Legal Knowledge and Information Systems
S. Villata et al. (Eds.)

doi:10.3233/FAIA200869

Automatic Removal of Identifying Information in Official EU Languages for Public Administrations: The MAPA Project

Lucie GIANOLA [a,1], Ēriks AJAUSKS [b], Victoria ARRANZ [c], Chomicha BENDAHMAN [c], Laurent BIÉ [d], Claudia BORG [e], Aleix CERDÀ [d], Khalid CHOUKRI [c], Montse CUADROS [f], Ona DE GIBERT [g], Hans DEGROOTE [d], Elena EDELMAN [c], Thierry ETCHEGOYHEN [f], Ángela FRANCO TORRES [d], Mercedes GARCÍA HERNANDEZ [d], Aitor GARCÍA PABLOS [f], Albert GATT [e], Cyril GROUIN [a], Manuel HERRANZ [d], Alejandro Adolfo KOHAN [d], Thomas LAVERGNE [a], Maite MELERO [g], Patrick PAROUBEK [a], Mickaël RIGAULT [c], Mike ROSNER [e], Roberts ROZIS [b], Lonneke VAN DER PLAS [e], Rinalds VĪKSNA [b] and Pierre ZWEIGENBAUM [a]

[a] *Université Paris Saclay, CNRS, LIMSI, Orsay, France*
[b] *Tilde, Riga, Latvia*
[c] *ELDA/ELRA, Paris, France*
[d] *Pangeanic – PangeaMT, Valencia, Spain*
[e] *University of Malta, Msida, Malta*
[f] *Vicomtech, Donostia, Gipuzcoa, Spain*
[g] *Barcelona Supercomputing Center, Barcelona, Spain*

AbstractThe European MAPA (Multilingual Anonymisation for Public Administrations) project aims at developing an open-source solution for automatic de-identification of medical and legal documents. We introduce here the context, partners and aims of the project, and report on preliminary results.

Keywords. automatic de-identification, legal documents, open-source, multingual

1. Introduction

Interpreting European guidelines for data sharing implies to resolve the conflicting objectives stated, on the one hand in the PSI (Public sector information) directive which encourages administrations to share as much data as possible for re-use in an open-data perspective, and on the other hand, in the General Data Protection Regulation (GDPR), which requires the protection of personal data. As the GDPR becomes an obstacle to data sharing, removing personal information allows to share data. Nevertheless, removing identifying information from documents is a challenge that public administrations (PA) face in order to fulfill their open-data commitment, in every European language.

[1]Corresponding Author: Lucie Gianola, Université Paris Saclay, CNRS, LIMSI; lucie.gianola@limsi.fr

2. Context & objectives

Multilingual Anonymisation for Public Administrations (MAPA) is a project[2] funded by the Connecting Europe Facility (CEF)[3]. Academic and industrial partners from four countries are working together to develop, test and evaluate an open-source de-identification toolkit, fully customizable by end users for legal and medical domains, in all official European Union (EU) languages. The project aims to improve data sharing opportunities for PA.

De-identification consists in hiding directly-identifying information items: name, date of birth, address, contact information, etc. [1]. Anonymization consists in making it impossible to find out who a document is about. Text anonymization is extremely difficult to achieve automatically because the facts discussed in a document may be sufficiently eloquent to reveal indirectly the identity of the involved persons, for example in high-profile criminal cases where the general public is familiar with the broad outlines of the case. With medical documents, it is sometimes the mention of rare diseases combined with other criteria that makes identification possible. Both operations conflict with the need to maintain the legal relevance [2] of the document: for example, concealing all references to legal texts invoked in a judgment may compromise its logical structure, and concealing the facts discussed may render the text incomprehensible.

The project targets de-identification because even though it is not considered as effective as anonymisation, as a minimal approach it is sufficient in many cases, technically more achievable, and it can be evaluated more formally [3]. The most straightforward way to de-identify a document is to remove all identifying data such as names, addresses, phone numbers, etc., while retaining as much as possible from the original material: otherwise what will be left will be useless, in particular for Artificial Intelligence or Natural Language Processing (NLP). It is important to replace the removed language elements by something that hints at their type (e.g., someone's name with an identifier like Person), in a consistent fashion throughout the document (e.g. if several people are mentioned, all the text spans referring to their respective names ought to be replaced by the same identifier, all occurrences of Mr Doe, John Doe, John, etc., ought to be replaced by Person_1), to preserve the internal logic of the original document. NLP has seen the emergence of the concept of Named Entity for Information Extraction applications, which has been popularized among others by the DARPA/NIST MUC series of evaluation campaigns on natural language understanding, where they appeared in the 6th venue in 1995 [4]. The entity extraction process relies on a NLP method called *Named Entity Recognition* (NER), which spots in a text all mentions of information elements of pre-defined types, such as person names, dates, etc. While the de-identification of medical documents has already been the subject of much research [5], the de-identification of legal documents has received less attention so far [2,6].

Developing a de-identification system requires to define the different types of language entities potentially subjected to removal, as well as to annotate a sufficient amount of documents by hand, in order to obtain the training material required by the neural Machine Learning approaches that have become state of the art in NLP. The annotation process requires writing precise annotation guidelines that explain to human annotators how to identify and classify the relevant text elements.

[2] https://mapa-project.eu/

[3] https://ec.europa.eu/inea/en/connecting-europe-facility

3. Preliminary tests

The project foresees the use of word-embeddings trained through manually annotated examples. As proof of concept, we trained a model by fine-tuning BERT multilingual embeddings [7]. This has been done on a dataset combining data from the CoNLL Named Entity shared tasks: 2002 (Spanish) [8] and 2003 (English) [9]. As a result, this model allows to perform named entity recognition on four basic entity types (Location, Organization, Person, and Miscellaneous). The preliminary evaluation results were 93.4% of weighted F-score for entity recognition and classification among entity types PER, LOC, ORG, and MISC, and 97.98% of weighed F-score for binary entity classification, i.e., deciding whether or not words must be de-identified. We ran an experiment on the same sentence available in 23 official languages (translation in Gaelic language is missing):

> *whereas the founder of Charter 97, Aleh Byabenin, was found hanged at his home near Minsk in September 2010; whereas Belarus-born Pavel Sheremet, a spokesperson for the organisation behind Charter 97, was killed in a car bombing in Kiev, the capital of Ukraine, in July 2016;*

This extract from a European joint motion for a resolution[4] contains 11 named entities of several types (Date, Location, Person, Organization, as well as nationality and job). Note that for person and location names, the translations take into account the writing form used in each language (French: "Pavel Cheremet", German: "Pawel Scheremet", including declination forms: "Pavelas Šeremetas" for Latvian). Although the system was only trained on English and Spanish data, it succeeded in identifying named entities in all languages.

4. Annotation scheme

The annotation guidelines define six implicit top-level entity types: Person, Time, Location, Organisation, Amount, Vehicle. These entities encapsulate other explicit or implicit Level 2 entity types: a Person entity will contain Name, Age, Profession, etc. Level 2 entities are themselves made up of components and types, which are always explicit: a Name entity may contain a Title, Given name, Family name, etc. In order to make the annotation work easier, implicit entities are inferred from either their Level-2 entities or their Level-3 components/types. As the project deals with documents from two fields of expertise (legal and medical), annotation modalities need to be adapted to be relevant for both domains. This is in line with Nadeau et Sekine [10] who point out that the performance of NER improves when the domain and textual genre are considered. Given that we deal with texts from institutions, the use of a "role" tag is intended to annotate the "side" on which the mentioned person stands: from the institution (carers, members of court) or from the public (patient, plaintiff, defendant). Vehicle has been added as a Level 1 and 2 entity type (and its components: License plate number, Colour, Model, etc.) since it may be identifying in legal texts. Manual annotation tests were carried out on a corpus that includes, for each language, 12 documents of European case law[5], totalling approximately 2,000 sentences, via the INCePTION platform [11]. A baseline

[4] `https://www.europarl.europa.eu/doceo/document/RC-8-2018-0451_FR.html`
[5] `https://eur-lex.europa.eu/`

has been established from the first annotated datasets (Spanish, French, Croatian, Latvian, Romanian), and two models, monolingual and multilingual, have been trained. The multilingual model is slightly outperformed by the monolingual model (0,82 F1 to 0,85 F1 for French data). Discrepancies were discovered in the annotations across languages, which have allowed to adjust both annotations and guidelines.

5. Conclusion & perspectives

As the need for de-identifying texts will continue to grow in the European context, so will the need for automatic and multilingual de-identification solutions. To this end, we plan to evaluate its results not only in terms of performance (precision, recall, F-score), but also in terms of adaptability across text genres. Indeed, medicine and law produce very diverse types of texts (e.g., in the legal domain, case-law is very different from police interviews; see Zweigenbaum et al [12] for the medical domain). This remark will be essential if the application of the tool is to be extended to other administrations (social, agricultural, financial, etc.). We also rely on the collaboration of PA to test the solution during its development to ensure the best possible fit with the needs of end users.

References

[1] Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. J Med Internet Res. 2019 May 31;21(5):e13484.

[2] Plamondon L, Lapalme G, Pelletier F. Anonymisation de décisions de justice. In: 11e Conférence sur le Traitement Automatique des Langues Naturelles. Fès, Morocco: Bernard Bel et Isabelle Martin (eds); 2004. p. 367–376.

[3] Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology. 2010 Aug;10(1):70. Available from: https://doi.org/10.1186/1471-2288-10-70.

[4] Grishman R, Sundheim B. Design of the MUC-6 evaluation. In: Proceedings of the 6th conference on Message understanding. Stroudsburg, PA: Association for Computational Linguistics; 1995. p. 1–11.

[5] Uzuner O, Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. Journal of the American Medical Informatics Association. 2007;14:550–563.

[6] Tamper M, Oksanen A, Tuominen J, Hyvönen E, Hietanen A. Anonymization Service for Finnish Case Law: Opening Data without Sacrificing Data Protection and Privacy of Citizens. In: International Conference on Law via the Internet, LVI; 2018. .

[7] Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR. 2018;abs/1810.04805.

[8] Tjong Kim Sang EF. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: COLING-02: The 6th Conference on Natural Language Learning; 2002. .

[9] Tjong Kim Sang EF, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003; 2003. p. 142–147.

[10] Nadeau D, Sekine S. A survey of named entity recognition and classification. Linguisticae Investigationes. Linguisticae Investigationes. 2007;30(1):3–26.

[11] Klie JC, Bugert M, Boullosa B, Eckart de Castilho R, Gurevych I. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. Santa Fe, NM: ACL; 2018. p. 5–9.

[12] Zweigenbaum P, Jacquemart P, Grabar N, Habert B. Building a Text Corpus for Representing the Variety of Medical Language. In: Patel VL, Rogers R, Haux R, editors. Proceedings of the 10th World Congress on Medical Informatics. London, UK; 2001. p. 290–294.