Comparison of Named Entity Recognition tools for raw OCR text

Kepa Joseba Rodriquez
Göttingen State and University Library
rodriguez@sub.uni-goettingen.de

Mike Bryant
Centre for e-Research King's College London
michael.bryant@kcl.ac.uk

Tobias Blanke
Centre for e-Research King's College London
tobias.blanke@kcl.ac.uk

Magdalena Luszczynska
University College London
zclhd0@ucl.ac.uk

Abstract

This short paper analyses an experiment comparing the efficacy of several Named Entity Recognition (NER) tools at extracting entities directly from the output of an optical character recognition (OCR) workflow. The authors present how they first created a set of test data, consisting of raw and corrected OCR output manually annotated with people, locations, and organizations. They then ran each of the NER tools against both raw and corrected OCR output, comparing the precision, recall, and F1 score against the manually annotated data.

1 Introduction[1]

While the amount of material being digitised and appearing online is steadily increasing, users' ability to find relevant material is constrained by the search and retrieval mechanisms available to them. Free text search engines, while undoubtedly of central importance, are often more effective and friendly to users when combined with structured classification schemes that allow faceted search and iterative narrowing-down of large result sets. Such structured classification, however, is costly in terms of resources. Named Entity Recognition (NER) holds the potential to lower this cost by using natural language processing tools to automatically tag items with people, places, and organizations that may be used as access points.

The European Holocaust Research Infrastructure (EHRI)[2] project aims to create a sustainable research infrastructure bringing together resources from dispersed historical archives across different countries. As part of this effort, we are investigating if and how we can enhance the research experience by offering tools to EHRI project partners that enable them to semantically enrich sources that they contribute to the wider infrastructure. A component of these services is currently planned to be a flexible OCR service based on the workflow tools developed at King's College London for the JISC-funded Ocropodium project (Blanke et al., 2011). While OCR itself is a complex process that produces hugely variable results based on the input material, modern tools are capable of producing reasonable transcripts from mid 20th-century typewritten material that is quite common among the EHRI project's primary sources. The quality of these transcripts prior to correction would be considered quite low for human readers, but even when uncorrected can offer value for search indexing and other machine processing purposes (Tanner et al., 2009). This experiment was undertaken to evaluate the efficacy of some available tools for accurately extracting semantic entities that could be used for automatically tagging and classifying documents based on this uncorrected OCR output.

Section 2 briefly summarises the existing literature covering the use of OCR and named entity extraction in the field of historical research. Section 3 describes our methodology for scoring each of the trialled NER tools in terms of pre-

[1] A version of this paper was presented in the proceedings of DH2012.

[2] http://www.ehri-project.eu

cision, recall, and F1. Section 4 describes the materials on which we conducted the experiment. Section 5 presents the results of our experiment and some analysis of those results, while section 6 concludes and discusses the research we intend to conduct in the future.

2 Literature Review

Grover et al. (2008) describe their evaluation of a custom rule-based NER system for person and place names on two sets of British Parliamentary records from the 17th and 19th centuries. They describe many of the problems encountered by NER systems that result from both OCR artefacts and the archaic nature of the sources themselves, such as conflation of marginal notes with body text, multi-line quoting rules, and capitalization of common nouns.

Packer et al. (2010) tested three different methods for the extraction of person names from noisy OCR output, scoring the results from each method against a hand-annotated reference. They noted a correlation between OCR word error rate (WER) and NER quality did exist, but was small, and hypothesised that errors in the OCR deriving from misunderstanding of page-level features (i.e. those that affect the ordering and context of words) have a greater impact on NER for person names than character-level accuracy.

3 Methodology

The goal of our experiment is the evaluation of the performance of existing NER tools on the output of an open source OCR system. We have selected four named entity extractors:

a) OpenNLP [3]
b) Stanford NER (Finkel et al., 2005)[4]
c) AlchemyAPI [5]
d) OpenCalais [6]

The entity types we chose to focus on were person (PER), location (LOC) and organization (ORG). These entity types are the most relevant for the extraction of metadata from documents in the domain of the EHRI project, and have a good coverage by the selected tools.

The named entity extractors use different tagsets for the annotation of named entities. We have selected the different categories and normalized the annotation as follows:

a) Stanford NER and OpenNLP: Person, location and organization categories have been annotated for all used models.
b) OpenCalais: we have merged the categories ”Country”, ”City” and ”NaturalFeature” into the category LOC, and the categories ”Organization” and ”Facility” into the category ORG. This second merging was possible because in our domain almost all facilities that appear were official buildings and public facility buildings referenced in an organizational capacity.
c) Alchemy: We have merged the categories ”Organization”, ”facility” and ”Company” into the category ORG and ”Country” , ”City” and ”Continent” into LOC.

To our knowledge there is at the moment no freely available corpus that can be used as gold standard for our purposes. For this reason we have produced a manually annotated resource with text extracted from images of documents related to the domain of the EHRI project as described in section 4. We have compared the output of the tools with this standard and computed the precision, recall and F1.

4 Experimental Setup

The material used for our initial experiments was obtained from the Wiener Library, London, and King’s College London’s Serving Soldier archive.

The Wiener Library material consisted of four individual Holocaust survivor testimonies, totalling seventeen pages of type-written monospaced text. Since the resolution of the scans was low and the images quite “noisy” (containing hand-written annotations and numerous scanning artifacts) we did not expect the quality of the OCR to be high, but felt they were suitably representative of material with which we expected to be dealing with in the EHRI project.

[3] http://opennlp.apache.org
[4] http://nlp.stanford.edu/software/CRF-NER.shtml
[5] http://www.alchemyapi.com
[6] http://www.opencalais.com

The OCR workflow used only open-source tools, employing several preprocessing tools from the Ocropus toolset (Breuel, 2008) for document cleanup, deskew, and binarisation, and Tesseract 3 (Smith, 2007) for character recognition. We also found that scaling the images to four times their original size using an anti-aliasing filter made a big positive difference to the quality of the OCR output. Ultimately, the word accuracy of the OCR averaged 88.6%, although the title pages of each survivor testimony, which resembles an archival finding aid with brief summaries of the contents in tabular form, were considerably worse, with only one exceeding 80% word accuracy. Character accuracy was somewhat higher, averaging 93.0% for the whole corpus (95.5% with the title pages excluded.)

The Serving Soldier material consisted of 33 newsletters written for the crew of H.M.S. Kelly in 1939. Like the Wiener Library documents, they also comprise type-written monospaced text, but have a strong yellow hue and often-significant warping and skew artifacts. The OCR workflow also used Ocropus for preprocessing and Tesseract 3 for character recognition, with most attention being paid to the thresholding. Word accuracy averaged 92.5% over the whole set, albeit with fairly high variance due to factors such as line-skew.

The corpus was constructed by manually correcting a copy of the OCR output text, in order to evaluate the impact of the noise produced by the OCR in the overal results. Both copies of the corpus, corrected and uncorrected, were then tokenized and POS tagged using the TreeTagger[7] (Schmid, 1994) and imported in the MMAX2 annotation tool (Müller and Strube, 2006), before being delivered to human annotators. In order to support the human annotation process we have used the POS tags to produce a pre-selection of annotation markable candidates, selecting all words tagged as PPN[8] and building markables with chains of these words that are not separated by punctuation marks.

We have tested the reliability of our annotation carrying an agreement study. For the study we have used the Kappa coefficient (Carletta, 1996) between two native English speaker annotators, and we get a value of $K = .93$[9].

The composition of the corpus has been summarized in table 1.

	Corpus WL		Corpus KCL	
	Raw	Corr	Raw	Corr
Files	17	17	33	33
Words	4415	4398	16982	15639
PER	75	83	82	80
LOC	60	63	170	178
ORG	13	13	52	60
TOTAL	148	159	305	319

Table 1: Composition of test data

		not corrected			corrected			p
		P	R	F1	P	R	F1	
AL	PER	.51	.36	.42	.46	.32	.38	-
	LOC	.92	.43	.59	.96	.51	.67	-
	ORG	.36	.31	.33	.50	.23	.32	-
	TOTAL	.61	.38	**.47**	.63	.38	**.48**	-
OC	PER	.69	.30	.42	.62	.18	.28	-
	LOC	.78	.20	.32	.76	.49	.60	-
	ORG	.40	.15	.22	.50	.15	.24	-
	TOTAL	.75	.29	**.41**	.69	.30	**.42**	-
ON	PER	.33	.04	.08	.50	.06	.10	-
	LOC	.56	.20	.29	.62	.25	.35	-
	ORG	.27	.23	.25	.20	.08	.11	-
	TOTAL	.42	.12	**.19**	.53	.13	**.21**	-
ST	PER	.58	.55	.56	.62	.62	.62	-
	LOC	.85	.57	.68	.84	.68	.75	*
	ORG	.09	.15	.11	.12	.23	.16	-
	TOTAL	.57	.52	**.54**	.60	.61	**.60**	*

Table 2: Output of NERs the Wiener Library dataset

5 Results

The results of the different NER tools have been summarized in table 2 for the Wiener Library's dataset and in table 3 for the King's College London's dataset[10]. The last row shows the statistical significance[11] of the differences between the performance of the NER tools on corrected and uncorrected text calculated globally and for each entity types. The stars indicate degrees of significance: one star signifies $0.025 \leq p \geq 0.05$; two stars $0.01 \leq p < 0.025$; three stars $p < 0.01$. We use a dash when differences are not statistically significant.

A first observation of the results for corrected and non-corrected text shows that the correction

[7] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

[8] Proper Name

[9] Values of $K > .8$ show that the annotation is reliable to draw definitive conclusions.

[10] The meaning of the abbreviations is: AL Alchemy, OC for OpenCalais, ON for OpenNLP, ST for Stanford NER.

[11] We used the paired t-test to compute the statistical significance.

		not corrected			corrected			p
		P	R	F1	P	R	F1	
AL	PER	.15	.22	.18	.24	.32	.27	***
	LOC	.63	.46	.53	.74	.59	.65	***
	ORG	.12	.16	.13	.24	.27	.25	**
	TOTAL	.31	.33	**.32**	.43	.44	**.44**	***
OC	PER	.25	.11	.15	.35	.16	.22	-
	LOC	.65	.50	.57	.66	.57	.61	-
	ORG	.14	.20	.16	.22	.27	.18	-
	TOTAL	.42	.33	**.37**	.46	.39	**.42**	-
ON	PER	.25	.13	.17	.29	.11	.16	-
	LOC	.65	.29	.40	.69	.33	.44	-
	ORG	.10	.24	.14	.14	.27	.18	-
	TOTAL	.28	.23	**.25**	.33	.25	**.28**	-
ST	PER	.18	.25	.21	.29	.37	.33	***
	LOC	.52	.71	.60	.53	.77	.62	-
	ORG	.08	.29	.13	.17	.48	.26	*
	TOTAL	.28	.50	**.36**	.35	.60	**.44**	***

Table 3: Output of NERs on the King College London dataset

of the text by hand does not increase the performance of the tools by a significant amount[12]. On the other side the improvement is not statistically significant in a reliable way: there is not an entity type for which the performance of the NER increases in a significant way for both datasets for the same NER tool.

In both cases the performance of systems is in the best case modest and some way below the performance reported in the evaluation of the tools. One reason for this is that the kind of text that we use is quite different to the texts usually used to train and evaluate NER systems[13].

For instance, a significant number of the non-extracted entities of type PER can be explained by their being represented in a variety of different ways in the source text, as for instance [Last name, first name], use of parentheses together with initials in the name - as for instance "Captain (D)" - sometimes written all in capital, or in other non standard ways, and these variants have often been missed by the NER tools. In some cases we feel that these omissions can potentially be resolved with a fairly straightforward heuristic pre-processing of the text.

Another difficult case is when the names of persons or locations are used to annotate other cases of entities. For instance in one of our data sets the warship with name "Kelly", which appears very often in all the files, has consistently been annotated incorrectly as PER, and the same occurs with the name of other warships.

Perhaps unsurprisingly, entities of type ORG proved the most difficult to extract. Organizations referenced in our test datasets (particularly the H.M.S. Kelly newsletters) tended to be once highly relevant organizations that no longer exist, and organizations with abbreviated names. A way to improve the results here could be to employ more external sources of knowledge that can enrich the output of the NE extractors.

An additional, more general, factor that makes NE extractors difficult is the high quantity of spelling errors and irregular spelling in the original files. OpenCalais tries to solve this problem using its knowledge to reconstruct the non-interpretable output. In several cases this reconstruction seems to be successful, but with the risk of introducing entities that don't appear in the text[14]. That suggests that automatically extracted entities should be validated using controlled vocabularies or other kinds of domain knowledge.

In terms of relative results of the various NER tools, a few tentative conclusions can be drawn. Stanford NER gave overall the best performance across both datasets, and was most effective on PER and LOC types. Alchemy API achieved the best results for the ORG type, particularly on manually corrected text, but was not markedly better than OpenCalais or Stanford NER with this data. OpenNLP performed the least accurately of the tools we tested.

6 Conclusions and Future Work

The experiment we describe here is very much linked to practical needs. At this stage we have plans for the implementation of a system that provides OCR and NER workflow facilities for EHRI project partners, and which we hope will provide real-world usage data to augment the results of our trials.

The results indicate that manual correction of OCR output does not significantly improve the

[12] Although we find a small improvement for some of the tools and entity types, it is not enough if we take in account the amount of ours necessary for the correction task. That is a highly resources consuming task that goes beyond the objectives of our project and can not be founded by mass digitization projects

[13] In the case of the (open source) Stanford NER and OpenNLP we know that a mixture of the MUC-6, MUC-7, and CoNLL were used as gold standard corpora. This information is not available for closed source webservices like OpenCalais and AlchemyAPI

[14] For instance the entity "Klan, Walter" was extracted as "Ku Klux Klan".

performance of named-entity extraction (though correction may of course be valuable for other IR purposes), and that there is scope for improving accuracy via some heuristic preprocessing and the use of domain knowledge.

In the near term our work will likely focus on employing simple heuristics and pattern-matching tools to attempt to improve the quality of the NER for our specific domain. Longer term, we intend to exploit the domain-specific knowledge generated by the EHRI project (in the form of authority files describing people, places, and organizations) to validate and enhance the output of automatic entity recognition tools.

References

T. Blanke, M. Bryant, and M. Hedges. Ocropodium: open source OCR for small-scale historical archives. *Journal of Information Science*, 38(1):76–86, November 2011. ISSN 0165-5515, 1741-6485. doi: 10.1177/0165551511429418. URL http://jis.sagepub.com/cgi/doi/10.1177/0165551511429418.

Thomas Breuel. The OCRopus open source OCR system. In *Proceedings IS&T/SPIE 20th Annual Symposium*, 2008. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.99.8505.

Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22(2):249–254, June 1996. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=230386.230390.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219885. URL http://dx.doi.org/10.3115/1219840.1219885.

Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. Named entity recognition for digitised historical texts. 2008. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.166.68.

Christoph Müller and Michael Strube. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany, 2006.

Thomas L. Packer, Joshua F. Lutes, Aaron P. Stewart, David W. Embley, Eric K. Ringger, Kevin D. Seppi, and Lee S. Jensen. Extracting person names from diverse and noisy OCR text. page 19. ACM Press, 2010. ISBN 9781450303767. doi: 10.1145/1871840.1871845. URL http://portal.acm.org/citation.cfm?doid=1871840.1871845.

Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Ray Smith. An overview of the tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, 2007. URL http://dl.acm.org/citation.cfm?id=1304846.

Simon Tanner, Trevor Muoz, and Pich Hemy Ros. Measuring mass text digitization quality and usefulness. *D-Lib Magazine*, 15(7/8), July 2009. ISSN 1082-9873. doi: 10.1045/july2009-munoz. URL http://www.dlib.org/dlib/july09/munoz/07munoz.html.