Automatic Detection and Labelling of Personal Data in Case Reports from the ECHR in Spanish: Evaluation of Two Different Annotation Approaches

Maria Sierro
LASLAB
University of the Basque Country (UPV/EHU)
maria.sierro@ehu.eus

Begoña Altuna
HiTZ Center - Ixa
University of the Basque Country (UPV/EHU)
begona.altuna@ehu.eus

Itziar Gonzalez-Dios
HiTZ Center - Ixa
University of the Basque Country (UPV/EHU)
itziar.gonzalezd@ehu.eus

Abstract

In this paper we evaluate two annotation approaches for automatic detection and labelling of personal information in legal texts in relation to the ambiguity of the labels and the homogeneity of the annotations. For this purpose, we built a corpus of 44 case reports from the European Court of Human Rights in Spanish language and we annotated it following two different annotation approaches: automatic projection of the annotations of an existing English corpus, and manual annotation with our reinterpretation of their guidelines. Moreover, we employ Flair on a Named Entity Recognition task to compare its performance in the two annotation schemes.

1 Introduction

One of the reasons why research on the automatic detection and labelling of personal information in legal texts (such as case reports) is important is that many countries, including Spain, have the legal requirement of removing sensitive information from these texts before publishing them. However, Pilán et al. (2022) have argued that most research on sensitive entities detection and classification has focused on clinical data, and publicly available evaluation datasets outside this domain are scarce. Moreover, as indicated by Csányi et al. (2021), carrying out this process manually is extremely inefficient.

Regarding datasets for automatic detection and labelling of personal data in the legal domain in Spanish, few datasets have been published, and often they have been built from texts subjected to previous anonymization, thus making the evaluation performed on this data less realistic. This is the case of the work of Arranz et al. (2022), which introduces very detailed annotation guidelines (Arranz et al., 2020).

In contrast, the privacy-oriented annotated corpus in English built by Pilán et al. (2022) stands out due to the use of case reports from the European Court of Human Rights (ECHR), which are publicly available in the HUDOC database[1] in full-text with the consent of the applicants involved in the cases. Additionally, they annotated the corpus by taking into consideration not just direct identifiers, following the usual procedure, but also quasi or indirect identifiers. Nevertheless, their annotation approach presents some ambiguity, given that it leaves room for interpretation and some of their selected entities cover a broad variety of forms.

The ambiguity of annotation guidelines is an important factor to take into consideration when investigating automatic detection and labelling of personal information because it might cause annotations to be non-homogeneous. As argued by Benesty (2019a), non-homogeneous annotations may decrease the performance of Language Models (LMs) on Named Entity Recognition (NER) tasks (the tasks that precede the masking or replacement of sensitive information for anonymization or pseudonymization in legal texts).

Drawing from the work of Pilán et al. (2022), in this paper we intend to contribute to the field of automatic detection and labelling of personal information by (1) building one evaluation corpus of case reports from the ECHR in Spanish, and annotating it following two different annotation approaches; and (2) employing this corpus for assessing the performance of Flair (Akbik et al., 2018) on a NER task and comparing the results of the two annotation schemes. On the one hand, we annotate the corpus via automatic annotation projection, respecting the annotation approach of Pilán et al. (2022). Separately, we annotate the same corpus by following our own reinterpretation of their annotation approach, inspired by the guidelines of Arranz et al. (2020). Our goal is to observe the effects of the different level of ambiguity of the

[1] https://hudoc.echr.coe.int (accessed on July 2023)

Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization
CALD-pseudo 2024, pages 18–24

annotation approaches on the results of the NER task per entity type. We publish the code employed as well as the annotated datasets on GitHub[2] under a MIT license.

2 The TAB corpus

In order to build an evaluation corpus of case reports from the ECHR in Spanish, we departed from the test set of the English corpus built by Pilán et al. (2022). Their corpus is called the Text Anonymization Benchmark (TAB) corpus and it is available on GitHub[3] in json format under a MIT license. This json file contains both the annotations as well as the texts from the ECHR, taken from the HUDOC database. In regards to the reproduction of its website content, the EHCR (n.d.) claims that:

> The information and texts available on the Court's website may be reproduced provided the source is acknowledged (© ECHR-CEDH) and the reproduction is made for private use or for the purposes of information and education in connection with the Court's activities. This authorisation is subject to the condition that the source is indicated and that any such reproduction is free of charge.

As it is explained by Pilán et al. (2022), the texts included in their annotated corpus only contain judgments from the "Grand Chamber" and the "Chamber", and they are restricted to the document sections called "Introduction" and "Statement of Facts", given that they contain the largest quantity of personal identifiers. In addition to that, they selected the judgements by ensuring that their annotators would have knowledge of the national language of the country accused of human rights violations.

It is important to note that the TAB corpus was annotated by 12 annotators. In the work of Pilán et al. (2022), annotators were instructed to annotate all sensitive entities and their semantic types in a first step. In a second step, they were asked to use their interpretation to determine whether to mask each sensitive entity for protecting a person's identity while preserving data utility. Moreover, annotators were instructed to indicate whether the entities to be masked were direct or quasi-identifiers. In a third step, they added a second attribute to the entity mentions indicating whether they corresponded to confidential information (such as religious beliefs, ethnicity or health data). The TAB corpus maintains the masking decisions by all the annotators given that Pilán et al. (2022) consider that there are often multiple correct masking choices in the same text.

3 Dataset creation

In this section we explain our workflow for creating the Spanish corpus.

3.1 Data collection and translation

We extracted 44 random texts from the TAB test set and automatically translated them into Spanish with DeepL.[4] The use of Machine Translation (MT) for building our corpus implies that a number of translation errors are expected. In our corpus, the texts translated with DeepL were not post-edited, but they were inspected by native speakers during the review of the projected annotations. In general, the most common flaw in the Spanish automatic translations was the inconsistent translation of organization names (e.g. "Poole Magistrate's Court" sometimes translated as "*Tribunal de Magistrados de Poole*" and sometimes left untranslated in the same text).

3.2 Projection of annotations

Before projecting the annotations, we collapsed the annotations by all the annotators in the test set of the TAB corpus (they are all considered equally correct examples). When collapsing all the annotators' decisions, only the annotations of the spans to be masked (which are both direct and quasi-identifiers) were kept.

Furthermore, in order to project the annotations with the T-Projection method (García-Ferrero et al., 2023) we transformed the data to get a CoNLL file with IOB tags, with sentences separated by double-space, and only one layer of annotations. The downside of this process was that there was some loss of information. The entity types (shown in the first column of Table 1) of the spans to be masked were transferred. However, the additional labels (which include the distinction between direct and quasi-identifiers and the indication of confidential information) were lost. Consequently, our work

[2] https://github.com/mariasierrofer/sensitive-entity-detection-ECHR-Spanish

[3] https://github.com/NorskRegnesentral/text-anonymization-benchmark

[4] https://www.deepl.com/translator

is limited to the recognition of the different entity types.

After projecting the annotations of the selected texts into the Spanish translations, two persons (a Natural Language Processing (NLP) Master's student and a linguist) reviewed them with the INCEpTION tool.[5] We measured the Inter-Annotator Agreement (IAA) with the same tool at the entity level using the metric Cohen's Kappa, resulting in a value of 0.89.

3.3 Reinterpretation and reannotation

Our reinterpreted guidelines, which combine the annotation approach of Pilán et al. (2022) with the detailed guidelines of Arranz et al. (2020), are available in the appendix. Table 1 compares the entity types included by Pilán et al. (2022) with the entity types included in our reinterpretation of their guidelines. Our goal is to pave the path for reducing label ambiguity. The most relevant changes of our reinterpreted guidelines include:

- The replacement of the DEM entity by two new labels: NATIONALITY (referring to a person's demonym) and ETHNIC CATEGORY (covering the ethnic parameters of a person's identity, such as race, religion, language, and regional origin). A disadvantage of this replacement is that the new labels do not cover some information that was included by the DEM entity (such as health information, political and sexual orientation). Due to the small size of our corpus, adding detailed labels for all types of personal information would produce few occurrences of each one. For the same reason, the MISC label is not addressed in our reinterpreted guidelines.

- The split of the DATETIME entity into two new labels: DATE and TIME. In regards to these labels, our reinterpreted guidelines contain one specific adaptation for Spanish language: the annotation of dates and times covers their preceding articles (but not prepositions) in order to comply with the ISO-TimeML standard for temporal annotation (ISO, 2008) and to potentially ease its automatic detection with existing tools.

- A specification related to the ORG entity, which now only covers the spans which refer to distinct organizations and not to generic institutions (e.g. "High Court", "Supreme Court").

- The split of the PERSON entity into two new labels: PER and LEGAL PROFESSIONAL. These two labels make a distinction between the names of the people professionally involved in the cases and the names of the rest of the people mentioned in the texts. The reason for making this distinction is that in Spain (and other countries), the names of the people professionally involved in the cases do not have to be masked (van Opijnen et al., 2017).

- A difference regarding the QUANTITY entity, which now covers meaningful quantities (not directly deducible from the rest of the information of the text) without their units of measure. It also covers periods of time, which were previously included in the DATETIME label. Currency instead gets a separate treatment: these units of measure are annotated with the CURRENCY tag because they can reveal information about the locations involved in the cases.

Corpus with projected annotations		**Corpus with manual annotations**	
Entity	**nb. tags**	**Entity**	**nb. tags**
PERSON	355	PER LEGAL PROFESSIONAL	191 170
CODE	54	CODE	87
LOC	163	LOC	454
ORG	216	ORG	130
DEM	92	NATIONALITY ETHNIC CATEGORY	110 22
QUANTITY	55	QUANTITY CURRENCY	204 32
DATETIME	799	DATE TIME	786 5
MISC	50	-	-
total	1,784	total	2,191

Table 1: Comparison of the entity types and number of tags included in the corpora with projected and manual annotations.

It is also important to note that, as stated in Section 2, Pilán et al. (2022) included a second step of annotation where they instructed annotators to judge case by case which combination of sensitive entities to mask for protecting a person's identity while preserving data utility. We intend to simplify this process and avoid leaving any room for interpretation by annotating all the occurrences of the

[5] https://inception-project.github.io/

entity types included in our reinterpreted guidelines in a unique annotation step. With the intention of avoiding compromising the data utility of the texts, our strategy consisted in defining the entity types for targeting precise sensitive information.

Two persons (a NLP Master's student and a philologist) carried out the manual annotations with the INCEpTION tool by making modifications to the preceding projected annotations. In this case, we measured the IAA with the same tool at the entity level using the metric Cohen's Kappa, resulting in a value of 0.99. This higher agreement could indicate that the new annotation approach of the reinterpreted guidelines was indeed less ambiguous and the annotators had less room for interpretation.

4 Experimental setup

Once we built and annotated the corpus of legal texts in Spanish language, we used it for assessing Flair (Akbik et al., 2018), which has provided good results in the identification of sensitive information in legal texts in previous studies (Benesty, 2019a; Benesty, 2019b). We used Flair version 0.12.2. For all the experiments, the corpus was split into train, dev, and test set as shown in Table 2.

Set	nb. sents.	nb. tokens
train	1,245	34,924
dev	178	4,430
test	193	5,255

Table 2: Number of sentences and tokens of the train, dev, and test sets.

We trained a bi-LSTM-CRF sequence tagger[6] with default hyper parameters for 18 epochs in both our experiment on the corpus with projected annotations as well as our experiment with the manually annotated corpus. We used pre-trained embeddings (Akbik et al., 2019) from Flair "ner-multi" model.[7]

The metrics used in the assessment of Flair are precision, recall, and F1-score, computed at the mention level, but for brevity we will only focus on the F1-scores in Section 5 and Section 6.

5 Evaluation on corpus with projected annotations

On the corpus with projected annotations, Flair achieves a micro average F1 of 0.73.

[6] https://github.com/flairNLP/flair/blob/master/flair/models/sequence_tagger_model.py

[7] https://huggingface.co/flair/ner-multi

Entity	F1-score
PERSON	0.73
CODE	0.92
LOC	0.67
ORG	0.39
DEM	0
QUANTITY	0.50
DATETIME	0.95
MISC	0
micro avg	0.73

Table 3: F1-scores per entity type and micro average F1 calculated on the test set of the corpus with projected annotations.

By looking at the results per entity type (shown in Table 3), it can be observed that there is a particularly stark contrast between Flair's performance on the DATETIME and CODE labels (F1-score over 0.9) vs. the DEM and MISC labels (0 F1-score). This difference could indicate that the DEM and MISC labels were more widely defined than the DATETIME and CODE labels and their annotations were less homogeneous. Pilán et al. (2022) noticed a similar difference in the performance of their selected LMs, and they stated that the reason could be related to the broad variety of forms that the DEM and MISC labels can take. Moreover, it could be argued that such an imbalanced performance might be due to a dissimilar number of tags for each label. However, while it is true that the DATETIME label presents the larger number of tags (799 tags in the corpus with projected annotations), the labels CODE, MISC, and DEM all have a similar number of tags (54, 50, and 92 tags respectively), and still the performance of Flair on the CODE label is much higher.

6 Evaluation on corpus annotated with our reinterpreted guidelines

On the corpus annotated with our reinterpreted guidelines, Flair outperforms the results of the previous experiment, achieving a micro average F1 of 0.80.

By looking at the results per entity type (shown in Table 4), it can be observed that the TIME and the ETHNIC CATEGORY labels obtained a 0 F1-score, likely due to the scarcity of tags of these types (5 and 22 tags respectively in the whole corpus).

On the other hand, the performance of Flair on the DATE label (0.98 F1-score) is slightly higher than it was on the DATETIME label (0.95 F1-score). With a similar number of tags of this type,

Entity	F1-score
PER	0.67
LEGAL PROFESSIONAL	0.46
CODE	1
LOC	0.87
ORG	0.20
NATIONALITY	0.85
ETHNIC CATEGORY	0
QUANTITY	0.75
CURRENCY	0.60
DATE	0.98
TIME	0
micro avg	0.80

Table 4: F1-scores per entity type and micro average F1 calculated on the test set of the corpus annotated with our reinterpreted guidelines.

the increase in performance could indicate that it was beneficial to make a distinction between the annotation of dates, times, and durations. In particular, the annotation of durations was covered by the DATETIME label according to the guidelines created by Pilán et al. (2022). On the contrary, in our reinterpretation of their guidelines, durations were covered by the QUANTITY label, which also shows an increase in performance (0.75 F1-score vs. 0.50 F1-score in the previous experiment).

There is also a slight increase in Flair's performance on the CODE label (1 F1-score vs. 0.92 F1-score in the previous experiment). Regarding the labels PER and LEGAL PROFESSIONAL, the performance of Flair decreases when compared to the PERSON label.

In regards to the NATIONALITY label included in the manually annotated corpus, which replaces the previous DEM label, the performance of Flair is higher (0.85 F1-score). This is especially interesting considering that the DEM label included in the corpus with projected annotations got a 0 F1-score and the number of tags is similar in both corpora.

Furthermore, while the performance of Flair on the ORG label decreases, its performance on the LOC label is much higher (0.87 F1-score vs. 0.67 F1-score in the previous experiment). In this case, the main reason for the increase seems to be related to the larger number of LOC tags in the corpus annotated according to our reinterpreted guidelines (454 tags vs. 163 tags). The larger number of LOC tags is due to our indication of annotating all the occurrences of the entity types included in our reinterpreted guidelines.

Finally, the performance of Flair on the CURRENCY label is low (0.60 F1-score). Other than having few occurrences (32 tags in the whole corpus), this low performance could also indicate that this label is still ambiguous.

7 Conclusions and future work

Throughout this paper, we have evaluated two annotation approaches for the automatic detection and labelling of personal information in case reports from the ECHR in Spanish language. Our goal was to observe the differences in the performance of Flair (Akbik et al., 2018) in relation to the ambiguity of the selected entity types. We performed this evaluation by building one evaluation corpus of case reports from the ECHR in Spanish, and annotating it by following two different annotation approaches: automatic projection of the annotations of the English corpus built by Pilán et al. (2022), and manual annotation with our reinterpretation of their guidelines, also based on the work of Arranz et al. (2020). We used this newly-built corpus for assessing Flair on a NER task and comparing the results of the two annotation schemes. We make both the corpus and the code public under a MIT license to encourage research on automatic detection and labelling of personal data in legal texts in Spanish.

The results showed that our reinterpreted guidelines partly succeeded in getting less ambiguous labels and more homogeneous annotations. This idea is reinforced by the higher IAA obtained with our reinterpreted guidelines, which suggests that the more detailed approach of our guidelines might also help human annotators to be consistent in their annotations. As we mentioned, the manual annotation of entities may be very time-consuming. An automatic system that yields a good performance in the task will help decreasing the burden. Trying to make a more consistent annotation has proven to be a sensible approach to improve the performance of Flair. In the near future an anonymization analysis should be conducted to see whether our approach effectively reduces the risk of reidentification while not compromising the readability of the document.

In the future research, we will expand our corpora, adapt and apply our reinterpreted guidelines to other languages, and include new specific labels. We also plan to test other Language Models and other techniques such as zero-shot or few shot. Additionally, we intend to test privacy models such as C-sanitized (Sánchez and Batet, 2016) for a comprehensive risk analysis.

Limitations

We evaluated the performance of Flair (Akbik et al., 2018) on a NER task with one corpus of 44 case reports from the ECHR in Spanish. The texts of our corpus were translated from English into Spanish via MT (using DeepL) without post-editing. In future work, it would be interesting to either employ professional translations or post-edit the automatic translations. Additionally, our work could be extended to other languages. It would also be interesting to carry out similar experiments on larger corpora and add labels covering other types of information that we could not cover due to the size of our corpus, including a deeper treatment of quasi-identifiers. On the other hand, it should be noted that our work is restricted to the recognition of sensitive entities on legal texts and it does not reflect on the masking operations following this task. Moreover, since we do not annotate pronouns and possessive adjectives, our corpus is suited for anonymization rather than pseudonymization. Lastly, there is no comprehensive risk analysis which examines the connection between the detected sensitive entities and external knowledge bases, as recommended by Csányi et al. (2021).

Acknowledgments

Maria Sierro receives funding from the LASLAB group (IT-1426-22) funded by the Basque Government.

Begoña Altuna is supported by the Basque Government postdoctoral grant POS 2022 2 0024.

We also acknowledge the funding from the following projects: a) Ixa group A type research group (IT-1805-22) funded by the Basque Government. b) DeepKnowledge (PID2021-127777OB-C21) project funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe. c) AWARE: Commonsense for a new generation of natural language understanding applications (TED2021-131617B-I00): funded by MCIN/AEI /10.13039/501100011033 by the European Union NextGenerationEU/ PRTR. d) DeepR3 (TED2021-130295B-C31) funded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Multilingual sequence labeling with one model.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Victoria Arranz, Chomicha Bendahman, Elena Edelman, Mickael Rigault, and Khalid Choukri (ELDA). 2020. Annotation Guidelines for Named Entities in MAPA.

Victoria Arranz, Khalid Choukri, Montse Cuadros, Aitor García Pablos, Lucie Gianola, Cyril Grouin, Manuel Herranz, Patrick Paroubek, and Pierre Zweigenbaum. 2022. MAPA project: Ready-to-go open-source datasets and deep learning technology to remove identifying information from text documents. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 64–72, Marseille, France. European Language Resources Association.

Michael Benesty. 2019a. NER algo benchmark: spaCy, Flair, m-BERT and camemBERT on anonymizing French commercial legal cases. *Towards Data Science*.

Michael Benesty. 2019b. Why we switched from Spacy to Flair to anonymize French case law. *Towards Data Science*.

Gergely M. Csányi, Dániel Nagy, Renátó Vági, János P. Vadász, and Tamás Orosz. 2021. Challenges and Open Problems of Legal Document Anonymization. *Symmetry*, 13(8):1490.

EHCR. n.d. Copyright and Disclaimer.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2023. T-Projection: High Quality Annotation Projection for Sequence Labeling Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15203–15217, Singapore. Association for Computational Linguistics.

ISO. 2008. *ISO DIS 24617-1: 2008 Language Resource Management - Semantic Annotation Framework - Part 1: Time and Events*. International Organization for Standardization, Geneva, Switzerland.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

David Sánchez and Montserrat Batet. 2016. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

Marc van Opijnen, Ginevra Peruginelli, Eleni Kefali, and Monica Palmirani. 2017. On-Line Publication of Court Decisions in the EU: Report of the Policy Group of the Project 'Building on the European Case Law Identifier'.

A Reinterpreted guidelines

This appendix contains the annotation guidelines for the detection and labelling of personal information of case reports from the ECHR in Spanish language. The set of entity types to be annotated are:

- PER: this label comprises the names, initials, titles and honorifics (e.g. "*Mr.*", "*Dr.*") of people who are not legal professionals involved in the cases.
- LEGAL PROFESSIONAL: this label comprises the names, initials, titles and honorifics (e.g. "*Mr.*", "*Dr.*") of people who are legal professionals involved in the cases.
- CODE: the CODE label covers all types of identification numbers (e.g. passport numbers, phone numbers, report identifiers, etc.). Nevertheless, even if the CODE label includes case numbers (because they make reference to the cases being treated and can consequently be considered a direct identifier), this label does not comprise any other numbers making reference to legal texts involved in the cases (e.g. convention and law articles, protocols, rules, paragraphs, etc.).
- LOC: covers all types of geographical locations (e.g. countries, cities, addresses, etc.).
- ORG: covers the names of distinct organizations (with the exception of the "ECHR" and the "European Commission on Human Rights", which should not be annotated), and not generic institutions (e.g. "High Court", "Supreme Court", etc.). Still, if address information (e.g. city, country, etc.) is comprised within the expression of a generic institution (e.g. "Supreme Court of *Sweden*"), the address information (e.g. "*Sweden*") should be annotated using the LOC label.
- ETHNIC CATEGORY: covers the ethnic parameters of a person's identity, such as race, religion, language and regional origin.
- NATIONALITY: refers to a person's demonym (e.g. "*French*", "*Swedish*", "*Norwegian*").
- DATE: this label makes reference to dates (days, months, and years) including articles (but not prepositions) in order to comply with the ISO-TimeML standard for temporal annotation. As it happened with the CODE label, the DATE label does not apply to dates that serve to identify legal texts (with the exception of case reports) involved in the case (e.g. convention and law articles, protocols, rules, paragraphs, etc.).
- TIME: corresponds to hours (e.g. "at *4 p.m.*"; or in Spanish "a *las 4 horas*"), expressed in figures or in words (e.g. "*morning*", "*evening*", etc.). It does not include durations, since these are covered by the label QUANTITY.
- QUANTITY: covers quantities (e.g. surface areas, distances, percentages, etc.) without their units of measure. In this way, the QUANTITY label targets meaningful quantities (not directly deducible from the rest of the information of the text), including figures associated to periods of time (e.g. "it lasted for *9* years and *9* months"), which were previously covered by the DATETIME label, as well as ages (e.g. "she was *19* years old").
- CURRENCY: covers currency types (e.g. "*euro*", "*pound*", "*dollar*", etc.).

The general principles that should be kept in mind when annotating are:

- Annotate all the entities in all the selected texts that correspond to the selected entity types.
- Do not annotate pronouns and possessive adjectives revealing gender information, since they imply a low re-identification risk.
- Annotate all the mentions pertaining to the same entity.