A Tool for Pseudonymization of Textual Documents for Digital Humanities Research and Publication

Arttu Oksanen[1,2], Minna Tamper[2], Eero Hyvönen[2,3], Jouni Tuominen[2,3,4], Henna Ylimaa[5], Katja Löytynoja[5], Matti Kokkonen[5] and Aki Hietanen[6]

[1]*Edita Publishing Ltd.*
[2]*Aalto University, Dept. of Computer Science*
[3]*University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG)*
[4]*University of Helsinki, Helsinki Institute for Social Sciences and Humanities (HSSH)*
[5]*Statistics Finland*
[6]*Ministry of Justice in Finland*

Abstract

Manual pseudonymization of documents is a costly and time-consuming procedure that can be automated by applying natural language processing methods. This paper introduces the Anoppi tool developed for automatic pseudonymization of Finnish texts. The tool can be used both as a web application and programmatically through a REST API. Evaluation shows that Anoppi performs well with different types of documents, however, further improving the performance of the named entity recognition and disambiguation methods would enhance the usefulness of the software and motivate organizations to bring Anoppi into use.

Keywords

pseudonymization, anonymization, named entity recognition

1. Introduction

Many Cultural Heritage (CH) texts and legal documents of interest to a wider audience, such as interviews of people and court decisions, contain sensitive personal data. This makes it difficult to publish and use them for research given the EU General Data Protection Regulation (GDPR)[1], unless personal data contained is disguised. However, manual pseudonymization or anonymization of the documents is a costly and time-consuming procedure.

This paper presents the Anoppi tool and web service for automatic and semi-automatic pseudonymization of Finnish documents. Utilizing both machine learning (ML) and rule-based named entity recognition methods (NER) and morphological analysis, Anoppi is able to automatically or semi-automatically pseudonymize documents written in Finnish while preserving their readability and layout. Anoppi is the first pseudonymization tool developed

DHNB 2022: the 6th Digital Humanities in the Nordic and Baltic Countries Conference, March 15–18, 2022, Uppsala, Sweden

arttu.oksanen@aalto.fi (A. Oksanen); minna.tamper@aalto.fi (M. Tamper); eero.hyvonen@aalto.fi (E. Hyvönen); jouni.tuominen@helsinki.fi (J. Tuominen); katja.loytynoja@stat.fi (K. Löytynoja); matti.kokkonen@stat.fi (M. Kokkonen); aki.hietanen@om.fi (A. Hietanen)

[1]https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/

for Finnish. The tool was initially developed for automatic pseudonymization of court decisions [1] in the Anoppi project[2] funded by the Ministry of Justice in Finland but can and has been used in other contexts, too.

This paper first introduces the Anoppi tool and the underlying method for automatic pseudonymization. Then the results, obtained using Anoppi to pseudonymize investigation reports and court decisions, are presented. Finally we conclude with discussion on related work and future improvements.

2. Anoppi Tool

The Anoppi tool consists of a language analysis component (LAC) and a web-based user interface (WUI). LAC performs the actual automatic pseudonymization by carrying out NER and disambiguation on the document while WUI allows the user to modify the result of the automatic pseudonymization. LAC can also be used programmatically through a REST interface[3], enabling integration of Anoppi into external software systems.

LAC REST API can be used either as an annotation tool or as a full pseudonymization pipeline. The annotation tool returns an annotated version of the text with mentions of the named entities highlighted using special tags along with a list of suggested pseudonyms for each of the named entities, whereas the pseudonymization pipeline returns only the completely pseudonymized version of the text.

For named entity recognition Anoppi uses a combination of ML and rule-based methods and eventually combines the results obtained from different methods. The ML-based NER component is a model trained using FinBERT NER[4] [2]. The rule-based methods include regular expressions (to find specific surface forms such as property codes and vehicle registration plates) and dictionaries such as the Finnish person name ontology[3].

The pseudonymization is done by replacing the recognized names with grouped sequential identifiers, such as 'person A' or 'place B', keeping track of the entities throughout the text. By performing morphological analysis on the original text the software is also able to inflect the generated pseudonyms correctly to improve the readability of the pseudonymized text. Turku Neural Parser[4] is used to perform morphological analysis of the text and UralicNLP[5][5] to correctly inflect the derived pseudonyms.

LAC aims to maximize recall in the named entity recognition phase because it is easier for the user of WUI to delete suggested entities than to manually pick new entities from the text, if LAC did not automatically recognize them. It is also possible to restrict pseudonymization to certain types of entities (for example only person names). Further fine-tuning can be done by whitelisting certain names and contexts, for example to keep names of the judges overt in court judgments.

The WUI is a separate web-based WYSIWYG type editor tool. User uploads a document in Word format to the application and the LAC performs the NER phase in the background. Once

[2] https://seco.cs.aalto.fi/projects/anoppi/
[3] https://nlp.ldf.fi/anoppi
[4] https://turkunlp.org/fin-ner.html
[5] https://uralicnlp.com/

the NER phase is complete the user can review the resulting annotations using the WUI and correct them, preview the resulting pseudonymized version of the text and finally export it in Word format with the original layout preserved.

3. Evaluation

To test the performance of the LAC in pseudonymization of person names we created a tailor-made test dataset that consists of text in Finnish with person names added in grammatically appropriate places. The names were handpicked from the Population Information System data with emphasis on selecting both common and rare names. Attention was paid to select names including both traditional Finnish names and foreign names as well as names with two parts and names with a common meaning, for example Karhu (bear). Investigation reports from Safety Investigation Authority of Finland (SIAF) were used as base text for the test data. These reports do not originally contain any names but only references to people in the form of pronouns and job titles that were replaced randomly by the selected names and their combinations.

Eventually the test data contained 152 added names and name combinations from which Anoppi identified and pseudonymized 136 (89,5 %) correctly. In total, Anoppi identified 141 names or name combinations from the test data, but five of them were false positive words. A total of 16 names remained unidentified. However, most of the unidentified names were located in parts of the text where names are not common which would not be an issue with real data.

In addition to evaluating the performance of the LAC, the Anoppi tool as a whole has been evaluated by measuring and comparing the time it takes to pseudonymize a court decision both semi-automatically using the WUI and manually using only word processor software. The results obtained in this manner so far show that on average it takes about half the time to pseudonymize a court order using Anoppi as compared to manual pseudonymization. Anoppi makes some mistakes especially by confusing person names with place names and vice versa and spotting and correcting all of these incorrect categories using the WUI slows down the process. Moreover, in order to verify the correctness of the pseudonymization result a human expert still has to get acquainted with the content of the document regardless of the pseudonymization being automatic.

4. Related Works and Discussion

Automatic or semi-automatic pseudonymization methods are already in use in several European judicial systems [6]. Recent projects similar to ours focusing on automatic pseudonymization of court orders using ML methods possibly in combination with rule-based ones have been conducted for example in Poland, Austria, Germany, Latvia and France[6].

Evaluation of Anoppi shows promising results in locating the names of persons, organizations, places, and different types of identifiers of specific form. Still, it is difficult to build a general solution for pseudonymization as the sufficiency of de-identification varies in each case. The

[6]Based on oral presentations at https://ec.europa.eu/info/policies/justice-and-fundamental-rights/digitalisation-justice/conferences-and-events_en#webinarsontheuseofartificialintelligenceinthejusticefield (1st Webinar, 26 and 29 March 2021)

category-based selection of named entities used in the current model is not sufficient if for example names of small companies should be pseudonymized but large ones should not. Another issue in the Anoppi project is the lack of task-specific training data as we are not able to store and make use of real production data in order to continuously train ML models due to restrictions imposed by the GDPR. That is why we ended up using a general NER model for Finnish language along with configurable case-based rules.

The Anoppi service is currently in pilot testing in the Ministry of Justice of Finland for pseudonymization of Finnish court decisions in order to make them available on the Web and for data analysis in the forth-coming public LawSampo data service and portal for publishing and studying Finnish legislation and case law. Future work focuses on further improving the performance of the named entity recognition algorithm and including identification of new entity types such as rare diseases or unique job titles that make re-identification of people straightforward.

Acknowledgments This work is part of Finnish AI special funding program by the Ministry of Finance, for experiments that promote productivity. CSC – IT Center for Science, Finland provided computational resources.

References

[1] A. Oksanen, M. Tamper, J. Tuominen, A. Hietanen, E. Hyvönen, Anoppi: A pseudonymization service for Finnish court documents, in: Legal Knowledge and Information Systems. JURIX 2019: The Thirty-second Annual Conference (Araszkiewicz, M. and Rodríguez-Doncel, V. (eds.)), IOS Press, 2019, pp. 251–254.

[2] J. Luoma, M. Oinonen, M. Pyykönen, V. Laippala, S. Pyysalo, A broad-coverage corpus for Finnish named entity recognition, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4615–4624. URL: https://aclanthology.org/2020.lrec-1.567.

[3] M. Tamper, P. Leskinen, J. Tuominen, E. Hyvönen, Modeling and publishing finnish person names as a linked open data ontology, in: 3rd Workshop on Humanities in the Semantic Web (WHiSe 2020), CEUR Workshop Proceedings, vol. 2695, 2020, pp. 3–14. URL: http://ceur-ws.org/Vol-2695/paper1.pdf.

[4] J. Kanerva, F. Ginter, N. Miekka, A. Leino, T. Salakoski, Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task, in: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, 2018.

[5] M. Hämäläinen, UralicNLP: An NLP library for uralic languages, Journal of Open Source Software 4 (2019) 1345. doi:10.21105/joss.01345.

[6] M. van Opijnen, G. Peruginelli, E. Kefali, M. Palmirani, On-Line Publication of Court Decisions in the EU: Report of the Policy Group of the Project 'Building on the European Case Law Identifier', 2017. Available at SSRN: https://ssrn.com/abstract=3088495.