliographical References

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, pp. 59--66. http://dx.doi.org/10.3115/1067807.1067817

Faruqui, M. and Padò, S. (2010). Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

Finke, J.R., Grenager, T. and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Stroudsburg, PA, USA, pp. 363--370. http://dx.doi.org/10.3115/1219840.1219885

Galibert, A., Rosset, S., Grouin, C., Zweigenbaum, P. and Quintard, L. (2012). Extended Named Entities Annotation on OCRed Documents: From Corpus Constitution to Evaluation Campaign. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, pp. 3126--3131.

Gooding, P. (2014). Exploring Usage of Digital Newspaper Archives through Web Log Analysis: A Case Study of Welsh Newspapers Online. Digital Humanities 2014, Lausanne, Switzerland.

Grover, C., Givon, S., Tobin, R. and Ball, J. (2008). Named Entity Recognition for Digitised Historical Texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, pp. 1343--1346.

Mosallam, Y., Abi-Haidar, A. and Ganascia, J.G. (2014). Unsupervised named entity recognition and disambiguation: An application to old French journals. In *Advances in Data Mining. Applications and Theoretical Aspects*, pp. 12--23. http://dx.doi.org/10.1007/978-3-319-08976-8_2

Neudecker, C., Faber, W.J., Wilms, L. and van Veen, T. (2014). Large scale refinement of digital historical newspapers with named entity recognition. In *Proceedings of the IFLA 2014 Newspaper Section Satellite Meeting*, Geneva, Switzerland.

Rodriguez, K.J., Bryant, M., Blanke, T. and Luszczynska, M. (2012). Comparison of named entity recognition tools for raw OCR text. In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, pp. 410--414.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. and Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*, Stroudsburg, PA, USA, pp. 102--107.

Tjong Kim Sang, E.F. and Veenstra, J. (1999). Representing text chunks. In *Proceedings of the 9th Conference on European chapter of the Association for Computational Linguistics (EACL '99)*, Stroudsburg, PA, USA, pp. 173--179. http://dx.doi.org/10.3115/977035.977059

Yimam, S.M., Eckart de Castilho, R., Gurevych, I. and Biemann C. (2014). Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, Baltimore, MD, USA, pp. 91--96.

Van Veen, T., Lonij, J. and Koppelaar, H. (2015). Semantic Enrichment: a Low-barrier Infrastructure and Proposal for Alignment. In *D-Lib Magazine, Volume 21, Number 7/8 (July 2015)*. http://dx.doi.org/10.1045/july2015-vanveen

10. Language Resource References

Europeana Newspapers Named Entities Corpus. (2015). Europeana Newspapers Project, https://github.com/EuropeanaNewspapers/ner-corpora.

[14] http://www.eurac.edu/en/research/projects/pages/projectdetails.aspx?pid=11263.