Spanish Datasets for Sensitive Entity Detection in the Legal Domain
Ona de Gibert1, Aitor García-Pablos2, Montse Cuadros3, Maite Melero1
1. Barcelona Super Computing Center (BSC)
2. SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)

ABSTRACT
The deidentification of sensible data, also known as automatic textual anonymisation, is essential for data sharing and reuse. The first step for data anonymisation is the detection of sensible entities. In this work, we present four new datasets for named entity detection in Spanish in the legal domain. In order to assess the quality of the generated datasets, we have used them to fine-tune a battery of entity-detection models, using as foundation different pre-trained language models. We compare the results obtained, which validate the datasets as a valuable resource to fine-tune models for the task of named entity detection. We further explore the proposed methodology by applying it to a real use case scenario.

MOTIVATION
- With the encouragement of data sharing in the EU and the compliance of the GDPR, data anonymization is essential.
- The first step for identification of sensible data is the detection of sensible entities.
- The MAPA project aims at addressing this issue for the 24 languages of the EU.
- In this work, we present the first open resources annotated for NERC in Spanish in the legal domain.

DATASETS
- Corpus de Procesos Penales (CPP)
  - 10 court cases from Taranilla (2012)
- Dicámenes del Estado (DE)
  - Opinions from the State Council from Samy et al. (2020)
- 71,667 pseudo-anonymised documents
- Insertion of fake named entities

ANNOTATION
- Two-level annotation hierarchy
- Proposed by Gianola et al. (2021)
- INCePTION platform (Klie et al, 2018)

EXPERIMENTS
- Pre-trained Language Models
  1. mBERT (Devlin et al. 2019): multilingual, general domain
  2. BETO (Canete et al. 2020): monolingual, general domain
  3. RoBERTa-b (Gutierrez-Fandiño et al. 2022): monolingual, general domain
  4. RoBERTaLex (Gutierrez-Fandiño et al. 2021): monolingual, domain-specific

CONCLUSIONS
- We present new available datasets
- Combination of synthetic and manually annotated data is beneficial
- Multilingual models can be very useful
- Domain adaptation is tricky

DATASETS
- Corpus de Procesos Penales (CPP)
  - 10 court cases from Taranilla (2012)
  - Conversion from PDF to text

MOTIVATION
- Need of automatic anonymisation tools

DATASET
- Pseudo-anonymised corpus of 120 documents of judicial resolutions, including court rulings, court orders and decrees

ANNOTATION
- Adapted set of entities

CONCLUSIONS
- We present new available datasets
- Combination of synthetic and manually annotated data is beneficial
- Multilingual models can be very useful
- Domain adaptation is tricky