Oksanen, Arttu; Tamper, Minna ; Tuominen, Jouni; Hietanen, Aki; Hyvönen, Eero

ANOPPI

Published: 01/01/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Oksanen, A., Tamper, M., Tuominen, J., Hietanen, A., & Hyvönen, E. (2019). *ANOPPI: A Pseudonymization Service for Finnish Court Documents*. 251-254. Paper presented at International Conference on Legal Knowledge and Information Systems, Madrid, Spain.

September 2019

ANOPPI: A Pseudonymization Service for Finnish Court Documents

Arttu OKSANEN [a,c], Minna TAMPER [a], Jouni TUOMINEN [a,b]
Aki HIETANEN [d], and Eero HYVÖNEN [a,b]

[a] *Semantic Computing Research Group (SeCo), Aalto University, Finland*
http://seco.cs.aalto.fi, firstname.lastname@aalto.fi
[b] *HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki*
http://heldig.fi
[c] *Edita Publishing Ltd.*
http://www.editapublishing.fi
[d] *Ministry of Justice, Finland*
http://oikeusministerio.fi, firstname.lastname@om.fi

Abstract. To comply with the EU General Data Protection Regulation (GDPR) publishing court judgments online requires that personal data contained in them must be disguised. However, anonymizing the documents manually is a costly and time-consuming procedure. This paper presents ANOPPI service for automatic and semi-automatic pseudonymization of Finnish court judgments. Utilizing both statistics- and rule-based named entity recognition methods and morphological analysis, ANOPPI is able to automatically pseudonymize documents written in Finnish preserving their readability and layout. The service is currently still in development but pilot tests are going to be carried out in Finnish courts in 2020.

Keywords. automatic pseudonymization, case law, named entity recognition

1. Introduction

Publishing court decisions openly on the web, either as human-readable documents or machine-readable data, enhances the legal protection of citizens by making the administration of justice more transparent. Open electronic access to case law can also be useful to decision-making and research concerning legal practice. In Finland, case law is published publicly online as HTML documents in the Finlex data bank[1] [3] and as linked open data in the Semantic Finlex service[2] [4].

Unfortunately, due to issues of data protection and privacy and the requirement to comply with the EU General Data Protection Regulation (GDPR), currently only a minor part of all the Finnish court judgments is published online. For example, currently none of the judgments of the district courts are available. Pub-

[1] http://www.finlex.fi
[2] http://data.finlex.fi

lishing court judgments online requires that the documents are pseudonymized so that identifying named entities appearing in the document, such as persons, companies and geographical locations, are replaced with referent identifiers. However, currently all of the pseudonymization work is done manually in the courts by experts which is costly and time-consuming. Therefore a tool that automates the process of pseudonymization is highly desired.

This paper presents ANOPPI, a web service for semi-automatic pseudonymization of documents written in Finnish. In the on-going project, we are focusing on case law documents as a first use case. However, the purpose of the ANOPPI service is to be a general-purpose domain-agnostic pseudonymization tool. The service is currently still under development, but a first demonstrator has already been created, and pilot tests will be carried out in Finnish courts in 2020. The source code will be published with an open license once the service is ready to be brought into real use. We will discuss the underlying ideas of the service as well as the first demonstrator in more detail in the following sections, starting with a description of the pseudonymization method in Section 2, followed by an overview of the application user interface in Section 3, and finally concluding with related work and discussion in Section 4.

2. Pseudonymization Method

The court orders are available in electronic format either as plain text, XML, HTML, or DOCX files. Based on [7], we have developed a tool that is able to find the named entities from these documents and annotate the occurrences of the named entities with special tags. The tool can be used as a RESTful web service that takes as input the document and produces as output the annotated document with a separate list of all the named entities found in the document.

To find the named entities the tool uses multiple different named entity recognizers and combines the results from those. First of all we use ready-made statistics- and rule-based named entity recognition (NER) software such as FiNER[3], a rule-based named entity recognizer for Finnish language, and Stanford NER [6]. Secondly, we have developed our own set of regular expression patterns to recognize things such as vehicle registration plates and property identifiers. In addition, we use an all-inclusive Finnish person name ontology that is based on the open data published by the Population Register Centre[4] to look up person names appearing in the court cases. Finally, we use the Finnish dependency parser [2] to support deciding if a term appearing in the text is a name.

After finding the named entities and their occurrences in the text the occurrences are replaced with pseudonyms. To assign a reasonable pseudonym for a given named entity its category must also be resolved. For example, we must be able to differentiate towns and corporations so that a pseudonym can be correctly determined as either "town A" or "corporation A". Categorical disambiguation is based on a scoring scheme that weighs the results obtained from the different named entity recognizers.

[3] `https://github.com/Traubert/FiNer-rules/blob/master/finer-readme.md`
[4] `https://vrk.fi/en/`

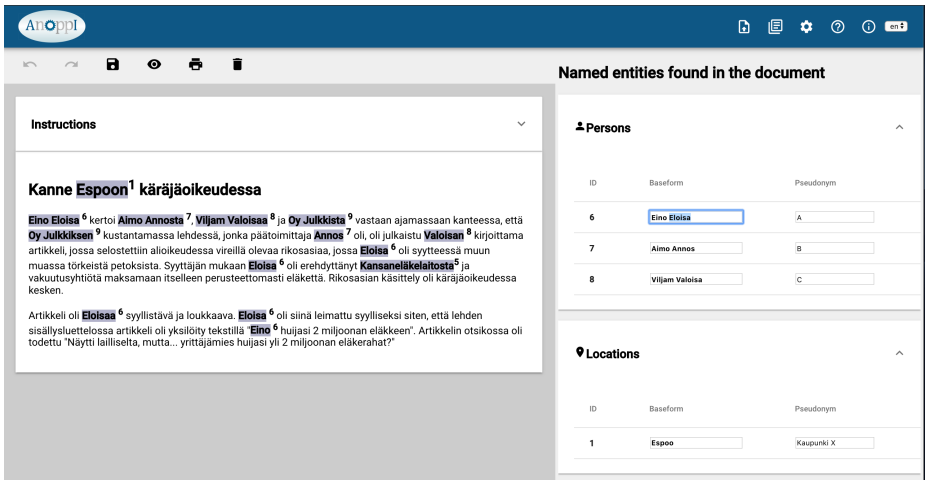

Figure 1. User interface of the ANOPPI application.

As Finnish is a highly inflected language we must also derive the correct inflected form for the pseudonym so that the pseudonymized text stays readable. To achieve this, we use morphological analysis to be able to distinguish, for example, the case and possessive suffix of a noun.

3. User Interface

As we do not expect the result of the automatic pseudonymization to be perfect, a web-based user interface, shown in Figure 1, is provided where the user of the service can make further modifications to the proposed named entities and their pseudonyms. The text with identified entities is shown on the left. On the right hand column, the user interface allows the user to add new named entities and also edit and remove the existing ones. In addition, it is possible to remove entire phrases from the text if de-identification requires it. Once the editing is complete, the user can preview and export the resulting document that aside from the pseudonyms and text removals should be identical to the original one.

Evaluation of the user interface is underway by usability tests, first within the project team and later in 2019 and 2020 in the courts where the service is eventually going to be brought into use.

4. Related Work and Discussion

Automatic or computer-aided pseudonymization is already utilized in judiciaries of various European countries [8]. As an example, in Denmark an anonymization tool for court orders was implemented using solely manually crafted grammar rules to find the named entities in the texts [5]. On-going development projects similar to ours, in which the focus of the automatic pseudonymization is on court

orders and where machine learning-based methods are used, are being carried out in France and Austria[5].

For the moment, the Finnish public sector utilizes hardly at all automatic anonymization or pseudonymization tools, and it is difficult to evaluate the sufficiency of de-identification for different types of data and requirements [1]. ANOPPI aims to change the situation by enabling organizations to deploy automatic pseudonymization in their processes cost-effectively using an open source solution. However, the usefulness of the ANOPPI service will eventually be largely dependent on the precision and recall of the NER methods as well as the applicability of the user interface. Edita Publishing Ltd. has previously estimated that on average it takes approximately 38 minutes to pseudonymize a precedent of the Supreme Court manually. In order for ANOPPI to be successful, pseudonymization of the precedents using the service should be more efficient.

Acknowledgments This work was funded by the the Ministry of Justice in Finland. We thank Saara Packalén, Tiina Husso, and Oili Salminen of the Ministry of Justice, and Risto Talo, Jari Linhala, and Sari Korhonen of Edita Publishing Ltd. for collaboration. CSC – IT Center for Science, Finland, provided us with computational resources.

References

[1] A. Bäck and J. Keränen. Anonymisointipalvelut. tarve ja toteutusvaihtoehdot, 2017. Liikenne- ja viestintäministeriön julkaisuja 7/2017.

[2] K. Haverinen, J. Nyblom, T. Viljanen, V. Laippala, S. Kohonen, A. Missilä, S. Ojala, T. Salakoski, and F. Ginter. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48:493–531, 2014. Open access.

[3] A. Hietanen. Free access to legislation in Finland: Principles, practices and prospects. In M. Ragona, editor, *Law via the Internet. Free Access – Quality of Information – Effectiveness of Rights*. European Press Academic Publishing, Florence, 2009.

[4] A. Oksanen, J. Tuominen, E. Mäkelä, M. Tamper, A. Hietanen, and E. Hyvönen. Semantic Finlex: Transforming, publishing, and using Finnish legislation and case law as linked open data on the web. In G. Peruginelli and S. Faro, editors, *Knowledge of the Law in the Big Data Age*, volume 317 of *Frontiers in Artificial Intelligence and Applications*, pages 212–228. IOS Press, 2019.

[5] C. Povlsen, B. Jongejan, D. H. Hansen, and B. K. Simonsen. Anonymization of court orders. In *11th Iberian Conference on Information Systems and Technologies (CISTI)*, Las Palmas, Spain, June 2016. IEEE.

[6] J. Rose Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 363–370, January 2005.

[7] M. Tamper, E. Hyvönen, and P. Leskinen. Visualizing and analyzing networks of named entities in biographical dictionaries for digital humanities research. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICling 2019)*. Springer-Verlag, April 2019. Forthcoming.

[8] M. van Opijnen, G. Peruginelli, E. Kefali, and 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', February 2017. Available at SSRN: https://ssrn.com/abstract=3088495.

[5] Based on oral presentations at `https://eu2019.fi/en/events/2019-09-05/workshop-anonymisation-of-court-judgements-challenges-and-solutions`