The Multilingual Anonymisation Toolkit for Public Administrations (MAPA) Project

Ē. Ajausks†, V. Arranz‡, L. Bié*, A. Cerdà-i-Cucó*, K. Choukri‡, M. Cuadros§, H. Degroote*, A. Estela*, T. Etchegoyhen§, M. García-Martínez*, A. García-Pablos§, M. Herranz*, A. Kohan*, M. Melero, M. Rosner¶, R. Rozis†, P. Paroubek◇, A. Vasiļevskis†, P. Zweigenbaum◇ ***

†Tilde {eriks.ajausks,roberts.rozis,arturs.vasilevskis}@tilde.lv

‡ELDA/ELRA {arranz,choukri}@elda.org

*Pangeanic - PangeaMT {l.bie,a.cerda,h.degroote,a.estela,m.garcia,m.herranz,a.kohan}@pangeanic.com

**Barcelona Supercomputing Center maite.melero@bsc.es ◇ Université Paris-Saclay, CNRS, LIMSI {pap,pz}@limsi.fr

¶University of Malta mike.rosner@um.edu.mt §Vicomtech {agarciap,mcuadros,tetchegoyhen}@vicomtech.com

Abstract

We describe the MAPA project, funded under the Connecting Europe Facility programme, whose goal is the development of an open-source de-identification toolkit for all official European Union languages. It will be developed since January 2020 until December 2021.

1 Introduction

De-identification may provide the means to share language data while also protecting private or sensitive data by spotting then deleting, obfuscating, pseudoymising or encrypting personally identifiable 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.

The Multilingual Anonymisation toolkit for Public Administrations (MAPA) project aims to leverage natural language processing tools to develop an open-source toolkit for effective and reliable text de-identification, focusing on the medical and legal domains. The project is funded by the Connecting Europe Facility (CEF) programme, under grant N^{o} A2019/1927065, and will run from January 2020 until December 2021.

The toolkit developed by the MAPA partners (Pangeanic[1], Tilde[2], CNRS[3], ELDA[4], University of Malta[5], Vicomtech[6] and SEAD[7]) will address all official EU languages, including under-resourced ones such as Latvian, Lithuanian, Estonian, Slovenian and Croatian, and severely under-resourced ones such as Irish and Maltese.

As a part of the project, a connection to eTranslation,[8] an online machine translation service provided by the European Commission, will be established to foster the provision of multilingual datasets by public administrations that may in turn improve the coverage and quality of machine translation systems.

2 Approach

At its core, the MAPA anonymisation toolkit will rely on Named Entity Recognition and Classification (NERC) techniques using neural networks and deep learning techniques. The latest deep learning architectures and the availability of pre-trained multilingual language models, such as BERT (Devlin et al., 2019) have pushed the state of the art in NERC to new levels of performance.

In addition, thanks to the transfer learning capabilities shown by this type of deep learning models, new systems can be trained using smaller datasets of manually labelled data, and the knowledge acquired for a given domain or language can be reused in a cross-domain or cross-language setting (García-Pablos et al., 2020). MAPA will leverage the most innovative technology to provide robust models for the 24 official European languages, trained to detect named entities that involve sensitive information, depending on the application do-

*All authors have contributed equally to this work.

[1]https://pangeamt.com/

[2]https://www.tilde.com/

[3]http://www.cnrs.fr/

[4]http://www.elra.info/en/

[5]https://www.um.edu.mt/

[6]https://www.vicomtech.org/en/

[7]https://avancedigital.gob.es/

[8]https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en

main (e.g., medical, legal).

MAPA will contain a general NERC model, that will be further fine-tuned to detect domain-specific entities. The system will then be tailored to fulfil the specific needs of each use case. Since some severely under-resourced languages such as Maltese, one of the official EU languages, are not included in the pre-trained multilingual BERT model, a separate solution will be developed in this case.

The deep learning NERC approach will be complemented with other configurable mechanisms, such as pattern detection based on regular expressions, to deal with pattern-based entities: email addresses, ID numbers, telephone numbers, bank accounts, etc. It will also be capable of using user-defined dictionaries to detect specific uses of entity names known in advance.

All these subsystems will be seamlessly combined into an integrated system that will provide a powerful and customisable de-identification engine. For each EU language, a separate docker image will be published, which will take text as input and return it in de-identified form.

3 Use cases

The project includes two specific deployment cases for public institutions in an EU country: one for the health domain and one for the legal domain. Both domains were selected given their strong de-identification requirements prior to any sharing of the data. In each deployment case, the system will be tailored to the specific needs of the relevant institution.

L-Università ta' Malta (University of Malta) will take care of the deployment case for the MAPA toolkit in Malta. In Spain, the deployment case will be executed under the umbrella of the Language Technology Plan[9], which is already running actions in the Health sector in close collaboration with the Ministry and regional institutions, and is willing to expand its activities to the Legal public sector.

4 Data Collection

MAPA will count on a data collection activity to provide the necessary training and testing data for the toolkit development. Data is currently being identified and collected for the 24 relevant European languages. One million sentence corpora are targeted per language, prioritising both medical and legal data, but also containing some general-language data for training. Testing will make use of sample data sets which will be manually annotated with named entities addressing the de-identification needs of the covered domains in the 24 languages. Specific annotation guidelines are currently being defined for that purpose.

The performance of the produced system will be evaluated for each language on held-out sample data sets for each of the two prioritized domains. This evaluation will inform use case designers and users about the expected performance of the base system so that they can assess their need for further adaptation.

5 Conclusion

The MAPA project will develop an open-source anonymisation toolkit for all official EU languages, which will support public administrations in sharing their data while complying with the GDPR requirements. The toolkit will be publicly available and particularly targeted to public administrations in the health and legal domains, as a result of the specific use cases addressed during the development of the project.

References

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

García-Pablos, Aitor, Naiara Perez, and Montse Cuadros. 2020. Sensitive Data Detection and Classification in Spanish Clinical Text: Experiments with BERT. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*.

[9]https://www.plantl.gob.es/tecnologias-lenguaje/PTL/Paginas/plan-impulso-tecnologias-lenguaje.aspx