CALL FOR PAPERS

Workshop on

Multilingual de-identification of (sensitive) language resources

To be held in conjunction with the 13th International Language Resources and Evaluation Conference (LREC 2022)

20 June 2022, Le Palais du Pharo, Marseille, France

http://workshops.elda.org/XXXX/ (under construction)

Deadline for submission: 10 April 2022

Description

The General Data Protection Regulation (GDPR - Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016) ensures the protection of natural persons with regard to the processing of personal data and on the free movement of such data. The GDPR outlines a specific set of rules that protect citizens and user data and create transparency in information sharing. GDPR is the strictest data privacy regulation in the world, and considerable work is taking place to develop techniques and deploy systems that help comply with this regulation while rendering data accessible and, thus, usable for further processing.

Different techniques are studied to guarantee such compliance, implying different levels of sensitive content protection and with a short- or long-term guarantee depending on whether we may have access to additional related information. In this regard, we can read about work on anonymization, de-identification and pseudonymization. While anonymization implies a zero re-identification risk, which is extremely difficult to secure, de-identification and pseudonymization represent an attainable target under the GDPR, given that this regulation defines pseudonymization as "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."

Bearing this context in mind, multilingual approaches and kits for (sensitive) language resources de-identification may provide the means to share language data while also protecting private or sensitive data by spotting then deleting, obfuscating, pseudonymizing or encrypting person identifying information.

De-identification is typically performed for the purpose of protecting an individual's private activities while maintaining the usefulness of the gathered data for research and development purposes. This workshop aims at discussing the various approaches to effective and reliable text de-identification, focusing on some sensitive domains such as the medical and legal domains, but not only.

Based on these premises a consensus emerges that shows a clear situation and needs:

1. Tools for the multilingual de-identification of (sensitive) language resources are becoming essential to ensure that such resources can be shared.
2. De-identification is crucial to ensure that all legal & ethical considerations are taken into account during the production/repurposing phases but also that the quality/nature of the de-identified data sets remains appropriate to conduct research activities.
3. European Public Administrations need personal data processing tools to handle the extremely large amounts of data they manage.

4. Europe's multilingual context will benefit from approaches and tools that can support the European Digital Market in their multilingual data exchanges.

Workshop Objectives and Topics of Interest

This workshop is organised by members of the MAPA project, funded by the EU Connecting Europe Facility (CEF) program (https://mapa-project.eu/). This project has developed a toolkit for the de-identification of texts in the medical and legal fields which addresses all EU official languages. It has followed a BERT-based Named Entity Recognition approach for personal information identification. A wide range of topics have been considered and are hot topics open for discussion to all participants of this workshop. Among them, we have the following:

1. Sensitive personal information, domains and services that require de-identification
2. Corpora annotation and/or creation
3. Annotation guidelines and platforms
4. De-identification tools, data and/or applications
5. De-identification and minority languages
6. Multi-domain and/or multilingual processing
7. NLP techniques and tools used for de-identification
8. Multimodal de-identification
9. Validation and benchmarking of de-identified resources
10. Evaluation of de-identification tools and applications
11. Evaluation protocols: how to evaluate, metrics, approaches, data, experiences
12. Best practices
13. Approaches, activities and systems addressing "anonymization" are also welcome to share their experience.
14. Any other topic related to de-identification

This workshop will also be a good forum to discuss the possibility to design and initiate a new (annual) Challenge (evaluation campaign) on this important topic.

We invite submissions for full papers and system demonstrations that address these questions and other related issues relevant to the workshop.

Workshop Programme and Audience Addressed

This full-day workshop aims at bringing together technology oriented working groups as well as institutions requiring de-identification support that can present their cases. Being de-identification a multi-topic and multi-problem technique, the workshop aims to get researchers, developers and groups needing their services together to discuss approaches, techniques, capabilities and potential collaborations.

Organising Committee

Victoria Arranz (ELDA/ELRA, France)
Montse Cuadros (Vicomtech, Spain)
Aitor Garcia Pablos (Vicomtech, Spain)
Cyril Grouin (LISN-CNR, France)
Manuel Herranz (Pangeanic, Spain)

Programme Committee

Khalid Choukri (ELDA/ELRA, France)
Hercules Dalianis (Stockholm University, Sweden) -TBC
Thierry Etchegoyhen (Vicomtech, Spain)
Albert Gatt (Malta University, Malta)
Lucie Gianola (LISN-CNR, France)
Ona de Gibert (BSC, Spain)
Marwa Hadj Salah (ELDA/ELRA, France)
Udo Hahn (University of Jena, Germany) -TBC
Thomas Kleinbauer (COMPRISE project)
Maite Melero (BSC, Spain)
Stéphane Meystre (Medical University of Southern Carolina, USA) - TBC
Mickaël Rigault (ELDA/ELRA, France)
Patrick Paroubek (LISN-CNR, France)
Naiara Perez (Vicomtech, Spain)
Stelios Piperidis (Athena Research & Innovation Center, Greece)
Prokopis Prokopidis (Athena Research & Innovation Center, Greece)
Mike Rosner (Malta University, Malta)
Roberts Rozis (TILDE, Latvia)
Özlem Uzuner (George Mason University, USA) – TBC
Emmanuel Vincent (Inria Nancy - Grand Est, France)
Rinalds Vīksna (TILDE, Latvia)
Pierre Zweigenbaum (LISN-CNR, France)

Important dates

Submission of full papers: Sunday 10 April 2022
Notification of acceptance of papers and demonstrations: Tuesday 3 May 2022
Submission of camera-ready version: 23 May 2022
Workshop: Monday 20 June 2022

Submission

Authors should use the START system accessible and the LREC author's kit for submitting their papers. Full details will be provided with the updated CFP that will be circulated early 2022.

For further queries, please contact Victoria Arranz at arranz@elda.org.

LRE 2022 Map and "Share your LRs!" initiative

When submitting a paper from the START page, authors will be asked to provide essential information about resources (in a broad sense, i.e. also technologies, standards, evaluation kits, etc.) that have been used for the work described in the paper or are a new result of your research. Moreover, ELRA encourages all LREC authors to share the described LRs (data, tools, services, etc.) to enable their reuse and replicability of experiments (including evaluation ones).

Identify, Describe and Share your LRs!

- Describing your LRs in the LRE Map is now a normal practice in the submission procedure of LREC (introduced in 2010 and adopted by other conferences). To continue the efforts initiated at LREC 2014 about "Sharing LRs" (data, tools, web-services, etc.),

authors will have the possibility, when submitting a paper, to upload LRs in a special LREC repository. This effort of sharing LRs, linked to the LRE Map for their description, may become a new "regular" feature for conferences in our field, thus contributing to creating a common repository where everyone can deposit and share data.

- As scientific work requires accurate citations of referenced work so as to allow the community to understand the whole context and also replicate the experiments conducted by other researchers, LREC 2022 endorses the need to uniquely Identify LRs through the use of the International Standard Language Resource Number (ISLRN, www.islrn.org), a Persistent Unique Identifier to be assigned to each Language Resource. The assignment of ISLRNs to LRs cited in LREC papers will be offered at submission time.