MAPA Project: Ready-to-Go Open-Source Datasets and Deep Learning Technology to Remove Identifying Information from Text Documents

Victoria Arranz, Khalid Choukri, Montse Cuadros, Aitor García-Pablos, Lucie Gianola, Cyril Grouin, Manuel Herranz, Patrick Paroubek, Pierre Zweigenbaum

The Legal and Ethical issues Workshop @LREC2022, pages 64–72
Marseille, 20 June 2022

MAPA Project: Ready-to-Go Open-Source Datasets and Deep Learning Technology to Remove Identifying Information from Text Documents

Victoria Arranz,[1] Khalid Choukri,[1] Montse Cuadros,[2]
Aitor García-Pablos,[2] Lucie Gianola,[3] Cyril Grouin,[3]
Manuel Herranz,[4] Patrick Paroubek,[3] Pierre Zweigenbaum[3]
[1]ELDA/ELRA, Paris, France; [2]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Donostia,Spain [3]Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique (LISN), 91405 Orsay, France; [4]Pangeanic – PangeaMT, Valencia, Spain
{arranz,choukri}@elda.org, {mcuadros,agarciap}@vicomtech.org,
{lucie.gianola,cyril.grouin,patrick.paroubek,pierre.zweigenbaum}@lisn.upsaclay.fr, manuel@pangeanic.com

Abstract

This paper presents the outcomes of the MAPA project, a set of annotated corpora for 24 languages of the European Union and an open-source customisable toolkit able to detect and substitute sensitive information in text documents from any domain, using state-of-the art, deep learning-based named entity recognition techniques. In the context of the project, the toolkit has been developed and tested on administrative, legal and medical documents, obtaining state-of-the-art results. As a result of the project, 24 dataset packages have been released and the de-identification toolkit is available as open source.

Keywords: anonymisation, de-identification, sensitive information, deep learning, BERT, NER, annotated data

1. Introduction

Computing Technology, Artificial Intelligence, and Machine Learning have made tremendous progress in recent years. However, a large part of the lore of electronic documents produced for work, administration, entertainment, public services, public communication or personal expression on social-media cannot be used and shared easily for research purposes: this is caused by the presence of sensitive information identifying individuals, like person names, phone numbers, etc. The laws and regulations that protect the private life of individuals[1] require anonymising a document if it is to be disclosed and shared. In many cases, this prevents researchers from using this document for their experiments. Performing such anonymisation task manually is costly. Besides, ever more documents are needed to train the algorithms of all kinds that are used nowadays to process documents automatically. This started a quest for automatic anonymisation, which starts by first addressing the detection and then the removal of any identifying information, a task called de-identifying a document. Anonymisation implies that not only identifying information is not present in a document anymore, but also that it is impossible to infer the identity of a person from the material preserved after the de-identification of the document.

Developing a language- and domain-independent system that detects information in text documents is already a challenge, because access to the full original document with its identifying information is needed as training data to feed machine-learning algorithms. Such algorithms currently provide state-of-the-art performance. What is expected more precisely is to reach the highest possible recall (detecting all that information that needs to be found in the document). Furthermore, de-identifying a document imposes the extra constraint that enough material of the original document should be preserved for the document to remain usable for research purposes. This requires to adopt methods that not only yield a high recall, but also a high precision, since too many false positive alerts would result in a document with insufficient material left for any research purpose.

1.1. The MAPA Project

MAPA[2] (Multilingual Anonymisation for Public Administrations) (Gianola et al., 2020) is an integration project aiming to introduce Natural Language Processing (NLP) tools to develop a toolkit for effective and reliable de-identification of documents in the medical and legal fields. It addresses all EU official languages, including under-resourced ones, such as Latvian, Lithuanian, Estonian, Slovenian and Croatian.

The project has built a deployable, docker-ready, open-source fully multilingual de-identification toolkit able to detect personal data (for instance: person names, addresses, emails, credit card numbers, bank accounts, etc.) as defined by deployment cases in different Member States. The open source toolkit and resources for 24 EU languages are intended to help public administrations to comply with both GDPR[3] and the PSI Di-

[1]General Data Protection Regulation (GDPR): `https://gdpr.eu/`

[2]`https://mapa-project.eu/`

[3]`https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679`

rective[4], particularly in the health and legal fields. The toolkit contributes to the promotion of Public Administration data sharing that is fully de-identified and thus not traceable to personal details, making it GDPR compliant. As a result, data that now remains in vaults and cannot be shared will be able to be re-used in European initiatives such as *NEC TM*[5] (data coming from Public Administrations for translation whose source language contains personal details), *ELRC*[6] (Public Administrations can share data that otherwise they would not be able to), and potentially *eTranslation*[7] (offering anonymisation services as a separate service or embedding it as part of its translation service), etc. MAPA's toolkit is easily customisable as it has pre- and post-processing modules available in the form of an API-ready toolkit dockerised version. This will ease integration and deployment as an isolated I/O module not disturbing current digital infrastructures. Furthermore, adaptation to specific terminology or regulation/language specific entities is made easier by the existence of the entry point offered by these pre- and post-processing modules with an intuitive interface based on regular expressions and add-on lexical resources. The de-identification toolkit is based on state-of-the-art Named-Entity Recognition (NER) (Yadav and Bethard, 2018; Huang et al., 2021), applicable to 24 EU languages. It is not restricted to names and surnames of European origin but addresses those mostly common in all EU countries, and with eTranslation in view, irrespective of whether the text is monolingual, bilingual or a patchwork of languages.

The remaining of this article is organised as follows. Section 2 describes the type of sensitive information that MAPA is targetting, with the hierarchy of named entities defined. In Section 3, we describe the MAPA open-source toolkit. Section 4 details the data production efforts and Section 5 reports on the toolkit evaluation approaches and results.

2. Sensitive Information to be De-Identified

The objective of MAPA is to build a multilingual de-identification toolkit that can de-identify personal and sensitive data referring to some person. For that purpose, multilingual language data in all 24 EU languages covered by the project needed to be annotated with the named entities to be detected, thus providing material for the development and evaluation of the system.

2.1. Named Entity Hierachy

The underlying model of the Named Entity (NE) hierarchy has been defined bearing in mind the needs of the de-identification tool. The objective has been to define a rich hierarchy with the entities that may be found in the different documents that need to be processed. The design of this hierarchy has focused on the legal and medical fields in particular but also targeting a general domain that can accommodate the multi-disciplinary application of the system once developed.

The MAPA NE hierarchy has three levels (as illustrated in Figure 1):

- Level 1 entities (in orange): implicit entities that can be inferred from their annotated elements.
- Level 2 entities (in blue): either explicit or implicit entities that may comprise some level-3 components and types to be annotated. They are also semantic classifiers for the lower level elements.
- Level 3 entity components and types (in green): these are either components within an entity or types of entity.

Despite having such a detailed hierarchy, not all elements are annotated. We benefit from the inferring capacity of some of them to reduce the annotation load (see Section 2.2) .

2.2. Annotation Guidelines

MAPA's annotation guidelines[8] explain the entity structure, relations and annotation definitions in detail. Given that several entities can be fully inferred either from their lower-level entities or from their level-3 components/types, not all elements within the hierarchy are annotated. This would be repetitive and very time consuming.

The annotation task has been carried out with the *INCEpTION*[9] annotation platform (Klie et al., 2018), an open-source tool which allows for a rich and flexible annotation of the data. For instance, an entity may be annotated with elements from any of the 3 MAPA levels if this is allowed by the annotation schema.

A series of general principles are stated in the guidelines defining what needs to be annotated. Sometimes these followed ambiguities and discussions with the annotators themselves. Establishing annotation principles that covered several domains while addressing domain-specific entities to be de-identified turned out to be rather complex. For that reason, guidelines were fine tuned after several annotation tests and inter-domain discussions took place to coordinate this fine-tuning. Some examples of general principles are as follows:

[4]Public Sector Information Directive`https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L1024&from=EN`

[5]`https://www.nec-tm.eu/`

[6]`https://lr-coordination.eu/`

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

[8]`http://www.elra.info/media/filer_public/2022/05/10/mapa_annotation-guidelines-v6.pdf`

[9]`https://inception-project.github.io/`

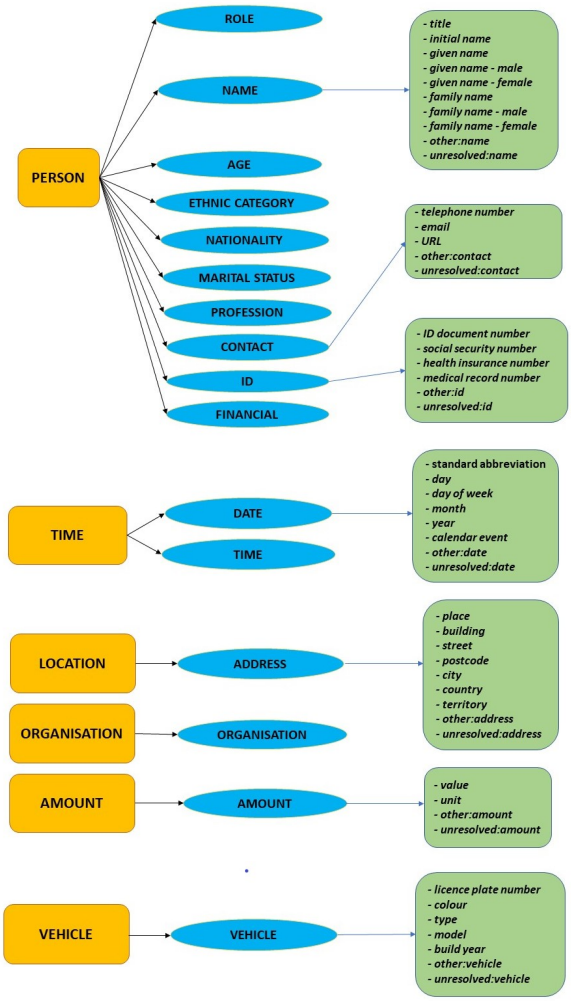


Figure 1: Named Entity Hierarchy

- Annotation needs to consider an entity's domain: for example, legal and administrative data contain many general references like Directives, Decisions, OJ numberings, etc. that will not be annotated.

- Annotation needs to consider an entity's nature: for instance, **Samsung** (referring to the organisation) should be annotated as ORGANISATION. However, **Samsung** (referring to a smartphone) will not be annotated ("phone" is not an entity type).

- The often-ambiguous distinction between **ROLE** and **PROFESSION** needs to consider the domain and the entity's function within a sentence (e.g., **judge** may indicate a ROLE if this refers to the judge taking care of some court ruling or to PROFESSION if the entity does not act as such in that context); etc.

Besides the general principles, annotation definitions are provided for annotators to understand some further entity cases and terminology in their annotation (e.g., how to deal with elements between entities, how to use the **other** and **unresolved** components, etc. Finally, specific instructions are provided for all entities defined within MAPA[10]

3. Open-Source Toolkit

MAPA is a de-identification toolkit for the detection and substitution of sensitive information in text. It is designed to work with potentially any language, provided it is trained/configured properly (Ajausks et al., 2020). In order to perform its task, MAPA relies on different components and approaches, which are configured and integrated into a simple web-service.

At its core, MAPA relies on Deep Learning, using Transformers based neural-networks, in particular BERT(Devlin et al., 2019). Since MAPA is meant to deal with multilingual content, it is developed using the multilingual pre-trained BERT model from Google. However, other BERT models (such as BETO(Cañete et al., 2020) for Spanish) can be used just by changing the name (file) of the base model when training, as long as the model to be used remains compatible with the Transformers library and follows the same "BERT" conventions (special BERT tokens, WordPiece tokenisation, etc.).

In order to achieve the anonymisation of the documents, MAPA performs two kind of tasks: sensitive entity detection/classification (cf. Section 3.1), and detected entities replacement. The latter can be of three types, depending on the user's needs (cf. Section 3.2).

3.1. Sensitive Entities Detection and Classification

The detection of sensitive entities is the task of selecting which entities bear the information that needs to be hidden, removed or replaced. Besides detection, the entities are classified into different types, since that information may help in later steps to perform the information removal/replacement.

The detection can be seen as a regular NERC (Named Entity Recognition and Classification) task, only that the targeted entities depend on your anonymisation use case. In MAPA, the provided datasets and models target a variety of entities, such as "PERSON", "ORGANISATION", "PROFESSION", "AGE", "GENDER", "DATES", "COUNTRIES", "CITIES", etc. (as seen in Section 2.1). These entities are arranged in a two-level hierarchy[11], to have more fine-grained entities if necessary (e.g. a mention of a PERSON containing a "first-name" and a "family-name"). The detection is performed by two different modules that complement each other.

3.1.1. Deep Learning Model Based Detection

The main technology used to provide the entity detection is based on a Transformers model. In MAPA we do

[10] Please refer to the guidelines for full details.

[11] The third level is inferred, as seen in Section 2.1.

provide several pre-trained models, for a few languages and domains, with different levels of performance. But MAPA also offers the capability to train new models provided that you have labelled data in a suitable format (cf Section 4).

3.1.2. Regular-Expression Based Detection

For certain entity types, it is easier to rely on patterns and regular expressions rather than on a Deep Learning based model. The Deep Learning detection can deal with everything provided enough training data is available, but there is no point in using it to detect entities such as phone-numbers, email addresses, URLs, or some identification numbers that can be easily matched using a regular expression. MAPA allows you to configure regular expressions and assign them a meaningful label. That label (entity-type) will be attached to any match occurred in the text.

3.2. Sensitive Entities Replacement

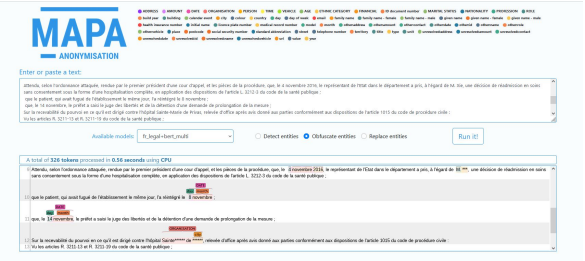
Once the relevant entities have been detected and classified, the anonymisation task requires one further step. The information is still there and needs to be removed or replaced. The simplest way to remove the information is by replacing the detected entities with a symbol like '*' (e.g. "The judge Robinson was in the room" becomes "The judge ******** was in the room"). However, depending on the intended usage for the resulting texts, this is not a suitable approach, because the texts become unnatural and hard to read. This is particularly a problem when intending to share the data for further processing by other systems.

MAPA allows to obfuscate the text with that simple approach, and allows to replace the information by other similar entities, leading to still-readable documents that no longer contain the original sensitive information (cf. Figure 2). For that, MAPA has different modules that complement each other.

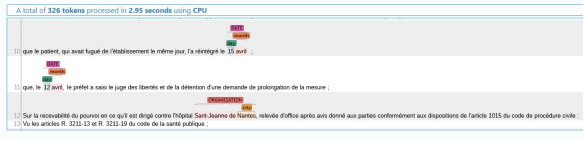
3.2.1. Neural LM Based Replacement

The Neural Language Model replacement is based on a Deep Learning language model. Again the multilingual BERT model is used as the base, but any other BERT model could work. When using this approach, the words that form the sensitive entity are replaced using a neural language model to predict a suitable entity to fill the gap. There are several extra heuristics and filters applied to avoid getting the very same word that we want to have replaced, and to avoid sampling unsuitable words (like a word when it was a number and vice-versa).

Using the Neural LM replacement has some advantages. Firstly, no pre-compiled list of names needs to be used for replacement. Secondly, ideally the Neural LM should pick words that contextually fit better, for example to match the gender of the names, or to deal with morphological inflexion in certain languages. In any case, MAPA provides users with other two replacement types.



(a)



(b)

Figure 2: MAPA demo interface with obfuscation functionality (a) and enlarged view of the entity replacement (with a type plausible substitute) result on a sample legal text in French (b).

3.2.2. Random Character Replacement

This replacer is suitable for identifiers, number strings, or other arbitrary sequence of alpha-numerical characters. It simply replaces each character with another, randomly sampled, character of the same type (digit or letter). For example, an identification code such as X-6543-432 would become something like Z-3234-768.

3.2.3. Dictionary-Based Replacement

This is the classic replacement strategy, using a pre-compiled set of name lists, for the different entity types (people names, city names, etc.). When an entity of a certain type is found, a random name from the pairing list can be sampled as the replacement. This is simple, and for certain languages and domains it is effective. However, one needs to gather meaningful names for every entity type (and language). Further, nothing guarantees the coherence if the sampled entity does not match the context exactly (e.g., a female name in a context that requires a male name, or a French location name in a context that requires a German location name).

MAPA allows configuring specific replacement strategies for each entity type, and even for each language. So one can decide which kind of replacement is more suitable for which type of entity, or leave certain entity types untouched, without replacing them.

3.3. Integration and Deployment of the Toolkit

MAPA is an open source toolkit. That means that one can use it as-is, and also use it to plug one's own resources, including the training of new Deep Learning entity detection models.

The tool can be deployed as a web service. There is a configuration file that allows a fine-grained con-

trol over which models, which detection/replacement strategies and which resources are used when the tool is launched. The exposed web service receives texts of arbitrary length, and returns the list of detected entities, together with the resulting text, with the corresponding entity replacements applied.
The toolkit is offered with a ready-to-use Docker integration, so it is easy to deploy. Even without the Docker wrapping, the tool is based on Python, and uses pretty standard Python dependencies such as Pytorch and Transformers, so it should work on any environment.
The software produced by MAPA has been made available to the community through Gitlab[12].

4. MAPA Data Production

MAPA carried out the production of both Named-Entity (NE) annotated and unannotated datasets for all EU official languages. The objective was to produce (collect and annotate) relevant GDPR-compliant data in the 24 languages for system training, development, and evaluation. In addition, it also produced lists of person names for the 24 languages (cf. Section 4.5). Early stages of the project confirmed the great difficulty to obtain data with sensitive content (in both legal and medical domains) and a new strategy was defined that is further detailed in the sections below:

1. We would focus on other relevant NE rich data sources from related fields like administrative-legal (see Section 4.1).

2. We would enrich available medical data with named entities so as to use clinical documents with relevant sensitive information (cf. Section 4.2).

3. We would process already anonymised data by de-anonymising it first: this would allow us to work on real sensitive data with de-identification needs (cf. Section 4.3).

4. We would explore the production of synthetic data by translation means (cf. Section 4.4).

All annotated datasets, raw corpora and name lists produced within MAPA can be downloaded per language package through the ELRC-SHARE repository[13] and the ELRA Catalogue[14].

4.1. Data from other Relevant Sources

As part of the new strategy, MAPA produced the following resources:

- 24 corpora on the legal-administrative domain from EUR-LEX[15]: these comprised over 2000 sentences per language and the choice for this data was based on the availability of NE-relevant parallel texts. These texts were available for all languages except Irish[16]. The Irish version has been produced with the EC's eTranslation platform[17] and manually revised. All 24 datasets were annotated with MAPA's Named Entities for initial system development, providing good results and the output is a parallel NE annotated set in 24 languages.

- 24 1-Million sentence raw corpora were produced for potential further training, from the following sources and languages:

  - Court of Justice of the European Union[18] (CS, DA, DE, EL, EN, ET, FI, FR, IT, LT, LV, PL, PT, SV).
  - Spanish Council of State (ES)[19].
  - Malta Government Gazzette, Malta Law Courts online and EU documents (MT).
  - Wikipedia, news, web crawling and data augmentation (GA).
  - Wikipedia, news and web crawling (BG, HR, HU, NL, RO, SK, SL).

4.2. Medical Domain Datasets

For the clinical domain, since it was not possible to access real clinical data with sensitive content, we designed an alternative solution consisting in using a corpus of 485 clinical cases written in French from a larger dataset (Grabar et al., 2018; Grabar et al., 2019). We automatically reintroduced nominative data within the texts: to this end, we replaced some occurrences of pronouns or key phrases *("he", "she", "the patient")* by sequences of randomly selected first name and last name, and we also incorporated dates, and either full addresses (hospital name, street name, post code, city name) or basically only city names within sentences. The final corpus is composed of 2279 entities.[20] and

[12] https://gitlab.com/MAPA-EU-Project/mapa_project

[13] https://elrc-share.eu/repository/search/?q=MAPA

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

[15] https://eur-lex.europa.eu/

[16] Although Irish became a full EU official language on 1 January 2007, EU institutions have profited from temporary derogations not to produce all acts in Irish on a transitional basis.

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

[18] https://curia.europa.eu/

[19] For experiments on the Spanish legal and EUR-LEX data the reader can check (de Gibert et al., 2022).

[20] The final corpus is composed of 1019 person names, 539 dates, 524 ages, 62 organisations, 60 city names, 43 country names, 15 zip codes, 8 profession names, 7 addresses, and 2 places, for 485 files.

it has been manually revised to correct any wrong automatic replacement. This corpus has then been translated into the other 23 languages to be used in order to train the MAPA system in all languages (cf. Section 4.4). Even if clinical cases are not patient reports, they are composed of words belonging to the clinical domain.

4.3. Legal Domain Datasets

Obtaining legal text that could be annotated and used for system development and evaluation also proved to be a challenge. Some countries provide their court decisions as public data that can be used for language processing (this was the case for the Italian *Corte di Cassazione* [21]), but this is not the case for many Courts of Justice or Supreme Courts. As a consequence, we ventured into a task of de-anonymising to help us create and annotate legal data from the French *Cour de Cassation* and from the Greek Supreme Court *Areiospagos* [22]. The task consisted in identifying the elements that needed de-anonymising within the text (e.g., *à l'égard de M. X...*, and *de Mme X... Viviane, épouse Y..., actuellement hospitalisée à l'unité [...]-[...]*) and replacing them with the same type of entities from some available lists (containing person given names and family names, hospital names, addresses, etc.), and then, checking them to avoid any remaining wrong replacements. Once the data were de-anonymised, they were annotated with NEs and prepared for training, development and evaluation. Around 2,000 sentences were completed per language.

In addition to these court decision data, we also collected legal data from the Spanish Ministry of Justice, Court Cases from Maltese jurisprudence and we annotated some legal-administrative data from the ELRC-SHARE repository.

4.4. Synthetic Data

In order to increase the size of our annotated corpora, synthetic data were produced. All already annotated data was MT translated by exporting annotation first, translating raw text and then re-inserting annotation into the translated output. A pipeline was developed for that purpose and although it presented some drawbacks in terms of tag export and noise from MT, results were very interesting (cf. Section 5) and it is a path worth pursuing by tackling the detected shortcomings.

4.5. Lists of Person Names and Surnames

In addition to the annotated and raw corpora produced, lists of ∼10,000 person names were built per language/country. These comprised commonly used names in the searched country, sometimes being of foreign origin too. In order to do so, first names and surnames were collected from different sources following availability. This was country and language dependent as some countries provide lists in their institutes of statistics or similar organisms and they are very helpful and willing to disseminate and share them. Generally, first names and surnames were combined to provide lists with full forms. However, whenever linguistic constraints were imposed to perform these combinations (e.g., for Baltic languages and Irish) separate lists have been compiled for given names and family names.

5. Toolkit Evaluation

The goal of the evaluation performed during MAPA was to evaluate the quality of the entity detection for de-identification.

5.1. Corpus

The MAPA data production activity collected and annotated data, which was split into training, development and test data. These data were used with Version 2 of the MAPA system to produce the evaluation results described here. The MAPA Consortium members first prepared annotated documents in their native languages. Training and testing was then performed on data splits of the "native language corpus" (e.g., Figure 3 shows the results for French medical data). In addition, this corpus was machine-translated to all the other languages addressed by the project (Figures 4 & 5 for medical translated and legal translated data, respectively). The goal was to examine whether the additional data produced this way would be suitable for training and testing the MAPA de-identification system.

5.2. Measures

Various considerations need to be taken into account with respect to the aim of MAPA, which is to provide a de-identification functionality for text documents:

- **Behaviour on whole text** (accuracy) vs **signal detection** (true positives). Accuracy computes the rate of agreement on every input word. Signal detection focuses on the correct detection of target entities (true positives). While accuracy is a convenient metric for use-agnostic, global system behavior, signal detection is more closely related to the intended use of the system.
- **Granularity: word-level** (correct entity type) vs **entity-level** (the detection of boundaries). Word-level evaluation focuses on the correct prediction of whether or not a word is part of an entity, and of which type. Its unit of measurement is the word. Entity-level evaluation additionally aims at determining the correct boundaries of each entity. Its unit of measurement is the entity (that can encompass several words). Entity-level detection, with both types and boundaries, is relevant when post-detection processing depends on the recognition of well-formed, full entities, while word-level evaluation is for other cases.

[21] https://www.cortedicassazione.it/corte-di-cassazione/

[22] http://www.areiospagos.gr/

- **Information Retrieval** (precision, recall, F1-score) vs **Test** (sensitivity, specificity). The main information retrieval measures (precision, recall, and F1-score) used for NER focus on the signal detection task. They measure the rate of correct detection (true positives) against system detection (precision) or against gold standard annotations (recall), and can be summarized with F1-score (the harmonic mean of precision and recall). Test measures (sensitivity and specificity) are typically used to interpret diagnostic tests in medicine. Here, sensitivity (equal to recall) is the probability that an entity in the gold standard is correctly detected by the system. Specificity is the probability that a non-entity according to the gold standard is correctly ignored (not detected) by the system. If sensitivity and specificity can be measured in a continuous way, they can be summarized by the area under the receiver operating characteristic (ROC) curve (AUC). Recall (or sensitivity) relates to the rate of de-identification and is thus important in the present context. Specificity relates to the preservation of information carried by the text beyond directly identifying entities, and is therefore a useful complementary metric.

5.3. Uniform Weights vs Different Weights

Uniform weights (plain named entity recognition) vs **different weights** (related to identifying power) are used for contrasting entity types, for balancing recall vs precision or sensitivity vs specificity. Some entity types (e.g., person name) have higher identifying power than others (e.g., age), it may be relevant to give them a higher weight. Consequently, a high recall (detecting as many identifying entities as possible) is more important here than a high precision (having the highest possible proportion of actually identifying entities among those predicted as identifying) because the consequences of a miss (false negative) have much more impact than removing a general word. Similarly, a high sensitivity (= recall: detecting as many identifying entities as possible) is more important than a high specificity (marking as non-identifying as many non-identifying words as possible) since identifying entities are always less frequent than general words. The importance of recall makes F2-score a more relevant alternative than the balanced F1-score.

5.4. Level of Importance of the Identification of Fine-Grained Entity Types

The MAPA annotation schema is a hierarchy with three levels of entity types. The lower levels define finer-grained types (e.g., *street* or *building* below *address*). Making the right distinction between lower-level entity types (e.g., *building* vs *street*) is less important than detecting the higher-level entity types (e.g., *address* vs *person*). In a more lenient evaluation mode, lower-level types may be converted into levels higher in the entity type hierarchy.

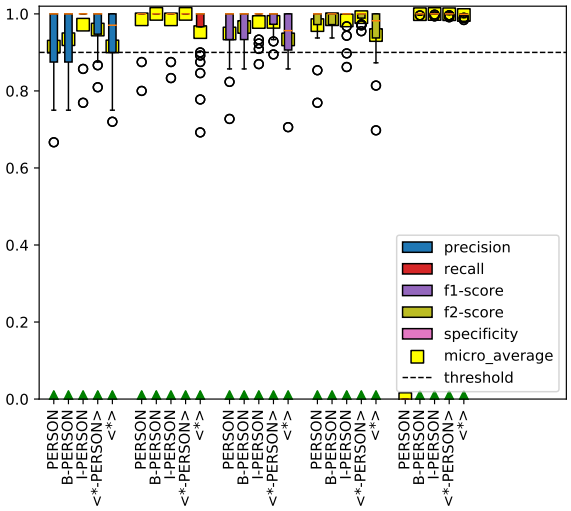


Figure 3: Distribution of precision, recall, F1-score, F2-score, and specificity of PERSON detection evaluation on native documents: French medical, at the entity level (PERSON) and at the word level (boundary-related word labels <B−PERSON> and <I−PERSON>, generic recognition of either of them <∗−PERSON>, word-level generic recognition of any type of level-1 entity <∗>), for each test document. Specificity is not computed for entity-level labels.

5.5. Distribution of Scores and Aggregation

The MAPA system was tested on a large set of documents of varied nature in different domains and languages. For a given metric, this results in a distribution of scores over individual documents. This distribution can be displayed or summarized in various ways. An average score can be computed globally based on individual results (micro-average) or after computing per-class results (macro-average). Weights can be applied as mentioned above. Because the various selected entity types have different levels of importance for de-identification, computing a plain macro-average is not necessarily optimal. Information on the distribution of values can be obtained through their standard deviation, quartiles, and more generally a histogram of values.

Based upon these considerations, Figures 3–5 show distributions of metrics across documents. The project targeted a threshold of 0.895 in general: a score above this threshold is signaled by a green, up-pointing arrow on the x axis. The figures show that most metrics were above threshold for most documents. The task was more difficult on translated documents and scores were accordingly lower, but recall and F2-score were still above or very close to threshold for most languages in both medical and legal documents.

5.6. Difficulty of Examples

While such scores provide a general idea of the behaviour of these systems, they ignore a key piece of in-

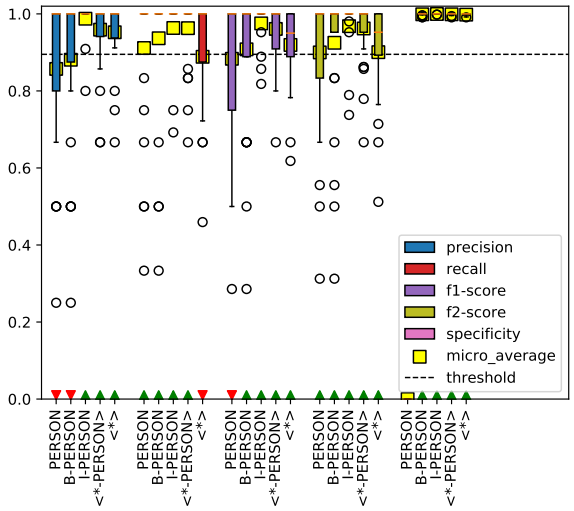


Figure 4: Distribution of precision, recall, F1-score, F2-score and specificity of PERSON detection in documents translated into Bulgarian, with same information as in Fig 3.

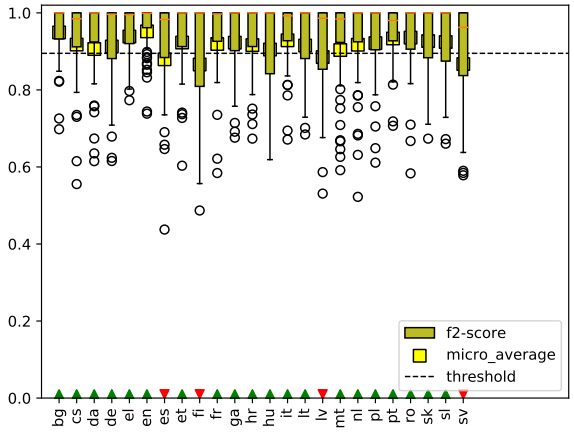


Figure 5: Distribution of F2-score of <*-PERSON> detection across documents for each language (translated documents).

formation that can be useful for assessing progress and discerning remaining challenges: the relative difficulty of test instances. To address this shortcoming, MAPA designed the notion of differential evaluation which effectively defines a pragmatic partition of instances into gradually more difficult bins by leveraging the predictions made by a set of systems. Comparing systems along these difficulty bins enables us to produce a finer-grained analysis of their relative merits. The methodology is described in (Gianola et al., 2021) with two illustrative examples: a multi-label text classification task (Névéol et al.,) and a comparison of neural models trained for biomedical entity detection (Wei et al., 2016).

6. Conclusions

In this paper we have presented the most relevant outcomes of the MAPA project involving datasets and technology. The project has produced corpora and tools that are available for the community:

- The tools software package produced is shared through Gitlab[23] under an Apache 2.0 licence [24].
- All annotated datasets, raw corpora and name lists produced within MAPA can be downloaded per language package from the ELRA Catalogue[25] and the ELRC-SHARE repository[26], and they will be shortly linked through the ELG catalogue[27].
- The annotation guidelines can be downloaded from the ELRA manual's library[28].

Regarding toolkit evaluation, this paper has focused on the evaluation of entity detection quality on a) a native language French medical data and b) datasets obtained through machine translated data (i.e., using synthetic data). Despite the shortcomings of noise inherited from the synthetic data creation process, both recall and F2-score on held-out data were above a 0.895 threshold for most documents in both legal and medical corpora.

7. Acknowledgements

This work was partially funded by the MAPA project. The MAPA project was an INEA-funded Action for the European Commission under the Connecting Europe Facility (CEF) – Telecommunications Sector with Grant Agreement No INEA/CEF/ICT/A2019/1927065.
The authors wish to acknowledge all MAPA consortium members for their work and contributions.
We would also like to thank (a) the *Cour de Cassation* and their anonymisation team for their fruitful discussions, and (b) the INCEpTION project group for their help with the annotation platform.
Finally, a very big thank you to our colleagues from DCU (Jane Dunne, Teresa Lynn, and Jane O'Connor) and from ILC-CNR (Monica Monachini, Claudia Soria and Tommaso Lo Sterzo) for their very valuable help with the Irish data and the Italian names, respectively.

[23] https://gitlab.com/MAPA-EU-Project/mapa_project
[24] http://www.apache.org/licenses/LICENSE-2.0
[25] http://www.elra.info/en/
[26] https://elrc-share.eu/repository/search/?q=MAPA
[27] https://live.european-language-grid.eu/
[28] http://www.elra.info/media/filer_public/2022/05/10/mapa_annotation-guidelines-v6.pdf

8. Bibliographical References

Ajausks, Ē., Arranz, V., Bié, L., Cerdà-i Cucó, A., Choukri, K., Cuadros, M., Degroote, H., Estela, A., Etchegoyhen, T., García-Martínez, M., García-Pablos, A., Herranz, M., Kohan, A., Melero, M., Rosner, M., Rozis, R., Paroubek, P., Vasiļevskis, A., and Zweigenbaum, P. (2020). The Multilingual Anonymisation Toolkit for Public Administrations (MAPA) Project. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 471–472, Lisboa, Portugal, November. European Association for Machine Translation.

Cañete, J., Chaperon, G., Fuentes, R., and Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. In *Proceedings of the Practical ML for Developing Countries Workshop at the Eighth International Conference on Learning Representations (ICLR 2020)*, pages 1–9.

de Gibert, O., García-Pablos, A., Cuadros, M., and Melero, M. (2022). Spanish datasets for sensitive entity detection in the legal domain. In Nicoletta Calzolari, et al., editors, *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22)*, Marseille, France, june. European Language Resource Association (ELRA).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Gianola, L., Ēriks Ajausks, Arranz, V., Bendahman, C., Bié, L., Borg, C., Cerdà, A., Choukri, K., Cuadros, M., Gibert, O. D., Degroote, H., Edelman, E., Etchegoyhen, T., Ángela Franco Torres, Hernandez, M. G., Pablos, A. G., Gatt, A., Grouin, C., Herranz, M., Kohan, A. A., Lavergne, T., Melero, M., Paroubek, P., Rigault, M., Rosner, M., Rozis, R., Plas, L. V. D., Vīksna, R., and Zweigenbaum, P., (2020). *Legal Knowledge and Information Systems*, volume 334 of *Frontiers in Artificial Intelligence and Applications*, chapter Automatic Removal of Identifying Information in Official EU Languages for Public Administrations: The MAPA Project, pages 223–226. IOS Press. DOI 10.3233/FAIA200869.

Gianola, L., El Boukkouri, H., Grouin, C., Lavergne, T., Paroubek, P., and Zweigenbaum, P. (2021). Differential Evaluation: a Qualitative Analysis of Natural Language Processing System Behavior Based Upon Data Resistance to Processing. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 1–10, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Huang, J., Li, C., Subudhi, K., Jose, D., Balakrishnan, S., Chen, W., Peng, B., Gao, J., and Han, J. (2021). Few-Shot Named Entity Recognition: An Empirical Baseline Study. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10423, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., and Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico, August. Association for Computational Linguistics.

Yadav, V. and Bethard, S. (2018). A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

9. Language Resource References

Grabar, N., Claveau, V., and Dalloux, C. (2018). CAS: French Corpus with Clinical Cases. In *Proc of LOUHI*, Brussels, Belgium.

Grabar, N., Grouin, C., Hamon, T., and Claveau, V. (2019). Recherche et Extraction d'Information dans des Cas Cliniques. Présentation de la Campagne d'Évaluation DEFT 2019. In *Actes de DEFT*, Toulouse, France.

Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikán, L., Ramadier, L., Rey, G., , and Zweigenbaum, P.). CLEF eHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian.

Wei, C.-H., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., Wiegers, T. C., and Lu, Z. (2016). Assessing the State of the Art in Biomedical Relation Extraction: Overview of the BioCreative V Chemical-Disease Relation (CDR) Task. *Database: The Journal of Biological Databases and Curation*, PMCID:PMC4799720. doi:10.1093/database/baw032.