Spanish Datasets for Sensitive Entity Detection in the Legal Domain

Ona de Gibert[1], Aitor García-Pablos[2], Montse Cuadros[2], Maite Melero[1]
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.**

New resources!

MOTIVATION

- With the encouragement of **data sharing** in the EU and the compliance of the **GDPR**, data anonymization is essential.
- The first step for deidentification 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**.

MAPA ANONYMISATION

DATASETS

EUR-Lex

- Multilingual corpus of court decisions from Baisa et al. (2016)

Corpus de Procesos Penales (CPP)

- 10 court cases from Taranilla (2012)
- Conversion from PDF to text

Dictámenes del Estado (DE)

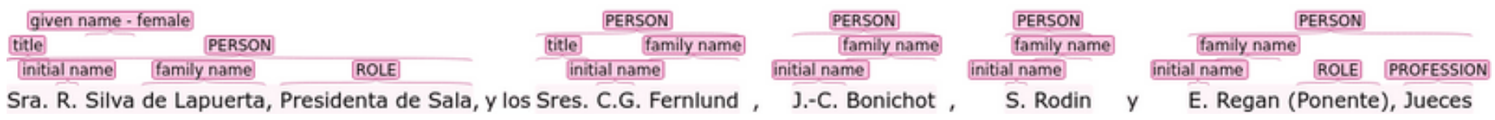
- Opinions from the State Council from Samy et al. (2020)
- 71,667 pseudo-anonymised documents
- Insertion of fake named entities

14,2% error rate

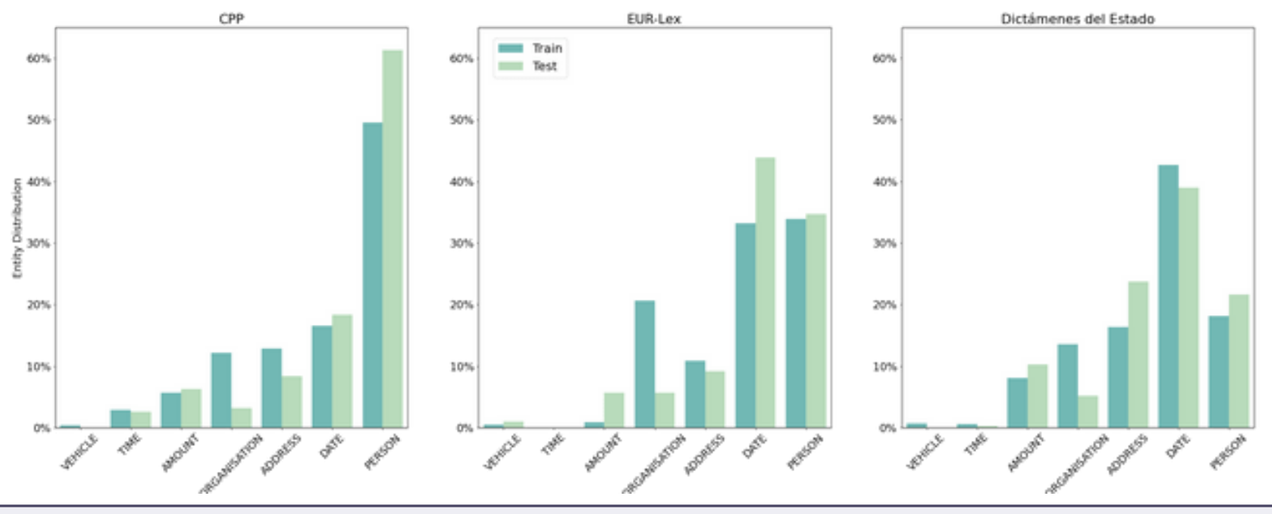
	CPP		EUR-Lex		DE	
	Train	Test	Train	Test	Train	Test
# docs	32	3	9	3	5,000	10
# tokens	40,573	7,445	62,705	27,454	10,943,332	16,592
# level 1 tags	3,189	578	2,837	1,306	907,920	1,367
# level 2 tags	2,463	486	1,721	910	604,598	1,039
# level 1 ents	921	191	1,076	493	266,336	388
# level 2 ents	1,555	338	1,495	826	471,335	716

ANNOTATION

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



AGE **ADDRESS** **AMOUNT** **DATE** **ETHNIC CATEGORY**
MARITAL STATUS **NATIONALITY** **ORGANISATION**
PERSON **PROFESSION** **ROLE** **TIME** **VEHICLE**
city country day family name given name - female
given name - male ID document number initial name
licence plate number model month place
postcode standard abbreviation street territory
title type unit url value year



EXPERIMENTS

Pre-trained Language Models

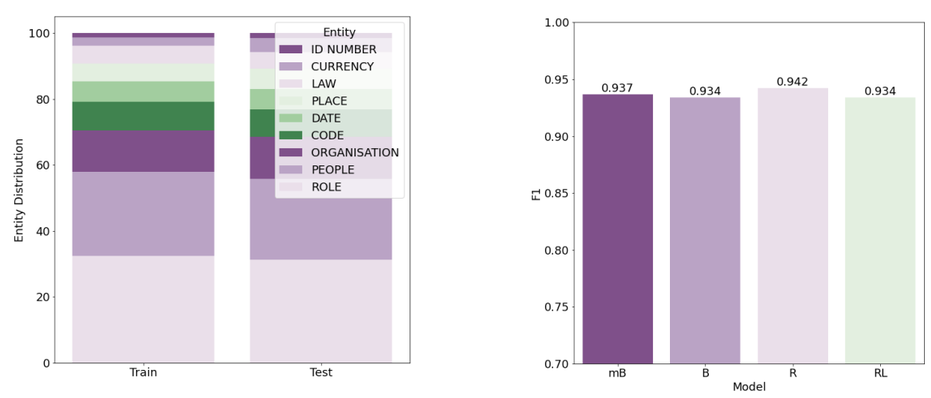
1. mBERT (Devlin et al. 2019): multilingual, general domain
2. BETO (Canete et al. 2020): monolingual, general domain
3. RoBERTa-b (Gutierrez-Fandiño et al. 2022): monolingual, general domain
4. RoBERTaLex (Gutierrez-Fandiño et al. 2021): monolingual, domain-specific

A Real Use Case: the Spanish Ministry of Justice

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

READ OUR PAPER!